SPREADSHEET MODELING & SIMULATION OF PROBLEMS IN PROBABILITY

Jonaki B Ghosh
Department of Elementary Education
Lady Shri Ram College
jonakibghosh@lsr.edu.in

4th January 2019

MTA(I) Inaugural Conference
HBCSE, TIFR, Mumbai

The Menu

- Fibonacci Sequence & Golden Ratio
- Exploring Chaos
- Birthday Paradox
- Monty Hall Problem
- Probability Distributions
- Central Limit Theorem
- Regression
- Hill Cipher

The Fibonacci Puzzle

The Fibonacci puzzle was posed by Leonardo of Pisa who considered the growth of an idealized rabbit population. A newly born pair of rabbits are put in the field, are able to mate at the age of one month so that at the end of the second month a female can produce a pair of rabbits. Rabbits never die and a mating pair always produces a new pair every month from the second month on. How many pairs will there be in one year?

The Fibonacci Sequence

The Fibonacci sequence can be generated by selecting two initial values (say 1 and 1), and adding them is to get the 3rd term of the sequence. Each subsequent term is obtained by adding the two previous terms. Thus the sequence is 1,1,2,3,5,8,13,21,……….

The sequence may be written recursively as

$$F_0 = F_1 = 1, \quad F_{n+2} = F_{n+1} + F_n$$

The Fibonacci Sequence

Generate the first 50 Fibonacci Numbers

Step 1: Enter 1 in cell A2 and =AI + 1 in A3 and drag till A51.

Step 2: Enter 1 in B2 and B3. Enter = B2 + B3 in B4 and double click in the corner of cell B4.

Step 3: Enter = B3/B2 in C3 and double click in the corner of cell C3.

Step 4: Enter = B4/B2 in D4 and double click in the corner of cell D4.

Step 5: Enter = B5/B2 in E5 and double click in the corner of cell E5.

Understanding Chaos

To consider the chaotic behaviour of iterative procedures, consider the function

$f(x) = ax(1 - x)$

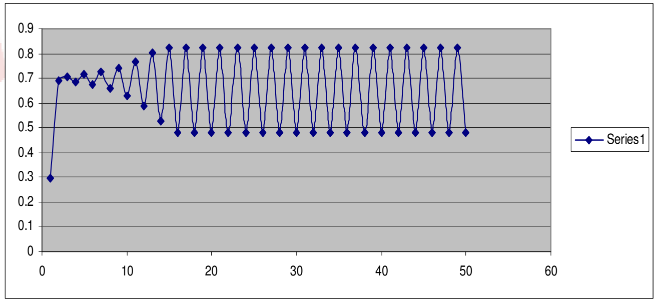
Let us observe the iterations of this function using

$a = 3.3$ …..taking the initial value $x_0 = 0.1, 0.6, 0.7, \ldots\ldots\ldots$

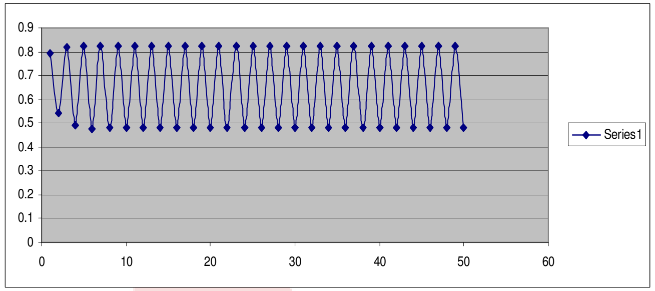
$a = 3.8$ …..taking the initial value $x_0 = 0.2, 0.21, 0.22, \ldots\ldots$

Sensitivity to initial conditions

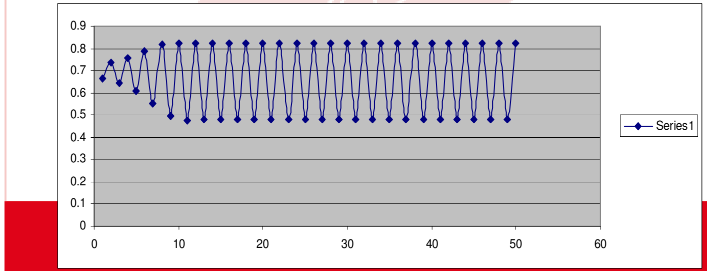
$a = 3.3, x_0 = 0.1$



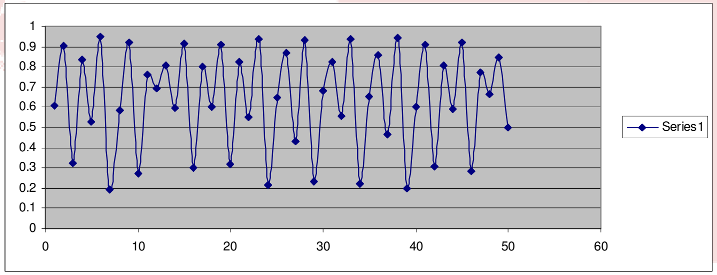
$a = 3.3, x_0 = 0.6$



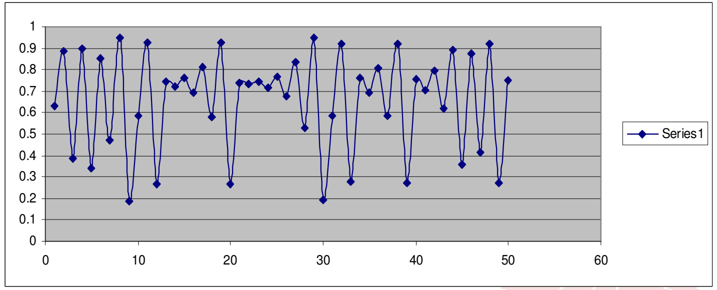
$a = 3.3, x_0 = 0.72$



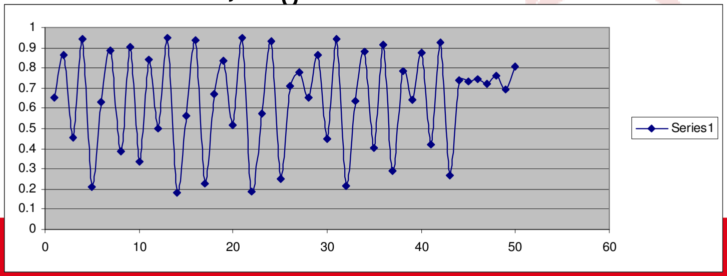
$a = 3.8, x_0 = 0.2$



$a = 3.8, x_0 = 0.21$



$a = 3.8, x_0 = 0.22$



SIMULATING PROBLEMS IN PROBABILIY

What is Simulation?

