Arab Academy for Science, Technology & Maritime Transport

Statistical and probability

Course Codes: BA203

Chapter 1

Basics of Statistics

1.1 Inroduction

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. Statistics plays a vital role in every fields of human activities such as the existing position of per capita income, unemployment, population growth rate, housing, schooling medical facilities etc.... Statistics holds a central position in Industry, commerce, trade, physics, chemistry, economics, mathematics, biology, botany, psychology, astronomy etc...

1.2 Data Type

When working with statistics, it's important to recognize the different types of data: numerical (discrete and continuous), and categorical. Data are the actual pieces of information that you collect through your study. For example, if you ask five of your friends how many pets they own, they might give you the following data: 0, 2, 1, 4, 18. (The fifth friend might count each of her/his aquarium fish as a separate pet). Not all data are numbers; let's say you also record the gender of each of your friends, getting the following data: male, male, female, male, female.

1.2.1 Qualitative Data

Qualitative data represents characteristics such as a person's gender, marital status, hometown, or the types of movies they like. Categorical data can take on numerical values (such as "1" indicating male and "2" indicating female), but these numbers don't have mathematical meaning. For example, you couldn't add them up together. Other names for categorical data are qualitative data, or Yes/No data.

1.2.2 Quantitative Data

Quantitative (Numerical) data have a meaning as a measurement, such as a person's height, weight, IQ, or blood pressure; or they're a count, such as the number of stock shares a person owns, how many teeth a dog has, or how many pages you can read of your favorite book before you fall asleep. Numerical data can be further broken into two types: discrete and continuous.

1.2.2.1 Discrete Data

Discrete data represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (also called finite); or it may go from 0, 1, 2, on to infinity (making it countably infinite). For example, the number of heads in 100 coin flips

takes on values from 0 through 100 (finite case), but the number of flips needed to get 100 heads takes on values from 100 (the fastest scenario) on up to infinity (if you never get to that 100th head). Its possible values are listed as 100, 101, 102, 103,... (representing the countably infinite case).

1.2.2.2 Continuous Data

Continuous data represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line. For example, the exact amount of gas purchased at the pump for cars with 20-gallon tanks would be continuous data from 0 gallons to 20 gallons, represented by the interval [0, 20], inclusive. You might pump 8.40 gallons, or 8.41, or 8.414863 gallons, or any possible number from 0 to 20. In this way, continuous data can be thought of as being uncountably infinite. For ease of record keeping, statisticians usually pick some point in the number to round off. Another example would be that the lifetime of a C battery can be anywhere from 0 hours to an infinite number of hours (if it lasts forever), technically, with all possible values in between. Granted, you don't expect a battery to last more than a few hundred hours, but no one can put a cap on how long it can go.

1.3 Categorizing Statistics

The study of statistics can be categorized into two main branches. These branches are descriptive statistics and inferential statistics. To collect data for any statistical study, a population must first be defined. **Population** indicates a group that has been designated for gathering data from. The **data** is information collected from the population. A population is not necessarily referring to people. A population could be a group of people, measurements of rainfall in a particular area or a batch of batteries. The **sample** is a set of data taken from population to represent the population.

1.3.1 Inferential Statistics

Suppose collected data come from a very large population. Suppose we want to know the average height of all men in a city with a population of so many million residents. It isn't very practical to try and get the height of each man. This is where inferential statistics comes into play. Inferential statistics makes inference about populations using data from the population. Instead of using the entire population to gather the data. The statistician will collect a sample or samples from the millions of residents and make inferences about the entire population using the sample. A decision, estimate, prediction, or generalization about a population are all based on a selected sample.

1.3.2 Descriptive Statistics

Descriptive statistics give information that describes the data in some manner. For example, suppose a pet shop sells cats, dogs, birds and fish. If 100 pets are sold and 40 out of the 100 were dogs, then one description of the data on the pets sold would be that 40%. This same pet shop may conduct a study on the number of fish sold each day for one month and determine that an average of 10 fish were sold each day. The average is an example of descriptive statistics. Some other measurements in descriptive statistics answer questions such as 'How widely dispersed is this data?', 'Are there a lot of different values?' or 'Are many of the values the same?', 'What value is in the middle of this data?', 'Where does a particular data value stand with respect to the other values in the data set?'

A graphical representation of data is another method of descriptive statistics. Examples of this visual representation are histograms, bar graphs and pie graphs. Using these methods, the data

is described by compiling it into a graph, table or other visual representation.

This provides a quick method to make comparisons between different data sets and to spot the smallest and largest values and trends or changes over a period of time. If the pet shop owner wanted to know what type of pet was purchased most in the summer, a graph might be a good medium to compare the number of each type of pet sold and the months of the year.

1.3.2.1 Data Organization and Data Presentation

This session discusses how to do the following:

1. Organize data in a frequency table.
2. Organize data in a cumulative frequency table.
3. Present data in a histogram figure.
4. Present data in a frequency polygon figure.
5. Present data in a ogive figure.

Example 1.1 ***(Categorical Example)*** **Twenty elementary school children were asked if they live with both parents (B), father only (F), mother only (M), or someone else (S). The responses of the children are as follows.**

M B B M F S B M F B

B F B M M B B F B M

(i) Construct a frequency distribution table.

