

1. INTRODUCTION

1.1 Definition of Statistics

Statistics can be seen in a **plural** and **singular** sense.

In a **plural sense**, statistics refers to statistical data, which is the numerical description of quantitative aspect of things.

In a **singular sense**, statistics refers to statistical methods or a subject of statistics. This is the focus of this course!

Statistics is a branch of applied mathematics concerned with the collection and interpretation of quantitative data and the use of probability theory to estimate population parameters. It is a science that helps us make better decisions in business and economics as well as in other fields. **Statistics** also teaches us how to summarize, analyze, and draw meaningful inferences/conclusions from data that then lead to improve our decisions. In short, statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

1.2 Classification of Statistics

There are two categories of statistics: **descriptive statistics and inferential statistics**

Descriptive statistics deals with data collection, summarization, analysis and interpretation.

Descriptive statistics can only be used to describe the sample data. That is, the results cannot be generalized to the population (no conclusion or inference about population characteristics). In other words, we use descriptive statistics simply to describe what is going on in our sample data we have at hand. It is making aggregation from the sample data but not generalization about the population. **Frequency distributions, graphs (like bar histogram, pie, line, etc.), measures of central tendencies (e.g. mean, median and mode), measures of variations (standard deviation, range, CV) are examples of descriptive statistics.**

Example: suppose that a sample of marks of 6 students were 45, 60, 72, 80, 85 and 93. “The average score of the six students is 72.5” or “The range of the six students is 48” are descriptive statistics.

- **Descriptive Statistics**

- ✓ **Collect**
- ✓ **Organize**
- ✓ **Summarize**
- ✓ **Display**
- ✓ **Analyze**
- ✓ **Interpret**

Inferential statistics it makes conclusion or inference about population parameter based on sample statistics. That is, it makes generalization about the population using sample information. In other words, we can take the results of an analysis using a sample and can generalize it to the larger population that the sample represents. Note that the sample must be the representative of the population about which we make conclusion. **Regression analyses, correlation analyses, ANOVA, etc. are examples of inferential statistics.**

Inferential statistics involves:

1. **estimation** of the unknown population **parameters**
2. **Testing hypotheses** about values of population parameters
3. **Decision-making**

What is the difference between Descriptive and Inferential Statistics?

- Descriptive statistics is focused on summarizing the data collected from a sample. The technique produces measures of central tendency and dispersion which represent how the values of the variables are concentrated and dispersed.
- Inferential statistics generalizes the statistics obtained from a sample to the general population to which the sample belongs. The measures of the population are termed as parameters.
- Descriptive statistics make only summarization of the properties of the sample from which data were acquired, but in inferential statistics, the measure from the sample is used to infer properties of the population.
- In inferential statistics, the parameters were obtained from a sample, but not the whole population; therefore, always some uncertainty exists compared to the real values.

1.3 Applications of Statistics

In the fields of economics it is almost impossible to think of a problem which does not require an extensive use of statistical methods and statistical data. For instance,

- At macro levels: government uses statistics to formulate sound policies (monetary, fiscal, etc.), depicting the trends of its GDP using graphs, and for planning purposes, which needs statistics for assessment of various resources both human and material of the country in order to make proper planning
- At sectoral levels: describing growth in agricultural and industrial sectors, depicting trends in fertilizers and improved seeds consumption over the last 15 years
- At micro levels: statistics is required to describe, compare and correlate about production, distribution and consumption, trend and fluctuation analysis prices, sales and advertising using time series data
- At subject level, econometrics applies statistical methods to the empirical study of economic theories and relationships

Thus, statistics is indispensable in economic analysis.

In other sciences, application of statistics is vital:

- Physical science. e.g.: Meteorology (for predicting weather condition)
- Natural science. e.g.: -Botany (the effect of temperature on the health of plants).

- Social science e.g.: Psychology (human characteristics)
- Medical science. e.g.: -the effectiveness of a new drug.
- Pharmaceuticals. e.g.: the percentage composition of a certain drug.

Limitations:

- Statistics is not suitable to study qualitative phenomenon like beauty, good, bad, honesty, courage, intelligence, etc.
- It doesn't study individual item.
- Its laws are not as perfect as laws in natural sciences.
- Its results are true only on average.
- Its interpretation requires a high degree of skill.
- Mayn't necessarily bring out the cause and effect relationships between the variables.
- Its methods are biased.
- Possibilities of misuses; like using a data with no source, defective data, unrepresentative data, inadequate data, unfair comparison, and mistakes in arithmetic.

2. SAMPLING THEORY

2.1. Basic Concepts ((Population, Universe, Census, Sample, Parameter, Statistic, Sampling frame, Sampling Unit, Sampling error, Sample size...))

Population (universe): is a set of large number of objects or the total number of conceivable observation of any kind or possessing some common specified characteristics that a researcher is interested in to investigate. The population may contain living or non-living things. The entire lot of anything under study is called population. For instance, in identifying factors that affect performance of first year students of College of business and economics, our population is the entire first year students of the college. Types of population are:

- a) **Finite vs. infinite population:** if it is possible to count all individuals in the population, then it is finite population. However, if the elements are infinite and uncountable, then it is infinite.
- b) **Real vs. hypothetical population:** real population exists in reality but hypothetical population does not exist (or it is simply imaged).
- c) **Target vs. studied population:** The target population is the entire group a researcher is interested in or the group about which the researcher wishes to draw conclusions. However, not all target population are willing to be investigated. The population out of which the sample is selected is called or studied population (population that is willing to be investigated).

Sample is a set of measurement taken from a process or series of experiments or a population indicating that sample is the subset of the population.

Population parameters: the descriptive characteristics (or summary measures) obtained from the entire population are termed as population parameters examples population mean and population variance.

Sample statistics are the summary measures obtained from the sample example sample mean and sample variance.

Census is the complete enumeration of all items in the population.

Sampling implies the selection of few items from a given population.

- **Sampling with replacement:** an item, after being picked up and included into the sample, is replaced back into the population so that it can be picked again.
- **Sampling without replacement:** if an item is picked up once, it is no more available to be picked up again.

Sample size- the sample size (denoted by n) is the number of elements included in the sample for investigation

Sampling frame is the source material or device from which a sample is drawn. It is a list of all those within a population who can be sampled, and may include individuals, households or institutions.

Sampling unit is one of the units into which an aggregate is divided for the purpose of sampling, each unit being regarded as individual and indivisible when the selection is made.

Sample design is the process of selecting sample elements from the sampling frame.

Sampling error is incurred when the statistical characteristics of a population are estimated from a subset, or sample, of that population. Since the sample does not include all members of the population, statistics on the sample, such as sample mean generally differs from population mean. That is, it occurs as we are working with the sample but not with the population and hence sample statistics differ from population parameters. In other words, the discrepancy between a sample statistic and its population parameter is called **sampling error**. It arises due to inappropriate sampling techniques.

Non-sampling error is a catch-all term for the deviations of estimates from their true values that are not a function of the sample chosen, including various: coverage errors ((non-representative sample), non-response errors (respondents give false information), data entry error, misunderstanding of questions being asked etc.

2.2.Reasons for Sampling

There are several reasons for sampling:

- (1) Costs/economy
- (2) Timeliness
- (3) Large size of many population
- (4) Inaccessibility of the entire population
- (5) Destructive nature of many tests
- (6) Reliability or accuracy.

(1) Cost/Economy

Unit cost of collecting data in the case of census is significantly less than in the case of sampling. However, due to the larger number of items in the population, the total cost involves in the case of is significantly higher than in the case of sampling. Suppose it takes Birr 200 per unit to make a census of 10,000,000 individuals but the unit cost of sampling 5000 individuals is Birr 1000. Thus, the total cost is: $10,000,000 \times 200 = 2,000,000,000$ but that of sample is $5,000 \times 1000 = 5,000,000$

(2) Timeliness

Due to the larger size of population total time involves in the case of census is significantly higher than that of sampling (i.e., the sample may provide us with necessary information quickly).

(3) Large population size

Some times, many populations about which inferences must be made are quite large implying that it is impossible to cover all the items in the population. Thus, the solution is to take sample from such a population.

(4) Inaccessibility of the entire population:

In some cases the entire population may not be accessible due to diseases, death, conflict, mental abnormality, prisoners, etc. In that case sampling is necessary.

(5) Destructive nature of many tests:

Due to destructive nature of many tests, the resources are completed to collect information only from part of the population. For example: blood test for a patient, life hours of a tube light, strength of wires, etc.

(6) Accuracy:

Non-sampling error in the case of census is higher than the non-sampling error committed in the case of a sample survey (as less qualified investigator are involve in the case of census and the supervision, monitoring and quality control mechanism in the case of census may be poor). The higher the degree of non-sampling error, the less reliable your result may be.

2.3. A Review of Methods of Sampling (Probability Vs Non-Probability Sampling Techniques)

There are two principal methods of drawing a sample from a population.

- A. Probability Sampling:** in this case each observation in the population has an equal chance of being selected to become part of the sample. Sampling techniques such as simple **random sampling**, **stratified sampling**, **cluster sampling** and **systematic sampling** are probability sampling.
- B. Non-probability Sampling** is **no way of estimating the probability** that each individual will be included in the sample. **Quota sampling** and **judgmental sampling** are examples of non-probability sampling.

A. Probability Sampling

- 1. Simple Random Sample:** is a method of probability sampling in which every unit in the population has an equal nonzero chance of being selected (or part of the sample). In other

words, each element of the population has an equal and independent chance of being included into the sample. The probability is given by n/N .

How to select a simple random sample?

There are two methods:

Lottery method- In this method, each population item is numbered 1 to N on slips of identical cards (size, shape and color). Then place numbered cards in a bowl, mix them thoroughly, and select as many cards as needed in a blind fold selection. The subjects whose numbers are selected constitute the sample. Since it is difficult to mix the cards thoroughly, there is a chance of obtaining a biased sample. Thus we need other method of selecting sample elements.

Random Number method- due to the problem of lottery method, statisticians use another method known as the **random number method** where numbers are generated using computers or they are available on the annex of text books. Suppose the two-digit number table is given as follows:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	37	75	10	49	98	66	03	86	34	80	98	44	22	22	45	83	53	86	23	51
2	50	91	56	41	52	82	98	11	57	96	27	10	27	16	35	34	47	01	36	08
3	99	14	23	50	21	01	03	25	79	07	80	54	55	41	12	15	15	03	68	56
4	70	72	01	00	33	25	19	16	23	58	03	78	47	43	77	88	15	02	55	67
5	18	46	06	49	47	32	58	08	75	29	63	66	89	09	22	35	97	74	30	80
6	65	76	34	11	33	60	95	03	53	72	06	78	28	14	51	78	76	45	26	45
7	83	76	95	25	70	60	13	32	52	11	87	38	49	01	82	84	99	02	64	00
8	58	90	07	84	20	98	57	93	36	65	10	71	83	93	42	46	34	61	44	01

How to Use Random Number Tables

- a) Assign a unique number to each population element in the **sampling frame**. Start with serial number 1, or 01, or 001, etc. depending on the number of digits required.
- b) Choose a random starting position by closing your eyes (blind fold selection) and placing your finger on a number in the table.
- c) Select serial numbers across rows or down columns or diagonally from the starting point.
- d) Discard numbers that are not assigned to any population element and ignore numbers that have already been selected.
- e) Repeat the selection process until the required number of sample elements is selected.

Example

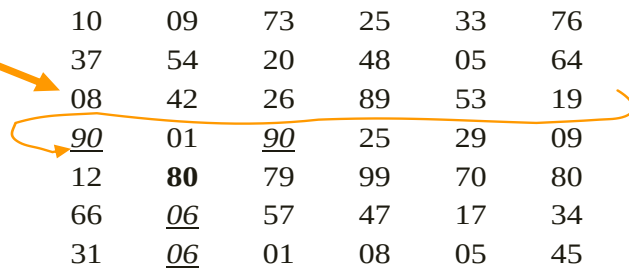
A lecturer wants to randomly select 20 students from a class of 100 students. Here is how he can do it using a random number table.

Step 1: Assign all the 100 members of the population a unique number. You may identify each element by assigning a two-digit number. Assign 01 to the first name on the list, and 00 to the last name. If this is done, then the task of selecting the sample will be easier as you would be able to use a 2-digit random number table.

NAME	NUMBER	NAME	NUMBER
Amare	01	Temesgen	42
			...
Chala	08	Tsion.....	61
	...		...
Genet	18	Tura	87
	...		...
Lelise	26		...
	...	Zeelam	99
Nuru	30	Zinesh	00

Step 2: Select any starting point in the **Random Number Table** and find the first number that corresponds to a number on the list of your population. In the example above, # 08 has been chosen as the starting point and the first student chosen is Chala.

Starting point:
move right to the end of the row, then down to the next row; move left to the end, then down to the next row, and so on.



10	09	73	25	33	76
37	54	20	48	05	64
08	42	26	89	53	19
<u>90</u>	01	<u>90</u>	25	29	09
12	80	79	99	70	80
66	<u>06</u>	57	47	17	34
31	<u>06</u>	01	08	05	45

Step 3: Move to the next number, 42 and select the person (Temesgen) corresponding to that number into the sample

Step 4: Continue to the next number that qualifies and select that person into the sample.

26 -- Lelise, followed by #89, #53 and #19

Step 5: After you have selected the student # 19, go to the next line and choose #90. Continue in the same manner until the full sample is selected. If you encounter a number selected earlier (e.g., 90, 06 in this example) simply skip over it and choose the next number.

Cost and feasibility can be problems, especially if the population is large.

2. Stratified sampling it is a three-step process:

- Step 1- Divide the population into **homogeneous**, mutually exclusive and collectively exhaustive **groups** or **strata** using some stratification variable (e.g. income level, sex, education level, etc.);
- Step 2- Select an independent simple random sample from each stratum (using simple random sample);
- Step 3- Form the final sample by consolidating all sample elements chosen in step 2.
- May yield smaller standard errors of estimators than does the simple random sampling. Thus precision can be gained with smaller sample sizes.
- Note that it is applied if the population is heterogeneous in some attributes

Stratified samples can be:

- **Proportionate:** involving the selection of sample elements from each stratum, such that the ratio of sample elements total number of population elements (n/N) is constant/equal for all strata.
- **Disproportionate:** the sample is disproportionate when the above mentioned ratio is unequal.

Example: To select a proportionate stratified sample of 20 households from Addis Ababa that belong to three income groups: low (50), middle (30) and high (20) ($N=50+30+20=100$).

- Sub-divide the club members into three homogeneous sub-groups or strata by the income groups: low, middle and high.
- Calculate the overall sampling fraction, f , in the following manner: $f=n/N=20/100=0.2$
Where n = sample size and N = population size: $n_1=0.2*50=10$, $n_2=0.2*30=6$ and $n_3=0.2*20=4$. Thus, $n=n_1+n_2+n_3=10+6+4=20$

3. Systematic Sampling:

In systematic sampling only one random number is needed throughout the entire sampling process. To use systematic sampling, a researcher needs:

- [i] a sampling frame of the population;
- [ii] a skip interval (K) calculated as follows:

$$\text{Skip interval } (k) = \frac{\text{population list size } (N)}{\text{Sample size } (n)} \text{ or } \frac{N}{n} = K$$

The first element (number), which is between 1 and K , is determined using simple random sampling and then the next items are selected using the skip interval. For instance, The j^{th} unit is selected at first and then $(j+k)^{\text{th}}$, $(j+2K)^{\text{th}}$, ..., etc until the required sample size is obtained.

Example: if a **lecturer wants to randomly select 20 students from a class of 100 students using systematic sampling**, she can take the first element between 1 and 5 using simple random sampling and then select every 5th element starting from the initial number.

4. Cluster Sampling

It is a type of sampling in which is divided into non-overlapping **heterogeneous** groups called clusters or groups and clusters/groups of elements are sampled as the sampling units using sample random sampling technique in the first phase (if it is the two-phase cluster sampling). In other words, cluster sampling is a type of sampling which involves dividing the population into groups (or clusters). Then, one or more clusters are chosen at random and individual within the chosen cluster is sampled.

A two-step-process:

- **Step 1-** Defined population is divided into number of mutually exclusive and collectively exhaustive heterogonous groups or clusters;
- **Step 2-** Select an independent simple random sample of clusters using sample random sampling.
- One special type of cluster sampling is called area sampling, where pieces of geographical areas are selected.
- A common form of cluster sampling where clusters consist of geographic areas, such as districts, housing blocks or townships. Area sampling could be one-stage, two-stage, or multi-stage.

How to Take an Area Sample Using Subdivisions

Step 1: Determine the geographic area to be surveyed, and identify its subdivisions. Each subdivision cluster should be highly similar to all others. For example, choose ten housing blocks within 2 kilometers of the proposed site [say, Model Town] for your new retail outlet; assign each a number.

Step 2: Decide on the use of one-step or two-step cluster sampling. Assume that you decide to use a two-stage cluster sampling.

Step 3: Using random numbers, select the housing blocks to be sampled. Here, you select 4 blocks randomly, say numbers #102, #104, #106, and #108.

Step 4: Using some probability method of sample selection, select the households in each of the chosen housing block to be included in the sample. Identify a random starting point (say, apartment no. 103).

Stratified Sampling vs Cluster Sampling

Stratified Sampling	Cluster Sampling
1. The target population is sub-divided into a few subgroups or strata, each containing a large number of elements. 2. Within each stratum, the elements are homogeneous. However, high degree of heterogeneity exists between strata. 3. A sample element is selected each time 4. Individuals are sampled from all strata 5. Less sampling error. 6. Objective is to increase precision	1. The target population is sub-divided into a large number of sub-population or clusters, each containing a few elements. 2. Within each cluster, the elements are heterogeneous. Between clusters, there is a high degree of homogeneity. 3. A cluster is selected each time 4. Individuals are sampled from selected clusters 5. More prone to sampling error. 6. Objective is to increase sampling efficiency by reducing costs.

