LECTURE NOTES OF STAT 3202

Dr. Pratheesh P. Gopinath

2022-01-30

Contents

Welcome

Welcome to the book **LECTURE NOTES ON STATISTICAL METHODS AND APPLICATIONS**.

Preface

Note: This book is published in MeLoN (Module for e-Learning & Online Notes) . The online version of this book is free to read here.

If you have any feedback, please feel free to contact Dr.Pratheesh P. Gopinath. E-mail: `pratheesh.pg@kau.in` Thank you!

This book is a collection of all lecture notes covering the syllabus of statistics course in B.Sc.(Hons.) Agriculture under Kerala Agricultural University

1

Introduction

In this lecture we will have the introduction, which includes, definition of statistics, collection and classification of data, formation of frequency distribution.(?) (?)

1.1 Origin of the word "Statistics"

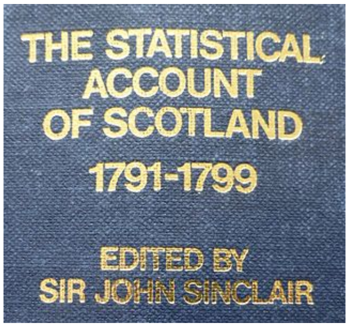
The term statistics was derived from the Neo-Latin word `statisticum collegium` meaning "council of state" and the Italian word `statista` meaning "statesman" or "politician".

A German word `Statistik`, got the meaning "collection and classification of data" generally in the early 19th century. This word was first introduced by Gottfried Achenwall (1749). `Statistik` was originally designated as a term for analysis of data about the state (data used by government or other administrative bodies). The term `Statistik` was introduced into English in 1791 by Sir John Sinclair when he published the first of 21 volumes titled "Statistical Account of Scotland" (?). The first book to have 'Statistics' in its title was "Contributions to Vital Statistics" (1845) by Francis GP Neison, actuary[1] to the Medical Invalid and General Life Office.

1.2 Statistics and Mathematics

Mathematics follows a rigid theorem and proof. Mathematical theories involve well-defined and proven facts which has the minimal scope of change. However, Statistics is a discipline where real-life data is handled. This factor makes this

[1] actuary: A person who compiles and analyses statistics and uses them to calculate insurance risks and premiums.



THE STATISTICAL
ACCOUNT
OF SCOTLAND
1791-1799

EDITED BY
SIR JOHN SINCLAIR

Figure 1.1: Statistical Account of Scotland by Sir John Sinclair (1791)

field of study more abstract, where individuals have to develop newer solutions to problems that was new and not observed before. Statistics is an applied science; in mathematics the goal is to prove theorems. In statistics, the main goal is to develop good methods for understanding data and making decisions. Statisticians often use mathematical theorems to justify their methods, but theorems are not the main focus. Statistics is now considered as an independent field which uses mathematics to solve real life problems.

1.3 Definition of Statistics

Statistics is the science which deals with the

- Collection of data
- Organization of data or Classification of data
- Presentation of data
- Analysis of data
- Interpretation of data

Two main branches of statistics are:

Descriptive statistics, which deals with summarizing data from a sample using indexes such as the mean or standard deviation etc.

Inferential statistics, use a random sample of data taken from a population to describe and make inferences about the population parameters.

1.4 Data

Data can be defined as individual pieces of factual information recorded and used for the purpose of analysis. It is the raw information from which inferences are drawn using the science "STATISTICS".

Example for data

- No. of farmers in a block.
- The rainfall over a period of time.
- Area under paddy crop in a state.

1.5 Use and limitations of statistics

Functions of statistics: Statistics simplifies complexity, presents facts in a definite form, helps in formulation of suitable policies, facilitates comparison and helps in forecasting. Valid results and conclusion are obtained in research experiments using proper statistical tools.

Uses of statistics: Statistics has pervaded almost all spheres of human activities. Statistics is useful in the administration, Industry, business, economics, research workers, banking,insurance companies etc.

Limitations of Statistics

- Statistical theories can be applied only when there is variability in the experimental material.
- Statistics deals with only aggregates or groups and not with individual objects.
- Statistical results are not exact.
- Statistics are often misused.

1.6 Population and Sample

Consider the following example.Suppose we wish to study the body masses of all students of College of Agriculture, Vellayani. It will take us a long time to measure the body masses of all students of the college and so we may select 20 of the students and measure their body masses (in kg). Suppose we obtain the measurements like this

49 56 48 61 59 43 58 52 64 71 57 52 63 58 51 47 57 46 53 59

In this study, we are interested in the body masses of all students of College of Agriculture, Vellayani. The set of body masses of all students of College of Agriculture, Vellayani is called the **population** of this study. The set of 20 body masses, $W = \{49, 56,48, ..., 53, 59\}$, is a **sample** from this population.

1.6.1 Population

A population is the set of all objects we wish to study

1.6.2 Sample

A sample is part of the population we study to learn about the population.

1.7 Variables and constants

1.7.1 Variables

Any type of observation which can take different values for different people, or different values at different times, or places, is called a variable. The following are examples of variables:

- family size, number of hospital beds, number of schools in a country, etc.
- height, mass, blood pressure, temperature, blood glucose level, etc.

Broadly speaking, there are two types of variables – **quantitative** and **qualitative** (or categorical) variables

1.7.2 Constants

Constants are characteristics that have values that do not change. Examples of constants are: pi ( ) = the ratio of the circumference of a circle to its diameter (= 3.14159...) and *e*, the base of the natural or (Napierian) logarithms (e=2.71828).

1.8 Types of variables

1.8.1 Quantitative variables

A quantitative variable is one that can take numerical values. The variables like family size, number of hospital beds, number of schools in a country, height, mass, blood pressure, temperature, blood glucose level, etc. are examples of quantitative variables. Quantitative variables may be characterized further as to whether they are discrete or continuous

1.8.2 Discrete variables

The variables like family size, number of hospital beds, number of schools in a country, etc. can be counted. These are examples of discrete variables. Variables that can only take on a finite number of values are called ”discrete variables.” Any variable phrased as “the number of ...”, is discrete, because it is possible to list its possible values {0,1, ...}. Any variable with a finite number of possible values is discrete. The following example illustrates the point. The number of daily admissions to a hospital is a discrete variable since it can be represented by a whole number, such as 0, 1, 2 or 3. The number of daily admissions on a given day cannot be a number such as 1.8, 3.96 or 5.33.

1.8.3 Continuous variables

The variables like height, mass, blood pressure, temperature, blood glucose level, etc. can be measured. These are examples of continuous variables. A continuous variable does not possess the gaps or interruptions characteristic of a discrete variable. A continuous variable can assume any value within a specific relevant interval of values assumed by the variable. Notice that age is continuous since an individual does not age in discrete jumps. Weight can be measured as 35.5, 35.8 kg etc so, it is a continuous variable.

1.8.4 Categorical variables

A variable is called categorical when the measurement scale is a set of categories. For example, marital status, with categories (single,married, widowed), is categorical. Whether employed (yes, no), religious affiliation (Protestant, Catholic, Jewish, Muslim, others, none), colours etc. Categorical variables are often called qualitative. It can be seen that categorical variables can neither be measured nor counted.

1.9 Measurement scales

Variables can further be classified according to the following four levels of measurement: nominal, ordinal, interval and ratio.

1.9.1 Nominal scale

This scale of measure applies to qualitative variables only. On the nominal scale, no order is required. For example,gender is nominal, blood group is nominal, and marital status is also nominal. We cannot perform arithmetic operations on data measured on the nominal scale.

1.9.2 Ordinal scale

This scale also applies to qualitative data. On the ordinal scale, order is necessary. This means that one category is lower than the next one or vice versa. For example, Grades are ordinal, as excellent is higher than very good, which in turn is higher than good, and so on. It should be noted that, in the ordinal scale, differences between category values have no meaning.

1.9.3 Interval scale

This scale of measurement applies to quantitative data only. In this scale, the zero point does not indicate a total absence of the quantity being measured. An example of such a scale is temperature on the Celsius or Fahrenheit scale. Suppose the minimum temperatures of 3 cities, A, B and C, on a particular day were 0^0C, 20^0C and 10^0C, respectively. It is clear that we can find the differences between these temperatures. For example, city B is 20^0C hotter than city A. However, we cannot say that city A has no temperature. Moreover, we cannot say that city B is twice as hot as city C, just because city B is 20^0C and city C is 10^0C. The reason is that, in the interval scale, the ratio between two numbers is not meaningful.