(ii) Write the relative frequency.

(iii) Draw a bar graph for the frequency distribution for the relative frequency.

Solution.

The relative frequency $= \dfrac{F}{\sum F}$ (Relative frequencies are useful for comparing distributions of different sizes.)

Table: The frequency table

Class	Tally	Frequency	Relative Frequency
M	𝍸 /	6	0.3
F	////	4	0.2
B	𝍸 ////	9	0.45
S	/	1	0.05

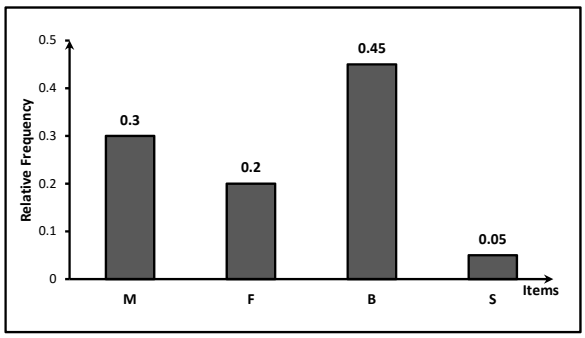

Fig: The bar graph

Example 1.2 ***(Categorical Example)*** **The following data give the number of computer course taken by 30 business major who recently graduated from university.**

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

(i) Construct a frequency distribution table.

(ii) Write the relative frequency for all categories.

(iii) Draw a bar graph for the frequency distribution for the relative frequency.

Solution.

Table: The frequency table

Class	Tally	Frequency	Relative Frequancy
1	𝍸 ///	8	0.2666
2	𝍸 𝍸 /	11	0.3666
3	𝍸 //	7	0.2333
4	////	4	0.1333

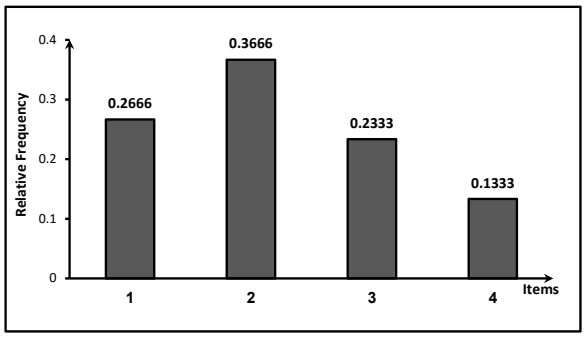

Fig: The bar graph

Example 1.3 ***(Group Example)*** **The following data give the number of computer keyboars assembled at Twentieth Century Electronic Company for a sample of 25 days**

45	52	48	41	56	46	44	42	48	53	51	53	51
48	46	43	52	50	54	47	44	47	50	49	52	

(i) Construct a frequency distribution table if the number of classes is 6.

(ii) Write the relative frequency for all categories.

(iii) Draw a histogram and ogive graph for the relative frequency distribution and class cumulative frequency respectively.

(iv) Draw the polygon graph.

Solution.

The minimum value is **41**. The maximum value is **56**.

Range= maximum value - minimum value = 56-41=15

Class width $= \dfrac{\text{Range}+1}{\text{no. of classes}} = \dfrac{15+1}{6} = 2.66 \approx 3$ (round it up).

Table: The frequency table

Class	Tally	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency	x_m
41-44	///	3	3	0.12	0.12	42.5
44-47	////	5	8	0.20	0.32	45.5
47-50	//// /	6	14	0.24	0.56	48.5
50-53	//// //	7	21	0.28	0.84	51.5
53-56	///	3	24	0.12	0.96	54.5
56-59	/	1	25	0.04	1	57.5

