(Section 1)

A (Basic vocabulary)

***** What is statistics

The mathematics of the collection, organization and interpretation of numerical data especially the analysis of population characteristics by inference form sampling.

Types of Statistical Applications:

The field of statistics consists of two branches – **Descriptive statistics** focuses on collection, summarization, presentation and analysis of the data using suitable numerical and graphical methods to look for patterns in a data set.
Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data (population).

***what is interval**

an interval is a range of values. suppose 20-30 is the range of a class. here 20-30 is the interval

***what is population?**

population is everything (or everyone or every element) like investigating college students so all college students are population.

***what is parameter?**

parameter is some characteristic of the population like average GPA of the whole college students. For representing Parameter, we use Greek letter mew, sigma

***what is sample?**

Sample is a small portion of the total population

***what is statistic (not statistics)?**

A statistic is a characteristic of the sample like average GPA of the sample students. We use English letter x bar, s or others to represent statistic. It's like parameter of sample not the whole population

***what is variable?**

Variable is any characteristic of interest gathered from each item in the sample. Like if we ask college students what's their GPA then the GPA is the variable. It can also be represented as the question

***what is data?**

Data is the actual value of the variable like GPA of a student is 2 so here 2 is data of the variable GPA

***what is primary data?**

A data is said to be primary data if it is obtained from an investigation conducted for the first time. Thus, the data collected for the first time by the investigator as original data are known as primary data

***What Is secondary data**?

When a statistical analysis is conducted on a data set available from a prior investigation is called a secondary data.
Example: National income data collected by the government are primary data but they become secondary data for those who use them.

*What is raw data?

In any statistical investigation, when data first collected usually appear in raw form where, information has been recorded merely in arbitrary order in which they happened to occur. This is known as the raw data set.

*How to define population

Population is defined with N and sample is defined with n

(Data and Sampling Techniques)

***How is statistical data collected**

A (Measurement levels)

Data can be measured on several different levels

Levels of measurement

1.**Nominal** – which means we put things in categories like different colors, opinions like yes or no or gender etc.

2.**Ordinal** – nominal data which can be put in order like that is more than that, that is less than that like these order but there can be no clear space like z is greater than a but we don't care how much space is between a and z or like top 5 cooks in America in here we can put them order of betterment but we don't know how much

one is better than other but we only know 2 no cook is better than 3 that's ordinal data

3.**Interval** – it has space between the numbers with meaning and it has no "zero" point like 90-degree temperature is 10 more than 80-degree temperature. This space has meaning which is 10. 0 can be also in temperature but that does not mean it has absent value but rather it can be a point in the number line.

another important thing to note is that sometimes we see that two class boundaries upper and lower limit are the same like one class is 120-125 and other is 125-130. In that case, we use exclusive form of interval and in that form, we include the lowest boundary of data in that class frequency which is 120 and 125 but not the highest boundary which is 125 and 130

And there is other form which is inclusive form where every number including upper and lower limit basically every data in that class is included in the class frequency

4.**ratio** – if we want clearly defined zero that's when we have ratio data and it can have absolute zero that means

none. like test score can be zero or someone having zero children etc.

B (Data Types)

***What are the types of data?**

2 types of data available

1.**Qualitative data** – this data focuses on categories or qualities but not on numbers. like types of car or some type of ethnic group etc.

2.**Quantitative data** – this data focuses on quantities, numbers which can be measured or counted like how many, how much, how far etc. like number of students or distance to school and such

Types of quantitative data

quantitative data can also be divided in two parts.

1.**Discrete** – this data is countable. Data which can be counted in integral and normally not in decimals are

discrete data. Like number of shoes, total students in a school etc.

2.**Continuous data** – this data is measured also this data values fall on a continuum. Like one's height can be from something to something inch, books on a shelf can be from 15 -30 cm, if we see this example have measured data which have ranges between them it can take any value between those two interval.so they are continuous data

C (Sampling)

Sampling – sampling should be random to avoid any bias which means all kind of specific characteristic things from the whole population should be included in sample, not just one type

***How to do random sampling(methods)?**

1.**Simple random sampling** – it is based on random selection methods like random numbers or draw out of a hat. like assign students' number and pick the numbers randomly to include in the study

2.**Stratified random sampling** – in this method we divide the population in two groups where each group is called strata and then select a proportion number from each group. like if a state has 40 percent dem, 35 percent rep and 25 percent independent, then we select 80 dem,70 rep and 50 independent population in my sample

3.**Cluster random sampling** – in this method we divide the population in different groups and randomly select a group which will have full population of that group

4.**Systematic random sampling** – in this method, we start with a random person or item and choose every nth person after that like if I select 2nd person to interview and plan to interview every 5^{th} person then after that then after 2^{nd} I interview 7^{th} , 12^{th},17^{th} and so on until the circle goes to the beginning or in other word the sample ends

(Section 2)

(Distribution, Data Visualizing and Box Plot)

***how do we display data visually?**

A (Contingency Table)

Contingency table is a table that lists results in relation to two variables. If we work with big data then listing all those will be a big mess and that's where contingency table come. This table makes calculate probabilities or compare relations easier and to do so we add column and row for totals. Example is

Example:

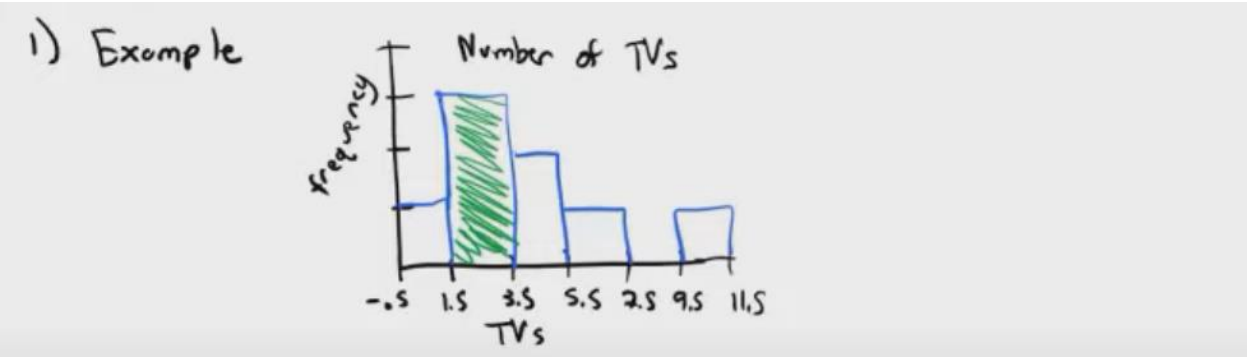
	under 21	21 - 25	over 25	Totals
Speeding ticket	82	39	18	139
no ticket	17	27	61	105
Totals	99	66	79	244

B (Histogram)

***How to use histogram or frequency histogram**

Using histogram or frequency histogram, we can display data which shows us frequency over intervals

Important things to note that the bars touch in histogram, come down on the interval numbers and they show frequency in a range. If possible(discrete) never have a bar come down on a value.in other words, bars should be constructed as made in the table made for constructing the histogram. Another thing is we should add title to the start of histogram and also use labels like x and y axis to make it understand more clear



Here the color part means 3 people have between 1.5 – 3.5 tv

Making Histogram:

1.**Decide on number of bars** – normally the highest number is represented at the last axis of x and the lowest at the start of x and based on that we make an interval and decide how many bars can we make. But its best to take .5 less than the lowest number as the starting interval of the graph and work through the whole graph. The y axis will be about the frequency

2.**Decide on width of bars** – after deciding number of bars, to decide the width of bars, we use a formula which is (highest number – lowest number) / total bars

After doing it, we have to round the number to the next value always. Like if its 3.1 then the answer will be 4 if its 3.2 or anything which. Something then the answer will be the next integral part. even if its 3 or pure integral then always have to add 1 like if result is 3 the real answer would be 4 and such or else the last element won't be added as per stat rule

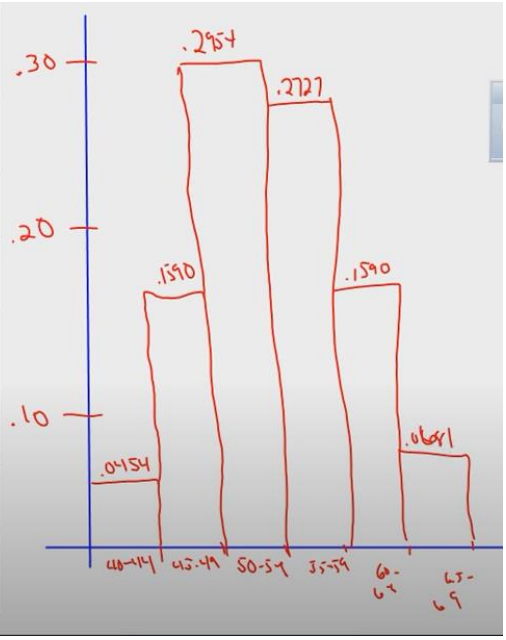
3.**Making frequency table with ranges –** taking frequency or observations with in the limit intervals

And based on the above 3 things, we can build a histogram table or graph

Relative Frequency Histogram

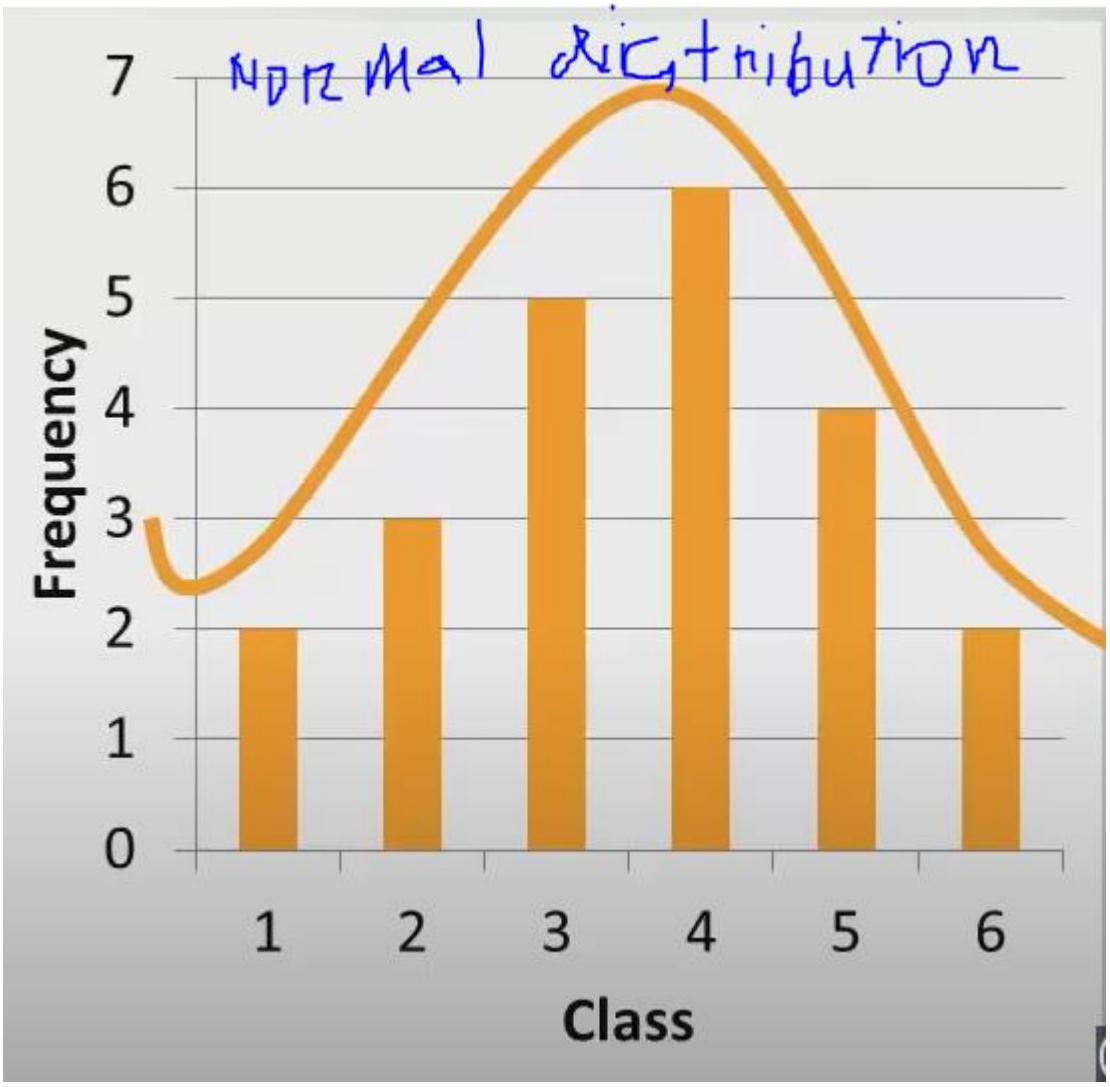
Relative frequency histogram is as same as frequency histogram just the y axis contains relative frequency values rather than frequency values

Age	Frequency	Relative Class Frequencies
40 – 44	2	2/44 .0454
45 – 49	7	7/44 .1590
50 – 54	13	13/44 .2954
55 – 59	12	12/44 .2727
60 – 64	7	7/44 .1590
65 – 69	3	3/44 .0681
Total	44	≈ 1



Describing shape of histogram

1. **Uniform shape** – in a histogram where height of each bars is almost equal or equal are called uniform shape
2. **Normal shape** – in this histogram, the bar (only one) in the middle is taller and the bars on the sides or side its body are shorter
3. **V-shape** – opposite of normal shape
4. **symetrical** – if it's the same in both sides of the middle then its symmetrical
5. **skewed right** – in this histogram, there is extra bars on right side of the middle bar
6. **skewed left** – opposite of skewed left
7. **Normal Distribution** – if the graph looks like a mountain then it's a normal distribution

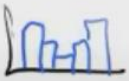

Normal distribution
Frequency
0
1
2
3
4
5
6
7
1
2
3
4
5
6
Class

4) Describe Shape:

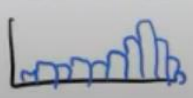
a) uniform - bars about the same 

b) normal - taller in middle, shorter on edges 

c) V-shape - shorter in middle, taller on edges 

d) symetrical - same on both sides 

e) skewed right - extra "stuff" on right 

f) skewed left - extra "stuff on left 

C (Ogive)

Normally ogive graph is about the cumulative frequency. Classes along the x axis and cumulative frequency along the y axis.

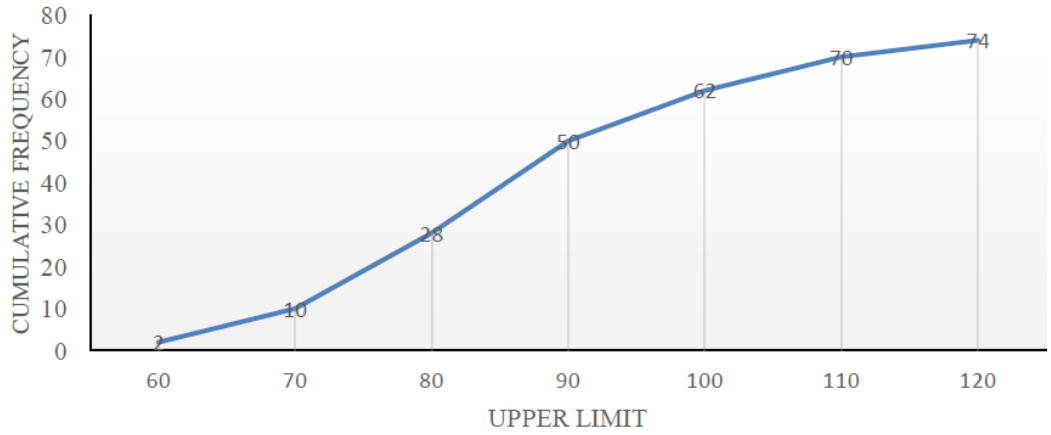
Making an ogive graph:

1.Make a frequency table showing upper class boundaries and cumulative frequencies

2.For each class, make a dot over the upper-class boundary at the height of the cumulative class frequency. The coordinates of the dots are (upper class boundary, cumulative class frequency). Connect these dots with line segment

Class interval of the amount of sell	upper limit	frequency	Cumulative frequency
50 - 60	60	2	2
60 - 70	70	8	10
70 - 80	80	18	28
80 - 90	90	22	50
90 - 100	100	12	62
100 - 110	110	8	70
110 - 120	120	4	74

Figure: Cumulative frequency curve

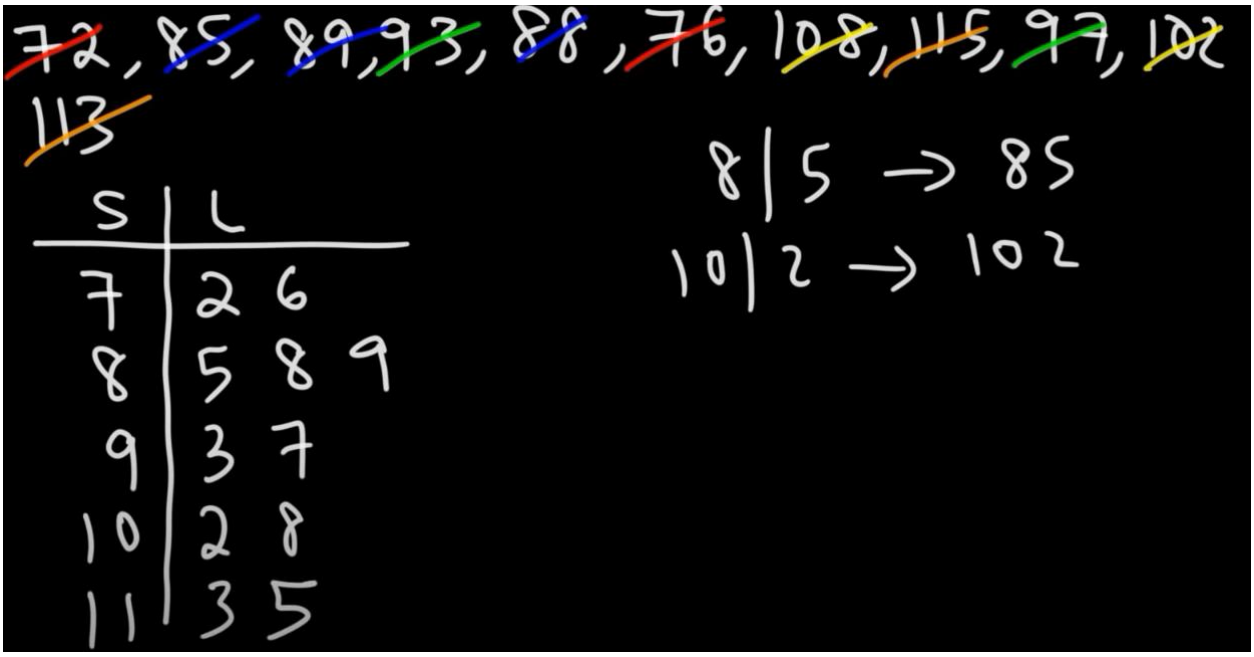


D (Stem And leaf)

Stem and leaf plot are a graphical technique of representing quantitative data that can be used to examine the shape of a frequency distribution, the range of the values and point of concentration of the values.

