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

Suppose a big building constructing takes 100 days on average after an update occur find the probability that it will construct in more than 110 days and so on as in data ?

λ	0.045					
		x	110	30	170	50
		Exponential distributio	0.00708	0.25924	0.00048	0.1054

Normal Distribution

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.

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

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

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, standard deviation, cumulative)

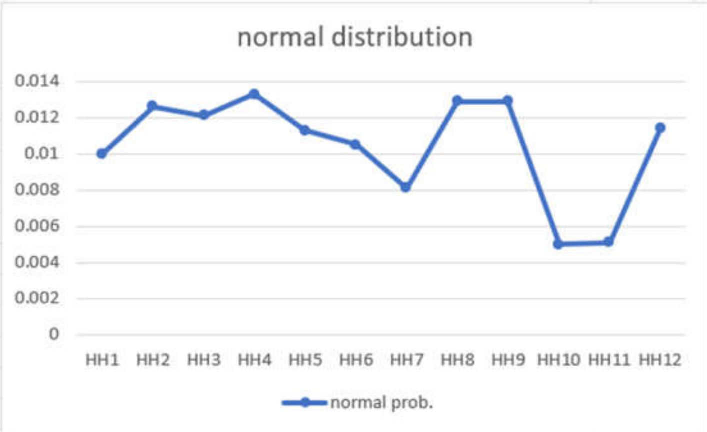
NORM.DIST(x, mean, standard deviation, FALSE)

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

Plot the normal graph according to the data of households and their water usage as shown below .

Household	Water Usage (liters)	Normal Probability Density
HH1	280	0.01
HH2	240	0.0126
HH3	260	0.0121
HH4	250	0.0133
HH5	270	0.0113
HH6	230	0.0105
HH7	220	0.0081
HH8	255	0.0129
HH9	245	0.0129
HH10	300	0.005
HH11	210	0.0051
HH12	235	0.0114
Mean (μ)	250 liters	
Std. Deviation (σ)	30 liters	

Normal Distribution: Daily Water Usage of 20 Households



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

Calculate normal graph and exponential cumulative distribution according to the data of as students and their exam scores shown below .

Delivery Time (min)	Normal Distribution (Cumulative Probability)			Time Between Calls (sec)	Exponential Distribution (Cumulative Probability)
20	0.0062			60	0.2592
30	0.0668			90	0.3624
35	0.1587			120	0.4512
40	0.3085			150	0.5276
45	0.5			180	0.5934
50	0.6915			210	0.6502
55	0.8413			240	0.699
60	0.9332			270	0.7411
65	0.9772			300	0.7774
Mean (μ): 45 minutes				$\lambda = 0.00556$	
Standard Deviation (σ): 10 minutes					

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.

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

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))***

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

In our case , we take data from the binomial and poisson distribution to find the eucliean distance between them.

trials	bionmial distribution	poisson distribution		distance			
1	0.301989888	0.259867336		0.099021679			
2	0.301989888	0.049374794		0.09724737			
3	0.176160768	0.006254141		0.033432984		prob.	0.2
4	0.066060288	0.000594143		0.004564722		k	9
5	0.016515072	4.51549E-05		0.000278906		λ	0.38
6	0.002752512	2.85981E-06		7.64781E-06			
7	0.000294912	1.55247E-07		8.7221E-08			
8	0.000018432	7.37422E-09		3.39467E-10			

10. Application problems based on the Binomial distribution.

A customer support agent can handle a maximum of 8 support tickets per day. For each ticket, the probability of successfully resolving it is 0.6. The number of tickets resolved in a day follows a binomial distribution with:

- n = 8 trials (maximum tickets per day)
- p = 0.6 (probability of resolving a single ticket)

Based on this information:

a) What is the probability that the agent resolves exactly 5 tickets in a day?