1.9.4 Ratio scale

This scale of measurement also applies to quantitative data only and has all the properties of the interval scale. In addition to these properties, the ratio scale has a meaningful zero starting point and a meaningful ratio between 2 numbers. An example of variables measured on the ratio scale, is weight. A weighing scale that reads 0 kg gives an indication that there is absolutely no weight on it. So the zero starting point is meaningful. If Ram weighs 40 kg and Laxman weighs 20 kg, then Ram weighs twice as Laxman. Another example of a variable measured on the ratio scale is temperature measured on the Kelvin scale. This has a true zero point.

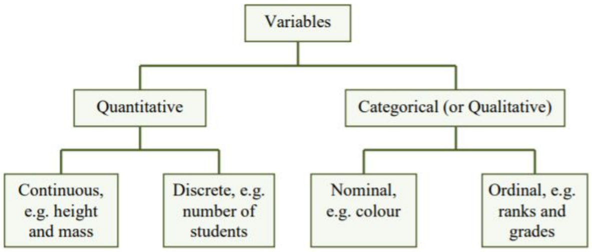

Variables
Quantitative
Categorical (or Qualitative)
Continuous, e.g. height and mass
Discrete, e.g. number of students
Nominal, e.g. colour
Ordinal, e.g. ranks and grades

Figure 1.2: Classification of variables

1.10 Collection of Data

The first step in any enquiry (investigation) is the collection of data. The data may be collected for the whole population or for a sample only. It is mostly collected on a sample basis. Collecting data is very difficult job. The enumerator or investigator is the well trained individual who collects the statistical data. The respondents are the persons from whom the information is collected.

1.10.1 Types of Data

There are two types (sources) for the collection of data: * Primary Data
* Secondary Data

1.10.1.1 Primary Data

Primary data are the first hand information which is collected, compiled and published by organizations for some purpose. They are the most original data in character and have not undergone any sort of statistical treatment.

Example: Population census reports are primary data because these are collected, complied and published by the population census organization.

1.10.1.2 Secondary Data

The secondary data are the second hand information which is already collected by an organization for some purpose and are available for the present study. Secondary data are not pure in character and have undergone some treatment at least once.

Example: An economic survey of England is secondary data because the data are collected by more than one organization like the Bureau of Statistics, Board of Revenue, banks, etc.

1.11 Methods of Collecting Primary Data

Primary data are collected using the following methods:

1.11.1 Personal Investigation

The researcher conducts the survey him/herself and collects data from it. The data collected in this way are usually accurate and reliable. This method of collecting data is only applicable in case of small research projects.

1.11.2 Through Investigation

Trained investigators are employed to collect the data. These investigators contact the individuals and fill in questionnaires after asking for the required information. Most organizations utilize this method.

1.11.3 Collection through Questionnaire

Researchers get the data from local representations or agents that are based upon their own experience. This method is quick but gives only a rough estimate.

1.11.4 Through the Telephone

Researchers get information from individuals through the telephone. This method is quick and gives accurate information.

1.12 Methods of Collecting Secondary Data

Secondary data are collected by the following methods:

1.12.1 Official

Publications from the Statistical Division, Ministry of Finance, the Federal Bureaus of Statistics, Ministries of Food, Agriculture, Industry, Labor, etc.

1.12.2 Semi-Official

- Publications from State Bank, Railway Board, Central Cotton Committee, Boards of Economic Enquiry etc.

- Publication of Trade Associations, Chambers of Commerce, etc.

- Technical and Trade Journals and Newspapers.

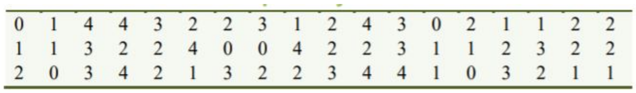
- Research Organizations such as universities and other institutions.

1.13 Difference Between Primary and Secondary Data

The difference between primary and secondary data is only a change of hand. Primary data are the first hand information which is directly collected form one source. They are the most original in character and have not undergone any sort of statistical treatment, while secondary data are obtained from other sources or agencies. They are not pure in character and have undergone some treatment at least once.

1.14 Frequency distribution

Table shows the number of children per family for 54 families selected from a town in India. The data, presented in this form in which it was collected, is called raw data.

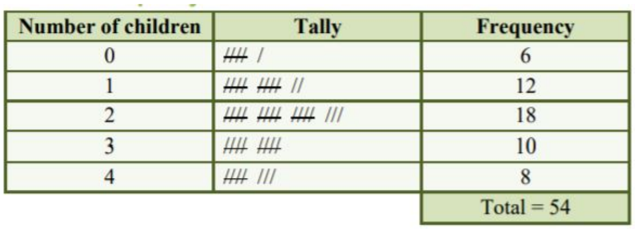


0	1	4	4	3	2	2	3	1	2	4	3	0	2	1	1	2	2
1	1	3	2	2	4	0	0	4	2	2	3	1	1	2	3	2	2
2	0	3	4	2	1	3	2	2	3	4	4	1	0	3	2	1	1

Figure 1.3: raw data set of No. of children in 54 families

It can be seen that, the minimum and the maximum numbers of children per family are 0 and 4, respectively. Apart from these numbers, it is impossible, without further careful study, to extract any exact information from the data. But by breaking down the data into the form below

Now certain features of the data become apparent. For instance, it can easily be seen that, most of the 54 families selected have two children because number of houses having 2 children is 18. This information cannot easily be obtained



Number of children	Tally	Frequency
0	卌 /	6
1	卌 卌 //	12
2	卌 卌 卌 ///	18
3	卌 卌	10
4	卌 ///	8
		Total = 54

Figure 1.4: Frequency distribution table

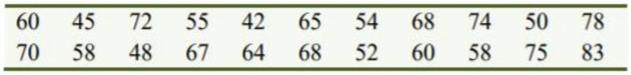
from the raw data. The above table is called a frequency table or a frequency distribution. It is so called because it gives the frequency or number of times each observation occurs. Thus, by finding the frequency of each observation, a more intelligible picture is obtained.

1.14.1 Construction of frequency distribution

1. List all values of the variable in ascending order of magnitude.
2. Form a tally column, that is, for each value in the data, record a stroke in the tally column next to that value. In the tally, each fifth stroke is made across the first four. This makes it easy to count the entries and enter the frequency of each observation.
3. Check that the frequencies sum to the total number of observations

1.15 Grouped frequency distribution

Data below gives the body masses of 22 patients, measured to the nearest kilogram.

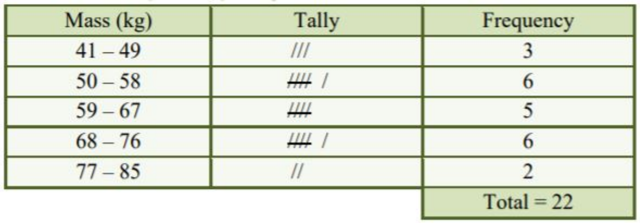


60	45	72	55	42	65	54	68	74	50	78
70	58	48	67	64	68	52	60	58	75	83

Figure 1.5: Body masses of 22 patients

It can be seen that the minimum and the maximum body masses are 42 kg and 83 kg, respectively. A frequency distribution giving every body mass between 42 kg and 83 kg would be very long and would not be very informative. The problem is to overcome by grouping the data into classes.
If we choose the classes
41 – 49
50 – 58

59 – 67
68 – 76 and 77 – 85, we obtain the frequency distribution given below:



Mass (kg)	Tally	Frequency
41 – 49	///	3
50 – 58	~~////~~ /	6
59 – 67	~~////~~	5
68 – 76	~~////~~ /	6
77 – 85	//	2
		Total = 22

Figure 1.6: Grouped Frequency distribution table

Above table gives the frequency of each group or class; it is therefore called a grouped frequency table or a grouped frequency distribution. Using this grouped frequency distribution, it is easier to obtain information about the data than using the raw data. For instance, it can be seen that 17 of the 22 patients have body masses between 50 kg and 76 kg (both inclusive). This information cannot easily be obtained from the raw data.
It should be noted that, even though above table is concise, some information is lost. For example, the grouped frequency distribution does not give us the exact body masses of the patients. Thus the individual body masses of the patients are lost in our effort to obtain an overall picture.

1.16 Terms used in grouped frequency tables.

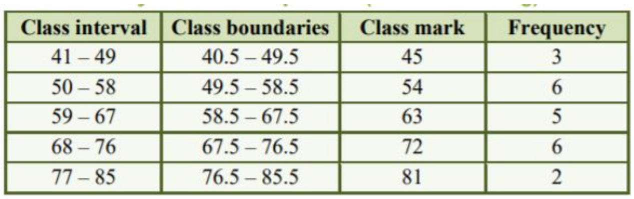
Class limits

The intervals into which the observations are put are called class intervals. The end points of the class intervals are called class limits. For example, the class interval 41 – 49, has lower class limit 41 and upper class limit 49.

