Mathematics Practical

UNIVERSITY OF DELHI

Ramanujan College

DSC 06 : Probability For Computing

Semester-2

(2023-24)

Submitted By:

~Ratnesh Kumar

20231432(23020570034)

Submitted To:

~Dr. Aakash

Acknowledgement

I would like to take this opportunity to acknowledge everyone who has helped us in every stage of this project.

I am deeply indebted to my mathematics Professor, Dr Aakash for his guidance and suggestions in completing this project. The completion of this project was possible under his guidance and support.

I am also very thankful to my parents and my friends who have boosted me up morally with their continuous support.

At last but not least, I am very thankful to God almighty for showering his blessings upon me.

Index

S.No	Topic
1.	Plotting and fitting of Binomial distribution and graphical representation of probabilities.
2.	Plotting and fitting of Multinomial distribution and graphical representation of probabilities.
3.	Plotting and fitting of Poisson distribution and graphical representation of probabilities.
4.	Plotting and fitting of Geometric distribution and graphical representation of probabilities.
5.	Plotting and fitting of Uniform distribution and graphical representation of probabilities.
6.	Plotting and fitting of Exponential distribution and graphical representation of probabilities.
7.	Plotting and fitting of Normal distribution and graphical representation of probabilities.
8.	Calculation of cumulative distribution functions for Exponential and Normal distribution.
9.	Given data from two distributions, find the distance between the distributions.
10.	Application problems based on the Binomial distribution.
11.	Application problems based on the Poisson distribution.
12.	Application problems based on the Normal distribution.
13.	Presentation of bivariate data through scatter-plot diagrams and calculations of covariance.
14.	Calculation of Karl Pearson's correlation coefficients.
15.	To find the correlation coefficient for a bivariate frequency distribution.
16.	Generating Random numbers from discrete (Bernoulli, Binomial, Poisson) distributions.
17.	Generating Random numbers from continuous (Uniform, Normal) distributions.
18.	Find the entropy from the given data set.

1. Plotting and fitting of Binomial distribution and graphical representation of probabilities.

Binomial Distribution:

The binomial distribution is a discrete probability distribution. It describes the outcome of binary scenarios, e.g. toss of a coin.

Binomial Distribution Formula:

$$P(x;n,p) = {}^{n}c_{x}\, p^{x}(1-p)^{n-x}$$

Where:

P(x; n, p) is the probability of x successes in n trials in an experiment which can result in exactly two outcomes (success or failure).
p is the probability of success on an individual trial.
n is the number of trials.
x is the total number of successes.

Mean(μ=np)

Variance(σ^2=npq)

Implementation in Excel:

=BINOM.DIST(number_s, trials, probability_s, cumulative)

Where:

number_s: number of successes.
trials: total number of trials.
probability_s: probability of success on each trial.

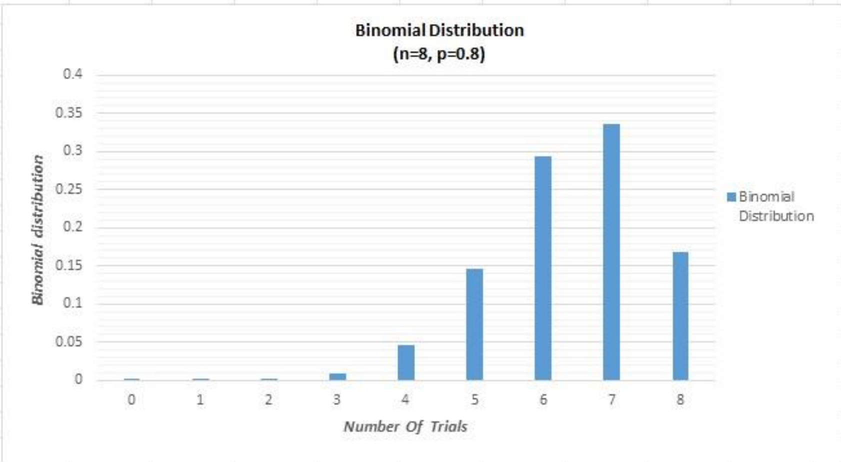
cumulative: TRUE returns the cumulative probability; FALSE returns the exact probability

Skewness In Case Binomial Distribution Can Be Defined As Follows:

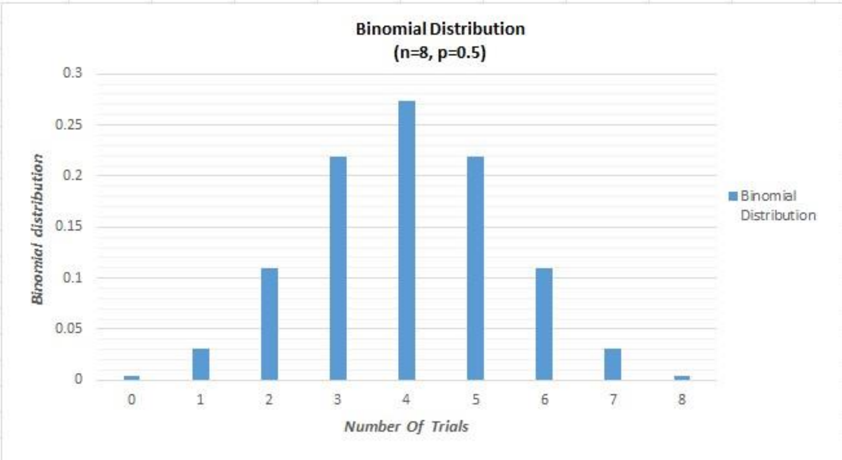
If $p = 0.5$, the binomial distribution will be symmetrical, regardless of the value of n.
If $p \neq 0.5$, the distribution will be skewed.
If $p < 0.5$, the distribution will be positively skewed or right-skewed. This means the bulk of the probability falls in the smaller numbers and the distribution tails off to the right.
If $p > 0.5$, the distribution will be negatively skewed or left-skewed. This means the

bulk of the probability falls in the larger numbers and the distribution tails off to the left.

