Raw data: are collected data that have not been organized numerically.

Arrays: is an arrangement of raw numerical data *in ascending or descending* order of magnitude.

The range: is the difference between the largest and smallest numbers of the raw data.

Number of classes or categories: There are different methods to determine the number of categories (Some of the methods will be studied later). Here, the number should not be less than **FIVE** and no more than **20**, depending on the nature, number of data and its repetition.

Ex: The table below represents vehicle speed data for a section of highway measured to the nearest 1km/hr. Find the frequency distribution table.

37	61	76	40	54	74	32	48	47	53
40	63	63	68	57	55	59	54	52	56
87	74	51	54	57	59	46	41	44	58
65	67	64	60	82	51	50	54	51	55
67	57	59	84	66	50	50	56	46	32
47	45	61	40	63	60	53	54	62	51
70	45	73	76	67	43	50	61	71	55
57	53	65	61	55	41	77	56	64	52
36	50	59	62	42	72	73	68	48	69
46	55	60	70	70	58	65	53	71	78

Ans.

First step is to arrange the raw data in ascending or descending order, as below

32	36	37	40	40	40	41	41	42	43
44	45	45	46	46	46	47	47	48	48
50	50	50	50	50	51	51	51	51	52
52	53	53	53	53	54	54	54	54	54
55	55	55	55	55	56	56	56	57	57
57	57	58	58	59	59	59	59	59	60
60	60	61	61	61	61	62	62	63	63
63	64	64	65	65	65	66	67	67	67
68	68	69	70	70	70	71	71	72	73
73	74	74	76	76	77	78	82	84	87

Finding the range, Range=largest value-smallest value=87-32=55

Choosing the number of categories (classes), let’s take 12 classes.

Finding the class interval, which is the difference between two successive lower class limits or two successive upper class limits.

Class interval= Range/number of classes or categories=55/12=4.58=5

*Always rounded to the highest integer number.

Finding midpoints of the class=(lower limit+upper limit)/2

Finding class boundaries, or true class limits,

lower value-0.5

upper value-0.5

Finding the frequency for each class. The total number of frequencies should be equal to the number of total readings (observations).

$$\sum_{i=1}^{k} \mathrm{fi} = \mathrm{n}$$

Finding relative frequency (probability of frequency) to each class.

$$Pi = \frac{fi}{n}$$

Where $\sum_{i=1}^{k} \mathrm{Pi} = 1$

The frequency table will be:

Class intervals	Class boundaries (true limits)	Class midpoints X_i	Frequency (f_i)	Relative frequency (P_i)
31-35	30.5-35.5	33	1	0.01
36-40	35.5-40.5	38	5	0.05
41-45	40.5-45.5	43	7	0.07
46-50	45.5-50.5	48	12	0.12
51-55	50.5-55.5	53	20	0.20
56-60	55.5-60.6	58	17	0.17
61-65	60.5-65.5	63	14	0.14
66-70	65.5-70.5	68	10	0.10
71-75	70.5-75.5	73	7	0.07
76-80	75.5-80.5	78	4	0.04
81-85	80.5-85.5	83	2	0.02
86-90	85.5-90.5	88	1	0.01
			∑=100	∑=1

Representation of statistical distribution

There are two graphic representation of frequency distribution:

Histograms,

They consist of:

Horizontal axis (X-axis) with centers representing class midpoints and lengths equal to class intervals

Areas proportional to class frequencies.

As in the figure below:

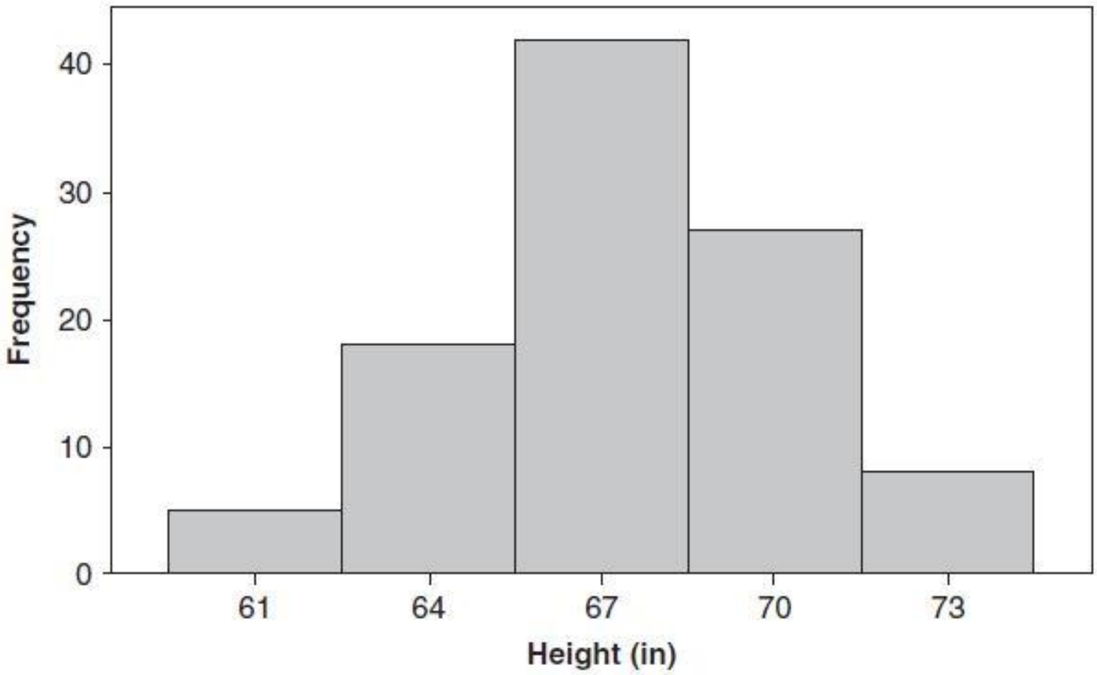


Figure: A histogram of students' heights in inches.

Polygon,

It is a line graph of the class frequencies plotted against class midpoints. It can be obtained by connecting the midpoints of the tops of the rectangles in the histogram.

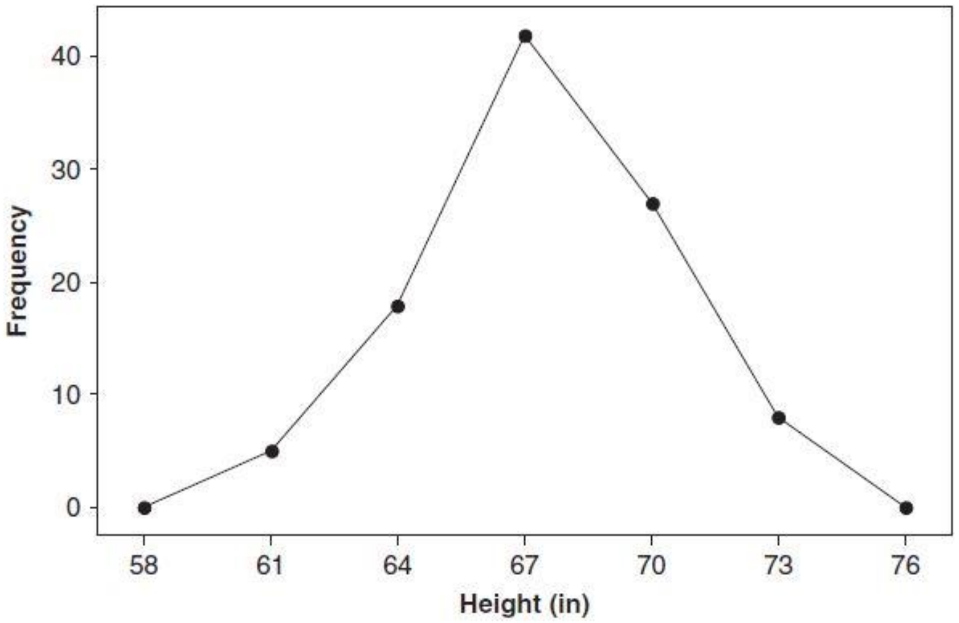
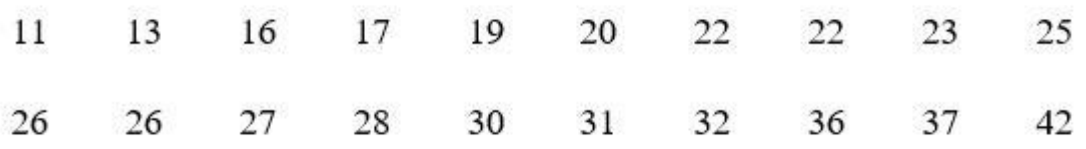


Figure: A polygon of students’ heights.

Cumulative frequency and probability curves

An explanation of frequency and probability distribution with X-axis representing the true points of a class, and Y-axis representing the cumulative values of frequency/probability of classes.

Ex:/ Precipitation in a desert was recorded in a year time, and the readings were rounded to the nearest 1mm, as follows:



11	13	16	17	19	20	22	22	23	25
26	26	27	28	30	31	32	36	37	42

Create the frequency table starting with the class (10-14), then:

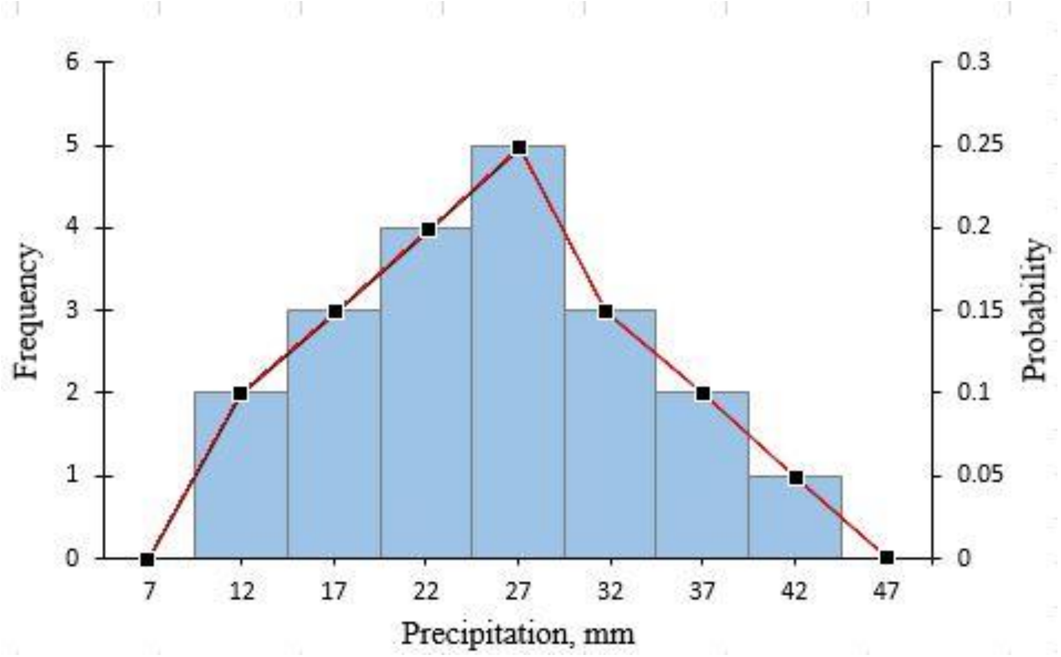
1- Histogram and polygon of frequency, and the ascending cumulative probability distribution curve.
2- The probability of precipitation less than 24.5mm.
3- The probability of precipitation between (24.5 - 34.5) mm.
4- The probability of 34mm precipitation.