5. Multi-stage and multi-phase sampling (optional)

In multi-stage sampling: several levels of nested clusters are involved where sample units are clusters at each stage except the final stage. It is known as 'multistage' because there are multiple stages, or steps, to creating the sample. The first stage in multistage sampling is the same as cluster sampling. Thus, it is a complex form of cluster sampling. It often includes both stratified and cluster sampling techniques. E.g. cluster Ethiopia into regions, zones, Woredas, kebeles and finally take households from sampled kebeles using SRS or stratified sampling.

Multi-phase sampling: is designed to make use of the information collected in one phase to develop sampling design for the next phase. For instance, in the double phase sampling, the first phase may consider relationship between income and expenditure and using information obtained in the first phase, surveyed households divided into groups based on income levels (strata). Or it is sometimes convenient and economical to collect certain items of information from the whole of the units of a sample and other items of usually more detailed information from a sub-sample of the units constituting the original sample. This may be termed two-phase sampling, e.g. if the

collection of information concerning variate, y , is relatively expensive, and there exists some other variate, x , correlated with it, which is relatively cheap to investigate, it may be profitable to carry out sampling in two phases.

B. Non-probability sampling: no way to assign nonzero equal chance

Convenience sampling

- Drawn at the convenience of the researcher. Common in exploratory research. Does not lead to any conclusion.

Judgmental sampling

- Sampling based on some judgment, gut-feelings or experience of the researcher. Common in commercial marketing research projects. If inference drawing is not necessary, these samples are quite useful.

Quota sampling

- In this method, the decision maker requires the sample to contain a certain number of items with a given characteristic. It is something like judgmental sampling. Individuals are selected from each quota using SRS.

Snowball sampling

- Used in studies involving respondents who are rare to find. To start with, the researcher compiles a short list of sample units from various sources. Each of these respondents is contacted to provide names of other probable respondents.

3. Data Collection and Presentation

3.1.Data types and sources

Definition of Data: data refer to the numerical description of quantitative aspect of things. In other words, data are the facts and figures that are collected, summarized, analyzed, and interpreted. The singular form of data is datum.

Types of data and data measurements

Data can be divided into quantitative and qualitative.

Qualitative data are data that can be placed into **distinct categories**, according to some characteristics or attributes. Qualitative data could be nominal or ordinal. Examples of qualitative data are gender (Male and Female), income categories (low, middle, high), education level (diploma, degree, masters, PhD), etc.

Quantitative data are numerical and can be ordered or ranked and they are closure under mathematical operations. Examples of quantitative data are age, height, weight, income, expenditure, profit, price, household size, inflation rate, interest rate, etc. Quantitative data can be **discrete** or **continuous**. **Discrete data** assume values that can be counted or within a given interval they assume some fixed values. Examples are household size, number of customers, number of patients, etc. **Continuous data** can assume an infinite number of values between any two specific values. They often include fractions and decimals. Examples are income, profit, price, inflation rate, interest rate, etc.

Data can also be categorized based on type of **measurement scale**:

- **Nominal data:** classifies data into mutually exclusive (non-overlapping), collectively exhaustive categories in which no order or ranking can be imposed on the data. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories (possible values or attributes can be distinguished only by their names), but we cannot quantify or even rank/order those categories. Typical examples of nominal variables are gender, race, colour, city, etc.

- **Ordinal data:** classify data into categories that can be ranked/ordered; however, precise differences between the ranks do not exist. That is, allow us to rank/order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say "how much more." Examples quality of instructors: superior, average, poor; grades: A, B, C, D, F; quality: excellent, very good, good, poor; etc.
- **Interval data:** allow us not only to rank or order the items that are measured, but also to quantify and compare the sizes of differences between them. In other word, the **interval level of measurement** ranks data, and precise differences between units of measure do exist; however, there is **no meaningful zero (i.e., zero is arbitrarily located in the interval scale)**. For example, year of birth, temperature as measured in degrees Fahrenheit or Celsius, constitutes an interval scale. However, for temperature, 0°F does not mean no heat at all.
- **Ratio data:** possess all the properties of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population. Examples of ratio scales are those used to measure height, weight, area, and number of phone calls received. Ratio scales have differences between units (1 cm, 1 kg, etc.) and a true zero. In addition, the ratio scale contains a true ratio between values. For example, if one person can lift 100kg and another can lift 50 kg, then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person.

Note that most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

Summary

Measurement scale	Distinguished by	Ranked/ ordered	True differences exist	True ratio and true zero exists	Example
Nominal scale	Their names only	no	no	No	gender
Ordinal scale	Their names only	Yes	No	No	Grades
Interval scale	Their names only	Yes	Yes	No	Year of birth
Ratio scale	Their names only	yes	Yes	Yes	Age

Sources of data: primary and secondary sources

Primary data are directly obtained from the respondents or from the collecting organization like CSA (i.e., it is unprocessed or raw data)

Secondary data: are processed that that are obtained from the secondary sources including various organizations, books, reports, journals etc.

3.2.Methods of data collection (from respondents)

A number of methods can be used to gather data from the so called 'respondents'. Some of the methods of data collection are:

i) Direct personal interview-as the name indicates directly ask the respondents

One advantage is obtaining in-depth responses to questions from the person being interviewed.

Disadvantage: costly to train enumerators and collect data

ii) Mail questionnaire: data obtained via mail survey

Advantages: cover a wider geographic area than telephone surveys or personal interviews since mailed questionnaire surveys are less expensive to conduct. Also, respondents can remain anonymous if they desire. Disadvantages:

- Lack of mail address
- Misunderstanding of questions
- Less response rates

- Difficulty in reading and writing (literacy related problems)

iii) Telephone interviews

An advantage over personal interview surveys is that they are less costly. Also, people may be more frank in their opinions since there is no face-to-face contact. A major drawback to the telephone survey is that some people in the population will not have phones or will not answer when the calls are made; hence, not all people have a chance of being surveyed.

vi) E-mail interviews : questionnaires are sent to respondents and they give responses.

Advantages: costs are nearly zero

Other methods of data collection: **direct personal observation**, **Focus Group Discussions** (FGDs) and **Key Informant Interview** (KII)-asking knowledgeable person(s). However, they are not used in scientific research.

A method of data collection depends on:

- Nature and scope of the problem
- Cost, time and resources
- Degree of accuracy desired

3.3. Tabular methods of data presentation and frequency distribution

Once data are generated from representative sample using appropriate data collection tools or obtained from secondary sources, they must be presented in a meaningful manner. Presenting data in tabular form has some advantages such as simplicity, save time and space, ease to handle, analyze and interpret and facilitate comparison.

3.3.1. Tabular method of data presentation for ungrouped data: frequency distribution

Nominal data: Here the construction is straight forward: count the occurrences in each category and find the totals. **Example:** The marital status of 60 adults classified as single, married, divorced and widowed is given below:

Marital status	Single	Married	Divorced	Widowed	Total
Frequency	25	20	8	7	60

Ordinal data: The construction is identical to the nominal case. However, the categories should be put in an ordered manner. **Example:** Satisfaction on teaching method in a class size of 40.

Satisfaction	Very Satisfied	Satisfied	Dissatisfied	Very dissatisfied	Total
Frequency	11	24	3	2	40

Ratio data: suppose the raw data, which are in original form before statistical techniques are applied, for the heights of 33 students are given as follows (in cm):

154	160	154	160	152	162	164	170	166
168	154	160	162	158	156	164	152	
168	160	150	164	154	158	160	160	
160	152	156	168	152	154	164	154	

What is the height of the tallest student?

What is the height of the shortest student?

Which is the most frequent height?

It is not ease to answer these and related questions from the above raw data. Thus, it must be presented in such a way that it is possible to make meaningful analysis and interpretation from such dataset.

Definitions:

- i. An arrangement of numerical data in an ascending or descending order is known as an **array**.
- ii. The number of times an item or a particular observation occurs/repeats itself is known as **frequency**.
- iii. **Frequency distribution:** let f_i be frequency of X_i (where X_i is value of each case/observation). Then the value of X_i associated with the corresponding values of f_i is

known as frequency distribution. That is, **frequency distribution** associates each value of X_i to its corresponding frequency (f_i).

- iv. The **relative frequency** is the ratio of frequency of a case to the total number of observation (sample size = n). That is, $rf = \frac{f}{n}$ or $rf = \frac{f}{\sum_{i=1}^n f_i}$ where f is the number of

times a given element repeats itself (absolute frequency) and n is the # of obs.

Examples:

- v. **Percentage frequency**-when the relative frequency is expressed in terms of percentage.

$$\Rightarrow \text{Percentage freq} = \text{rel. freq.} \times 100\%$$

(Absolute) Frequency Distribution

i	Value of x (xi)	Absolute Frequency (fi)
1	150	1
2	152	4
3	154	6
4	156	2
5	158	2
6	160	7
7	162	2
8	164	4
9	166	1
10	168	3
11	170	1

Answers to the above questions:

- The height of the tallest student is 170cm
- The height of the shortest student is 150cm
- The most frequent height is 160cm

Relative Frequency Distribution

i	Value of x(xi)	Relative Frequency (rf)
1	150	1/33
2	152	4/33
3	154	6/33
4	156	2/33
5	158	2/33
6	160	7/33
7	162	2/33
8	164	4/33
9	166	1/33
10	168	3/33
11	170	1/33

3.3.2. Tabular method of data presentation for grouped data: frequency distribution

Note, however, that when the number of possible values of our variable is very large the discrete frequency distribution will no more be a condensed presentation. Then data have to be handled as continuous and distributed in to classes.

Consider the following table:

Class limit	Class boundary	frequency
24-30	23.5-30.5	3
31-37	30.5-37.5	1

38-44	37.5-44.5	5
45-51	44.5-51.5	9
52-58	51.5-58.5	6
59-65	58.5-65.5	1

Second example is on the costs of spare parts for certain consumers.

n Relative Frequency and Percent Frequency Distributions

	Relative	Percent
<u>Cost (\$)</u>	<u>Frequency</u>	<u>Frequency</u>
50-59	.04	4
60-69	.26	26

Insights Gained from the Percent Frequency Distribution

- Only 4% of the parts costs are in the \$50-59 class.
- 30% of the parts costs are under \$70.
- The greatest percentage (32% or almost one-third) of the parts costs are in the \$70-79 class.
- 10% of the parts costs are \$100 or more.

Some concepts and terminologies

- **Class Frequency (or simply frequency):** refers to the number of items that belong to a class or number of observations in a particular class.

- **Class limits (C.L.):** the lowest and highest values that can be included in a class such that there is gap between successive classes are **called class limits**. The lower class limit (L.C.L.) of a class is a value such that no lower value can fall in to that class, whereas the upper class limit (U.C.L.) of a class is a value such that no upper value can fall into that class.
- **Class Boundary (C.B) or Real class limits:** class boundaries are the lowest and the highest values in each class when there is no gap between successive classes. To work with the distribution of a variable as if it was continuous, we make use of these real class limits (also known as class boundaries).

Finding class boundaries

Let $d = \text{LCL of a class} - \text{UCL of preceding/previous class}$. Add half of this difference to all upper class limits to get the upper class boundary (UCB), and subtract it from all lower class limits to get the lower class boundary (LCB). That is,

$$\text{UCB}_i = \text{UCL}_i + \frac{1}{2}d \quad \text{and} \quad \text{LCB}_i = \text{LCL}_i - \frac{1}{2}d$$

For instance, $31 - 30 = 1$, and $\frac{1}{2}d = 0.5$. For the first class,

$\text{UCB} = \text{UCL} + 0.5 = 30 + 0.5 = 30.5$, and $\text{LCB} = \text{LCL} - 0.5 = 24 - 0.5 = 23.5$; continuing in such a way, we get the class boundaries in the second column of the example in the above table.

- **Class width (w) or class interval**

The **class width** for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class. For example, the class width of the above example is seven (7). That is, $31 - 24 = 7$ or $37 - 30 = 7$.

- **Class Mark (C.M.) or class midpoint**

The class mark is the mid-point of the class interval or is a value which lies mid way between the lower and upper limits of the class. It is obtained as:

$$CM = \frac{LCL + UCL}{2}, \text{ or } CM = \frac{LCB + UCB}{2}.$$

Note that: In further analysis of the data (measures of central tendencies, measures of variations, etc.), a CM is used to represent all the items in that class.

- **Number of class** is the possible number of groups of data.
- **Inclusive method:** according inclusive method the upper class limit of one class is included into that class itself.
- **Exclusive method** (continuous case): according this method the upper class limit of one class is included into the lower class limit of the next class.

Practical steps in constructing continuous frequency distribution:

1. Decide the number of classes (k): Select the number of classes desired, usually between 5 and 20 or use Sturge's rule $K = 1 + 3.322 \log n$ where "*k*" is number of classes desired and *n* is total number of observation. **For example;** If $n = 10$, $k = 4.32 \approx 4$; if $n = 100$, $k = 7.644 \approx 8$; if $n = 1000$, $k = 10.96 \approx 11$.

2. Compute the Range(R) = Maximum value- Minimum value

3. Determine the Class Width (w): If the number of classes is known and if it is decided to use a uniform class width, we use $w \approx \frac{\text{Range}}{k}$ and rounded up to the nearest integer, where Range is the difference between the highest and the smallest value of the data.

Note: As far as possible, a class width of 5 or a multiple of 5 is convenient and facilitates computations.

4. Determine the Class Limits: Pick a suitable starting point less than or equal to the minimum value. The starting point is called the **lower limit** of the first class. Continue to add the class width to this lower limit to get the rest of the lower limits.

5. Determine the Upper class limits: To find the upper limit of the first class, subtract 1 (one) from the lower limit of the second class. Then continue to add the class width to this upper limit to find the rest of the upper limits.

6. Determine the frequency of each class

Frequency of each class can be determined simply by counting the number of observations belonging to each class.

7. Sum up the frequency of each class to check whether it is equal to the total number of data collected from the field or not.

Example: Construct a continuous frequency distributions for the following raw data on marks (out of 100) obtained by 50 students in Statistics.

57, 53, 65, 55, 50, 45, 64, 52, 16, 46, 42, 63, 33, 64, 53, 25, 54, 35, 48, 55, 70, 47, 39, 58, 52, 36, 65, 75, 26, 20, 55, 60, 83, 61, 45, 63, 49, 42, 35, 18, 51, 45, 42, 65, 39, 59, 45, 41, 30, 40.

Solution:

i. Since $n = 50$, using the Sturge's rule, the **number of classes** is:

$k = 1 + 3.322 \log 50 = 6.64 \approx 7$. Thus, the number of class is 7.

ii. Range = highest value – lowest value = $83 - 16 = 67$.

iii. Class width = $w = \frac{\text{Range}}{k} = \frac{67}{7} = 9.57 \approx 10$.

iv. Since the smallest value is 16, the LCL_1 can be 15 and the UCL_1 should be 24; and the frequency distribution would look like:

Marks	frequency
15 - 24	3
25 – 34	4
35 – 44	10
45 -54	15
55 – 64	12
65 – 74	4
75 – 84	2
Total	50

Note: 1. For the class boundaries, see the nature of the data.

- If there is no decimal point, then $d = 1$.
- If there is one digit after the decimal, then $d = 0.1$.
- If there are two digits after the decimal, then $d = 0.01$.

Cumulative Frequency Distribution

The cumulative frequency of a class tells us how often the values fall below or above that class. Or as the name indicates it cumulates frequencies starting at the lowest or the highest class boundary. There are two types of cumulative frequency distributions: the “less than” and the “more than” cumulative frequency distributions.

- The “**less than**” cumulative frequency distribution is obtained by adding the frequency of all the preceding (previous or earlier) classes including **the class** against which it is written or including the frequency of that class. In other words, it is obtained by adding successively the frequencies of all the previous classes including **the class** under consideration. The cumulate is started from the lowest to the highest size.
- The “**more than**” cumulative frequency distribution is obtained by adding the frequency of the succeeding (later) classes including the frequency of that class. In other words, it is obtained by finding the cumulate total of frequencies starting from the highest to the lowest class. Example consider the distribution of marks of 50 students:

Class limits	fi	Class boundary	Marks of students	Less than cumulative	Marks of students	More than cumulative
10-19	2	9.5-19.5	Less than 19.5	2	9.5 or more	$48+2=50$
20-29	4	19.5-29.5	Less than 29.5	$2+4=6$	19.5 or more	$44+4=48$
30-39	7	29.5-39.5	Less than 39.5	$6+7=13$	29.5 or more	$37+7=44$
40-49	10	39.5-49.5	Less than 49.5	$13+10=23$	39.5 or more	$27+10=37$
50-59	16	49.5-59.5	Less than 59.5	$23+16=39$	49.5 or more	$11+16=27$
60-69	8	59.5-69.5	Less than 69.5	$39+8=47$	59.5 or more	$3+8=11$
70-79	3	69.5-79.5	Less than 79.5	$47+3=50$	69.5 or more	3

Interpretation of **less than** cumulative frequency: for instance, 23 out of 50 students have marks less than 49.5.

Interpretation of **more than** cumulative frequency: for instance, 27 out of 50 students have marks 49.5 or more.

Relative frequencies for less than and more than cumulative frequencies

3.4. Graphical methods of data presentation

It is the second method of presenting data in the form of graphs.

Presenting graphically or diagrammatically helps to

- understand the information easily.
- make the data attractive to the eye.
- make comparisons of items easily.
- draw attention of the observer.