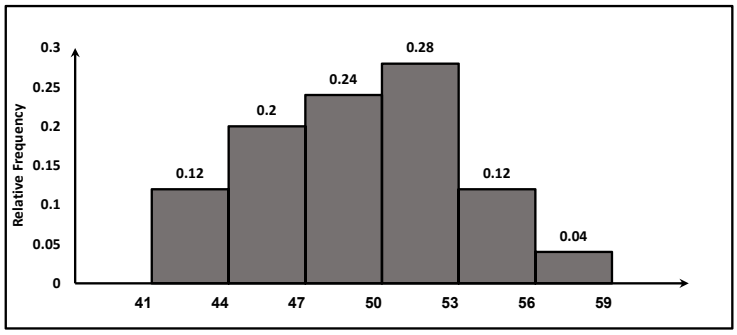

Fig: The histogram graph

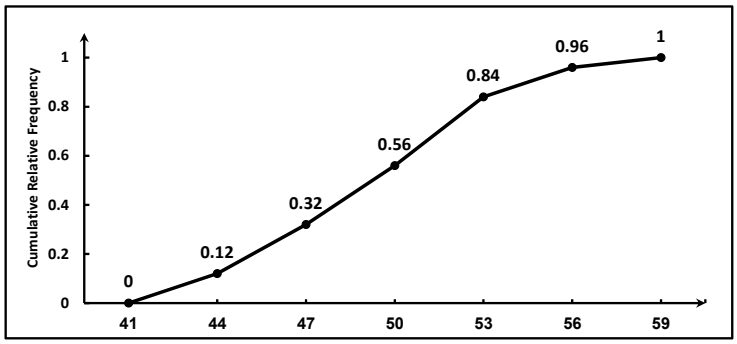

Fig: The Ogive graph

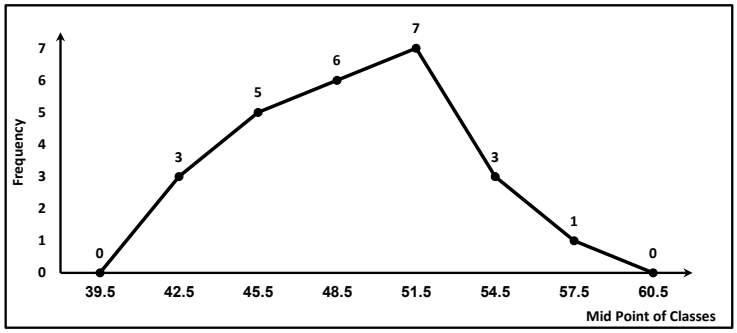

Fig: The Frequency Polygon graph

Example 1.4 ***(Group Example)*** **The following data give the number of computer terminals produced at the company for a sample of 30 days**

24	32	27	23	33	33	29	25	23	28
21	26	31	22	27	33	27	23	28	29
31	35	34	22	26	28	23	35	31	27

(i) Construct a frequency distribution table if the number of classes is 5.

(ii) Write the relative frequency for all categories.

(iii) Draw a histogram and ogive graph for the relative frequency distribution and class cumulative frequency respectively.

(iv) Draw the polygon graph.

Solution.

The minimum value is **35**. The maximum value is **21**.

Range= maximum value - minimum value = 35-21=14

$$\text{Class width} = \frac{\text{Range}+1}{\text{no. of classes}} = \frac{14+1}{5} = 3.$$

Table: The frequency table

Class	Tally	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency	x_m
21-24	𝍸 //	7	7	0.2333	0.2333	22.5
24-27	////	4	11	0.1333	0.3666	25.5
27-30	𝍸 ////	9	20	0.3000	0.6666	28.5
30-33	////	4	24	0.1333	0.8	31.5
33-36	𝍸 /	6	30	0.2000	1	34.5

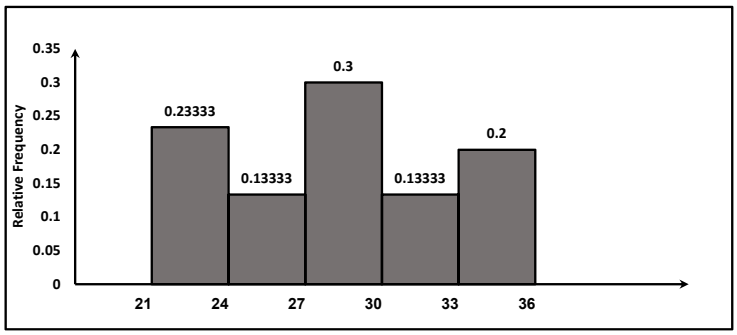

Fig: The histogram graph

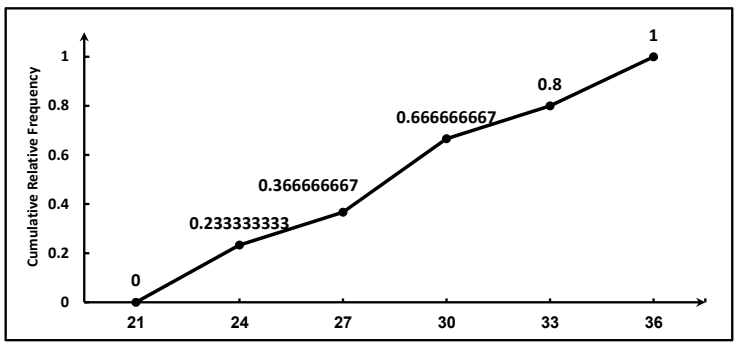

Fig: The Ogive graph

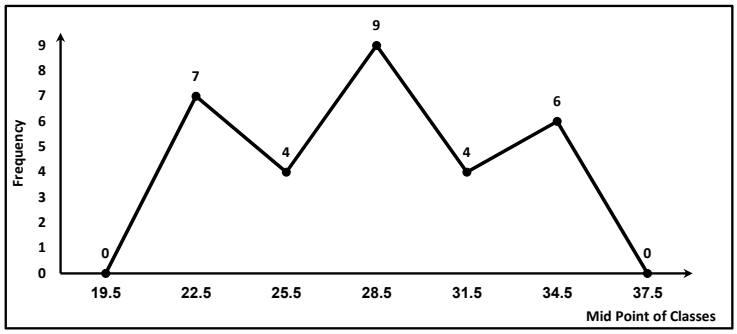

Fig: The Frequency Polygon graph

1.3.2.2 Stem-and-Leaf Plot

Stem-and-leaf plot is a statistical technique **used if the number of classes is not given** to present a set of data by splitting each numerical value into 2 parts:

- **The Stem**: **the leading left most digit(s).**
- **The leaf**: **the remaining "trailing" right most digits.**

Example 1.5 ***(Stem-and-leaf plot)*** **Construct Stem-and-leaf plot for the following data**

37	**33**	**33**	**32**	**29**	**28**	**28**	**23**
22	**22**	**22**	**21**	**21**	**21**	**20**	**20**
19	**19**	**18**	**18**	**18**	**18**	**16**	**15**
14	**14**	**14**	**12**	**12**	**9**	**6**	

Solution. In this example, we will explain how to construct and interpret this kind of graph. A stem and leaf display of the data is shown in the following table.