Sol./ We start with frequency distribution curve, as below:

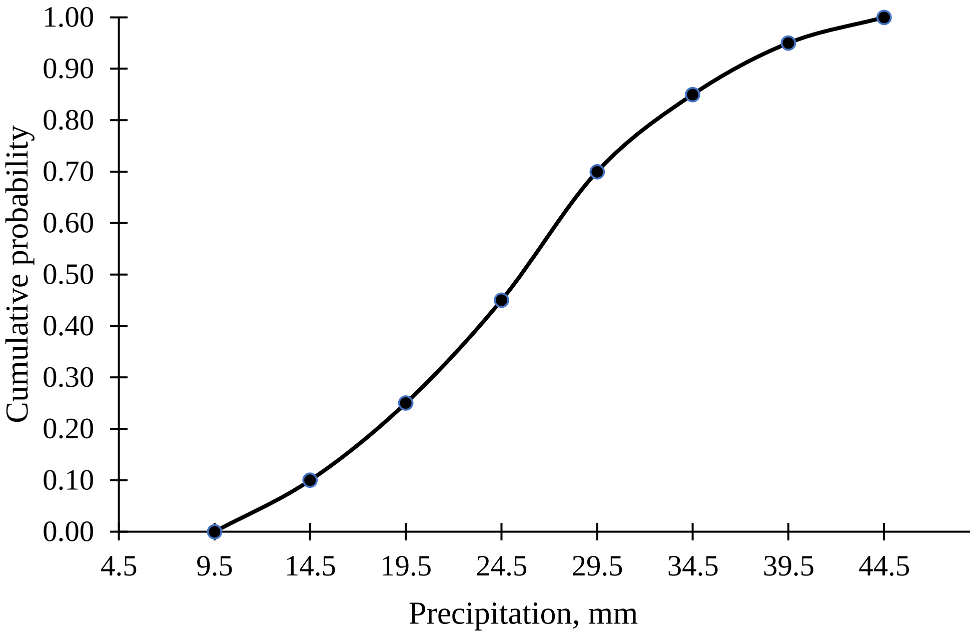
Class intervals	Class true limits	Class midpoints	Frequency f_i	Probability P_i	Category less than	Cumulative frequency (F_x)	Cumulative probability (P_x)
					< 9.5	0	0.00
10-14	9.5-14.5	12	2	0.10	< 14.5	2	0.10
15-19	14.5-19.5	17	3	0.15	< 19.5	5	0.25
20-24	19.5-24.5	22	4	0.20	< 24.5	9	0.45
25-29	24.5-29.5	27	5	0.25	< 29.5	14	0.70
30-34	29.5-34.5	32	3	0.15	< 34.5	17	0.85

35-39	34.5-39.5	37	2	0.10	< 39.5	19	0.95
40-44	39.5-44.5	42	1	0.05	< 44.5	20	1.00
∑			20	1.00			

1- Histogram and polygon of the frequency and the probability distribution.



The ascending probability distribution curve is as follows:



2- The probability of precipitation less than 24.5mm, is as follows:
From the probability histogram:

$$p(x < 24.5) = \frac{from\ x = 24.5\ and\ the\ area\ to\ the\ left}{Total\ area}$$

$$= \frac{(0.1 + 0.15 + 0.2) * 5}{1 * 5} = 0.45$$

Or from the cumulative probability curve:

$p(x < 24.5) = 0.45$

3- The probability of precipitation between (24.5 - 34.5) mm:

From the histogram of probability:

$$p(24.5 < x < 34.5) = \frac{Areas\ between\ 34.5\ ,24.5}{Total\ area}$$

$$= \frac{(0.25 + 0.15) * 5}{1 * 5} = 0.4$$

Or from the cumulative probability curve:

$$p(24.5 < x < 34.5) = p(x < 34.5) - p(x < 24.5)$$

= 0.85 - 0.45 = 0.4

4- The probability of 34mm precipitation:

From the histogram of probability:

$$p(x = 34) = p(33.5 < x < 34.5) = \frac{Area\ between\ 34.5\ ,33.5}{Total\ area}$$

$$= \frac{(0.15) * 1}{1 * 5} \quad = 0.03$$

Or from the cumulative probability curve:

$$p(x = 34) = p(33.5 < x < 34.5) = p(x < 34.5) - p(x < 33.5)$$

= 0.85 – 0.82 = 0.03

Ex:/ The smallest of 150 measurements is 5.18 in, and the largest is 7.44 in. Determine a suitable set of:

(a) class intervals.

(b) class boundaries.

Sol/

a- The range=7.44-5.18 = 2.26

For a minimum number of classes

Number of classes =5

Interval =2.26/5 = 0.45

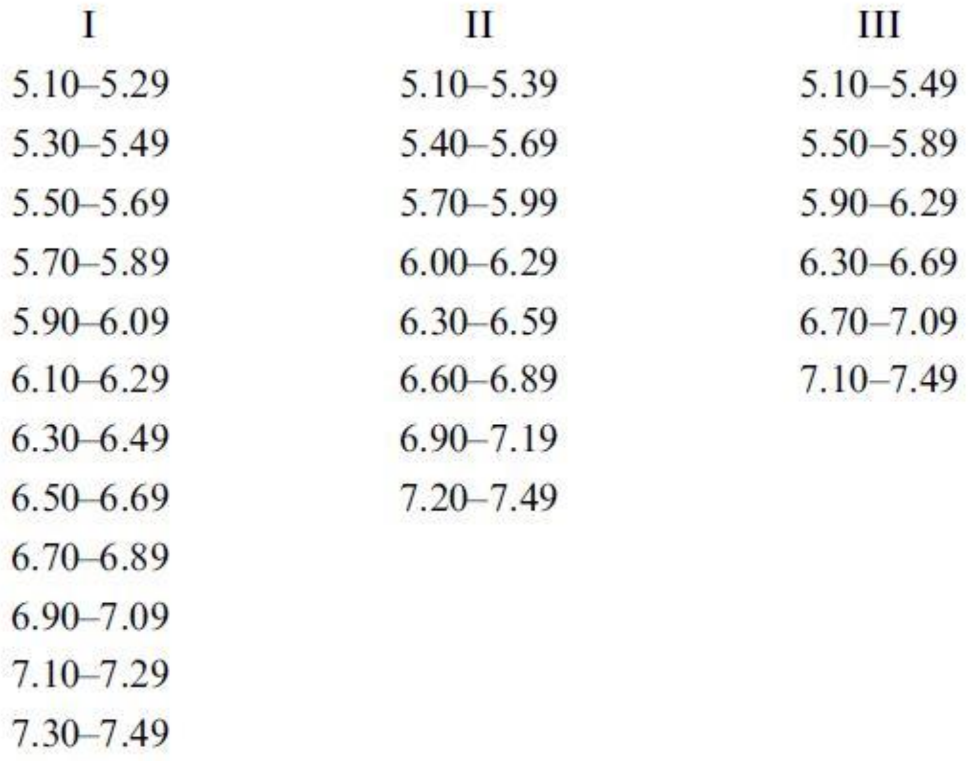
For a maximum number of classes

Number of classes = 20

Interval =2.26/20 = 0.11

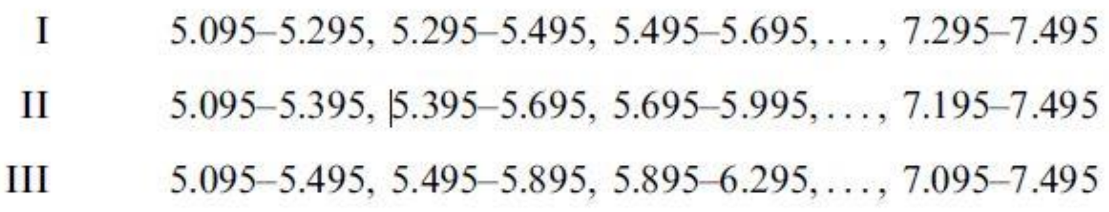
The convenient choice of class intervals is lying between 0.11 and 0.45, which is 0.20, 0.30 or 0.40

The col umns I, II and III below show suitable classes with intervals 0.2, 0.3 and 0.40, respectively.