Class boundaries

The raw data in the above example were recorded to the nearest kilogram. Thus, a body mass of 49.5kg would have been recorded as 50 kg, a body mass of 58.4 kg would have been recorded as 58 kg, while a body mass of 58.5 kg would have been recorded as 59 kg. It can therefore be seen that, the class interval 50 – 58, consists of measurements greater than or equal to 49.5 kg and less than 58.5 kg. The numbers 49.5 and 58.5 are called the lower and upper boundaries of the class interval 50 – 58. The class boundaries of the other class intervals are given below:

Note:
Notice that the lower class boundary of the ith class interval is the mean of the lower class limit of the class interval and the upper class limit of the (i-1)th



Class interval	Class boundaries	Class mark	Frequency
41 – 49	40.5 – 49.5	45	3
50 – 58	49.5 – 58.5	54	6
59 – 67	58.5 – 67.5	63	5
68 – 76	67.5 – 76.5	72	6
77 – 85	76.5 – 85.5	81	2

Figure 1.7: Class boundary and class limits

class interval (i = 2, 3, 4, ...). For example, in the table above the lower class boundaries of the second and the fourth class intervals are (50 + 49) /2 = 49.5 and (68 + 67)/2 = 67.5 respectively.
It can also be seen that the upper class boundary of the ith class interval is the mean of the upper class limit of the class interval and the lower class limit of the (i+1)th class interval (i = 1, 2, 3, ...). Thus, in the above table the upper class boundary of the fourth class interval is (76 + 77)/2 = 76.5.

Class mark
The mid-point of a class interval is called the class mark or class mid-point of the class interval. It is the average of the upper and lower class limits of the class interval. It is also the average of the upper and lower class boundaries of the class interval. For example, in the table, the class mark of the third class interval was found as follows: class mark =(59+67) /2 = (58.5 + 67.5)/2= 63.

Class width
The difference between the upper and lower class boundaries of a class interval is called the class width of the class interval. Class widths of class intervals can also be found by subtracting two consecutive lower class limits, or by subtracting two consecutive upper class limits.

Note:

The width of the ith class interval is the numerical difference between the upper class limits of the ith and the (i-1)th class intervals (i = 2, 3, ...). It is also the numerical difference between the lower class limits of the ith and the (i+1) th class intervals (i = 1, 2, ...).

In grouped frequency table above the width of the first class interval is |41-50| = 9. This is the numerical difference between the lower class limits of the first and the second class intervals. The width of the second class interval is |50-59|= 9. This is the numerical difference between the lower class limits of the second and the third class intervals. It is also equal to |58-49| the numerical, difference between the upper class limits of the first and the second class intervals.

1.17 Construction of frequency distribution table

Step 1. Decide how many classes you wish to use.
Step 2. Determine the class width
Step 3. Set up the individual class limits **Step 4**. Tally the items into the classes
Step 5. Count the number of items in each class

Consider the example
An agricultural student measured the lengths of leaves on an oak tree (to the nearest cm). Measurements on 38 leaves are as follows
9,16,13,7,8,4,18,10,17,18,9,12,5,9,9,16,1,8,17,1,10,5,9,11,15,6,14,9,1,12,5,16,4,16,8,15,14,17

Step 1. Decide how many classes you wish to use.

H.A. Sturges provides a formula for determining the approximation number of classes. $\mathbf{k = 1 + 3.322}\log\mathbf{N}$**.** Number of classes should be greater than calculated k
In our example N=38, so k=1+3.322×log(38) = 1+3.322×1.5797 = 6.24 = approx 7

So the approximated number of classes should be not less than 6.24 *i.e.* $k^{'}$ =7

Step 2. Determine the class width

Generally, the class width should be the same size for all classes. C= | max − min|/ k. Class width $C^{'}$should be greater than calculated C. For this example, C = | 18− 1|/**6.24** = 2.72, so approximately class width$C^{'}$ = 3 (Note that k used here is the calculated value using Struges formula not the approximated).

Step 3. To set up the individual class limits, We need to find the lower limit only

$$L = min - \frac{C^{'} \times k^{'} - (max - min)}{2}$$

where C and k here are final approximated class width and number of classes respectively in our example $L = 1 - \frac{3\times7-(18-1)}{2}$=1-2=-1; since there is no negative values in data = 0.

Class	Frequency
0-3	3
3-6	5
6-9	5
9-12	9
12-15	5
15-18	9
18-21	2

Even though the student only measured in whole numbers, the data is continuous, so ”4 cm” means the actual value could have been anywhere from 3.5 cm to 4.5 cm.

1.18 Cumulative frequency

In many situations, we are not interested in the number of observations in a given class interval, but in the number of observations which are less than (or greater than) a specified value. For example, in the above table, it can be seen that 3 leaves have length less than 3.5 cm and 9 leaves (i.e. 3 + 6) have length less than 6.5 cm. These frequencies are called cumulative frequencies. A table of such cumulative frequencies is called **a cumulative frequency table** or **cumulative frequency distribution**.

Cumulative frequency is defined as a running total of frequencies. Cumulative frequency can also defined as the sum of all previous frequencies up to the current point. Notice that the last cumulative frequency is equal to the sum of all the frequencies. Two types of cumulative frequencies are Less than cumulative frequency and Greater than cumulative frequency. Less than cumulative frequency (LCF) is the number of values less than a specified value. Greater than cumulative frequency (GCF) is the number of observations greater than a specified value.

The specified value for LCF in the case of grouped frequency distribution will be upper limits and for GCF will be the lower limits of the classes. LCF’s are obtained by adding frequencies in the successive classes and GCF are obtained by subtracting the successive class frequencies from the total frequency.

1.19 Relative frequency

It is sometimes useful to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the frequency of the particular class interval by the total number of observations.

Relative frequency of a class is the frequency of class / total observation. Relative frequencies all add up to 1.

Class	Frequency	A	B	C
0.5 – 3.5	3	3	38	0.078947
3.5 – 6.5	6	9	35	0.157895
6.5 – 9.5	10	19	29	0.263158
9.5 – 12.5	5	24	19	0.131579
12.5 – 15.5	5	29	14	0.131579
15.5 – 18.5	9	38	9	0.236842

[1] “Note: A= Less than cumulative frequency; B= Greater than cumulative frequency, C = Relative frequency”

Data is the sword of the 21st century, those who wield it well, the Samurai.” - Jonathan Rosenberg, former Google SVP!

2

Graphical representation of data

We found that information given in a frequency distribution is easier to interpret than raw data. Information given in a frequency distribution in a tabular form is easier to grasp if presented graphically. Many types of diagrams are used in statistics, depending on the nature of the data and the purpose for which the diagram is intended.

2.1 Histogram

A histogram consists of rectangles with:

- Bases on a horizontal axis, centres at the class marks, and lengths equal to the class widths.
- Areas proportional to class frequencies.

Note:
If the class intervals are of equal size, then the heights of the rectangles are proportional to the class frequencies and it is then customary to take the heights of the rectangles numerically equal to the class frequencies. If the class intervals are of different widths, then the heights of the rectangles are proportional to $\frac{\text{Class Frequency}}{\text{Class Width}}$. This ratio is called **frequency density**.

Table below shows the frequency distribution of the body masses of 50 AIDS patients. Draw a Histogram.

Mass	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
Frequency	3	6	17	13	8	3

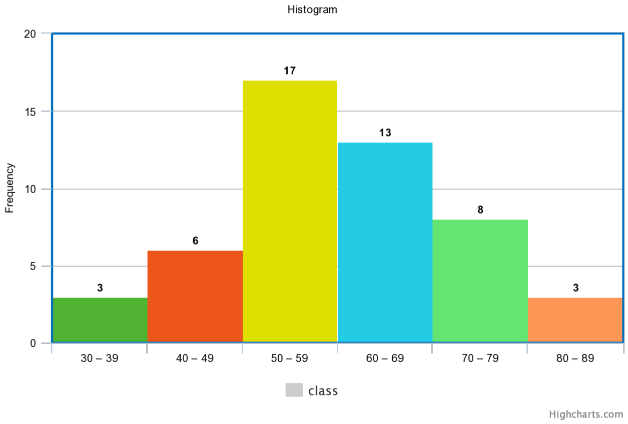


Figure 2.1: Histogram

2.2 Cumulative frequency curve (Ogive)

A graph obtained by plotting a cumulative frequency against the class boundary and joining the points by a smooth curve, is called a cumulative frequency curve. It is also called as Ogive. Two types of ogive are there, Less Than Type Cumulative Frequency Curve (Less than Ogive) and Greater Than Type Cumulative Frequency Curve (Greater than Ogive).