Stem	leaf												
3	2	3	3	7									
2	0	0	1	1	1	2	2	2	3	8	8	9	
1	2	2	4	4	4	5	6	8	8	8	8	9	9
0	6	9											

The left portion of the table contains the stems. They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. These number is 10's digits. A stem of 3, for example, can be used to represent the 10's digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leafs, and they represent the 1's digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

Example 1.6 ***(Group Example - Stem-and-leaf plot)*** **The following data give the annual incomes (an thousands of dollars) for 40 production managers randomly selected from a large companies.**

57.6	**63.3**	**47.3**	**72.5**	**41.2**	**66.1**	**59.6**	**68.5**
73.3	**39.4**	**44.15**	**84.9**	**53.7**	**37.7**	**63.3**	**77.4**
60.2	**55.9**	**43.1**	**35.6**	**49.3**	**67.4**	**79.2**	**71.9**
48.8	**73.2**	**76.0**	**64.3**	**51.8**	**73.5**	**48.8**	**63.5**
81.5	**72.7**	**69.4**	**51.5**	**77.5**	**67.9**	**46.1**	**65.1**

(i) Construct a frequency distribution table.

(ii) Write the relative frequency for all categories.

(iii) Draw a histogram and ogive graph for the relative frequency distribution and class cumulative frequency respectively.

(iv) Draw the polygon graph.

Solution.

Stem	leaf										
3	9.4	7.7	5.6								
4	7.3	1.2	4.15	3.1	9.3	8.8	8.8	6.1			
5	7.6	9.6	3.7	5.9	1.8	1.5					
6	3.3	6.1	8.5	3.3	0.2	7.4	4.3	3.5	9.4	7.9	5.1
7	2.5	3.3	7.4	9.2	1.9	3.2	6.0	3.5	2.7	7.5	
8	4.9	1.5									

Table: The frequency table

Class	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency	x_m
30-40	3	3	0.075	0.075	35
40-50	8	11	0.2	0.275	45
50-60	6	17	0.15	0.475	55
60-70	11	28	0.275	0.7	65
70-80	10	38	0.25	0.95	75
80-90	2	40	0.05	1	85

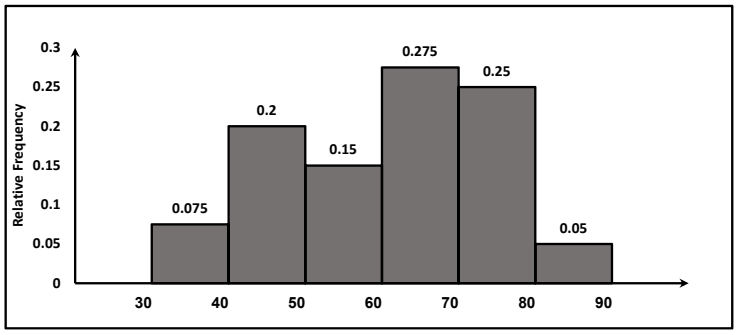

Fig: The histogram graph

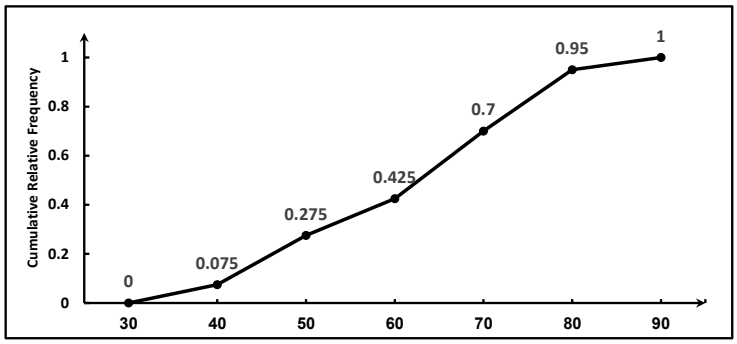

Fig: The Ogive graph

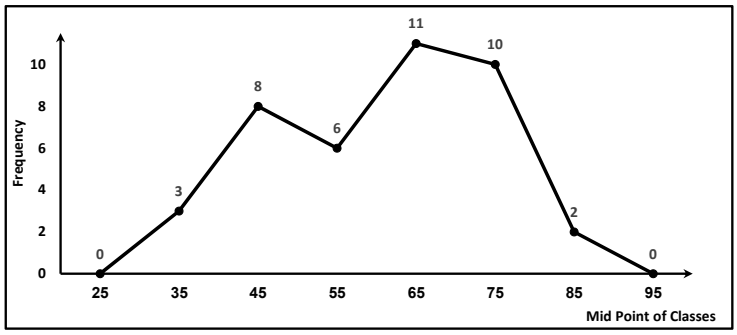

Fig: The Frequency Polygon graph

Example 1.7 ***(Group Example - Stem-and-leaf plot)*** **The following are the GPA of 30 students who signed up for a graduate course at a university.**