Constructing stem and leaf

To construct stem and leaf, the numerical value is divided in two parts where the first part is stem and the 2nd part is leaf

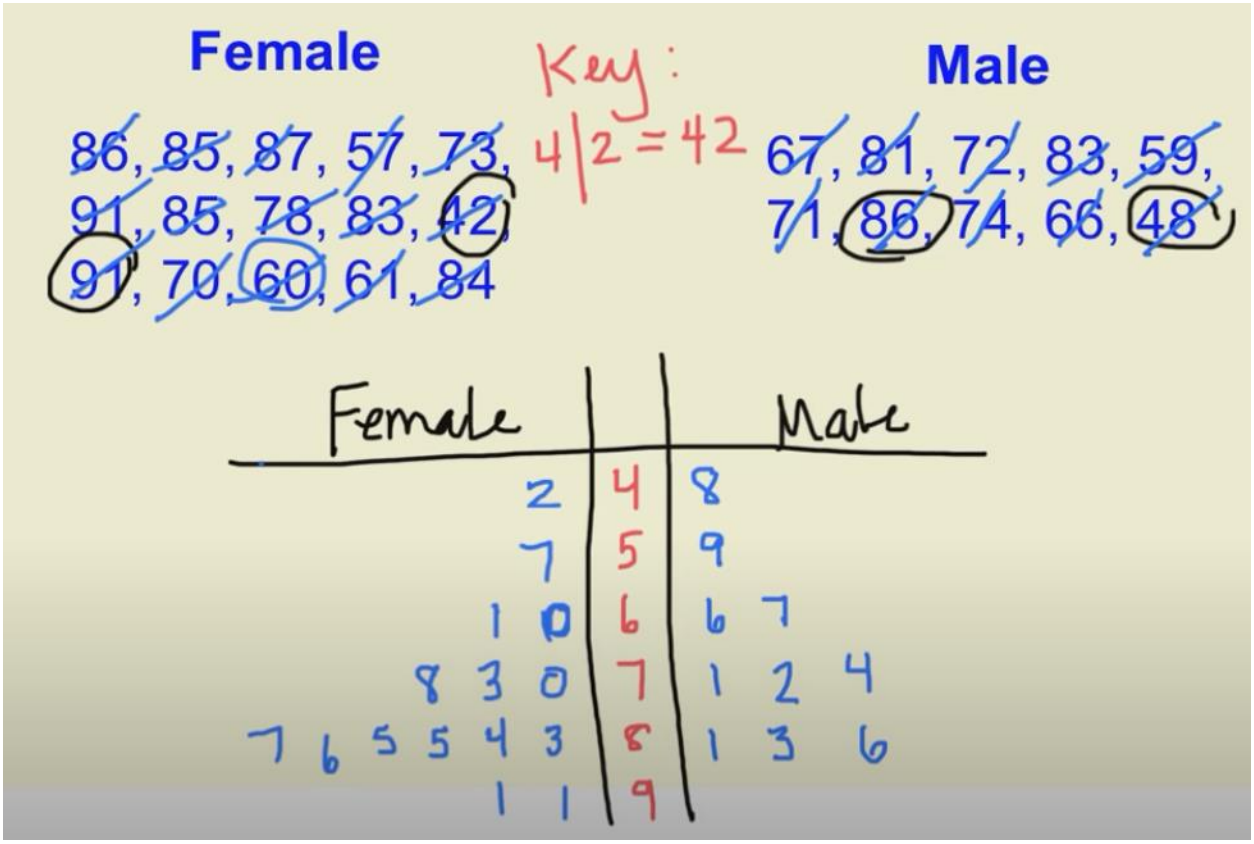


It's also the same if it has decimal value like 2.3 where stem will be 2 and leaf will be 3

Back to back stem and leaf

There is another stem and leaf which compares two dataset and that is back to back method

To do this, we have to find the largest and smallest data from each set and put them in the stem accordingly and use the stem and leaf method for each data set



E (Frequency)

***What is frequency?**

Frequency is how often a value occurs.

Frequency table

we often organize frequency in things like frequency table. now frequency tables have some vocabulary too

relative frequency: proportion of times (frequency of that class/total frequency in whole)

cumulative frequency: sum of all previous frequency entries

cumulative relative frequency: sum of all previous relative frequency entries

frequency table:

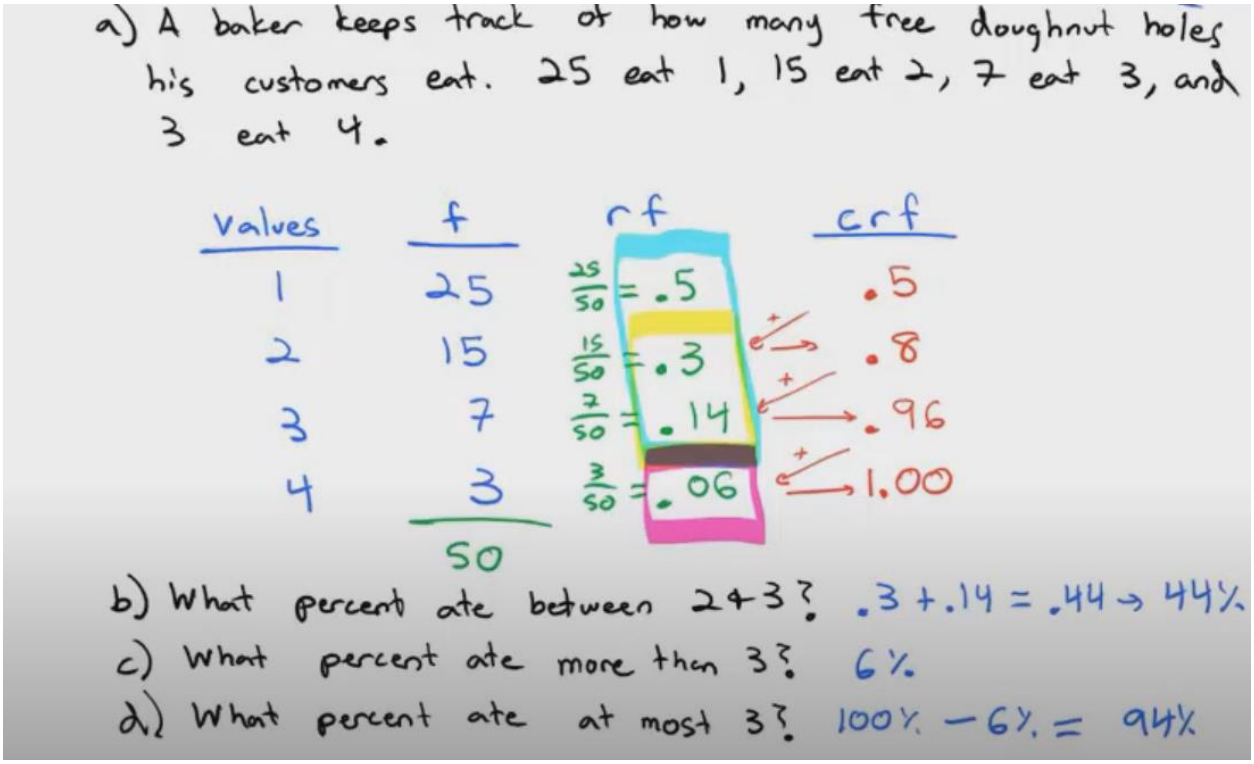
frequency table will generally have 4 columns.1 for data value,1 for frequency,1 for relative frequency and other for cumulative relative frequency. cumulative relative frequencies total sum should always be 1

class width formula of frequency table:

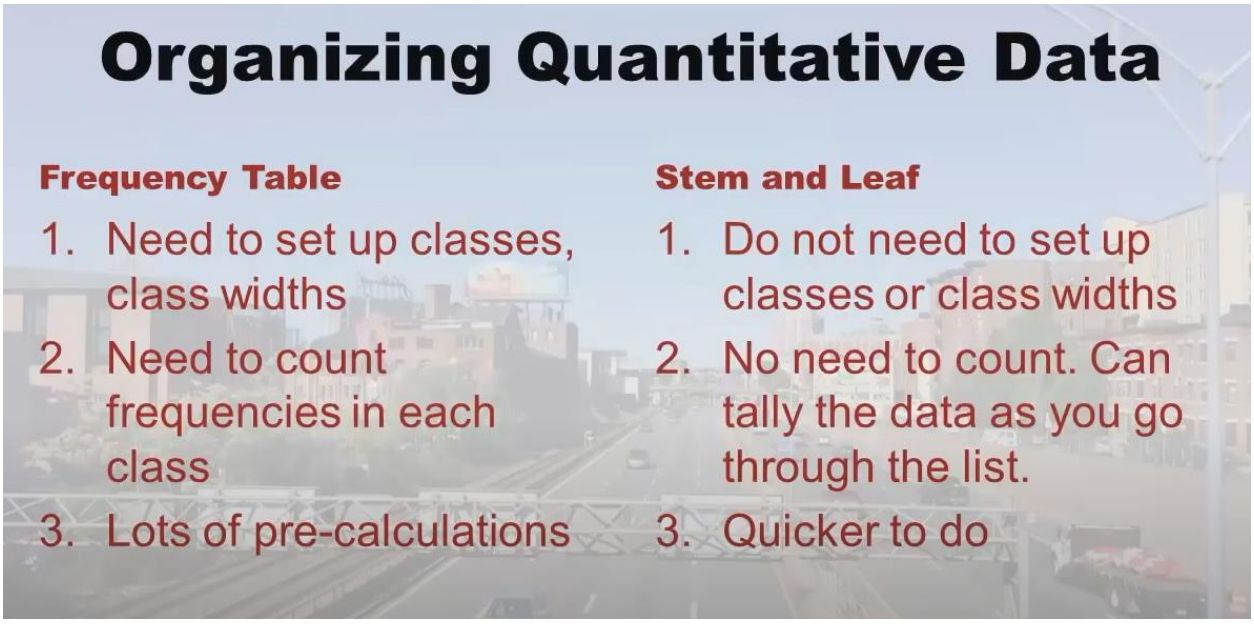
1.Calculate the maximum – minimum (Ex: 47 – 1 = 46)

2.Divide this by the number of classes desired (Ex: If we want 6 classes then 46/6 = 7.7

3.Increase this to the next whole number (round 7.7 to 8)



Differences between frequency table and Stem and Leaf



An important thing to note that we should always make stem and leaf data in order after making the unordered version

F (Polygon)

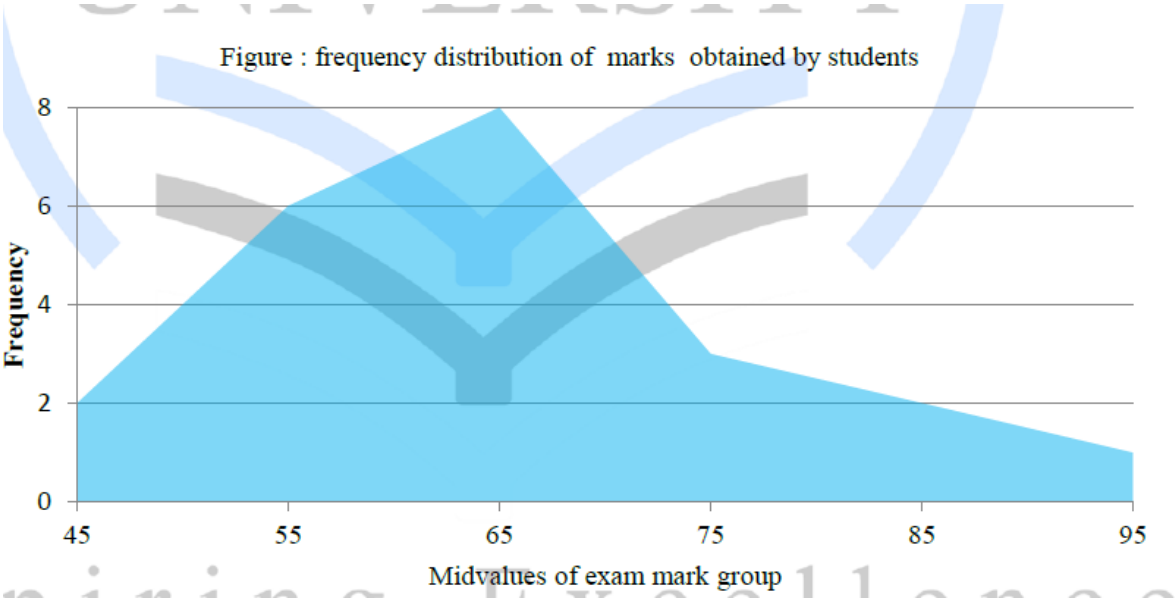
The Frequency Polygon

Frequency polygon is basically the midpoint of each class bars. To draw frequency polygon, we have to first build a table with class intervals, class mid value and its frequency

Table 1.4: Frequency distribution of students by age group

Marks	Mid value	Frequency
40-50	45	2
50-60	55	6
60-70	65	8
70-80	75	3
80-90	85	2
90-100	95	1

and then we do the following



The percentage polygon

Constructing multiple histograms on the same graph to compare two or more data sets often gets confusing. Super imposing the vertical bars of one histogram on another histogram makes interpretation difficult. When there are two or more groups, one should use a percentage polygon

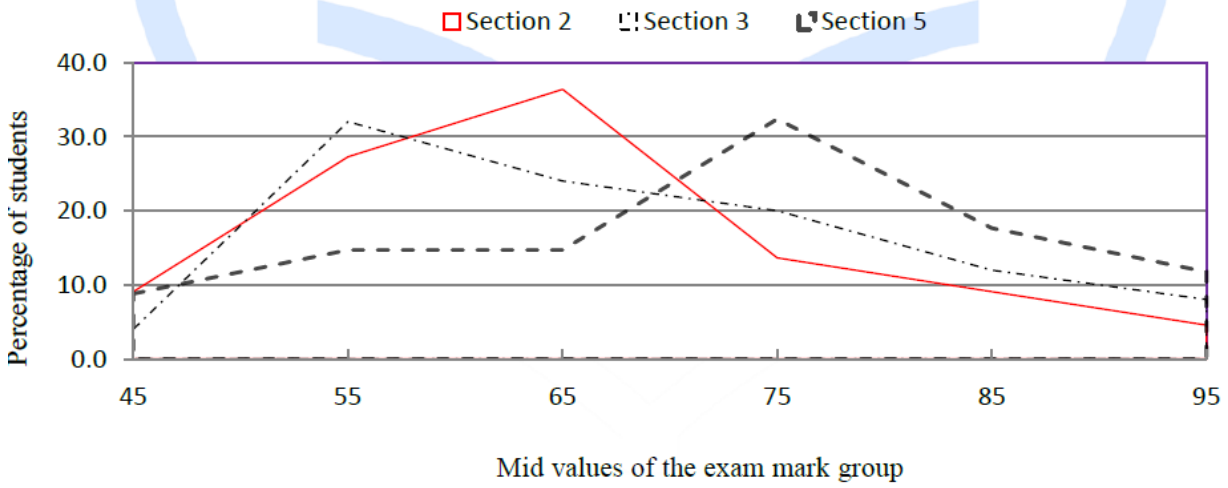
A percentage polygon is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoint at their respective class percentages. The following table and figure

illustrate the construction of the percentage polygon

Table: Frequency distribution of Marks obtained by students taught by "X"

Mark Group	Mid value	Frequency of students			Percentage of Students		
		Section 2	Section 3	Section 5	Section 2	Section 3	Section 5
40-50	45	2	1	3	9.1	4.0	8.8
50-60	55	6	8	5	27.3	32.0	14.7
60-70	65	8	6	5	36.4	24.0	14.7
70-80	75	3	5	11	13.6	20.0	32.4
80-90	85	2	3	6	9.1	12.0	17.6
90-100	95	1	2	4	4.5	8.0	11.8
Total		22	25	34	100	100	100

Figure: Comparison of percentage distribution of grades obtained by students of taught by "X"



G (Box Plot)

Box plot shows the spread of data with 5 number summaries. 5 number summaries are made of 5 pieces.