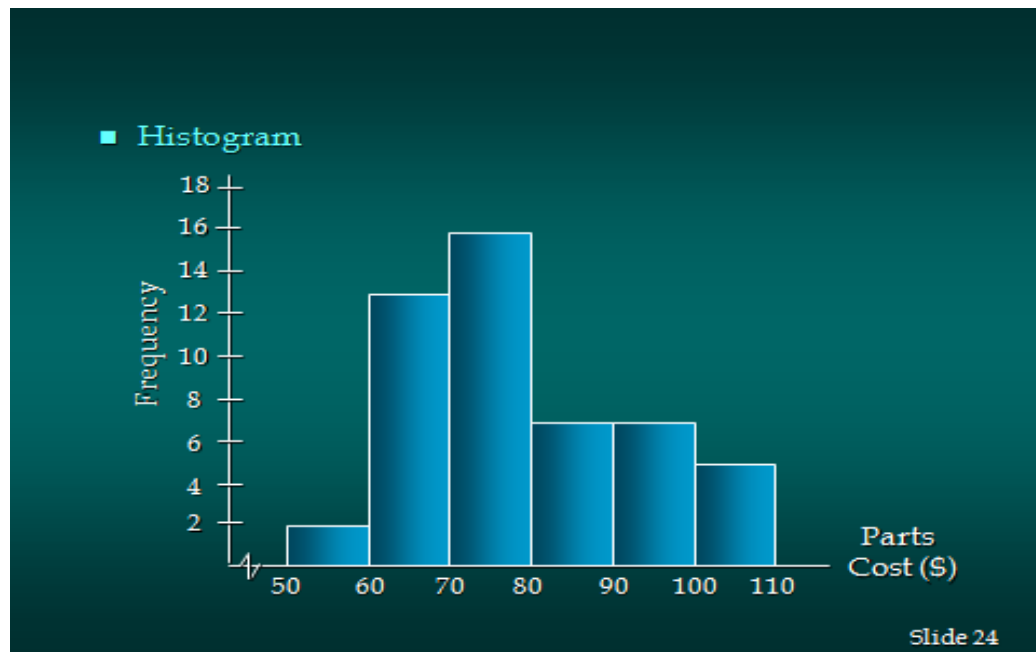
Some of the graphs are presented below:

I. The histogram

The histogram is the set of adjoining vertical bars whose width is the class interval and whose vertical heights represent the frequency of that class. On the x-axis, the class boundaries or limits are presented and on the y-axis, we present the frequency of each class. That is, the histogram is a graph consisting of a series of adjacent rectangles whose bases are equal to the class width of the corresponding classes and whose heights are proportional to the corresponding class frequencies with no gap between rectangles. In short, the **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

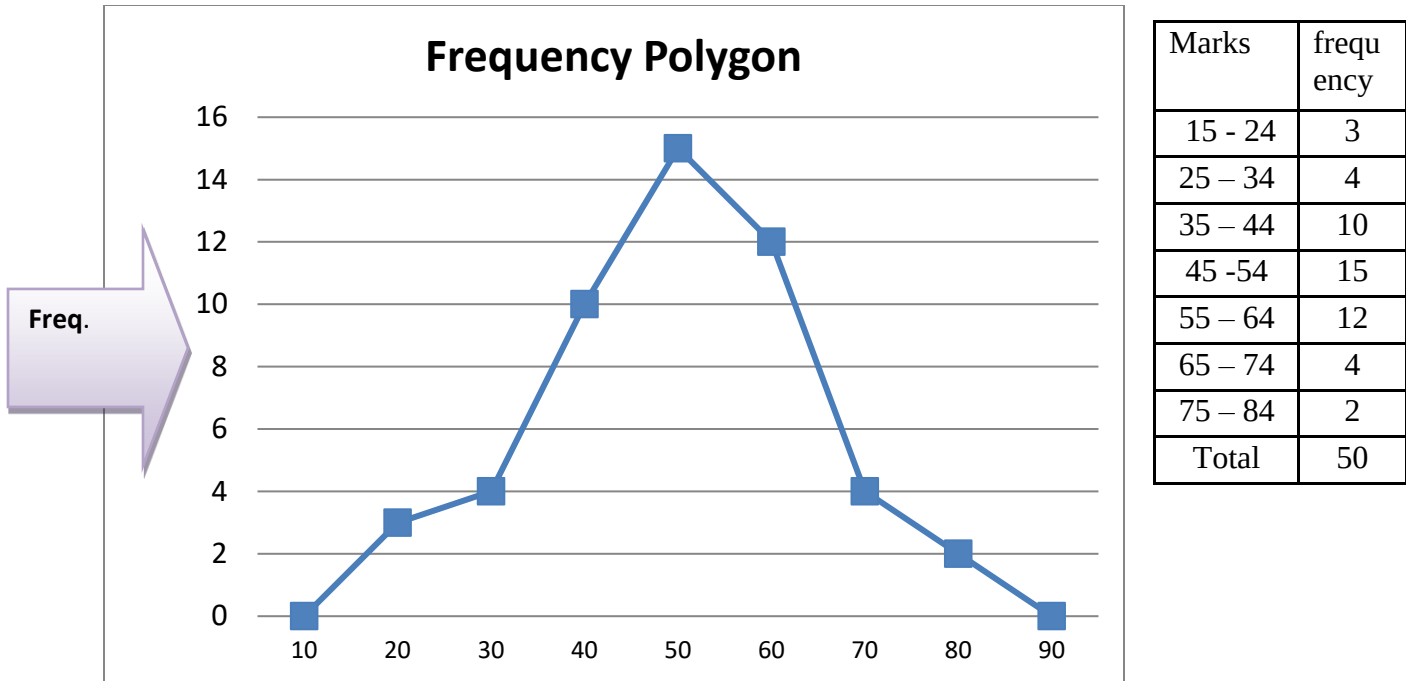
To construct a histogram, the class boundaries, the class marks or the class limits are plotted on the horizontal axis and the class frequencies are plotted on the vertical axis. The absolute (actual) frequencies may also be indicated on the top of each rectangle.

Consider the distribution for costs of customers on certain item:



II. Frequency polygon

A frequency polygon is a line graph where class frequencies are plotted against the class marks (or mid points) and the successive points are connected by straight lines. As in the histogram, the class marks are plotted along the X-axis and frequencies along the Y-axis. In other words, the **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points. Two classes with zero frequencies at both ends must be added to tie down the graph to the X-axis. **Example:**



III. Ogive curves

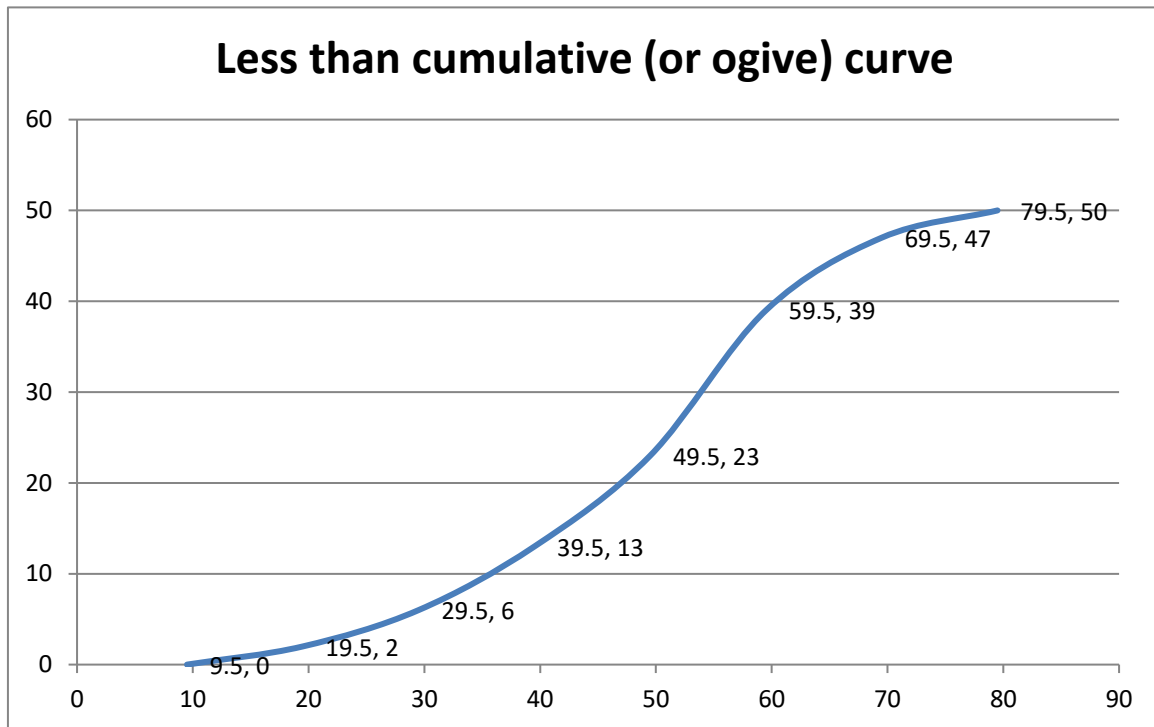
The **ogive curve** is a graph that represents the cumulative frequencies for the classes in a frequency distribution. It is also known as the cumulative frequency curve, which is a graphical representation of a cumulative frequency distribution. There are two types of Ogive curves: the “less than Ogive” and the “or more Ogive”.

- i. **The “less than” Ogive** – the less than cumulative frequencies are plotted against upper class boundaries of their respective classes and they are joined by either straight lines or smooth curves.
- ii. **The “more than” Ogive**- in this case, the “ or more than” cumulative frequencies are plotted against the lower class boundaries of their respective classes, and the connections may be by straight lines or smooth curves.

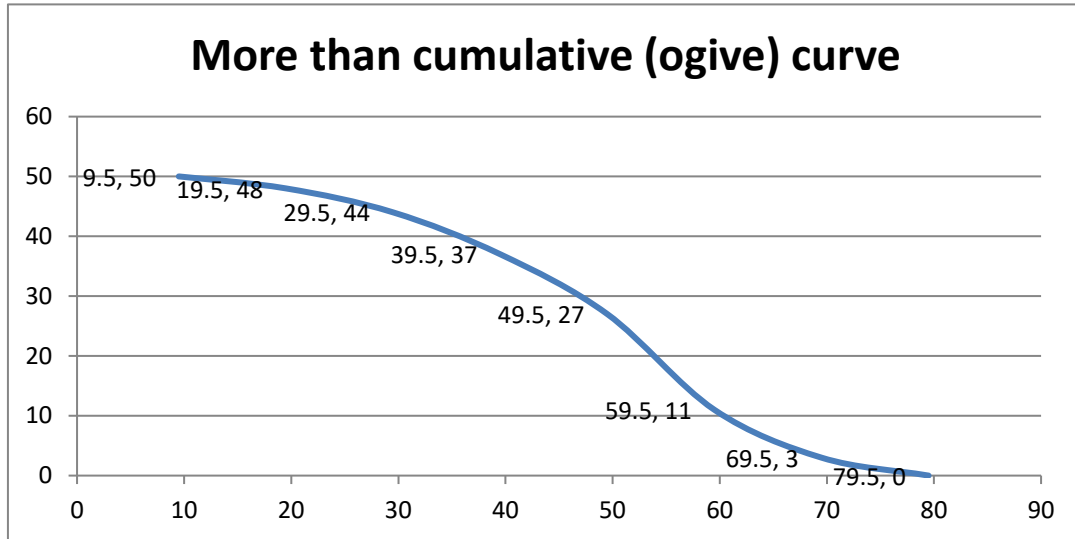
Example :

Class boundary	marks	Less than cumulative	marks	More than cumulative
	Less than 9.5	0		
9.5-19.5	Less than 19.5	2	9.5 or more	50
19.5-29.5	Less than 29.5	6	19.5 or more	48
29.5-39.5	Less than 39.5	13	29.5 or more	44
39.5-49.5	Less than 49.5	23	39.5 or more	37
49.5-59.5	Less than 59.5	39	49.5 or more	27
59.5-69.5	Less than 69.5	47	59.5 or more	11
69.5-79.5	Less than 79.5	50	69.5 or more	3
			79.5 or more	0

C u m u l a t i v e F r e q u e n c y



C u m u l a t i v e F r e q u e n c y



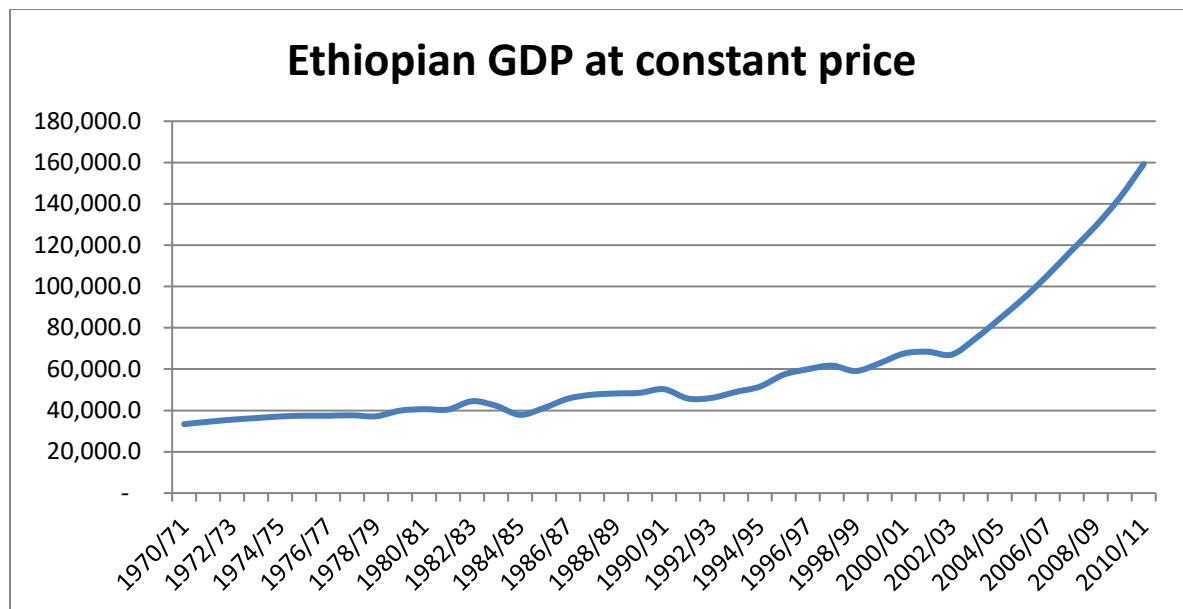
Note that the Ogives can help us in estimating the number of observations falling either above, below or in between values.

IV. The line graphs

It is the graph that displays the item on the X-axis and the value on the Y-axis.

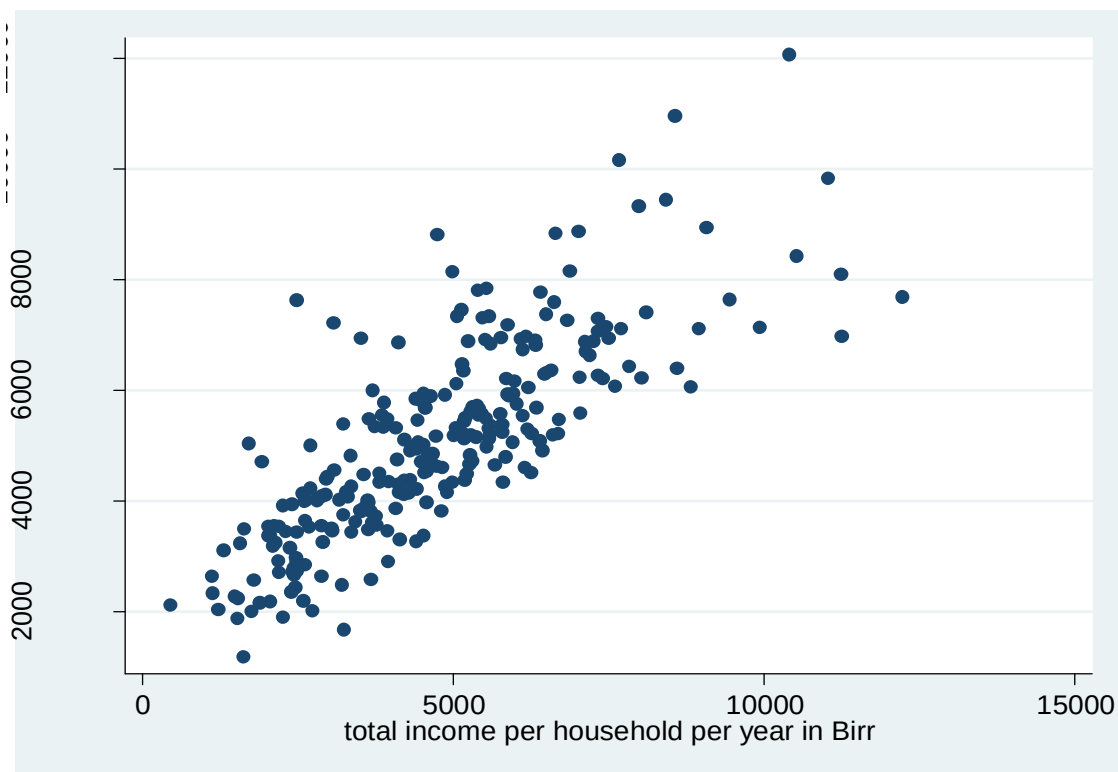
-used to present discrete frequency distributions.

Example: consider the Ethiopian GDP at constant market for the last 40 years (1971-2011).



V. Scatter plots (two variables) (optional)

Scatter plots show the relationship between two variables. For instance, income and expenditure of households can be presented using a scatter plot.



2.4 Diagrammatical Presentation of Data

Diagrammatical presentation of data is usually used to present categorical data. The two most commonly used charts (usually for qualitative data are bar diagram (Bar chart) and Pie diagram (Pie chart).

