QM Notes Sajin J

Quantitative Methods (Birla Institute of Technology and Science, Pilani)

QUANTITATIVE METHODS

MBAZC417

SAJIN JOHN
2020HB58042

Table of Contents

Data

QM – Application and Techniques

Business applications
- Casino game design, KBC
- Insurance
- Email spam filters
- Opinion/Exit polls
- Marketing research
- Warranty policies
- Emergency services
- Quality control
- Portfolio management
- Options/Futures/Derivatives
- Risk management
- Bundling- Data mining

Other fields
- Clinical trials
- Fertilizers- Design of Experiments
- Court judgements
- Meteorology
- Dam design and reservoir operation
- Statistical mechanics
- Genetics- Mendel's laws
- Quantum theory
- Econometrics
- Radar- Aircraft detection
- Image/Signal processing
- Theory building and testing

Techniques
- Classification
- Clustering
- Association
- Regression
- Forecasting
- Decision tree analysis
- Discriminant analysis
- Singular Value Decomposition (SVD)
- Principal Component Analysis (PCA)
- Factor analysis
- Markov process
- Monte Carlo simulation

Data in Business

- *Marketing*
  Sales and usage (Point of Sale data), Order booking, Credit Sales, Sales Persons, Dealers, Returns, Customers Satisfaction, Loyalty Schemes, Marketing Research…
- *Production and Operations*
  Production, Machine Utilization, Inventory, Productivity, Quality,…
- *Purchase*
  Prices, Delivery, Suppliers, Prices, Inventory, Supplier Rating, Returns, …
- *HR*
  Employee records: Employee satisfaction, Performance, Manpower Planning, …
- *Finance*
  Invoices, Receivables, Assets, Budgeting, Interest Rates, Stock Market Index, …
- *Maintenance*
  Breakdown, MTTR, MTFF, …

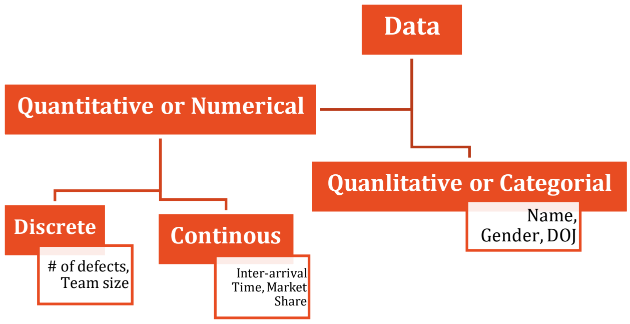
Data Analysis Stages

Statistics
- Descriptive
  - Summarise, Presentation of Data
- Inferential
  - Estimation, Hypothesis Testing, Regression, Correlation

Data Organization:
Raw Data ➔ Organized into tables ➔ Pictorial Representation

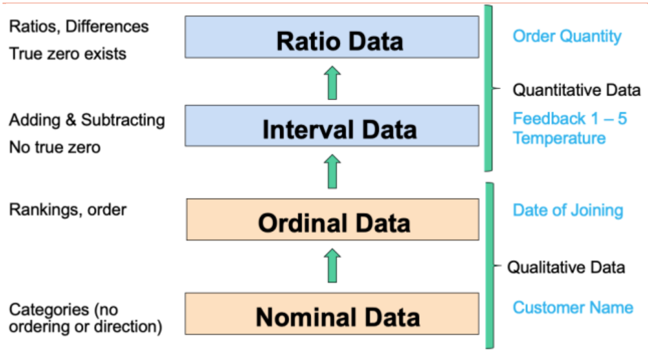
Types of Data

Statistics begins with data. Data can be Qualitative or Quantitative.



Data
- ***Qualitative or Categorical Data***
  - Could be a Nominal Data or Ordinal Data.
  - Arithmetic operation do not make sense.
  - Name, Gender, Car Colour, Date of Joining
- ***Quantitative or Numerical Data***
  - ***Discrete***
    - # of defects, # of Children
  - ***Continuous***
    - Inter-arrival Time, Market Share, PE Ratio

Measurement Scale



Let's consider a HR database. Variables of interest could be Employee Name, Date of Joining, Appraisal rating & Annual CTC

Nominal Data
- It's a Qualitative data (like employee name)
- Data will be either one of the option.
- Data can be categorized and counted cannot be measured or ranked.
- Bar Charts & Pie Charts
- E.g.:
- *Engineer? – Yes/No*
- *Gender – Male/Female*
- *Material –*
- *Wood/Metal/Plastic/Steel/Bronze*

- *Industry – Oil/Mining/Automobile/Media/IT*
- *TV Channels – Music/News/Kids/Others*
- *Marital Status – Married/Single/Unmarried*
- *IPL Teams – KKR/CSK/MI/SRH/...*

Ordinal/Ranked Data

- It's a Qualitative data which can be sorted
- Date of Joining
- Measuring instrument may not be required to rank the data.
- Frequency & Cumulative frequency.
- E.g.:
- *Tall, taller, tallest*
- *Big, bigger, biggest*
- *Olympics: First, second, third, fourth*
- *Thickness: very thick, thick, thin.*
- *Taste: Good, average, below average, bad*
- *Temperature: freezing, cool, warm, hot.*

Interval Data

- Data is represented in a specific interval
- There won't be any true zero
- Numerical data. Where zero is arbitrary chosen.
- Histogram, Cumulative frequency & Pie Chart
- i.e. *Zero degree Centigrade/Fahrenheit is not zero temperature*
  Zero customer satisfaction is arbitrary.
- E.g.:
- *Appraisal rating (A-D ➔ 1,2,3,4 or A-D ➔ -2,-1,1,2)*
- *Customer satisfaction measured from 1 to 5*
- *Men's Shoe Size (zero is arbitrary)*
- *Garment Size*

Ratio Data

- A ratio scale allows all arithmetic operation
- Salary of a person is double the other person.
- True zero exists
- Salary 0 has a meaning
- E.g.:
- *Height in mm/cm/m*
- *Weight in g/kg/tons*
- *Time in sec/min/hr*
- *Temperature in Kevin (0 K exists)*
- *Humidity in %*
- *Sales (numbers, tons, or Rs)*
- *No. of patients*
- *Average time to serve, minutes*
- *Profit, Cost (Rs)*

Cross Sectional Vs Time Series Data

Cross Sectional Data

- A snapshot of the system
- Closing Values of Indices on November 17, 2016

Time Series Data

- A variable of Interest is measured periodically.
- In this case we may also depict the time series by a trend line where the X Axis is Time and the Y Axis is the value of the variable

Closing Values on November 17, 2016	
Index Name	Index Value
Nifty 50	8079.95
Nifty Next 50	21231.55
Nifty 100	8281.95
Nifty 200	4306.75
Nifty 500	6903.45
Nifty Midcap 50	3641.4
Nifty Free Float Midcap 100	14290.15
Nifty Free Float Smallcap 100	5611.55
Nifty50 Dividend Points	90.12
Nifty Auto	8915.6
Nifty Bank	19087.85
Nifty Energy	9678.4
Nifty Financial Services	7642.3
Nifty FMCG	20154
Nifty IT	9496.2
Nifty Media	2641.9
Nifty Metal	2662.95

NIFTY Closing Price	
Date	Close
01-Jan-16	7963.2
04-Jan-16	7791.3
05-Jan-16	7784.65
06-Jan-16	7741
07-Jan-16	7568.3
08-Jan-16	7601.35
11-Jan-16	7563.85
12-Jan-16	7510.3
13-Jan-16	7562.4
14-Jan-16	7536.8
15-Jan-16	7437.8
18-Jan-16	7351
19-Jan-16	7435.1
20-Jan-16	7309.3
21-Jan-16	7276.8
22-Jan-16	7422.45
25-Jan-16	7436.15

Displaying Qualitative Data

Qualitative Data also known as Categorical Data as these are grouped by *specific categories*.
Such date could be Nominal or Ordinal scale.
Examples for Qualitative/Categorial Data are:

- Location of customers
- Age groups: 0-5, 6-10, …
- Dates when orders where shipped

Qualitative Data may be summarized by:

- Frequency Distribution
- Relative Frequency Distribution
- Cumulative Frequency Distribution
- Percent Frequency Distribution
- Bar Charts / Graph
- Pie Charts
- Cross Tabulation summarizes the data of two variables

Application (examples) of Qualitative data

- An e-commerce firm would like to know the distribution of its customers to optimally locate its distribution centres.
- Patterns in shipping dates may decide staffing patters.
- An ice-cream company may be interested in knowing whether "Flavour Preference" depends on "Age Group". This would decide what should be stocked at a stall near a school.

Analysing e-Commerce Customer Location

Data – Extracted from the Orders placed by the customer

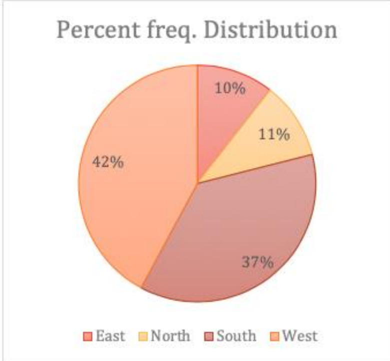
Zone	Product Category
South	Office Supplies
South	Office Supplies
West	Office Supplies
West	Furniture
North	Office Supplies
West	Office Supplies
West	Office Supplies
North	Technology
East	Office Supplies
West	Office Supplies
West	Furniture
East	Furniture
South	Office Supplies
South	Technology
South	Office Supplies
West	Office Supplies
West	Office Supplies
South	Office Supplies
South	Technology

Questions:

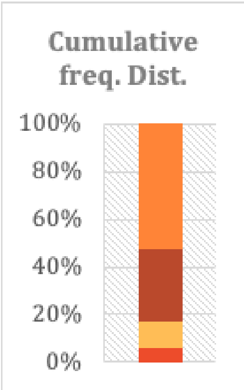
1. Where should you advertise?
2. In a restructuring exercise, where should you locate the warehouses? Where should you rent pay-per-use?

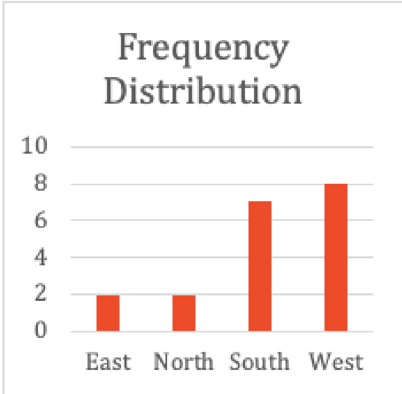
Analysis of Data:

Zone	Frequency Distribution (Count of Data)	Cumulative freq. Distribution (Summation to adjacent)	Relative freq. Distribution (ratio wrt total) ==> n/Total	Percent freq. Distribution % of total
East	2	2	0.11	11%
North	2	4	0.11	11%
South	7	11	0.37	37%
West	8	19	0.42	42%



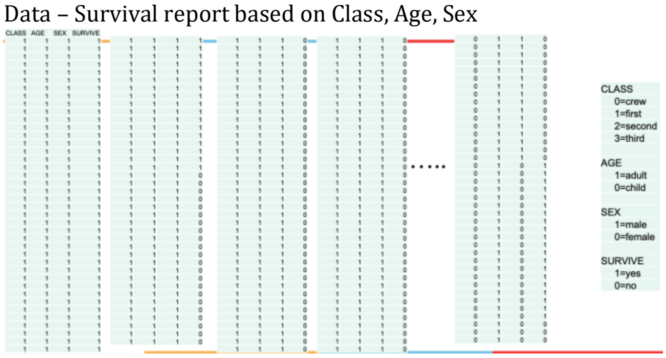
Charts:





Analysing Titanic Data

Data – Survival report based on Class, Age, Sex



Analysis of Data –

	Crew	First Class	Second Class	Third Class	Grand Total
Dead	**673**	**122**	**167**	**528**	**1490**
Female	3	4	13	106	126
Male	670	118	154	422	1364
Alive	**212**	**203**	**118**	**178**	**711**
Female	20	141	93	90	344
Male	192	62	25	88	367
Grand Total	**885**	**325**	**285**	**706**	**2201**

Cross Tabulating the data:

	First	Second	Third	Crew	Total
Alive	62%	41%	25%	24%	32%
Dead	38%	59%	75%	76%	68%
Total	**100%**	**100%**	**100%**	**100%**	**100%**

Displaying Quantitative Data

Quantitative Data is also known as Numerical Data.
These are either Interval or Ratio Scale.
Example of Quantitative or Numerical Data:

- Time taken for Level 1 Support
- Arrival times of customers
- Feedback 1 – 5

Possible questions that could be resolved/analysed using Quantitative data analysis:

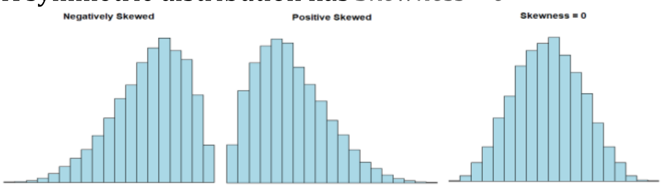
- Can additional training reduce the average service time
- Can we identify peak hours for optimal staffing?
- Which unit has the lowest employee satisfaction?