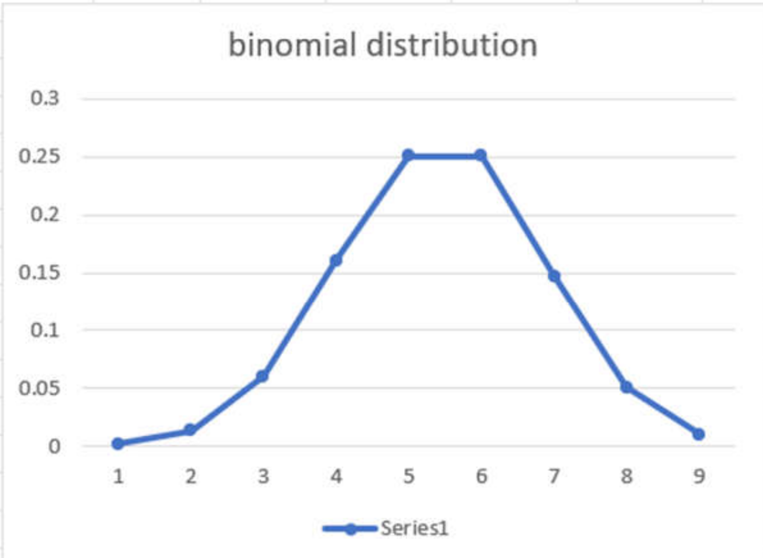
b) What is the probability that the agent resolves at least 6 tickets in a day?

c) Is it likely for the agent to resolve 2 or fewer tickets in a day? Justify your answer using the binomial probabilities.

k (Tickets Resolved)	Binomial Probability
0	0.00168
1	0.0135
2	0.0605
3	0.1608
4	0.2508
5	0.2508
6	0.1469
7	0.0506
8	0.0101

n (number of trials) = 8 (max tickets per day)

p (success probability) = 0.6 (chance of resolving one ticket)



11. Application problems based on the Poisson distribution.

Ques:. : A network server receives data packets at an average rate of **10 packets per second**. The arrival of packets follows a **Poisson distribution** with a mean (λ) of **10**.

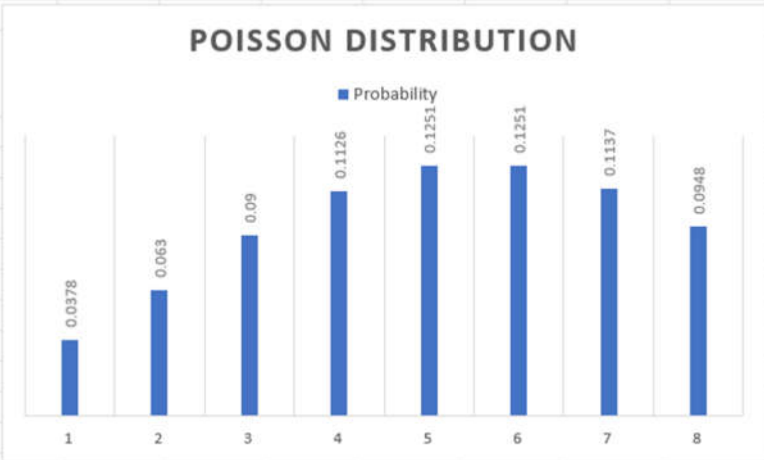
Using this model:

a) What is the probability that exactly **9 packets** are received in a given second?

b) What is the probability that the server receives **between 7 and 11 packets inclusive** in a second?

c) Based on the distribution, would receiving **only 5 packets** in a second be considered unusual? Explain using the probability values.

λ (Lambda)	10
X (Packets per second)	**Probability**
5	0.0378
6	0.063
7	0.09
8	0.1126
9	0.1251
10	0.1251
11	0.1137
12	0.0948



12. Application problems based on the Normal distribution.

Ques:: The systolic blood pressure of adults in a certain population is normally distributed with:

- Mean (μ) = 120 mm Hg
- Standard deviation (σ) = 10 mm Hg

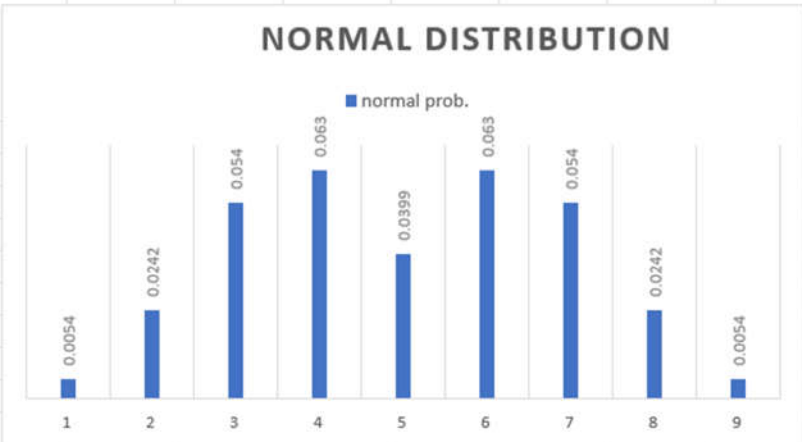
Using this distribution:

a) What is the probability density of a person having a systolic BP of 110 mm Hg?

b) Compare the likelihood (based on probability density) of systolic BP being 115 mm Hg versus 130 mm Hg.

c) Is it more common for individuals to have a systolic BP between 110 and 130 mm Hg, or outside this range? Justify your answer using the data and graph.

Systolic BP (mm Hg)	Normal Distribution (Probability Density)
90	0.0054
100	0.0242
110	0.054
115	0.063
120	0.0399
125	0.063
130	0.054
135	0.0242
140	0.0054
Mean (μ)	120 mm Hg
Standard Deviation	10 mm Hg



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* $-\infty$ *to* $+\infty$ *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

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

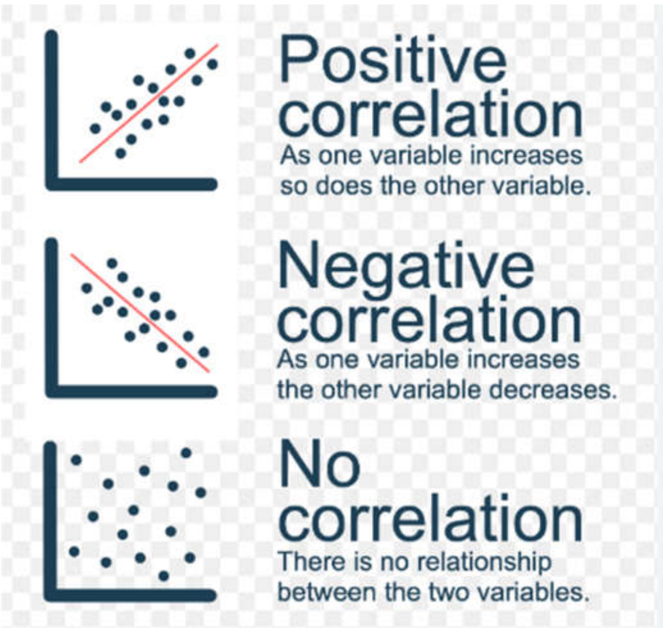
r = correlation coefficient

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*

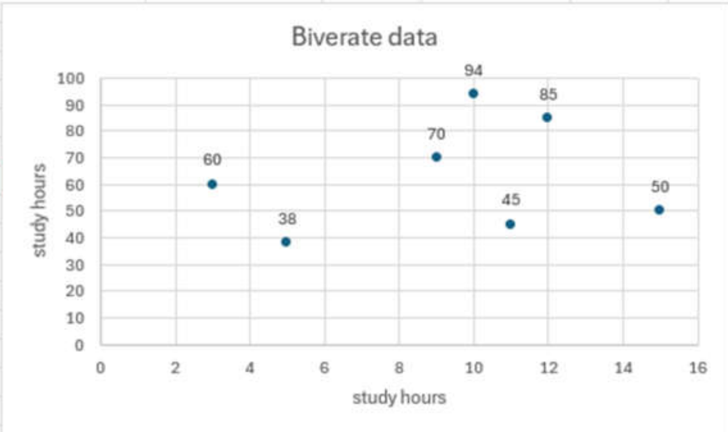
13. Presentation of bivariate data through scatter-plot diagrams and calculations of covariance.