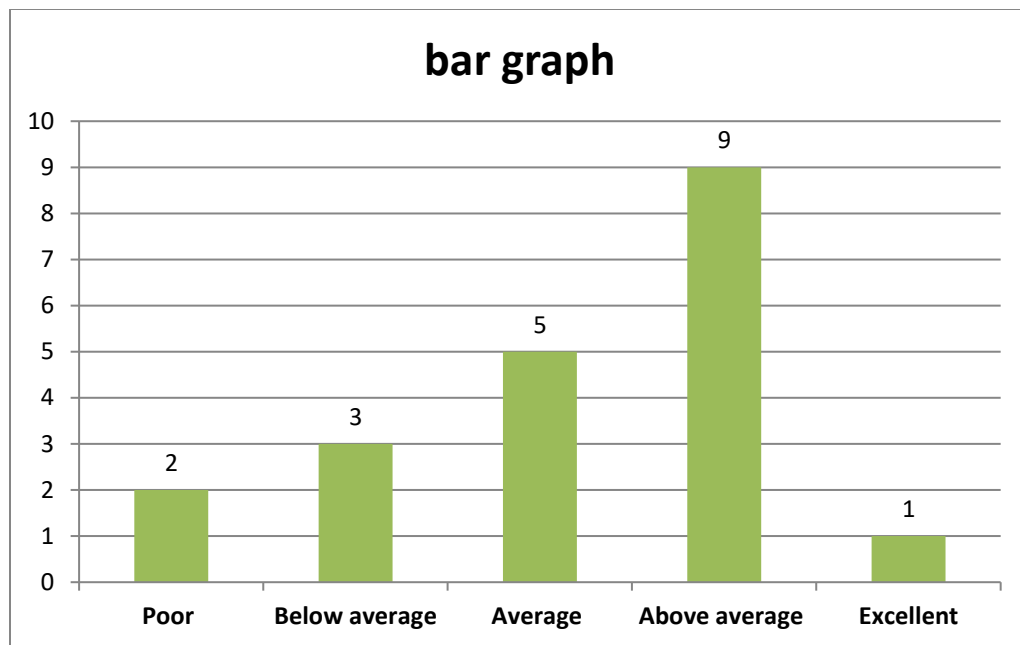
a) Bar chart

A bar graph is a graphical device for depicting qualitative data. On the horizontal axis we specify the labels that are used for each of the classes. A frequency, relative frequency, or percent frequency scale can be used for the vertical axis. The bars are separated to emphasize the fact that each class is a separate category. It is the simplest and most frequently used graphical methods of data presentation. It uses a series of equally spaced bars of uniform width. The bases of the bars are the categories on the horizontal axis while the height of each bar represents the absolute frequency of the particular category. It is of simple, multiple, component and percentage component types.

Example 1: Rating of students for introduction to statistics

rating	frequency	Relative frequency	%
Poor	2	0.1	10
Below average	3	0.15	15
Average	5	0.25	25
Above average	9	0.45	45
Excellent	1	0.05	5

A **simple bar** chart of the number of students by their performance:

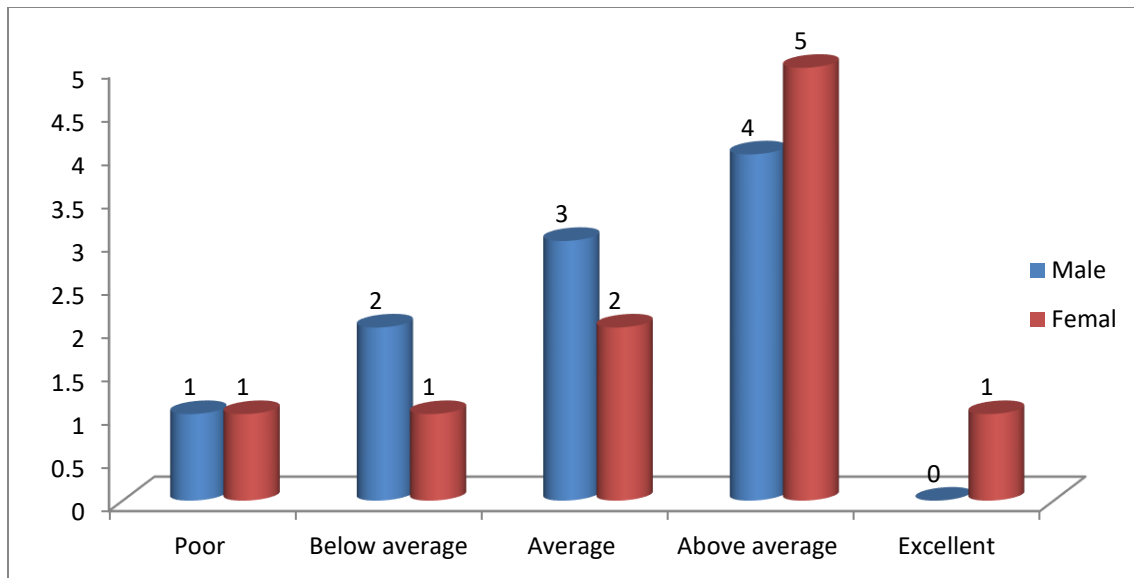


The bar graph for performance of students that can be sub divided into components (say, by sex) as:

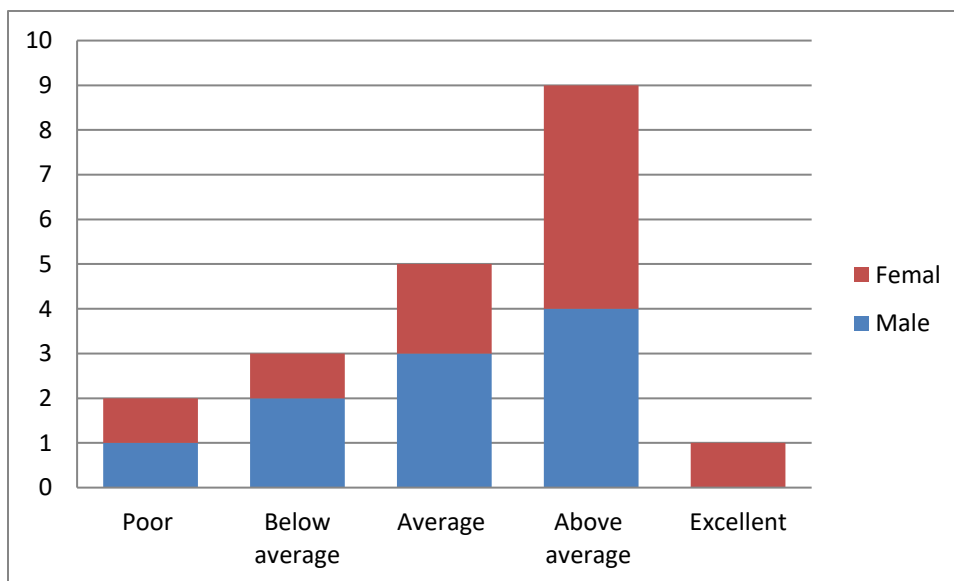
Multiple Bar chart and **sub-divided** bar chart or **component** bar chart.

rating	total	Male	Female
Poor	2	1	1
Below average	3	2	1
Average	5	3	2
Above average	9	4	5
Excellent	1	0	1

Multiple Bar charts



The **sub-divided (component)** bar chart would look like:

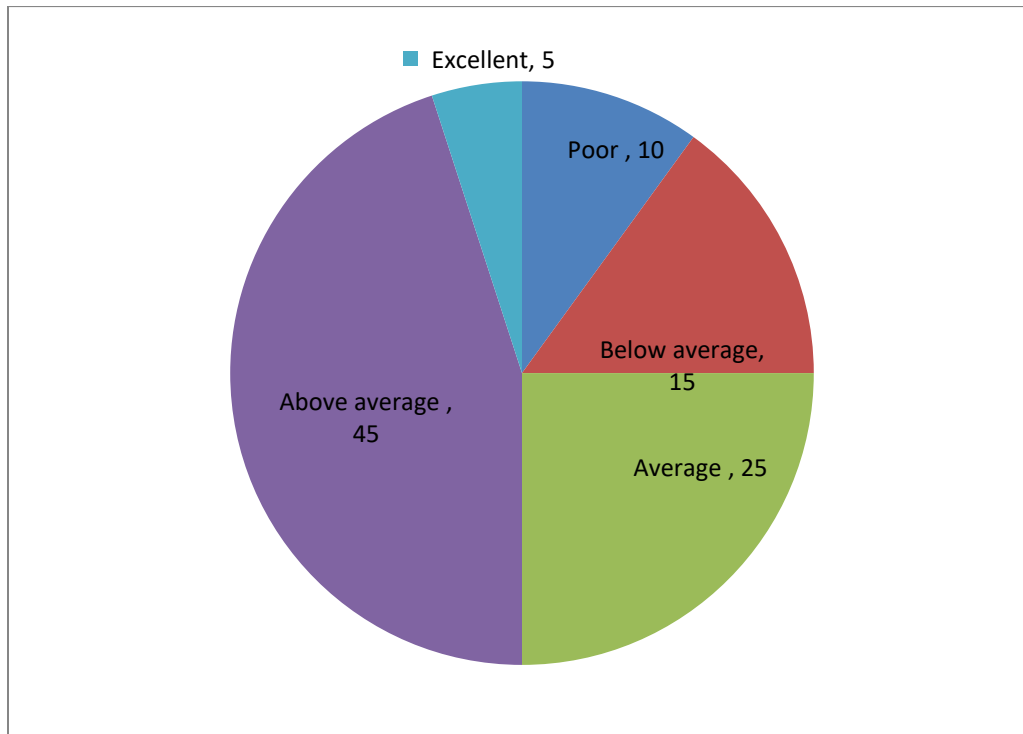


b) Pie-chart (diagram)

Pie- charts are very popularly used in practice to show percentage breakdowns of a categorical data. A pie-chart is a circle representing a set of data by dividing a circle into sectors proportional to the number of items in the categories. To construct a pie-chart, we need to find the relative frequencies; compute the central angles: $\theta_i = \frac{f_i}{n} \times 360^\circ$, and finally draw a circle partitioned according to the central angles.

Example:

rating	frequency	Relative frequency	%	size of central angle
Poor	2	0.1	10	36
Below average	3	0.15	15	54
Average	5	0.25	25	90
Above average	9	0.45	45	162
Excellent	1	0.05	5	18



Pie-charts are useful to make comparisons as long as the number of components is not large (usually, five or six components are used).

4. Measures of Central Tendency

4.1. Definition and objectives of the chapter

The most important aspect of studying the distribution of a sample measurement is the position of the central value, that is, a representative value about which the measurements are distributed and when it is convenient. This figure is known as the *average* of the group. If the numbers of the group are arranged in order of magnitude, the averages tend to fall around the central position in the group, so averages are called measures of central tendency. In short, any measure intended to represent the center of data set is called **a measure of location or central tendency**. Measure of central tendencies is statistical figures which indicate the average or the center of a given distribution (data). It enables us to compare two or more series pertaining to the same time period (e.g. student's result in a semester) or within the same distribution through time (e.g. one's GPA in semesters). The three most common measures of central tendencies are: mean, median, and mode.

The most important objectives of measuring central tendency are:

- To determine a single value around which the other data will concentrate,
- To summarize/condense the volume of the data, and
- To facilitate comparison within one group or between groups of data.

Desirable properties of measure of central tendency

A measure of central tendency is best if it possess most of the following properties:

- be simple to understand and easy to calculate/interpret,
- exist and be unique,
- be rigidly defined by mathematical formula,
- be based on all observations,
- not be seriously affected by extreme observations,
- have capable of further statistical analysis and/or algebraic manipulation.

The Summation Notation (Σ)

Let a data set consists of a number of observations, represents by $x_1, x_2, \dots, x_n$ where n (the last subscript) denotes the number of observations in the data and x_i is the i^{th} observation. Then the sum is:

$$x_1 + x_2 + \dots + x_n = \sum_{i=1}^n x_i$$

For instance, a data set consisting of six measurements 21, 13, 54, 46, 32 and 37 is represented by x_1, x_2, x_3, x_4, x_5 and x_6 where $x_1 = 21$, $x_2 = 13$, $x_3 = 54$, $x_4 = 46$, $x_5 = 32$ and $x_6 = 37$.

Their sum becomes $\sum_{i=1}^6 x_i = 21 + 13 + 54 + 46 + 32 + 37 = 208$.

$$\text{Similarly } x_1^2 + x_2^2 + \dots + x_n^2 = \sum_{i=1}^n x_i^2$$

Some Properties of the Summation Notation

1. $\sum_{i=1}^n a = n.a$ where a is a constant number.
2. $\sum_{i=1}^n b.x_i = b \sum_{i=1}^n x_i$ where b is a constant number
3. $\sum_{i=1}^n (a + b.x_i) = n.a + b \sum_{i=1}^n x_i$ where a and b are constant numbers
4. $\sum_{i=1}^n (x_i \pm y_i) = \sum_{i=1}^n x_i \pm \sum_{i=1}^n y_i$

3.2. The Mean

I. Arithmetic Mean (A.M) or simple mean

The A.M is the best known and most commonly used average. Sometimes it is referred as simply “the mean”. The A.M can be defined for ungrouped data (with a single and/or more frequencies)

a. A.M. for ungrouped data: when f_i (or frequency) is 1 for each observation

Suppose that $x_1, x_2, \dots, x_n$ are n observed values in a sample of size n from a population of size N , $n < N$. Then, the sample mean ($\bar{x}$, read as "x - bar"), and the population mean (μ , read as "mu") are defined as:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} ; \text{ For sample data}$$

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N} ; \text{ For population data}$$

Recall that $\bar{x}$ is a statistic used to estimate μ (a parameter).

Example 1: Find the mean of the scores (out of 35) of a population of 12 students in a test:

20, 23, 26, 17, 30, 10, 8, 34, 25, 28, 29 and 14.

Solution: Since $N = 12$, and $x_1 = 20, x_2 = 23, \dots, x_{12} = 14$, we have

Solution: Since $N = 12$, and $x_1 = 20, x_2 = 23, \dots, x_{12} = 14$, we have $\mu = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{264}{12} = 22$.

b. A.M. for ungrouped frequency distribution

If X_1 occurs f_1 times, X_2 occurs f_2 times, ..., X_n occurs f_n times, then the A.M. will be

$$\bar{X} = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i}.$$

Example 2: Find the mean of the following ungrouped sample data:

x_i	1	2	3	4	5	6	7	Total
f_i	5	9	12	17	14	10	6	73
$f_i x_i$	5	18	36	68	70	60	42	299

$$\text{Therefore, } \bar{X} = \frac{\sum_{i=1}^7 f_i X_i}{\sum_{i=1}^7 f_i} = \frac{299}{73} = 4.09$$

c. A.M. for a grouped frequency distribution (continuous Data)

In case of grouped continuous data, the mid-points (class marks) of the various classes are taken as representatives of their respective classes to estimate the mean with the assumption that the values in a class are concentrated around the center. Therefore, for grouped continuous data with m classes where the i^{th} class has frequency f_i and class mark x_i , the mean is given by:

$$\bar{X} = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} \quad \text{Where: } X_i = \text{class-mark/midpoint of } i^{\text{th}} \text{ class, and } f_i = \text{Frequency of } i^{\text{th}} \text{ class.}$$

Example: Find the mean for the following continuous distribution.

C.I.	1 - 5	6 - 10	11 - 15	16 - 20	21 - 25	26 - 30	31 - 35
Freq.	4	8	12	6	3	4	3

Solution: To calculate the mean, we have to construct the following table along with the class marks for each class.

Class mark (x_i) or midpoint	3	8	13	18	23	28	33	Total
Frequency (f_i)	4	8	12	6	3	4	3	40
$f_i x_i$	12	64	156	108	69	112	99	620

$$\text{Thus, } \bar{X} = \frac{\sum_{i=1}^7 f_i X_i}{\sum_{i=1}^7 f_i} = \frac{620}{40} = 15.5.$$

Finding missing value in the frequency distribution and correcting the wrong mean

Missing frequency (given the mean)

Exercises:

1. If the mean of the following data is 32, find the missing freq., f

Marks of students	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59
Number of students	5	15	20	f	20	10

Answer: f=30

2.

<u>Class A</u>	<u>Class B</u>
$n = 20$	$n = 30$
$\bar{X} = 70$	$\bar{X} = 80$

- (1) Find ΣX for class A.
- (2) Find ΣX for class B.
- (3) Find ΣX for the combined classes.
- (4) Find $\bar{X}$ for the combined classes.

3. (i) A student made 65, 78, and 85 on three tests. What must she make on the fourth test in order to have an average of 80?

$$80 = \frac{65 + 78 + 85 + X}{4}$$

$$320 = 228 + X \Rightarrow X = 320 - 228 = 92$$

- (ii) A student had an average of 75 on four tests. If he makes 95 on the fifth test, what is his average for the five tests?

$$\bar{X} = \frac{4 * 75 + 95}{5} = 79$$

Wrong values

If a wrong figure has been used in calculating the mean, we can correct it if we know the correct figure that should have been used. Let X_{wr} denote the wrong figure used in calculating the mean and X_c be the correct figure that should have been used. Then $\bar{X}_{wr}$ be the wrong mean calculated using X_{wr} , then the correct mean, $\bar{X}_{correct}$, is given by

$$\bar{X}_{correct} = \frac{n\bar{X}_{wr} + X_c - X_{wr}}{n}$$

Examples:

1. An average weekly wage of 25 workers of XYZ Company is Birr 378.4. It was later discovered that one figure was entered wrongly as Birr 160 instead of the correct value of Birr 200. Find the correct mean.
2. Suppose the mean of 200 observations was 50. Later on it was discovered that two observations were wrongly read as 92 and 8 instead of the correct values of 192 and 88. Find the correct mean.