- **Simulation** is a modeling tool which is used to imitate a real-world process in order to understand system behavior.

- **Monte Carlo simulation** - A problem solving technique used to approximate the probability of certain outcomes by running multiple trial runs, called simulations, using random variables.

- The true behavior of a system is estimated using **distributions**.

- **Random numbers** from these distributions can be generated to evaluate multiple strategies and predict future performance.

The Birthday Paradox

How many people do you need in a group to ensure that there are at least two people having the same birthday?

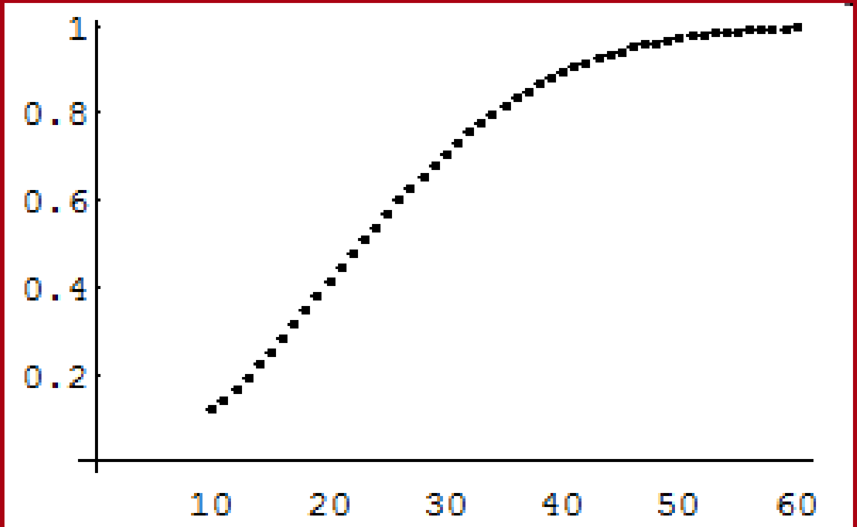
How many do you need so that the probability of at least two of them having the same birthday is about half?

Generalizing the Probability of a Birthday Match

Group Size	Probability of a Repeated Birthday
3	$1-\frac{365\times364\times363}{365^3}=0.0082$
4	$1-\frac{365\times364\times363\times362}{365^4}=0.0164$
5	$1-\frac{365\times364\times363\times362\times361}{365^5}=0.0271$
n	$1-\frac{365\times364\times.....\times(365-(n-1))}{365^n}=1-\frac{365!}{(365-n)!\times365^n}$

Mathematica Program for generating probabilities of a repeated birthday for group sizes (*n*) varying from 10 to 60

```
prob[n_]:=1-((365!/(365-n)!)*(1/365^n))
Table[{n,N[prob[n]]},{n,10,60}]//TableForm
ListPlot[prob,PlotStyle->PointSize[0.03]]
```



The output indicates that the probability of at least one repeated birthday is 0.507 for a group of 23 people. The probability approaches 1 as *n* approaches 50.

Simulating the Birthday Problem using Mathematica

```
month=Table[Random[Integer,{1,12}],{i,1,23}];
day=Table[Random[Integer,{1,31}],{i,1,23}];
bday=100*month+day;
Sort[bday]//TableForm
```

This program outputs 3 or 4 digit numbers. The first one or two digits indicates the month and the last two digits the day of the month.

e.g 127 (27th January), 1120 (20th November)

Use this program to simulate 10 sets of 23 birthdays each. At least 5 out of 10 should contain a match.

Simulation	Repeated birthday/No match
1	No match
2	No match
3	No match
4	524 (24th May), 1210 (10th December)
5	530 (30th May)
6	903 (3rd September), 1201 (1st December)
7	716 (16th July)
8	621 (21st June)
9	No match
10	No match

Spreadsheet Simulation

Step 1: Enter =INT(12*RAND()+1) in cell B2

Step 2: Enter = INT(31*RAND()+1) in cell C2

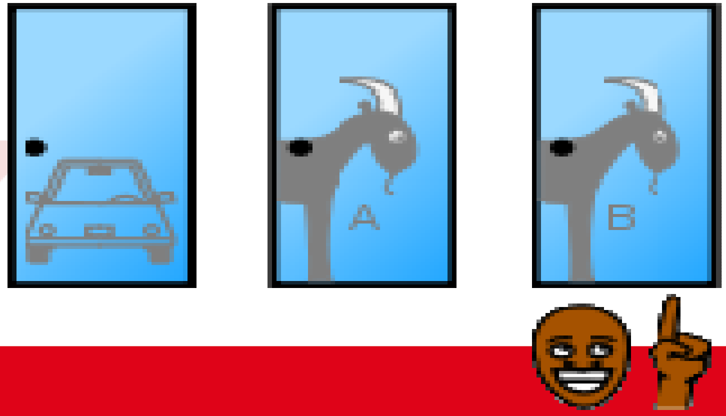
Step 3: Enter =100*B2+C2 in cell D2.

Step 4: Drag B2, C2 and D2 till cells B101,C101,D101.