How to perform variant analysis and experiment using the following dataset having two variables .

a) hours spent studying and

b) exam score receive by students

study hours	10	12	15	3	9	11	5
marks	94	85	50	60	70	45	38
CORRELATION	0.19533913						
COVARIANCE	14.3877551						



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

CORRELATION in excel-

= CORREL(hours, score)

COVARIANCE in excel –

COVARIANCE.P(hours, score)

14. Calculation of Karl Pearson's correlation coefficients.

X	Y1	Y2	Y3
2	38	63	10
3	56	59	35
8	74	68	15
11	33	79	24
10	40	85	12
Karl pearson's correlation	XY1	-0.14493936	
	XY2	0.889803569	
	XY3	-0.140953713	

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

Marks	16_18	18_20	20-22	22_24	total
10_20	5	1	1	0	7
20_30	2	0	4	2	8
30_40	1	3	5	3	12
40_50	2	2	2	1	7
50_60	3	0	3	1	7
60_70	0	1	0	5	6
total	13	7	15	12	

correlation coffecient along X axis

-0.109

corelation cofficient along y-axis

0.15028

margerial frequency distribution of X:

marks	total
15	7
25	8
34	12
33	7
53	7
42	6

age in years	total
18	13
12	7
11	15
14	12

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

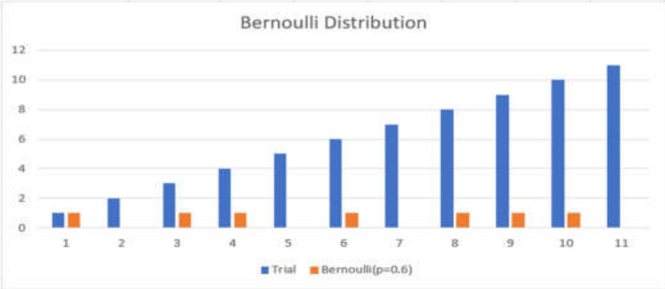
How to implement in excel-

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

Bernoulli Distribution

Trial	Bernoulli(p=0.6)
1	1
2	0
3	1
4	1
5	0
6	1
7	0
8	1
9	1
10	1
11	0

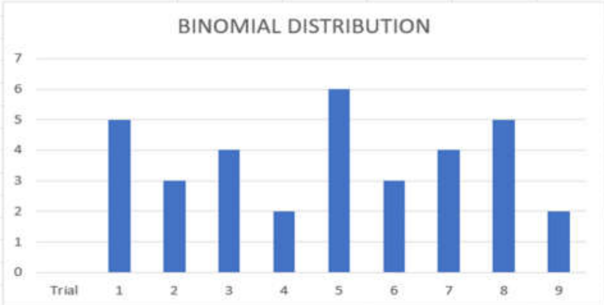
p=0.6



Binomial Distribution