Solution

$$\bar{X}_c = \frac{n\bar{X}_w + X_c - X_w}{n}$$

$$1. \quad \bar{X}_c = \frac{25(378.4) + 200 - 160}{25} = 380$$

$$\bar{X}_c = \frac{n\bar{X}_w + X_{c1} + X_{c2} - X_{w1} - X_{w2}}{n}$$

$$2. \quad \bar{X}_c = \frac{200(50) + 192 + 88 - 92 - 8}{200} = 50.9$$

Properties of A.M

1. The algebraic sum of deviations from the mean is always zero: $\sum_{i=1}^n (x_i - \bar{x}) = 0$. For

$$\sum x_i = n\bar{x}$$

2. The sum of squares of deviations from the mean is minimum. That is, $\sum_{i=1}^n (x_i - A)^2$ is minimum when $A = \bar{x}$.

3. Let $y = a \pm bx$ be a linear function of x , then $\bar{y}$ is also same linear function of $\bar{x}$, i.e. $\bar{y} = a \pm b\bar{x}$

That is, if the mean of $x_1, x_2, \dots, x_n$ is $\bar{x}$, then

a) the mean of $x_1 \pm k, x_2 \pm k, \dots, x_n \pm k$ will be $\bar{x} \pm k$

b) The mean of $kx_1, kx_2, \dots, kx_n$ will be $k\bar{x}$.

Merits of the A.M.

- Easy to calculate and understand
- Rigidly defined by mathematical formula given
- Is based on all the observations class
- Is a stable average

Demerits of the A.M.

- affected by extreme values
- cannot be calculated if all the values are not
- cannot be used when there is an open-ended
- cannot be determined for qualitative data

b) Weighted Mean

While calculating the simple arithmetic mean, we had given equal importance to all values. But there are cases where the relative importance is not the same for all items. When this is the case, it is necessary to assign their weights (i.e., relative importance), and then calculate a weighted mean.

The Weighted mean ($\bar{x}_w$) of a set of values $x_1, x_2, \dots$, and x_n whose relative importance (weight) $w_1, w_2, \dots$, and w_n , is given by:

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum w_i x_i}{\sum w_i}$$

Examples:

1. A typical example of the weighted mean is computation of your GPA. Suppose that a student was registered for five courses with 4, 4, 3, 2 and 3 credit hours and she obtained grades B, A, C, D and A, respectively. Find her GPA.

Solution:

Here, the numerical equivalences of the grades are the values and the corresponding credit hours are their respective weights. So, we have

Values (x_i)	3	4	2	1	4	Total
Weights (w_i)	4	4	3	2	3	16

$w_i x_i$	12	16	6	2	12	48
-----------	----	----	---	---	----	----

and $\text{GPA} = \frac{48}{16} = 3.0$.

2. In a certain English class, quizzes make up 15% of the final average, major tests make up 35%, papers make up 20%, and the final exam makes up 30%. If a student has an average of 90 on quizzes, 80 on major tests, 75 on papers, and 85 on the final exam, what is his final average?

Solution:

$$\bar{X}_w = \frac{0.15 * 90 + 0.35 * 80 + 0.2 * 75 + 0.3 * 85}{0.15 + 0.35 + 0.2 + 0.3}$$

$$\bar{X}_w = 82$$

c) Combined (Grand) Mean.

The combined mean for a combined set of data may be obtained from the separate means if their sample or population sizes are given. Suppose that the combined set was formed by combining two sets of data with means $\bar{X}_1$ and $\bar{X}_2$ with n_1 and n_2 observations, respectively. Then, the combined mean or grand mean is given by:

$$\bar{x}_c = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} .$$

Examples:

3. If the mean weight of 50 women working in a factory is 48 kg and the mean weight for men is found to be 58kg, and the total number of workers in the factory was 125, find the mean weight of all workers in the factory.

Solution: Let $\bar{x}_1 = 48\text{kg}$, $n_1 = 50$, $\bar{x}_2 = 58\text{kg}$, $n_2 = 75 (= 125 - 50)$. Then,

$$\bar{x}_c = \frac{50(48\text{kg}) + 75(58\text{kg})}{50 + 75} = 54\text{kg} .$$

2. Suppose the mean grades/marks in statistics of 100 students of a class is 72. The mean marks of boys is 75 Find the mean marks of girls in the class if the number of boys is 70.

n_1 =number of boys

$\bar{X}$ -bar(1) is mean of boys

$\bar{X}$ -bar is combined mean

n_2 =number of girls (100-70=30)

$$\bar{X}_c = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

$$72 = \frac{70(75) + 30(\bar{X}_2)}{70 + 30}$$

$$72(100) = 5250 + 30(\bar{X}_2)$$

$$\Rightarrow \bar{X}_2 = 65$$

In general, if there are k-different groups of data with $n_1, n_2, \dots, n_k$ numbers of observations with means $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$, respectively, then the combined mean is:

$$\bar{X}_c = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \dots + n_k \bar{X}_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum n_i \bar{X}_i}{\sum n_i}.$$

d) Geometric Mean (G.M.)

It is the mean of n positive numbers $x_1, x_2, \dots, x_n$ and is the n^{th} root of their product. It is used in averaging: **ratios, rate of change/growth rate, economic indices** (like Consumer price index), **compound interest, discounting**, etc.

$$\Rightarrow G.M. = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}.$$

$$\text{If } n = 2, \text{ then } G.M. = \sqrt{x_1 \cdot x_2}$$

Examples: 1. The G.M. of 3 & 12 is $\sqrt{3 \cdot 12} = \sqrt{36} = 6$

2. The G.M. of 2, 3 & 36 is $\sqrt[3]{2 \cdot 3 \cdot 36} = \sqrt[3]{216} = 6$

3. The G.M. of 1, 2, 8 & 16 is $\sqrt[4]{1 \cdot 2 \cdot 8 \cdot 16} = 4$

4. An economy has grown for the last five consecutive years as 5%, 6%, 4.5%, 3% and 7.5%. What was the average growth rate during these years?

$$\text{Solution: Rate of growth} = \text{G.M.} = \sqrt[5]{(0.05)(0.06)(0.045)(0.03)(0.075)} \approx 4.97\%$$

For grouped data: it is given by $\log \text{GM} = \frac{1}{n} (\sum_{i=1}^n f_i x_i)$

e) Harmonic Mean (H.M.)

It is the mean of n numbers $x_1, x_2, \dots, x_n$ and is defined as n divided by the sum of the reciprocals of the n numbers. It is appropriate for situations when the average of **rates** is desired (e.g., it helps to find the average speed of a trip over a route divided into constant speed segments (of distance), ..). For example, if one travels half-way to a destination at 20 mi/hr, and then goes 60 mi/hr for the second half of the distance. The average speed is 30mi/hr.

$$\Rightarrow H.M. = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_i}} \dots\dots\dots \text{for ungrouped data}$$

Example:-1. The H.M. of 6 & 4 is $\frac{2}{\frac{1}{6} + \frac{1}{4}} = \frac{24}{5}$

2. The H.M. of the first six natural numbers is $\frac{6}{1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6}} = \frac{360}{147}$

$$\Rightarrow H.M. = \frac{n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \dots + \frac{f_i}{x_n}} = \frac{n}{\sum (\frac{f_i}{x_i})} \dots\dots\dots \text{for grouped data}$$

Note! $AM \geq GM \geq HM$.

Example

CB	f_i	x_i	$X_i * f_i$	$\log x_i$	$(f_i * \log x_i) * 1/n$	f_i/x_i
2-4	20	3	60	0.5	0.095	6.7
4-6	40	5	200	0.7	0.28	8.0
6-8	30	7	210	0.8	0.254	4.3
8-10	10	9	90	1.0	0.095	1.1
sum	100		560		0.724	20.1

$$\bar{X} = \frac{560}{100} \text{ -----A.M.}$$

$$\bar{X} = 5.6$$

$$G.M. = \text{antilog}\left(\frac{1}{n} \sum_{i=1}^n f_i x_i\right) = \text{antilog}\left(\frac{72.4}{100}\right) = 100.724 = 5.29$$

$$H.M. = \frac{n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \dots + \frac{f_i}{x_n}} = \frac{n}{\sum \left(\frac{f_i}{x_i}\right)} = \frac{100}{20.1} = 4.98$$

Thus, A.M. > G.M. > H.M.

Exercise: Find two natural numbers whose A.M. is 10 and G.M. is 8.

The Median

It has been pointed out that whenever there is a frequency distribution with open-end intervals, the arithmetic mean cannot be calculated. Also the mean is greatly affected by extremely large or small values. Hence, in such cases, the mean cannot be a good representative. Instead, other measures are used to describe the data. In this section, we will discuss the most popular measure of position, the median, and other related measures known as quintiles. Positional measures are chosen because of their positions.

The **median** is, as its name indicates, the middle most value in the **arrangement** in an ascending or descending order of magnitude, which divides the data in to two equal parts. It is the value which exceeds, and is exceeded by, equal number (i.e., half) of the observations. That is, the median is found by arranging the data in an increasing or decreasing order of magnitude. We can consider the following three cases in finding the median:

Case 1: Ungrouped data (with frequency=1 for all x_i): also known as **individual series**. In such cases, the median is the value of the middle term when the data are arranged in order of magnitude. When the number of observations is odd, there will always be a single value in the middle of the arrayed data. When n is even, however, there will be two middle observations, and the median is the mean of these two values. Let $x_1, x_2, \dots, x_n$ be n ordered observations. Then, the median value is

$$\text{a) the } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ value, if } n \text{ is odd and} \quad \text{b) the } \frac{\left(\frac{n}{2}\right)^{\text{th}} + \left(\frac{n}{2} + 1\right)^{\text{th}}}{2} \text{ value, if } n \text{ is even.}$$

Example 1: Find the median for the following data: 5, 2, 7, 1, 9, 10, 12. $N=7$

Solution: $n = 7$ is odd, we use above formula, and the median is the 4th item in the array. i.e. median = 7.

Example 2: Find the median value of the population figures (in thousands) of 10 cities: 2000, 1180, 1785, 1500, 560, 782, 1200, 385, 1123, 222.

Solution: The arrayed data is: 222, 385, 560, 782, 1123, 1180, 1200, 1500, 1785, 2000, since $n = 10$ is even, the median is the mean of the 5th, and 6th values; i.e., Median = $\frac{1123+1180}{2} = 1151.5$ thousands.

Case 2: Ungrouped data (with frequency= f_i for x_i): In this case also, the median is obtained by the same formula. Only one more step, finding the less than cumulative frequencies, is added, because cumulative frequency distribution is itself an arrangement of values in an order:

- Find the less than cumulative frequencies.
- Look at the cumulative frequency and find that total which is either equal to or next higher to $\left(\frac{n+1}{2}\right)^{\text{th}}$ obs. when n =odd and the average of the two middle values when= n =even, and the corresponding value is the median.

Example: Find the median of the data given below

x_i	3	5	6	8	10
f_i	4	4	7	9	5

Solution: First construct the cumulative frequency distribution:

x_i	f_i	Cum. Freq ($\leq$)
-------	-------	----------------------

3	4	4
5	4	8
6	7	15
8	9	24
10	5	29

Since $n = 29$, $\left(\frac{n+1}{2}\right)^{th} = 15^{th}$ value, the median is 6.

Example 4: Suppose that the frequencies of the above example are 4, 6, 5, 9, 6, respectively. Find the median.

Solution: The cumulative frequencies are:

x_i	f_i	Cum. Freq. ($\leq$)
3	4	4
5	6	10
6	5	15
8	9	24
10	6	30

Here, $n = 30$, and median = 15.5^{th} obs. Thus, it is the average of value = 15^{th} and 16^{th} obs. That is, $(6+8)/2 = 7$

Case 3: Grouped continuous

For continuous grouped data, the exact median cannot be obtained unless the original data has been retained. Hence, the median has to be interpolated (or estimated) from the median class. An interpolation formula which is based on the assumption that classes are uniformly distributed is:

$$\text{Median} = L + \frac{w}{f_{\text{med}}} \left(\frac{n}{2} - CF \right) = \tilde{x}$$

Where: L = the lower class boundary of the median class;

w = the class width of the median class;

f_{med} = the frequency of the median class; and

CF = the cum. Freq. corresponding to the class preceding the median class.

That is, the sums of the frequencies of all classes lower than the median class. Where the median class is the class which contains the $\left(\frac{n}{2}\right)^{th}$ observation whether n is odd or even, since the items have already lost their originality once they are grouped into continuous classes.

Example: For the following distribution, find the median.

Grade	Frequency
40 – 49	5
50 – 59	18
60 – 69	27
70 – 79	15
80 – 89	6

Solution: Construct the less than cumulative frequency distribution as follows

C.B's	Freq. (f_i)	CF(<)
39.5 – 49.5	5	5
49.5 – 59.5	18	23
59.5 – 69.5	27	50
69.5 – 79.5	15	65
79.5 – 89.5	6	71

Since $n = 71$, $71/2 = 35.5$, and the smallest CF greater than or equal to 35.5 is 50; thus, the median class is the third class. And for this class, $L = 59.5$, $w = 10$, $f_{med} = 27$, $CF = 23$. Then applying Formula, we get:

$$\tilde{x} = 59.5 + \frac{10}{27}(35.5 - 23) = 64.13.$$

Example:- Find the median of the following open-ended distribution.

IQ	Freq.
< 90	3
90 – 99	14
100 – 109	22
110 – 119	19

> 119	7
-------	---

Solution: The less than cumulative frequency distribution is:

Class boundary	Freq.	CF
< 89.5	3	3
89.5 - 99.5	14	17
99.5 - 109.5	22	39
109.5 - 119.5	19	58
> 119.5	7	65

Since $n = 65$, $n/2 = 32.5$, the median class is the third class; and

$L = 99.5$, $w = 10$, $f_{med} = 22$, $CF = 17$. Then,

$$\tilde{x} = 99.5 + \frac{(32.5 - 17)10}{22} = 99.5 + 7.05 = 106.55.$$

Merits of the Median

- easy to calculate and understand
- rigidly defined by mathematical formula
- can be computed when there is an open-ended class
- not affected by extreme values treatments

Demerits of the Median

- is based only on the middle item
- is relatively less stable than the mean
- doesn't consider all the item
- not suitable for further mathematical

Some measures of non- central locations or Quintiles

Quintiles are the value of the variable that divides the whole series in to a number of equal parts. The most commonly used ones include **quartiles, deciles, and percentiles**. These measures are computed in the same way as the median, but the data must be arranged only in increasing order. As their names indicate, the quartiles divide a set of ordered data in to four equal parts, deciles in to ten equal parts, and percentiles in to hundred equal parts.

1. Quartiles

The quartiles, usually denoted by Q_1 , Q_2 and Q_3 , are known as the first (or lower), second, and third (or upper) quartiles, respectively. The first quartile (Q_1) is the value for which one-fourth of the data lie below it and the rest three-fourth lie above it. The second quartile (Q_2) is the value for which half

of the data lie below or above it. Since this is the property of the median, Q_2 is equal to the median. The third quartile (Q_3) is the value for which three-fourth of the arranged items lie below it, or one-fourth lie above it.

Like the median, we can identify three cases in locating the quartiles:

Case 1: Quartiles in ungrouped data. In this case, we proceed as follows: Let $x_1, x_2, \dots, x_n$ be n ordered observations. The i^{th} quartile (Q_i) is the value of the item corresponding with the

$\left(\frac{i}{4}(n+1)\right)^{\text{th}}$ position, for $i = 1, 2, 3$.

That is, after arranging the data in ascending order, Q_1 , Q_2 , & Q_3 are, respectively

$\left(\frac{1}{4}(n+1)\right)^{\text{th}}$, $\left(\frac{2}{4}(n+1)\right)^{\text{th}}$, and $\left(\frac{3}{4}(n+1)\right)^{\text{th}}$ values.

Example 1: Find Q_1 and Q_3 for 1080, 1100, 1120, 1150, 1160, 1200, 1400.

Solution: The data is already arranged: 1080, 1100, 1120, 1150, 1160, 1200, and 1400.

Then, $Q_1 = \frac{1}{4}(n+1)^{\text{th}} = \frac{1}{4}(7+1)^{\text{th}} = 2^{\text{nd}}$ value = 1100; and

$$Q_3 = \frac{3}{4}(n+1)^{\text{th}} = \frac{3}{4}(7+1)^{\text{th}} = 6^{\text{th}} \text{ value} = 1200.$$

Note that (when $n = 7$), the median value was 1150, which is nothing but Q_2 is the median of the values below the median and Q_3 is the median of the values above the median.

Case 2: Quartiles for discrete data in frequency distribution.

Arranged in a frequency distribution this case also, we will follow the same procedure as the median.

That is, we construct the less than cumulative frequency distribution and apply the above formula.

To find Q_1 , compute $\frac{1}{4}(n+1)$, and find the minimum cumulative frequency greater than or equal to this value. Then, the value corresponding to this cumulative frequency is the value of Q_1 .

To find Q_3 , compute $\frac{3}{4}(n+1)$ and locate it in a similar fashion.

Example 2: Find the median and the two quartiles from the following table.

x_i	5	7	9	11	13	15	17	19
f_i	1	2	7	9	11	8	5	4

Solution: We first find the less than cum. freq. and proceed as discussed above.

x_i	5	7	9	11	13	15	17	19
f_i	1	2	7	9	11	8	5	4
Cum. freq. ($\leq$)	1	3	10	19	30	38	43	49

We see that since $n = 49$,

$$Q_1 = \frac{1}{4}(49+1)^{th} = 12.5^{th} \text{ value} = 11; Q_2 = \frac{2}{4}(49+1)^{th} = 25^{th} \text{ value} = 13; \text{ and}$$

$$Q_3 = \frac{3}{4}(49+1)^{th} = 37.5^{th} \text{ value} = 15.$$

Case 3: Quartiles in grouped continuous data.