1.Minimum

2.Q1 or first quartile which is the middle of the lower half

3.Median or the middle when the data is in order which cuts the middle of the data and we get the top and bottom half

4.Q3 or third quartile which is the middle of the upper half

5.Maximum

Important to note that if there are 2 middle values then we have to add them and divide them by 2

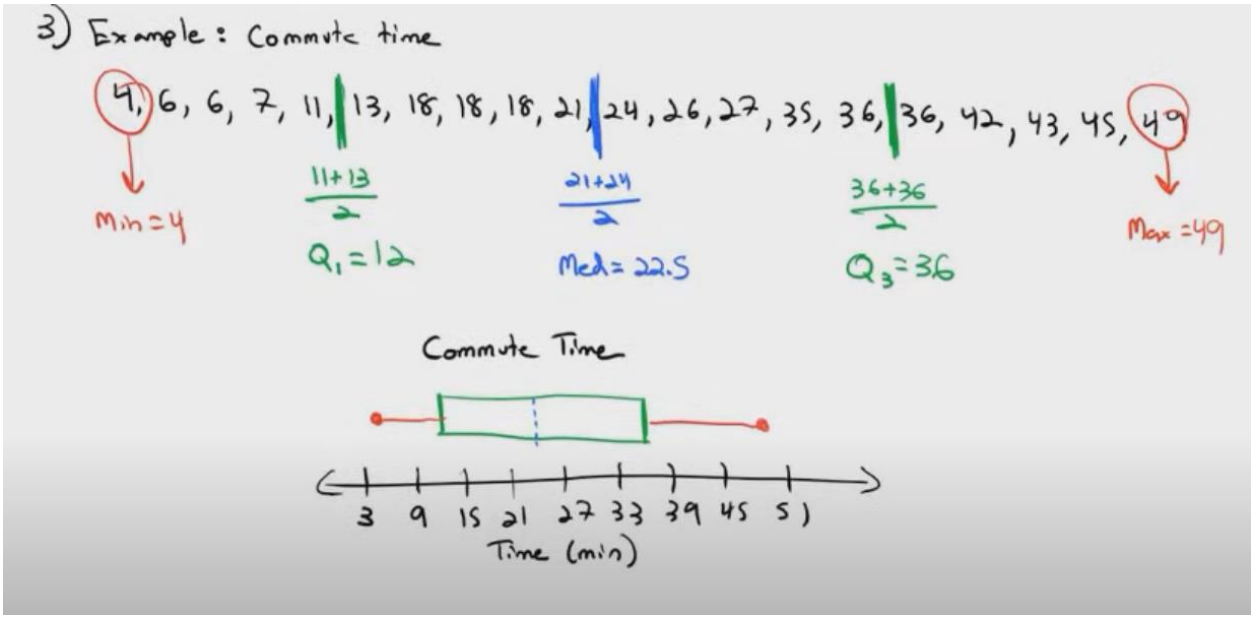
After we get this 5-number summary, we represent it visually with box plot. To box plot -

1.Split the data into quarters

2.q1 and q3 edge of the box

3.whiskers out to min/max

4.dotted line for median

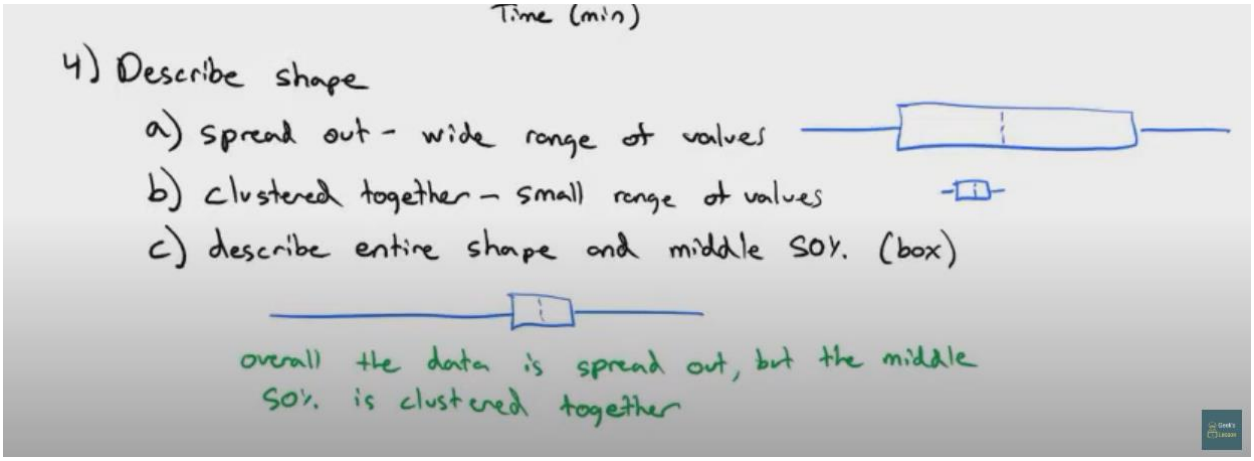


Describing shape of box plot

1. **Spread out** – here we have wide range of values where its box plot is large and covers a large range of values

2. **Clustered together** – here we have small range of values and its box plot is really tiny where everything is really close

3. **describe entire shape and middle 50%** - where value is large and spread out but the middle box plot is small or clustered together

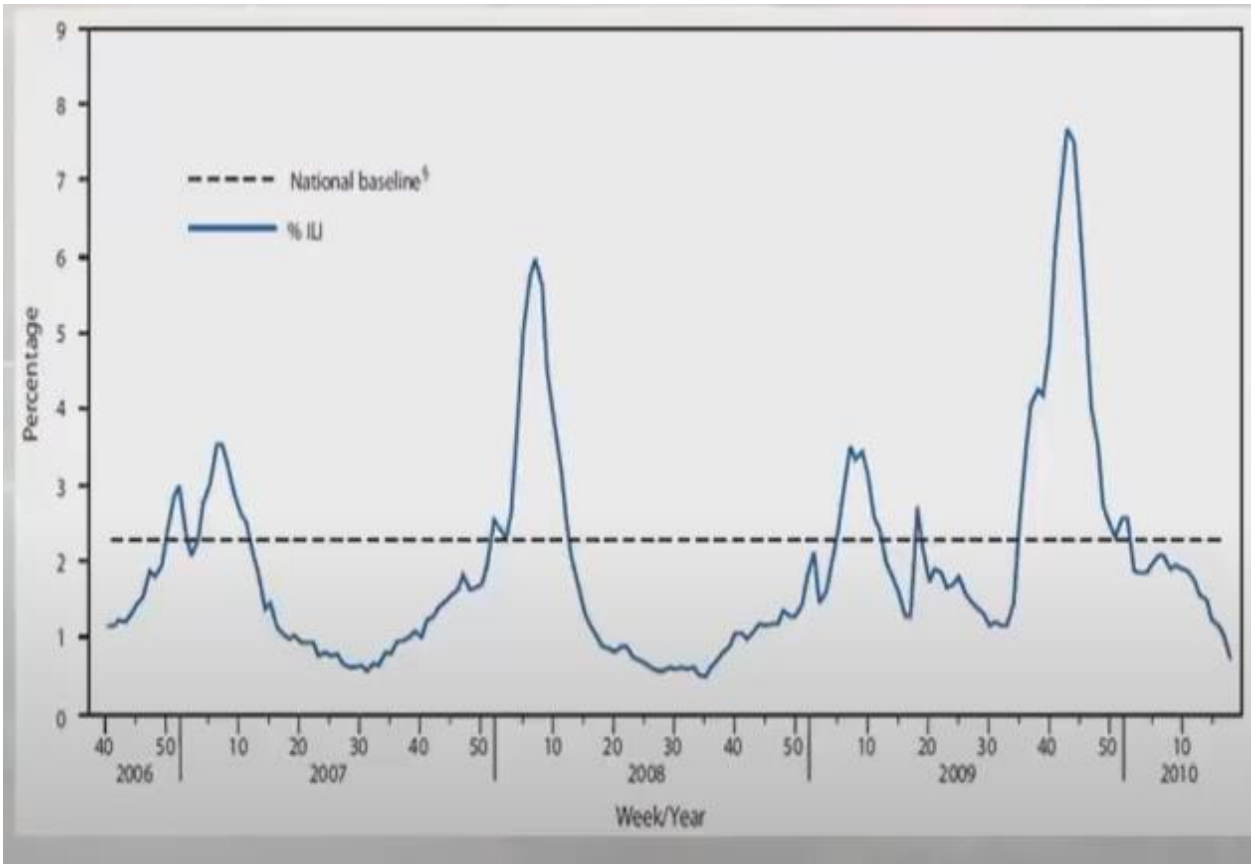


(Graphs)

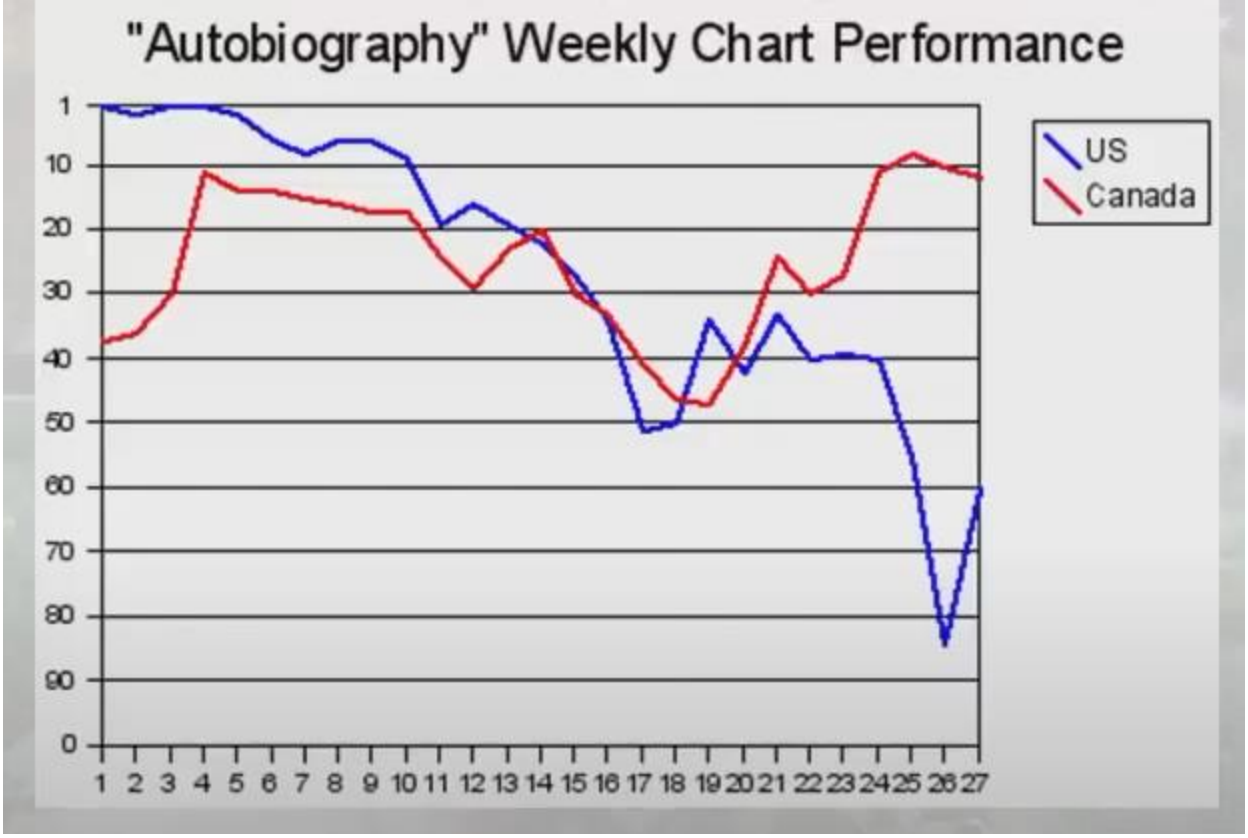
A (Time Series graph)

Time series data are made of measurements for the same variable for the same individual taken at intervals over a period of time and it uses quantitative data where time is plotted on x axis and collected data along y. And to make this graph, we have to build table where time and data related to that time are paired in a class

This kind of time graph where time and value is lot less will look simple whereas if time and data is much, it will look complicated like this



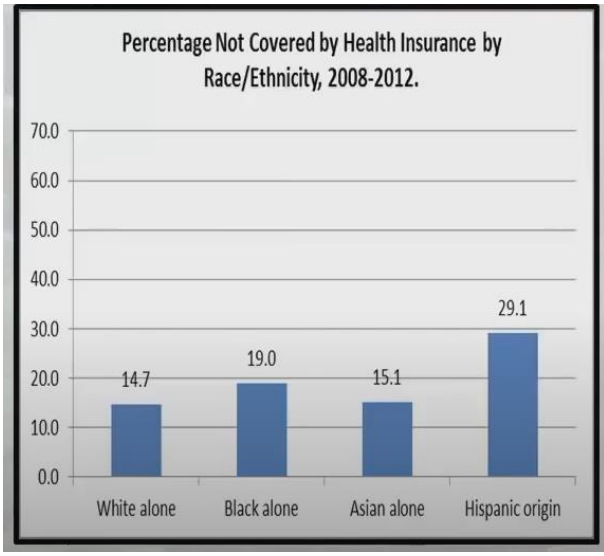
Also, there can be two graphs of different two things plotted at same graph two represent two different values which can help us compare sets or values of two different things

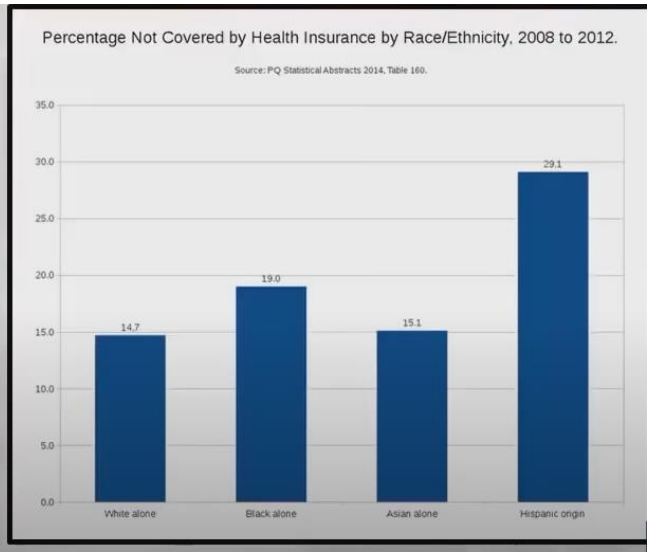


The graphs upward, backward or still trends depend on the graph. This time series graph is used basically for knowing trend over a particular thing over some few periods of time

B (Bar Graph)

Bar charts can be very helpful for visualizing and comparing qualitative and quantitative data. If data is not mutually exclusive then bar graph is a good option. In bar graph, Bars can be vertical or horizontal which has uniform width and spacing (uniform means kind of same space and width between each bar and the bars) and its length represent frequency or percentage of occurrence. The measurement scale for the bars should be same and the graph should include title, bar labels and scale labels on axis or values for each bar so viewer don't get confused. And an important thing to note is that we should always try to take big scales (basically graph max should a bit more than the data max) on y axis to make it look more clear and good though we can also do the other way.

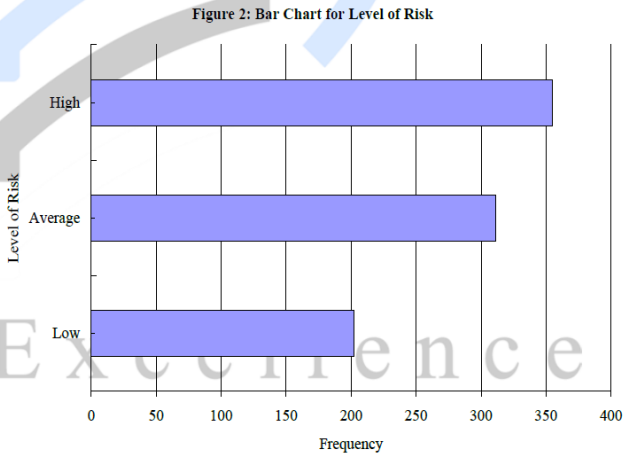




We can do this bar chart like sideways also

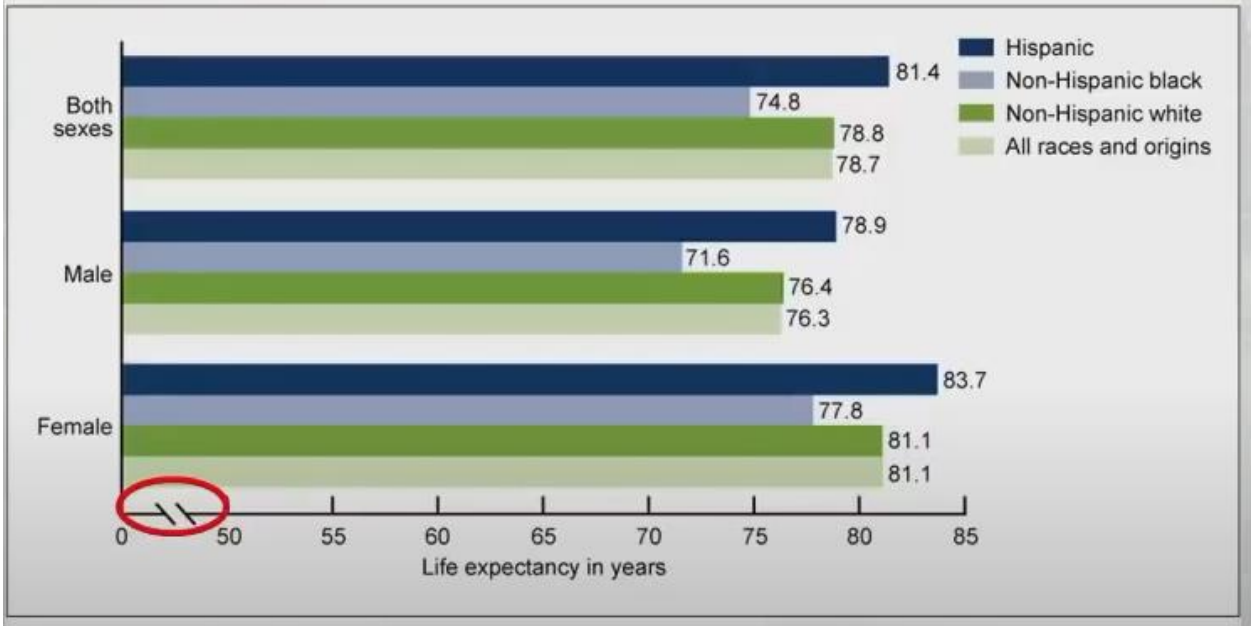
Table 2: Frequency and Percentage Summary table of Risk Level for 868 Mutual Funds

Fund Risk Level	Number of funds	Percentage of funds (%)
Low	202	23.37
Average	311	35.83
High	355	40.89
Total	864	100.00



There can also be clustered bar graph which means more than one bar is graphed for each category

Figure 1. Life expectancy at birth, by Hispanic origin, race for non-Hispanic population, and sex: United States, preliminary 2011



NOTE: Life expectancies for the Hispanic population are adjusted for underreporting of Hispanic ethnicity but are not adjusted to account for the potential effects of reverse migration.
SOURCE: National Vital Statistics System, Mortality.