2.2.1 Less than Ogive

Also known as less than type cumulative frequency curve. Here we use the upper limit of the classes and the less than cumulative frequency to plot the curve. Let us see for the example of the body masses of 50 AIDS patients.

Upper limit	39	49	59	69	79	89
Less than Cumulative frequency	3	9	26	39	47	50

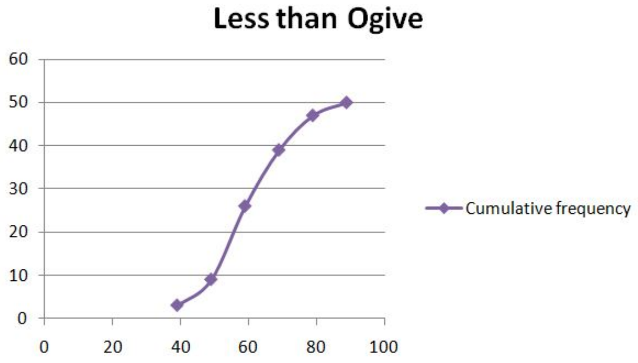


Figure 2.2: Less than ogive

2.2.2 Greater than Ogive

Also known as greater than type cumulative frequency curve Here we use the lower limit of the classes and the greater than cumulative frequency to plot the curve.

Lower Limit	30	40	50	60	70	80
Greater than Cumulative frequency	50	47	41	24	11	3

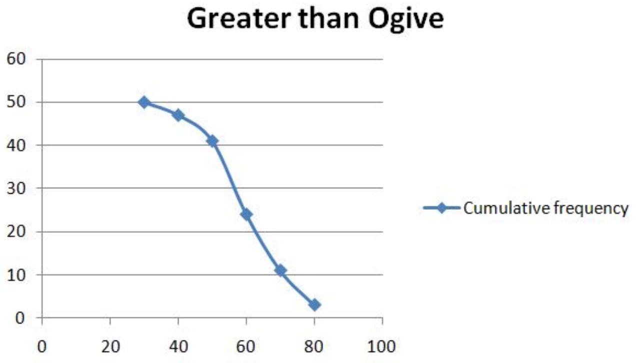


Figure 2.3: greater than ogive

Note:
Intersection of both ogives gives the median

2.2.3 Frequency polygon

A grouped frequency table can also be represented by a frequency polygon, which is a special kind of line graph. To construct a frequency polygon, we plot a graph of class frequencies against the corresponding class mid-points and join successive points with straight lines. Frequency polygon is also obtained by joining the midpoints of a histogram as shown in Fig 2.5.

Class Midpoints	34.5	44.5	54.5	64.5	74.5	84.5
Frequencies	3.0	6.0	17.0	13.0	8.0	3.0

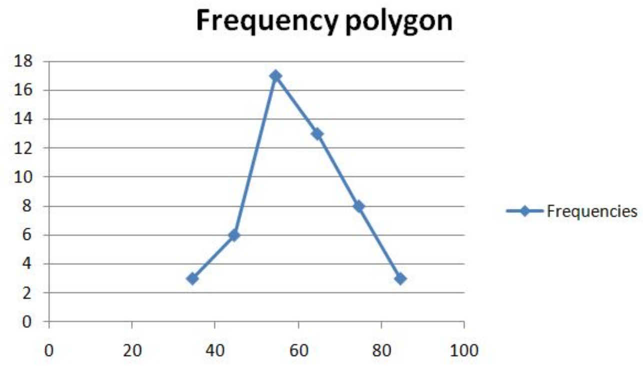


Figure 2.4: Frequency polygon

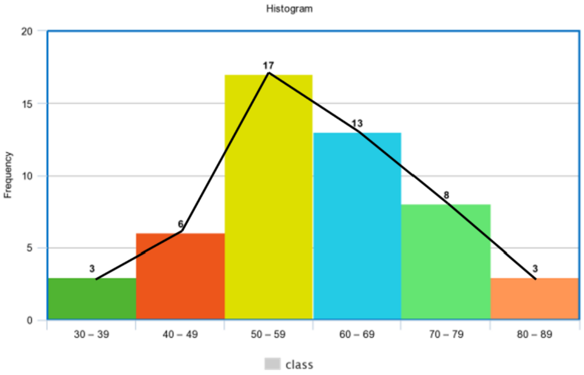


Figure 2.5: Frequency polygon and histogram

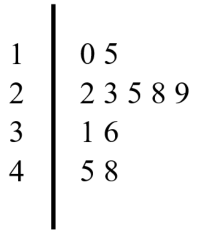
leaf plot, we partition each measurement into two parts. The first part is called the stem, and the second part is called the leaf. Here each numerical value is divided into two parts: The leading digits become the stem the trailing digits become the leaf. One advantage of the stem-and-leaf display over a frequency distribution is that we retain the value of each observation. Another is the distribution of the data within each groups is clear. A stem-and-leaf plot conveys similar information as a histogram. Turned on its side, it has the same shape as the histogram. In fact, since the stem-and-leaf plot shows each observation,it displays information that is lost in a histogram. A properly constructed stem-and-leaf plot, like a histogram, provides information regarding the range of the data set, shows the location of the highest concentration of measurements, and reveals the presence or absence of symmetry.

Consider the example

10,15,22,25,28,23,29,31,36,45,48

stem and leaf plot can be drawn as shown below.

Stem	Leaf
1	0 5
2	2 3 5 8 9
3	1 6
4	5 8



1	0 5
2	2 3 5 8 9
3	1 6
4	5 8

Figure 2.6: Stem and Leaf plot

2.4 Bar chart

A bar chart or bar graph is a diagram consisting of a series of horizontal or vertical bars of equal width. The bars represent various categories of the data. There are three types of bar charts, and these are simple bar charts, component bar charts and grouped bar charts.

2.4.1 Simple bar chart

In a simple bar chart, the height (or length) of each bar is equal to the value of category in the y-axis it represents. For example data below shows the production of timber in five districts of Kerala in a certain year.

District	Production
Alappuzha	600
Kannur	900
Trissur	1800
Ernakulam	1500
Wayanad	2400

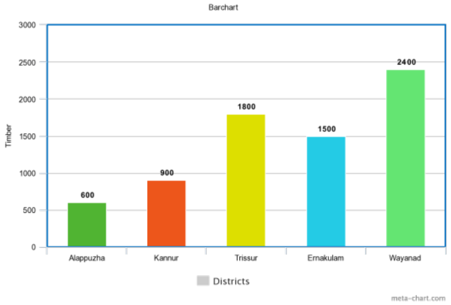


Figure 2.7: Barchart

2.4.2 Component bar chart

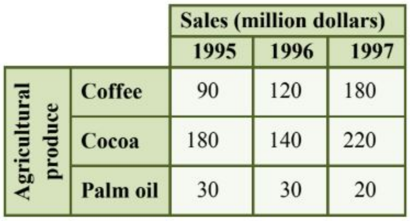
In a component bar chart, the bar for each category is subdivided into component parts; hence its name. Component bar charts are therefore used to show the division of items into components. This is illustrated in the following example.

Example shows the distribution of sales of agricultural produce from a Farm in 1995, 1996 and 1997.

The component bar chart shows the changes of each component over the years as well as the comparison of the total sales between different years.

2.4.3 Grouped bar chart

For a grouped bar chart, the components are grouped together and drawn side by side. We illustrate this with the above example.