For continuous grouped data, like the median, the exact quartiles cannot be obtained and thus the following interpolation formula gives approximate values;

$$Q_i = L + \frac{w}{f_{Q_i}} \left(\frac{in}{4} - CF \right)$$

Where $i = 1, 2, 3$, and L, w, f_{Q_i} and CF are defined in the same way as the median. The class under question is the one including $(\frac{i}{4}n)^{th}$ value. That is, the class with the minimum frequency greater than or equal to $(\frac{i}{4}n)$ is the class of the i^{th} quartile.

Example 3: Calculate the first and the third quartiles for the following data.

Grade	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
Frequency.	5	18	27	15	6	3

Solution: The first step is to construct the cumulative frequency distribution.

Class boundary	Frequency	Cum. freq. (less than)
39.5 - 49.5	5	5
49.5 - 59.5	18	23
59.5 - 69.5	27	50
69.5 - 79.5	15	65
79.5 - 89.5	6	71
89.5 - 99.5	3	74

To find Q_1 : find its class, determine L, w, f_{Q_1} , and CF, and apply the formula.

$\frac{n}{4} = 18.5$, the minimum cumulative frequency greater than or equal to 18.5 is 23. So, the class corresponding to this cumulative frequency is the second class. For this class of Q_1 , $L = 49.5$, $w = 10$, $f_{Q_1} = 18$, and $CF = 5$.

Now applying Formula 2.17, $Q_1 = 49.5 + \frac{(18.5-5) \times 10}{18} = 49.5 + 7.5 = 57$.

To calculate Q_3 , following similar procedures,

$\frac{3n}{4} = 55.5$, so it's in the 4th class, and $L = 69.5$, $w = 10$, $f_{Q_3} = 15$, and $CF = 50$.

Therefore, $Q_3 = 69.5 + \frac{(55.5-50) \times 10}{15} = 69.5 + 3.67 = 73.17$.

2. Deciles:

The deciles divide a given set of data in to ten equal parts. There are nine deciles, usually denoted by $D_1, D_2, \dots, D_9$; generally, D_k is used to denote the k^{th} decile. The first decile (D_1) is the value for which $(1/10)^{\text{th}}$ of the arranged data lies below it, $(2/10)^{\text{th}}$ of these items lie below $D_2, \dots$ and so on.

When $x_1, x_2, \dots, x_n$ are n ordered observations, the k^{th} decile (D_k) is the value corresponding to the

$\left(\frac{k}{10}(n+1)\right)^{\text{th}}$ position, $k = 1, 2, \dots, 9$. That is, after arranging the data in an increasing order,

$$D_1 = \frac{1}{10}(n+1)^{\text{th}} \text{ value}, D_2 = \frac{2}{10}(n+1)^{\text{th}} \text{ value}, \dots, D_9 = \frac{9}{10}(n+1)^{\text{th}} \text{ value}.$$

Example 1: Find D_4 and D_6 for the following set of numbers ($n = 19$).

2, 9, 7, 6, 1, 5, 8, 10, 6, 11, 6, 14, 7, 1, 2, 7, 8, 9, 7.

Solution: After arranging the numbers in ascending order of magnitude, we get: 1, 2, 2, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 10, 11, 14. Then, $D_4 = \frac{4}{10}(19+1)^{th} = 8^{th}$ value = 6; and $D_6 = \frac{6}{10}(19+1)^{th} = 12^{th}$ value = 7.

Deciles for discrete data:

In this case, the Formula is also applicable. To find the k th decile, $(k = 1, 2, \dots, 9)$, find $\frac{k}{10}(n+1)$ and look for the minimum cumulative frequency greater than or equal to this value. Then, the value associated with this cumulative frequency is the k th decile, D_k .

Example 2: For the data given in Example on page 53 (copy here!), calculate D_3 and D_8 .

x_i	5	7	9	11	13	15	17	19
f_i	1	2	7	9	11	8	5	4

Solution: The cumulative frequency for the problem given will be our reference.

To find D_3 , $\frac{3}{10}(n+1) = 15$, the minimum cumulative frequency greater than or equal to 15 is 19, and the corresponding value is 11 = D_3 .

To find D_8 , $\frac{8}{10}(n+1) = 40$, the minimum cumulative frequency greater than or equal to 40 is 43, and the corresponding value to this is 17 = D_8 .

Deciles for grouped continuous data

In a similar way we use the interpolation formula:

$$D_k = L + \frac{w}{f_{D_k}} \left(\frac{kn}{10} - CF \right), \text{ where } k = 1, 2, 3, \dots, 9.$$

Example 3: Find D_4 and D_7 for the F.D of marks of 74 students given below.

Grade	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	> 74
Frequency	2	3	8	10	15	12	14	10

Solution: The less than cumulative frequency with the class boundaries is:

Class boundaries	Frequency	Cum. freq. (less than)
39.5 – 44.5	2	2
44.5 – 49.5	3	5
49.5 – 54.5	8	13
54.5 – 59.5	10	23
59.5 – 64.5	15	38
64.5 – 69.5	12	50
69.5 – 74.5	14	64
≥ 74.5	10	74

To find D_4 : $n = 74$, $\frac{4(74)}{10} = 29.6 \leq 38$, shows D_4 is found in the 5th class, where $L = 59.5$, $w = 5$,

$$f_{D_4} = 15, CF = 23, \text{ and, } D_4 = 59.5 + \frac{5}{15}(29.6 - 23) = 61.7.$$

To find D_7 also, $\frac{7(74)}{10} = 51.8 \leq 64$, and the class of D_7 is the 7th class, where $L = 69.5$, $w = 5$, $f_{D_7} = 14$,

$$CF = 50. \text{ Thus, } D_7 = 69.5 + \frac{5}{14}(51.8 - 50) = 70.14.$$

Percentiles

Generally, p_m is used to denote the m^{th} percentile

Percentiles for ungrouped data

When $x_1, x_2, \dots, x_n$ are n ordered observations, the m^{th} percentile (p_m) is the value of the item

corresponding to the $\left(\frac{m}{100}(n+1)\right)^{\text{th}}$ position, $m = 1, 2, \dots, 99$.

Exercise: find p_{45} and p_{90} for the data in example 1, above

Percentiles for discrete data arranged in a frequency distribution

Like the fractiles, the above formula is also applicable following similar procedures. The m^{th} percentile, $p_m (m = 1, 2, \dots, 99)$, is obtained by first computing $\frac{m}{100}(n+1)$ and then looking for the minimum cumulative frequency greater than or equal to this value. Then, the value associated with this cumulative frequency is p_m . Exercise: find p_{40} & p_{70} . For the data in example 2, above

Percentiles for grouped continuous data

To identify the m^{th} percentile class, find $\frac{mn}{100}$ (for $m = 1, 2, \dots, 99$) and look for the minimum cumulative frequency greater than or equal to this value. The class corresponding to this cumulative frequency is the class of the m^{th} percentile. The desired percentile is then approximated by:

$$p_m = L + \frac{w}{f_{p_m}} \left(\frac{mn}{100} - CF \right) \quad \text{where } m = 1, 2, 3, \dots, 99.$$

Exercise: Find p_{35} and p_{80} for the data given in Example 3 above.

THE MODE

The mode of a set of data is defined as the value with the highest frequency, and which occurs more than once”.

Mode for ungrouped data

The mode or the modal value of a raw data is simply obtained by locating the observation with the maximum frequency (if there exists such a value).

Example 1: Find the mode(s) for each of the following sets of data:

- a) 3, 5, 5, 4, 6, 5, 4, 7, 8, 7.
- b) 3, 8, 7, 8, 4, 7, 8, 7, 4, 5.
- c) 3, 4, 6, 8, 7, 2, 1, 9, 5, 12.

Solution: In order to detect the mode(s), it is better if we collect values of the same magnitude together.

- a) 3, 4, 4, 5, 5, 5, 6, 7, 7, 8. The modal value is 5.
- b) 3, 4, 4, 5, 7, 7, 7, 8, 8, 8. The modes are 7 and 8.
- c) 5, 6, 7, 8, 9, 12. The mode is undefined.

From these examples, we note that a given set of data may have a single mode (called **uni-modal**), it may have two modes (called **bimodal data**), or it may have no mode at all (it does not exist or the mode is **undefined**). Generally, if a set of data has two or more modes, then it is said to be **multimodal (ill-defined)**.

Mode for grouped discrete data

In the case of discrete grouped data, the mode is determined just by inspection, i.e., by looking to that value (s) having the highest frequency.

Mode for grouped continuous data

In such cases, one can only determine the modal class easily: the class with the highest frequency.

After locating this class, the mode is interpolated using:

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times w = L + \frac{f_{\text{mod}} - f_1}{(f_{\text{mod}} - f_1) + (f_{\text{mod}} - f_2)} \times w$$

Where L = the lower class boundary of the modal class;

$$\Delta_1 = f_{\text{mod}} - f_1,$$

$$\Delta_2 = f_{\text{mod}} - f_2,$$

w = the common class width

f_1 = frequency of the class preceding the modal class;

f_2 = frequency of the class succeeding the modal class; and

f_{mod} = frequency of the modal class.

Example 2: Calculate the modal age for the age distribution of 228 housewives.

Class Interval	Number of women
15 – 19	6
20 – 24	19
25 – 29	50
30 – 34	57
35 – 39	48
40 – 44	27
45 – 49	21
Total	228

Solution: By inspection, the mode lies in the fourth class, where $L = 29.5$, $f_{\text{mod}} = 57$, $f_1 = 50$, $f_2 = 48$, $w = 5$, and $\Delta_1 = 57 - 50 = 7$, $\Delta_2 = 57 - 48 = 9$.

Therefore, $\hat{x} = 29.5 + \frac{7}{7+9} \times 5 = 29.5 + 2.2 = 31.7$ years.

Merits of the Mode

-easy to calculate and understand

-can be computed when there is an open-ended class

Demerits of the Mode

- not rigidly defined by mathematical formula

-doesn't consider all the items

- not affected by extreme values
treatments

-not suitable for further mathematical

-can be used for qualitative data

-is an unstable average

Relationship between the Mean, Median and Mode

In a perfectly symmetrical distribution, they all coincide. $\Rightarrow \bar{X} = \text{median} = \text{mode}$.

If a distribution is moderately symmetrical, then **mean, median and mode** are related by the empirical formula: $\text{Mode} = 3\text{median} - 2\text{mean}$.

Example:-If the mode and mean of a moderately symmetrical distribution are 60 and 66, respectively, then what is the median value of this distribution?

Solution: $\text{mode} = 3\text{median} - 2\text{mean} \Rightarrow \text{Median} = \frac{1}{3}(\text{mode} + 2\text{mean}) = \frac{1}{3}(60 + 132) = 64$

A distribution having a tail in any one side is called a skewed distribution. Thus,

- if a distribution is skewed to the **right**, then $\text{mean} > \text{median} > \text{mode}$ and
- if a distribution is skewed to the **left**, then $\text{mean} < \text{median} < \text{mode}$.

When to Use the Different Averages?

Mean is appropriate if the data is quantitative and there is no extreme (abnormal) observation(s). For the data having extreme value(s) (or for qualitative data having ordinal measurement scale) it is better to use median as measure of central tendency. It is largely used measure of central tendency in psychology, education and other social sciences. On the other hand, mode is best measure of central tendency for qualitative data with nominal scale of measurement. It can also be used as a quick measure of central tendency for both qualitative and quantitative data.

5. Measures of Variation

5.1.Introduction

Definition: Measures of dispersion (variation) are statistical figures which indicate the spread or scatterings of a given distribution about the center. The highest figure indicates high inequality between the observations. Some of the common measures of dispersion are: Positional measures (range, quartile deviation); Mathematical Measures (mean deviation, standard deviation); and Measures of relative variation (Coefficient of Variation and Standard score). The positional measures and mathematical measures are known as absolute measures of variation.

5.2.Positional measures of Dispersion

5.2.1. The Range

The range is the simplest and crudest measure of variation.

For a discrete frequency distribution, **Range** = Largest value- Smallest value. $\Rightarrow R = L - S$.

Example:1 Data: 12,3,65,27,45,51,36,48,29,32,15

$$\therefore R = L-S \Rightarrow R = 65-3 = 62$$

Example2:

x_i	5	7	9	11	13	15	17	19
f_i	1	2	7	9	11	8	5	4

$$\therefore R = L-S = 19-5 = 14.$$

For a grouped discrete data, **Range** = upper class boundary of the highest class-lower class boundary of the lowest class or the difference between the largest class mark and the smallest class mark.

Example 3:

Grade	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	
Frequency	2	3	8	10	15	12	14	

$$\therefore R = L-S = 74-40 = 34.$$

The coefficient of range: it is the relative measure of variation, which is unit free

The coefficient of range is a **relative measure** of dispersion and is calculated by the formula

$$\frac{L - S}{L + S}.$$

For the above examples, 1. $\text{coef. } R = \frac{L - S}{L + S} = \frac{65 - 3}{65 + 3} = \frac{62}{68} = 0.912.$

$$2. \text{coef. } R = \frac{L - S}{L + S} = \frac{19 - 5}{19 + 5} = \frac{14}{24} = 0.583.$$

$$3. \text{coef. } R = \frac{L - S}{L + S} = \frac{74 - 40}{74 + 40} = \frac{34}{114} = 0.298.$$

Merits: easy to understand and calculate.

Demerits: Considers only the two extreme values and not possible to determine it in case of open-ended distributions.

Note! Range uses in quality control, weather forecasting . . .

5.2.2. Quartile Deviation (Q.D.)

The difference between the third quartile and the first quartile is called the inter quartile range (I.R.). Usually, the I.R. is reduced to the form of semi- inter quartile range (S.I.R.), which is called quartile deviation and it describes the variability of the middle 50% of the data.

$$\text{I.R.} = (Q_3 - Q_1), \quad \text{Q.D.} = \frac{Q_3 - Q_1}{2} \quad \text{and coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Example: For a certain continuous frequency distribution, the first and the last class limits are 0-9, and 60-69, and the lower and the upper quartiles are 26.32, and 39.25. Find a) the range; b) the I.R. and Q.D. c) coeff. Q.D.

Solution: a) Range = 69-0 = 69 or 64.5-4.5 = 60 or 69.5 - (-0.50) = 70.

b) I.R. = 39.25 - 26.32 = 12.93, and Q.D. = 12.93/2 = 6.465.

$$\text{c) coef. Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{26.32 - 39.25}{26.32 + 39.25} = \frac{12.93}{65.57} = 0.197.$$

Merits: easy to understand and calculate, not affected by extreme values, can be determined in case of open-ended distributions.

Demerits: not based on all the values.

5.3. Mathematical measures of dispersion

5.3.1. Mean Deviation (M.D.)

The mean Deviation (M.D.) is the arithmetic mean of the absolute deviations of the values from the mean or median.

If the deviations are taken from the mean then it is called M.D around the mean, if the deviations are taken from the median, it is called M.D around the median, or if the deviations are taken from the mode, it is called M.D around the mode.

$$\text{M.D. about the mean} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}, \quad \text{M.D. about the median} = \frac{\sum_{i=1}^n |x_i - \text{median}|}{n}$$

$$\text{M.D. about the mode} = \frac{\sum_{i=1}^n |x_i - \text{mode}|}{n}$$

$$\text{The coefficient of M.D.} = \frac{M.D._{\bar{x}}}{\bar{x}} \text{ or } \frac{M.D._{\tilde{x}}}{\text{median}} \text{ or } \frac{M.D._{\hat{x}}}{\text{mode}}$$

.

While dealing with population values, it is adjusted accordingly.

Example : Find the mean deviations around the mean and around the median of the following marks of 10 students (out of 50): 21, 23, 25, 28, 30, 32, 38, 39, 46, 48

Solution: The mean = 33 and median = 31. Then, we get

x_i	21	23	25	28	30	32	38	39	46	48	Total
$ x_i - 33 $	12	10	8	5	3	1	5	6	13	15	78
$ x_i - 31 $	10	8	6	3	1	1	7	8	15	17	76

Therefore, the M.D around the mean = $78/10 = 7.8$, and $76/10 = 7.6$ around the median. That means, the marks of the students deviate, on the average, by 7.8 from the mean 33, (or by 7.6 from the median 31).

Mean Deviation for Grouped data (discrete or continuous)

For grouped data, the mean deviations around the mean, median, and mode are obtained, respectively, as follows:

$$\frac{\sum_{i=1}^m f_i |x_i - \bar{x}|}{n}, \quad \frac{\sum_{i=1}^m f_i |x_i - \text{median}|}{n}, \quad \text{and} \quad \frac{\sum_{i=1}^m f_i |x_i - \text{mode}|}{n}$$

Where m = number of classes, and x_i = class mark of the i^{th} class.

Example 2: Calculate the two mean deviations for the data given below

Class Interval (C.I)	1-5	6-10	11-15	16-20
Frequency	4	1	2	3

Solution: Mean = $\frac{3 \times 4 + 8 \times 1 + 13 \times 2 + 18 \times 3}{10} = 10$, and Median = $5.5 + \frac{(5-4)5}{1} = 10.5$.

C.I	x_i	f_i	$f_i x_i - 10 $	$f_i x_i - 10.5 $
1-5	3	4	28	30
6-10	8	1	2	2.5
11-15	13	2	6	5
16-20	18	3	24	22.5
Total		10	60	60

Therefore, MD around the mean and around the median are $60/10 = 6$.