3.46	**3.72**	**3.95**	**3.55**	**3.62**	**3.80**	**3.86**	**3.71**	**3.56**	**3.49**
3.96	**3.90**	**3.70**	**3.61**	**3.72**	**3.65**	**3.48**	**3.87**	**3.82**	**3.91**
3.69	**3.67**	**3.72**	**3.66**	**3.79**	**3.75**	**3.93**	**3.74**	**3.50**	**3.83**

(i) Construct a frequency distribution table.

(ii) Write the relative frequency for all categories.

(iii) Draw a histogram and ogive graph for the relative frequency distribution and class cumulative frequency respectively.

(iv) Draw the polygon graph.

Solution.

Stem	leaf							
3.4	6	8	9					
3.5	0	5	6					
3.6	1	2	5	6	7	9		
3.7	0	1	2	2	2	4	5	9
3.8	0	2	3	6	7			
3.9	0	1	3	5	6			

Table: The frequency table

Class	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency	x_m
3.4-3.5	3	3	0.1	0.1	3.45
3.5-3.6	3	6	0.1	0.2	3.55
3.6-3.7	6	12	0.2	0.4	3.65
3.7-3.8	8	20	0.2667	0.6667	3.75
3.8-3.9	5	25	0.1667	0.8334	3.85
3.9-4.0	5	30	0.1667	1	3.95

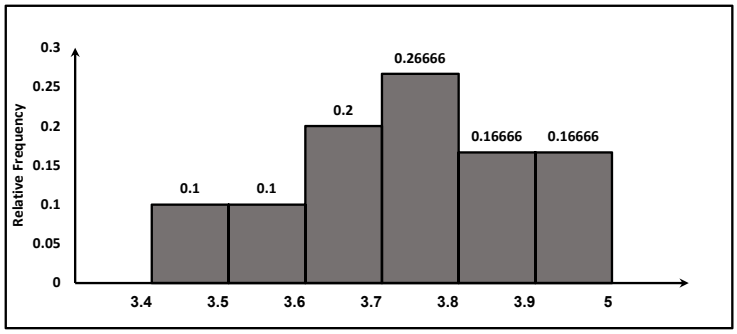

Fig: The histogram graph

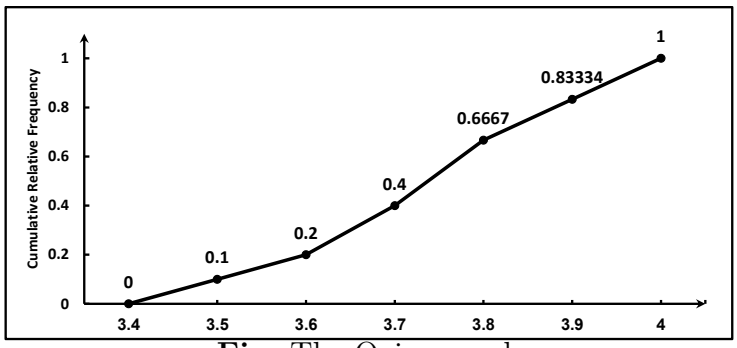

Fig: The Ogive graph

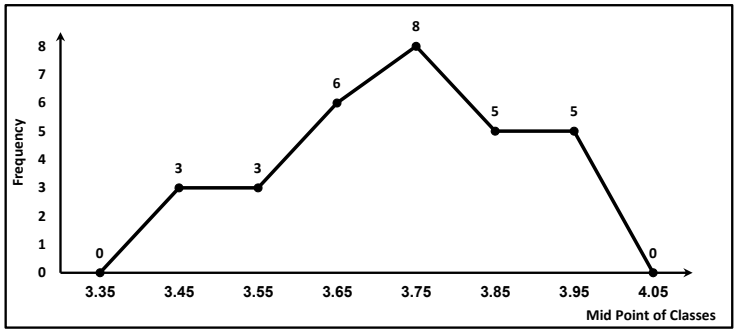

Fig: The Frequency Polygon graph

1.3.2.3 Numerical Descriptive Measures

In this part we introduce the subject matter of descriptive statistics and in doing so learn ways to describe and summarize a set of data.

Measures of Central Tendency

In statistical terms we are trying to find a measure of central tendency. The question we are now faced with is: what is the central position in our frequency distribution? One answer is simply to select the most frequent mark, the longest bar in the histogram. This statistic is called the **mode**.

Another measure of central tendency that is used more often than the mode is the **median or the second quartile** (Q_2) . This is the score that comes in the middle of the list when we have ordered it from lowest to highest. The median of the data value below (Q_2) is called **the first quartile** (Q_1) . The median of the data value above (Q_2) is called **the third quartile** (Q_3).