Quantitative Data could be summarized by using:

- Frequency Distribution
- Relative Frequency Distribution
- Percentage Relative Frequency Distribution
- Cumulative Frequency Distribution
- Histogram
- Ogive
- Dot Plot

Skewness

Skewness measures the asymmetric nature of the distribution.
A distribution with the peak towards the right and a longer left tail is *skewed left* or *negatively skewed*.
A distribution with the peak towards the left and a longer right tail is *skewed right* or *positively skewed*.
A symmetric distribution has *Skewness = 0*



Analysing Kentucky Derby Finishing Times

Data

:

Year	Time
1875	158
1876	158
1877	158
1878	157
1879	147
1880	148
1881	160
1882	160
1883	163
1884	160
1885	157
1886	157
1887	159
1888	158
1889	155
1890	165
1891	172
1892	162
1893	159
1894	161
1895	158
1896	128
1897	133
1898	129
1899	132
1900	126
1901	128
1902	129
1903	129
1904	129
1905	131
1906	129
1907	133
1908	135
1909	128
1910	126
1911	125
1912	129
1913	125
1914	123
1915	125
1916	124
1917	125
1918	131
1919	130
1920	129
1921	124
1922	125
1923	125
1924	125
1925	128
1926	124
1927	126
1928	130
1929	131
1930	128
1931	122
1932	125
1933	127
1934	124
1935	125
1936	124
1937	123
1938	125
1939	123
1940	125
1941	121
1942	124
1943	124
1944	124
1945	127
1946	127
1947	127
1948	125
1949	124
1950	122
1951	123
1952	122
1953	122
1954	123
1955	122
1956	123
1957	122
1958	125
1959	122
1960	122
1961	124
1962	120
1963	122
1964	120
1965	121
1966	122
1967	121
1968	122
1969	122
1970	123
1971	123
1972	122
1973	119
1974	124
1975	122
1976	122
1977	122
1978	121
1979	122
1980	122
1981	122
1982	122
1983	122
1984	122
1985	120
1986	123
1987	123
1988	122
1989	125
1990	122
1991	123
1992	123
1993	122
1994	124
1995	121
1996	121
1997	122
1998	124
1999	123
2000	121
2001	120
2002	121
2003	121
2004	124
2005	123
2006	121
2007	122
2008	122
2009	123
2010	124

Kentucky Derby FinishingTimes

Time Taken – Frequency Distribution (FOUR BINS)

FOUR BINS

Class	Freq	Cumul	Relative Frequency
[115,130)	106	106	0.779
[130,145)	9	115	0.066
[145,160)	13	128	0.096
[160,175)	8	136	0.059

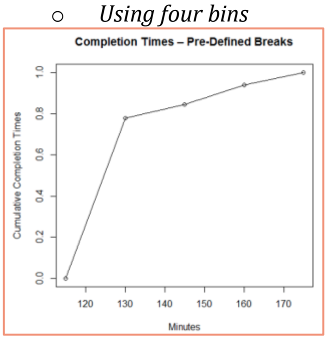
Time Taken – Frequency Distribution (NINE BINS)

NINE BINS

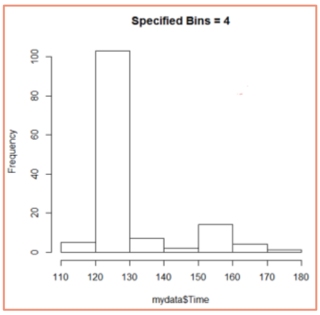
Class	Freq	Cumul	Relative Frequency
[119,125)	73	73	0.5368
[125,131)	35	108	0.2574
[131,137)	7	115	0.0515
[137,143)	0	115	0.0000
[143,148)	2	117	0.0147
[148,154)	0	117	0.0000
[154,160)	14	131	0.1029
[160,166)	4	135	0.0294
[166,172)	1	136	0.0074

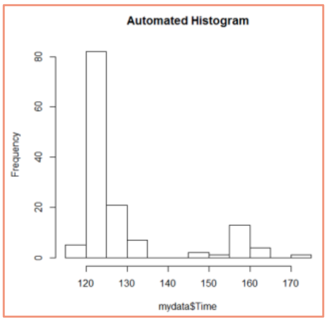
*Time Taken – **The Ogive***

- *Using four bins*

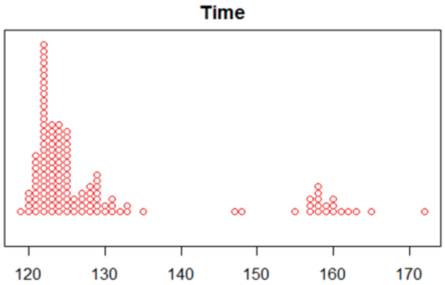


*Time Taken – **Histograms***





*Time Taken – **Dot Plot***



*Time Taken – **Scatter Plot***

Stem-and-Leaf diagram

Stem-and-Leaf diagram is like a histogram without losing the data.

153
154
154
162
165
169
172
176
176
176
177
180
182
186
187
190

Range	Frequency
150-160	3
160-170	3
170-180	5
180-190	4
190-200	1
Total	**16**

Stem	Leaf
15	3 4 4
16	2 5 9
17	2 6 6 6 7
18	0 2 6 7
19	0

Summary Tables

1. ***Frequency Table***
   One nominal variable – Blood Group

Blood group	Tally marks	Number of students (Frequency)
A	卌 卌 II	12
B	卌 III	8
AB	IIII	4
O	卌 I	6
Total		30

2. ***Contingency Table – 2 variables***
   Two nominal variables – Blood Group & Ethnicity

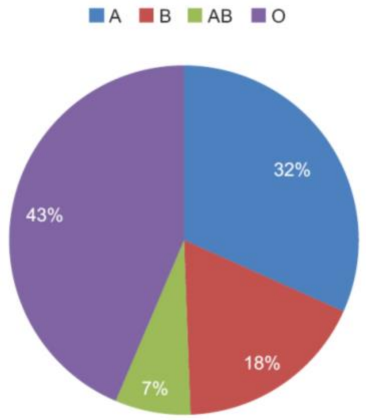
	Caucasians	African American	Hispanic	Asian
O +	37%	47%	53%	39%
O -	8%	4%	4%	1%
A +	33%	24%	29%	27%
A -	7%	2%	2%	0.5%
B +	9%	18%	9%	25%
B -	2%	1%	1%	0.4%
AB +	3%	4%	2%	7%
AB -	1%	0.3%	0.2%	0.1%

3. ***Contingency Table – 3 variables***
   Three nominal variables – Blood Group, Gender & Rh

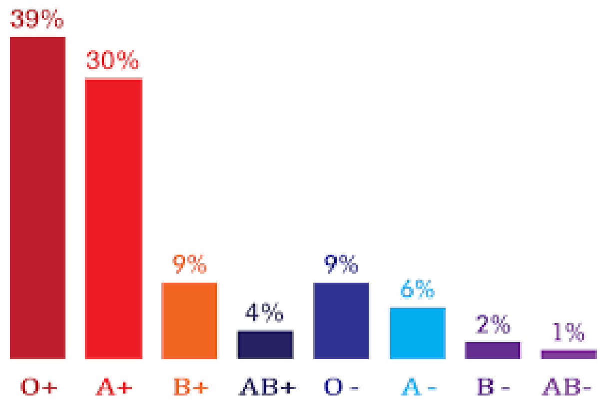
ABO blood group	Rh blood group *n* (%)				Total *n* (%)
	Rh-positive		Rh-negative		
	Male	Female	Male	Female	
A	1753 (19.53)	14 (0.16)	98 (1.09)	0 (0)	1865 (20.78)
B	1993 (22.21)	25 (0.28)	90 (1.00)	1 (0.01)	2109 (23.50)
AB	331 (3.69)	6 (0.07)	5 (0.06)	0 (0)	342 (3.81)
O	4481 (49.93)	54 (0.60)	118 (1.31)	6 (0.07)	4659 (51.91)
Total	8558 (95.35)	99 (1.10)	311 (3.47)	7 (0.08)	8975 (100)

Visual Representations – Charts

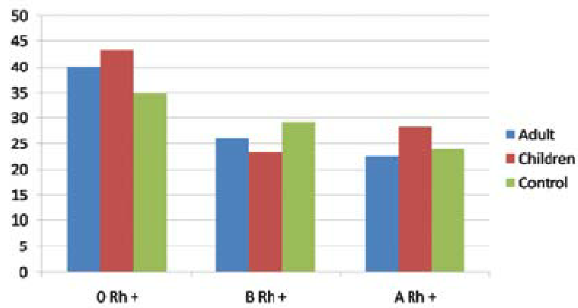
1. ***Pie Chart***



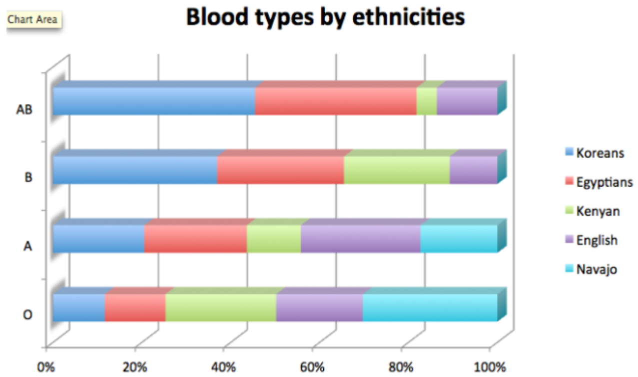
2. ***Column Chart***



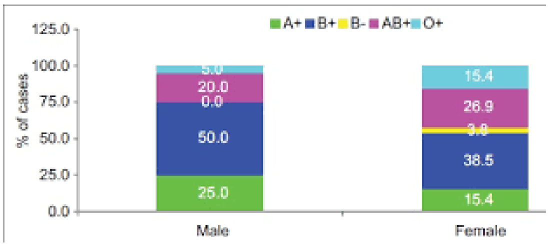
3. ***Side-by-side Chart***



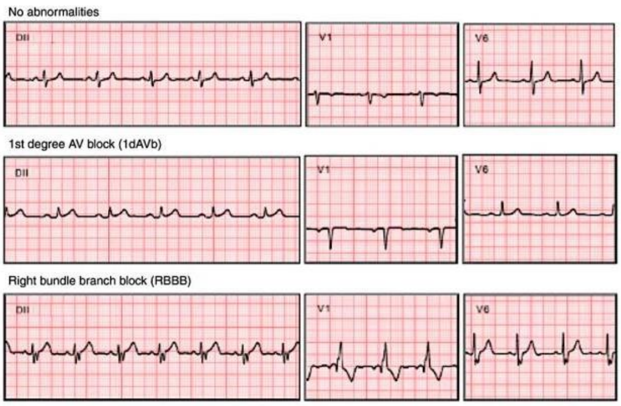
4. ***Stacked Row Chart***



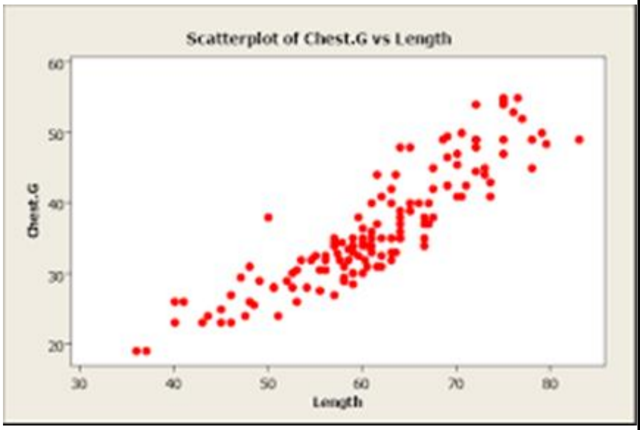
5. *Stacked Column Chart*



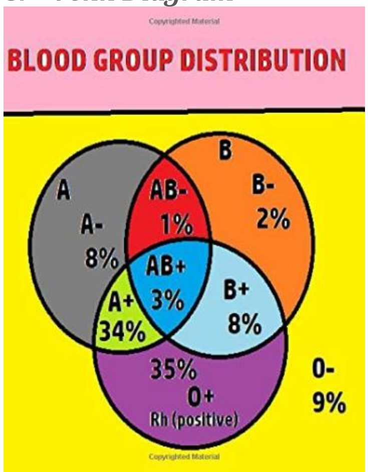
6. *Time Series Chart*



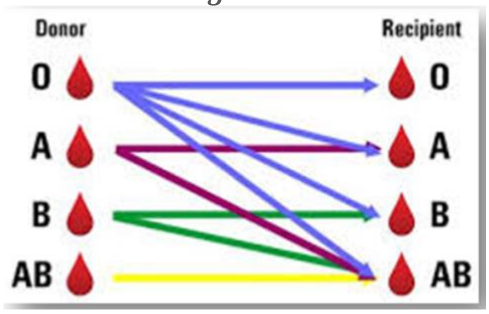
7. *Scatter Plot*



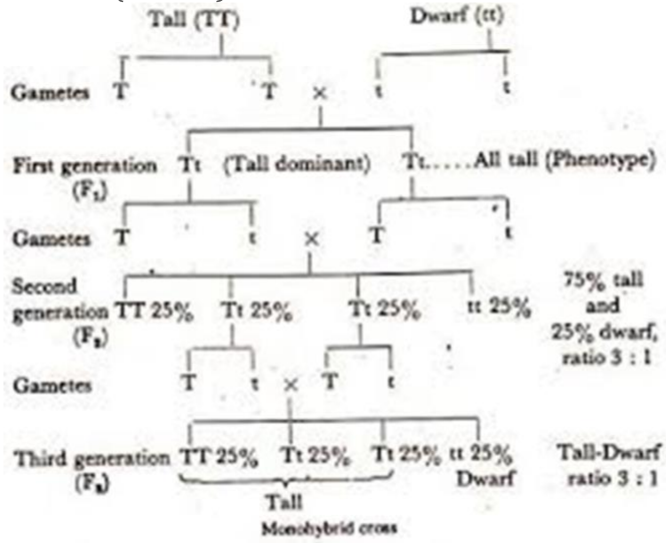
8. *Venn Diagram*



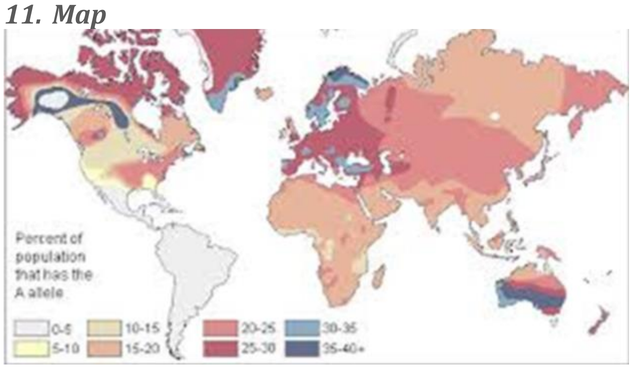
9. *Network Diagram*



10. *Tree (Mendel)*



11. *Map*



Numerical Measures

Graphical and Tabular presentations pictorially summarizes the etire data set. But business may require a measure that summarizes the data or the spread of data with a single number.