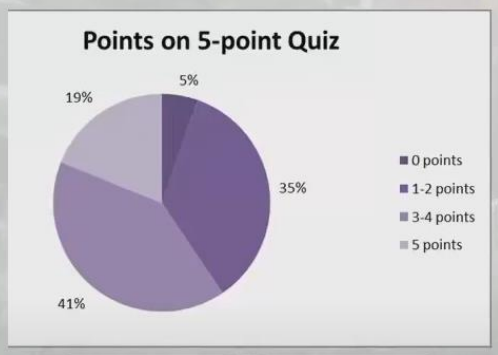
C (Pie Chart)

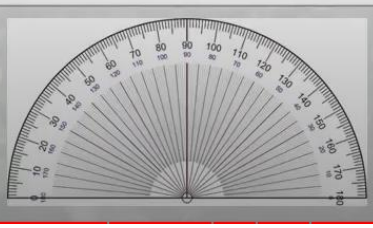
Pie chart (also circle graph) is used with counts of "mutually exclusive" frequencies which is often made in graphing programs. It is used for data's which only fall in one category. Two events are mutually exclusive if they cannot occur at the same time. Suppose someone asks me if I am eating rice, the answer would be either yes or no. Here yes or no is mutually exclusive as both cannot happen at same time

Features of pie chart

Every individual must be put in only one category and it can be qualitative or quantitative variable. If quantitative then first we have to make classes based on those quantitative data then make pie chart graph based on those

Making a Pie Chart





Class	Frequency	Proportion of Circle	Degrees
0 points	2	2/37=0.05	0.05*360° =19°
1-2 points	13	13/37=0.35	0.35*360° =126°
3-4 points	15	15/37=0.41	0.41*360° =146°
5 points	7	7/37=0.19	0.19*360° =68°
Total	37	1.00	360°

Important point to note for pie chart is that its categories must be mutually exclusive and another thing is it is better to showcase percentage using frequencies than showing frequencies directly in pie chart.

D (Choosing graphs for particular data)

Choosing the Right Kind of Graph

Type of Graph	Cases Where Graph is Useful
Frequency Histogram	For quantitative data, when you want to see the distribution.
Relative Frequency Histogram	For quantitative data, when you want to see the distribution. Also, good for comparing to other data.
Stem-and-leaf Display	For quantitative data, when you want to see the distribution. Easier to make by hand than histogram.
Time series graph	For graphing a variable that changes over time and is measured at regular intervals.
Bar graph	For qualitative or quantitative data, and for displaying frequency or percentage.
[illegible]	[illegible]
Pie Graph	For mutually-exclusive categories (quantitative or qualitative).

(Section 3)

(Measures of Centre And Spread)

***How do we summarize data numerically?**

A (Measure of center)

Measure of center means the middle of the data. The measure of center will be either median, mean or mode

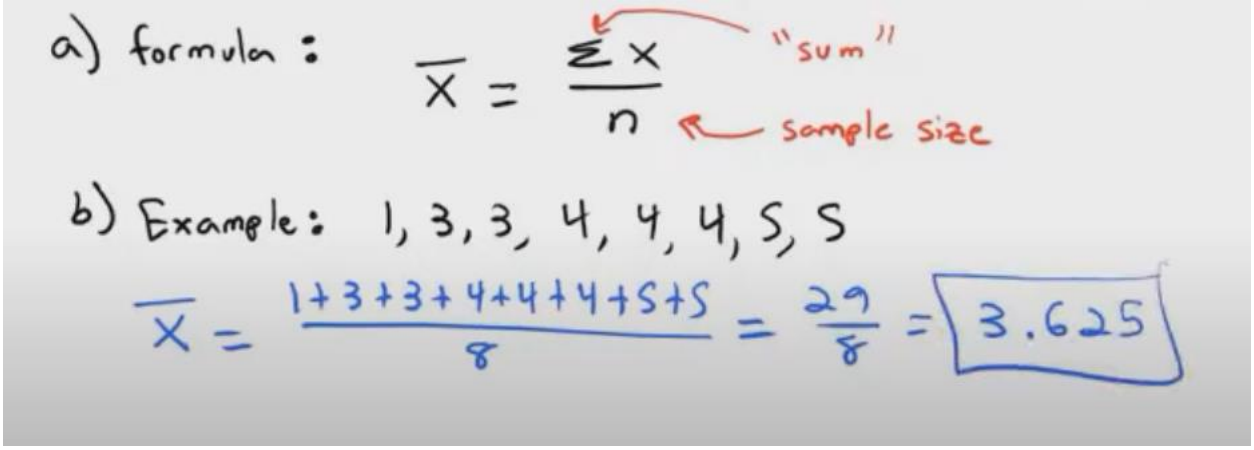
1.**Mean** – mean typically means average. Like if we use population mean we use Greek letter mew and if we use sample mean then we use x bar.

Now there are some different kind of techniques to determine mean from sample population.

Arithmetic mean (Ungrouped data)

In this mean, we simply add all sample population value and divide them by total sample

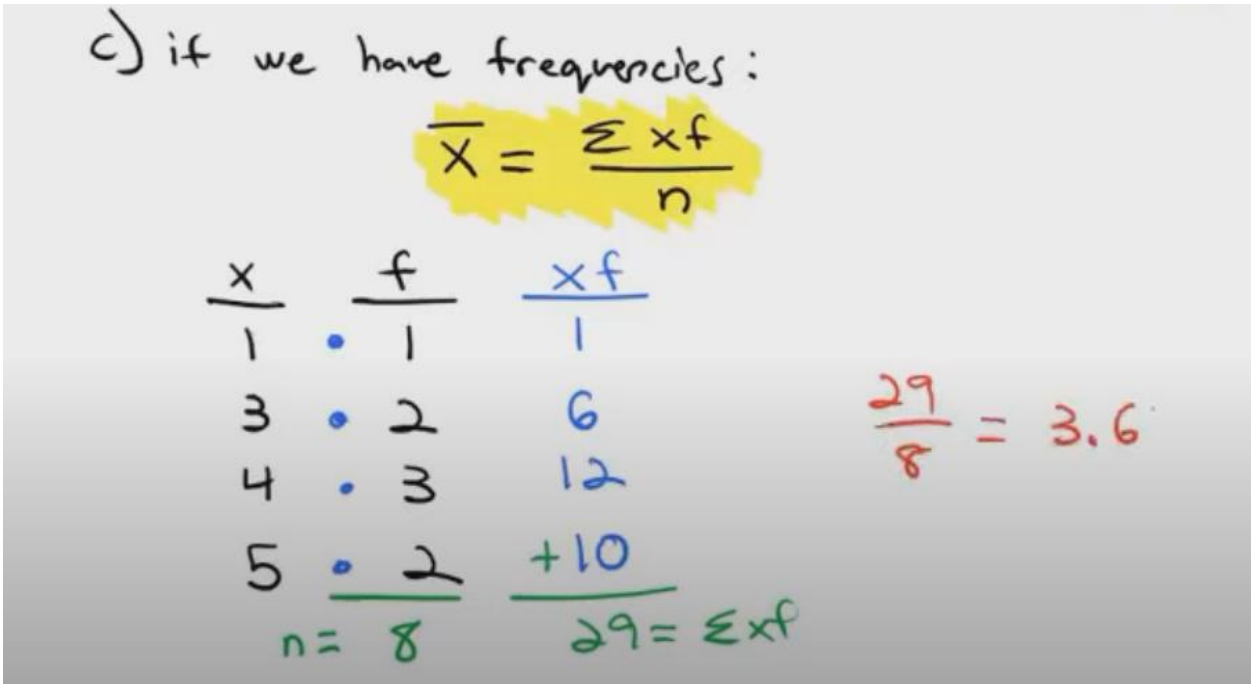
Formula



So, mean equals sum of all values in the sample divide by total sample size.

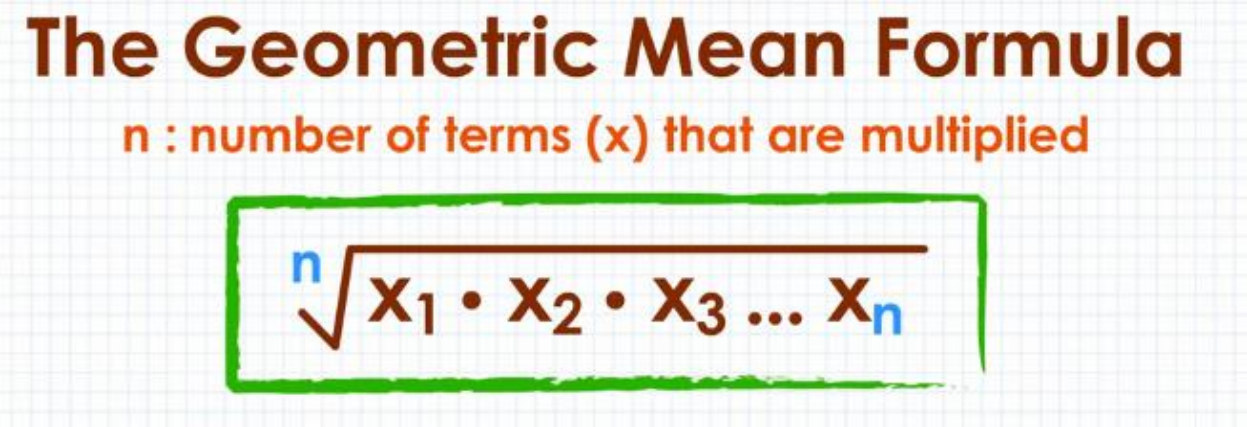
Arithmetic mean (grouped data with frequency)

if frequency is given then we have to first multiply the all the value with its frequency and add them and then divide it with sample size



Geometric Mean (Ungrouped data)

The geometric mean is basically the n root square of n values multiply.



An important thing to note that the ratio of the values with the geometric mean will always be really close values or equal values.

Another thing is that geometric mean will always be less or equal to arithmetic mean

Geometric mean (grouped data)

Class	f	Mid x	log x	f log x
0-10	4	5	0.6990	2.7960
10-20	8	15	1.1761	9.4088
20-30	10	25	1.33969	13.9790
30-40	6	35	1.5441	9.2646
40-50	7	45	1.6532	11.5724
Total	35			47.0208

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

$$= antilog \left[\frac{47.0208}{35} \right]$$

$$= antilog \{1.34345\}$$

$$G = 22.0521$$

2. Find the geometric mean of the following data

x_i	50	63	65	130	135
f_i	5	10	5	15	15

Answer: 96.43

Problem 2.

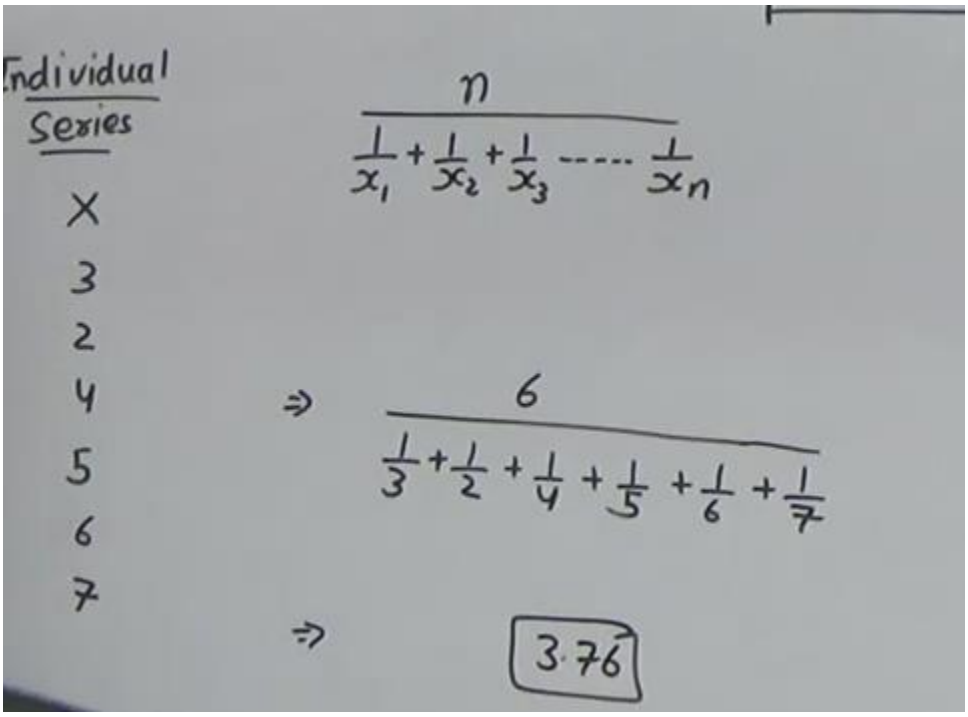
x_i	f_i	$logX$	$f_i * logx_i$
50	5	1.69	8.49
63	10		
65	5		
130	15		
135	15	2.13	31.95
Total	50		99.21

$$GM = Antilog \frac{\sum_{i=1}^{k} f_i log x_i}{n} = 96.43$$

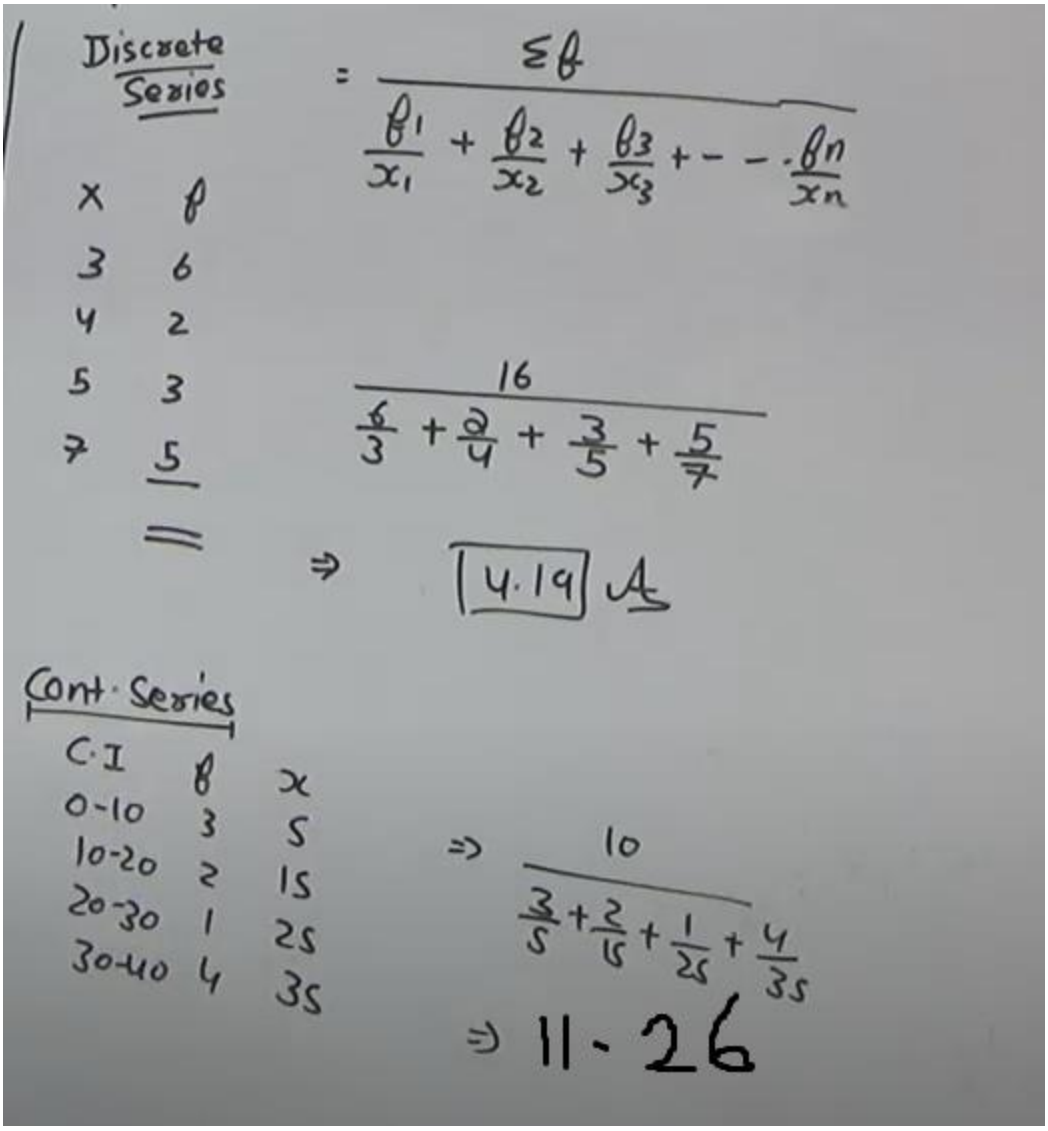
Geometric mean is usually used for dealing with data related to growth rates (like population growth etc.) or interest rates, index number and such

Harmonic Mean (Ungrouped data)

For ungrouped data, we will simply use the formula as shown below in the pic. The basic formula is (n/(1/x1+1/x2+1/x3+1/xN))



Harmonic mean (for grouped data)



Harmonic means are often **used** in averaging things like rates (e.g., the **average** travel speed given a duration of several trips). The weighted **harmonic mean** is **used** in finance to **average** multiples like the price-earnings ratio because it gives equal weight to each data point.

Weighted Mean

The weighted mean is a special case of the arithmetic mean. It occurs when there are several observations of the same value.

A **weighted mean** is a kind of **average**. Instead of each data point contributing equally to the final **mean**, some data points contribute more "weight" than others. If all the weights are equal, then the **weighted mean** equals the arithmetic **mean**.

Formula

$$\overline{X}_w = \frac{\sum(WX)}{\sum W} = \frac{W_1X_1 + W_2X_2 + \ldots + W_nX_n}{W_1 + W_2 + \ldots + W_n}$$

Example:
Madina Construction Company pays its part time employees hourly basis. For different level of employee, the hourly rate are Tk. 50, Tk. 75 and Tk. 90. There are 260 hourly employees, 140 of which are paid at Tk. 50 rate, 100 at Tk. 75 and 20 at the Tk. 90 rate. What is the mean hourly rate paid to the employees?