Trial	Binomial(n=10, p=0.4)
1	5
2	3
3	4
4	2
5	6
6	3
7	4
8	5
9	2

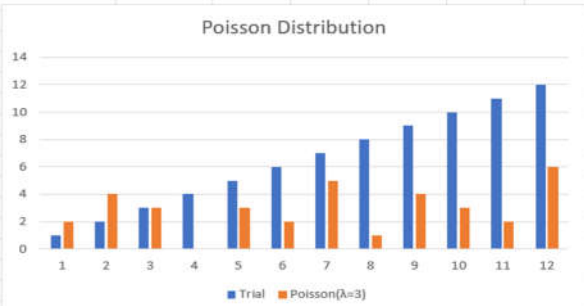
n=10, p=0.4



Poisson Distribution

Trial	Poisson(λ=3)
1	2
2	4
3	3
4	0
5	3
6	2
7	5
8	1
9	4
10	3
11	2
12	6

λ=3

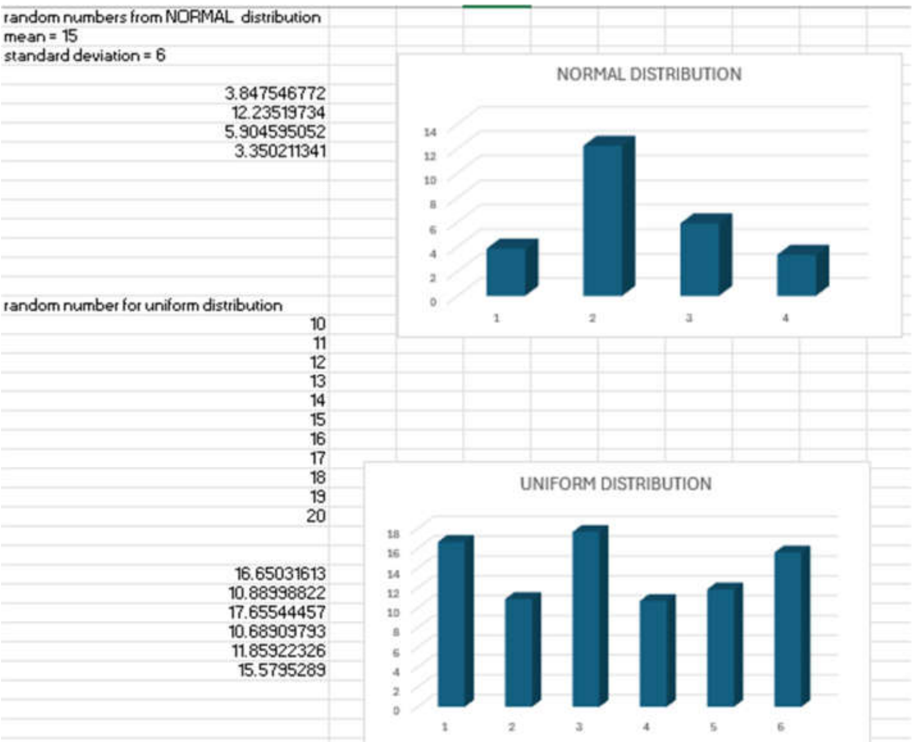


17. Generating Random numbers from continuous (Uniform, Normal) distributions.

How to implement in excel-

= NORMINV(RAND(),B2,C2)

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



Entropy

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 .

Two bits of entropy

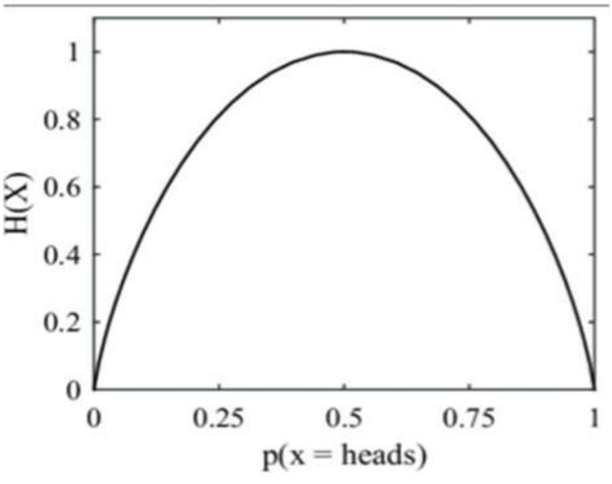
In case of a fair coin tosses, the information entropy enbits base 2 logarithm the number of possible outcomes with two coin, there are 4 possible outcomes and two bits of entropy

{HH,HT,TH,TT}

Generally information entropy is the average amount of information converged by an event, when considering all possible outcomes.

Example –

Entropy H(X) of a coin flip measured in bits, graph versus the bias of the coin where X=1 represent entropy is the result of heads.



Here the entropy is at most 1 bit and to communicate the outcome of a coin flip (2 possible values – H or T) which requires an average of 1 bit (exactly 1 bit for a fair coin)

The result of a fair dice (6 possible values) have entropy log 6 bits.

18. Find the entropy from the given data set.

HH	1	0	0
HT	0.5	-1	0.5
TH	0	0	0
TT	0.5	-1	0.5

P(X)	H(X)
1	0
0.5	1
0	0