Statistics & Parameter measure

- If the measure summarizes a sample data, it is referred to as a ***Statistic***. So, statistuc is a number that represents a property of the sample.
- If the measure summarizes the population, it is referred to as a ***Parameter***. So, a parameter is a numerical characteristic of the whole population that can be estimated by a statistic.

Measure of Location

Data Sets ➔ A = {-2, -2, -1, 0, 1, 5}
Data Sets ➔ B = {-2, -2, -1, 0, 1, 5, 5}

Mean

- The sum of all data points divided by the total number of observations.
- The sample mean $\bar{X}$ is said to be a point estimator of the population mean μ.

$$\bar{X} = \frac{\sum x_i}{n} \quad , \quad \mu = \frac{\sum x_i}{N}$$

- The words “mean” and “average” are often used interchangeable.
- The technical term is “*arithmetic mean*” (AM).
- AM is always unique for a data set.
- Change in any value of the observations always affects AM.
- AM cannot be computed if all values are not available.
- *for A → 1/6 = 0.167*
  for B → 6/7 = 0.857

Median

- The median is a number that measures the “***centre***” / “***mid-point***” of the sorted data.
- Its like a “middle value”… like a median of a triangle.
- So, sort the data and find the middle value if the number of data (n/N) is odd else take the average of the two middle values (in case of even data sets)
- Visually, Median is preferred when histograms are negatively or positively skewed – i.e. histograms are far from symmetric
- Median may not be unique.
- Changes in extreme values may not affect median
  1 2 3 4 5 6 7 8 9
  1 2 3 4 5 6 7 8 10000
- Median may be computed even if values of all observations are not available.
  1 2 3 4 5 6 7 8 X
- *for A → (-1 + 0)/2 = -0.5*
  for B → 0

Mode

- The mode is the most frequent value in a data.
- There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest.
- A data set with two modes is called bimodal.
- Mode can be calculated for both quantitative or quanlitative data.
- Mode may or may not exist or can have multiple modes
  11 12 13 14
  12 12 13 14 15 15 16
- Change in the value of an observation may not affect Mode
  Blue Green Red Red Red Yellow
  Blue Green Red Red Red Blue
- Mode may be computed even if values of all observations are not available
  1 2 3 3 3 3 4 X
- *for A → -2*
  for B → -2, 5

Percentiles

- Percentiles divide ordered data into hundredths.
- Percentiles are useful for comparing values.
- p^{th} percentile:

  $$i = \left(\frac{p}{100}\right) n$$

  if, $i \neq integer$, $x_{roundup(i)}$
  else, $Avg(x_i, x_{(i+1)})$

Quartiles

- Quartiles are the numbers that separate the data into quarters.
- Quartils may or may not be part of the data.
- Median is the second quartile (Q_2)
- The first Quartile, Q_1, is the middle value of the lower half of the data. This is same as the 25^{th} percentile.
- The third Quartile, Q_3, is the middle value of the upper half of the data. This is same as the 75^{th} percentile.
- *for A →*
  *Q_1 -- 25%(n) = 0.25*6 = 1.5 ~ 2nd element.*
  $Q_1 = -2$
  *Q_2 -- 50%(n) = 0.5*6 = 3, Avg(3rd & 4th)*
  $Q_2 = Avg(-1,0) = -0.5$
  *Q_3 -- 75%(n) = 0.75*6 = 4.5 ~ 5th element.*
  $Q_3 = 1$

- *for B →*
  *Q_1 -- 0.25*7 = 1.75 ~ 2*
  Q_1 = -2
  *Q_2 -- 0.5*7 = 3.5 ~ 4*
  Q_2 = 0
  *Q_3 -- 0.75*7 = 5.25 ~ 6*
  Q_3 = 5

Percentile Rank

$$\%\,rank = \left[\frac{(\#of\ values\ below\ x) + 0.5}{Total\ \#\ of\ values}\right] \times 100$$

- for A, if x=0, # of values below x is 3
  $$\%rank = \left[\frac{(3)+0.5}{6}\right] \times 100 = \left(\frac{350}{6}\right) = 58.3\%$$
- for B, if x=0, # of values below x is 3
  $$\%rank = \left[\frac{(3)+0.5}{7}\right] \times 100 = \left(\frac{350}{7}\right) = 50\%$$

Measures of Dispersion/Variation

High Variation may mean…

- High Inconsistent
- Poor Quality
- Low Reliability
- High Uncertainty
- High Risk
- High Volatility
- High Fluctuations
- Low Predictability
- Not Steady
- Highly Uneven

Range

- Difference between the maximum and minimum value.
- Range = Max Val – Min Val
- *for A → Range = 5 - (-2) = 7*
  for B → Range = 5 - (-2) = 7

IQR

- The Inter Quartile Range is a number that indicates the spread of the middle half or the middle 50% of the data.
- It is the difference between the third quartile (Q_3) and the first quartile (Q_1).
  IQR = Q3 – Q1

- *for A → IQR = 1 - (-2) = 3*
  for B → IQR = 5 - (-2) = 7

Outliers

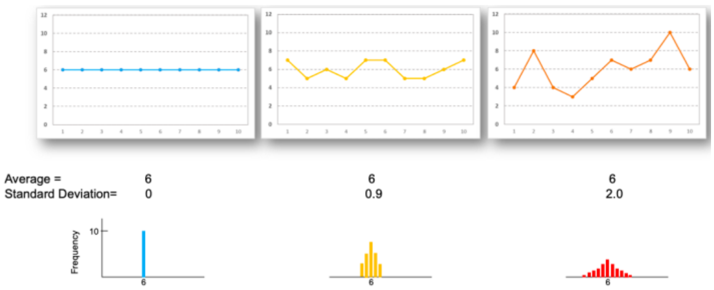
- A value is suspected to be a potential outlier if it is less than *(1.5) x (IQR)* below the first quartile or more than *(1.5) x (IQR)* above the third quartile.
- A potential outlier is a data point that is significantly different from the other data points.
- A data point that is an abnormal distance from other values in a data set
- These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.
- E.g.: In a class, all the students scored marks between range 50-70 except one who secured 95%. So, the one scored 95 is an outlier; thus, need to understand what helped the student to score more so that the same can be made available to remaining students to increase the marks.

Variance

- The variance is the average of the squares of the deviations.
- for samples $(x - \bar{X})$ & for population $(x - \mu)$
- for population, $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$
- for sample, $s^2 = \frac{\sum(x_i-\bar{X})^2}{n-1}$
- *for A → s2 = 6.97*
  for B → s2 = 9.14

Standard Deviation

- The sample standard deviation **s** is said to be a point estimator of the population standard deviation **σ**.
- The average squared distance from the mean.
- The standard deviation is the square root of the variance.



- Square Root of Mean of Square of Error (RMSE).
- *for A* → $s = \sqrt{6.97} = 2.64$
  for B → $s = \sqrt{9.14} = 3.02$

Variance and Standard Deviation of 1, 2, 3, 4, 10?

Average = 20/5 = 4
Range = 10-1 = 9

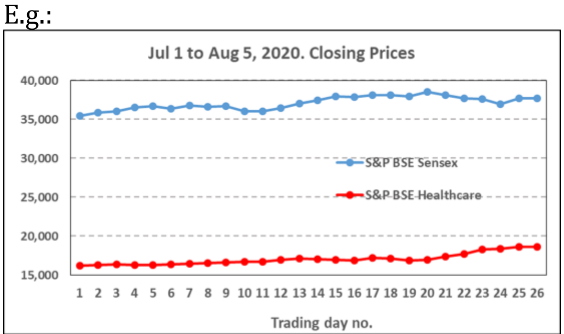
(RMSE)	Data	1	2	3	4	10
E - Error	Error or Deviation from	(1 - 4)	(2 - 4)	(3 - 4)	(4 - 4)	(10 - 4)
		-3	-2	-1	0	6
S - Square	Square the Error/Deviation	9	4	1	0	36
	Sum of Square of the Err	Sum = 9 + 4 + 1 + 0 + 36 = 50				
M - Mean	Average/Mean of Sum of Square	**Variance = Sum/5 = 50/5 = 10**				
R - Square Root	Square Root of the Mean	**Standard Deviation = sqrt(10) = 3.162**				

Coefficient of Variation, CoV

Co-efficient of Variation = Std Deviation/Average

$$CoV = \frac{\sigma}{\mu}$$

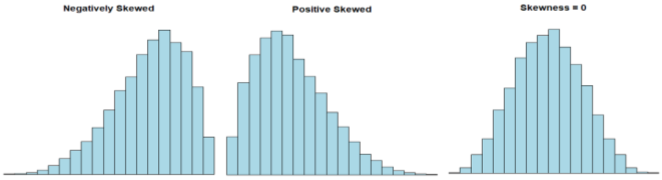
When averages of two data differ a lot, CoV may capture variation better than Variance.
E.g.:



	Sensex	Healthcare
Minimum	35,414	16,169
Maximum	38,493	18,630
Range	3,079	2,461
Std. Deviation	841	715
Average	37,077	17,026
CoV	0.023	0.042

In above example, CoV is able to capture the variation on Healthcare is more than that on Senxex.

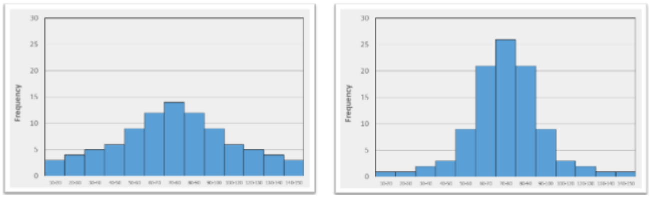
Skewness from Mean & Median



Suppose, ***Median > Mean*** ➔ Histogram may be skewed left
Suppose, ***Median < Mean*** ➔ Histogram may be skewed right. (i.e. More than 50% population is to the left of mean)
Suppose ***Median = Mean*** ➔ We may have a symmetric distribution

Peakedness (or Flatness) of a frequency distribution

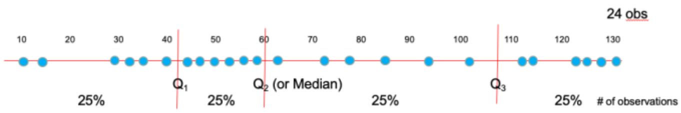
Kurtosis measures Peakedness of a distribution.



Frequency distribution on the left is flatter than the one on the right.
Kurtosis measures flatness of a frequency distribution.

Five-Number Summary

Five-Number Summary consist of *Minimum Value, Q_1, Q_2 (Median), Q_3, Maximum Value*



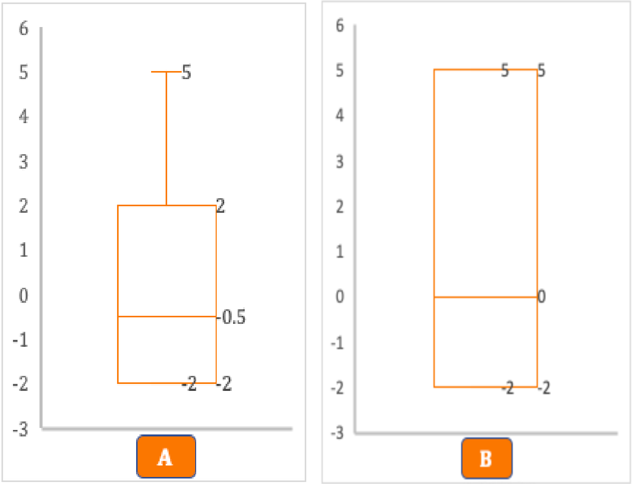
- Minimum [10], Q_1 [42], Q_2 (Median) [60], Q_3 [105] and Maximum [130].

Five-Number Summary		
	A	**B**
	-2, -2, -1, 0, 1, 5	**-2, -2, -1, 0, 1, 5, 5**
Minimum	-2	-2
Q1	-2	-2
Q2	-0.5	0
Q3	1	5
Maximum	5	5

Outliners		
IQR	3	7
1.5*IQR	4.5	10.5
Upper Limit Q3 + 1.5*IQR	5.5	15.5
Lower Limit Q1 - 1.5*IQR	-6.5	-12.5

Boxplot

Boxplots are also called Whisker plots



Excel Formula:

Minimum	*=Min(range)*
Maximum	*=Max(range)*
Q1	*=Quartile.Inc(range,1)*
Q2	*=Quartile.Inc(range,2)*
Q3	*=Quartile.Inc(range,3)*
Mean	*=Average(range)*
Median	*=Median(range)*
Mode	*=Mode(range)*
Population Variance	*=Var.p(range)*
Population Stdev	*=Stdev.p(range)*
Sample Variance	*=Var.s(range)*
Sample Stdev	*=Stdev.s(range)*

Standard Deviation as a Ruler

z-score

z-score is the statistical distance from the mean. i.e. how far is an observation from the average, in terms of standard deviation.
z-score is the standardized value which measures how many standard deviations is the data point from the mean.

$$z = \frac{x-\overline{X}}{s} \quad , \quad z = \frac{x-\mu}{\sigma}$$

If we assume that anything beyond $\pm k\sigma$ is an outliner, then we have another tool to analyse data.
E.g.: *You have a job offer of Rs. 8.2 lakhs in Hyderabad and another of Rs. 8.6 lakhs in Bangalore. The mean and variance for similar jobs in Hyderabad were Rs. 7 lakh and Rs. 9 lakhs, while for Bangalore, these were Rs. 7.4 lakhs and Rs. 12.25 lakhs. Which is better job offer?*
For Hyderabad,

$$x = 8.2, \quad \mu = 7, \quad \sigma^2 = 9$$
$$Z_{hyd} = \frac{8.2-7}{\sqrt{9}} = \frac{1.2}{2} = 0.4$$

For Bangalore,

$$x = 8.6, \quad \mu = 7.4, \quad \sigma^2 = 12.25$$
$$Z_{bgl} = \frac{8.6-7.4}{\sqrt{12.25}} = \frac{1.2}{3.5} = 0.3428$$

The offer in Hyderabad is 0.4 σ's from the mean.
The offer in Bangalore is 0.34 σ's from the mean.
Hence, the offer in Hyderabad is better than in Bangalore.
E.g.: *The Feb "high" temperature averaged 30°C with variance 100°C, while in May these were 40°C and 64°C. When is it more unusual to have a high of 35°C?*
During February,

$$x = 35, \quad \mu = 30, \quad \sigma^2 = 100$$
$$Z_{feb} = \frac{35-30}{\sqrt{100}} = \frac{5}{10} = 0.5$$

During May,

$$x = 35, \quad \mu = 40, \quad \sigma^2 = 64$$
$$Z_{may} = \frac{35-40}{\sqrt{64}} = \frac{-5}{8} = -0.625$$

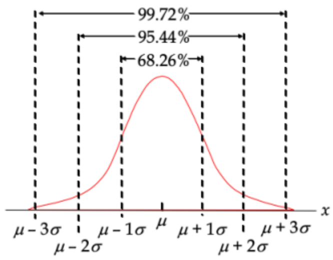
So, 35°C in February is 0.5 σ's from the mean & in May is 0.625 σ's from mean.
Hence, 35°C is more unusual in the month of May than in February.