Answer:
To find the mean hourly rate, we multiply each of the hourly rates by the number of employees earning that rate as follows -

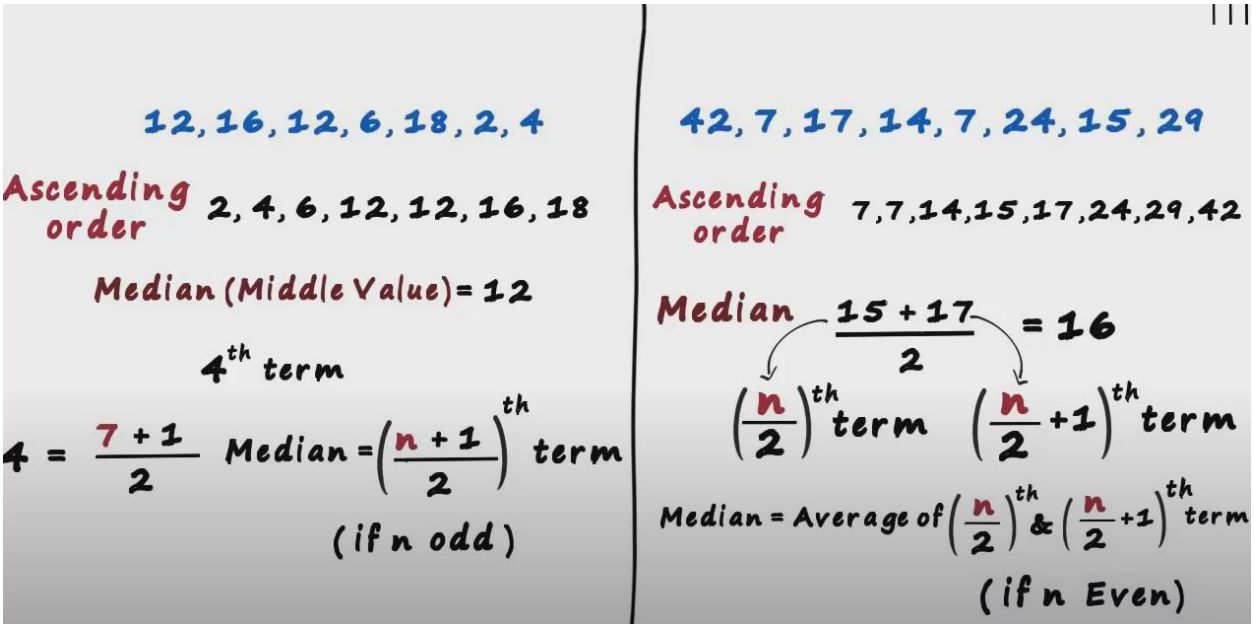
$$\overline{X}_w = \frac{\sum(WX)}{\sum W} = \frac{140 * 50 + 100 * 75 + 20 * 90}{140 + 100 + 20} = \frac{16300}{260} = Tk.\ 62.69\ .$$

The weighted mean hourly wage is Tk. 62.69 or Tk. 63.00 (approximately).

2.**Median** – median is also called “middle” number (after putting in order).

Formula (for ungrouped data)

When n (total sample number) is odd then formula for median is (n+1)/2 when n is even then formula is ((n/2) + (n/2) + 1 term) /2



A good thing about median is one extreme value does not impact median as significantly as mean

Formula (for grouped data)

Median for grouped data:

Me is given by the formula, $Me = L_0 + \frac{\left(\frac{n}{2} - F_{-Me}\right)}{f_{Me}} * W_{Me}$

Where

Me	= Median	f_{Me}	= Frequency of the median class
L_0	= Lower Limit of the median class	W_{Me}	= Width of the median class
F_{-Me}	= Cumulative frequency of the pre median class	n	= Total number of observation

MEDIAN CLASS *is the class that contains* $\frac{n}{2}$ *th observation of the given data.*

3.**Mode** – mode is which thing occur most or the most frequent one. Usually it is used for categories.

Formula (ungrouped data)

like in number 1,3,3,4,4,4,5 4 is mode as 4 occurs most 3 time.

Formula (grouped data)

For grouped data mode is obtained by using the following formula

$$Mo = L_0 + \left\{ \frac{(f_0 - f_{-1})}{(f_0 - f_{-1}) + (f_0 - f_1)} \right\} * W$$

Where,

Mo = Mode

L_0 = Lower Limit of the Modal class

f_0 = Frequency of the modal class

f_{-1} = Frequency of the pre modal class

f_1 = Frequency of post modal class

W = Width of the modal class

B (Measure of Spread or Variation or Dispersion)

Measure of spread tells us not only the middle but how the values are spread also and how the data varies from one data set to another.

Which measures of Dispersion to choose

- When dealing with data, ones' **objective** is ***"only to determine"*** the variation of ***single set of variable / information*** - s/he can / will choose to use **Absolute measure of dispersion**.

- When dealing with data, ones' **objective** is ***"to determine and compare"*** the variations of ***multiple set of variables / information*** having ***expressed*** in ***same / different unit(s)*** - s/he can / will choose to use **Relative measure of dispersion**.

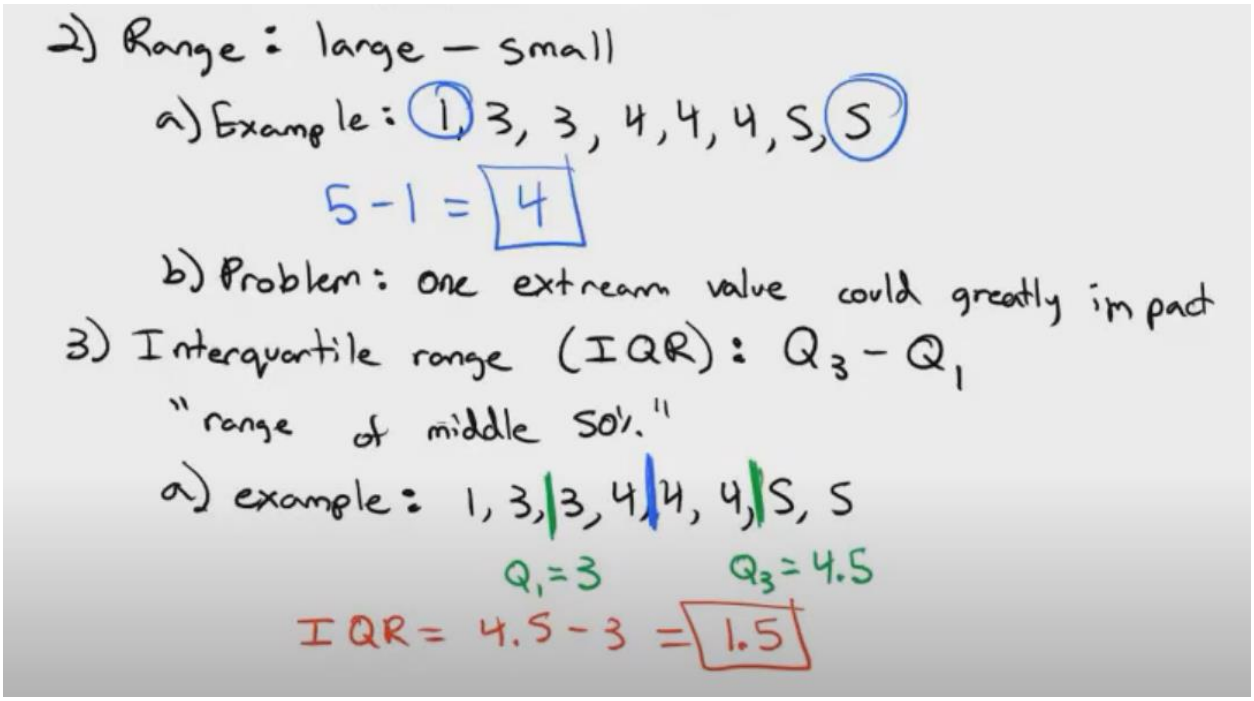
---

ifferent types of Absolute and Relative measure of dispersion are listed below:

Absolute measure of dispersion	Relative measure of dispersion
1. Range 2. Quartile deviation 3. Variance and Standard deviation	1. Coefficient of range 2. Coefficient of quartile deviation 3. Coefficient of variation and standard deviation

1.**Range** – Range tells us about the space between the largest and smallest number. Normally we take the large number and subtract number from it to check how far the numbers are split. Suppose in a class, the highest value is 63 and lowest in 12. So, the range would be 63-12 = 51. The problem with this is one extreme value could greatly impact.

2.**Interquartile range (IQR)** – it basically means the range of middle 50%. We take Q3 and subtract Q1 value from it



However, there is a problem and that is this IQR formula only considers 2 values which are Q1 and Q3. Quartile divide the observations into four equal parts, when observations are arranged in order of magnitudes median, denoted by Q2, is the middle most observation or it is the median and Q1 and Q3 are the median of the lower and upper half respectively

Important to note that

- ***Median*** is the quantity of the variable that ***divides the into 2 equal halves.***
- ***Quartiles (Denoted by $Q1$, $Q2$ and $Q3$)*** are the quantities of the variable that ***divides the total frequencies in to 4 equal parts.***

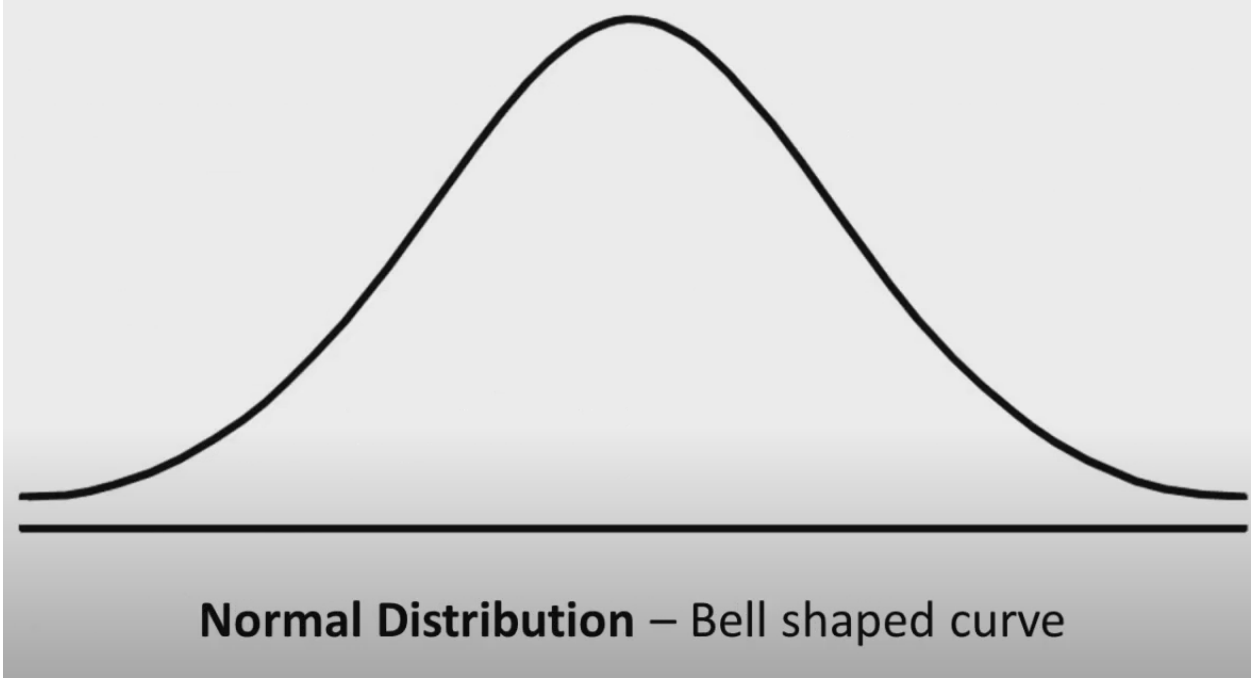
Similarly

- ***Deciles (Denoted by $D1$, $D2$, ..., ..., ..., $D9$)*** are the quantities of the variable that ***divides the total frequencies into 10 equal parts.***

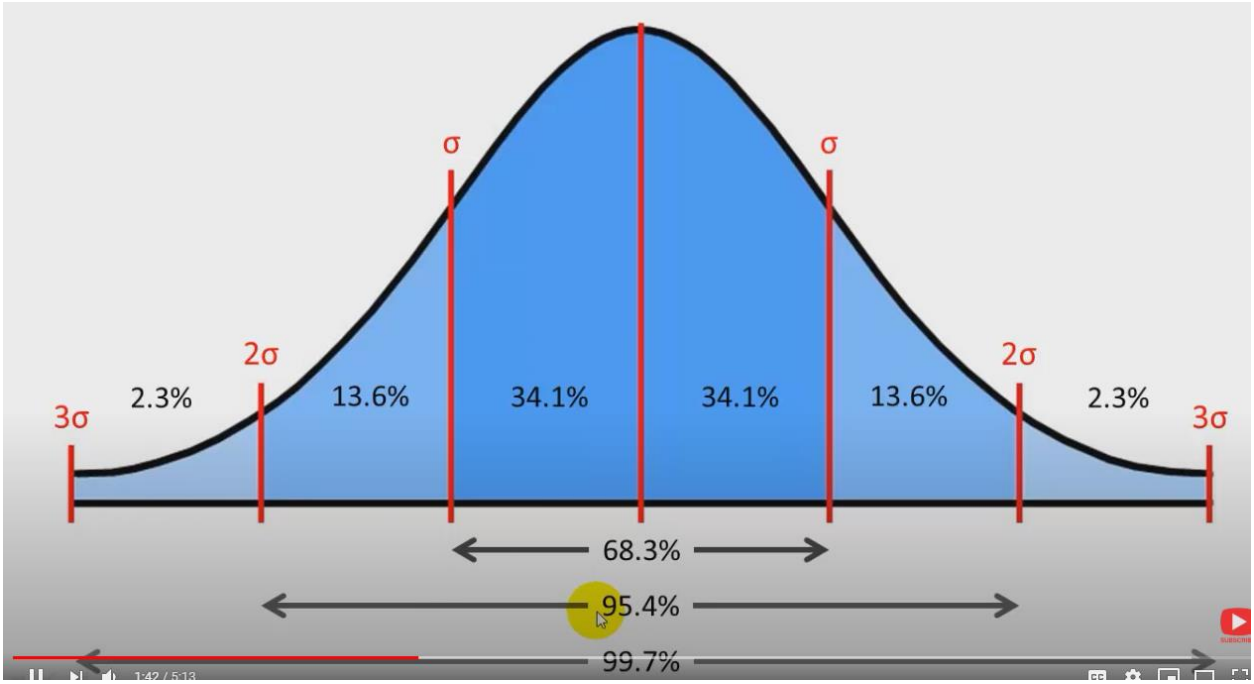
And

- ***Percentiles (Denoted by $P1$, $P2$, ..., ..., ..., $P99$)*** are the quantities of the variable that ***divides the total frequencies in to 100 equal parts.***

3.**Standard Deviation** – it attempts to measure the average distance a point is from the mean. How spread out is the data considering all the data values on average how far are they from the mean. For population we use sigma to express standard deviation and for sample we use s. Low SD means the data are clustered around the mean and high SD means the data is widely spread out. Normal distribution can be called as Bell shaped curve



In normal distribution, the mean, median and the mode is normally all the same



This means +-1 SD normally represents 68.3% of the data, +-2 = 95.4 and +-3 = 99.7 % (+- SD always counted from mean)

The sigma depends on SD. Normally value +- is considered what the value of SD is. An example is

Problem:

The following data give the number of passengers travelling by airplane from one city to another in one week.

115 122 129 113 119 124 132 120 110 116

Calculate themean and standard deviation and determine the percentage of class that lie between $(i)\mu \pm \sigma$, $(ii)\mu \pm 2\sigma$ and $(iii)\ \mu \pm 3\sigma$. What percentage pf cases lie outside these limits?