		Sales (million dollars)		
		1995	1996	1997
Agricultural produce	Coffee	90	120	180
	Cocoa	180	140	220
	Palm oil	30	30	20

Figure 2.8: Sales data of agricultural produce

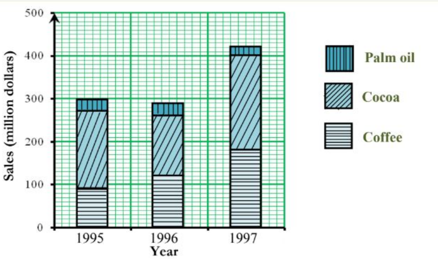


Figure 2.9: Component bar chart

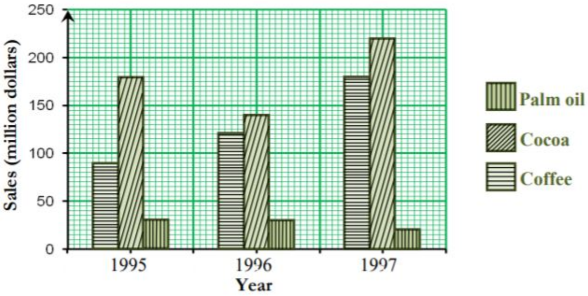


Figure 2.10: Grouped bar chart

2.5 Histogram and Bar chart

Items	HISTOGRAM	BAR GRAPH
Meaning	Histogram refers to a graphical representation, that displays data by way of bars to show the frequency of numerical data.	Bar graph is a pictorial representation of data that uses bars to compare different categories of data.
Indicates	Distribution of non-discrete variables	Comparison of discrete variables
Presents	Quantitative data	Categorical data
Spaces	Bars touch each other, hence there are no spaces between bars	Bars do not touch each other, hence there are spaces between bars.
Elements	Elements are grouped together, so that they are considered as ranges.	Elements are taken as individual entities.
Can bars be re-ordered?	No	Yes
Width of bars	Need not to be same	Same

2.6 Pie Charts

A pie chart is a circular graph divided into sectors, each sector representing a different value or category. The angle of each sector of a pie chart is proportional to the value of the part of the data it represents. The bar chart is more precise than the pie chart for visual comparison of categories with similar relative frequencies.

2.6.1 Steps for constructing a pie chart

1. Find the sum of the category values.

2. Calculate the angle of the sector for each category, using the following formula.Angle of the sector for category A = $\frac{\text{value of category A}}{\text{sum of category values}} \times 360$

3. Construct a circle and mark the centre.

4. Use a protractor to divide the circle into sectors, using the angles obtained in step 2.

5. Label each sector clearly.

See the example:
A lady spent the following sums of money on buying ingredients for a family Christmas cake.

Ingredients	Price	Angle
Flour	24	(24/240)×360= 36
Margarine	96	144
Sugar	18	27
Eggs	60	90
Baking powder	12	18
Miscellaneous	30	45
Total	**240**	**360**

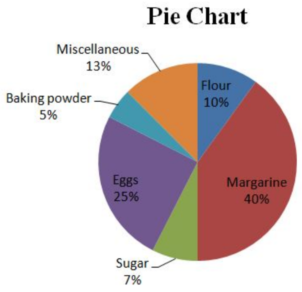


Figure 2.11: Pie chart

Statistics is the grammar of science." - Karl Pearson!

3

Measures of central tendency - I

In the previous lecture, you have learnt how data can be summarised in the form of tables and presented in the form of graphs so that important features can be illustrated easily and more effectively. In this Lecture, we consider statistical measures which can be used to describe the characteristics of a set of data.

We are interested in a single value that serves as a representative value of the overall data. A **measure of central tendency** is a summary statistic that represents the centre point or typical value of a dataset.

There are five averages. Among them mean, median and mode are called **simple averages** and the other two averages geometric mean and harmonic mean are called **special averages**. These measures reflect numerical values in the centre of a set of data and are therefore called measures of central tendency.

Requisites of a Good Measure of Central Tendency:

- It should be rigidly defined.
- It should be simple to understand & easy to calculate
- It should be based upon all values of given data
- It should be capable of further mathematical treatment.
- It should have sampling stability.
- It should be not be unduly affected by extreme values

The main objectives of Measure of Central Tendency

- To condense data in a single value.
- To facilitate comparisons between data.

3.1 Arithmetic Mean

This is what people usually intend when they say ”average”. Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. Mean of set of numbers x_1 , $x_2, \ldots, x_n$ is denoted as $\overline{x}$. It is given by the formula

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$= \frac{1}{n}\sum_{i=1}^{n} x_i$

Example 3.1 Find the mean of the numbers 2, 4, 7, 8, 11, 12

$$\overline{x} = \frac{2+4+7+8+11+12}{6} = \frac{44}{6} = 7.33$$

3.1.1 The mean of a frequency distribution

3.1.1.1 Direct method

If the numbers x_1 , $x_2, \ldots, x_n$ occur with frequencies f_1 , $f_2, \ldots, f_n$ respectively then

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 \quad + \ldots + x_n f_n}{f_1 + f_2 + \ldots f_n}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Example 3.2 Table below shows the body masses of 50 men. Find the mean body mass.

Table 3.1: Body masses of 50 men.

Mass(kg)	59	60	61	62	63
Frequency	3	9	23	11	4

Solution 3.2

The calculation can be arranged as shown

Mass(x)	Frequency(f)	fx
59	3	177
60	9	540
61	23	1403
62	11	682
63	4	252
	$\sum_{i=1}^{n} f_i$= 50	$\sum_{i=1}^{n} f_i x_i$= 3054

$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{3054}{50}$= 61.08 kg

3.1.1.2 Assumed mean method (Indirect method)

The amount of computation involved above can be reduced by using the following formula:

$$\overline{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}$$

Where A is the assumed mean, which can be any value in x. $d_i = x_i - A$, f_i is the frequency of x_i

Consider the Example 2.2 see Table: 3.1

let $A = 61$; it can be any number in x

Mass(x)	Frequency(f)	$\mathbf{d_i{=}x_i{-}61}$	$\mathbf{f_i d_i}$
59	3	-2	-6
60	9	-1	-9
61	23	0	0
62	11	1	11
63	4	2	8
	$\sum_{i=1}^{n} f_i$= 50		$\sum_{i=1}^{n} f_i d_i$= 4

$\overline{x} = 61 + \frac{4}{50} = 61.08$ kg

The mean mass is 61.08 kg

3.1.2 Mean of Grouped Data

3.1.2.1 Direct method

The mean for grouped data is obtained from the following formula:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

Where x_i = the mid-point of i^{th} class (i^{th} class mark); f_i= the frequency of i^{th} class; n = the sum of the frequencies or total frequencies in a sample. Note that i =1,2..., k, *i.e.* there are k classes.

Example 3.3 Shows the distribution of the marks scored by 60 students in a Physics examination. Find the mean mark.

Table 3.4: Distribution of the marks scored by 60 students

Mark (%)	60-65	65-70	70-75	75-80	80-85
Number of students	2	15	25	14	4

Solution 3.3

The solution can be arranged as shown

Marks	Class mark($\mathbf{x_i}$)	Frequency($\mathbf{f_i}$)	$\mathbf{f_i x_i}$
60-65	62.5	2	125
65-70	67.5	15	1012.5
70-75	72.5	25	1812.5
75-80	77.5	14	1085
80-85	82.5	4	330
		$\sum_{i=1}^{n} f_i$= 60	$\sum_{i=1}^{n} f_i x_i$= 4365

$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{4365}{60}$= 72.75

The mean mark is 72.75%

3.1.2.2 Coding method (Indirect method)

If all the class intervals of a grouped frequency distribution have equal size C (class width); then the following formula can be used instead of direct method above. This formula makes calculations easier.

$$\overline{x} = A + C\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}$$

Where A is the he class mark with the highest frequency, $u_i = \frac{x_i - A}{C}$, f_i is the frequency of x_i, C is the class width

This is called the "coding" method for computing the mean. It is a very short method and should always be used for finding the mean of a grouped frequency distribution with equal class widths.

Consider the Example 3.3 see Table:3.4

A=72.5, class mark with highest frequency; C =5

Marks	Class mark($\mathbf{x_i}$)	Frequency($\mathbf{f_i}$)	$\mathbf{u_i = \frac{x_i - 72.5}{5}}$	$\mathbf{f_i u_i}$
60-65	62.5	2	-2	-4
65-70	67.5	15	-1	-15
70-75	72.5	25	0	0
75-80	77.5	14	1	14
80-85	82.5	4	2	8
		$\sum_{i=1}^{k} f_i$= 60		$\sum_{i=1}^{k} f_i u_i$=3

$\overline{x} = 72.5 + 5 \times \left(\frac{3}{60}\right)$= 72.75

The mean mark is 72.75%

3.2 Merits and demerits of Arithmetic mean Merits

Merits

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement *i.e.* Intelligence, beauty, honesty etc.
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

3.3 The median

The median of a set of data is defined as the middle value when the data is arranged in order of magnitude. If there are no ties, half of the observations will be smaller than the median, and half of the observations will be larger than the median. The median can be the middle most item that divides the group into two equal parts, one part comprising all values greater, and the other, all values less than that item. It is a positional measure.

3.3.1 Median of ungrouped or raw data

Arrange the given n observations x_1 , $x_2, \ldots, x_n$ in ascending order. If the number of values is odd, median is the middle value. If the number of values is even, median is the mean of middle two values.