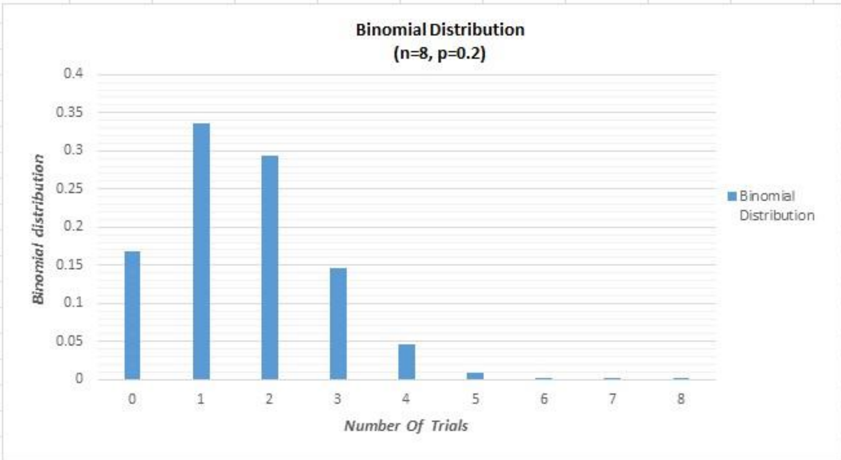
n(Number Of Trials)	8
p(Probability of Success on given Trails)	0.8
k(Number of success)	Binomial Distribution
0	0.00000256
1	0.00008192
2	0.00114688
3	0.00917504
4	0.0458752
5	0.14680064
6	0.29360128
7	0.33554432
8	0.16777216



n(Number Of Trials)	8
p(Probability of Success on given Trails)	0.5
k(Number of success)	Binomial Distribution
0	0.00390625
1	0.03125
2	0.109375
3	0.21875
4	0.2734375
5	0.21875
6	0.109375
7	0.03125
8	0.00390625



n(Number Of Trials)	8
p(Probability of Success on given Trails)	0.2
k(Number of success)	Binomial Distribution
0	0.16777216
1	0.33554432
2	0.29360128
3	0.14680064
4	0.0458752
5	0.00917504
6	0.00114688
7	0.00008192
8	0.00000256



2. Plotting and fitting of Multinomial distribution and graphical representation of probabilities.

*The multinomial distribution is a multivariate generalization of the binomial distribution. Consider a trial that results in exactly one of some fixed finite number k of possible outcomes, with probabilities $p_1, p_2, \ldots, p_k$ (so that $p_i \geq 0$ for $i = 1, \ldots, k$ and ∑i=1kpi=1), and there are n independent trials. Then let the random variables X_i indicate the number of times outcome number i was observed over the n trials. Then $X = (X_1, X_2, \ldots, X_k)$ follows a multinomial distribution with parameters n and **p**, where **p** $=(p_1, p_2, \ldots, p_k)$.*

Multinomial Distribution Formula

$$p(x_1, x_2 \ldots \ldots x_k) = \left[\frac{n!}{x_1!\,.\,x_2! \ldots\ldots x_k!}\right] . p_1^{x_1} . p_2^{x_2} \ldots \ldots . p_k^{x_k}$$

$$\mathrm{Cov}(X_i, X_j) = -np_i p_j \quad (i \neq j)$$

When $X = (x_1, x_2, \ldots, x_k)$ follows a multinomial distribution with the PMF given above, X_i follows a binomial distribution with n trials and success probability p_i.

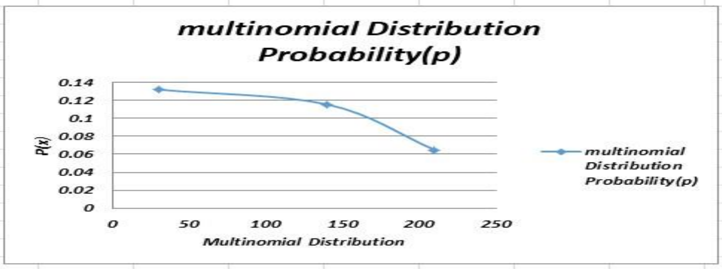
how to implement in excel

Multinomial = MULTINOMIAL(X1,X2,X3)

Probability = MULTINOMIAL*PRODUCT(p1^X1,p2^X2,p3^X3)

		1st Case	2nd Case	3rd Case
Red	X1	4	3	2
Green	X2	0	1	1
Blue	X3	6	3	2
P_Red	P1	0.375	0.375	0.375
P_Green	P2	0.125	0.125	0.125
P_Blue	P3	0.5	0.5	0.5
	Multinomial	210	140	30
	P1^X1	0.019775391	0.052734375	0.140625
	P2^X2	1	0.125	0.125
	P3^X3	0.015625	0.125	0.25
	Probability	0.064888	0.115356445	0.131835938

Multinomial	multinomial Distribution Probability(p)
210	0.064888
140	0.115356445
30	0.131835938



3. Plotting and fitting of Poisson distribution and graphical representation of probabilities.

The Poisson distribution is a type of discrete probability distribution that determines the likelihood of an event occurring a specific number of times (k) within a designated time or space interval. This distribution is characterized by a single parameter, λ (lambda), representing the average number of occurrences of the event.