Solution:

The calculation for mean and standard deviation are given in the following table

x	$x-\mu$	$(x-\mu)^2$
115		
122		
129		
113		
119		
124		
132		
120		
110		
116		

$$\mu = \frac{\Sigma x}{N} = \frac{????}{??} = 120 \text{ and } \sigma^2 = \frac{\Sigma(x-\mu)^2}{N} = ??? = 43.6$$

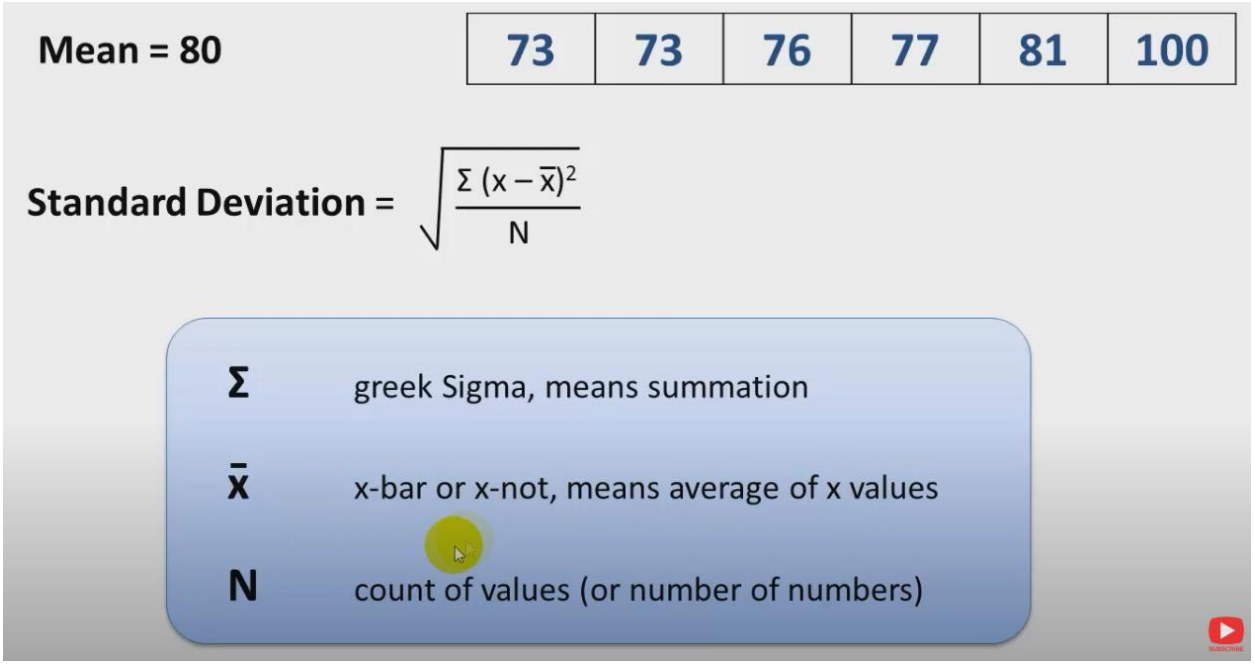
Therefore, $\sigma = \sqrt{\sigma^2} = \sqrt{43.3} = 6.60$

The percentage of cases that lie between a given limit are as follows:

Interval	Values within Interval	Percentage of population	Percentage falling Outside
$\mu \pm \sigma = 120 \pm 6.60$ = 113.4 and 126.6	115, 116, 119, 120, 122, 124	60%	40%
$\mu \pm 2\sigma$ =120 $\pm$ 2(6.60) = 106.80 and 133.20	110, 113, 115, 116, 119, 120, 122, 124, 129, 132	100%	Nil

Formula of standard deviation

First, we need to count mean (x bar) and then the formula is



Mean = 80

73	73	76	77	81	100

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma (x-\bar{x})^2}{N}} = \sqrt{\frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots (x_n-\bar{x})^2}{N}}$$

$$= \sqrt{\frac{(73-80)^2 + (73-80)^2 + (76-80)^2 + (77-80)^2 + (81-80)^2 + (100-80)^2}{6}}$$

$$= \sqrt{\frac{524}{6}} = 9.3$$

if we have frequency formula would be

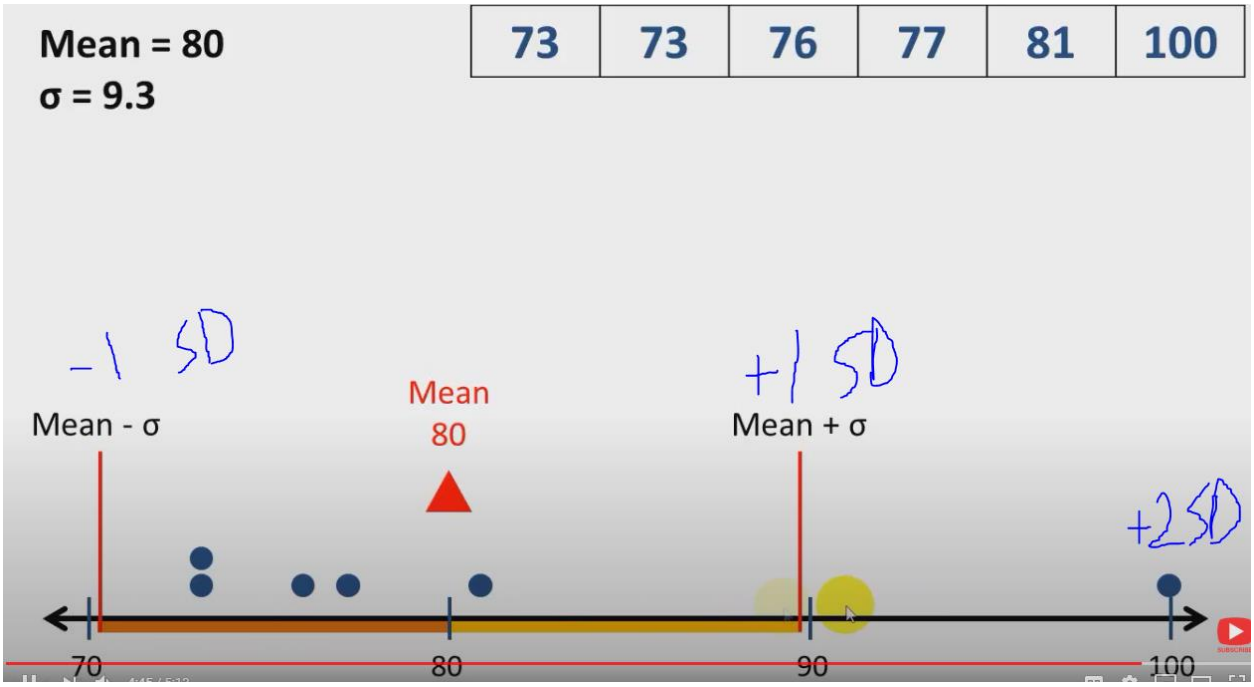
c) If we have frequencies: $s = \sqrt{\frac{\Sigma (x-\bar{x})^2 f}{n-1}}$

d) example:

x	f	$(x-\bar{x})$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
1	1	-2.625	6.89	6.89
3	2	-0.625	.39	.78
4	3	0.375	.14	.42
5	2	1.375	1.89	3.78

$(\bar{x} = 3.625)$

$\Sigma (x-\bar{x})^2 f = 11.87$



So, it does not follow normal standard deviation curve and within the +-1 SD range from mean, most data in presented

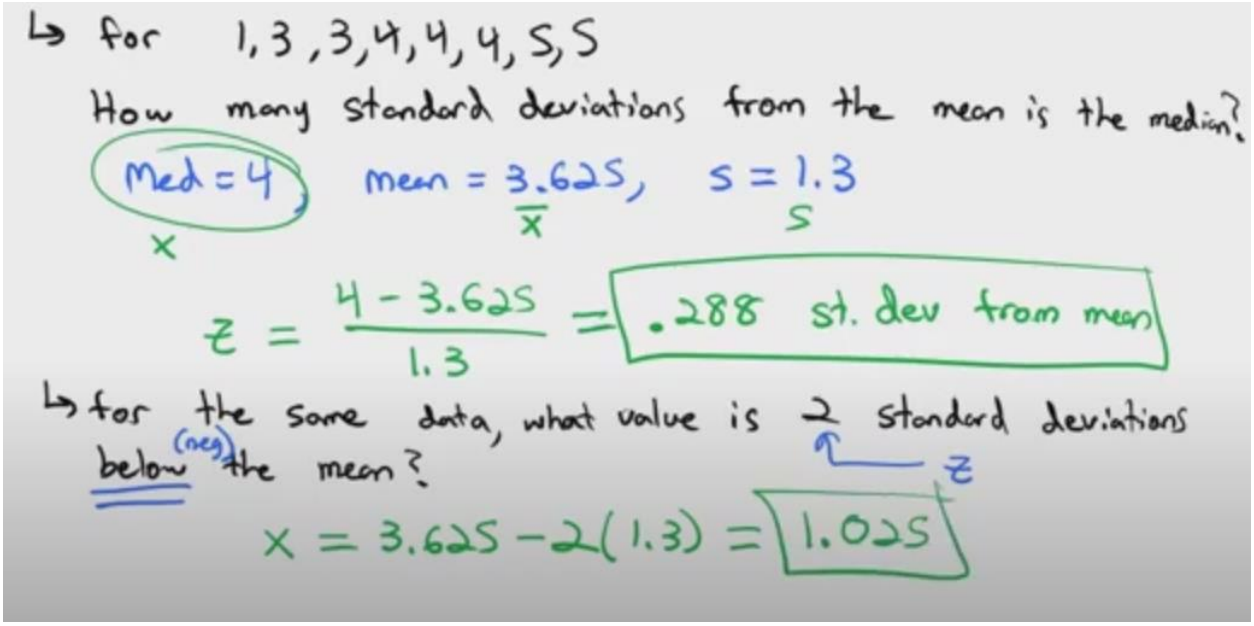
Standard deviations measure distance from the mean

In statistic, to represent the number of standard deviations from the mean, we use z.

Formula

The left side formula is to find the number of deviations from the mean so it takes the value we working with so we know how many standard deviations x is from the mean

The right side formula is to find particular standard deviations from the mean like what is 3 SD , 4 SD etc deviations from the mean. and x is that value



Here in the second example, z is given and the sign before (ZS) is considered as negative because it tells to show SD below of the mean but if it was above then the sign would have been positive or plus

Skewness

The term skewness means the lack of symmetry. The skewness may be either positive or negative. When the skewness is positive the associated distribution is called positively skewed. When the skewness is negative the associated distribution is negatively skewed.

Method 1	If for a distribution $Mean > Median > Mode \Rightarrow$ The distribution is positively skewed $Mean < Median < Mode \Rightarrow$ The distribution is negatively skewed $Mean = Median = Mode \Rightarrow$ The distribution is symmetric
Method 2	Pearson's coefficient of skew ness $Sk_p = \frac{Mean - Mode}{SD} = \frac{3(Mean - Median)}{SD}$ Then if $Sk_p > 0 \Rightarrow$ The distribution is positively skewed $Sk_p < 0 \Rightarrow$ The distribution is negatively skewed $Sk_p = 0 \Rightarrow$ The distribution is symmetric.

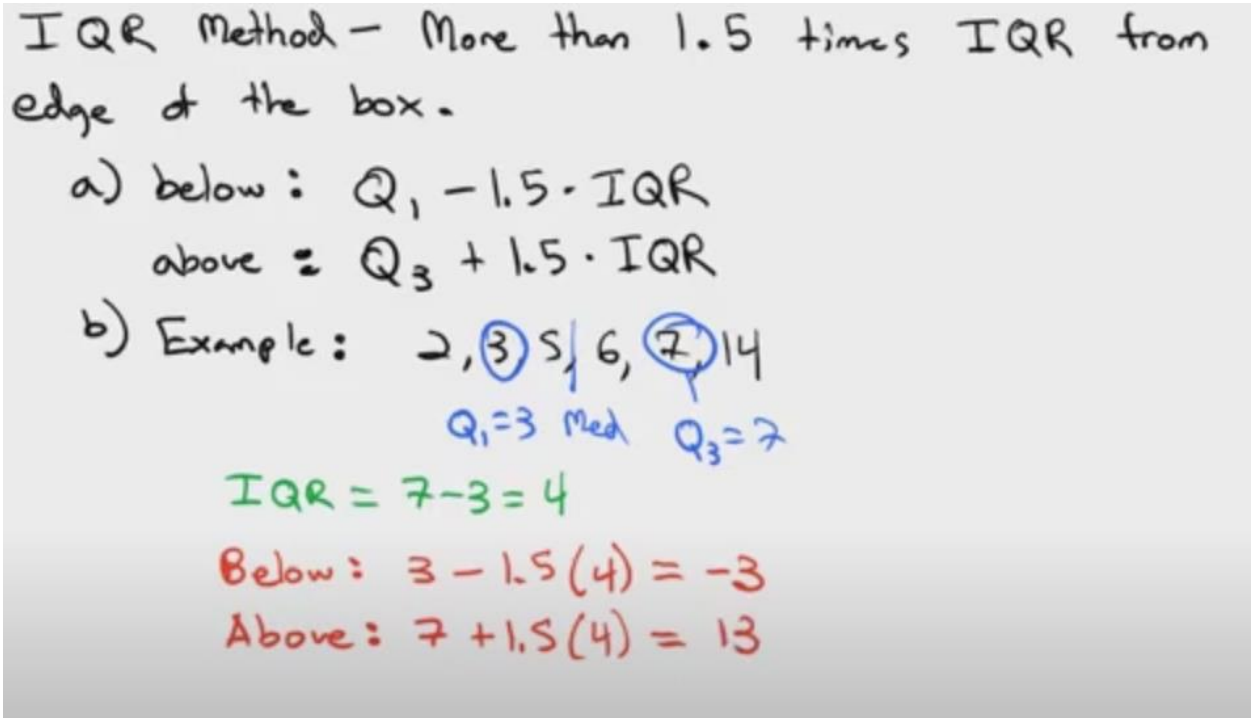
Always try to use method 1 as method 2 is a bit lengthy and it creates hassle sometimes

4.Outliers

As we know from above, 95 percent data falls into two SD of the mean but if it does not then those values are unusual or extreme which are called outliers. In other word, values far removed from the rest of data
like in the number 1,2,3,4,4,4,4,5,86 here 86 is outlier

*How to determine outliers?

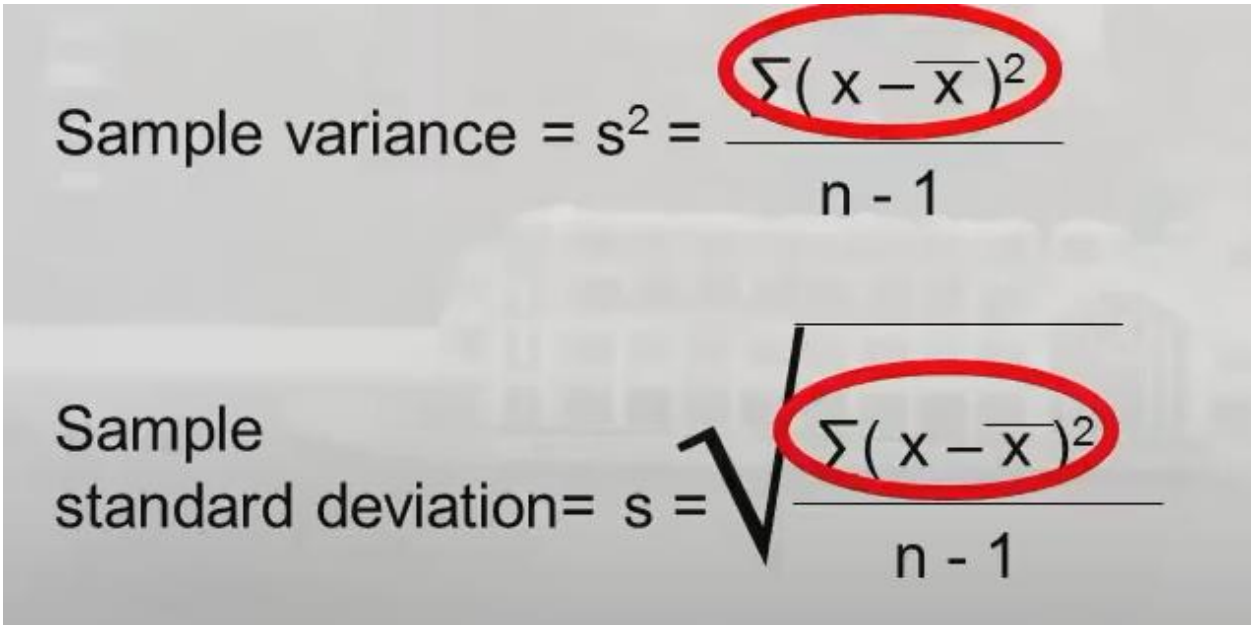
There are two ways but I would only talk about IQR method as both methods provide almost same results



Here the formula is if the data is below Q1 – 1.5 X IQR or above Q3 + 1.5 X IQR then the data is considered as outlier. Here below -3 and above 13 data are considered as outlier and we have 14 in our data set. So here 14 is outlier data

5.Sample Variance

There is another thing called sample variance which is nothing but the square of standard deviation

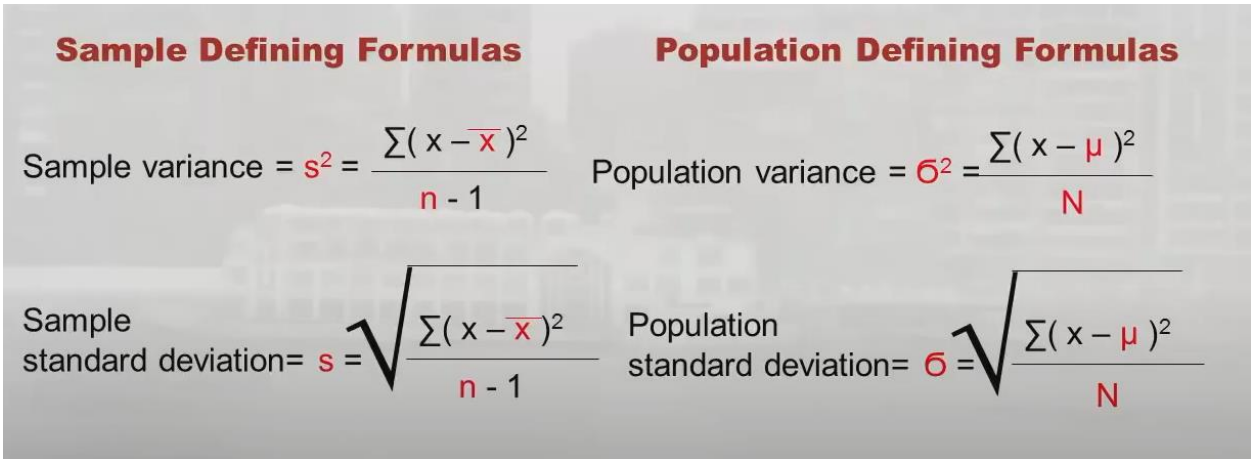


6.Population variance and deviation

Now as we know normally, we do most of the work with samples in statistics but sometimes if we work with population then the standard deviation and population variance is the same as sample just, we use mew in change of x bar.