Whilst we might regard the median as a better choice of a central value than the mode, as it finds the score at the middle position rather than the most frequent score, there is a third measure of central tendency that is used far more often than either of the above two measures. This is the **mean**. The mean represents the average value of the given scores.

Measures of Central Tendency for Ungrouped Data

How to determine the mode for Ungrouped Data?

The mode is a data value that has the highest frequency in a data set. A distribution may have one, more than one, or no mode at all.

How to determine the Median for Ungrouped Data?

1. Sort the data values.
2. Determine the order of the median such that

(a) If n is odd, the order of the median number is the next integer up to $(\frac{n}{2})$-th, where n is data size.

(b) If n is even, the median number is the average between the two numbers have orders $(\frac{n}{2})$ and $(\frac{n+2}{2})$.

How to determine the 1-st Quartile and 3-rd Quartile for Ungrouped Data?

1. Sort the data values.
2. The 1-st Quartile Q_1 is the median of the data values that fall below Q_2. The 3-rd Quartile Q_3 the median of the data values that fall above Q_2.

Interquartile Range (IQR)

Interquartile range is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by $Q1, Q2$, and $Q3$, respectively. So, the interquartile can be calculated as

$$IQR = Q_3 - Q_1$$

How to determine the Mean for Ungrouped Data?

When we talk of an "average" we are usually referring to the mean (although the word "average" is often used much more loosely than the word "mean" (μ) which has its statistical definition). To calculate the mean we add up all the marks and divide them by the number of data.

1. For a **population** of size N, the mean (μ) is given by:

$$\mu = \frac{\sum x_i}{N}$$

2. For a **sample** of size n, the mean ($\bar{x}$) is given by:

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

Measures of Variation (Spread) for Ungrouped Data

Measures of central tendency locate the center of a distribution. They do not indicate how the values are distributed around the center. Measuring variation examine the spread, or variation, of data values around the center. For example, the following two groups illustrate the meaning of variation

Group I : 65 67 68 72 75 80 85 88
Group II : 10 20 30 40 110 120 130 140

$$\mu_I = \frac{65+67+68+72+75+80+85+88}{8} = 75$$
$$\mu_{II} = \frac{10+20+30+40+110+120+130+140}{8} = 75$$

The two distributions have the same mean. In group I, the data values are clustered closer to the mean but in group I is more consistent. The mean is therefore considered representative of the data. Conversely, a large measure of dispersion indicates that the mean is not reliable. A

second reason for studying the dispersion in a set of data is to compare the spread in two or more distributions.

How to determine the Variance and Standard Deviation for ungrouped data?

1. For a **population** of size N, the variance (σ^2) and standard deviation (σ) are given by:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}, \quad \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

We can write simple formulas of the variance and standard deviation in the forms

$$\sigma^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N}, \quad \sigma = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N}}$$

2. For a **sample** of size n, the variance (S^2) and standard deviation (S) are given by:

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

Also, we can write simple formulas of the variance and standard deviation in the forms

$$S^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}, \quad S = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}}$$

Example 1.8 ***(Ungrouped Example)*** **A sample of 7 business statistics books produced the following data on their prices.**

56 47 68 55 71 52 62

(i) Calculate the mean, median, mode, range, variance and standard deviation.

(ii) Evaluate the first, third quartiles and interquartile range (IQR).

Solution. $n = 7$

									$\sum$
Rank	x	47	52	55	56	62	68	71	411
	x^2	2209	2704	3025	3136	3844	4624	5041	24583

Mean : $\bar{x} = \frac{\sum x}{n} = \frac{411}{7} = 58.74.$

Median : Median order $= \frac{n}{2} = 3.5$ (Round it up $\Rightarrow 4$) $\Rightarrow Q_2 = 56.$

Mode : No mode (Each number appears once)

Range : Range = Max. value − Min. value $= 71 - 47 = 24.$

Variance : $S^2 = \dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \dfrac{24583 - \frac{(411)^2}{7}}{6} = 75.23.$

Standard Deviation : $S = \sqrt{\dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\dfrac{24583 - \frac{(411)^2}{7}}{6}} = 8.6739.$

The first quartile : The median of the data values that fall below $Q_2 \Rightarrow$

: $Q_1 = 52.$

The third quartile : The median of the data values that fall above $Q_2 \Rightarrow$

: $Q_3 = 68.$

Interquartile range : IQR $= Q_3 - Q_1 = 68 - 52 = 16.$

Example 1.9 ***(Ungrouped Example)*** **The following table gives the 1992 gross sales (rounded to billions of dollars) for a sample of eight U.S. companies.**

(i) Calculate the mean, median, mode, range, variance and standard deviation

(ii) Evaluate the first, third quartiles and interquartile range (IQR).