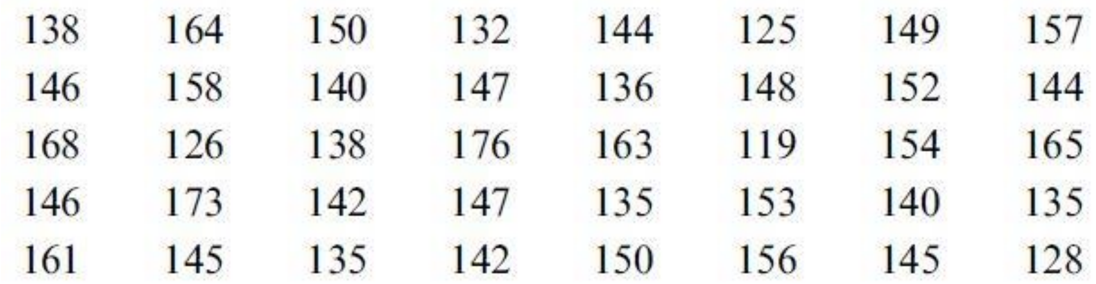
I	II	III
5.10–5.29	5.10–5.39	5.10–5.49
5.30–5.49	5.40–5.69	5.50–5.89
5.50–5.69	5.70–5.99	5.90–6.29
5.70–5.89	6.00–6.29	6.30–6.69
5.90–6.09	6.30–6.59	6.70–7.09
6.10–6.29	6.60–6.89	7.10–7.49
6.30–6.49	6.90–7.19	
6.50–6.69	7.20–7.49	
6.70–6.89		
6.90–7.09		
7.10–7.29		
7.30–7.49		

b- The class boundaries (true limits) corresponding to columns I, II, and III of part (a) are given, respectively, by:



I	5.095–5.295, 5.295–5.495, 5.495–5.695, . . . , 7.295–7.495
II	5.095–5.395, 5.395–5.695, 5.695–5.995, . . . , 7.195–7.495
III	5.095–5.495, 5.495–5.895, 5.895–6.295, . . . , 7.095–7.495

H.W/ In the following table the weights of 40 male students at a University are recorded to the nearest pound. Construct a frequency distribution.



138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	176	163	119	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

Then, find histogram and polygon of frequency, and the ascending cumulative probability distribution curve.

Measures of Central Location

These measures are typical, or representative, of a set of data. Since such typical values tend to lie centrally within a set of data arranged according to magnitude, central location measures are also considered ***averages***.

All these measures could deal with data in two types; Classified (with frequency), and unclassified (without frequency).

1- The Mean

It is the most common measure of central tendency, which has several types:

- The Arithmetic Mean

It is denoted by ($\overline{\text{X}}$) and calculated by:

$$\bar{X} = \frac{\sum_{i=1}^{n} xi}{n} \qquad For\ unclassified\ data\ (No\ frequency)$$

Where: x_i, the readings (data)

n, the total number of readings.

Or for classified data (having frequencies), as follows:

$$\overline{X} = \frac{\sum_{i=1}^{k} xi\ fi}{\sum_{i=1}^{k} fi}$$

Where:

$x_{i:}$ the midpoint of each class

f_i: the frequency of each class

k: number of classes.

For example, the arithmetic mean of the numbers 8, 3, 5, 12, and 10 is:

$\overline{X} = \frac{8+3+5+12+10}{5} = 7.5$

Ex. Data below represent the compressive strength of concrete samples tested in MPa: 40, 38, 45, 35, 52

Calculate the arithmetic mean of these values.

Ans.

$\overline{X} = \frac{40+38+45+35+52}{5} = 42$

Ex. Calculate the arithmetic mean of the vehicle speeds (km/h) from the table below:

Class	31 – 40	41 – 50	51 – 60	61 – 70	71 – 80	81 – 90	91 – 100
Frequency (f_i)	1	2	5	15	25	20	12

Ans.

Class	Frequency (fi)	Centre of class or midpoints (xi)	xi * fi

31 – 40	1	35.5	35.5
41 – 50	2	45.5	91
51 – 60	5	55.5	277.5
61 – 70	15	65.5	982.5
71 – 80	25	75.5	1887.5
81 – 90	20	85.5	1710
91 – 100	12	95.5	1146
	$\sum fi = 80$		$\sum xifi = 6130$

$$\bar{\mathrm{X}} = \frac{\sum_{\mathrm{i}=1}^{k} xi\ fi}{\sum_{i=1}^{k} fi} = \frac{6130}{80} = 76.63\ km/h$$

- The Geometric Mean

Is calculated from the following equations:

$$\bar{X}_g = (x_1 * x_2 * x_3 * \ldots\ldots\ldots x_n)^{1/n}$$

Or

$$\log \bar{X}_g = \frac{1}{n}\sum_{i=1}^{n} \log xi$$ For unclassified data

Where: $\bar{X}_g$ is the geometric mean; xi, the readings; n, total readings.

*For classified data, it is calculated as follows:

$$\bar{X}_g = ({x_1}^{f1} * {x_2}^{f2} * {x_3}^{f3} \ldots\ldots\ldots * {x_k}^{fk})^{1/n}$$

Or

$$\log \bar{X}_g = \frac{1}{n}\sum_{i=1}^{k} fi\ \ \log xi$$

Where: $\bar{X}_g$ is the geometric mean; $x_1, x_2 \ldots\ldots x_k$ are centres of classes; fi is the frequency; k is the number of classes, n is the number of readings.

Ex. For the readings: 3, 5, 8, 3, 7, 2. Find the geometric mean.

Ans.

$$\bar{\mathrm{X}}_g = (x_1 * x_2 * x_3 * \ldots\ldots\ldots x_n)^{1/n} \ = \sqrt[6]{3 * 5 * 8 * 3 * 7 * 2} = \sqrt[6]{5040} = 4.14$$

Or:

$\log \bar{\mathrm{X}}_g = \frac{1}{n} \sum_{i=1}^{n} \log xi$ = . the geometric mean.

ns follows:

ions:

h of concrete samples tested in MPa:

$$= \frac{1}{6} (\log 3 + \log 5 + \log 8 + \log 3 + \log 7 + \log 2)$$

= 0.617

$$\therefore \bar{\mathrm{X}}_g = 10^{0.617} = 4.14$$

Ex. Find the geometric mean for the following table.

Class	Frequency (f_i)
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8

Ans.

Class	Frequency (fi)	c- Midpoints

		(xi)
60 – 62	5	61
63 – 65	18	64
66 – 68	42	67
69 – 71	27	70
72 – 74	8	73
	$\sum fi = 100$	

$$\bar{X}_g = (x_1^{f1} * x_2^{f2} * x_3^{f3} \ldots\ldots\ldots * x_k^{fk})^{1/n}$$

$$\bar{X}_g = (61^5 * 64^{18} * 67^{42} * 70^{27} * 73^8)^{1/100} = 67.38$$

- The Harmonic Mean

It is calculated from the following equations:

$$\bar{X}_h = \frac{n}{\sum_{i=1}^{n} 1/xi} \qquad For\ unclassified\ data$$

Where: $\bar{X}_h$:The harmonic mean

xi: The readings

n: The total number of readings

And for classified data, it is calculated as:

$$\bar{X}_h = \frac{\sum_{i=1}^{k} fi}{\sum_{i=1}^{k} fi/xi}$$

where: $\bar{X}_h$ is the harmonic mean; x_i, midpoints of classes;

f_i, the frequency; *k*, number of classes.

Ex. Find the harmonic mean for the following data: 3, 5, 6, 6, 7, 10, 12.

Ans.

$$\bar{X}_h = \frac{n}{\sum_{i=1}^{n} 1/xi}$$,

$$\bar{X}_h = \frac{7}{\frac{1}{3}+\frac{1}{5}+\frac{1}{6}+\frac{1}{6}+\frac{1}{7}+\frac{1}{10}+\frac{1}{12}} = 5.87$$

Ex. Find the harmonic mean for the following table.

Class	Frequency (f_i)
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8

Ans.

Class	Frequency (f_i)	Midpoints (x_i)	(f_i/x_i)
60 – 62	5	61	0.082
63 – 65	18	64	0.281
66 – 68	42	67	0.627
69 – 71	27	70	0.386
72 – 74	8	73	0.109
	$\sum fi = 100$		$\sum fi/xi = 1.485$

$$\bar{X}_h = \frac{\sum_{i=1}^{k} fi}{\sum_{i=1}^{k} fi/xi} = \frac{100}{1.485} = 67.34$$

2- The Median

The median of a set of numbers arranged in order of magnitude (i.e., in an array) is either the middle value (*n* is an odd number) or the arithmetic mean of the two middle values (*n* is an even number).

For example: The median for the set of numbers 3, 4, 4, 5, 6, 8, 8, 8,10 = 6.

Where median is the value in the order of $\frac{n+1}{2}$, *n* is the number of data

Or it is just the middle value in case of *n* is an odd number.

Ex. Find the median for the following data: 2, 9, 12, 3, 7, 8, 4, 5.

Ans.

Arrange in an array: 2, 3, 4, 5, 7, 8, 9, 12. *n= 8* is an even number.

The median is the arithmetic mean of the middle values = $\frac{5+7}{2} = 6$.

*In the case of classified data, the median is calculated from:

$$M = a + \frac{n/2 - n_1}{fm} * \Delta$$

Where:

M, the median;

a, the lower class boundary of the median class (i.e., the class containing the median).

n, number of items in the data (i.e., total frequency).

n_1, sum of frequencies of all classes lower than the median class.

f_m, frequency of the median class.

Δ, class interval.

Ex. Find the median for the table below:

Class	Frequency (f_i)

60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8

Ans.

Class	Frequency (f_i)	Cumulative ascending frequency (f_i)		
60 – 62	5	Less than 62.5	5	
63 – 65	18	Less than 65.5	23	
66 – 68	42	Less than 68.5	65	First cumulative frequency higher than middle value of n.
69 – 71	27	Less than 71.5	92	
72 – 74	8	Less than 74.5	100	
	$\sum \mathrm{fi} = 100$			

a= 65.5 , n = 100 , n_1= 23 , f_m=42 , Δ =3

$$M = a + \left[\frac{n/2 - n_1}{fm}\right] * \Delta$$

$$= 65.5 + \left[\frac{100/2 - 23}{42}\right] * 3 = 67.43$$

3- The Mode

The mode of a set of numbers is that value which occurs with the greatest frequency; that is, it is the most common value.

For example: The mode of the set 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12,18 = 9.

And the set 3, 5, 8, 10, 12, 15,16 has no mode!

The mode here 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9 is **4** and **7**, and is called *bimodal*

A distribution having only one mode is called *unimodal*.

*For the case of classified data, the mode is calculated as following:

$$Mo = a + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) * \Delta$$

Where: M_o, is the mode;

a, lower class boundary of the modal class (the class with higher frequency);

Δ_1, the difference between modal frequency and the frequency of the preceding class;

Δ_2, the difference between modal frequency and the frequency of next class;

Δ, class interval.