Formula

The below picture is for ungrouped data



But if we have grouped data then

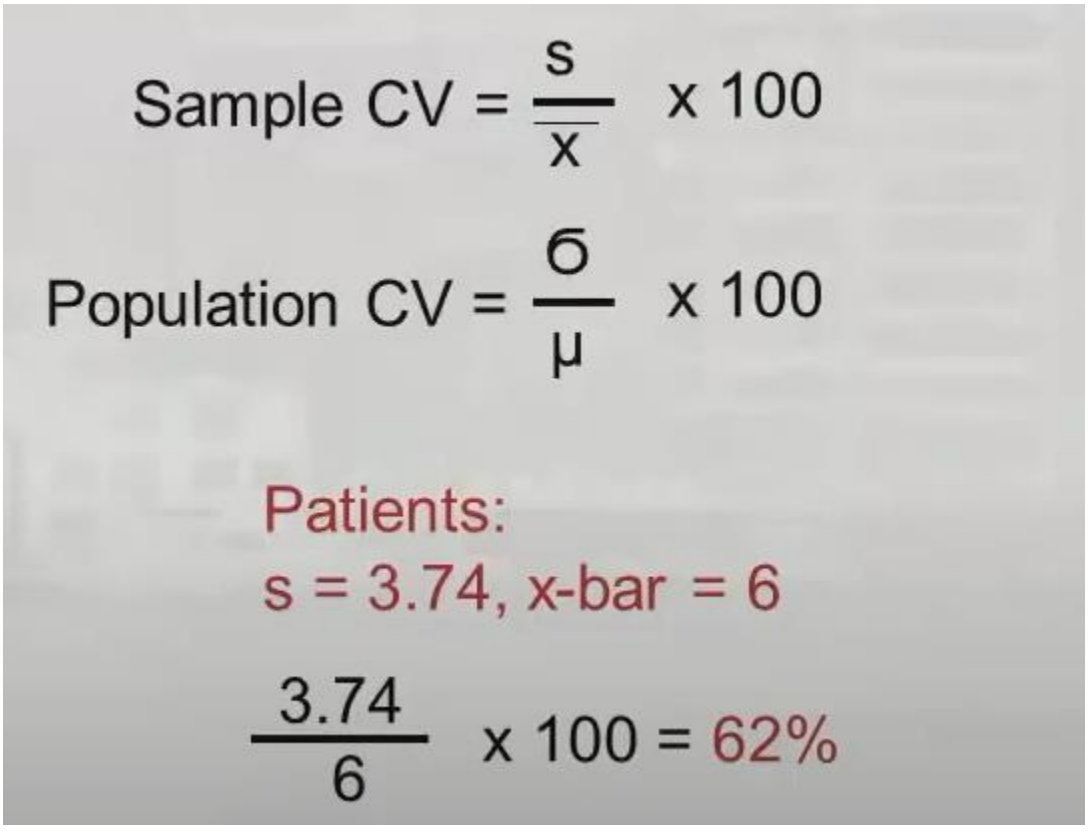
$$S^2 = \frac{\sum f_i \left(X_i - \bar{x}\right)^2}{n-1}$$

7.Coefficient of Variation-

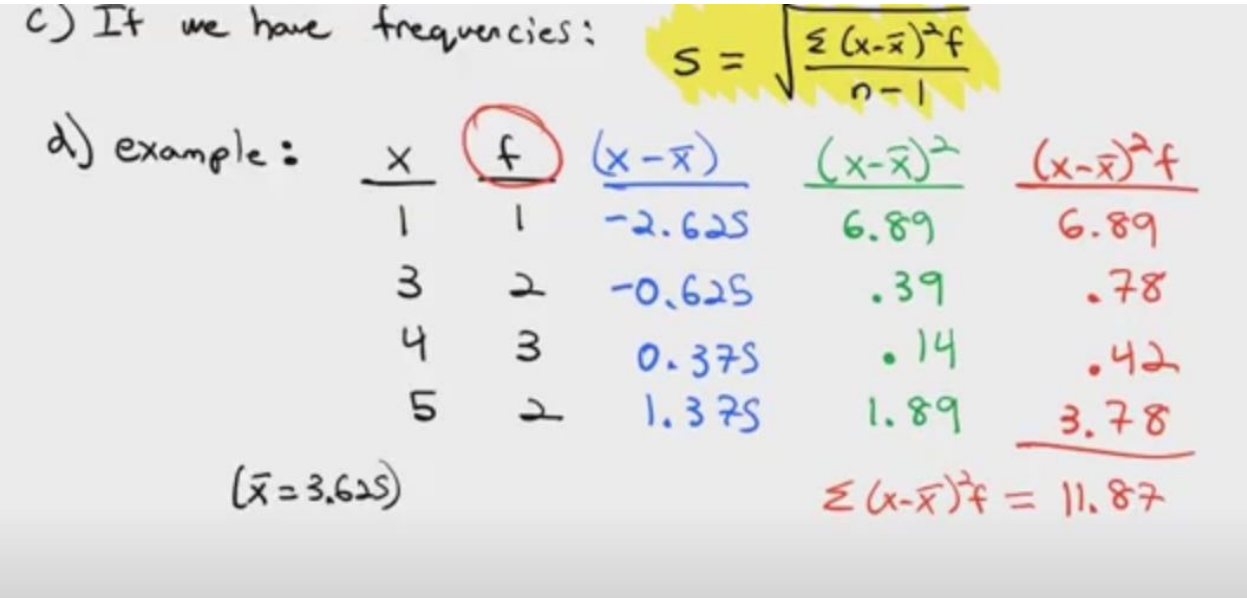
CV is the short form of Coefficient of variation. It shows how much the data varies compared to the mean. It should always be expressed in %. There are other coefficients but for now we will learn only about “CV”

Formula

The below formula is for ungrouped data



But if we have a grouped data then we have to determine SD as per rule of grouped data shown before but I again will show it how to get SD of grouped data



Then using that SD, we will determine the CV but the formula will be the same which is (SD/N-1). Don't forget x bar here is the mean data

The CV is the measure of the spread of the data relative of the average of the data.

In the first sample, the s (sample deviation) is only 50% of the mean and in the second sample it is 62%

8. Coefficient of Range-

It is the ratio of difference between two extreme items which are largest and the smallest value of the distribution to their sum

Formula

$$\text{Coefficient of range} = \frac{X_l - X_s}{X_l + X_s}\text{; Where } X_l = \text{Largest value and } X_s = \text{Smallest value}$$

9. Coefficient of Quartile Deviation-

I have talked about quartile deviation above in the IQR (inter quartile range) part. I will just showcase the formula here

$$\text{Coefficient of } QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

10.**Kurtosis**-

It is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. It helps to determine if the curve is more or less high as compared to the normal curve

(Section 4)

(Correlation Analysis and Regression)

A (Correlation)

Correlation analysis is a method used to find the relationship between variables. It enables us to measure or quantify the relationship between the pairs of variables. Correlation comes from two words where “CO” means “Together” and “relation” means “connection”. It also shows whether and how strongly pairs of variables (two variables) are related

Important thing to note

1. The value of the correlation coefficient r lies between -1 and +1;

2. The closer the value of r to either +1 or -1, the more strongly the two variables are related.

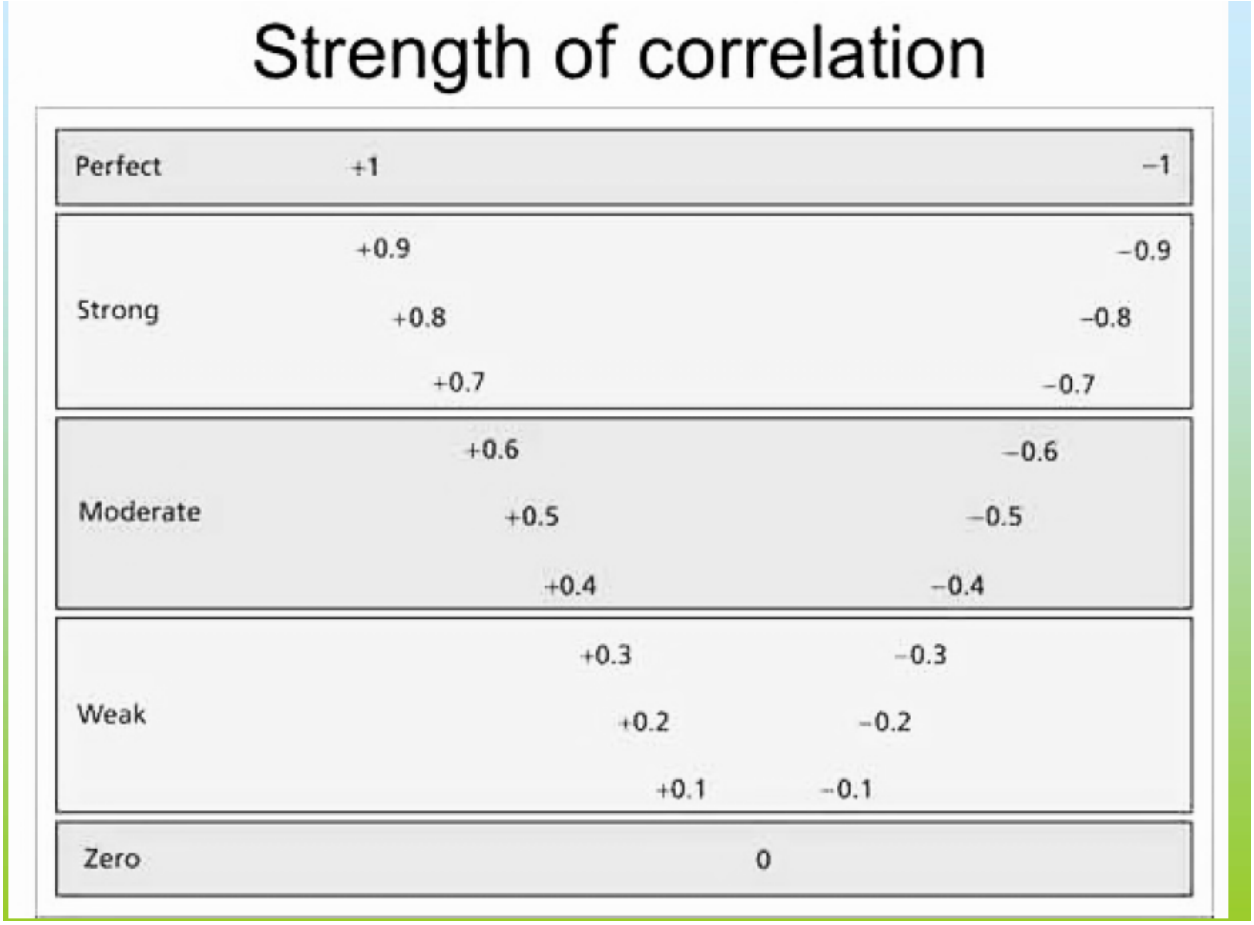
3. When r is positive, one variable tends to increase as the other increases or vice versa;

4. When r is negative, one variable tends to decrease as the other increases or vice versa

5. When r = +1, it means there is perfect positive correlation between the variables;

6. When r = -1, it means there is perfect negative correlation between the variables

7. When r = 0, the variables are said to be uncorrelated.



If its + then its positive strong, moderate, weak correlation or whatever the number table says about its bond and if its – then its negative. Suppose if CR (correlation) is .56 then its positive moderate CR but if its -.9 then its negative strong CR (I used CR to keep it short but you should use the full term always).

Formula

The formula for correlation is

$$r = \frac{\sum x_i y_i - n\overline{xy}}{\sqrt{\left(\sum x_i^2 - n\overline{x}^2\right)\left(\sum y_i^2 - n\overline{y}^2\right)}}$$

Now let's check an example using this formula

Find the correlation between age and weight and comment

Age (years)	Height (inches)
6	40
7	42
8	45
9	47
10	49
11	52
12	55
13	59
14	62
15	65

Age (years) x	Height (inches) ,y	xy	x^2	y^2
6	40	6×40=240	6×6=36	40×40=1600
7	42	7×42=294	7×7=49	42×42=1764
8	45	360	64	2025
9	47	423	81	2209
10	49	490	100	2401
11	52	572	121	2704
12	55	660	144	3025
13	59	767	169	3481
14	62	868	196	3844
15	65	975	225	4225
$\sum x=105$	$\sum y=516$	$\sum xy=5649$	$\sum x^2=1185$	$\sum y^2= 27278$

n =10

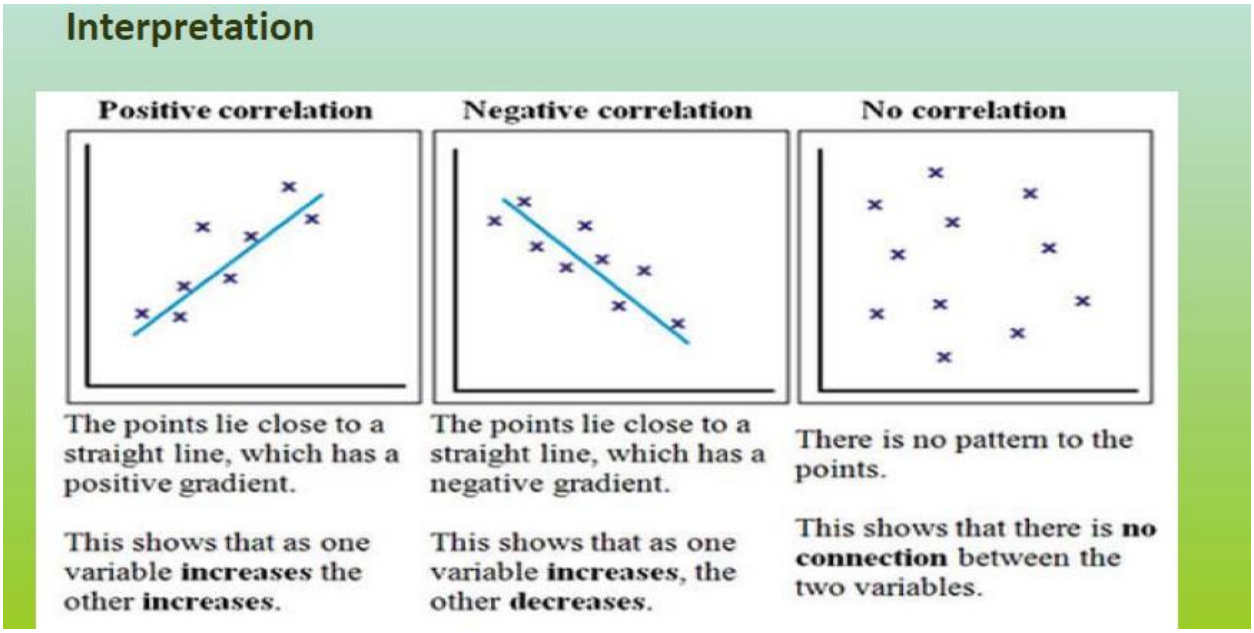
$$\bar{x} = \frac{\sum x}{n} = \frac{105}{10} = 10.5 \text{ and } \bar{y} = \frac{\sum y}{n} = \frac{516}{10} = 51.6$$

Now if we use these data in the formula, r will be .93 which is a positive strong correlation

(Scatter diagram)

Scatter diagram or scatter plot, is a helpful way to visualize a relationship to identify patterns between two variables.

In scatter diagram the independent variable x scaled on the horizontal axis and the dependent variable y on the vertical axis



Scatter plot or graph and correlation chart have same connection. We use CR to detect the connection and use scatter plot or graph to visualize the pattern or connection

Now let's check two examples for scatter graph (first example for positive CR and second for negative)

Example 1*:** Let ***x and ***y*** represent the two variables. Examine the relationship by scatter diagram.

Income ('000' tk) X	Expenditure('000' tk) y
5	4
10	8
15	12
20	18
30	20
50	45

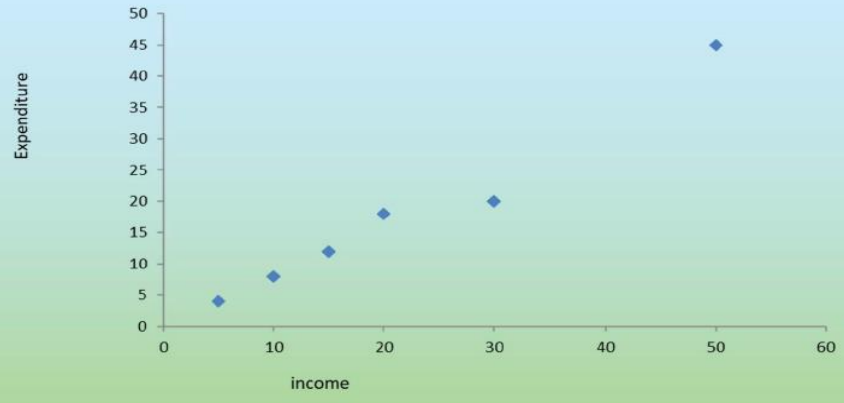


Figure: Scatter diagram between income and expenditure

The scatte diagram shows that there exists a ***positive*** correlation. We say that there is positive relationship between income, ***x*** and the expenditure ***y*** which means for higher income expenditure will also be higher or vice-versa.