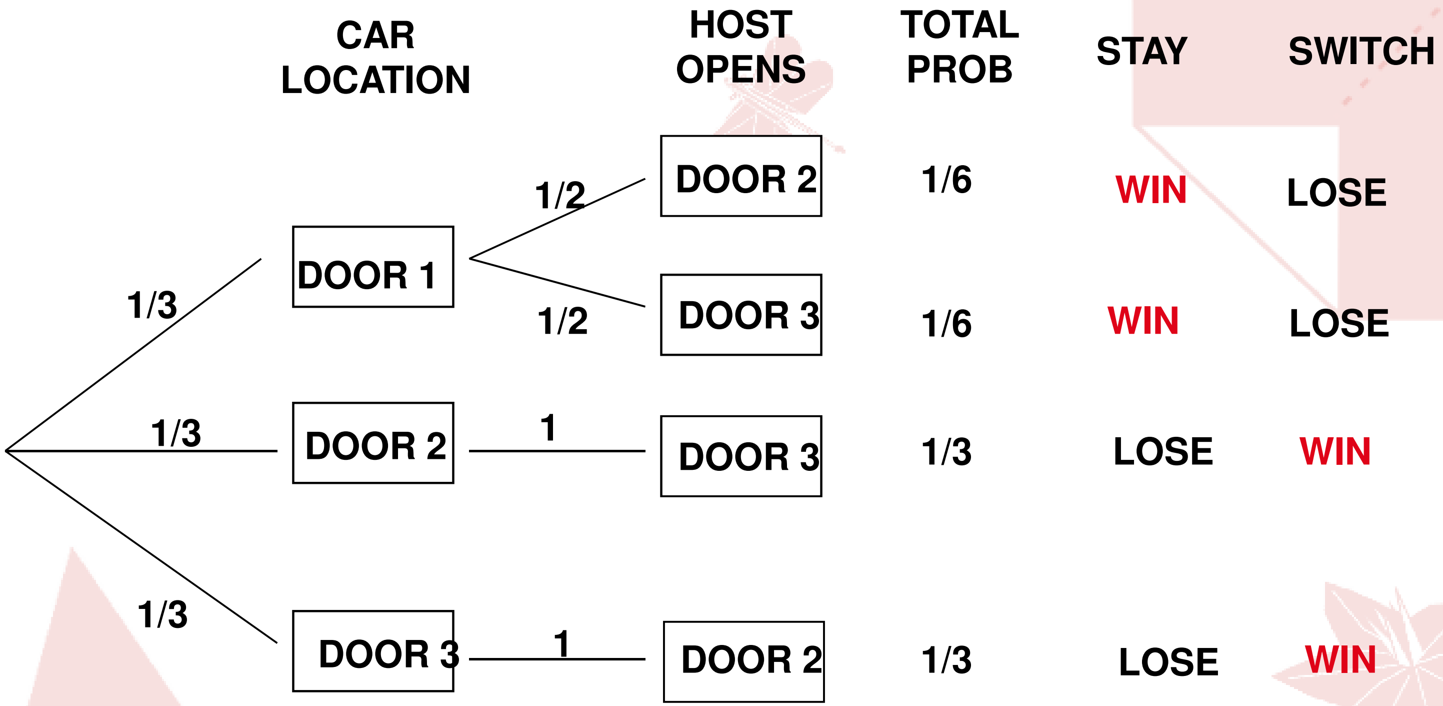
Step 5: Copy the contents of column D into any other column and sort.

The Monty Hall Problem

Suppose you are on a game show and you are given the choice of three doors. Behind one door is a car and behind the other two are goats. You pick a door (say no. 1) and Monty (the host), who knows what's behind the doors, opens one of the other two doors (say no. 3) which reveals a goat. He then asks you if you would like to stick to your original choice (no. 1) or switch (to no. 2). Is it better to stick to your original choice or to switch?