Ex. Find the mode for the following table

Class	Frequency (f_i)
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8

The modal class is the one with higher frequency, which is 66-68.

$$Mo = a + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) * \Delta$$

(a=65.5 , $\Delta_{\mathbf{1}}$ =(42-18)=24 , $\Delta_{\mathbf{2}}$ = (42-27)=15 , Δ=3)

$$\therefore Mo = 65.5 + \left(\frac{24}{24+15}\right) * 3$$

$$= 67.35$$

Measures of Variation (Dispersion) from the average

The variation or dispersion of the data is the degree to which numerical data tend to spread about the average.

Measuring the average does not give a complete description of the data, for instance, it is important to know the dispersion of these data from the middle value.

For example: the average of this set 1,2,3,4,5,6,7,8,9,10 is = **5**

and the average of this set 0, 10 is **5**

The average here is more indicative to the first set that the second.

Therefore, various measures of this dispersion (or variation) are available, the most common being the *range*, *mean deviation*, and *standard deviation*.

The Range

The range of a set of numbers is the difference between the largest and smallest numbers in the set.

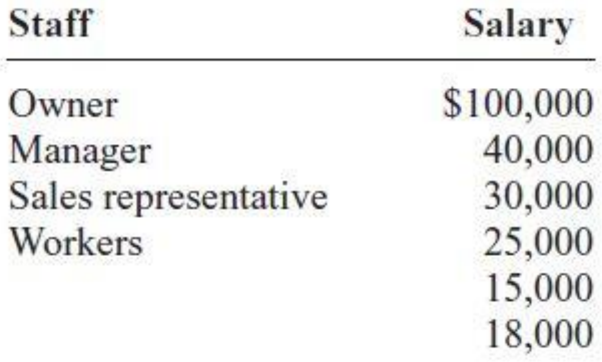
R= highest value – lowest value

For example: The range of the set 2, 3, 3, 5, 5, 5, 8, 10, 12 is 12 - 2 = 10.

Sometimes the range is given by simply quoting the smallest and largest numbers; in the above set, for instance, the range could be indicated as 2 to 12, or 2–12.

*Extremely high and extremely low values have an influence on the range value, as in the example below:

Ex. The salaries for the staff of the XYZ Manufacturing Co. are shown here. Find the range.



Staff	Salary
Owner	$100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
	15,000
	18,000

Ans. The range is R = $100,000 - $15,000 = $85,000.

Since the owner’s salary is included in the data of the example, the range is a large number.

*To have a more meaningful statistic to measure the variability, it is important to measure the dispersion by other methods.

The Mean Deviation

The *mean deviation*, or *average deviation*, is used to describe the dispersion around the middle value, and it is for a set of *N* numbers $X_1, X_2, \ldots, X_N$ is abbreviated MD and is defined by:

$$\mathrm{M.D} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad \text{For unclassified data,}$$

Where:

M.D: Mean deviation; x_i: readings; $\bar{x}$: the arithmetic mean; n: total number of data.

Ex. Find the mean deviation for the data: 9, 8, 6, 5, 7

Ans.

$$\mathrm{M.D} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

$$\bar{X} = \frac{9+8+6+5+7}{5} = 7$$

Xi	$Xi-\bar{X}$	$\lvert Xi-\bar{X}\rvert$
9	2	2
8	1	1
6	-1	1
5	-2	2
7	0	0
$\sum x_i = 35$		$\sum \lvert Xi-\bar{X}\rvert = 6$

$$\mathrm{M.D} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad = \frac{6}{5} = 1.2$$

For classified data: $$\mathrm{M.D} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}|\, fi}{\sum_{i=1}^{k} fi}$$

where:
x_i: the midpoints of classes; f_i: the frequency for each class; k: the number of classes.

Ex. Find the mean deviation for the table below:

Class	Frequency f_i
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8
	$\sum fi = 100$

Ans.

$$\mathrm{M.D} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}|\, fi}{\sum_{i=1}^{k} fi} \quad , \quad \bar{\mathrm{X}} = \frac{\sum_{i=1}^{k} xi\ fi}{\sum_{i=1}^{k} fi} = \frac{6745}{100} = 67.45$$

Class	Frequency f^i	Midpoint xi	fi* xi	$\lvert xi - \bar{x} \rvert$	$\lvert xi - \bar{x} \rvert$fi
60 – 62	5	61	305	6.45	32.25
63 – 65	18	64	1152	3.45	62.1
66 – 68	42	67	2814	0.45	18.9
69 – 71	27	70	1890	2.55	68.85
72 – 74	8	73	584	5.55	44.4
	$\sum fi = 100$		6745		226.5

$$\text{M. D} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}|\, fi}{\sum_{i=1}^{k} fi} = \frac{226.5}{100} = 2.265$$

The Variance and the Standard Deviation

The *variance* is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2.

The *standard deviation* is the square root of the variance. The symbol for the population standard deviation is σ.

To eliminate the negative signs of the dispersion, each value is squared to calculate the variance from the average value, then a square root is calculated to put the standard deviation in the same units as the raw data.

The formulas for the variance and the standard deviation *for a sample* (denoted by *S*):

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

For unclassified data

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

Where:

S^2: the variance of a sample; S: the standard deviation of a sample; $\bar{x}$: the arithmetic mean; n: total number of readings.

For *classified* data:

$$S^2 = \frac{\sum_{i=1}^{k} (xi - \bar{x})^2 fi}{\sum_{i=1}^{k} fi}, \qquad S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{k} (xi - \bar{x})^2 fi}{\sum_{i=1}^{k} fi}}$$

Where: x_i: the midpoints of classes; f_i: the frequency for each class; k: the number of classes.

In addition, the symbol for *population* variance is denoted by σ^2, and the standard deviation is denoted by σ. As in the equations below:

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} \qquad \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Where: X=individual values; μ= population mean; N=population size.

Same rule of classified data could be applied to the above formulas.

Ex. Find the variance and the standard deviation for the set of data: 9,8,6,5,7

Ans.

$$\bar{X} = \frac{\sum_{i=1}^{n} xi}{n} = \frac{9+8+6+5+7}{5} = \frac{35}{5} = 7$$

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

$$= \frac{(9-7)^2 + (8-7)^2 + (6-7)^2 + (5-7)^2 + (7-7)^2}{5} = 2$$

$$\therefore S = \sqrt{S^2} = \sqrt{2} = 1.41$$

Ex. Find the variance and standard deviation for the table below:

Class	Frequency f_i
60 – 62	**5**
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8
	$\sum fi = 100$

Ans.

$$\bar{X} = \frac{\sum_{i=1}^{k} xi\, fi}{\sum_{i=1}^{k} fi} = \frac{6745}{100} = 67.45$$

Class	Frequency f_i	xi	fi* xi	$(xi-\bar{x})^2 fi$
60 – 62	5	61	305	208.01
63 – 65	18	64	1152	214.25
66 – 68	42	67	2814	8.51
69 – 71	27	70	1890	175.57
72 – 74	8	73	584	246.42
	$\sum fi = 100$		6745	852.76

$$S^2 = \frac{\sum_{i=1}^{k}(xi-\bar{x})^2 fi}{\sum_{i=1}^{k} fi} = \frac{852.76}{100} = 8.53$$

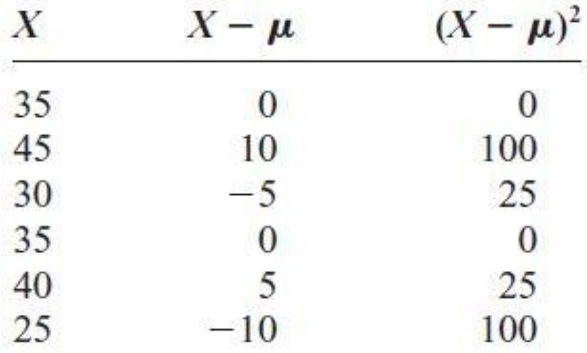
$$\therefore S = \sqrt{S^2} = \sqrt{8.53} = 2.92$$

Ex. The following data represent the variation of paint (in months):

35, 45, 30, 35, 40, 25. Find the variance and standard deviation.

Ans.

$$\mu = \frac{\Sigma X}{N} = \frac{35+45+30+35+40+25}{6} = \frac{210}{6} = 35$$



X	$X-\mu$	$(X-\mu)^2$
35	0	0
45	10	100
30	−5	25
35	0	0
40	5	25
25	−10	100

$$\Sigma(X-\mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} = \frac{250}{6} = 41.7$$

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}} = \sqrt{41.7} = 6.5$$

The mean deviation

$$\mathrm{M.D} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}|\, fi}{\sum_{i=1}^{k} fi}$$

Elementary Probability Theory

The classical definition of probability assumes that all outcomes in the sample space are equally likely to occur. It uses sample spaces to determine the numerical probability that an event will happen.

For example, when a single dice is rolled, each outcome has the same probability of occurring. *Since there are six outcomes* (*1, 2, 3, 4, 5, 6*), each outcome has a probability of $\frac{1}{6}$.

When a card is selected from an ordinary deck of 52 cards, you assume that the deck has been shuffled, and each card has the same probability of being selected, which is $\frac{1}{52}$.

*The probability of occurrence of the event (called its *success*) is denoted by p= $P(E)$ = $\frac{h}{n}$; where h is the ways the event E can happen, and n refers to the total possible ways.

*The probability of nonoccurrence of the event (called its failure) is denoted by

$$q = P(\overline{E}) = \frac{n-h}{n} = 1 - \frac{h}{n} = 1 - p = 1 - P(E)$$

Where p+q=1, which means P(E)+ P($\overline{E}$)= 1

The odds in favour of E is $\frac{P(E)}{P(\bar{E})} = \frac{p}{q}$, while the odds against E is $\frac{q}{p}$.

Probability rule: The probability of any event E is a number (either a fraction or decimal) between and including 0 and 1. This is denoted by $0 \leq P(E) \leq 1$.

In general, there are three basic interpretations of probability:

1- Classical probability: This approach is appropriate only for modeling chance experiments with equally likely outcomes.