Example 2: The price and amount of demand is given below. Draws scatter diagram and comment

Price ($)	Demand (Kg)
20	4
19	4
18	5
16	6
14	6
12	8
10	9
10	9
9	10
8	10

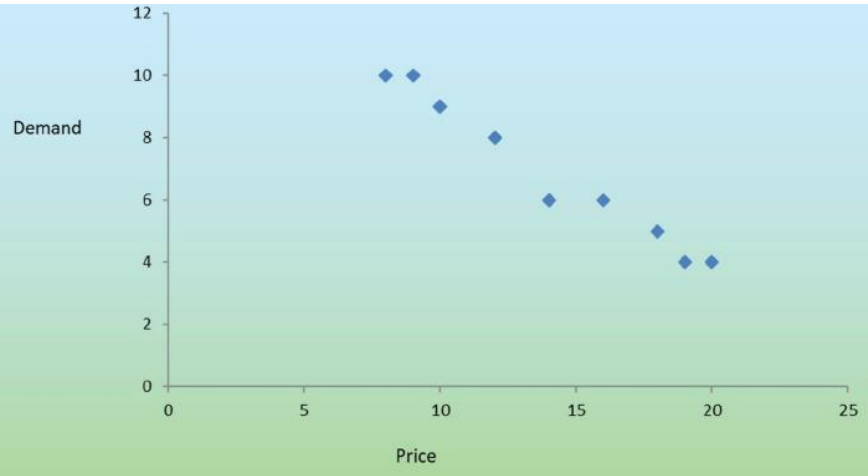


Figure: Scatter diagram between price and demand

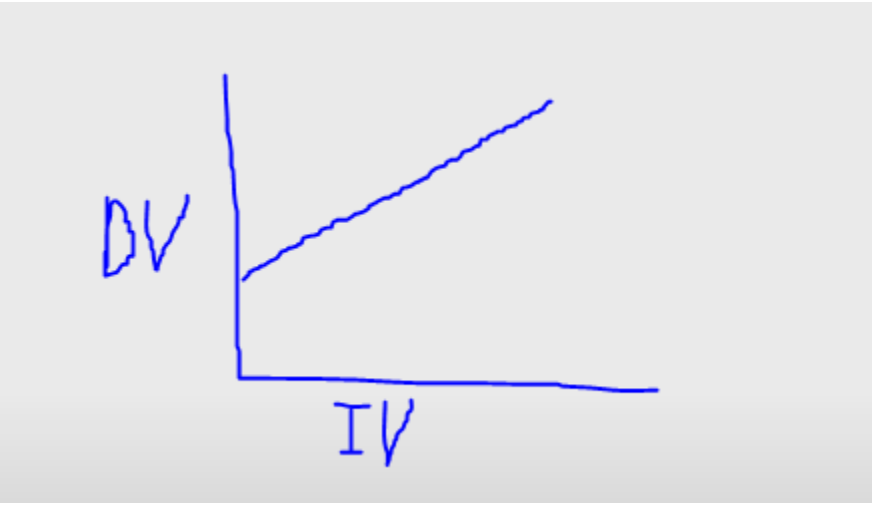
In this case, the correlation is ***negative***. There is negative relationship between the price, ***x*** and the amount demand, ***y***, when price increase, demand decreases

B (Regression)

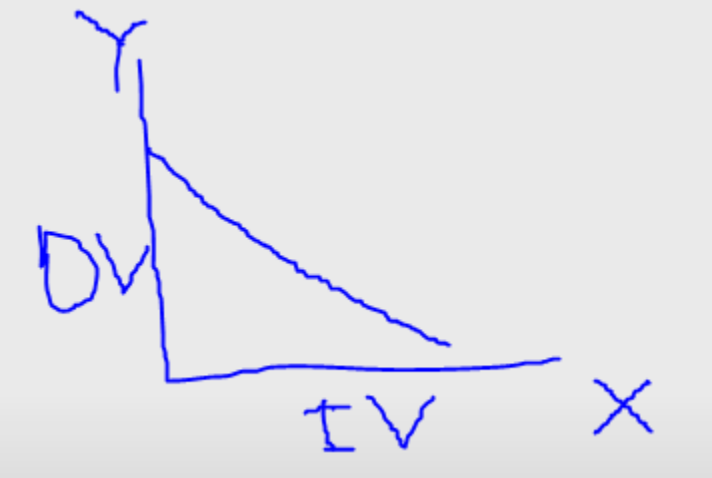
Regression describes the relationship between two variables based on observed data. It is the study of dependency. The dependent variable is always drawn on the y axis and the independent on the x axis.

Important thing to note

If the independent variable increases as with as the dependent variable or vice versa then we can say the relationship between the two is positive and the graph will look like this (here iv means independent variable and dv means dependent variable)



On the other hand, if the independent variable increases and the dependent decreases then it has a negative relationship and the graph will look like this



Regression model

The linear regression model is the single most useful model for prediction. The basic linear regression model of y on x can be written as

$$y = \alpha + \beta x + \varepsilon$$

1.Here y is dependent or explained and x is Independent or explanatory variable

2. α is intercept, value (or average value) of y when x is absent (zero)

3. β is the slope coefficient, measures the average change (increase/decrease) in y for a unit change (increase/decrease) in x.

4. the coefficients α and β are unknown parameters, known as regression coefficients

5. ε is known as random error term. The disturbance term ε represents all those factors that affect the dependent variable but are not taken into account.

Formula

$$\hat{\beta} = \frac{\sum xy - n\overline{x}\,\overline{y}}{\sum x^2 - n(\overline{x})^2}$$

And $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

The estimated/fitted regression equation is

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Let's check an example

Cigarettes	5	23	25	48	17	8	26	35	4	23	11
Longevity	80	78	60	53	85	84	79	72	92	65	81

i. Can you establish the relationship between number of cigarettes and longevity ?

ii. Fit a regression line of longevity on number of cigarettes

Here for question 2 no. answer

ii. It is required to fit a regression model where

Dependent variable y=longevity and

Independent variable, x= Cigarettes

By the method of ordinary least square (OLS)

$$\hat{\beta} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{15683 - 11 \times (20.45)(75.36)}{6403 - 11 \times (20.45)^2} = -0.704$$

and

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ = 75.36 – (-0.704)(20.45) = 89.76

Thus the fitted regression model is

$\hat{y} = \hat{\alpha} + \hat{\beta}x$ = 89.76 + (-0.704).(x)

And the interpretation or meaning of this is

Interpretation of the model

$\hat{\alpha}$=89.76 means (y) the longevity will be 89.76 years, when the number of cigarettes smoked per day (x) is 0

It implies that who did not smoke cigarette at all, on an will live up to 89.76 years

$\hat{\beta}$ =-0.704,

As it is negative, the longevity (y) will decrease

So $\hat{\beta}$ =-0.704means (y) the longevity will decreased 0.704 years, when the number of cigarettes smoked per day (x) is increased by 1

iv.

if the number of cigarettes smoked per day is 5 , that is x=5, then the longevity (y)is

$$\hat{y} = 89.76 + (-0.704)x = 89.76 + (-0.704) \times 5 = 86.27\ years$$

Similarly you can check if the number of cigarettes smoked per day is 10 , that is x=10, then the longevity (y) is

$$\hat{y} = 89.76 + (-0.704)x = 89.76 + (-0.704) \times 10 = 82.72\ years$$

Coefficient of determination formula

The coefficient of determination is calculated as

$$R^2 = r^2$$

Where r is correlation coefficient.

The coefficient of determination R2 is one of the important tools to verify the strength or fitness of the model. We got that r= -0.805, then R2 = 0.65. We can say that 65% of the variation in the longevity that is age at death is explained by taking into account the average number of cigarettes smoked per day. Or, we can say that 65% of the variation in age at death is explained by the average number of cigarettes smoked per day

(Section 5)

(Probability)

*How do we calculate basic probabilities?

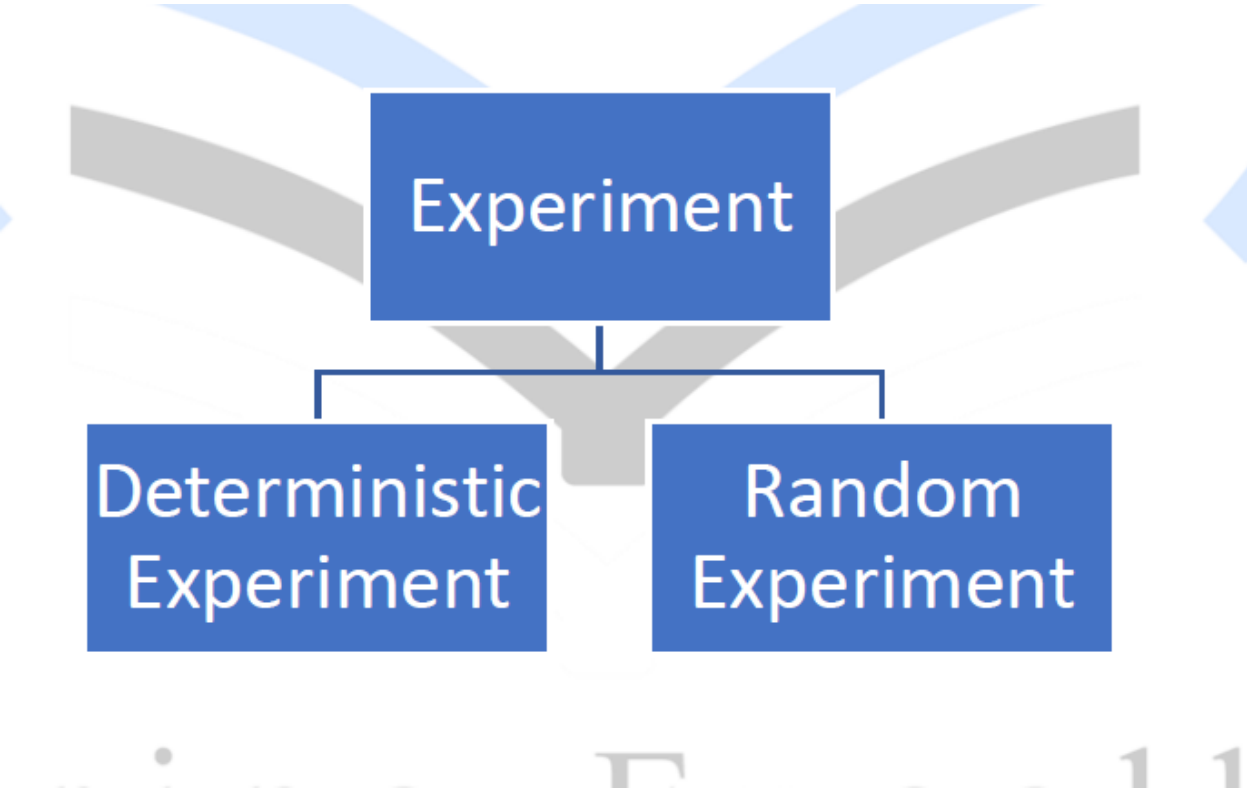
A (Basic Vocabulary)

*what is an experiment?

An experiment is a procedure which can repeated infinitely but it has a well-defined set of possible outcomes.
Tossing a coin is an example of experiment since it can be infinitely repeated and it has defined set of possible outcomes which are “heads” & “tails”.

Measuring the lifetime of an automobile is also an example of experiment and its sample space consists of all nonnegative real numbers. That is,
$S = [\ 0, \infty)$

*How many experiments are there?



1.Deterministic Experiment:

The experiments which have only one possible result or outcome, that is, whose result is certain or unique are called deterministic or predictable experiments.
The results of these experiments are known with certainty and is known prior to its conduct

An experiment conducted to verify the Newton's law of motion and an experiment conducted to verify the Economic Law of Demand are examples of deterministic or predictable experiment.

2.Random Experiment:

A Random Experiment is an experiment, trial, or observation that can be repeated numerous times under the same conditions. The outcome of an individual random experiment must be independent and identically distributed. It must in no way be affected by any previous outcome and cannot be predicted with certainty.

Examples-

- Tossing a coin.
- Rolling a dice.
- The selection of a numbered ball (1-50) in an urn

*what are disjoint events?

Two events are mutually exclusive if they cannot occur at the same time. Another word that means mutually exclusive is disjoint.

If two events are disjoint, then the probability of them both occurring at the same time is 0.

Disjoint: P (A and B) = 0

*what are joint events?

Joint probability is a statistical measure that calculates the likelihood of two events occurring together and at the same point in time. Joint probability is the probability of event Y occurring at the same time that event X occurs.

Joint : P (A and B) = 1

*What is Sample Space?

The set of all possible outcomes of an experiment is known as the sample space of the experiment and it is denoted by S.
The sample space of rolling a dice experiment is, S = { 1, 2, 3, 4, 5, 6 }

*What is Outcome?

An outcome is the result of an experiment. In other words, an outcome is a particular result of an experiment.
In a tossing coin experiment “Head” is an outcome of the experiment.

*what is an event?

Any subset of the sample space S is defined as an event.
An event is the set of outcomes of an experiment to which a probability is assigned.
In a rolling dice example, which has sample space, S = { 1, 2, 3, 4, 5, 6 }; occurrence of even numbers, E1 = { 2, 4, 6 } and occurrence of odd numbers E2 = { 1, 3, 5 } can be two possible events.

B (Basic Concepts of Probability)

*what is probability?

The probability of an event measures the likelihood of the occurrence of that event. Its scale is from 0 to 1. 0 means the incident will not going to happen and number 1 means that its obviously going to happen. And decimal means it can or cannot happen. Like .5 means 50-50 percent chance it will happen.

$$Probability\ of\ an\ event = \frac{Number\ of\ favorable\ outcomes}{Total\ number\ of\ possible\ outcomes}$$

- The probability of event A is denoted by P(A)
- Probability is quantified as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty.
- The higher the probability of an event, the more likely it is that the event will occur.
- The sum of probabilities of all sample points in a sample space is equal to 1.
- The probability of event A is the sum of the probabilities of all the sample points in event A.

Let's see an example

Example 2: Suppose a coin is flipped 3 times. What is the probability of getting two tails and one head?

Solution: For this experiment, the sample space consists of 8 sample points.

$$S = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$$

Each sample point is equally likely to occur, so the probability of getting any particular sample point is 1/8. The event "getting two tails and one head" consists of the following subset of the sample space.

$$A = \{TTH, THT, HTT\}$$

The probability of Event A is the sum of the probabilities of the sample points in A. Therefore,

$$P(A) = 1/8 + 1/8 + 1/8 = 3/8$$

Axioms of Probability:

Axioms of Probability:

Consider an experiment whose sample space is S. For each event E of the sample space, S, we assume that a number P(E) is defined and satisfies the following three conditions:

(i) (Axiom of positivizes): $0 \leq P(E) \leq 1$.

(ii) (Axiom of certainty): $P(S) = 1$.

(iii) (Axiom of additivity): For any sequence of events $E_1, E_2, \ldots$ that are mutually exclusive, that is, events for which $E_n E_m = \emptyset$ when $n \neq m$, then,

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n)$$

Note: Well many won't understand what is axiom of additivity is and I wont also go into details here about that because this topic is really complex and I could not find any good videos or notes about it. It is the heavy fundament of artificial intelligence and also MIT has an open course on this which includes 20 minutes of lecture which included some complex examples and topics to cover this particular thing which I also didn't understand. So, anyone who wants to go deep into it can search on his own

Types of events

1. Independent (independent event means each event is not affected by other events. Example: Tossing a coin, throwing a die etc.)

2. Dependent (dependent events indicate that they can be influenced by the previous events. Example: After taking a card from a deck of card the probability changes since there are less cards available then.)

3. Mutually Exclusive (mutually exclusive events mean both the events cannot occur at the same time. For example, in a coin tossing experiment both head and tail cannot occur at the same time so the occurrences of head or tail is are mutually exclusive events, turning left or right at the same time are also mutually exclusive events.)

4. Disjoint events (disjoint events do not have any common elements)

5.Joint events (joint events have common elements. For example, hearts and kings are joint events.)

C (Probability Laws)

Additional Law of probability

1.For disjoint events A and B, the probability that, either event A or event B will occur is,

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

2. For disjoint events A, B, C, …, Z The probability that, either event A or event B or event C or … or event Z will occur is,

$$P(A \cup B \cup C \cup \cdots \cup Z) = P(A) + P(B) + P(C) + \ldots + P(Z)$$

3. For joint events A and B The probability that, either event A or event B or both will occur is,

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

4. For joint events A, B, and C The probability that, either event A or event B or event C or any two of them or all will occur is,

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

5.For joint events A and B The probability that, either event A or event B or both will occur is,

$$P(+P(A \cap B \cap C)$$

6.For joint events A and B The probability that, either event A or event B or both will occur is,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\ 8$$

7. For joint events A, B, and C The probability that, either event A or event B or event C or any two of them or all will occur is,

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

Multiplication laws of probability

1. For two independent events A and B The probability that, both event A and event B will occur simultaneously is,

$$P(A \cap B) = P(A)\,P(B)$$

2.For two dependent events A and B The probability that, both event A and event B will occur simultaneously is,

$$P(A \cap B) = P(A|B)\,P(B).$$

Here, occurrence of event A depends on occurrence of event B.

Independence:

Two events are known as independent events if the occurrence of one event does not affect the probability of occurring another event.

For two independent events A and B, the probability that A and B will both occur is found by multiplying the two probabilities.

$$P(A \text{ and } B) = P(A)\,P(B)$$

Similarly, for three independent events, A, B, and C, the special rule of multiplication used to determine the probability that all three events will occur is:

$$P(A \text{ and } B \text{ and } C) = P(A)\,P(B)\,P(C)$$

Conditional Probability:

If the probability of a particular event occurring, given that another event has occurred, then it is known as conditional probability. In other words, **conditional probability** is the probability of one event occurring with some relationship to one or more other events.
The formula for conditional probability is:

$$P(B|A) = P(A \text{ and } B) / P(A)$$

which you can also rewrite as:

$$P(B|A) = P(A \cap B) / P(A)$$

Let's check an example

Example 3:

Your neighbor has 2 children. You learn that he has a son, Joe. What is the probability that Joe's sibling is a brother?

Solution: Let us consider the experiment of selecting a random family having two children and recording whether they are boys or girls. Then, the sample space is S = {BB, BG, GB, GG}, where, e.g., outcome "BG" means that the first-born child is a boy and the second-born is a girl. Assuming boys and girls are equally likely to be born, the 4 elements of S are equally likely. The event, E, that the neighbor has a son is the set E = {BB, BG, GB}. The event, F, that the neighbor has two boys (i.e., Joe has a brother) is the set F = {BB}. We want to compute,

$$P(F|E) = \frac{P(F \cap E)}{P(E)}$$

$$= \frac{P(\{BB\})}{P(\{BB,BG,GB\})}$$

$$= \frac{1/4}{3/4}$$

$$= \frac{1}{3}$$

Probability using contingency table

Contingency table is a power tool in data analysis for comparing categorical variables. Although it is designed for analyzing categorical variables, this approach can also be applied to other discrete variables and even continuous variables

A general 2×2 contingency table will be like the follows:

X \ Y	Y_1	Y_2	Total
X_1	a	b	a + b
X_2	c	d	c + d
Total	a + c	b + d	a + b + c + d

Here the two variables are X and Y and each of them have two possible categories.

Let's see an example

Example 5:

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

Analyzing this we get the following question answers

a. Find P (Person is a car phone user).

Ans. P (person is a car phone user) $= \frac{\text{number of car phone users}}{\text{Total number of users in the study}} = \frac{305}{755}$

b. Find P (person had no violation in the last year)

Ans. P (person had no violation in the last year) $= \frac{\text{number of car phone users that had no violation}}{\text{Total number of users in the study}} = \frac{685}{755}$

c. Find P (Person is a car phone user | person had a violation in the last year)

Ans. P (Person is a car phone user | person had a violation in the last year)

$$= \frac{\text{number of car phone users that had violation in the last year}}{\text{Total number of users in the study that had violation in the last year}} = \frac{25}{70}$$

Total Probability Law

https://www.youtube.com/watch?v=U3_783xznQI&t=43s

Its an easy thing to catch but to express it clearly, it would take a lot of time, at least for me that's why I am putting the link that I think is the best to understand this topic