Poisson Distribution Formula

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Mean	$\mu = E(X) = \lambda$
Variance	$\sigma^2 = V(X) = \lambda$
Standard Deviation	$\sigma = \sqrt{\sigma^2} = \sqrt{\lambda}$

Where:

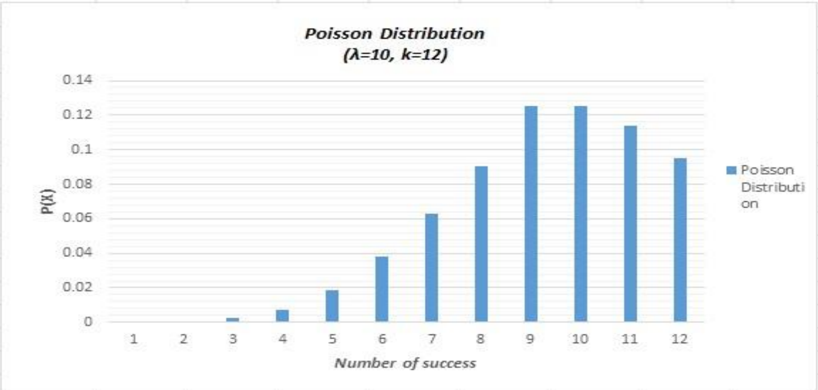
- $P(X=k)$ is the probability of observing k events
- e is the base of the natural logarithm (approximately 2.71828)
- λ mean number of success that occur during a specific interval, $\lambda = np$
- k is the number of success

how to implement in excel

POISSON.DIST(number_s,average,cumulative)

POISSON.DIST(k, λ ,FALSE)

λ(Mean number Of Success)	10
k(Number of success)	Poisson Distribution
0	4.53999E-05
1	0.000453999
2	0.002269996
3	0.007566655
4	0.018916637
5	0.037833275
6	0.063055458
7	0.090079226
9	0.125110036
10	0.125110036
11	0.113736396
12	0.09478033



4. Plotting and fitting of Geometric distribution and graphical representation of probabilities.

In a Bernoulli trial, the likelihood of the number of successive failures before success is obtained is represented by a geometric distribution, which is a sort of discrete probability distribution. A Bernoulli trial is a test that can only have one of two outcomes: success or failure. In other words, a Bernoulli trial is repeated until success is obtained and then stopped in geometric distribution.

A geometric distribution is a discrete probability distribution that indicates the likelihood of achieving one's first success after a series of failures. The number of attempts in a geometric distribution can go on indefinitely until the first success is achieved. Geometric distributions are probability distributions that are based on three key assumptions.

- *The trials that are being undertaken are self-contained.*
- *Each trial may only have one of two outcomes: success or failure.*
- *For each trial, the success probability, represented by p, is the same*

Geometric Distribution formula

$$P(X=k) = (1-p)^k p$$

Mean:	$\mu = E(X) = \frac{1}{p}$
Variance:	$\sigma^2 = V(X) = \frac{(1-p)}{p^2}$

where:

- **k:** number of failures before first success

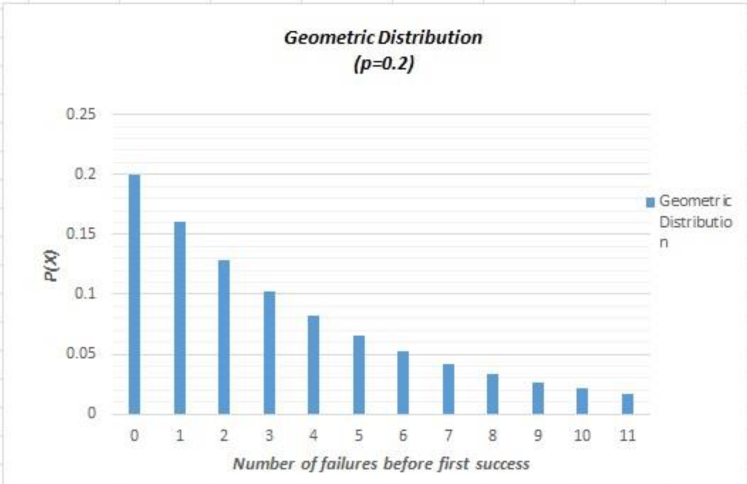
- **p:** probability of success on each trial

The chance of a trial's success is denoted by p, whereas the likelihood of failure is denoted by q, q = 1 – p in this case. X ~ G (p) represents a discrete random variable, X, with a geometric probability distribution.

how to implement in excel

Probability = (1-p)^k*p

Prob. Of Success on Given trial(p)	0.2
Number of failures before first success	Geometric Distribution
0	0.2
1	0.16
2	0.128
3	0.1024
4	0.08192
5	0.065536
6	0.0524288
7	0.04194304
8	0.033554432
9	0.026843546
10	0.021474836
11	0.017179869



5. Plotting and fitting of Uniform distribution and graphical representation of probabilities.

A uniform distribution is a distribution that has constant probability due to equally likely occurring events. It is also known as rectangular distribution (continuous uniform distribution). It has two parameters a and b: a = minimum and b = maximum. The distribution is written as U (a, b).