2- Empirical or relative frequency probability: An estimate is based on an accumulation of experimental results. This estimate, usually derived empirically, presumes a replicable chance experiment.

3- Subjective probability: In this case, the probability represents an individual's judgment based on facts combined with personal evaluation of other information.

*All three types of probability (classical, empirical, and subjective) are used to solve a variety of problems in business, engineering, and other fields.

Ex. When a single dice is rolled, what is the probability of:

1- Getting a 3.
2- Not getting 6.
3- Getting a 9.
4- Getting a number less than 7.
5- The odds in favour of 6.
6- The odds against 6.

Ans./

1- $P(3) = \frac{1}{6}$

2- $P(\overline{6}) = q(6) = 1 - \frac{1}{6} = \frac{5}{6}$

3- $P(9) = \frac{0}{6} = 0$

4- $P(<7) = \frac{6}{6} = 1$

5- The odds in favour of 6 = $\frac{P(6)}{P(\overline{6})} = \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{5}$ (it is said 1:5)

6- The odds against 6 = $\frac{\frac{5}{6}}{\frac{1}{6}} = \frac{5}{1} = 5$ (Or 5:1)

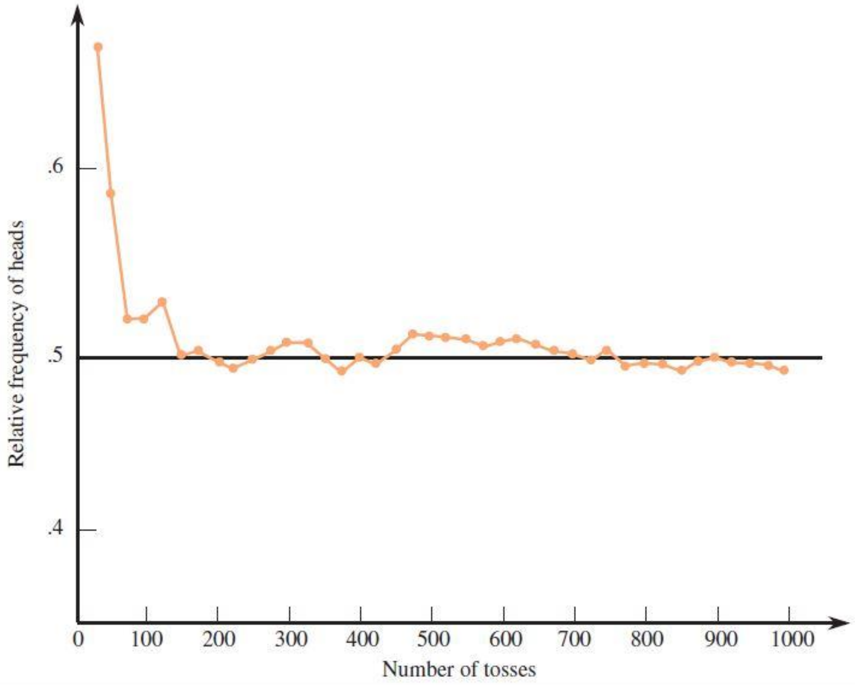
Statistically, the probability of an event is the relative frequency of the occurrence of that event divided by the total number of observations, as studied in the frequency distribution tables earlier.

Ex. If 1000 tosses of a coin result in 529 heads, the relative frequency of heads is 529/1000 = **0.529**.

If another 1000 tosses results in 493 heads, the relative frequency in the total of 2000 tosses is:

(529 + 493) /2000 = **0.511**.

According to the statistical definition, this should be **0.5** and to obtain that, further observations must be made, as shown in the figure below:



Ex. If a family has three children, find the probability that two of the three children are girls.

Ans.\

The possibility of the three children is: BBB, BBG, BGB, GBB, GGG, GGB, GBG, and BGG.

*Since there are three ways to have two girls, namely, GGB, GBG, and BGG

Then, P(2 girls) = $\frac{3}{8}$

Conditional probability: Dependent and Independent Events

If E_1 and E_2 are two events, the probability that E_2 occurs given that E_1 has occurred is denoted by $\boldsymbol{P(E_2/E_1)}$ **or** $\boldsymbol{P(E_2\ given\ E_1)}$.

This is called the *conditional probability* of E_2 given that E_1 has occurred.

*If the occurrence or nonoccurrence of E_1 does not affect the probability of occurrence of E_2, then $\boldsymbol{P(E_2/E_1) = P(E_2)}$ and we say that E_1 and E_2 are *independent events*; otherwise, they are *dependent* events.

If we denote by E_1E_2 to the event that ‘‘both E_1 and E_2 occur,’’ sometimes called a *compound event*, then:

$P(E_1E_2) = P(E_1)\ P(E_2/E_1)$

Or $P(E_1E_2) = P(E_1)\ P(E_2)$ for independent events

For three events E_1, E_2, and E_3, we have:

$P(E_1E_2E_3) = P(E_1)\ P(E_2/E_1)\ P(E_3/E_1E_2)$

Or $P(E_1E_2E_3) = P(E_1)\ P(E_2)\ P(E_3)$ for independent events

Ex. If the probability that A will be alive in 20 years is 0.7 and the probability that B will be alive in 20 years is 0.5. Find the probability that they both be alive in 20 years.

Ans.

Since they are independent evens, then

$P(E_1E_2) = P(E_1)\ P(E_2) = 0.7 * 0.5 = 0.35$

Ex. Suppose that a box contains 3 white balls and 2 black balls. What is the probability that both balls drawn are black?

Ans.

Let E_1 be the event first ball drawn is black and E_2 the event ‘‘second ball drawn is black. Here E_1 and E_2 are dependent events.

$P(E_1)= \frac{2}{3+2} = \frac{2}{5}$, $P(E_2)= \frac{1}{3+1} = \frac{1}{4}$

$P(E_1E_2) = P(E_1)\ P(E_2/E_1)$

$P(E_1E_2) = \frac{2}{5} * \frac{1}{4} = \frac{1}{10}$

Mutually Exclusive Events

Two or more events are called *mutually exclusive* if the occurrence of any one of them excludes the occurrence of the others. Thus, if E_1 and E_2 are mutually exclusive events, then:

$P(E_1 + E_2) = P(E_1) + P(E_2)$ - $P(E_1E_2)$, here $P(E_1E_2)= 0$

❖ $P(E_1 + E_2) = P(E_1) + P(E_2)$ For mutually exclusive events.

Where $P(E_1 + E_2)$ denotes that either E_1 or E_2 (***or both***) occur.

As an extension of this, if $E_1, E_2, \ldots, E_n$ are *n* mutually exclusive events having respective probabilities of occurrence $p_1, p_2, \ldots, p_n$, then the probability of occurrence of either E_1 or E_2 or …. E_n is $p_1 + p_2 + \ldots\ldots p_n$

Ex. A ball is drawn at random from a box containing 6 red balls, 4 white balls, and 5 blue balls. Determine the probability that the ball drawn is (a) red, (b) white, (c) blue, (d) not red, and (e) red or white.

Ans.\ Let R, W, and B denote the events of drawing a red ball, white ball, and blue ball, respectively. Then:

a) $P(R)= \frac{6}{6+4+5} = \frac{6}{15} = \frac{2}{5}$

b) $P(W)= \frac{4}{15}$

c) $P(B)= \frac{5}{15} = \frac{1}{3}$

d) P(not R) or $P(\bar{R}) = 1- \frac{2}{5} = \frac{3}{5}$

e) **P(R+W) = P(R)+P(W)** $= \frac{2}{5} + \frac{4}{15} = \frac{2}{3}$

Ex. For the same example above, if three balls are drawn successively from the box. Find the probability that they are drawn in the order red, white, and blue if each ball is (a) replaced and (b) not replaced.

Ans.\

a) If each ball is replaced, then R, W, and B are independent events and

$$P(RWB) = P(R)\,P(W)\,P(B) = \left(\frac{2}{5}\right)\left(\frac{4}{15}\right)\left(\frac{1}{3}\right) = \frac{8}{225}$$

b) If each ball is not replaced, then R, W, and B are dependent events and

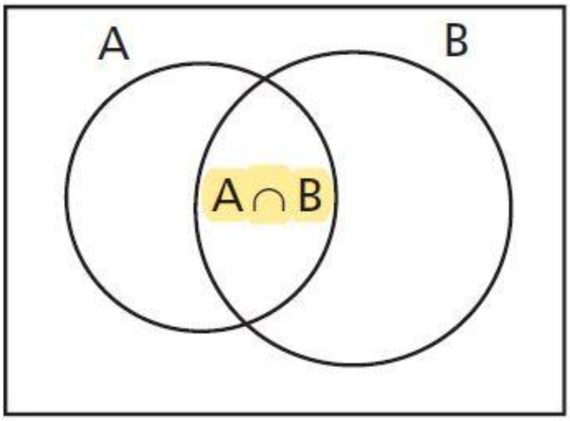
$$P(RWB) = P(R)\,P(W/R)\,P(B/WR) = \left(\frac{6}{6+4+5}\right)\left(\frac{4}{5+4+5}\right)\left(\frac{5}{5+3+5}\right) = \left(\frac{4}{91}\right)$$

H.W/ For the example above: If the three balls were all whites or all blues?

Venn Diagram

When events are not mutually exclusive, there can be ***overlap*** between them. This can be visualized using a *Venn Diagram*.

The probability of overlap must be subtracted from the sum of probabilities of the separate events (i.e., we must not count the same area on the Venn Diagram twice).



Venn Diagram

In the figure above, the intersection **A ∩ B** (A and B) represents the overlap between events A and B.

Sample Space: The set consisting of all possible outcomes of a particular experiment is called the sample space of that experiment. Thus, the rectangle on the Venn diagram corresponds to the sample space.

Some set of notations:

P(A or B) = **P**(occurrence of **A** or **B** or **Both**), the *union* of the two events A and B.

P(A and B) = **P**(occurrence of both **A** and **B**), the *intersection* of the two events A and B.

Ex: A fair six-sided dice is tossed twice. What is the probability that a five will occur at least once?

Ans:

Method 1:

Let say the event of getting 5 on the first toss is E_1, and the event of getting a 5 on the second toss is E_2.