Empirical Rule

Symmetrically distributed data follows a pattern whereby most data points fall within 3 standard deviations of the mean.
Only works with a well-centred, symmetrical bell-shaped curve.

Empirical Rule = Three Sigma Rule
Sigma = Standard deviation

- ***68.26 %*** of data points are within *1 standard deviation*.
- ***95.44%*** of data points are within *2 standard deviation*
- ***99.72%*** of data points are within *3 standard deviation*.



E.g.: *Your team member says he has received a job offer of Rs. 15 lakhs in Hyderabad and would like to put in his papers. Should you negotiate with him and try to increase his salary or tell him you cannot match that offer and he should put in his papers.*
HR says that for his level in Hyderabad,
μ = Rs. 8 lakhs & σ = Rs. 2 lakhs
So, z-score of the offer = (15 – 8) / 2 = 3.5
Suppose the salaries for his profile and level are bell shaped then his offer is an Outliers as its outside of the 99.72% of the data points (within 3 StDev)

Chebyshev's Theorem

At least $(1 - 1/z^2)$ of the items in any data set will be within z standard deviations of the mean, where z is any value greater than 1.
Chebyshev's theorem **requires z > 1**, but z need not be an integer.

- At least 75% of the data values must be within 2 standard deviations from the mean.
$$\left(1-\frac{1}{2^2}\right) = \left(1-\frac{1}{4}\right) = \frac{3}{4} = 0.75$$
- At least 89% of the data values must be within 3 standard deviations from the mean.
$$\left(1-\frac{1}{3^2}\right) = \left(1-\frac{1}{9}\right) = \frac{8}{9} = 0.89$$
- At least 94% of the data values must be within 4 standard deviations from the mean.
$$\left(1-\frac{1}{4^2}\right) = \left(1-\frac{1}{16}\right) = \frac{15}{16} = 0.9375$$

E.g.: *Applying Chebyshev's Theorem on above example (Rs. 15 lakhs offer in Hyderabad)/*
We know,
μ = Rs. 8 lakhs & σ = Rs. 2 lakhs
So, z-score of the offer = (15 – 8) / 2 = 3.5
Considering the population of all employees in Hyderabad at that level.
Using Chebyshev's Theorem,

$$\left(1-\frac{1}{3.5^2}\right) = \left(1-\frac{1}{12.25}\right) = \frac{45}{49} = 0.9184$$

So, ~92% of the population of all employees in Hyderabad at that level will be within 3.5σ of the mean.
i.e. only 8% lies outside this interval.

Introduction to Probability

***Probability** is a mathematical tool used to study randomness.*

Probability measures uncertainty.

It deals with the chance (the likelihood) of an event occurring.

An ***experiment*** is a planned operation carried out under controlled conditions.

A result of an experiment is called an ***outcome***.

The ***sample space*** of an experiment is the set of all possible outcomes. E.g.: S = {H,T}

An ***event*** is any combination of outcomes. Upper case letters like A and B represent events.

Equally likely means that each outcome of an experiment occurs with equal probability. So, the probability of each such event can be calculated by dividing count of outcomes of an event to total number of outcomes in the sample space.

"∪" Event: The Union

An outcome is in the event A ∪ B if the outcome is in A or is in B or is in both A and B.

"∩" Event: The Intersection

An outcome is in the event A ∩ B if the outcome is in both A and B at the same time.

Two events are ***mutually exclusive*** when both the event cannot occur at the same time. i.e. P(A ∩ B)=0

The ***complement*** of event A is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in A. Notice that $P(A) + P(A') = 1$.

The ***conditional probability*** of A given B is written $P(A|B)$. $P(A|B)$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space.**

$$\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)}, where\ P(B) > 0}$$

Two events A and B are ***independent*** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events.

The ***odds*** of an event presents the probability as a ratio of success to failure. This is common in various gambling formats. Mathematically, the odds of an event can be defined as:

$$\frac{P(A)}{1 - P(A)}$$

A ***contingency table*** provides a way of portraying data that can facilitate calculating probabilities.

A ***tree diagram*** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labelled with either frequencies or probabilities.

Example:

The sample space S is the whole numbers starting at one and less than 20.

Let event A = the even numbers and event B = numbers greater than 13.

$S = \{1,2,3,4,5,6,7,8,9,10,11,12,13,1,15,16,17,18,19\}$

$A = \{2,4,6,8,10,12,14,16,18\}$

$B = \{14,15,16,17,18,19\}$

$A' = \{1,3,5,7,9,11,13,15,17,19\}$

$A \cap B = \{1416,18\}$

$A \cup B = \{2,4,6,8,10,12,14,15,16,17,18,19$

$$P(A) = \frac{9}{19}, P(B) = \frac{6}{19}, P(A \cap B) = \frac{3}{19},$$

$$P(A \cup B) = \frac{10}{19}$$

$$P(A') = \frac{10}{19}, \quad so \longrightarrow P(A) + P(A') = 1$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3}{6}, P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{3}{9}$$

Rules of Probabilities

1. Probability of an event must be between *0 and 1* (inclusive)
   $$0 < P(A) < 1$$
2. A event that A does not occur is called A *complement* or simply not A.
   $$P(A') = 1 - P(A)$$
3. If two events A and B are *mutually exclusive*, the probability of both events A & B occurring is 0.
   $$P(A \cap B) = 0$$
4. If two events A & B are *mutually exclusive*, the probability of either A or B is sum of their separate probability
   $$P(A \cup B) = P(A) + P(B)$$
5. If events in a set are *mutually exclusive* and collective exhaustive, the sum of their probabilities must add up to 1
   $$P(Heads) + P(Tails) = 1$$
6. If two events A & B are *not mutually exclusive*, the probability of either event A or event B occurring is the sum of their separate probabilities minus the probability of their simultaneous occurrence.
   $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
7. If two events A and B are *independent*, the probability of both events A & B occurring is equal to the product of their individual probabilities.
   $$P(A \cap B) = P(A)P(B)$$
8. If two events A & B are *not independent*, the probability of both events A & B occurring is the product of the probability of event A multiplied by probability of event B occurring, given that event A has occurred.
   $$P(A \cap B) = P(A)P(B|A)$$

Getting values of Probability

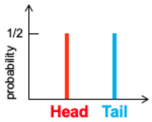
A priori/classical Probability

- Probability is assumed.
- Equal probability is assigned to each outcome.

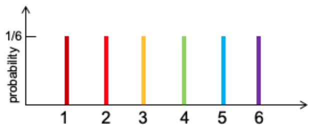
$$Probability = \frac{No. of\ outcomes\ in\ which\ the\ event\ occurs}{Total\ no. of\ possible\ outcomes}$$

Example:

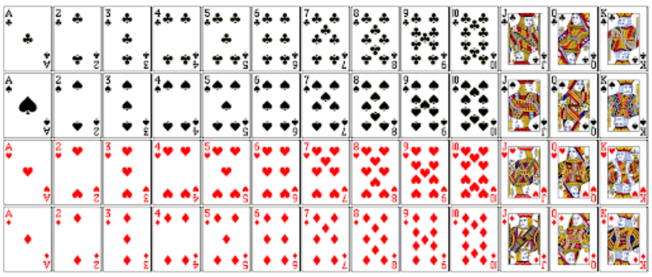
- Tossing a coin
  P(H) = ½
  P(T) = ½

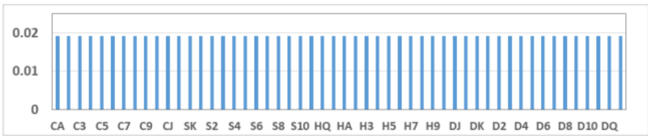


- Gender of the baby
  P(Boy) = ½
  P(Girl) = ½
- Tossing a dice
  P(1) = 1/6
  P(2) = 1/6
  P(3) = 1/6
  P(4) = 1/6
  P(5) = 1/6
  P(6) = 1/6



- Pack of 52 Cards
  P(any card) = 1/52 = 0.019231





Short coming of a priori approach

- Probability does not give a proper idea if an event could actually occur.
  E.g.: Rain/NoRain – 50:50 chances

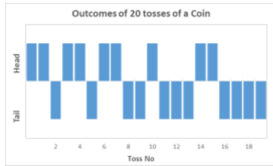
Empirical Probability

- From historical data or experiments or observation
- Life tables in insurance, earthquakes, rainfall, twins, quality, stock market, …

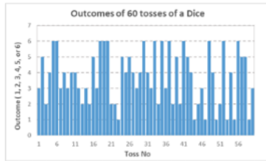
Item	Probability
Left-handed	1 : 10 persons
Twins	3 : 100 births
Vegetarian	38 : 100 persons
Aircraft crash	1 : 48 lakh flights
Boys to Girls ratio	51.2 : 48.8

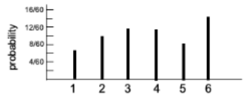
Example:

- Tossing a coin (20 times)
  P(H) = 9/20
  P(T) = 11/20



- Gender of the baby (40 births in a hospital)
  P(G) = 18/40
  P(B) = 22/40
- Tossing a dice 60 times
  P(1) = 7/60
  P(2) = 10/60
  P(3) = 11/60
  P(4) = 11/60
  P(5) = 8/60
  P(6) = 13/60





- Life table

Subjective Probability

- Personal judgement.
- Past experience, personal opinions & biases, etc
- Covid-19 vaccine by Dec20.

Example:

- P(Fire) = 0.001 --- Insurance Company A
  P(Fire) = 0.002 --- Insurance Company B
  P(Fire) = 0.010 --- Insured Person
- Sports betting
  P(India Wins) = 0.40 --- Bookie A
  P(India Wins) = 0.45 --- Bookie

Type of probability	How determined?	Examples
A priori	By assumption- all outcomes are equi-likely	Most examples in the textbooks and to explain the concepts.
Empirical	From observation, experiments, customer surveys, etc.	Life tables used by insurance companies; Clinical trial data used by drug approval agencies; Exit polls; Customer surveys; Customer data- in banking, ARPU in telecom; Life of automobile batteries, tires, electrical switches; Occurrence of earthquakes, rainfall, floods, fires, etc.
Subjective	Own belief	Everyday decisions- carry umbrella or not, buy or sell gold/oil/stocks by retail customers, participating in a closed-bid auction, etc.

Assigning Probabilities

Basic Requirements for Assigning Probabilities

1. $0 \leq P(E) \leq 1$ for any outcome E
2. $\sum P(E_i) = 1$

Equally Likely

Assigning probabilities based on the assumption of equal likely outcomes.

Example: Rolling a fair die and observing the top face
S={1,2,3,4,5,6}
Probabilities of each sample point is 1/6, i.e. each sample point has a 1/6 chance of occurring.

Properties of equally likely events:

- Complementary events
  P(A') = 1 – P(A)
- Intersection of events
  P(A & B) = $P(A \cap B)$

- Inclusive or Additional Law

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

Relative Frequency Method

Assigning probabilities based on historical data

Example:

The pizza eatery only has delivery service and has divided its catchment area into four zones. The following table gives the historical average number of orders received on a Saturday from each zone.

Zone	Freq	Relative Freq
East	20	0.10
North	30	0.15
South	70	0.35
West	80	0.40
TOTAL	**200**	**1.00**

Properties of relative frequency method:

- Each probability is between 0 & 1
  P(Cust from EAST)= P(E) = 0.1
- All the probabilities add up to 1
- Complementary Events
  P(Customer not from EAST)=P(E')=1-0.1=0.9
- Addition Law & Mutually exclusive events
  P(Cust is from North or West)

$$P(N \cup W) = P(N) + P(W) = 0.15 + 0.40 = 0.55$$

Analysing using ***contingency table***

		CLASS				
		First	Second	Third	Crew	Total
SURVIVAL	Alive	202	118	178	212	710
	Dead	123	167	528	673	1491
	Total	325	285	706	885	2201

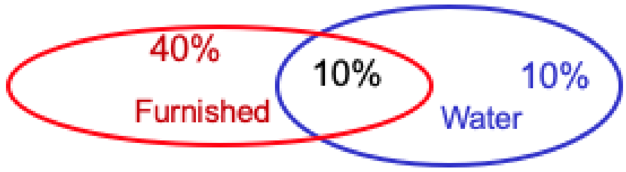
$$P((F \cap S) \cup (S \cap A)) = \frac{202}{2201} + \frac{118}{2201} = \frac{320}{2201}$$

$$P(!(F \cap A)) = 1 - \frac{202}{2201} = \frac{1999}{2201}$$

$$P(Crew \cup Survived) = P(C) + P(S) - P(C \cap S)$$

$$P(C \cup S) = \frac{885}{2201} + \frac{710}{2201} - \frac{212}{2201} = \frac{1383}{2201}$$

Analysing using ***Venn Diagram***



$$P(F) = 0.5, P(W) = 0.2, P(F \cap W) = 0.1$$
$$P(F \cup W) = P(F) + P(W) - P(F \cap W) = 0.6$$
$$P(!(F \cup W)) = 1 - P(F \cup W) = 0.4$$
$$P(W \cap F`) = P(W) - P(W \cap F) = 0.1$$

Subjective Method

Assigning probabilities based on judgement.

Often managers use their experience and intuition (and the data available) to assign probabilities. The probabilities represent their belief in the likelihood of the events

Usually, probability estimates are based on the Relative Frequency approach together with the subjective estimate.