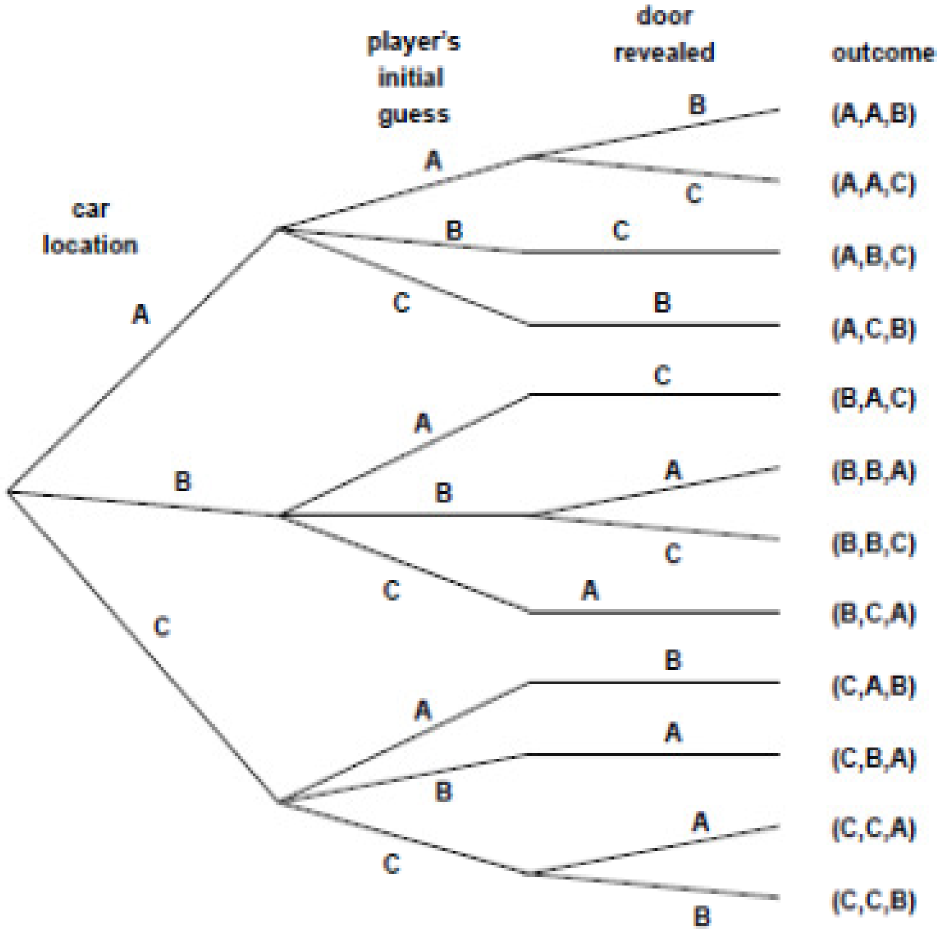
The Monty Hall Problem – Tree Diagram



	CAR LOCATION		HOST OPENS	TOTAL PROB	STAY	SWITCH
1/3	DOOR 1	1/2	DOOR 2	1/6	WIN	LOSE
		1/2	DOOR 3	1/6	WIN	LOSE
1/3	DOOR 2	1	DOOR 3	1/3	LOSE	WIN
1/3	DOOR 3	1	DOOR 2	1/3	LOSE	WIN

PLAYER CHOOSES DOOR 1

The Monty Hall Problem – sample space



The Monty Hall Problem – Bayesian Analysis

C_i: car is behind door i, i = 1,2,3 $P(C_i) = 1/3$

M_{ij}: Monty opens door j when player chooses door i

$$P(M_{ij}/C_k) = \begin{cases} 0, \mathit{if}\ i = j \\ 0,\ \mathit{if}\ \ j = k \\ 1/2,\ \mathit{if}\ \ i = k \\ 1,\ \mathit{if}\ \ i \neq k, j \neq k \end{cases}$$

$$\begin{aligned} P(M_{13}) &= P(M_{13} \cap C_1) + P(M_{13} \cap C_2) + P(M_{13} \cap C_3) \\ &= P(M_{13}/C_1)P(C_1) + P(M_{13}/C_2)P(C_2) + P(M_{13}/C_3)P(C_3) \\ &= \frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3} = \frac{1}{2} \end{aligned}$$

The Monty Hall Problem – Bayesian Analysis

$$P(C_1 / M_{13}) = \frac{P(M_{13} \cap C_1)}{P(M_{13})} = \frac{P(M_{13} / C_1)P(C_1)}{P(M_{13})} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(C_3 / M_{13}) = \frac{P(M_{13} \cap C_3)}{P(M_{13})} = \frac{P(M_{13} / C_3)P(C_3)}{P(M_{13})} = \frac{0 \times \frac{1}{3}}{\frac{1}{2}} = 0$$

$$P(C_2 / M_{13}) = 1 - \{P(C_3 / M_{13}) + P(C_3 / M_{13})\} = \frac{2}{3}$$

Spreadsheet Simulation

Step 1: Enter =INT(3*RAND()+1) in A2 and generate 100 entries till A101.

Step 2: Enter 1 in B2 and copy to all cells till B101.

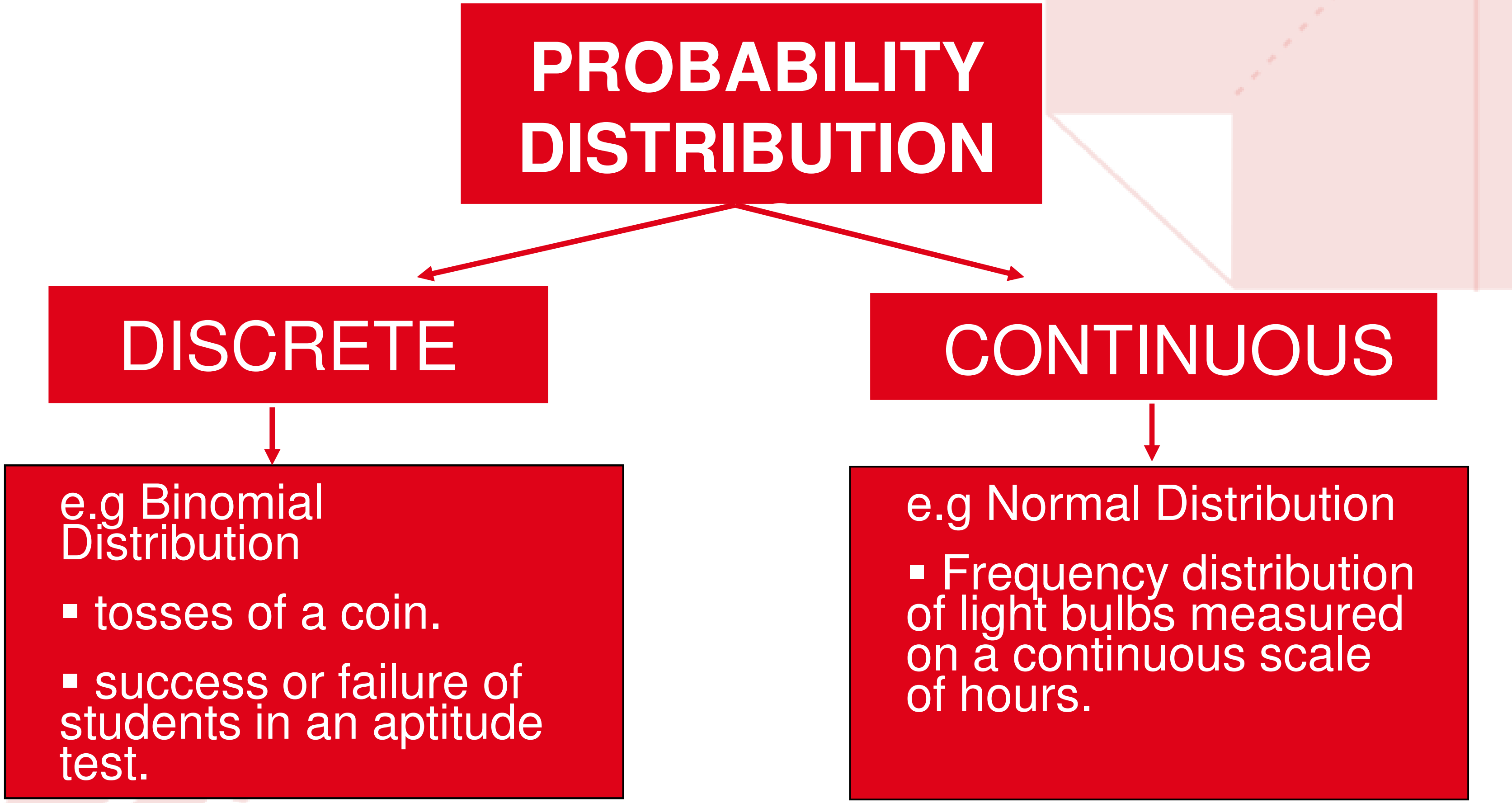
Step 3: Enter =IF(A2=1,IF(RAND()<0.5,"2","3"),IF(A2=2,"3","2")) in cell C2.