A uniform distribution is a type of probability distribution where every possible outcome has an equal probability of occurring. This means that all values within a given range are equally likely to be observed.

Uniform Distribution Formula

The probability density function (PDF) of a continuous uniform distribution defines the probability of a random variable falling within a particular interval. For a continuous uniform distribution over the interval [a,b].

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

$$\text{Mean } \mu = \frac{a+b}{2}$$

$$\text{Variance } \sigma^2 = \frac{(b-a)^2}{12}$$

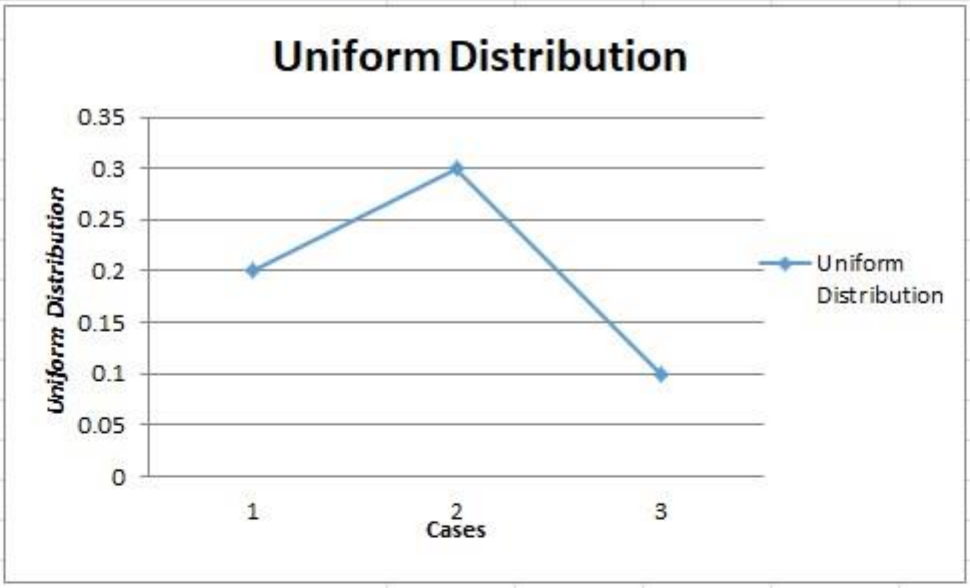
how to implement in excel

$$P = (x_2 - x_1) / (b - a)$$

For calculating probability, we need:

1. *a: minimum value in the distribution*
2. *b: maximum value in the distribution*
3. *x_1: the minimum value you're interested in*
4. *x_2: the maximum value you're interested in*

a	15			
b	25			
		case 1	case 2	case3
x1		16	19	23
x2		18	22	24
Uniform Distribution		0.2	0.3	0.1



6. Plotting and fitting of Exponential distribution and graphical representation of probabilities.

The support (set of values the Random Variable can take) of an Exponential Random Variable is the set of all positive real numbers. Suppose we are posed with the question- How much time do we need to wait before a given event occurs? The answer to this question can be given in probabilistic terms if we model the given problem using the Exponential Distribution. Since the **time** we need to wait is unknown, we can think of it as a Random Variable. If the probability of the event happening in a given interval is proportional to the length of the interval, then the Random Variable has an exponential distribution. The support (set of values the Random Variable can take) of an Exponential Random Variable is the set of all positive real numbers.

This distribution can be used to solve following type of real life problems-

- How long does a shop owner need to wait until a customer enter a shop.
- How long will a battery continue to work before it dies.
- How long will a computer continue to work before it breakdown.

$$f(x) = \begin{cases} \lambda e^{-\lambda * x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Here λ is the rate parameter and its effects on the density function .

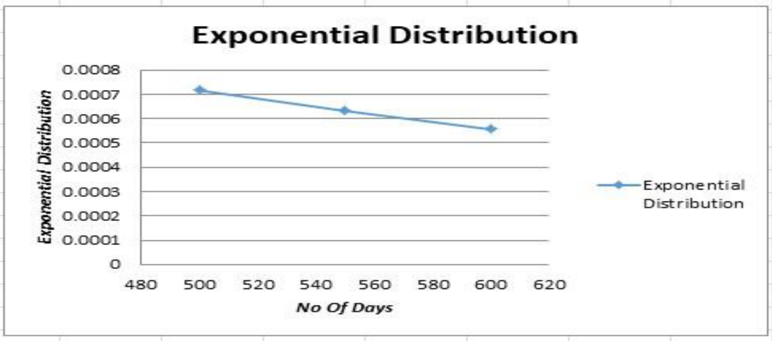
e is a constant roughly equal to 2.718

How to Implement in excel

EXPON.DIST(X,lambda,cumulative)

EXPON.DIST(X,lambda,FALSE)

				case 1	case 2	case 3
μ	400		x	500	550	600
λ	0.0025					
		Exponential Distribution		0.000716	0.000632	0.000558
		500	0.000716			
		550	0.000632			
		600	0.000558			



7. Plotting and fitting of Normal distribution and graphical representation of probabilities.

We define Normal Distribution as the probability density function of any continuous random variable for any given system. Now for defining Normal Distribution suppose we take f(x) as the probability density function for any random variable X.