Example: The firm will shortly launch a variant of the existing model.

R&A assigned the following probabilities to the possible market share by year-end: P(5%) = 20%, P(10%) = 55% and P(15%) = 25%.

VP Marketing modified the numbers as follows:

P(5%) = 25%, P(10%) = 40% and P(15%) = 35%

Conditional Probability

Using Contingency Table

		CLASS				
		First	Second	Third	Crew	Total
SURVIVAL	Alive	202	118	178	212	710
	Dead	123	167	528	673	1491
	Total	**325**	**285**	**706**	**885**	**2201**

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

$$P(Alive\ \&\ from\ First\ class) = P(A \cap B) = \frac{202}{2201}$$

So, probability of combination of either of the survival and any of the class together is joint probability. Using contingency table, we directly its value by dividing the count by total instances.

Marginal probability is the probability of an event irrespective of the outcome of another variable.

P(First)=325/2201, P(Alive)=710/2201

Conditional Probability is the likelihood that an event will occur given that another event has already occurred.
P(survived given he was in 3rd class) = 178/706
P(3rd class given he survived) = 178/710

$$P(A|T) = \frac{P(A \cap T)}{P(T)} = \frac{178/2201}{706/2201} = \frac{178}{706}$$

Additional (OR) Rule

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

$$P(A \cup F) = \frac{710}{2201} + \frac{325}{2201} - \frac{202}{2201} = \frac{833}{2201}$$

Example:

Frequency of parts, nos

	Defective	**Good**	Total
Supplier-A	5	10	15
Supplier-B	15	20	35
Total	20	30	50

Calculating probability in percentage:

Fraction of parts, % or probability in %

	Defective	**Good**	Total
Supplier-A	10	20	30
Supplier-B	30	40	70
Total	40	60	100

Joint Probability:

P(A and Defective) = 10
P(A and Good) = 20
P(B and Defective) = 30
P(B and Good) = 40

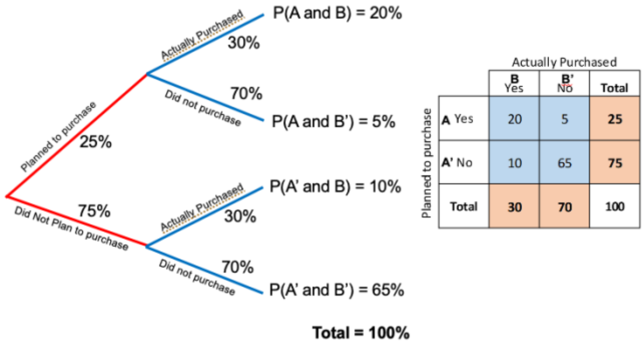
Marginal Probability:

P(A) = P(A and Defective) + P(A and Good) = 10 + 20 = 30
P(B) = P(B and Defective) + P(B and Good) = 30 + 40 = 70

P(D) = P(A and Defective) + P(B and Defective) = 10 + 30 = 40
P(G) = P(A and Good) + P(B and Good) = 20 + 40 = 60

Using Tree Diagrams

A ***tree diagram*** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labelled with either frequencies or probabilities.
For Instance:



Bayes' Theorem

Suppose E_i's are ALL the possible outcomes and the prior probabilities $P(E_i)$ have been assigned to them. The event F has occurred.

$$P(E_i|F) = \frac{P(E_i)P(F|E_i)}{P(E_1)P(F|E_1) + P(E_2)P(F|E_2) + \cdots + P(E_n)P(F|E_n)}$$

$$P(E_i|F) = \frac{P(E_i)P(F|E_i)}{P(E_1 \cap F) + P(E_2 \cap F) + \cdots + P(E_n \cap F)}$$

$$\boldsymbol{P(E_i|F) = \frac{P(E_i)P(F|E_i)}{P(F)}}$$

Required:

- The E_i's are mutually exclusive

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

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

- The E_i's are collectively exhaustive – i.e. are all the possible events.

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_3)$$

- $P(F|E_i)$'s are available

Tabular Approach Steps

Step 1

- Identify the mutually exclusive events (E's) that make up the Sample Space;
- Identify the Fact F.
- Note down the P(E) and the conditional probabilities (F | E)
- Prepare the table with 5 columns and (n+1) rows, where n is the size of the sample space

Step 2

- Enter
  - Column 1 The events E's
  - Column 2 The prior probabilities P(E)'s
  - Column 3 P(F | E)'s

Step 3

- Column 4 Compute the joint probabilities Using, $P(F \cap E) = P(E)P(F|E)$

Step 4

- Column 4. The last cell will contain $\Sigma P(E \cap F) = P(F)$

Step 5

- Column 5 Compute the posterior probabilities using

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

Example 1:

The MD's EA is often late returning from lunch. Based on observation, HR assigned the following probabilities:

Lunch Location	Probability he is late:
Out	40%
Company Canteen	19%
Cubicle	1%

HR knows that all locations are equally likely. Today the EA came back late from lunch.

This information has to be factored into the data above!

Events:

E_1: Lunch Out, E_2: Lunch at Canteen, E_3: Lunch in Cubicle

Assuming all the possibilities for the EA to have lunch are E_1, E_2, E_3.

F: EA came late today.

Prior Probabilities, $P(E_1) = P(E_2) = P(E_3) = {}^1/_3$

Conditional Probabilities,

$P(F|E_1) = 0.4, P(F|E_2) = 0.19, P(F|E_3) = 0.01$

Event (F)	P(E)	P(F|E)	P(F∩E)	P(E|F)
E_1	1/3	0.4	0.4/3	0.67
E_2	1/3	0.19	0.19/3	0.31
E_3	1/3	0.01	0.01/3	0.02
Total	**1.0**		**P(F)=0.2**	**1.00**

$$P(F \cap E) = P(E)P(F|E)$$

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

So,

P(Eating Lunch given EA came late) = $P(E_1|F)$=0.67

$P(E_2|F)$=0.31

$P(E_3|F)$=0.02

Using Contingency Table,

E_1: Lunch Out, E_2: Lunch at Canteen, E_3: Lunch in Cubicle

Assuming all the possibilities for the EA to have lunch are E_1, E_2, E_3.

F: EA came late today.

Prior Probabilities, $P(E_1) = P(E_2) = P(E_3) = {}^1/_3$

	Late	Not Late	Total
Eating Out	40%	?	1/3
Canteen	19%	?	1/3
Cubicle	1%	?	1/3
Total	?	?	1

Computing the remaining values:

	Late	Not Late	Total
Eating Out	40% (1/3)	60% (1/3)	1/3
Canteen	19% (1/3)	81%(1/3)	1/3
Cubicle	1% (1/3)	99%(1/3)	1/3
Total	?	?	1

Converting it based on Probabilities:

	Late (F)	Not Late (F')	Total
Eating Out (E_1)	0.4/3	0.6/3	1/3
Canteen (E_2)	0.19/3	0.81/3	1/3
Cubicle(E_3)	0.01/3	0.99/3	1/3
Total	0.2	0.8	1

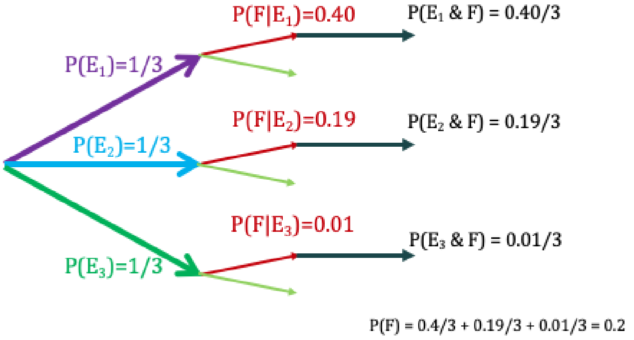
Now,

$P(E_1|F)$ = (0.4/3)/0.2 = 2/3 = 0.67

$P(E_2|F)$ = (0.19/3)/0.2 = 0.31

$P(E_3|F)$ = (0.01/3)/0.2 = 0.02

Using Tree Diagram,



P(E1|F) = (0.40/3) / 0.2 = 0.67
P(E2|F) = (0.19/3) / 0.2 = 0.31
P(E3|F) = (0.01/3) / 0.2 = 0.03

Example 2:

The disease is present in 0.5% of the population. It is a deadly disease and death is almost always inevitable.
But there is a test that can detect the disease.
The True Positive is 99% while the False Positive is 5%.
Question: If you test positive, should you panic?

	Disease	No Disease
Test Positive	True Positive	False Positive
Test Negative	False Negative	True Negative

Events:
E_1: Have Disease, E_2: No Disease
F: Tested Positive
Prior Probabilities,
$P(E_1) = 0.5\% = 0.005$, $P(E_2) = 0.995$
Conditional Probabilities,

$$P(F|E_1) = 0.99, P(F|E_2) = 0.05$$

Event	P(E)	P(F|E)	P(F∩E)	P(E|F)
E_1	0.005	0.99	0.00495	0.0905
E_2	0.995	0.05	0.04975	0.9095
	1		P(F)=0.0547	1

$$P(F \cap E) = P(E)P(F|E)$$

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

So, P(Having disease given tested positive)
$P(E_1|F)=0.0905$

Using Contingency table,

Events:
E_1: Have Disease, E_2: No Disease
F: Tested Positive
Prior Probabilities,
$P(E_1) = 0.5\% = 0.005$, $P(E_2) = 0.995$
Conditional Probabilities,

$$P(F|E_1) = 0.99, P(F|E_2) = 0.05$$

	Tested Positive	Tested Negative	Total
Have Disease	99%	?	0.50%
No Disease	5%	?	?
Total	?	?	100%

Computing remaining values:

	Tested Positive	Tested Negative	Total
Have Disease	99%(0.005)	1%(0.005)	0.005
No Disease	5%(0.995)	95%(0.995)	0.995
Total	?	?	1

Converting to Probabilities:

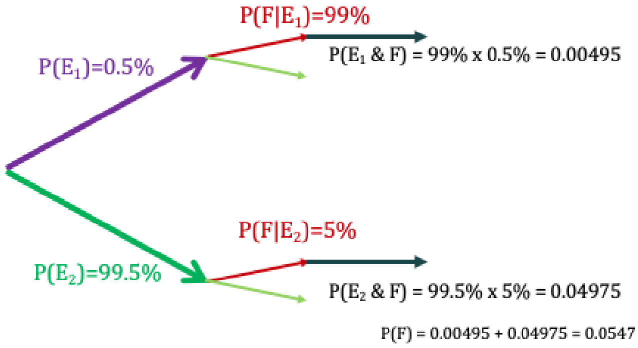
	Tested Positive	Tested Negative	Total
Have Disease	0.00495	0.00005	0.005
No Disease	0.04975	0.94525	0.995
Total	0.0547	0.9453	1

Now,
$P(E_1|F) = 0.00495/0.0547 = 0.0905$

Using Tree Diagram,

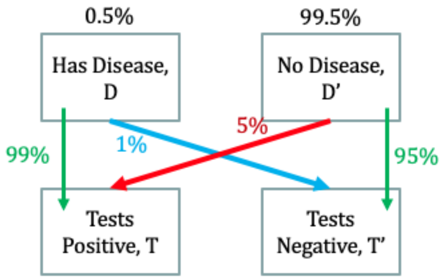
Events:
E_1: Have Disease, E_2: No Disease
F: Tested Positive
Prior Probabilities,
$P(E_1) = 0.5\% = 0.005$, $P(E_2) = 0.995$
Conditional Probabilities,

$$P(F|E_1) = 0.99, P(F|E_2) = 0.05$$



P(E1|F) = 0.00495 / 0.0547 = 0.0905

Using Classification Algorithm,



$$P(D|T) = \frac{(99 \times 0.5)}{(99 \times 0.5) + (99.5 \times 5)} = 0.0905$$

$$P(T) = P(T|D)P(D) + P(T|D')P(D')$$

$$P(T) = 99 \times 0.5 + 99.5 \times 5$$

Random Variables

A ***random variable*** is a numerical description of the outcome of a random experiment.

A ***discrete random variable*** may assume a countable number of values
The discrete random variables remain unknown till its completed. i.e. # of customers using the ATM in a day remains unknown until the day ends.
E.g.:
# of dependents of an employee
# of customers using the ATM in a day
# of sixes in a T20 match
# of owners who like the product

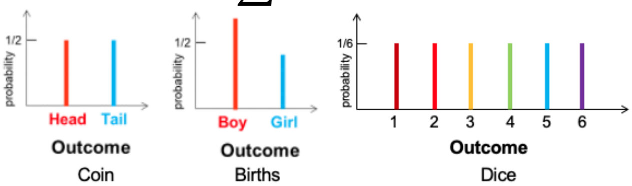
A ***continuous random variable*** mat assume any numerical value in an interval

The random variable inherits the probabilities of the events of the random experiment.
E.g.:
Life of a tire
Time between call at the call centre
Volume of water in a 1 litre mineral water bottle
% of owners who like the product

Probability function of a Discrete Random Variable

The ***Probability function*** lists the possible outcomes and their probabilities
Like frequency distribution, probability distribution have descriptive measures

$$0 \leq P(x) \leq 1$$
$$\sum P(x) = 1$$



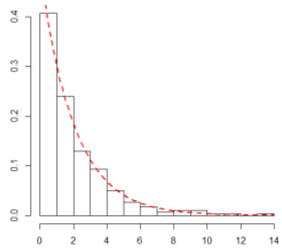
E.g.: # of dependents of an employee

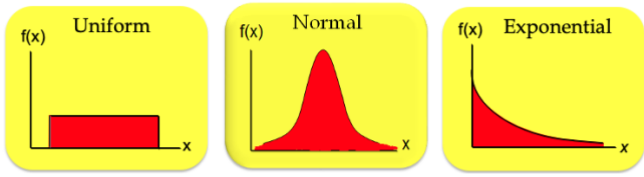
# of dependent (X)	0	1	2	3
Probability P(x)	0.1	0.4	0.4	0.1

Probability function of a Continuous Random Variable

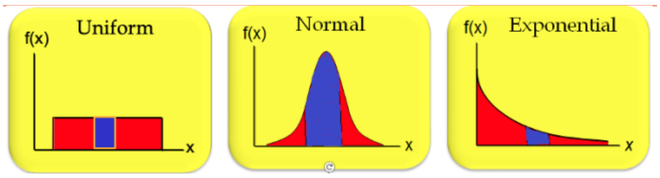
Probabilities assigned to intervals of numbers.
Such probability distributions have descriptive measures: μ, σ etc

Let X: *The time spend by a car at the toll booth*
A Histogram can be developed for the data.
A line joining the top of each rectangle at the enter will generate the empirical probability density function.