Step 4: Enter =IF(A2<>B2,"Yes","No") in cell D2.

Step 5: Enter =IF(A2=B2,"Yes","No") in cell E2.

Step 6: Enter =COUNTIF(D2:D101,"YES")/100 in cell H2 and =COUNTIF(E2:E101,"YES")/100 in cell I2

PROBABILIY DISTRIBUTIONS
A Graphical Exploration



PROBABILITY DISTRIBUTION

DISCRETE

e.g Binomial Distribution

- tosses of a coin.
- success or failure of students in an aptitude test.

CONTINUOUS

e.g Normal Distribution

- Frequency distribution of light bulbs measured on a continuous scale of hours.

Binomial Distribution

Characteristics of Bernoulli Process

1. Each trial has two possible outcomes: success or failure.
2. The probability of outcome is fixed over time.
3. Trials are statistically independent: outcome of one trial does not affect or depend on another.
4. Examples: In the random experiment of the throws of a dice ‘getting 6’ is a success and ‘not getting 6’ is a failure.

Binomial Formula

$$P(X=r) = {}^{n}C_{r}\, p^{r} q^{n-r} = \frac{n!}{r!(n-r)!} p^{r} q^{n-r}$$

p = probability of success,
q = 1- p = probability of failure,
r = number of successes,
n = total number of trials

Binomial Formula – Exercise

The incharge of the electronics section of a large departmental store has observed that the probability that a customer who is just browsing will buy something is 0.3. Suppose that 15 customers browse in the electronics section each hour, what is the probability that

(a) exactly 4 browsing customers will buy something in the specified hour?

(b) at least one browsing customer will buy something in the specified hour?

Binomial Formula – Exercise (solution)

$$(a)\quad P(X=4) = {}^{15}C_4(0.3)^4(0.7)^{11} = 0.2183$$

$$\begin{aligned}(b)\quad P(X\geq 1) &= 1 - P(X=0)\\ &= 1 - {}^{15}C_0(0.3)^0(0.7)^{15}\\ &= 0.9952\end{aligned}$$

(a) BINOMDIST(4,15,0.3,FALSE)

(b) 1 – BIMONDIST(1,15,0.3,FALSE)

Binomial Formula – self check

Which do you think is more likely: getting exactly 5 heads in 10 throws of a coin or getting exactly 50 heads in 100 throws of a coin?

BINOMDIST(5,10,0.5,FALSE)

BINOMDIST(50,100,0.5,FALSE)

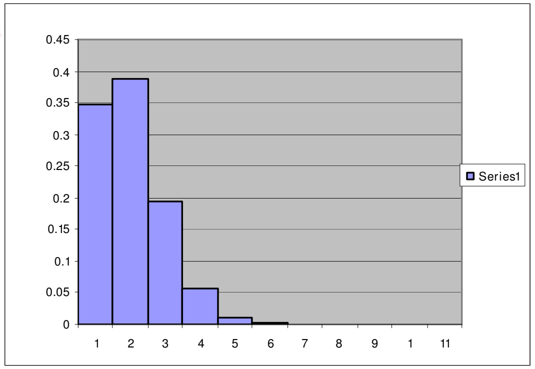
Binomial Distribution – A Graphical Exploration

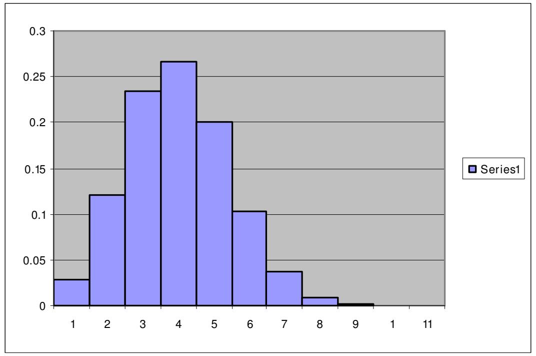
Draw probability histograms of the binomial distribution for

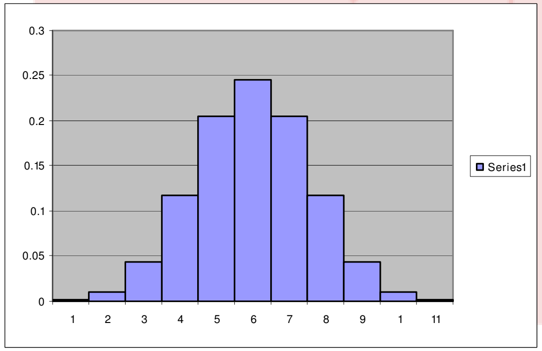
(a) $n = 10$ and $p = 0.1, 0.3, 0.5, 0.7, 0.9$