$f(x) \geq 0$ $x \in (-\infty,+\infty)$,

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

where,

- *x is **Random Variable***
- *μ is **Mean***
- *σ is **Standard Deviation***

Properties of Normal Distribution

- *For normal distribution of data, mean, median, and mode are equal, (i.e., Mean = Median = Mode).*
- *Total area under the normal distribution curve is equal to 1.*
- *Normally distributed curve is symmetric at the center along the mean.*
- *In a normally distributed curve, there is exactly half value to the right of the central and exactly half value to the right side of the central value.*
- *Normal distribution is defined using the values of the mean and standard deviation.*
- *Normal distribution curve is a Unimodal Curve, i.e. a curve with only one peak*

how to implement in excel

1. Input your data set into an Excel spreadsheet

2. Find the mean of your data set

=AVERAGE(cell range)

- *"cell range" is a required component and the range of cells where your data exists, such as cells A1 through A64. You can write this in the function as A1:A64.*

3. Find the standard deviation of your data set

=STDEV(cell range)

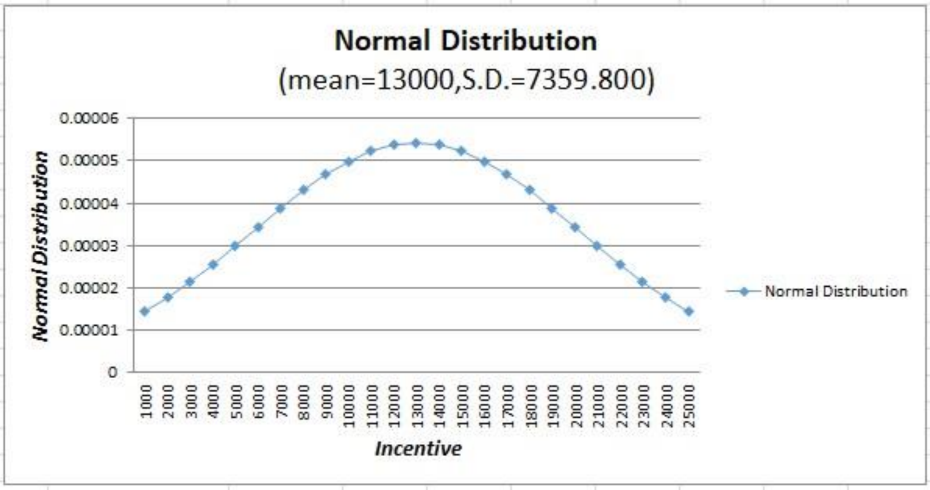
4. Select a value for the distribution

5. Type the NORM.DIST function and fill

NORM.DIST(x,mean,standarddeviation,cumulative)

NORM.DIST(x,mean,standarddeviation,FALSE)

Employee	incentive	Normal Distribution		
EMP1	1000	1.43472E-05		
EMP2	2000	1.77407E-05	mean	13000
EMP3	3000	2.15356E-05	standand Deviation	7359.800722
EMP4	4000	2.56641E-05		
EMP5	5000	3.00245E-05		
EMP6	6000	3.44833E-05		
EMP7	7000	3.88799E-05		
EMP8	8000	4.3035E-05		
EMP9	9000	4.6763E-05		
EMP10	10000	4.98843E-05		
EMP11	11000	5.22406E-05		
EMP12	12000	5.37075E-05		
EMP13	13000	5.42056E-05		
EMP14	14000	5.37075E-05		
EMP15	15000	5.22406E-05		
EMP16	16000	4.98843E-05		
EMP17	17000	4.6763E-05		
EMP18	18000	4.3035E-05		
EMP19	19000	3.88799E-05		
EMP20	20000	3.44833E-05		
EMP21	21000	3.00245E-05		
EMP22	22000	2.56641E-05		
EMP23	23000	2.15356E-05		
EMP24	24000	1.77407E-05		
EMP25	25000	1.43472E-05		



8. Calculation of cumulative distribution functions for Exponential and Normal distribution.

Normal Distribution					Exponential Distribution			
Employee	incentive	Normal Distribution			No Of Days	Exponential Distribution		
EMP1	1000	0.572137292			300	0.527633447		
EMP2	2000	0.641935216	mean	5500	320	0.550671036		
EMP3	3000	0.707279533	standand Deviation	3027.650354	350	0.58313798	λ	0.0025
EMP4	4000	0.766470549			400	0.632120559		
EMP5	5000	0.81834893			420	0.650062251		
EMP6	6000	0.862343557			450	0.675347533		
EMP7	7000	0.898442582			500	0.713495203		
EMP8	8000	0.92710243			520	0.727468207		
EMP9	9000	0.949118248			550	0.747160404		
EMP10	10000	0.965481826			600	0.77686984		