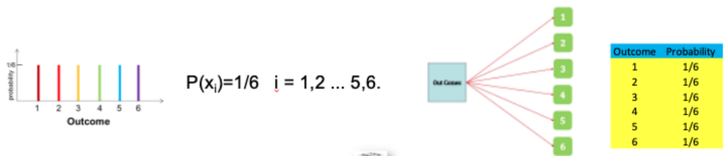
Area Under Curve



- X can assume any value in an interval on the real line or in an interval.
- P(X = a) = 0, for any number a
- P(a < X < b) = P(a ≤ X < b) = P(a < X ≤ b) = P(a ≤ X ≤ b)
- P(a < X < b) is the area under the graph of the probability density function between a and b.

Probability Distribution – 4 Ways

1. Graph
2. Equation
3. Probability Tree
4. Table



Probability distribution of a fair dice.

Probability Distribution Family (GETT):
Graph + Equation + Table + Tree

Discrete Probability Distribution

The probability function provides the probability for each value of the random variable.

The required conditions for probability function are:

$$0 \le f(x) \le 1$$
$$\sum f(x) = 1$$

The ***expected value***, or ***mean*** or average or error (deviation from average) of a random variable is a measure of its central location.

$$E(X) = \mu = \sum Xf(X)$$

The ***variance*** summarizes the variability in the values of a random variable.

$$Var(X) = V(X) = \sigma^2 = \sum (X-\mu)^2 f(X)$$

The ***standard deviation,*** σ is defined as the positive square root of the variance.

Root Weighted sum Square of Error - RW_SSE

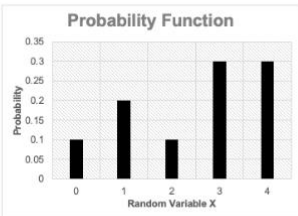
Example:

Constructing the Empirical Distribution

Below table provides past data of AC sales.

Calculating probability function f(x) – frequency distribution

Units Sold (x)	# of days	f(x)
0	10	0.1
1	20	0.2
2	10	0.1
3	30	0.3
4	30	0.3
TOTAL	**100**	**1.0**



Computing Expected value (μ) & Variance (σ^2)

X	f(X)	X*f(X)	(X-μ)	f(X) (X-μ)	(X-μ)²	f(X) (X-μ)²
0	0.1	0	-2.5	-0.25	6.25	0.625
1	0.2	0.2	-1.5	-0.30	2.25	0.45
2	0.1	0.2	-0.5	-0.05	0.25	0.025
3	0.3	0.9	0.5	0.15	0.25	0.075
4	0.3	1.2	1.5	0.45	2.25	0.675
T	**1.0**	**μ=2.5**		**0.0**		**σ^2=1.85**

So, mean daily sales = μ = 2.5 AC

Variance of daily sales = σ^2 = 1.85

Standard deviation of daily sales = σ = 1.36 AC

Now, if there is change in the number of AC sale suppose (Y=-3X), below date is computed:

X	f(X)	Y=-3X	f (Y)	Y* f(Y)	(Y-μ)	f(Y) (Y-μ)	(Y-μ)²	f(Y) (Y-μ)²
0	0.1	0	0.1	0	7.5	0.75	56.25	5.625
1	0.2	-3	0.2	0.2	4.5	0.90	20.25	4.050
2	0.1	-6	0.1	0.2	1.5	0.15	2.25	0.225
3	0.3	-9	0.3	0.9	-1.5	-0.45	2.25	0.675
4	0.3	-12	0.3	1.2	-4.5	-1.35	20.25	6.075
T	**1.0**		**1.0**	**μ=-7.5**		**0.0**		**σ^2=16.65**

So, mean daily sales = μ = -7.5 AC

Variance of daily sales = σ^2 = 16.65

Standard deviation of daily sales = σ = 5.08 AC

$$E(\alpha X) = \alpha E(X)$$
$$V(\alpha X) = \alpha^2 V(X)$$

Example: X gets a 10% discount on all orders, so here a=0.9

$$E(X \pm c) = E(X) \pm c$$
$$V(X \pm c) = V(X)$$

Example: Suppose X managed to convivence Ace to reduce charges on every order by Rs. 100. Here, c=100

Bivariate Discrete Probability Distribution

A probability distribution involving two random variables is called a ***bivariate probability distribution***.

Example:

HR ran a survey among 2000-strong staff on Job Satisfaction and Work Stress.

The cross-tabulation of the data is given below:

Job Satisfaction (x)	Work Stress (y)			Total
	Low	**Medium**	**High**	
Low	260	280	40	**580**
Medium	200	480	320	**1000**
High	40	240	140	**420**
Total	**500**	**1000**	**500**	**1000**

Computing it to probabilities… and creating Joint and Marginal probabilities.

Job Satisfaction (x)	Work Stress (y)			Total
	1	**2**	**3**	
1	0.13	0.14	0.02	**0.29**
2	0.10	0.24	0.16	**0.50**
3	0.02	0.12	0.07	**0.21**
Total	**0.25**	**0.50**	**0.25**	**1.00**

P(Job Satisfaction = Low & Work Stress),

P(x = 1 & y=3)=0.02 , P(x=1)=0.29

Computing the Joint distribution of two independent Random Variables

Ace Ad agency is a small start-up with only two clients, represented by X and Y, operating in different industries. For the two clients, Ace places ads in the local paper's classified section of the Saturday edition. X weekly ad spends in Rs.'000 is 0, 1, 2, 3, 4 (with some rounding off). Similarly, Y weekly ad spends in Rs.'000 is 0 and 1.
Based on past data, Ace has created the frequency distribution and subsequently the probability distributions for X and Y.

X	Data	f(X)
0	10	0.1
1	20	0.2
2	10	0.1
3	30	0.3
4	30	0.3

Y	Data	f(Y)
0	70	0.7
1	30	0.3

Since, X & Y operate in different industries, we may assume that the ad spends are independent of each other. So,
P(X = a & Y = b) = P(X = a)*P(Y = b)
So, we can create joint distribution as follows:

		X					
		0	1	2	3	4	
Y	0	0.07	0.14	0.07	0.21	0.21	0.7
	1	0.03	0.06	0.03	0.09	0.09	0.3
		0.1	0.2	0.1	0.3	0.3	1

Computing, Probability for X+Y

R=(X+Y)	Combination	P=f(X+Y)
0	y=0, x=0	0.07
1	y=0, x=1 y=1, x=0	0.17
2	y=0, x=2 y=1, x=1	0.13
3	y=0, x=3 y=1,x=2	0.24
4	y=0, x=4 y=1, x=3	0.30
5	y=1, x=4	0.09
	TOTAL	1

Calculating estimated value and variance for X & Y.

X	f(X)	X*f(X)	(X-μ)	f(X) (X-μ)	(X-μ)²	f(X) (X-μ)²
0	0.1	0	-2.5	-0.25	6.25	0.625
1	0.2	0.2	-1.5	-0.30	2.25	0.45
2	0.1	0.2	-0.5	-0.05	0.25	0.025
3	0.3	0.9	0.5	0.15	0.25	0.075
4	0.3	1.2	1.5	0.45	2.25	0.675
T	**1.0**	**μ=2.5**		**0.0**		**σ^2=1.85**

Y	f(Y)	Y*f(Y)	(Y-μ)	f(Y) (Y-μ)	(Y-μ)²	f(Y) (Y-μ)²
0	0.7	0	-0.3	-0.21	0.09	0.063
1	0.3	0.3	0.7	0.21	0.49	0.147
T	**1.0**	**μ=0.3**		**0.0**		**σ^2=0.21**

Calculating estimated value and variance for R=(X+Y)

R	P	R*P	(R-μ)	P (R-μ)	(R-μ)²	P (R-μ)²
0	0.07	0	-2.8	-0.0196	7.84	0.5488
1	0.17	0.17	-1.8	-0.3060	3.24	0.5508
2	0.13	0.26	-0.8	-0.1040	0.64	0.0832
3	0.24	0.72	0.2	0.0480	0.04	0.0096
4	0.30	1.20	1.2	0.3600	1.44	0.4320
5	0.09	0.45	2.2	0.1980	4.84	0.4356
T	**1.0**	**μ=2.8**		**0.0**		**σ^2=2.06**

$$\boldsymbol{E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)}$$

If X& Y are independent,

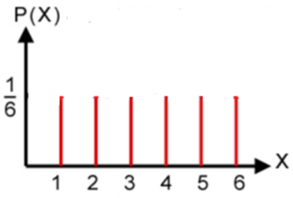
$$\boldsymbol{V(\alpha X + \beta Y) = \alpha^2 V(X) + \beta^2 V(Y)}$$

So, using above equation,

$$E(X - Y) = E(X) - E(Y)$$
$$E(X - Y) = 2.5 - 0..3 = 2.2$$
$$V(X - Y) = V(X) + V(Y)$$
$$V(X - Y) = 1.85 + 0.21 = 2.06$$

Discrete Uniform Probability Distribution

You are about to launch a new product. The product was test marketed, but preference for body colour was not.
There are six colours: Violet (1), Blue (2), Green (3), Yellow (4), Orange (5) and Red (6).
Initially it must be assumed that each body colour is equally preferred.
The probability function f(x) = 1/6



Poisson Probability Distribution ***Π(μ)***

A Poisson distributed random variable is used in estimating the number of occurrences in a **specified interval of time or space**

Application
Sizing the size of operations at a bank, call centre, service centre, petrol bunk, …
Examples
- # of vehicles arriving at a toll booth in one hour
- # of patients arriving in an emergency room between 11 and 12 pm
- # of typos in a page

Requirements
- Events occur independently.
- Two events cannot occur at exactly the same instant.
- The probability of an event in an interval is proportional to the length of the interval
- The probability that an event occurs is same in all intervals of equal size.

I: The specified interval
X = the number of occurrences in an interval
f(x) = the probability of x occurrences in an interval
μ = mean number of occurrences in an interval
X ~ Π(μ)

$$P(X=x)=\frac{\mu^x e^{-\mu}}{x!}=\frac{\lambda^x e^{-\lambda}}{x!}$$
$$\mu=\lambda=E(X)=V(X)$$

Poisson PD using MS Excel

=POISSON(X, AVERAGE, false)*100

x	Probability %	MS Excel formula
0	5.0	=POISSON(X,AVERAGE,false)*100
1	14.9	=POISSON(X,AVERAGE,false)*100
2	22.4	=POISSON(X,AVERAGE,false)*100
3	22.4	=POISSON(X,AVERAGE,false)*100
4	16.8	…
5	10.1	…
6	5.0	…
7	2.2	…
8	0.8	…
9	0.3	…
10	0.1	…
…		Average= 3
Total	100.0	

Poisson PD using Published Tables

Poisson Probabilities for Different Values of λ

Number of events	$\lambda=0.5$	$\lambda=1$	$\lambda=1.5$	$\lambda=2$	$\lambda=2.5$	$\lambda=3$
$x=0$	0.6065	0.3679	0.2231	0.1353	0.0821	0.0498
$x=1$	0.3033	0.3679	0.3347	0.2707	0.2052	0.1494
$x=2$	0.0758	0.1839	0.2510	0.2707	0.2565	0.2240
$x=3$	0.0126	0.0613	0.1255	0.1804	0.2138	0.2240
$x=4$	0.0016	0.0153	0.0471	0.0902	0.1336	0.1680
$x=5$	0.0002	0.0031	0.0141	0.0361	0.0668	0.1008
$x=6$	0.0000	0.0005	0.0035	0.0120	0.0278	0.0504
$x=7$	0.0000	0.0001	0.0008	0.0034	0.0099	0.0216
$x=8$	0.0000	0.0000	0.0001	0.0009	0.0031	0.0081

Example:
Employees visit the ATM at the average rate of 6 per hour in the post-lunch period. What is the probability of 2 arrivals in 30 minutes in the post-lunch period? What is the expected # of arrivals? Variance?

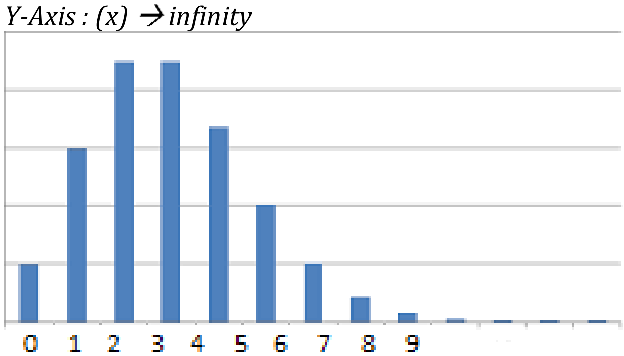
$$I = 30 \text{ minutes}$$
$$\mu = 6 \text{ per hour} = 3 \text{ per } 30 \text{ minutes}$$
$$x = 2$$
$$X : \#\ of\ arrivals\ in\ 30\ minutes\ period$$
$$P(X=2)=\frac{\mu^x e^{-\mu}}{x!}=\frac{3^2 e^{-3}}{2!}=0.2240$$

Expected # of arrivals & Variance is 3 per 30 minutes
P(2 arrivals in 30 minutes)=0.2240

Can be solve using Poisson Chart:

	Mean		
X	*2.9*	*3*	*3.1*
0	*0.0550*	*0.0498*	*0.0450*
1	*0.1596*	*0.1494*	*0.1397*
2	*0.2314*	*0.2240*	*0.2165*
3	*0.2237*	*0.2240*	*0.2237*
4	*0.1622*	*0.1680*	*0.1733*
5	*0.0940*	*0.1008*	*0.1075*

X – Axis : f(X) : Probability of X
Y-Axis : (x) → infinity



Binomial Distribution

Binomial Probability Distribution B(n,p)

The random variable X counts the number of successes in n trials.