$E_1 = \frac{1}{6}$, $E_2 = \frac{1}{6}$, and $\overline{E_1}$ and $\overline{E_2} = \frac{5}{6}$. Then:

P(E_1 and E_2) = $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$

P(E_1 and $\overline{E_2}$) = $\frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$

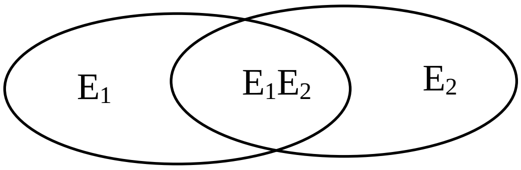
P($\overline{E_1}$ and E_2) = $\frac{5}{6} \cdot \frac{1}{6} = \frac{5}{36}$

P($\overline{E_1}$ and $\overline{E_2}$) = $\frac{5}{6} \cdot \frac{5}{6} = \frac{25}{36}$,

$\sum$**all = 1** For accuracy.

Then, P(*at least one* **5** in both tosses) = $\frac{1}{36} + \frac{5}{36} + \frac{5}{36} = \frac{11}{36}$

Method 2: From Venn diagram



P(at least one 5 in both tosses)⇨ P(E_1 or E_2) = P(E_1) + P(E_2) - P(E_1E_2)

$$\frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}$$

Ex: In a factory unit there are 8 engineers and 5 accountants; 7 engineers and 3 accountants are females. If a staff person is selected, find the probability that the subject is an engineer or a male. And the probability of obtaining an accountant or female?

Ans:

P(engineer or male) = P(engineer) + P(male) - P(male engineer)

$$= \frac{8}{8+5} + \frac{3}{8+5} - \frac{1}{8+5} = \frac{10}{13}$$

Ex: An oil company is bidding for the rights to drill a well in field A and a well in field B. The probability it will drill a well in field A is 40%. If it does, the probability the well will be successful is 45%. The probability it will drill a well in field B is 30%. If it does, the probability the well will be successful is 55%. Calculate each of the following probabilities:

a) probability of a successful well in field A,

b) probability of a successful well in field B,

c) probability of both a successful well in field A and a successful well in field B,

d) probability of at least one successful well in the two fields together,

e) probability of no successful well in field A,

f) probability of no successful well in field B,

g) probability of no successful well in the two fields together (calculate by two methods),

h) probability of exactly one successful well in the two fields together.

Ans:

a) P(a successful well in field A) = P(a well in A) × P(successful well in A)
= (0.40)(0.45)
= 0.18

b) P(a successful well in field B) = P(a well in B) × P(successful well in B)
= (0.30)(0.55)
= 0.165

c) P(both a successful well in field A and a successful well in field B)
= P(a successful well in field A) × P(a successful well in field B)
= (0.18)(0.165)
= 0.0297

d) P(at least one successful well in the two fields)
= P[(successful well in field A) **OR** (successful well in field B)]
= P(successful well in field A) + P(successful well in field B) – P(both successful)

= 0.18 + 0.165 – 0.0297
= 0.315 (remaining two ways of solutions?)

e) P(no successful well in field A)
= P(no well in field A) + P(unsuccessful well in field A)
= P(no well in field A) + P(well in field A)× P(**failed well in A**)
= **0.60** + (0.40)(**0.55**)
= 0.60 + 0.22
= 0.82

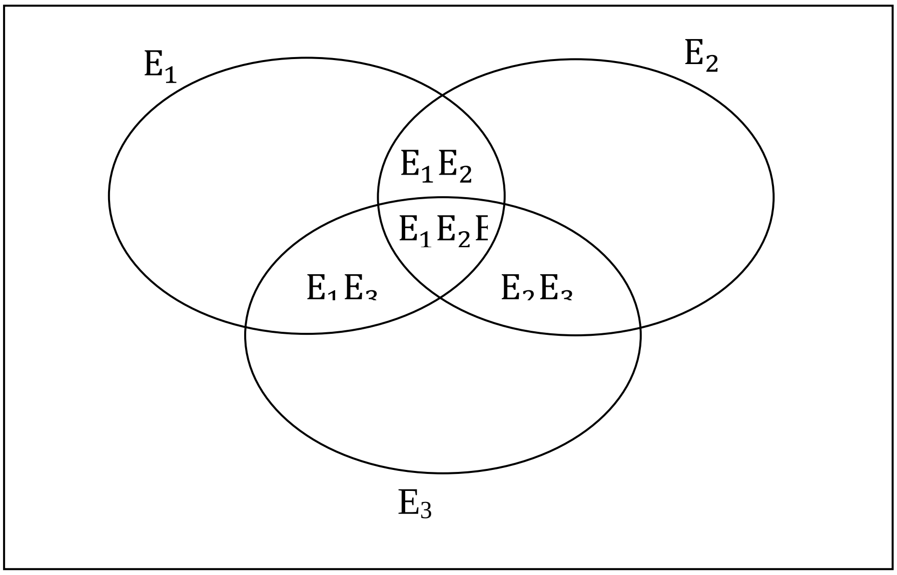
f) **P**(no successful well in the two fields) can be calculated in **two ways**.

Method 1:
P(no successful well in the two fields) = 1 – P(at least one successful well in the two fields)
= 1 – 0.3153
= 0.685

OR:
P(no successful well in the two fields)
= P(no successful well in field A) × P(no successful well in field B)
= (0.82)(0.835)
= 0.685

g) P(exactly one successful well in the two fields)
= P[(successful well in A) **and** (no successful well in B)] + P[(no successful well in A) **and** (successful well in B)]
= (0.18)(0.835) + (0.82)(0.165)
= 0.1503 + 0.1353
= 0.286

Venn diagram can combine three events as well, as shown in the figure below:



Then:

$$
\begin{aligned}
&p(E_1 + E_2 + E_3) \\
&\qquad = p(E_1) + p(E_2) + p(E_3) - p(E_1E_2) - p(E_2E_3) - p(E_1E_3) \\
&\qquad + p(E_1E_2E_3)
\end{aligned}
$$

Ex: Three junior civil engineers are required to design a highway section. The probability of accurate design for the first engineer is $\frac{4}{5}$, the probability for the second engineer is $\frac{3}{4}$, and the probability for the last one is $\frac{2}{3}$. Find the probability of obtaining an accurate design.

Ans:

$$
\begin{aligned}
&p(E_1 + E_2 + E_3) \\
&\qquad = p(E_1) + p(E_2) + p(E_3) - p(E_1E_2) - p(E_2E_3) - p(E_1E_3) \\
&\qquad + p(E_1E_2E_3) \\
&\qquad = \frac{4}{5} + \frac{3}{4} + \frac{2}{3} - \left(\frac{4}{5} * \frac{3}{4}\right) - \left(\frac{3}{4} * \frac{2}{3}\right) - \left(\frac{4}{5} * \frac{2}{3}\right) + \left(\frac{4}{5} * \frac{3}{4} * \frac{2}{3}\right) \\
&\qquad = \frac{59}{60} \simeq 0.98
\end{aligned}
$$

Another method:

Assuming all the designs are inaccurate:

$$p(\bar{E}_1\bar{E}_2\bar{E}_3) = \frac{1}{5} * \frac{1}{4} * \frac{1}{3} = \frac{1}{60}, \Longrightarrow \quad p(E_1 + E_2 + E_3) = 1 - \frac{1}{60} = \frac{59}{60}$$

Permutations and combinations

Permutations and combinations give us algebraic methods of arrangement of n objects in a specific order.

They are used in probability problems for two purposes:

To count the number of equally likely possible results for the classical approach to probability, and,
To count the number of different arrangements of the same items to give a multiplying factor.

Permutations

They refer to any of the various ways in which a set of things can be ordered

*Each separate arrangement of all or part of a set of items is called a ***permutation***.

The number of permutations is the number of different arrangements in which items can be placed.

Notice that if the order of the items is *changed*, the *arrangement is different*, so we have a *different permutation*.

Let’s say we have a total of $\boldsymbol{n}$ items to be arranged, the number of possible permutations of the items is:

nP_n= n!

and if we can choose $\boldsymbol{r}$ of those items at a time, where $r \leq n$. The number of permutations of n items chosen r at a time is written:

$$\boldsymbol{nPr} = \frac{\boldsymbol{n}!}{(\boldsymbol{n}-\boldsymbol{r})!} \quad ,$$

where $n!$ = (n)(n – 1)(n – 2)(n – 3)...(3)(2)(1),

(n – r)! = (n – r)(n – r – 1) ((n – r – 2)...(3)(2)(1).

If $r = n$, then: $\boldsymbol{nPr} = \frac{\boldsymbol{n}!}{\boldsymbol{0}!} = \boldsymbol{n}!$ As **0! = 1**.

For example, There are **120** permutations of the numbers 1, 2, 3, 4, 5. for example, 1, 3, 2, 4, 5 or 5, 1, 4, 2, 3.

Ex: If there are 50 students in a class, and only three students were allowed to ask questions. How many possible ways to ask the questions?

Ans.

$$nPr = \frac{n!}{(n-r)!}$$

$50P_3 = \frac{50!}{(50-3)!} = \frac{50!}{47!} = 50 * 49 * 48 = 117600$ possible way!

Ex: An engineer in technical sales must visit plants in Baghdad, Najaf, and Nasiriya. How many different sequences or orders of visiting these three plants are possible?

Ans.

The number of different sequences is equal to: nP_n= n!

$3P_3$ = 3! = 6 different permutations.

This can be verified by the following ***tree diagram***:

If some of them are indistinguishable from one another, the number of possible permutations is reduced. If ***n₁*** items are the same, and the remaining (***n–n₁***) items are the same of a different class, the number of permutations can be:

$$nP_n = \frac{n!}{n1!\,(n-n1)!}$$

Ex: A machinist produces 22 items during a shift. Three of the 22 items are defective, and the rest are not defective. In how many different orders can the 22 items be arranged if all the defective items are considered **identical** and all the nondefective items are identical of a different class?