9. Given data from two distributions, find the distance between the distributions.

Euclidean distance

Euclidean distance is the distance between two real distinct value .It is calculate by the square root of the sum of the squared difference elements in two vectors.

$$\text{Euclidean Distance} = |X - Y| = \sqrt{\sum_{i=1}^{i=n} (x_i - y_i)^2}$$

X: Array or vector X

Y: Array or vector Y

x_i: Values of horizontal axis in the coordinate plane

y_i: Values of vertical axis in the coordinate plane

n: Number of observations

how to implement in excel

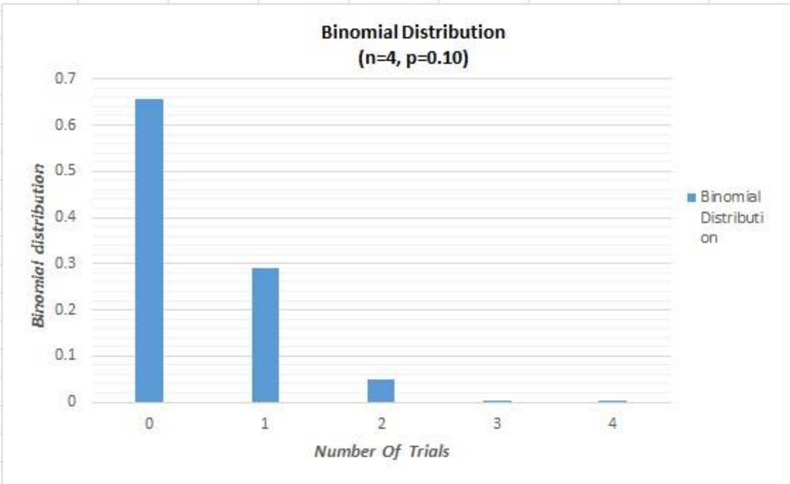
= ***SORT(SUM X MYZ(array_X,array_Y))***

Trials	Binomial Distribution	Poisson Distribution		distance			
1	0.08957952	0.268128018		0.178548		n	0.4
2	0.20901888	0.053625604		0.155393		k	8
3	0.27869184	0.00715008		0.271542		λ	0.4
4	0.2322432	0.000715008		0.231528			
5	0.12386304	5.72006E-05		0.123806			
6	0.04128768	3.81338E-06		0.041284			
7	0.00786432	2.17907E-07		0.007864			
8	0.00065536	1.08954E-08		0.000655			

10. Application problems based on the Binomial distribution.

Ques: Antibiotics occasionally cause nausea as a side effect. A major drug company has developed a new antibiotic called Phe-Mycin. The company claims that, at most, 10 percent of all patients treated with Phe-Mycin would experience nausea as a side effect of taking the drug. Suppose that we randomly select n = 4 patients and treat them with Phe-Mycin. Each patient will either experience nausea (which we arbitrarily call a success) or will not experience nausea (a failure). We will assume that p, the true probability that a patient will experience nausea as a side effect, is .10, the maximum value of p claimed by the drug company. Furthermore, it is reasonable to assume that patients' reactions to the drug would be independent of each other. Let x denote the number of patients among the four who will experience nausea as a side effect. It follows that x is a binomial random variable, which can take on any of the potential values 0, 1, 2, 3, or 4. That is, anywhere between none of the patients and all four of the patients could potentially experience nausea as a side effectSuppose that we wish to investigate whether p, the probability that a patient will experience nausea as a side effect of taking Phe-Mycin, is greater than.10, the maximum value of p claimed by the drug company. This assessment will be made by assuming, for the sake of argument, that p equals .10, and by using sample information to weigh the evidence against this assumption and in favor of the conclusion that p is greater than .10. Suppose that when a sample of n=4 randomly selected patients is treated with Phe-Mycin, three of the four patients experience nausea. Because the fraction of patients in the sample that experience nausea is 3/4 = .75, which is far greater than .10, we have some evidence contradicting the assumption that p equals .10.

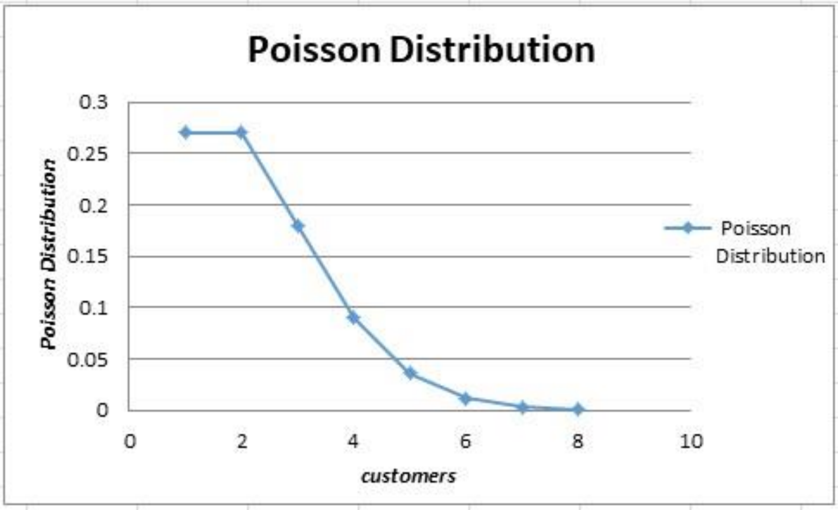
n(Number Of Trials)	4
p(Probability of Success on given Trails)	0.1
k(Number of success)	Binomial Distribution
0	0.6561
1	0.2916
2	0.0486
3	0.0036
4	0.0001



11. Application problems based on the Poisson distribution.

Ques: In a cafe, the customer arrives at a mean rate of 2 per min. Find the probability of arrival of 5 customers in 1 minute using the Poisson distribution formula.