Company	Gross Sales (Billions of dollar)
Philip Morris	50
General Electric	62
Pfizer	7
Merck	10
Coca-Cola	13
AT&T	65
Hewlett-Packard	17
Johnson & Johnson	13

Solution. $n = 8$

										$\sum$
Rank	x	7	10	13	13	17	50	62	65	237
	x^2	49	100	169	169	289	2500	3844	4225	11345

Mean : $\bar{x} = \dfrac{\sum x}{n} = \dfrac{237}{8} = 29.625.$

Median : Median order $= \dfrac{n}{2} = 4 \Rightarrow Q_2 = \dfrac{13+17}{2} = 15.$

Mode : Mode is AT&T

Range : Range $=$ Max. value $-$ Min. value $= 65 - 7 = 58.$

Variance : $S^2 = \dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \dfrac{11345 - \frac{(237)^2}{8}}{7} = 617.69.$

Standard Deviation : $S = \sqrt{\dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\dfrac{11345 - \frac{(237)^2}{8}}{7}} = 24.85.$

The first quartile : The median of the data values that fall below Q_2.

: $Q_1 = \dfrac{10+13}{2} = 11.5.$

The third quartile : The median of the data values that fall above Q_2.

: $Q_3 = \dfrac{50+62}{2} = 56.$

Interquartile range : $\text{IQR} = Q_3 - Q_1 = 56 - 11.5 = 44.5.$

Measures of Central Tendency for Grouped Data

How to determine the **Modal class** and **Mode** for **Grouped Data**?

There are two accepted technique for determining the approximate mode from grouped frequency distribution:

1. Determining the modal class (the class with the highest frequency) and then using its class mark as the approximate mode.
2. Using the following mode-locating formula for grouped frequency distributions.

$$\text{Mode} \approx l_{mo} + \left(\frac{d_1}{d_1+d_2}\right)(w_c),$$

where

l_{mo} : The lower boundary of the modal class.
d_1 : The difference between the modal class frequency and the frequency in the class preceding it in the distribution.
d_1 : The difference between the modal class frequency and the frequency in the class following it in the distribution.
w_c : The class width.

How to determine the Median for Grouped Data?

1. Construct the cumulative frequency distribution.
2. Decide the class that contain the median. Class Median is the first class with the value of cumulative frequency equal at least $\frac{n}{2}$.
3. Find the median by using the following formula:

$$\text{Median} = l_m + \left(\frac{\frac{n}{2} - cf}{f_m}\right)(w_c),$$

where

n : The total frequency.
l_m : the lower boundary of the class median.
cf : the cumulative frequency before class median.
f_m : the frequency of the class median.
w_c : The class width.

How to determine the Quartiles for Grouped Data?

Using the same method of calculation as in the Median, we can get Q_1 and Q_3 equation as follows:

$$Q_1 = l_{Q_1} + \left(\frac{\frac{n}{4} - cf}{f_m}\right)(w_c), \quad Q_3 = l_{Q_3} + \left(\frac{\frac{3n}{4} - cf}{f_m}\right)(w_c).$$

How to determine the Mean for Grouped Data?

1. For a **population** of size N, the mean (μ) is given by:

$$\mu = \frac{\sum(x_m f)}{\sum f} = \frac{\sum(x_m f)}{N}$$

where x_m is the midpoint of each class.

2. For a **sample** of size n, the mean $(\bar{x})$ is given by:

$$\bar{x} = \frac{\sum(x_m f)}{\sum f} = \frac{\sum(x_m f)}{n}$$

How to determine the Variance and Standard Deviation for Grouped Data?

1. For a **population** of size N, the variance (σ^2) and standard deviation (σ) are given by:

$$\sigma^2 = \frac{\sum f\,(x_m - \mu)^2}{N}, \quad \sigma = \sqrt{\frac{\sum f\,(x_m - \mu)^2}{N}}$$

We can write simple formulas of the variance and standard deviation in the forms

$$\sigma^2 = \frac{\sum f x_m^2 - \frac{(\sum f x_m)^2}{N}}{N}, \quad \sigma = \sqrt{\frac{\sum f x_m^2 - \frac{(\sum f x_m)^2}{N}}{N}}$$

2. For a **sample** of size n, the variance (S^2) and standard deviation (S) are given by:

$$S^2 = \frac{\sum f\,(x_m - \bar{x})^2}{n-1}, \quad S = \sqrt{\frac{\sum f\,(x_m - \bar{x})^2}{n-1}}$$

Also, we can write simple formulas of the variance and standard deviation in the forms

$$S^2 = \frac{\sum f x_m^2 - \frac{(\sum f x_m)^2}{n}}{n-1}, \quad S = \sqrt{\frac{\sum f x_m^2 - \frac{(\sum f x_m)^2}{n}}{n-1}}$$

Example 1.10 ***(Grouped Example)*** **A hardware distributor reports the following distribution of sales from a sample of 100 sales receipts.**

Dollar Values of Sales	Number of Sales
0–20	16
20–40	18
40–60	14
60–80	24
80–100	20
100–120	8

(i) Calculate the mean, median, mode, variance and standard deviation.

(ii) Evaluate the first, third quartiles and interquartile range (IQR).

Solution.