Merits: easy to understand and calculate, less affected by extreme values, considers every item.

Demerits: cannot be determined in case of open-ended distributions.

Note! M.D. is used in studying economic problems such as distribution of income and wealth in a society.

5.3.2. Variance and Standard Deviation

a) Population variance and S.D. (ungrouped data)

Definitions: **Variance** is the mean of the squared deviations about the A.M. and **Standard Deviation** is the positive square root of variance. S.D. is the most commonly used measure of dispersion and is also extremely useful in judging the representatives of the mean.

Suppose that $x_1, x_2, \dots, x_n$ are the values of the observation in a population of size N with mean μ . Then, the population variance and standard deviation are defined by:

$$\text{Variance} = \sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N} \dots\dots\dots (\#)$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{N}}$$

Example 1: Suppose that the ages of all patients in the recovery room of a certain Hospital are: 26,30,38,40,36,20,45, and 37 years. Find the population variance and standard deviation

Solution: Given $N = 8$, $x_1 = 26$, $x_2 = 30, \dots, x_8 = 37$.

The population mean is, $\mu = \frac{26+30+\dots+37}{8} = \frac{272}{8} = 34$.

Then, we construct the following table for the deviations and squared deviations from μ (the last column is for totals).

Age (x_i)	26	30	38	40	36	20	45	37	272
$x_i - \mu$	-8	-4	4	6	2	-14	11	3	0
$(x_i - \mu)^2$	64	16	16	36	4	196	121	9	462

Thus, using above formulae we have, $\sigma^2 = 462/8 = 57.75$; and $\sigma = \sqrt{57.75} = 7.599$.

$$\text{OR } \sigma^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2$$

The derivation of this computational formula is obtained by the above formulae (#):

$$\sum (x_i - \mu)^2 = \sum_{i=1}^N x_i^2 - \sum_{i=1}^N 2\mu x_i + \sum_{i=1}^N \mu^2 = \sum_{i=1}^N x_i^2 - 2\mu \sum_{i=1}^N x_i + N\mu^2$$

$$= \sum_{i=1}^N x_i^2 - 2N\mu^2 + N\mu^2 = \sum_{i=1}^N x_i^2 - N\mu^2. \text{ Hence, the resulting formula.}$$

Now, to solve Example 1 above using this formula, we have $\sum_{i=1}^8 x_i^2 = 9710$, and

$$\sum_{i=1}^8 x_i = 272 \Rightarrow \mu = 34, \sigma^2 = \frac{1}{8}(9710) - (34)^2 = 57.75$$

b) Population variance and S.D for grouped data

In a grouped discrete or continuous data where $X_1, X_2, \dots, X_m$ have their corresponding frequencies of $f_1, f_2, \dots, f_m$, the population variance is redefined as:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^m f_i (x_i - \mu)^2 \text{ for } N \text{ items and } m \text{ classes.}$$

Where x_i denote the class marks in case of continuous F.D.

Example 2.: Find the variance and S.D for the population values:

x_i	2	3	5	6	8
f_i	3	4	4	5	4

Solution: First find the mean,

$$\mu = \frac{1}{N} \sum f_i x_i = \frac{100}{20} = 5, \text{ and proceed as:}$$

x_i	f_i	$f_i x_i$	$(x_i - \mu)$	$(x_i - \mu)^2$	$f_i (x_i - \mu)^2$
2	3	6	-3	9	27
3	4	12	-2	4	16
5	4	20	0	0	0
6	5	30	1	1	5
8	4	32	3	9	36
Sum	20	100			84

Here, $N = \sum_{i=1}^5 f_i = 20$; then, using the definitional formula, the variance is

$$\sigma^2 = 84/20 = 4.2 \text{ and } \sigma = \sqrt{4.2} = 2.05.$$

c) Variance and S.D. for samples

Recall that a parameter is a summary measure computed from the population data. In the above cases, μ , σ , and σ^2 are the parameters. Whereas $\bar{x}$, s and s^2 are the statistics which are used to estimate the corresponding parameters. So far, the parameters σ^2 and σ have been discussed. Normally, one can say that if we replace N by n and μ by $\bar{x}$ the resulting formula would become

$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ and this would be used to estimate σ^2 . But, theoretically, it can be shown that this underestimates σ^2 . Instead, if we use the divisor to be $(n-1)$ instead of n , it will be an **unbiased estimator** of σ^2 . (i.e.; on the average, we will get a closer estimation). Thus, the sample variance for the values $x_1, x_2, \dots, x_n$ is given by:

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

Example 3: A sample of 5 students was taken from a class and their weights were found to be 48, 51, 52, 51, and 53kg. Find the variance and standard deviation.

Solution: The mean is $\bar{x} = 255/5 = 51$, and prepare the following table:

Weights (x_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
48	-3	9
51	0	0
52	1	1
51	0	0
53	2	4
Total	0	14

Thus, we have $n = 5$, and

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^5 (x_i - 51)^2}{5-1} = \frac{14}{4} = 3.5, \text{ and } s = \sqrt{3.5} = 1.87 \text{ is the sample standard}$$

deviation.

d) Sample variance and S.D for grouped data

If the values x_i have frequencies f_i ($i=1,2,\dots,m$), then the sample variance is given by:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^m f_i (x_i - \bar{x})^2$$

The above definition for sample variance also holds for grouped continuous distribution where x_i =class mark of the i^{th} class

Example 4: Find the sample variance and standard deviation for the distribution:

C.I	1-5	6-10	11-15	16-20
Freq.	4	1	2	3

Solution: In a continuous F.D., x_i is the class mark representing the i^{th} class.

C.I	x_i	f_i	$f_i x_i$	$f_i x_i^2$
1-5	3	4	12	36
6-10	8	1	8	64
11-15	13	2	26	338
16.20	18	3	54	972
Total		10	100	1410

Where $n = \sum f_i = 10$, $\bar{x} = \frac{\sum f_i x_i}{n} = \frac{100}{10} = 10$, $\sum f_i x_i^2 = 1410$, so that

$$s^2 = \frac{1}{9} (1410 - 10(10)^2) = \frac{410}{9} = 45.56, \text{ and } s = \sqrt{45.56} = 6.75.$$

Properties of S.D.

1. Adding or subtracting the same number on each observation doesn't change the S.D.

2. Multiplication and division on each number will affect the S.D. similarly.

Merits of the S.D.

-rigidly defined by mathematical formula

-is based on all the observations given

-is useful for further mathematical treatments. class

Demerits of the S.D.

-gives more weight for extreme values

-cannot be calculated if all the values are not

-cannot be used when there is an open-ended

-takes time to compute it

5.4. Measures of relative Variation

5.4.1. Coefficient of Variation (C.V.)

It is a relative measure of dispersion and is always a percentage. it is the best measure for comparing the variability of two series having different means and standard deviations.

$$\therefore C.V. = \frac{\sigma}{\mu} \times 100\% \quad \text{or} \quad C.V. = \frac{s}{\bar{x}} \times 100\% .$$

Example: Suppose typist A types out 30 pages per day on average with a S.D. of 6 and typist B types out 45 pages per day on average with a S.D. of 10. Which typist has shown greater consistency in her/his output?

$$\text{Solution: } C.V.(\text{Typist A}) = \frac{s}{\bar{x}} \times 100\% = \frac{6}{30} \times 100\% = 20\% .$$

$$C.V.(\text{Typist B}) = \frac{s}{\bar{x}} \times 100\% = \frac{10}{45} \times 100\% = 22.2\% .$$

$\therefore$ Typist A.

Exercise! Let $n = 10$, $\bar{x} = 12$ and $\sum x_i^2 = 1530$ for a certain data. Find the C.V.

4.4.2. Standard Scores (Z-Scores)

Standard scores is one of the applications of standard deviation which is a relative measure. It indicates the position of individual observations. Suppose that a student scored 66 in Statistics and 80 in Mathematics. The question is, in which course did he score better as compared to his classmates?

At first glance, it seems that he did much better in Mathematics. But suppose that all the students in the class averaged 51 points in statistics with a standard deviation of 12, and averaged 72 in Mathematics with a standard deviation of 16.

Thus, one can argue that the student's score in Statistics is: $\frac{66-51}{12} = 1.25$ standard deviations above the average, while his score in Mathematics is only $\frac{80-72}{16} = 0.50$ standard deviation above the average for the class. Here, the grades have been converted in to standard scores. Where as the original scores cannot be meaningfully compared, the standard scores expressed in terms of standard deviations can be compared. Thus, the student scored much higher in Statistics than in Mathematics compared to the rest of the class.

In general, we define the standard score as:

$$z = \frac{x - \bar{x}}{s} \text{ or } Z = \frac{x - \mu}{\sigma}$$

The Z- score tells us how many standard deviations a value lies above (if positive) or below (if negative) the mean of the set of data to which it belongs.

Example: If a set of measurements has the mean 48 with a S.D. of 12, convert each of the following in to

standard units: a) 54; b) 72 ; c) 78.

Solution: a) For $x = 54$, $z = \frac{54-48}{12} = -0.5$; that is, 54 is 0.5 S.D's below the mean .

b) For $x = 72$, $z = \frac{72-48}{12} = 2$; that is, 72 is 2 S.D's above the mean.

c) Similarly, 78 is 2.5 standard deviations above the mean.

Exercise! Convert the numbers 6,7,5,10,12 into standard scores.

6. Measures of Skewness and Kurtosis

6.1.Skewness

When referring to the shape of frequency or probability distributions, “skewness” refers to asymmetry of the distribution. More precisely, it measures lack of symmetry.

A distribution with the bell shaped and the mean= median =mode of the distribution is known as the symmetric distribution (or the normal distribution)

A distribution with an asymmetric tail extending out to the right is referred to as “positively skewed” or “skewed to the right”. In case the mean > the median > the mode

A distribution with an asymmetric tail extending out to the left is referred to as “negatively skewed” or “skewed to the left.” In this case, the mean < the median < the mode of the distribution

6.2. The measures of Skewness

The Pearsonian measure of skewness

$sk = \frac{3(\bar{X} - \text{median})}{s}$. If $sk > 0$, the distribution is right skewed. If $sk < 0$, it is left skewed and if $sk = 0$, it is normal distribution

Skewness has also been defined with respect to the **third moment** about the mean:

$\gamma_1 = \frac{\sum (X - \mu)^3}{n\sigma^3}$, which is simply the expected value of the distribution of cubed z scores.

Skewness measured in this way is sometimes referred to as “Fisher’s skewness.” When the deviations from the mean are greater in one direction than in the other direction, this statistic will deviate from zero in the direction of the larger deviations

Great skewness may motivate the researcher to investigate outliers. When making decisions about which measure of location to report (means being drawn in the direction of the skew) and which inferential statistic to employ (one which assumes normality or one which does not), one should take into consideration the estimated skewness of the population. Normal distributions have zero skewness. Of course, a distribution can be perfectly symmetric but far from normal. Transformations commonly employed to reduce (positive) skewness include square root, log, and reciprocal transformations.

6.3.Kurtosis

It measures the degree of peakdenss or flatness of a distribution at the top. There are three types of kurtosis:

- a. mesokurtic: it is normal of symmetric at the top (i.e., equally flat-topped distribution) when compared to the normal distribution.
- b. Platykurtic: it is flat at the top (i.e., more flat-topped distribution) when compared to the normal distribution.
- c. Leptokurtic: if is peak at the top (i.e., less flat-topped distribution) when compared to the normal distribution.

Karl Pearson (1905) defined a distribution's degree of kurtosis as $\eta = \beta_2 - 3$, where

$$\beta_2 = \frac{\sum(X - \mu)^4}{n\sigma^4},$$

He referred to more flat-topped distributions ($\eta < 0$) as “platykurtic,” less flat-topped distributions ($\eta > 0$) as “leptokurtic,” and equally flat-topped distributions as “mesokurtic” ($\eta \approx 0$).

7. Simple Correlation and Regression Analysis

7.1.Simple Correlation Analysis

Definition

Simple correlation coefficient (r) measures the degree of linear relationship (or association) between two variables only. Simple correlation coefficient (r) varies from negative one to positive one (i.e., $-1 \leq r \leq 1$).

$r = -1$ means there is perfect negative linear association between two variables

$r = 1$ means there is perfect positive linear association between two variables

$r = 0$ means there is no linear association between two variables

Note that the existence of correlation does not necessarily imply **causation**. However, causation necessarily leads to correlation. In general, correlation may be due to one of the following facts:

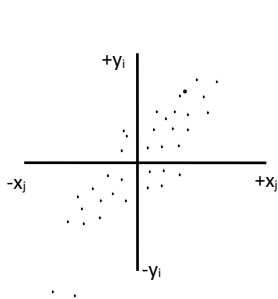
- a. One variable being the cause of the other variable. E.g. money supply causes inflation
- b. Both variables being the result of a common cause. E.g. yield of maize and sorghum may be positively correlated due to the fact that both are related to rainfall
- c. Chance factor

Limitations of simple correlation coefficient

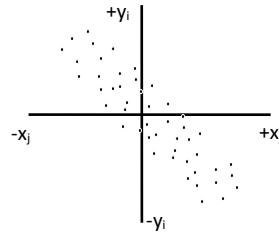
1. it considers only two variables and if the 3rd, 4th, etc variables exist, simple correlation coefficient is not appropriate
- 2.
3. it does not capture non-linear relationship between two variables

7.2.Measures of Simple Correlation

1. Scatter Diagrams



positive
correlation



negative
correlation

2. Karl Pearson's Correlation coefficient (r)

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$$r = \frac{\sum_{i=1}^n (x_i)(y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2 \cdot \sum_{i=1}^n (y_i)^2}}$$

The second equation is in deviation form.

3. The Rank Correlation Coefficient (ρ): it is used when our variables are qualitative in their nature.

$$\rho = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}$$

Where n=# of pairs ranks

D= the absolute difference between the ranks of an item.

7.3.Simple Linear Regression Analysis

Economics is often concerned with the relationship between economic variables. **Regression analysis** is basically concerned with examining these relationships.

Regression analysis is one of the most commonly used tools in econometric work.

What is Regression Analysis?

Regression Analysis is concerned with the study of the dependency of one variable, the **dependent or explained variable**, on one or more other variables called the **explanatory variable** or the **independent variables**.

EXAMPLE: $g_d = f(P)$ *demand function*
 $g_s = f(P, \text{Tax}, \text{Tech.})$ *supply function*

Conventionally the explained variable is designated by **Y** and the explanatory variables by **X_i**s.

$$Y = f(X_1, X_2, \dots, X_K)$$

⇒ If **K=1**, that is, there is only one X-variable, we have the **simple Regression**.

But

⇒ If **K>1**, i.e. more than one X-variable, we have the **multiple regression** case.

EXAMPLE: *If sales volume is only affected/influenced by advertising expenditure, then the regression is defined as:*

$Y = \text{sales}$
 $X = \text{advertising expenditure}$

$Y = f(X)$ simple Regression

But if: $Y = \text{consumption expenditure of a family is affected by } X_1, x_2, \text{ and } X_3.$

$X_1 = \text{family income}$
 $X_2 = \text{financial assets of the family}$
 $X_3 = \text{family size}$

We have the case of multiple regression.

$$Y = f(X_1, X_2, X_3) \quad \text{multiple Regression}$$

Estimation in the Simple Regression Model

Once the variables have been identified and the data are collected, the next step in regression analysis is to estimate the coefficients of the equation.

Thus it is often said that **the bread and butter** of regression analysis is **the estimation of the parameters** of the econometric model.

⇒ In other words, the main purpose of regression analysis is to take a purely theoretical equation like:

$$\Rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

and use a set of data to create an estimated equation like:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

⇒ In a sense, **Estimation is what regression is all about.**

The numerical values of the parameters (coefficients) of a regression equation can be estimated using different methods.

- The method of moments
- The ordinary least square (OLS) method
- The maximum likelihood method.

The most widely used method of obtaining the parameter estimates from a sample is called **ordinary least squares (OLS)**.

Method of Ordinary Least Squares (OLS)

The Ordinary Least Squares (OLS) was developed by Friedrich Gauss.

Starting point:

Population regression function (PRF): $Y_i = \alpha + \beta X_i + \varepsilon_i$

Sample regression function (SRF): $Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$ but $Yhat = \hat{\alpha} + \hat{\beta} X_i$

$$\Rightarrow Y_i = Yhat + e_i$$

$$\Rightarrow e_i = Y_i - Yhat$$

Thus, the residual is the difference between the actual and the estimated/predicted values of Y.

Now, given data on both X and Y, our objective is to find the SRF that best “fits” the PRF or the values of estimators, which are as close as possible to population parameters. However, what is the possible criterion to choose the SRF that best “fits” the actual PRF?

Criterion I: Minimize the sum of residuals

Choose the SRF in such a way that the sum of the residuals (i.e., $\sum_{i=1}^n e_i$) is as small as possible. If

the sum the residuals is zero (i.e., $\sum_{i=1}^n e_i = 0$), we will have the method of moments. However, this criterion is not good as it gives equal weight to all kinds of residuals (large, medium and small).

Criterion II: Sum Least Squares Criterion

According to this criterion, find the SRF, which minimize the sum of the squared residual (i.e.,

min. $\sum_{i=1}^n e_i^2$). Alternatively, we choose $\hat{\alpha}$ and $\hat{\beta}$ such that the sum of squares of errors is minimized. This criterion is very important for two basic reasons:

- a. it gives more weight to larger residuals and less weight to smaller residuals
- b. the estimators obtained through OLS method of estimation have some desirable statistical properties

Since different samples from a given population result in different estimators, which of these estimators minimize the sum of squared residuals? (Recall how to find extreme values of functions- the necessary and sufficient conditions).