(b) $p = 0.4$ and $n = 5, 10, 30$

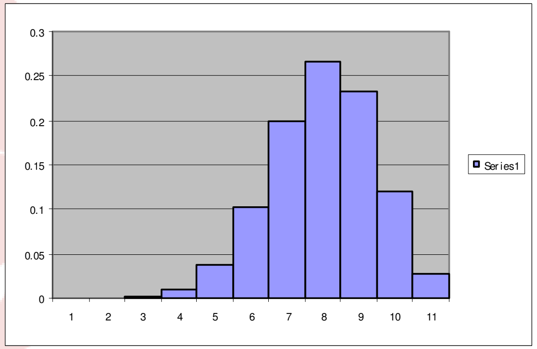
Binomial Distribution – A Graphical Exploration

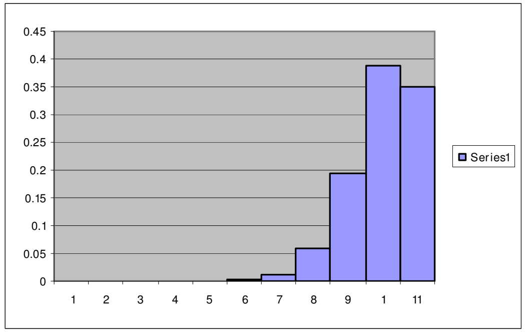




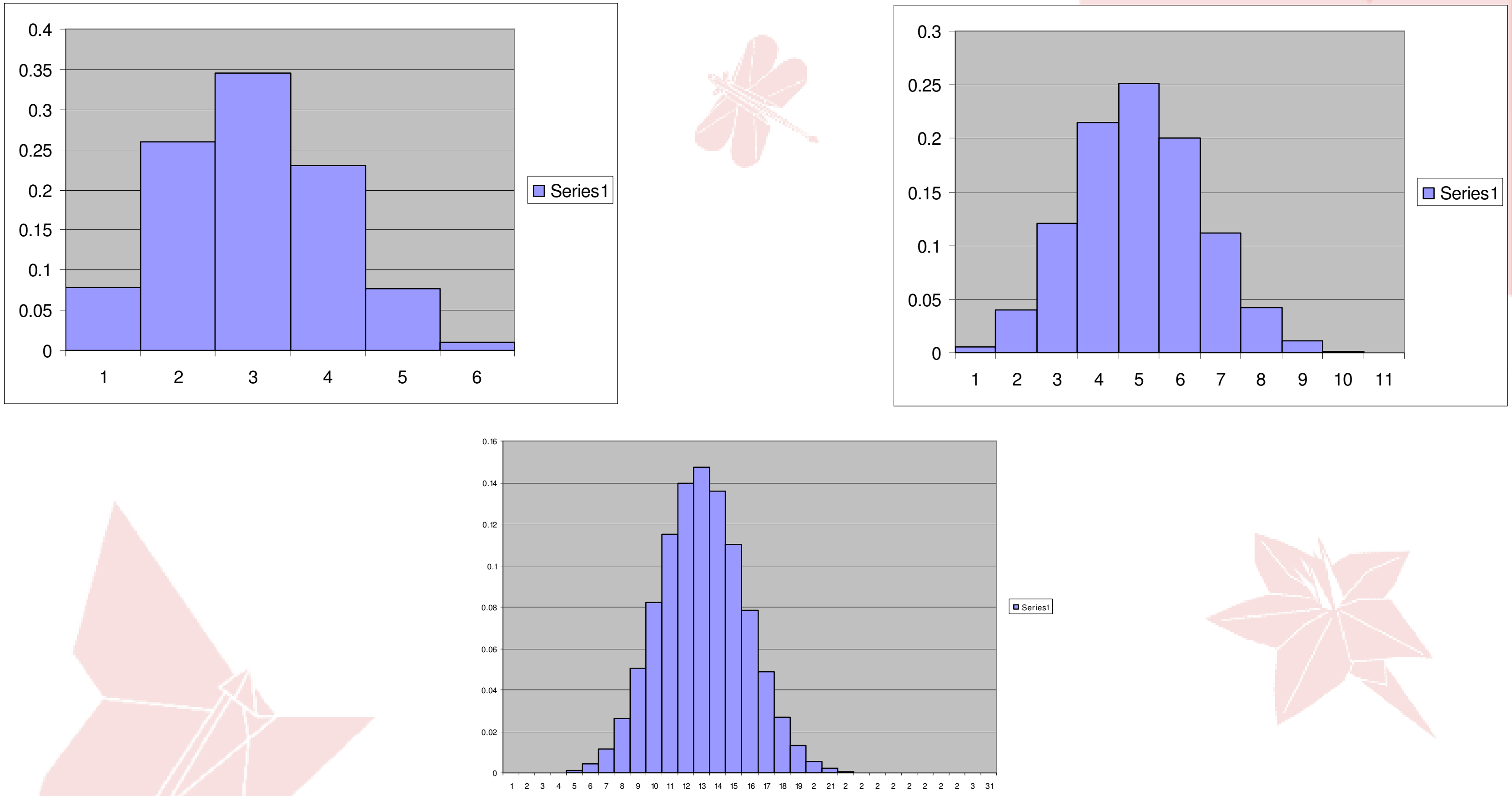


The binomial probability histogram for n = 10 and p = 0.1,0.3,0.5,0.7,0.9





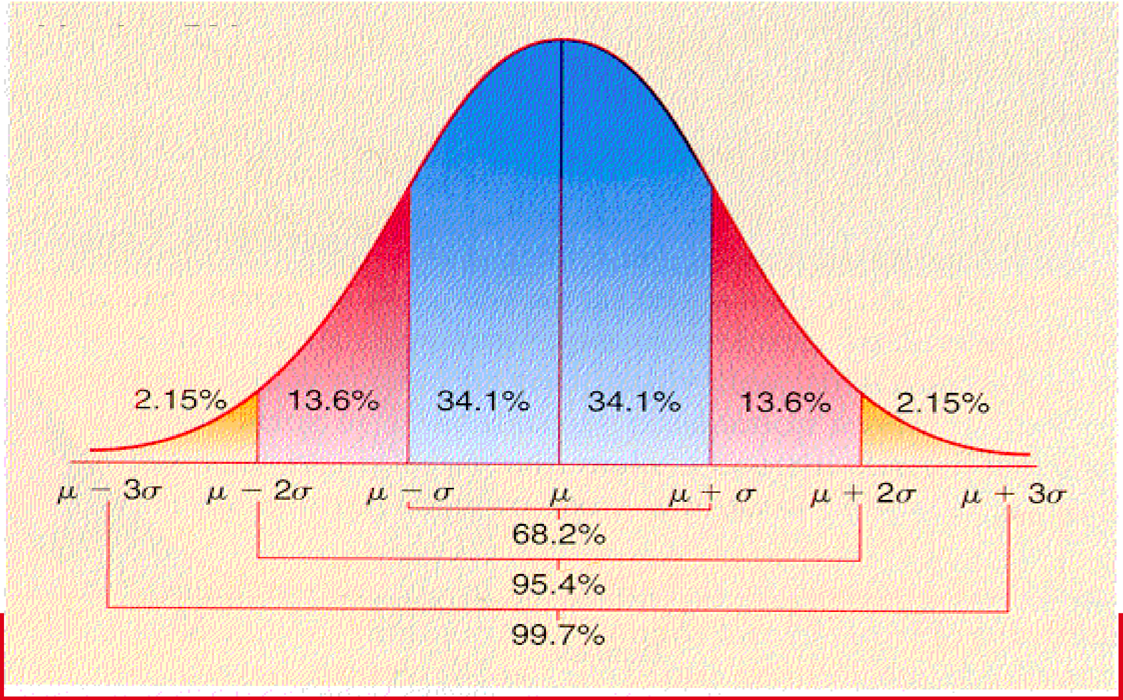
Binomial Distribution – A Graphical Exploration



The binomial probability histogram for p = 0.4 and n = 5,10,30

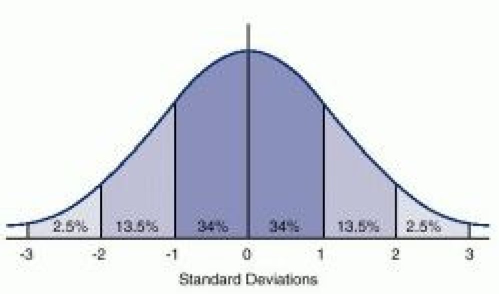
Normal Distribution

1. Approx 68% of all values in a normally distributed population lie within μ ± σ.
2. Approx 95.4% of all values in a normally distributed population lie within μ ± 2σ.
3. Approx 99.7% of all values in a normally distributed population lie within



Standard Normal Probability Distribution

1. The standard normal dist. helps us to compare two distributions.
2. μ = 0, σ = 1
3. $Z = \frac{X-\mu}{\sigma}$ (Standardizing a normal variable)



4. The standard normal prob. dist. Table shows the area under the normal curve **between the mean and positive values of z.**
5. Normally distributed random variables take different units of measure: dollars, inches etc. z denotes standard units (i.e standard deviations from x to the mean)

Standard Normal Probability Distribution Example

The life of electronic tubes of a certain type are assumed to be normally distributed with mean 155 hours and standard deviation of 19 hours. What is the probability that

(i) The life of a randomly chosen tube is less than 117 hours.

(ii) The life of a randomly chosen tube is between 136 and 174 hours.

Standard Normal Probability Distribution Example (Solution)

μ = 155 hours, σ = 19 hours

(i) $P(X < 117) = P(Z < -2) = P(Z > 2) = 0.5 - P(0 < Z < 2)$

$$= 0.5 - 0.4772 = 0.0228$$

(NORMSDIST(-2))

(ii) $P(136 < X < 174) = P(-1 < Z < 1) = 2P(0 < Z < 1)$

$$= 2 \times 0.3413 = 0.6826$$

(NORMSDIST(1)-NORMSDIST(-1))

Linear Regression

The Regression line

Method of Least Squares/Line of Best Fit Y = a + bX