Classes	Frequency f	Midpoint x_m	$x_m f$	$x_m^2 f$	Cumulative frequency
0–20	16	10	160	1600	16
20–40	18	30	540	16200	34
40–60	14	50	700	35000	48
60–80	24	70	1680	117600	72
80–100	20	90	1800	162000	92
100–120	8	110	880	96800	100
$\sum$	100		5760	429200	

Mean : $\bar{x} = \dfrac{\sum x_m f}{n} = \dfrac{5760}{100} = 57.6.$

Median : Median order is $\dfrac{n}{2} = 50 \Rightarrow 48 \leq 50 < 72.$

: Median class is $60 - 80.$

: $\text{Median } = l_m + \left(\dfrac{\frac{n}{2} - cf}{f_m}\right)(w_c) = 60 + \left(\dfrac{\frac{100}{2} - 48}{24}\right)(20)$

: $Q_2 \approx 61.666.$

Mode : Modal class is $60 - 80.$

: $\text{Mode } = l_{mo} + \left(\dfrac{d_1}{d_1 + d_2}\right)(w_c) = 60 + \left(\dfrac{10}{10+4}\right)(20) \approx 74.28.$

Variance :: $S^2 = \dfrac{\sum f x_m^2 - \frac{(\sum f x_m)^2}{n}}{n-1} = \dfrac{429200 - \frac{(5760)^2}{100}}{99} = 984.08.$

Standard Deviation : $S = \sqrt{\dfrac{\sum f x_m^2 - \frac{(\sum f x_m)^2}{n}}{n-1}} = \sqrt{\dfrac{429200 - \frac{(5760)^2}{100}}{99}} = 31.37.$

The first quartile : The first quartile order is $\dfrac{n}{4} = 25 \Rightarrow 16 \leq 25 < 34.$

: The first quartile class is $20 - 40.$

: $Q_1 = l_{Q_1} + \left(\dfrac{\frac{n}{4} - cf}{f_{Q_1}}\right)(w_c) = 20 + \left(\dfrac{\frac{100}{4} - 16}{18}\right)(20) \approx 30.$

The third quartile : The first quartile order is $\frac{3n}{4} = 10 \Rightarrow 72 \leq 75 < 92$.

: The first quartile class is $80 - 100$.

: $Q_3 = l_{Q_3} + \left(\frac{\frac{3n}{4} - cf}{f_{Q_3}}\right)(w_c) = 80 + \left(\frac{\frac{3\times 100}{4} - 72}{20}\right)(20) \approx 83.$

Interquartile range : $\text{IQR} = Q_3 - Q_1 = 83 - 30 = 53.$

Example 1.11 ***(Grouped Example)*** **The following table gives the frequency distribution of total hours spent the summer for a sample of 40 university students enrolled in an introductory during spring 2015.**

Hours of Study	Number of Students
24–40	3
40–56	5
56–72	10
72–88	12
88–104	5
104–120	5

(i) Calculate the mean, median, mode, variance and standard deviation.

(ii) Evaluate the first, third quartiles and interquartile range (IQR).

Solution.

Classes	Frequency f	Midpoint x_m	$x_m f$	$x_m^2 f$	Cumulative frequency
24–40	3	32	96	3072	3
40–56	5	48	240	11520	8
56–72	10	64	640	40960	18
72–88	12	80	960	76800	30
88–104	5	96	480	46080	35
104–120	5	122	560	62720	40
$\sum$	40		2976	241152	

Mean : $\bar{x} = \frac{\sum x_m f}{n} = \frac{2976}{40} = 74.4.$

Median : Median order is $\frac{n}{2} = 20 \Rightarrow 18 \leq 20 < 30$.

: Median class is $72 - 88$.

: $\text{Median} = l_m + \left(\frac{\frac{n}{2} - cf}{f_m}\right)(w_c) = 72 + \left(\frac{\frac{40}{2} - 18}{12}\right)(16)$

: $\text{Median} \approx 74.666.$

Mode : Modal class is $72 - 88$.

: $\text{Mode} = l_{mo} + \left(\frac{d_1}{d_1 + d_2}\right)(w_c) = 72 + \left(\frac{2}{2+7}\right)(16) \approx 75.5.$

Variance : $S^2 = \frac{\sum f x_m^2 - \frac{(\sum f x_m)^2}{n}}{n-1} = \frac{241152 - \frac{(2976)^2}{40}}{39} = 506.09.$

Standard Deviation : $S = \sqrt{\dfrac{\sum f x_m^2 - \frac{(\sum f x_m)^2}{n}}{n-1}} = \sqrt{\dfrac{241152 - \frac{(2976)^2}{40}}{39}} = 22.49.$

The first quartile : The first quartile order is $\dfrac{n}{4} = 10 \Rightarrow 8 \leq 10 < 18.$

: The first quartile class is $56 - 72.$

: $Q_1 = l_{Q_1} + \left(\dfrac{\frac{n}{4} - cf}{f_{Q_1}}\right)(w_c) = 56 + \left(\dfrac{\frac{40}{4} - 8}{10}\right)(16) \approx 59.2.$

The third quartile : The first quartile order is $\dfrac{3n}{4} = 10 \Rightarrow 30 \leq 30 < 35.$

: The first quartile class is $88 - 104.$

: $Q_3 = l_{Q_3} + \left(\dfrac{\frac{3n}{4} - cf}{f_{Q_3}}\right)(w_c) = 88 + \left(\dfrac{\frac{3\times 40}{4} - 30}{5}\right)(16) \approx 88.$

Interquartile range : $\text{IQR} = Q_3 - Q_1 = 83 - 30 = 53.$