Ans.

$$nP_n = \frac{n!}{n1!\,(n-n1)!}$$

$$22P_{22} = \frac{22!}{3!\,(22-3)!} = \frac{22!}{3!\,19!} = \frac{(22)(21)(20)}{(3)(2)(1)} = 1540$$

*If the case of similarities is repeated for more than one item, then:

$$nPn = \frac{n!}{n_1!\,n_2!\ldots\ldots\ldots n_k!}$$

Where: n_1 , n_2 , ….. n_k are number of group items have similarities.

and $n_1 + n_2 + n_3 + \ldots\ldots\ldots + n_k = n$

Ex: How many permutations can the letters AAABBCCDE have?

Ans.

There are 3A , 2B , 2C , 1D , 1E

Then:

$$9P_9 = \frac{9!}{3!\,2!\,2!\,1!\,1!} = \frac{60480}{4} = 15120$$

Ex: In a horse race, a field could have 20 horses. The first and the second horse in order are considered winners. If two horses were randomly selected for a bet, what is the probability that those horses could be winners?

Ans.

n= 20, *r*=2

$20P_2 = \frac{20!}{(20-2)!}$ = 380 different possible arrangements of two horses.

The probability of one of those arrangements are winning is:

P(**Any two winners**) = $\frac{1}{380}$.

Combinations

They are similar to permutations, but with the important difference that combinations take ***no account of order***.

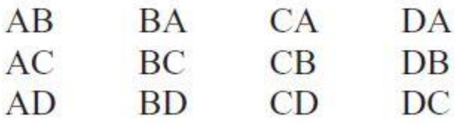
$$nCr = \frac{n!}{[(n-r)!\ r!}$$

Ex: Given the letters A, B, C, and D. If we select two letters, how many different groups are possible? Consider:

a) The letters are selected in sequence.
b) Letters are treated regardless of sequence.

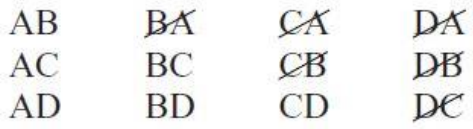
Ans.

a) Permutations, $4P_2 = \frac{4!}{2!}$ = 12, and the possible groups are as follows:



AB	BA	CA	DA
AC	BC	CB	DB
AD	BD	CD	DC

b) Combinations, $4C_2 = \frac{4!}{[(4-2)!\,2!]}$ = 6, and the possible groups are as follows:



AB	~~BA~~	~~CA~~	~~DA~~
AC	BC	~~CB~~	~~DB~~
AD	BD	CD	~~DC~~

In *permutations*, AB *is different* from BA. But in combinations, AB is the same as BA since the order of the objects *does not matter in combinations*. Therefore, duplicates are removed from the list.

Important notes:

When we use the terms **order**, **arrangement**, **sequence**. It means that permutation is taken into account.
When we talk about **selection, choice, to choose,** process, where the order or arrangement is not important. It means combinations are considered.

Ex: In a factory there are 7 engineers and 5 accountants. A committee of 3 engineers and 2 accountants is to be **chosen**. How many different possibilities are there?

Ans.

The selection of engineers, $7C_3 = \frac{7!}{[(7-3)!\,3!]}$ = 35

And the selection of accountants, $5C_2 = \frac{5!}{[(5-2)!\,2!]}$ = 10

Finally, by the fundamental counting rule,

The total number of different ways = 35 *10 = 350

H.W/ If there were 3 female engineers and 2 female accountants. What is the probability of choosing a committee of all females?

Probability Distributions

Probability distributions describe what will probably happen instead of what actually did happen, and they are often given in the format of a graph, table or formula.

In order to fully understand the probability distributions, we must first understand the concept of a random variable, and be able to distinguish between discrete and continuous random variables.

The random variable is typically represented by $\boldsymbol{X}$ and has a single numerical value ***determined by chance***, for each outcome of the procedure.

1- Binomial Probability Distributions

It is considered an important discrete probability distribution that is widely used in engineering applications.

This distribution applies in cases where there are only two possible outcomes: good or defective, success or failure, head or tail, rainy or not rainy, and many other possible pairs.

The probability of each outcome can be calculated using the multiplication rule, perhaps with a tree diagram, but *it is usually much faster and more convenient to use a general formula.*

The *requirements* for using the binomial distribution are as follows:

a- The outcome is determined completely by chance.
b- There are only two possible outcomes.
c- All trials have the same probability for a particular outcome in a single trial and they are independent of the outcome of a previous trials.
d- The number of trials, **n**, must be fixed, regardless of the outcome of each trial.

Let say ***p*** is the probability that an event will happen in any single trial (called success) and ***q*** = ***1 - p*** is the probability that the event will not happen in any single trial (called failure), and let say that ***x*** is a random variable representing the times the event will happen in ***n*** trials (i.e., ***X*** successes and

n - X failures). Then, the probability distribution of ***X*** values is called the Binomial probability distribution. As follows:

$p(x) = p^x(q)^{n-x} = p^x(1-p)^{n-x}$ (Events happening in sequence)

$p(x) = nCx \ p^x \ q^{n-x}$ (Sequence is not important)

Where, *p(x)* is the binomial probability formula,

$nCx = \frac{n!}{(n-x)!*x!}$ $(combination)$, where x is chosen from n

x: number of success

n: number of total trials

p: probability of success in any trial *x*

q: probability of failure in any trial *x*

And the rule is:

$$\sum_{x=0}^{n} p(x) = 1$$

The average: **μ=np = ∑x p(x)** which calculates the expected number of successes.

Variance: $\boldsymbol{\sigma^2} = \boldsymbol{npq}$ = $\sum x^2$ p(x) - μ^2

Standard deviation: $\boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma^2}}$

Ex: On the basis of past experience, the probability that a certain electrical component will be satisfactory is 0.98. The components are sampled item by item from continuous production. In a sample of five components, what are the probabilities of finding (a) zero, (b) exactly one, (c) exactly two, (d) two or more defectives?

Ans.

The requirements of the binomial distribution are met.

$n = 5$, $p = 0.98$, $q = 0.02$, where q is the probability that an item will be defective.

(a) p(0 defectives) = $(0.98)^5 = 0.904$.

(b) p(1 defective) = $5C_1\ (0.98)^4\ (0.02)^1 = (5)\ (0.98)^4(0.02)^1 = 0.092$.

(c) p(2 defectives) = $5C_2\ (0.98)^3(0.02)^2 = \frac{(5)(4)}{(2)}\ (0.98)^3(0.02)^2 = 0.0038$ or 0.004.

(d) p(2 or more defectives) = 1 – p(0 def.) – p(1 def.)

=1 – 0.9039 – 0.0922 = 0.0038 or 0.004.

Ex: A company is considering drilling four oil wells. The probability of success for each well is 0.40, independent of the results for any other well. The cost of each well is $200,000. Each well that is successful will be worth $600,000.

a) What is the probability that one or more wells will be successful?

b) What is the expected number of successes?

c) What is the expected gain?

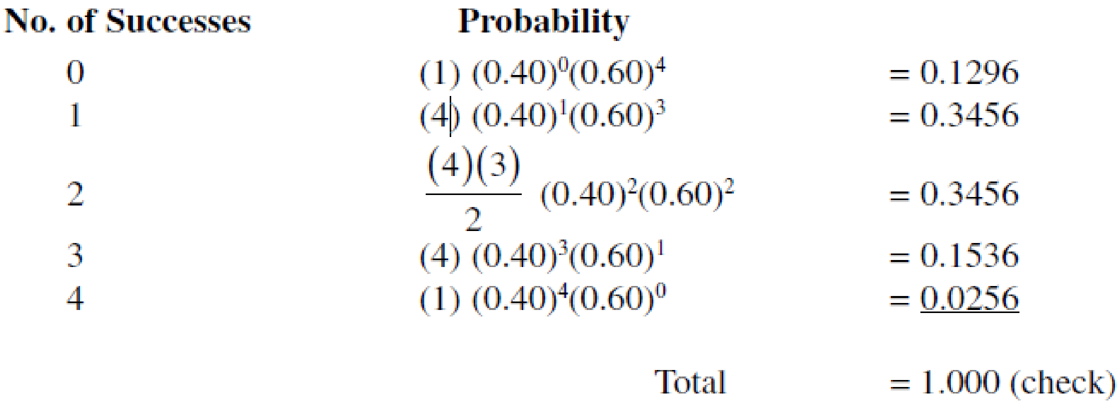
d) What will be the gain if only one well is successful?

e) Considering all possible results, what is the probability of a loss rather than a gain?

f) What is the standard deviation of the number of successes?

Ans.

The binomial distribution applies. Let us start by calculating the probability of each possible result. We use n = 4, p = 0.40, q = 0.60.



No. of Successes	**Probability**	
0	$(1)\ (0.40)^0(0.60)^4$	= 0.1296
1	$(4)\ (0.40)^1(0.60)^3$	= 0.3456
2	$\frac{(4)(3)}{2}\ (0.40)^2(0.60)^2$	= 0.3456
3	$(4)\ (0.40)^3(0.60)^1$	= 0.1536
4	$(1)\ (0.40)^4(0.60)^0$	= 0.0256
	Total	= 1.000 (check)

a) p(one or more successful wells) = 1– p(no successful wells)

= 1 – 0.1296 = 0.870.

b) Expected number of successes (μ) = (1)(0.3456) + (2)(0.3456) + (3)(0.1536) + (4)(0.0256)

= 1.600. **Or** μ=np = (4)(0.40) =1.60

c) Expected gain = (1.6)($600,000) – (4)($200,000) = $160,000.

d) If only one well is successful, gain = (1)($600,000) – (4)($200,000)

= –$200,000 (so it is a **loss**).

e) There will be a loss if 0 or 1 well is successful, so the probability of a loss is

(0.1296 + 0.3456) = 0.475

f) $\sigma^2 = \sum x^2\, p(x) - \mu^2$

where $\sum x^2\, p(x) = (1)^2\,(0.3456) + (2)^2\,(0.3456) + (3)^2\,(0.1536) + (4)^2\,(0.0256) =$ 3.5200

Then, $\sigma^2 = 3.5200 - (1.6)^2 = 0.96$.

Or $\sigma^2 = npq = (4)(0.40)(0.60) = 0.96$

The standard deviation of the number of successes is $\sqrt{0.96} = 0.98$

Ex: If the probability of rain in any day of September is 0.20, find the probability of:

a- The rain three consecutive days.
b- The rain in any three days of the month.
c- It will rain 27 days or more.
d- It will rain in two days or more.
e- No rain in this month.