Arrange data in ascending then use the following formula

When n is odd, Median = Md $=\left(\frac{n+1}{2}\right)^{\text{th}}$value

When n is even, Median = Md =Average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ value

Example 3.4 Find the median of each of the following sets of numbers.

(a) 12, 15, 22, 17, 20, 26, 22, 26, 12

(b) 4, 7, 9, 10, 5, 1, 3, 4, 12, 10

Solution 3.4

(a) Arranging the data in an increasing order of magnitude, we obtain 12, 12, 15, 17, 20, 22, 22, 26, 26. Here, N (= 9) is odd, and so, median $=\left(\frac{9+1}{2}\right)^{\text{th}} = 5^{\text{th}}$ ordered observation = 20.

Note: If a number is repeated, we still count it the number of times it appears when we calculate the median.

(b) Arranging the data in an increasing order of magnitude, we obtain 1, 3, 4, 4, 5, 7, 9, 10, 10, 12. Here, N(=10) is an even number and so median = $\frac{1}{2}\{5^{\text{th}}$ ordered observation + 6^{th} ordered observation$\} = \frac{1}{2}(5+7) = 6$.

Note: You can see in each case, the median divides the distribution into two equal parts, with 50% of the observations greater than it and the other 50% less than it.

3.3.2 Median of ungrouped frequency distribution

The median is the middle number is an ordered set of data. In a frequency table, the observations are already arranged in an ascending order. We can obtain the median by looking for the value in the middle position.

3.3.2.1 Median of a frequency table when the number of observations is odd

When the number of observations (n) is odd, then the median is the value at the $\left(\frac{n+1}{2}\right)^{\text{th}}$ positional value. For that we use less than cumulative frequency.

Example 3.5: The following is a frequency table of the score obtained in a mathematics quiz. Find the median score.

Table 3.7: Score obtained in a mathematics quiz.

Score	0	1	2	3	4
Frequency	3	4	7	6	3

Solution 3.5:

Total frequency = 3 + 4 + 7 + 6 + 3 = 23 (odd number). Since the number of scores is odd, the median is at $\left(\frac{23+1}{2}\right)^{\text{th}} = 12^{\text{th}}$ position. To find out the 12^{th} position, we use less than cumulative frequencies as shown:

Score	0	1	**2**	3	4
Frequency	3	4	**7**	6	3
less than cumulative frequency	3	7	**14**	20	23

The 12^{th} position is after the 7^{th} position but before the 14^{th} position. So, the median is 2.

3.3.2.2 Median of a frequency table when the number of observations is even

When the number of observations is even, then the median is the average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ position values.

Example 3.6: The table is a frequency table of the marks obtained in a competition. Find the median score.

Table 3.9: Distribution of marks obtained in a competition.

Mark	0	1	2	3	4
Frequency	11	9	5	10	15

Solution 3.6:

Total frequency = 11 + 9 + 5 + 10 + 15 = 50 (even number). Since the number of scores is even, the median is at the average of the values in $\left(\frac{n}{2}\right)^{th} = 25 \; and \; \left(\frac{n}{2}+1\right)^{\text{th}} = 26$ positions. To find out the 25th position and 26th position, we add up the frequencies as shown:

Mark	0	1	2	3	4
Frequency	11	9	5	10	15
less than cumulative frequency	11	20	25	35	50

The mark at the 25th position is 2 and the mark at the 26th position is 3. The median is the average of the scores at 25th and 26th positions = $\frac{2+3}{2} = 2.5$

3.3.3 Median of grouped frequency distribution

The exact value of the median of a grouped data cannot be obtained because the actual values of a grouped data are not known. For a grouped frequency distribution, the median is in the class interval which contains the $\left(\frac{N}{2}\right)^{\text{th}}$ ordered observation, where N is the total number of observations. This class interval is called the **median class**. The median of a grouped frequency distribution can be estimated by either of the following two methods:

3.3.3.1 Linear interpolation method for estimating the median

The median of a grouped frequency distribution can be estimated by linear interpolation. We assume that the observations are evenly spread through the median class. The median can then be computed by using the following formula:

$$median = L + \left(\frac{\frac{1}{2}N - F}{f_m} \right) C$$

where N = total number of observations, L = lower limit of the median class, F = sum of all frequencies below L(cumulative frequency), f_m = frequency of the median class, C = class width of the median class.

3.3.3.2 Estimation of the median from a cumulative frequency curve

The median of a grouped frequency distribution can be estimated from a cumulative frequency curve. A horizontal line is drawn from the point $\frac{\mathrm{N}}{2}$ on the vertical axis to meet the cumulative frequency curve. From the point of intersection, a vertical line is dropped to the horizontal axis. The value on the horizontal axis is equal to the median.

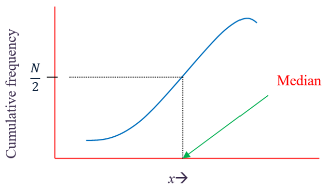


Figure 3.1: median from a cumulative frequency curve

Example 3.7 Table below gives the distribution of the heights of 60 students in a Senior High school. Find the median height of the students

Table 3.11: Distribution of heights of 60 students

Height(cm)	145-150	150-155	155-160	160-165	165-170	170-175
Number of students	3	9	16	18	10	4

Solution 3.7

(i) Linear interpolation method for estimating the median

$N = 60$

Median class= class interval which contains the $\left(\frac{N}{2}\right)^{\text{th}}$ordered observation; here $\left(\frac{60}{2}\right)^{\text{th}} = 30^{\text{th}}$ observation. Before the class 160-165 there are 3+9+16=28 observations so 30^{th} observation will be in the class 160-165, therefore it is the median class.

L = lower limit of the median class =160

F = sum of all frequencies below 160(cumulative frequency) = 16+9+3= 28

f_m = frequency of the median class=18

C = class width of the median class=5

$median = 160 + \left(\frac{\frac{1}{2}60-28}{18}\right)5 = 160.56$

(ii) Estimation of the median from a cumulative frequency curve

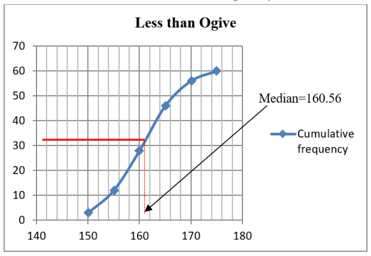


Figure 3.2: Median from a cumulative frequency curve Example 3.7

3.4 Merits and Demerits of Median

Merits

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals
3. Median can be located even if the data are incomplete.

Demerits

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in calculating mean deviation.
4. It does not take into account all the observations

3.5 The mode

The mode of a set of data is the value which occurs with the greatest frequency. The mode is therefore the most common value. The mode is an important measure in case of qualitative data. The mode can be used to describe both quantitative and qualitative data.

3.5.1 Mode of ungrouped or raw data

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example 3.8

(a) The mode of 1, 2, 2, 2, 3 is 2.

(b) The modes of 2, 3, 4, 4, 5, 5 are 4 and 5.

(c) The mode does not exist when every observation has the same frequency. For example, the following sets of data have no modes: (i) 3, 6, 8, 9; (ii) 4, 4, 4, 7, 7, 7, 9, 9, 9.

Note: It can be seen that the mode of a distribution may not exist, and even if it exists, it may not be unique. Distributions with a single mode are referred to as *unimodal.* Distributions with two modes are referred to as *bimodal.* Distributions may have several modes, in which case they are referred to as *multimodal.*

Example 3.9 20 patients selected at random had their blood groups determined. The results are given in the table below

Table 3.12: Blood groups of 20 patients

Blood group	A	AB	B	O
No. of patients	2	4	6	8

The blood group with the highest frequency is O. The mode of the data is therefore blood group O. We can say that most of the patients selected have blood group O. Notice that the mean and the median cannot be applied to the data. This is because the variable “blood group” cannot take numerical values. However, it can be seen that the mode can be used to describe both quantitative and qualitative data.

3.5.2 Mode of Grouped frequency distribution

$$mode = L + \left(\frac{f_s}{f_p + f_s} \right) C$$

Locate the highest frequency the class corresponding to that frequency is called the **modal class**.

Where L = lower limit of the model class; f_p= the frequency of the class preceding the model class; f_s= the frequency of the class succeeding the model class and C = class interval

Example 3.10 For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the mode.