Four Properties of a Binomial Experiment:

1. The experiment consists of a sequence of n identical trials.
2. Two outcomes, ***success*** and ***failure***, are possible on each trial.
3. The probability of a success, denoted by p, does not change from trial to trial.
4. The trials are independent.

$X \sim B(n, p)$

p: Probability of Success

q: Probability of Failure (q=1-p)

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{(n-x)} = C_x^n p^x q^{(n-x)}$$
$$= \frac{n!}{x!\,(n-x)!} p^x q^{(n-x)}$$

$$\mu = np$$
$$\sigma^2 = npq$$
$$\sigma = \sqrt{npq}$$

Rule of Thumb:

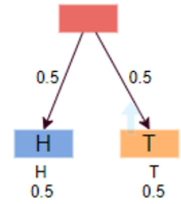
If N >> n, so that p does not change by much (if case of any situation where probability changes)

$$\frac{n}{N} < 5\%$$

Understanding Probability Tree for Binomial

Given: A fair coin tossed, P(H)=0.5 & P(T)=0.5

A Fair coin tossed ***two*** times:

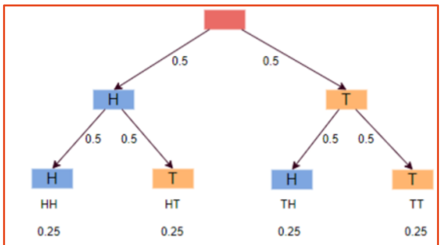


Given

P(H)= 0.5

P(T)= 0.5

Sum= 1.0



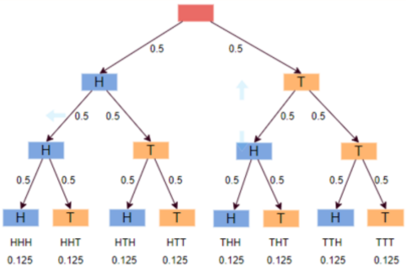
P(2 Heads) = 0.25

P(1 Head)= 2 x 0.25 = 0.50

P(0 Head) = 0.25

Sum = 1.00

A Fair coin tossed ***three*** times:



P(3 Heads) = 0.125

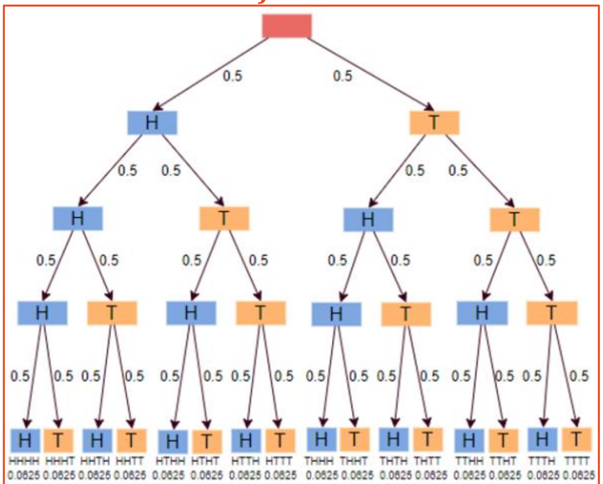
P(2 Heads)= 3 x 0.125 = 0.375

P(1 Head) = 3 x 0.125 = 0.375

P(0 Head) = 0.125

Sum= 1.000

A Fair coin tossed for ***four*** times:



P(4 Heads) = 0.0625

P(3 Heads)= 4 x 0.0625 = 0.2500

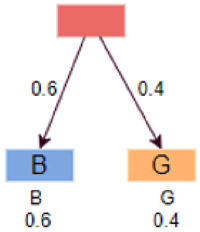
P(2 Heads)= 6 x 0.0625 = 0.3750

P(1 Head) = 4 x 0.0625 = 0.2500

P(0 Head) = 0.0625

Sum = 1.0000

Given: Probability distribution of Boys and Girls in 1, 2, and 3, children families?

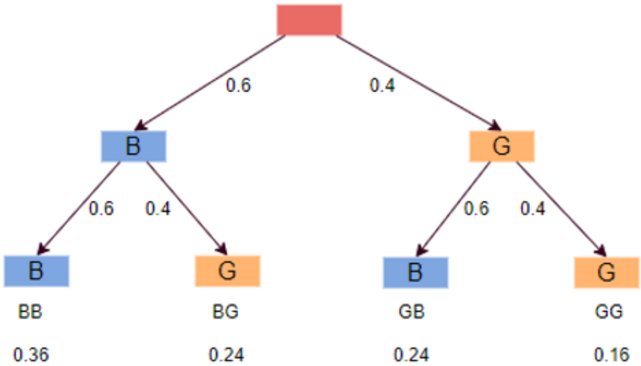


Given

P(B)= 0.6

P(G)= 0.4

Sum= 1.0



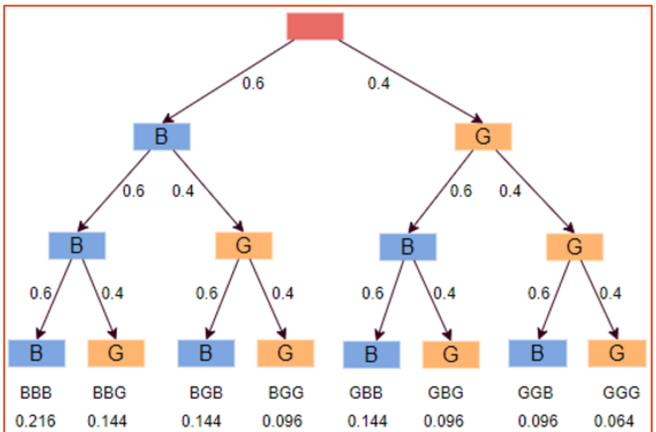
In **two-children** families-

P(2 Boys) = 0.36

P(1 Boy)= 2 x 0.24 = 0.48

P(0 Boy) = 0.16

Sum = 1.00



In **three-children** families-.

P(3 Boys) = 0.216

P(2 Boys)= 3 x 0.144 = 0.432

P(1 Boy) = 3 x 0.096 = 0.288

P(0 Girl) = 0.064

Sum = 1.000

Binomial Distribution using MS Excel

MS Excel generic formula:

=BINOM.DIST(x, NumberOfTrials, ProbabilityOfSuccess, FALSE)*100

For, Number if trials = 4 &

Probility of Success=0.1

Tagged, x	Probability x, %	MS Excel formula
0	65.61	=BINOM.DIST(0,4,0.1,FALSE)*100
1	29.16	=BINOM.DIST(1,4,0.1,FALSE)*100
2	4.86	=BINOM.DIST(2,4,0.1,FALSE)*100
3	0.36	=BINOM.DIST(3,4,0.1,FALSE)*100
4	0.01	=BINOM.DIST(4,4,0.1,FALSE)*100
Total	**100**	

Binomial Distribution using Published Tables

Table 1 Binomial distribution — probability fun

	x	p 0.01	0.05	0.10	0.15	0.20	0.25
n=1	0	.9900	.9500	.9000	.8500	.8000	.7500
	1	.0100	.0500	.1000	.1500	.2000	.2500
n=2	0	.9801	.9025	.8100	.7225	.6400	.5625
	1	.0198	.0950	.1800	.2550	.3200	.3750
	2	.0001	.0025	.0100	.0225	.0400	.0625
n=3	0	.9703	.8574	.7290	.6141	.5120	.4219
	1	.0294	.1354	.2430	.3251	.3840	.4219
	2	.0003	.0071	.0270	.0574	.0960	.1406
	3		.0001	.0010	.0034	.0080	.0156
n=4	0	.9606	.8145	.6561	.5220	.4096	.3164
	1	.0388	.1715	.2916	.3685	.4096	.4219
	2	.0006	.0135	.0486	.0975	.1536	.2109
	3		.0005	.0036	.0115	.0256	.0469
	4			.0001	.0005	.0016	.0039
n=5	0	.9510	.7738	.5905	.4437	.3277	.2373
	1	.0480	.2036	.3281	.3915	.4096	.3955

For, n=4 & p=0.1

P(0)=0.6561

P(1)=0.2916

P(2)=0.0486

P(3)=0.0036

P(4)=0.0001

Example 1:
Tossing a coin 10 times and we are interested in the number of heads

Here, X: # of heads
n=10
p=0.5 (success : heads)
q=0.5 (failure : tails)
All 4 properties satisfy for this example.

$$\mu = np = 10 * 0.5 = 5$$
$$\sigma^2 = npq = 10 * 0.5 * 0.5 = 2.5$$
$$\sigma = \sqrt{npq} = 1.58$$

suppose, we need to find P(# heads = 3)
x=3
X ~ B(10, 0.5)

$$f(x) = P(X = 3) = \binom{10}{3} p^3 q^{(10-3)}$$
$$= \frac{10!}{3!\,(10-3)!} 0.5^3 0.5^{(7)}$$
$$= 120 \times 0.5^{10} = 0.1171875$$

Example 2:
Indica sells encyclopaedias targeted towards children using door-to-door saleswomen. Ms Rita, a saleswomen with Indica, has randomly selected **20 houses** in the neighbourhood to sell the product. From past experience, Rita knows that the probability that a sale will be made is **0.1**.

Here, X:# of sale made in a day
n=20
p=0.1 (success: sale)
q=0.9 (failure: no sale)
Out of the 4 properties, 1,2 & 4 satisfies for this given situation.
Whereas, the probability of success may change. Initially p is 0.1, but as day progresses, Rita may get tired and the success rate may decrease.
So, we cannot use the Binomial Probability Distribution.

Example 3:
A 1000-strong IT firm is concerned about a low retention rate for its employees. In recent years, management has seen a turnover of 10% of the employees annually.
Thus, for any employee chosen at random, management estimates a probability of 0.1 that the person will not be with the company next year.
Three employees are selected at random. What is the probability that 1 of them will leave the company this year? m? s2?

Here, X : # of employees resigning
n=3
p=0.1 (Success: resign)
q=0.9 (failure: retained)
Out of all 4 properties, one might not satisfy here too, i.e the probability would change as the sampling is done without replacement. So… in first case the p=100/1000=0.1 & in second case p=99/999=0.099 and so on… as the sample space keeps reducing (without replacement), there is a change in probability but very minor change.
So, as per thumb rule. n/N = 3/1000 = 0.003 < 5%
As n<<N, we can consider above situation for binomial probability distribution.
x=1 *, P(1 of the employee will leave the company)*
X ~ B(3, 0.1)

$$\mu = np = 3 * 0.1 = 0.3$$
$$\sigma^2 = npq = 3 * 0.1 * 0.9 = 0.27$$
$$\sigma = \sqrt{npq} = 0.5196$$

$$f(x) = P(X = 1) = \binom{3}{1} p^1 q^{(3-1)}$$
$$= \frac{3!}{1!\,(3-1)!} 0.1^1 0.9^{(2)} = 0.3 \times 0.9^2$$
$$= 0.243$$

Example 3:
The Canteen Manager in a large factory estimated that 20% of the workers bring lunch from home. A random sample of 10 workers are taken.
Q. Compute the probability that exactly 4 bring lunch from home
Q. Compute the probability that at least 2 workers bring lunch from home

Here, X: # that bring lunch from home
X ~ B(10, 0.2)
Using below chart table:

10	p			
x	0.1	0.2	0.3	0.4
0	0.3487	0.1074	0.0282	0.0060
1	0.3874	0.2684	0.1211	0.0403
2	0.1937	0.3020	0.2335	0.1209
3	0.0574	0.2013	0.2668	0.2150
4	0.0112	0.0881	0.2001	0.2508
5	0.0015	0.0264	0.1029	0.2007
6	0.0001	0.0055	0.0368	0.1115
7	0.0000	0.0008	0.0090	0.0425
8	0.0000	0.0001	0.0014	0.0106
9	0.0000	0.0000	0.0001	0.0016
10	0.0000	0.0000	0.0000	0.0001

X ~ B(10, 0.2)
P(X = 4) = 0.0881
P(X >= 2) = 1 - P(X<2)
=1 – [P(X = 0) + P(X = 1)]
=1- (0.1074 + 0.2684)
=0.6242

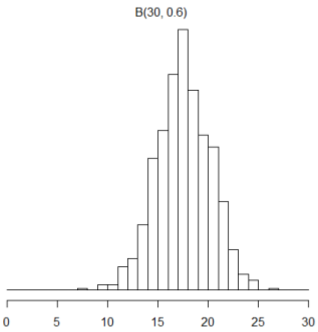
A Digression

Case 1:

X ~ B(30, 0.6)
*np = 30*0.6 = 18*
*nq = 30*0.4 = 12*

np >= 5,
nq >= 5

So, binomial distribution is symmetric.

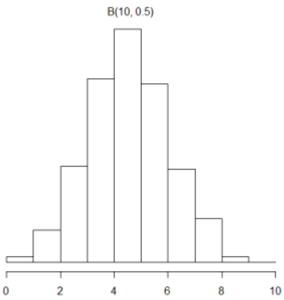


Case 2:

X ~ B(10, 0.5)
*np = 10*0.5 = 5*
*nq = 10*0.5 = 5*

np >= 5,
nq >= 5

So, binomial distribution is symmetric.

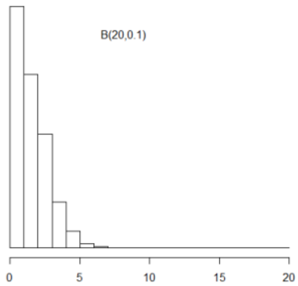


Case 3:

X ~ B(20, 0.9)
*np = 20*0.9 = 18*
*nq = 20*0.1 = 2*

np >= 5,
nq <= 5 (less)

So, binomial distribution is asymmetric.