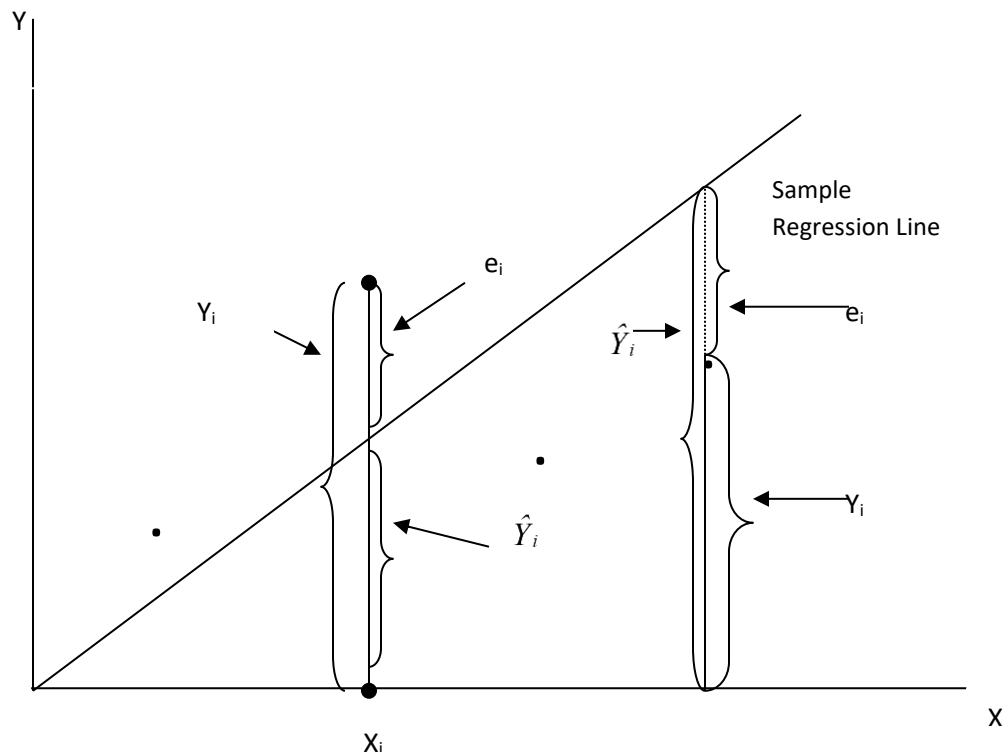
Now, given the population regression function as:

$Y_i = \alpha + \beta X_i + \varepsilon_i$, we want to substitute $\hat{\alpha}$ and $\hat{\beta}$ for α and β , and minimise

$$S = \sum_{i=1}^n \left(Y_i - \hat{\alpha} - \hat{\beta} X_i \right)^2$$

$$= \sum_{i=1}^n e_i^2$$

The intuitive idea behind the procedure of least squares is given by looking at the following figure.



The regression line passes through points in such a way that it is ‘as close as possible’ to the points of the data. Closeness could mean different things. The minimisation procedure of the OLS implies that we minimise the sum of squares of the vertical distances of the points from the line.

The Necessary Condition

$$\frac{\partial S}{\partial \hat{\alpha}} = \frac{\partial S}{\partial \hat{\beta}} = 0$$

This equation is known as the first order conditions for minimisation. This procedure yields:

$$\begin{aligned}
\frac{\partial S}{\partial \hat{\alpha}} &= 0 \Rightarrow \sum_{i=1}^n 2(Y_i - \hat{\alpha} - \hat{\beta} X_i)(-1) = 0 \Rightarrow \sum_{i=1}^n e_i = 0 \\
&\Rightarrow \sum_{i=1}^n Y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^n X_i \\
&\Rightarrow \bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X} \\
&\Rightarrow \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}
\end{aligned}
\tag{EQ 1}$$

$$\begin{aligned}
\frac{\partial S}{\partial \hat{\beta}} &= 0 \Rightarrow \sum_{i=1}^n 2(Y_i - \hat{\alpha} - \hat{\beta} X_i)(-X_i) = 0 \Rightarrow \sum_{i=1}^n X_i e_i = 0 \\
&\Rightarrow \sum_{i=1}^n X_i Y_i = \hat{\alpha} \sum_{i=1}^n X_i + \hat{\beta} \sum_{i=1}^n X_i^2
\end{aligned}
\tag{EQ 2}$$

Equations 1 and 2 are known as the normal equations. To solve for $\hat{\beta}$, we substitute for $\hat{\alpha}$ in equation 2 from equation 1 and we get

$$\begin{aligned}
\sum_{i=1}^n Y_i X_i &= (\bar{Y} - \hat{\beta} \bar{X}) \sum_{i=1}^n X_i + \hat{\beta} \sum_{i=1}^n X_i^2 \\
\sum_{i=1}^n Y_i X_i &= \bar{Y} \sum_{i=1}^n X_i - \hat{\beta} \bar{X} \sum_{i=1}^n X_i + \hat{\beta} \sum_{i=1}^n X_i^2 \\
\sum_{i=1}^n Y_i X_i - \bar{Y} \sum_{i=1}^n X_i &= \hat{\beta} \left(\sum_{i=1}^n X_i^2 - \bar{X} \sum_{i=1}^n X_i \right) \\
\sum_{i=1}^n Y_i X_i - \frac{1}{n} \sum_{i=1}^n Y_i \sum_{i=1}^n X_i &= \hat{\beta} \left(\sum_{i=1}^n X_i^2 - \frac{1}{n} \sum_{i=1}^n X_i \sum_{i=1}^n X_i \right) \\
\frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n} &= \hat{\beta} \left(\frac{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}{n} \right) \\
\Rightarrow \hat{\beta} &= \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}
\end{aligned}$$

Numerator of the last Equation can be simplified as follows: First, we take the numerator to get

$$\begin{aligned}
n \sum_{i=1}^n Y_i X_i - \sum_{i=1}^n Y_i \sum_{i=1}^n X_i &= n \sum_{i=1}^n Y_i X_i - n \bar{Y} n \bar{X} \\
&= n \left(\sum_{i=1}^n Y_i X_i - n \bar{Y} \bar{X} \right) \\
&= n \left(\sum_{i=1}^n (Y_i X_i - \bar{Y} \bar{X}) \right) \\
&= n \left(\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) \right)
\end{aligned}$$

Similarly the denominator of our last equation can be rewritten as:

$$\begin{aligned}
n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2 &= n \sum_{i=1}^n X_i^2 - (n \bar{X})^2 \\
&= n \left(\sum_{i=1}^n X_i^2 - (n \bar{X})^2 \right) \\
&= n \sum_{i=1}^n (X_i - \bar{X})^2
\end{aligned}$$

Therefore, we have:

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}, \text{ where } x_i = X_i - \bar{X} \text{ \& } y_i = Y_i - \bar{Y} \text{ the deviation form}$$

Note also that we can write this equation can be written as

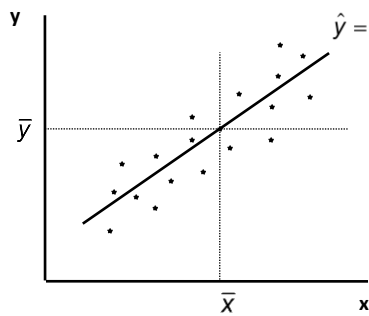
$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / (n-1)}{\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)} = \frac{Cov(X, Y)}{Var(X)}$$

The estimators obtained thus are known as the **Least Square Estimators**.

- (1) The estimators are expressed solely in terms of the observable quantities (sample)
- (2) They are point estimators: $\Rightarrow$ they provide a single (point) value of the relevant population parameter

The regression line thus obtained has the following properties:

1) It passes through the sample means of y and x.



$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_i \text{ since } \sum y_i = N\hat{b}_0 + \hat{b}_1 \sum x_i$$

dividing both sides by N

$$\Rightarrow \bar{y} = \hat{b}_0 + \hat{b}_1 \bar{x}$$

2) The mean value of the estimated y ($= \hat{y}_i$) is equal to the mean value of the actual y for

$$\begin{aligned} \bar{\hat{y}}_i &= \hat{b}_0 + \hat{b}_1 \bar{x}_i \\ &= (\bar{y} - \hat{b}_1 \bar{x}) + \hat{b}_1 \bar{x}_i \\ &= \bar{y} - \hat{b}_1 (\bar{x} - \bar{x}) \end{aligned}$$

Summing both sides of this last quantity over the sample values and dividing through by the sample values N gives $\bar{\hat{y}} = \bar{y}$ since $\sum (\bar{x} - \bar{x}) = 0$

3) The mean value of the residuals is equal to zero: $\sum e_i = 0$

4) The residuals are uncorrelated with the predicted $\hat{y}_i$: $\sum \hat{y}_i e_i = \hat{b}_1 \sum x_i e_i$

5) The residuals U_i are uncorrelated with x_i : $\sum e_i x_i = 0$

8. Introduction Probability Theory

8.1. Basic Concepts

Typically, the starting point of an investigation is an **experiment**. This may involve either a simple rolling of a pair of dice or as complicated as conducting a large-scale survey of households or firms.

a. Experiment: An experiment is defined as a process of observation or measurement.

Examples: 1. Tossing a coin

2. Rolling a die

3. Counting defective products of a company etc.

An experiment is said to be a **Random experiment** if the following three conditions are satisfied:

1. all possible distinct outcomes are known ahead of time
2. the outcome of a particular trial is not known a priori
3. the experiment can be duplicated, in principle, under ideal conditions.

b. Outcomes: The results one obtains from an experiment are known as outcomes. There are certain regularities that one may observe outcomes from an experiment; however, it is not possible to predict them. Example the outcomes are head or tail in tossing a coin only once

c. Sample space: The set of all possible outcomes (totality) of an experiment is referred to as the sample space. It is usually denoted by the letter S.

Examples: 1. In tossing a coin once the sample space is: $S = \{H, T\}$. Tossing the coin twice the sample space is $S = \{HH, HT, TH, TT\}$. Tossing the coin three times the sample space is:

$\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. In general tossing coin n times, the

number of elements in the sample space is 2^n

2. In rolling a die, the sample space is: $S = \{1, 2, 3, 4, 5, 6\}$. In rolling two dice at the same time $S = \{(x, y) \mid x = 1, 2, \dots, 6; y = 1, 2, \dots, 6\}$. The number of elements is $6 \times 6 = 36$. In general, rolling n dice at a time the number of elements is 6^n

Types Sample Space

- **Discrete Sample Space:** If the sample space contains a finite, or an infinite but countable number of elements, with a one to one correspondence to positive integers, it is called a discrete space.
- **Continuous Sample space:** If the sample space has infinite and uncountable number of elements (i.e., if it has as many elements as there are real numbers) it is said to be continuous.

d. Sample point or elementary event: Elementary events are singletons in a sample space or each distinct individual element of the sample space is known as the sample point (or elementary event).

Example: If a coin is tossed twice, the sample space consists of four sample points HH, HT, TH, and TT; $S = \{HH, HT, TH, TT\}$, where H = heads, and T = tails.

e. Events: Any subset of the sample space (which is also a set of sample points) that represents several outcomes of an experiment is known as an event.

Example: 1. in tossing two coins at the same time, event A contains at least one head.

$$A = \{HH, HT, TH\}$$

2. when a pair of dice is rolled, “total score of 7” is an event represented by the sample points in A as follows: $A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$, where the first number is the outcome of the first die and the second number is the outcome of the second die.

f. Null (impossible) event: the null or impossible event is represented by $\emptyset$. It refers to the event that never occurs at all.

g. Mutually Exclusive Events: Events **A** and **B** are said to be mutually exclusive if the occurrence of A precludes the occurrence of B and vice-versa. They cannot occur together. That is, $A \cap B = \emptyset$

h. Equally Likely Events: the outcomes of a random event are said to be equally likely iff occurrence of none of them can be expected in preference to another.

i. Collectively Exhaustive: two events A and B are said to be collectively exhaustive if they contain all possible outcomes of the sample space.

j. Independent Events: Events **A** and **B** are said to be independent if the occurrence of A has no effect on the occurrence of B and vice-versa.

k. Complement of an Event: If event A is a subset of sample space S, then its complement is **A'**, which contains elements in S but not in A. Thus, complement is S-A. In other words, for each set **A** we can define a unique set **A'** such that

$$A \cap A' = \emptyset \text{ and } A \cup A' = S.$$

8.2. Definitions of probability: Classical, Relative Frequency and Subjective Definitions

a. Classical probability concept. This concept applies when all possible outcomes are equally likely, as is presumably the case in most games of chances.

Definition: If a random experiment can result in N collectively exhaustive, mutually exclusive and equally likely outcomes and if “n” of these outcomes result in the occurrence of event A (or favourable outcomes or success), then the probability of event A is defined by: $P(A) = \frac{n}{N}$. That is, the probability of an event is given by the number of favourable outcomes divided by the total number of possible outcomes.

Where P(A) is the probability of event A or probability that event A occurs.

Example1: Toss a fair coin twice, then the sample space $S = \{HH, HT, TH, TT\}$; $N=4$

Let the event A be “observing at least one head”, then $A = \{HH, HT, TH\}$; $n = 3$

$$\therefore P(A) = \frac{3}{4}$$

Example 2: when a pair of dice is rolled, event A represents “total score of 7”. $N=36$ and $n=6$

$$\therefore P(A) = \frac{6}{36} = \frac{1}{6}$$

b. A Posteriori or Frequency interpretation or statistical interpretation of probability or empirical probability: suggested in an attempt to tackle the problem posed by the classical approach. The basic argument of the frequency approach is that probability does not have to be restricted to the situations of equally likely since the notion of probability should be interpreted as stemming from the observable stability of empirical frequencies. Let N_A be the number of times the same event (say event A) occurs in large number of N trials of an experiment. If there exists a real number p such that $p = \lim_{N \rightarrow \infty} \frac{N_A}{N}$, then p is called the probability of event A and is denoted by $P(A)$.

NB: The probability of an event, according to this definition, is the proportion of that event of the same kind will occur in the long run.

Despite improvement over the classical approach, there are some objections: what is meant by limit as N goes to infinity? How can we generate infinite sequence of trials? What happens to phenomena where repeated trials are not possible?

c. Subjective Definition: According to this approach, probability refers to the “degree of belief” of individuals in assessing the uncertainty of a particular situation. Since there is no past experience, it depends on educated guesses. The two approaches lead to alternative methods of statistical inference. The frequentist will conduct the discussion around what happens “in the long-run” or

“on average”, and attempt to develop “objective” procedures. On the other hand, a subjectivist will be concerned with the question of revising prior beliefs in the light of available information in the form of the observed data.

8.3. The Axioms of Probability and Some Theorems of Probability

Axiom 1: For an event A in S , the probability of A is a nonnegative real number. That is, $P(A) \geq 0$.

Axiom 2: The probability of the sample space S is unity, i.e.,

$$P(S) = 1.$$

Axiom 3: If $A_1, A_2, \dots$ is a finite or infinite sequence of mutually exclusive events in S then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Theorem 1.1: If A is an event in a discrete sample space S , then $P(A)$ equals the sum of the probabilities of the individual outcomes comprising A .

Theorem 1.2: If an experiment can result in any one of the N different equally likely outcomes, and n of these outcomes together constitute event A , then the probability of event A is given by

$$P(A) = \frac{n}{N}.$$

Note: $P(A) = 0$ does not imply $A = \emptyset$

Theorem 1.4: If A and B are events in a sample space S and $A \subset B$, then $P(A) \leq P(B)$ (Monotonicity of the probability function).

Theorem 1.5: If A and A^c are complementary events in a sample space S , then, $P(A^c) = 1 - P(A)$.

Theorem 1.6: $0 \leq P(A) \leq 1$

Theorem 1.7: If A and B are two events in a sample space S , then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Counting Procedures

Fundamental Rule of Counting- the multiplication Rule

Theorem 1: If an operation can be performed in n_1 different ways and if for each of these the second operation can be performed in n_2 ways, then the two operations when associated together can be performed in $n_1 n_2$ ways.

Example: How many sample points are in the sample space when a pair of dice is thrown once?

$$n_1 = 6, n_2 = 6 \quad \therefore n_1 n_2 = 36$$

Theorem 2: If an operation can be performed in n_1 ways and if for each of these the second operation can be performed in n_2 ways, and for each of the first two the third operation can be performed in n_3 ways, etc..., then the sequence of k operations can be performed in $n_1 n_2 \dots n_k$ ways.

Example: A college freshman must take an economics course, a management course, and a statistics course. If she may select any one of 3 economics courses, any of 4 management courses and any of 2 statistics courses, in how many ways can she arrange her program?

$$n_1 = 3, n_2 = 4, n_3 = 2 \quad \therefore n_1 n_2 n_3 = 24$$

Permutations: Refers to the ordered arrangements

The number of permutations is the product of n consecutively positioned integers. Since this number arises in many problems in mathematics, it is given a special name. If n is a positive integer we define

$n! = n(n-1)(n-2) \dots (2)(1)$ and is read as n factorial.

E.g. $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

$$2! = 2 \times 1 = 2$$

and $0! = 1$ and $1! = 1$ by definition.

Theorem: The number of ordered arrangements or permutations of n distinct objects taken all at a time (with out repetition) is given by $n!$ and denoted by ${}_nP_n = n!$

Theorem: The number of permutations of n distinct objects taken r at a time (with out repetition) is given by

$$\begin{aligned} {}_nP_r &= n(n-1)(n-2) \dots (n-r+1) \\ &= \frac{n!}{(n-r)!} \end{aligned}$$

Combinations – refers to unordered arrangements (i.e., no concern about the order of the arrangements)

The number of ways of choosing r objects from a total of n distinct objects, called the number of combinations of n objects taken r at a time, is given by (order is not important in this case).

$${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}.$$