Table 3.13: requency distribution of weights of sorghum ear heads

Weights of ear heads (g)	No of ear heads (f)
60-80	22
80-100	38
100-120	45
120-140	35
140-160	20

Modal class is **100-120**

$mode = 100 + \left(\frac{35}{38+35}\right) 20 = 109.589$

3.5.3 Mode using Histogram

Consider the figure below. The modal class is the class interval which corresponds to rectangle ABCD. An estimate of the mode of the distribution is the abscissa of the point of intersection of the line segments $\overline{\text{AE}}$and $\overline{\text{BF}}$in

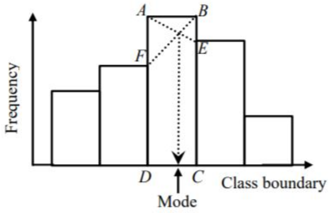


Figure 3.3: Median from a cumulative frequency curve for Example 3.10

3.6 Merits and Demerits of Mode

Merits

1. It is readily comprehensible and easy to compute. In some case it can be computed merely by inspection.
2. It is not affected by extreme values. It can be obtained even if the extreme values are not known.
3. Mode can be determined in distributions with open classes.
4. Mode can be located on the graph also.
5. Mode can be used to describe both quantitative and qualitative data.

Demerits

1. The mode is not unique. That is, there can be more than one mode for a given set of data.
2. The mode of a set of data may not exist
3. It is not based upon all the observation.

If the statistics are boring, you've got the wrong numbers!

4

Measures of central tendency -II

4.1 Geometric mean

The geometric mean is a type of average, usually used for growth rates, like population growth or interest rates. While the arithmetic mean adds items, the geometric mean multiplies items.

The geometric mean of a series containing n observations is the n^{th} root of the product of the values. If $x_1,\ x_2, \dots,\ x_n$ are observations then

$$\text{Geometric mean, } \mathbf{GM} = \sqrt[\mathbf{n}]{\mathbf{x_1 x_2 ... x_n}}$$

$$= \left(\mathbf{x_1 x_2 ... x_n}\right)^{\frac{\mathbf{1}}{\mathbf{n}}}$$

$$\text{logGM}=\frac{\mathbf{1}}{\mathbf{n}}\log\left(\mathbf{x_1 x_2 ... x_n}\right)$$

$$=\frac{\mathbf{1}}{\mathbf{n}}\left(\log\mathbf{x_1}+\log\mathbf{x_2}...+\log\mathbf{x_n}\right)$$

$$=\frac{\sum_{\mathbf{i=1}}^{\mathbf{n}}\log\mathbf{x_i}}{\mathbf{n}}$$

$$\mathbf{GM} = \mathbf{Antilog}\left(\frac{\sum_{\mathbf{i=1}}^{\mathbf{n}}\log\mathbf{x_i}}{\mathbf{n}}\right)$$

4.1.1 Geometric mean for grouped frequency table data

$$\mathbf{GM} = \mathbf{Antilog}\left(\frac{\sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{f_i}\mathrm{log}\mathbf{x_i}}{\mathbf{n}}\right)$$

where x_i is the mid-value, f_i is the frequency , k is the number of classes

Example 4.1: If the weight of sorghum ear heads are 45, 60, 48,100, 65 gms. Find the Geometric mean?

Weight of ear head (x)	log(x)
45	1.653
60	1.778
48	1.681
100	2.000
65	1.813
Total	**8.926**

Solution 4.1:

Here n =5

Geometric mean=

$$\text{Antilog}\left(\frac{\sum_{i=1}^{n} \log x_i}{n}\right) =$$

$$Antilog\left(\frac{8.926}{5}\right) =$$

$$Antilog(1.785) = 60.95$$

(note: here $\text{Antilog}\,(x) = 10^x$ *i.e.*

$$\text{Antilog}\,(1.785) = \ 10^{1.785} = 60.95$$

Example 4.2: Geometric mean of a Frequency Distribution

Weight of ear head (x)	Frequency(f)	log(x)	f[log(x)]
45	5	1.653	8.266
60	4	1.778	7.113
48	6	1.681	10.087
100	8	2.000	16.000

Weight of ear head (x)	Frequency(f)	log(x)	f[log(x)]
65	9	1.813	16.316
Total	**32**		**57.782**

Solution 4.2: Here n =32

$$GM = \ Antilog \left(\frac{\sum_{i=1}^{k} f_i \, \text{log} x_i}{n} \right)$$

$$\sum_{i=1}^{k} f_i \, \text{log} x_i = 57.782$$

$$\text{GM} = \ Antilog \left(\frac{57.782}{32} \right)$$

$$= Antilog \, (1.8056) = 10^{1.8056} = 63.92$$

Example 4.3: Geometric mean of a Grouped Frequency Distribution

Class	Mid value (x)	Frequency(f)	log(x)	f[log(x)]
60-80	70	5	1.845	9.225
80-100	90	4	1.954	7.817
100-120	110	6	2.041	12.248
120-140	130	8	2.114	16.912
140-160	150	9	2.176	19.585
	Total	**32**		**65.787**

Solution 4.4:
Here n =32

$$GM = \ Antilog \left(\frac{\sum_{i=1}^{k} f_i \, \text{log} x_i}{n} \right)$$

$$\sum_{i=1}^{k} f_i \, \text{log} x_i = 65.787$$

$$\text{GM} = \ Antilog \left(\frac{65.787}{32} \right)$$

$$= Antilog \, (2.0558) = 10^{2.0558} = 113.72$$

4.2 Merits and Demerits of Geometric mean

Merits

- It is rigidly defined.
- It is based on all the observations of the series.
- It is suitable for measuring the relative changes.
- It gives more weights to the small values and less weight to the large values.
- It is used in averaging the ratios, percentages and in determining the rate gradual increase and decrease.
- It is capable of further algebraic treatment.

Demerits

- It is not easy to understand.
- It is difficult to calculate.
- It cannot be calculated, if the number of negative values is odd.
- It cannot be calculated, if any value of a series is zero.
- At times it gives a value which may not be found in the series or impractical.

4.3 Harmonic mean

Harmonic means are often used in averaging things like rates (e.g. the average travel speed given duration of several trips). Harmonic mean (HM) of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given value.

If $x_1,\ x_2, \dots,\ x_n$ are n observations then

$$\text{H.M} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

In case of Frequency distribution

$$\text{H.M} = \frac{n}{\sum_{i=1}^{k} f_i \frac{1}{x_i}}$$

where x_i is the mid-value, f_i is the frequency , k is the number of classes

4.3.1 Steps in calculating Harmonic Mean (H.M)

1. Calculate the reciprocal (1/value) for every value.
2. Find the average of those reciprocals (just add them and divide by how many there are)
3. Then do the reciprocal of that average (=1/average)

Example 4.4: From the given data 5, 10, 17, 24, 30 calculate H.M

Solution 4.4:

Here $n = 5$

x	1/x
5	0.2
10	0.1
17	0.058824
24	0.041667
30	0.033333
Total	**0.433824**

$$\text{H.M} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{5}{0.433824} = 11.525$$

Example 4.5: Number of tomatoes per plant are given below. Calculate the harmonic mean.

No. of Tomato per plants	20	21	22	23	24	25
No. of Plants	4	2	7	1	3	1

Solution 4.5:

x	f	1/x	f.1/x
20	4	0.05	0.2
21	2	0.047619	0.095238
22	7	0.045455	0.318182
23	1	0.043478	0.043478
24	3	0.041667	0.125
25	1	0.04	0.04
	18		**0.821898**

Here n =18

$$\text{H.M} = \frac{n}{\sum_{i=1}^{n} f_i \frac{1}{x_i}} = \frac{18}{0.821898} = 21.90$$

4.4 Merits and Demerits of Harmonic mean

Merits

- It is rigidly defined.
- It is defined on all observations.
- It is amenable to further algebraic treatment.
- It is the most suitable average when it is desired to give greater weight to smaller and less weight to the larger ones.

Demerits

- It is not easily understood.
- It is difficult to compute.
- It is only a summary figure and may not be the actual item in the series
- It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.
- It is rarely used in grouped data.

4.5 Relation between AM, GM and HM

If AM stands for Arithmetic Mean, GM stands for Geometric Mean and HM stands for Harmonic Mean; then

$$\text{AM}\times\text{HM}=\text{GM}^{\mathbf{2}}$$

also

$$\mathbf{AM} \geq \mathbf{GM} \geq \mathbf{HM}$$

4.6 When to use AM, GM and HM?

A practical answer is that it depends on what your numbers are measuring.

If you are measuring units that add up linearly in a sequence; such as lengths, distances, weights, then an arithmetic mean will give you a meaningful average. For example, the arithmetic mean of the height or weight of students in a class represents the average height or weight of students in the class.