Ans.

September is 30 days, then n=30, p = 0.2, q = 0.8

a- $p(X = x) = p^x (q)^{n-x}$

$p(X = 3) = 0.2^3 (0.8)^{27} = 1.934 * 10^{-5}$ which is a very little probability.

b- $p(X = x) = nCx\, p^x (q)^{n-x}$

$p(X = 3) = 30C3\,(0.2)^3 (0.8)^{27} \cong 0.08 = 8\%$

c- $p(X \geq 27) = p(X = 27) + p(X = 28) + p(X = 29) + p(X = 30)$

$$= 30C27\,(0.2)^{27}(0.8)^3 + 30C28\,(0.2)^{28}(0.8)^2 + 30C29\,(0.2)^{29}(0.8)^1 + 30C30\,(0.2)^{30}(0.8)^0 = 0.008 = 0.8\%$$

d- $p(X \geq 2) = 1 - p(X < 2) = 1 - [p(X = 1) + p(X = 0)]$

$$= 1 - [30C1\,(0.2)^1(0.8)^{29} + 30C0\,(0.2)^0(0.8)^{30}]$$

$$= 0.988 = 98.8\%$$

e- $p(X = 0) = 30C0\ (0.2)^0(0.8)^{30} \cong\ 0.001 = 0.1\%$

2- Poisson Distribution

The distribution is named for Simeon-Denis Poisson, a French mathematician of the nineteenth century.

The Poisson distribution applies in a possible number of discrete occurrences in a much larger sample space and less probability of success (rare events) over a ***specified interval***. The interval can be ***time***, ***distance***, ***area***, ***volume***, *or some similar unit*. It is used, when certain conditions are met, with these characteristics:

- The random X is the number of possible occurrences of an event over some interval.
- The outcomes must occur randomly, that is, completely by chance.
- The outcomes are independent, which means they are not affected by whether they happened previously.
- The outcomes must be uniformly distributed over the interval being used.

Examples of occurrences to which the ***Poisson distribution*** often applies include collisions of cars at a specific intersection under specific conditions, telephone calls to a particular telephone or office under particular conditions, the number of plants growing per acre or the number of defects in a given length of videotape, chances of rain over a particular period of time or place, etc.

The probability of $\boldsymbol{X}$ occurrences ($p(X = x)$) in an interval of time, volume, area, etc., for a variable where

$$p(X = x) = p(x) = \frac{\lambda^x *\ e^{-\lambda}}{x!}$$

Where:

λ (*Lamda*), is the mean number of occurrences per unit.

X, number of successes.

e, constant approximately equals to **2.71828**
Other important equations:

The mean $\mu = \lambda = np$ which is the expectation of success.
Variance $\sigma^2 = \lambda$
Standard deviation $\sigma = \sqrt{\lambda}$

Ex: If there are 200 typographical errors randomly distributed in a 500-page manuscript, find the probability that a given page contains exactly 3 errors.

Ans. First, find the mean number of errors. Since there are 200 errors distributed over 500 pages, each page has an average of:

$\lambda = \frac{200}{500} = 0.4$, which means 0.4 error per page.

Since X = 3, substituting into the formula yields:

$$p(X = 3) = \frac{\lambda^x * e^{-\lambda}}{x!} = \frac{0.4^3 * 2.71828^{-0.4}}{3!} = 0.0072 = 0.72\%$$

Thus, there is less than a 1% chance that any given page will contain exactly 3 errors.

Ex: If approximately 2% of the people in a room of 200 people are left-handed, find the probability that exactly 5 people there are left-handed.

Ans. Since λ = *np*, then λ = (200) (0.02) = 4, *x* = 5.

$$p(X = 5) = \frac{\lambda^x * e^{-\lambda}}{x!} = \frac{4^5 * 2.71828^{-4}}{5!} = 0.1563$$

Ex: Assuming there are 20% of measurement items are defective, if we choose a sample of 20 items. What is the probability of 3 defective items?

a- Use Binomial distribution
b- Use Poisson distribution

Ans.

a- n = 20 , p = 0.2 , q = 0.8 , X= 3

$$p(X = 3) = 20C3\ (0.2)^3(0.8)^{17} \cong 0.205 = 20.5\%$$

b- n = 20 , p = 0.2 , X = 3

$$\therefore \lambda = n * p = 20 * 0.2 = 4$$

$$p(X = x) = \frac{\lambda^x * e^{-\lambda}}{x!}$$

$$\therefore p(X = 3) = \frac{4^3 * e^{-4}}{3!} = 0.195 = 19.5\%$$

H.W: The average number of collisions occurring in a week during the summer months at a particular intersection is 2.00. Assume that the requirements of the Poisson distribution are satisfied.

a) What is the probability of no collisions in any particular week?

b) What is the probability that there will be exactly one collision in a week?

c) What is the probability of exactly two collisions in a week?

d) What is the probability of finding not more than two collisions in a week?

e) What is the probability of finding more than two collisions in a week?

f) What is the probability of exactly two collisions in a particular two-week interval?

Ans.

a) 0.135, **b)** 0.271, **c)** 0.271, **d)** 0.677, **e)** 0.323, **f)** 0.147

Solution of the homework:

λ = 2.00.

a) p(X = 0) = $e^{-\lambda} = e^{-2.00}$ = 0.135

b) p(exactly one collision in a week)

= p(X = 1) = λ e $^{-λ}$ = 2.00 e$^{-2.00}$

= 0.271

c) p(exactly two collisions in a week)

$p(X = 2) = \frac{\lambda^{x} * e^{-\lambda}}{x!} = \frac{\lambda^{2} * e^{-2}}{2!}$ = 0.271

d) p(not more than two collisions in a week)

= p(X ≤ 2)

= p(X = 0) + p(X = 1) + p(X = 2)

= 0.135 + 0.271 + 0.271

= 0.677

e) p(more than two collisions in a week)

= p(X > 2)

= 1– p(X ≤ 2)

= 1 – 0.677

= 0.323

f) We have λ = 2.00/week, therefore, λ = 4.00 in 2 weeks

Then:

p(exactly two collisions in a two-week interval)

$p(X = 2) = \frac{\lambda^{x} * e^{-\lambda}}{x!} = \frac{4^{2} * e^{-4}}{2!}$ =0.147

3- Normal Probability Distribution

A normal distribution is a continuous, symmetric, bell-shaped distribution of a variable.

It is completely defined by the mean (μ) and the standard deviation (σ), as presented in the following equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where μ is the mean value of theoretical distribution and σ is the standard deviation. $\pi = 3.14159$

This density function extends from –∞ to +∞. Its shape is shown in the figure below.

The first scale on the figure gives values of $\frac{x-\mu}{\sigma}$, which gives corresponding values of x. Thus, $\frac{x-\mu}{\sigma} = 0$ corresponds to x = μ, and $\frac{x-\mu}{\sigma} = -3$ corresponds to x = μ – 3σ.

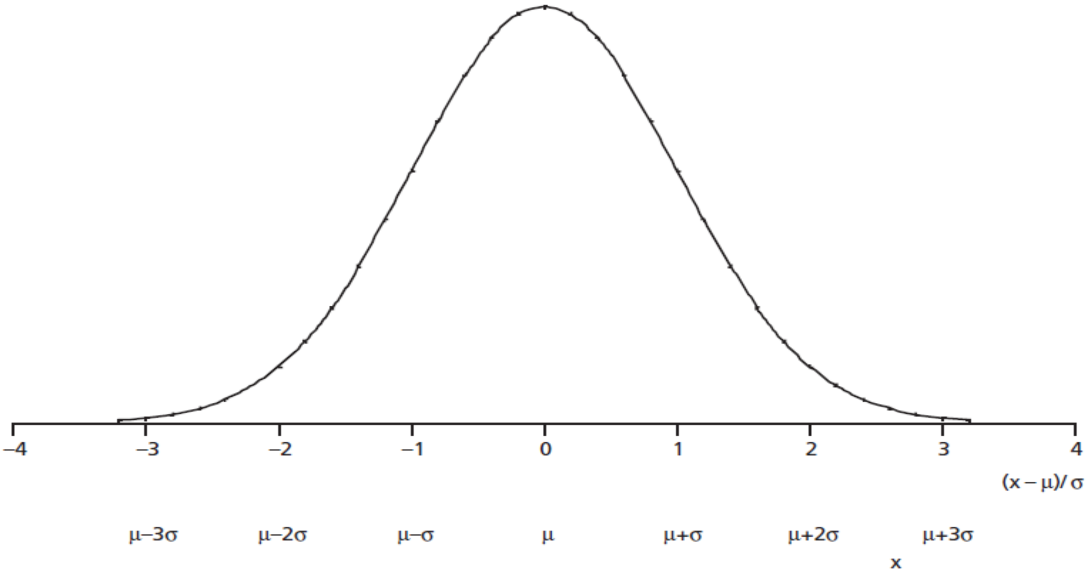


Figure: Shape of the normal distribution.

Because the normal probability density function is symmetrical, the mean, median and mode coincide at x = μ. Thus, the value of μ determines the location of the center of the distribution, and the value of σ determines its spread.

The properties of normal distribution:

1. A normal distribution curve is bell-shaped.

2. The mean, median, and mode are equal and are located at the center of the distribution.

3. A normal distribution curve is unimodal (i.e., it has only one mode).

4. The curve is symmetric about the mean, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center.

5. The curve is continuous; that is, there are no gaps or holes. For each value of X, there is a corresponding value of Y.

6. The curve never touches the x axis. Theoretically, no matter how far in either direction the curve extends, it never meets the x axis—but it gets increasingly closer.

7. The total area under a normal distribution curve is equal to 1.00, or 100%. This fact may seem unusual, since the curve never touches the x axis, but one **can prove it mathematically by using calculus**. (*The proof is beyond the scope of our course*)

Let $z = \frac{x - \mu}{\sigma}$,

Then z is a ratio between (x – μ) and σ. It represents the number of standard deviations between any point and the mean. This is used to simplify the integration of the probability function. The normal distribution becomes:

$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, $-\infty < z < \infty$