- The equation of the regression line is, $\hat{Y} = a + bX$ where $\hat{Y}$ is the estimated value of Y for a given value of X.
- If (X_i,Y_i) are the data points, then $Y_i - \hat{Y}$ represents the error which may be positive or negative.
- We compute the sum of $(Y_i - \hat{Y})^2$. The line having the least sum (i.e least error) is the line of best fit or the regression line.
- $b = \dfrac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$, $a = \bar{Y} - b\bar{X}$ b is the slope and a is the y-intercept of regression line.
- These are obtained by solving the normal equations

$$\sum y = na + b\sum x \quad and \quad \sum xy = a\sum x + b\sum x^2$$

The Regression line – Standard Error of Estimate

The standard error of estimate

- Helps us to check if the regression line is a good fit to the data i.e it helps us to check the reliability of the equation.
- Measures the variability, or scatter, of the observed values around the regression line.
- The standard error can be calculated using $$S_e = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$
- Coefficient of determination tells us how much of the variation is explained by the regression line

$$r^2 = \frac{a\sum Y + b\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}$$

Interpreting the Standard Error of Estimate

The standard error of estimate

- Is similar to standard deviation.
- measured along the y – axis.
- Assuming that the observed points are normally distributed around the regression line, we can expect to find 68% of the points within ±1S_e, 95.5% of the points within ±2S_e and 99.7% of the points within ±3S_e.

Using the Standard Error of Estimate for finding Prediction Intervals

If n > 30, then

- $\hat{Y} \pm 1S_e$ gives a 68% confidence interval of the predicted value i.e we can be 68% sure that the predicted value will be within this interval
- $\hat{Y} \pm 2S_e$ gives a 95.5%confidence interval of the predicted value.
- $\hat{Y} \pm 3S_e$ gives a 99.7%confidence interval of the predicted value.

If n < 30, we need to find the appropriate t value and $\hat{Y} \pm tS_e$ will give the required prediction interval.