Harmonic mean will give you a meaningful average, if you are measuring units that add up as reciprocals in a sequence; such as speed or distance travelled per unit time, capacitance in series, resistance in parallel. For example, the harmonic mean of capacitors in series represents the capacitance that a single capacitor would have if only one capacitor was used instead of the set of capacitors in series.

If you're measuring units that multiply in a sequence; such as growth rates or percentages, then a geometric mean will give you a meaningful average. For example, the geometric mean of a sequence of different annual interest rates over 10 years represents an interest rate that, if applied constantly for ten years, would produce the same amount growth in principal as the sequence of different annual interest rates over ten years did.

4.7 Positional Averages

Positional average of a series of values refers to the averages which are taken out from the series itself which represents the whole series or may have some positional properties.

In median, the middle most value of the series is taken as the representative value. Therefore, median is a positional average. Mode is also a positional average as modal values are the most frequently occurring values that are directly taken from the series itself. Other positional averages include **Percentiles**, **Quartiles** and **Deciles**

Note that Arithmetic mean, Harmonic mean and Geometric mean are termed as mathematical averages

4.8 Quartiles

The median divides a set of data into two equal parts. We can also divide a set of data into more than two parts. When an ordered set of data is divided into four equal parts, the division points are called **quartiles**.

The **first or lower quartile ($\mathbf{Q_1}$)** is a value that has one fourth, or 25% of the observations below its value.

The **second quartile** $(\mathbf{Q_2})$, has one-half, or 50% of the observations below its value. The second quartile is equal to the **median**.

The **third or upper quartile**, $(\mathbf{Q_3})$, is a value that has three-fourths, or 75% of the observations below it.

$\mathbf{Q_1}= \left(\frac{\mathbf{n+1}}{\mathbf{4}}\right)^{\text{th}}$**item**

$\mathbf{Q_3}= \left(\frac{\mathbf{3(n+1)}}{\mathbf{4}}\right)^{\text{th}}$**item**

Calculations of quartiles are explained using the example below. See in the example the procedure followed when a fraction appear in the calculation.

Example 4.6: Compute quartiles for the data 25, 18, 30, 8, 15, 5, 10, 35, 40, 45

Solution 4.6:

First arrange the data in ascending order

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

here $n = 10$

$\mathbf{Q_1}= \left(\frac{\mathbf{n+1}}{\mathbf{4}}\right)^{\text{th}}$**item**

i.e. $Q_1 = \left(\frac{10+1}{4}\right)^{th} = 2.75^{\text{th}}$ item; when such a fraction appears we use the following procedure

$Q_1 = 2.75^{\text{th}}$ item $= 2^{\text{nd}}$ item $+ 0.75(3^{\text{rd}}$ item – 2^{nd} item)

So from the given data Q_1= 8+0.75(10– 8) = **9.5**

$\mathbf{Q_2}$**= median**

here Q_2 = (18+25)/2 = **21.5**

$\mathbf{Q_3}= \left(\frac{\mathbf{3(n+1)}}{\mathbf{4}}\right)^{\text{th}}$**item**

i.e. $Q_3 = \left(3 \times \frac{(10+1)}{4}\right)^{th} = 8.25^{\text{th}}$ item $= 8^{\text{th}}$ item $+ 0.25(9^{\text{th}}$ item -8^{th} item) = 35+0.25(40-35) =**36.25**

4.8.1 Quartiles of a discrete frequency data

1. Find cumulative frequencies.
2. Find $\left(\frac{\mathbf{n+1}}{\mathbf{4}}\right)$
3. See in the cumulative frequencies, the value just greater than $\left(\frac{\mathbf{n+1}}{\mathbf{4}}\right)$, then the corresponding value of x is Q_1

4. Find $\left(\mathbf{\frac{3(n+1)}{4}}\right)$

5. See in the cumulative frequencies, the value just greater than $\left(\mathbf{\frac{3(n+1)}{4}}\right)$,then the corresponding value of x is Q_3

Example 4.7: Compute quartiles for the data given bellow

x	5	8	12	15	19	24	30
f	4	3	2	4	5	2	4

Solution 4.7:

x	f	cf
5	4	4
8	3	7
12	2	9
15	4	13
19	5	18
24	2	20
30	4	24

Here n =24

$(\mathbf{\frac{n+1}{4}}) = (\mathbf{\frac{n+1}{4}}) = (\mathbf{\frac{25}{4}})= 6.25$

The cumulative frequency value just greater than 6.25 is 7, the
x value corresponding to cumulative frequency 7 is 8. So $\mathbf{Q_1= 8.}$

$\left(\mathbf{\frac{3(n+1)}{4}}\right) = (\mathbf{\frac{3\times25}{4}})= 18.75$

The cumulative frequency value just greater than 18.75 is 20, the
x value corresponding to cumulative frequency 20 is 24. So $\mathbf{Q_3= 24.}$

4.8.2 Quartiles of a continuous frequency data

1. Find cumulative frequencies

2. Find $(\mathbf{\frac{n}{4}})$

3. See in the cumulative frequencies, the value just greater than $(\mathbf{\frac{n}{4}})$, and then the corresponding class interval is called **first quartile class**.

4. Find $3\,(\mathbf{\frac{n}{4}})$

5. See in the cumulative frequencies the value just greater than $3\left(\frac{\mathbf{n}}{\mathbf{4}}\right)$ then the corresponding class interval is called **3$^{\text{rd}}$ quartile class**. Then apply the respective formulae

$$\mathbf{Q_1{=}l_1{+}\frac{\frac{n}{4}{-}m_1}{f_1}{\times}c_1}$$

$$\mathbf{Q_3{=}l_3{+}\frac{3\left(\frac{n}{4}\right){-}m_3}{f_3}{\times}c_3}$$

Where l_1 = lower limit of the first quartile class

f_1 = frequency of the first quartile class

c_1 = width of the first quartile class

m_1 = cumulative frequency preceding the first quartile class

l_3= 1ower limit of the 3$^{\text{rd}}$ quartile class

f_3= frequency of the 3$^{\text{rd}}$ quartile class

c_3= width of the 3$^{\text{rd}}$ quartile class

m_3 = cumulative frequency preceding the 3$^{\text{rd}}$ quartile class

Example 4.8: Find the quartiles for the grouped frequency data given

Class	**frequency**	**cumulative frequency**
0-10	11	11
10-20	18	29
20-30	25	54
30-40	28	82
40-50	30	112
50-60	33	145
60-70	22	167
70-80	15	182
80-90	12	194
90-100	10	204

Solution 4.8:

$\left(\frac{n}{4}\right) = \frac{204}{4} = 51$

The cumulative frequency value just greater than 51 is 54 so the class 20-30 is the 1$^{\text{st}}$ quartile class

$$\mathbf{Q_1{=}l_1{+}\frac{\frac{n}{4}{-}m_1}{f_1}{\times}c_1}$$

$$=\mathbf{20}+\frac{\mathbf{51-29}}{\mathbf{25}}\times\mathbf{10}\ \ =\mathbf{28.8}$$

$3\left(\frac{n}{4}\right)=3\times\frac{204}{4}=153$

The cumulative frequency value just greater than 153 is 167 so the class 60-70 is the 3rd quartile class

$$\mathbf{Q_3{=}l_3{+}\frac{3\left(\frac{n}{4}\right)-m_3}{f_3}{\times}c_3}$$

$$=\mathbf{60}+\frac{\mathbf{153-145}}{\mathbf{22}}\times\mathbf{10}=\mathbf{63.63}$$

4.9 Percentiles

The percentile values divide an ordered set of data into 100 equal parts each containing 1 percent of the observations. The x^{th} percentile, denoted as P_x is that value below which x percent of values in the distribution fall. It may be noted that the median is the 50th percentile, 25th percentile is first quartile Q_1 and 75th percentile is Q_3

For raw data, first arrange the n observations in increasing order. Then the x^{th} percentile is given by

$\mathbf{P_x{=}\left(\frac{x(n+1)}{100}\right)^{th}item}$

For a frequency distribution the x^{th} percentile is given by following steps

1. Find cumulative frequencies
2. Find $\left(\frac{\mathrm{x.n}}{100}\right)$
3. See in the cumulative frequencies, the value just greater than$\left(\frac{\mathrm{x.n}}{100}\right)$and then the corresponding class interval is called **Percentile class**.
4. Use the following formula

$$\mathbf{P_x{=}\ l{+}\frac{\left(\frac{x\times n}{100}\right)-cf}{f}{\times}c}$$

Where

l = lower limit of the percentile class

cf = cumulative frequency preceding the percentile class