λ	2
x	5
x	Poisson Distribution
1	0.270670566
2	0.270670566
3	0.180447044
4	0.090223522
5	0.036089409
6	0.012029803
7	0.003437087
8	0.000859272

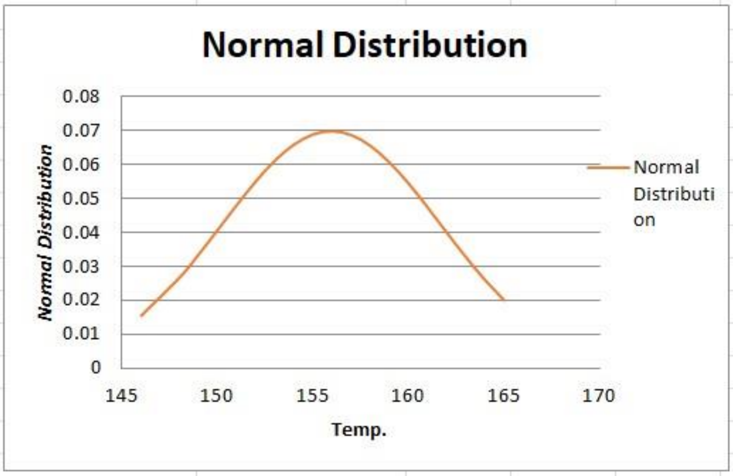


12. Application problems based on the Normal distribution.

Ques:: According to the website of the American Association for Justice, ^11 Stella Liebeck of Albuquerque, New Mexico, was severely burned by McDonald's coffee in February 1992. Liebeck, who received third-degree bums over 6 percent of her body, was awarded $160,000 in compensatory damages and $480,000 in punitive damages. A post-verdict investigation revealed that the coffee temperature at the local Albuquerque McDonald's had dropped from about 185 degree F before the trial to about 158 degree after the trial. This case concerns coffee temperatures at a fast-food restaurant. Because of the possibility of future litigation and to possibly improve the coffee's taste, the restaurant wishes to study the temperature of the coffee it serves. To do this, the restaurant personnel measure the temperature of the coffee being dispensed (in degrees Fahrenheit) at a randomly selected time during each of the 24 half-hour periods from 8 a.m. to 7:30 p.m on a given day. This is then repeated on a second day, giving the 48 coffee temperatures in excel..

Temp.	Normal Distribution	Temp.	Normal Distribution
146	0.015374765	166	0.014888007
148	0.026566325	167	0.010783701
149	0.033356691	168	0.007575753
150	0.04062198	169	0.00516191
151	0.047980623	171	0.002186551
152	0.054966405	173	0.000819625
153	0.06107387		
154	0.065817321		
155	0.068794167		
156	0.069741244		
157	0.068573201		
158	0.065395191		
159	0.060487252		
160	0.054263594		
161	0.047214991		
162	0.039845376		
163	0.03261389		
164	0.025891304		
165	0.019935719		

mean	155.947
standand Deviation	5.72008



13. Presentation of bivariate data through scatter-plot diagrams and calculations of covariance.

Bivariate Data/ Bivariate Analysis

Bivariate analysis is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes occurred between the two variables and to what extent.

The term bivariate analysis refers to as the analysis of two variables . the objective of bivariate analysis to understand the relationship between two variables. There are three common way to analysis the bivariate analysis –

1. Scatter plots

2. Correlation Coefficient

3. Simple linear Regression(SLR)

Bivariate frequency distribution

A series of statistical data showing the frequency of two variables simultaneously is called Bivariate frequency distribution. In other words, the frequency distribution of two variable is called Bivariate frequency distribution. For example: sales and advertisement expenditure , weight and height of an individual.

Why bivariate frequency distribution is significant in business research ?

1. Decision Making

2. Market-segmentation

3. Risk-assessment

4. Resource allocation

how to implement in excel

= COVARIANCE.P(array1,array2)

The COVARIANCE.P function used the following arguments array1, this is range or array of integer value. array2 is also the second range or values.

Few things to remember about argument

1. If the given array contain text or logical value then are ignore by the Covariance function in excel.

2. The data should contain numbers, names, array or references that are numeric .IF the some cell do not contain numeric data they are ignored.

3. The data set should be same size with the same number of data points.

4. The data set should not be empty nor should the standard Deviation of the value equal .

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

X and Y are the sample mean of the two set of values and n is the sample size.

5. Covariance is measure to indicate the extent to which two random variable in tandem.

6. Correlation is the measure used to represent how strongly two random variable are strongly related to each other.

7. Covariance is nothing but a measure of correlation.

8. Correlation referred to the scaled form of covariance.

9. Covariance can ary between −∞ to +∞ and correlation range between -1 to +1 .

10. Covariance indicate the direction of the linear relationship between variables .

11. Correlation on the other hand measure both the strength and direction of the linear relationship between two variables.