Regression Analysis – Example

Sales of major appliances vary with the new housing market: when new home sales are good, so are the sales of dishwashers, washing machines refrigerators etc. A trade association compiled the following historical data (in thousands of units) on major appliance sales and housing starts

Housing starts (in thousands)	2.0	2.5	3.2	3.6	3.3	4.0	4.2	4.6	4.8	5.0
Appliance sales (in thousands)	5.0	5.5	6.0	7.0	7.2	7.7	8.4	9.0	9.7	10.0

(a) Develop a equation for the estimating line for the relationship between appliance sales and housing starts.

(b) Compute and interpret the standard error of estimate.

Regression Analysis – Exercise (solution)

$\hat{Y} = 1.1681 + 1.715X$ where

$\hat{Y}$ denotes the estimated value of appliance sales and X denotes the housing starts.

S_e = 0.3737 i.e the standard deviation of the data points around the regression line is about 374 units.

Hill Ciphers

- Consider the substitution table

A	B	C	D	E	F	G	H	I	J	K	L	M
0	1	2	3	4	5	6	7	8	9	10	11	12
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
13	14	15	16	17	18	19	20	21	22	23	24	25

- . is 26, -- is 27, ? is 28.
- The encoding matrix is $E = \begin{bmatrix} 1 & 4 \\ 2 & 9 \end{bmatrix}$

Hill Cipher : The Encoding Process

- Convert the plaintext message say **MATH _ IS _ FUN.** to its substitution values 12 0 19 7 27 8 18 27 5 20 13 26

- Compute the product EM

$$\begin{bmatrix} 1 & 4 \\ 2 & 9 \end{bmatrix}\begin{bmatrix} 12 & 19 & 27 & 18 & 5 & 13 \\ 0 & 7 & 8 & 27 & 20 & 26 \end{bmatrix} = \begin{bmatrix} 12 & 47 & 39 & 126 & 85 & 117 \\ 24 & 101 & 126 & 279 & 190 & 260 \end{bmatrix}$$

- Reduce the product modulo 29 to obtain the Hill – 2 – cipher values.

$$\begin{bmatrix} 12 & 47 & 39 & 126 & 85 & 117 \\ 24 & 101 & 126 & 279 & 190 & 260 \end{bmatrix} = \begin{bmatrix} 12 & 18 & 1 & 10 & 27 & 1 \\ 24 & 14 & 10 & 18 & 16 & 28 \end{bmatrix} \pmod{29}$$

Hill Cipher: The Decoding Process

- Convert the ciphertext message say MYSOBKS _ QB to its Hill-2-cipher values 12 24 18 14 1 10 10 18 27 16 1 28
- Compute the product E^{-1} M

$$\begin{bmatrix} 9 & -4 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} 12 & 18 & 1 & 10 & 27 & 1 \\ 24 & 14 & 10 & 18 & 16 & 28 \end{bmatrix} = \begin{bmatrix} 12 & 106 & -31 & 18 & 179 & -103 \\ 0 & -22 & 8 & -2 & -38 & 26 \end{bmatrix}$$

- Reduce the product modulo 29 to obtain the substitution values.

$$\begin{bmatrix} 12 & 106 & -31 & 18 & 179 & -103 \\ 0 & -22 & 8 & -2 & -38 & 26 \end{bmatrix} = \begin{bmatrix} 12 & 19 & 27 & 18 & 5 & 13 \\ 0 & 7 & 8 & 27 & 20 & 26 \end{bmatrix} (\mathrm{mod}\, 29)$$

To multiply matrices use
=MMULT and within brackets select array 1, array2 and click on
CRTL+SHIFT followed by ENTER

Linear Programming

Linear programming was invented in the 1940s. ‘Programming’ refers to the process of choosing the best plan from a set of possible alternatives in a decision problem.

Three elements are involved in a LPP

- Decision variables
- Objective function
- Constraints

Insulation Production – Linear Programming

An insulation plant makes two types of insulation called type B and type R. Both types of insulation are produced using the same machine. The machine can produce any mix of output, as long as the **total weight** is no more than 70 tons per day. Insulation leaves the plant in trucks; the loading facilities can handle up to **30 trucks per day**. One truckload of type B insulation weighs 1.4 tons; one truckload of type R weighs 2.8 tons. Each truck can carry type B insulation, Type R insulation, or any mixture thereof. The insulation contains a flame retarding agent which is presently in short supply; the plant can obtain **at most 65 canisters** of the agent per day. One truckload of (finished) type B insulation requires an input of three canisters of the agent while one truckload of type R requires only one canister.

The plant manager has calculated that, at current prices, **the contribution from each truckload of type B is $950 and $1200 for type R**. There appears to be no difficulty in selling the entire output of the plant, no matter what production mix is selected.

How much of each type of insulation should be produced?

Insulation Production – Formulating the LPP

Suppose x truckloads of type B and y truckloads of type R are produced

Max 950 x + 1200 y (Objective function)

Subject to the constraints

$$1.4x + 2.8y \leq 70 \ (\textit{total wt. in tons})$$

$$x + y \leq 30 \ (\textit{no. of trucks})$$

$$3x + y \leq 65 \ (\textit{canisters of agent})$$

$$x, y \geq 0$$

Insulation Production – Linear Programming

Interpreting the Linear Programming Output

- The maximum possible contribution is $33,500 per day and is obtained by producing 10 truckloads of type B and 20 truckloads of type R.
- Slack – the amount of resource that remains unused. E.g the machine and truck capacity are being fully used whereas only 50 out of the 65 cannisters of flame retarding agent are being used.
- Shadow Price – rate at which the optimal objective value would change in response to small alterations in the available amount of the resource corresponding to that constraint. The machine constraint has a shadow price of $ 178.57. This means that if we could increase the capacity of the machine by one unit (one ton per day) we could rearrange the production schedules to increase the daily contribution by $178.57.
- The lower (upper) ranges denote the maximum decrease (increase) in the amount of resource for which the shadow price is valid.