12. Covariance is affected by change in scale.

13. Correlation is not affected by the change in scale.

Pearson Correlation Coefficient formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

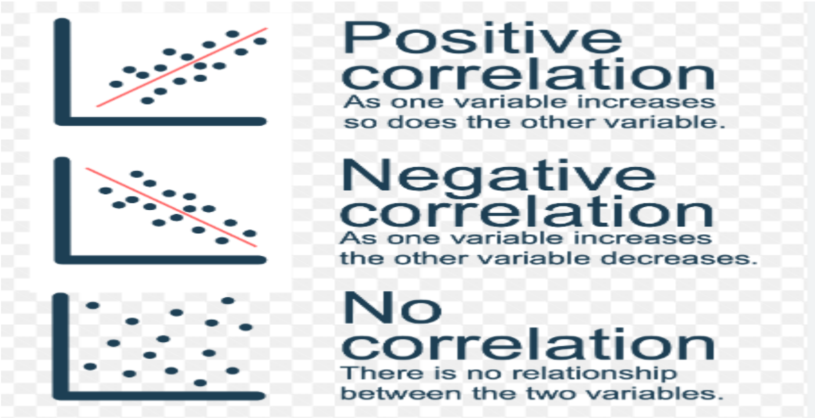
r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable



= PEARSON(arry:arry2)

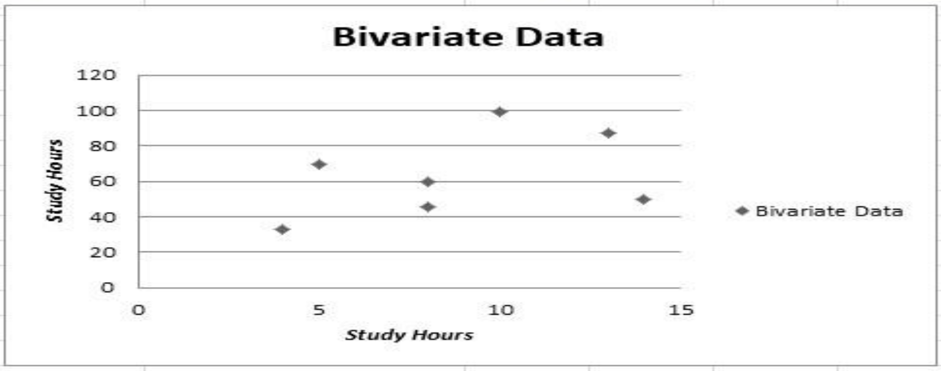
Scatter plots

Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis. These plots are often called scatter graphs or scatter diagrams.

Scatter plots instantly report a large volume of data. It is beneficial in the following situations –

- For a large set of data points given
- Each set comprises a pair of values
- The given data is in numeric form

Study Hours	10	13	14	8	5	4	8
Marks	99	87	50	60	70	33	46



covariance	34.92857143

Click insert tab along the top ribbon then click scatter chart within chart group.

CORRELATION in excel and COVARIANCE in excel –

= CORREL(hours,score)

= COVARIANCE.P(hours,score)

14. Calculation of Karl Pearson's correlation coefficients.

X	Y1	Y2	Y3
2	80	65	10
5	95	69	30
6	76	60	15
8	58	95	25
10	67	80	10
Karl Pearson's correlation	XY1	-0.62732	
	XY2	0.633225	
	XY3	0.018149	

15. To find the correlation coefficient for a bivariate frequency distribution.

	Age In Years				
Marks	16_18	18_20	20_22	22_24	Total
10_20	2	1	1	0	4
20_30	3	2	3	1	9
30_40	3	3	5	6	17
40_50	2	2	3	4	11
50_60	0	1	2	2	5
60_70	0	1	2	1	4

correlation coefficient along x-axis

-0.18773

correlation coefficient along Y-axis

0.774597

Marginal Frequency Distribution of X:

Marks	Total
15	4
25	9
35	17
45	11
55	5
65	4

Age In Years	Total
17	10
19	10
21	16
23	14

16. Generating Random numbers from discrete (Bernoulli, Binomial, Poisson) distributions.

How to implement in excel-

= BINOM.INV (1,P, RAND()) will generate 1 or 0 with chance of 1 being P random number

Random numbers from Binomial distributions
n=10
p=0.4
7
4
6
4
3

17. Generating Random numbers from continuous (Uniform, Normal) distributions.

= NORMINV(RAND(),B2,C2)

Where this RAND() function create your probability . B2 provides you mean , C2 refers your standand deviation.

Random numbers from Normal distributions
mean=20
standand Deviation=5
18.46540668
19.13569112
22.4762389
21.37881651
22.72814895

18. Find the entropy from the given data set.

The entropy of a random variable is the average level of information, surprise, or uncertainly inherent to the variable's possible outcomes. Given a discrete random random variable X which takes value in the alphabet x and distributed according to the P:x[0,1]. The entropy is H[X]

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

The choice of base for log varies for different applications.

Base 2 gives the unit of bits while base e gives ***natural units.***

Base o gives the units of H(X).

An equivalent definition of entropy is the expected value of the self information of a variable .

	P(x)	log2p(x)	H(x)
HH	1	0	0
HT	0.5	-1	0.5
TH	0.5	-1	0.5
TT	0	0	0

P(x)	H(x)
1	0
0.5	1
0	0