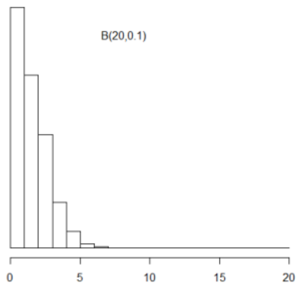


Case 4:

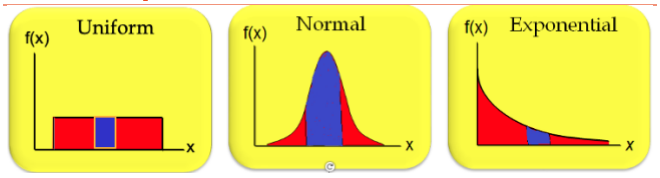
X ~ B(20, 0.1)
*np = 20*0.1 = 2*
*nq = 20*0.9 = 18*

np <= 5, (less)
nq >= 5

So, binomial distribution is asymmetric.



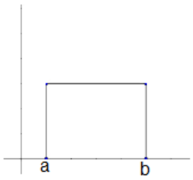
Probability = Area Under the Curve



- X can assume any value in an interval on the real line or in an interval.
- P(X = a) = 0, for any number a
  This is because, at a point the area will be 0
- P(a < X < b) = P(a ≤ X < b) = P(a < X ≤ b) = P(a ≤ X ≤ b)
  This is because, even if we consider or ignore the end points, the area remains the same
- P(a < X < b) is the area under the graph of the probability density function between a and b.
- For Normal distribution the end points goes to infinity on both sides

Uniform Probability Distribution U(a,b)

$$\boldsymbol{f(x) = \begin{cases} \frac{1}{b-a} & for\ a \le x \le b \\ 0 & else\ where \end{cases}}$$



Length = b-a, so **to** make the probability of 1, breath = 1/(b-a)
Where, b > a

$$E(U) = \frac{a+b}{2}$$

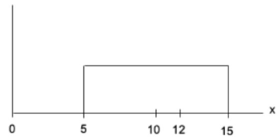
$$V(U) = \frac{(b-a)^2}{12}$$

So, considering $x_2 > x_2$

$$\boldsymbol{P(x_1 < U(a,b) < x_2) = \frac{x_2 - x_1}{b-a}}$$

Example:
Kumar is usually late to office. HR believes that his arrival time is uniformly distributed between 5m late and 15m late.
What is the probability density function?
What is the probability that today he will be late by at least 12m?
What is the average time he is late by? The variance?

Here,
a=5
b=15



So,

$$\boldsymbol{f(x) = \begin{cases} \frac{1}{10} & for\ 5 \le x \le a \\ 0 & else\ where \end{cases}}$$

Late by at least 12, x1=12 & x2=15, so,

$$P(12 < U(5{,}15) < 15) = \frac{15-12}{15-5} = \frac{3}{10} = 0.3$$

Average Time, $E(U) = \frac{5+15}{2} = 10$

Variance, $V(U) = \frac{(15-5)^2}{12} = \frac{100}{12} = 8.33$

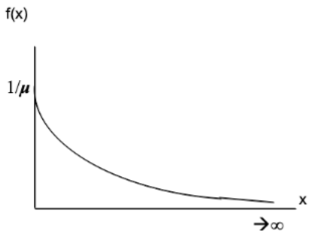
The Exponential Distribution

Exponential Probability Distribution – exp(μ)

- exp(μ) is useful in modeling
  - Time between vehicle arrivals at a toll booth
  - Time required to complete a questionnaire
  - Distance between major defects in a highway

- ***Exponential Distribution*** and the ***Poisson*** Distribution are related
  - Average time between vehicle arrivals is μ = 5m = 1/12 h
  - Average number of vehicles arriving in 1 hour is 12 (<-- μ for Poisson)

$$f(x) = \left(\frac{1}{\mu}\right) e^{-x/\mu}$$



Where,
μ *: Avg interval time*
: Avg time between events
1/ μ *: Arrival rate*

exp(μ) Properties

- μ is the average waiting time
- The ***mean and standard deviation*** are equal.
- The exponential distribution is ***skewed to the right***.
- $$P(X < x) = 1 - e^{-x/\mu}$$
- The distribution is memoryless!

Example 1:
At the Petrol Bunk,
The time between arrivals of cars follows an exponential probability distribution with a mean time of 3 minutes.
What is the probability that the time between two successive arrivals will be 2 minutes or less?
The clock is reset to 0, three minutes after the second car arrives. What is the probability that the waiting time for the next arrival is 2 minutes or less?!
Here,

μ=3
x=2

$$\mathrm{P(X < 2) = 1 - e^{-\frac{2}{3}} = 1 - 0.5135 = 0.4865}$$

The clock is reset after 3 minutes of second car arrival.
So, as per the property the distribution is memoryless.

Hence,
P(X < 2) = 0.4865

Example 2:
Suppose the rate at which cars cross the toll booth is 10 cars/h, and the arrival process can be described by a Poisson Distribution. Write down the Poisson & Exponential distributions that describe the process.
Here,
For Poisson distribution
X: # of cars that cross the toll booth

μ=10 ***cars/hr***

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} = \frac{10^x e^{-10}}{x!}$$

For Exponential distribution,
Y: Inter-arrival time : Time between cars

μ=1/10 ***hr/car***

$$exp(x) = \left(\frac{1}{\mu}\right) e^{-x/\mu} = 10e^{-10x}$$

Example 3:
Suppose calls on your cell phone follow an exponential distribution with the average time between calls being 10m. What are the Poisson & Exponential distributions that describe the process? (For the Poisson distribution, take the time period to be 1 h.) Find the probability that there will be no calls in the next 1 hour.
Here,
For Poisson distribution
X : # of calls in 1 hr
1 call in 10 minutes = 1/6 hour
So, in 1 hour, 6 calls

μ=6 ***calls/hr***

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} = \frac{6^x e^{-6}}{x!}$$

For no calls, x=0

$$P(X = 0) = \frac{6^0 e^{-6}}{0!} = e^{-6} = 0.002479$$

For Exponential distribution,
Y : Time interval between calls

μ=10 ***mins/call***

$$exp(x) = \left(\frac{1}{\mu}\right) e^{-x/\mu} = 0.1e^{-0.1x}$$

For no calls in next 1 hour, x>60 minutes

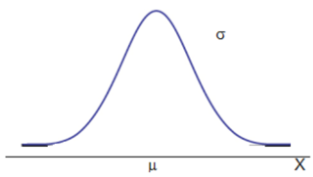
$$P(X > 60) = 1 - P(X \leq 60) = 1 - 1 - e^{-x/\mu}$$
$$= e^{-60/10} = 0.002479$$

Normal Probability Distribution N(μ, σ)

It is widely used in statistical inference.

It has been used in a wide variety of applications including:

- Heights of people
- Rainfall amounts
- Test scores



Equation of Normal distribution curve:

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

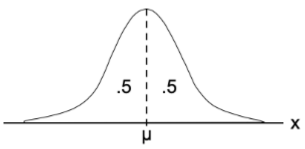
- Normal curve with mean=0 and standard deviation=1 is called z curve, or standard normal distribution.
- Other names- Gaussian, Bell curve, and Law of error.
- It is a continuous distribution- fractional values like 6.66, etc. (distance, temperature) are allowed on X-axis.
- Area under the curve is 1 (that is, probability of all events=1).
- The curve ranges from –infinity to +infinity.
- For Normal distribution, Mean=Median=Mode.

We will use this extensively while describing

- Distribution of sample mean
- Distribution of a sample proportion

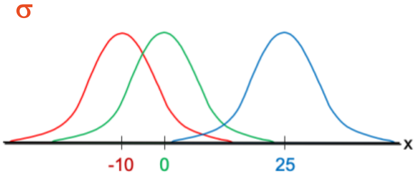
N(μ, σ) – Properties

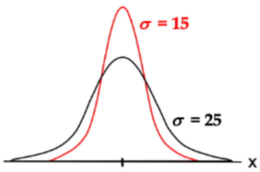
- The distribution is Symmetric So, probability of either side of the mean is 0.5
- The entire family of normal probability distributions is defined by its μ and its σ
- Z ≡ N (0,1) is the standard normal distribution



Change in Mean – μ

Change in Standard Deviation - σ





Normal Distribution using MS Excel

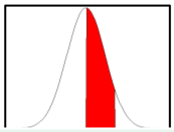
Find area from – infinity to X

= NORM.DIST(X, Mean, Sd, TRUE)*100

Find X, when area from infinity is given

= NORM.INV (Area, Mean, Sd)

Reading the Table for N(0,1)



NORMAL TABLES

Gives the area between Z = 0 and the specified Z-value

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.10	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.20	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.30	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.40	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.50	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.60	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.70	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.80	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.90	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389

$$P(0 < Z < 0.5) = 0.1915$$
$$P(0 < Z < 0.03) = 0.0120$$
$$P(0 < Z < 0.53) = 0.2019$$
$$P(0 < Z < 0.87) = 0.3078$$
$$P(Z < 0) = P(Z \leq 0) = 0.5$$
$$P(Z > 0) = P(Z \geq 0) = 0.5$$
$$\begin{aligned} P(Z < 0.85) &= P(-\infty < Z < 0.85) \\ &= P(Z < 0) + P(0 < Z < 0.85) \\ &= 0.5 + 0.3023 = 0.8023 \end{aligned}$$
$$\begin{aligned} P(Z < -0.85) &= P(-\infty < Z < -0.85) \\ &= P(Z < 0) - P(0 > Z > -0.85) \\ &= P(Z < 0) - P(0 < Z < 0.85) \\ &= 0.5 - 0.3023 = 0.1977 \end{aligned}$$
$$\begin{aligned} P(0.5 < Z < 0.85) &= P(Z < 0.85) - P(Z < 0.5) \\ &= 0.8023 - 0.1915 = 0.1108 \end{aligned}$$
$$P(-0.85 < Z < -0.5) = 0.1108$$

Transforming N(μ, σ) to N(0, 1)

X ~ N (μ, σ)

$$\mathbf{z = \frac{x - \mu}{\sigma}}$$

Z ~ N(0, 1)

z is the number of standard deviations x is from μ.

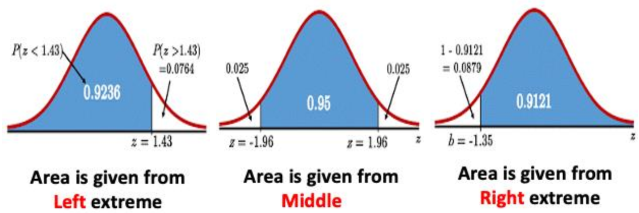
Example:

Given, X ~ N(4,2)

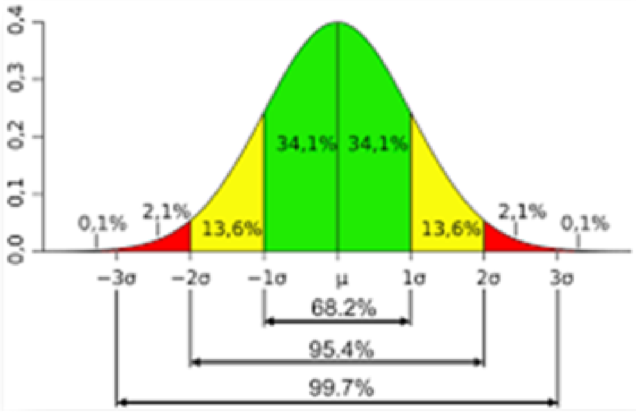
$$\begin{aligned} P(X > 5) &= P\left(\frac{x-4}{2} > \frac{5-4}{2}\right) = P(Z > 0.5) \\ &= P(Z > 0) - P(Z < 0.5) \\ &= 0.5 - 0.1915 = 0.3085 \end{aligned}$$

$$P(4 < X < 5) = P\left(\frac{4-4}{2} < \frac{x-4}{2} < \frac{5-4}{2}\right) = P(0 < Z < 0.5 = 0.1915$$

Three types of Tables



68-95-97 Rules from centre



- 68% within 1 standard deviations.
- 95% within 2 standard deviations.
- 99% within 3 standard deviations.

SK rule-

- 0.1 – 2 – 14 - 34 rule, from left side.
- 0.1 – 2.1 – 13.6 – 31.4

For 6-sigma enthusiasts (from Centre)

- 68.27% within 1 σ
- 95.50% within 2 σ
- 99.73% within 3 σ
- 99.994% within 4 σ
- 99.999 94% within 5 σ
- 99.999 999 8% within 6 σ
- 99.999 999 999 7 within 7 σ

Example:

The Engineering College has been conducting an entrance exam for the last 50 years. The scores on this test are normally distributed with a μ = 50 and a σ = 10. Kumar believes that he must do better than at least 80% of those who take the test. Kumar scores 58. Will he be selected?

Here, Given

X : Score of a randomly selected student

X~N(50, 10)

P(selected student scored less than 58),

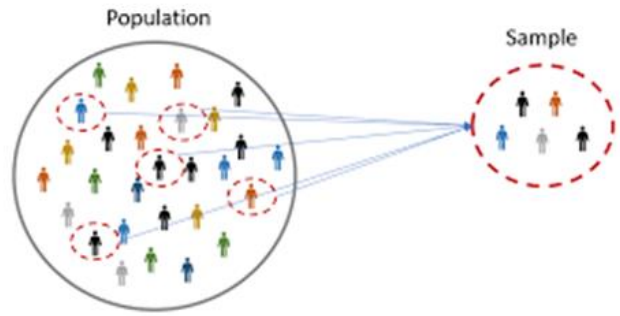
$$P(X < 58) = P\left(\frac{x - 50}{10} < \frac{58 - 50}{10}\right) = P(Z < 0.8)$$
$$= P(Z < 0) + P(0 < Z < 0.8)$$
$$= 0.5 + 0.2881 = 0.7881$$

This means, 78.8% students scored 58 or more marks but Kumar was hoping to score more than 80% of the students.

Confidence Interval Estimation

Parameters for Population
Proportion, π
Population size, N
Population mean, μ
Population Standard Deviation σ

Statistic for Sample
Proportion, p
(if 5% items are found defective in a sample of 100, then p=0.05)
Sample size, n
Sample mean, X'
Sample Standard Deviation. S



Population - Variance/Std Dev

Mean,

$$\mu = \frac{\sum x_i}{N}$$

Variance,

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Standard Deviation,

$$\sigma = \sqrt{\sigma^2}$$

For population, variance is divided by N
Excel Formula:
Population variance, =Var.p(range)
Population standard deviation, =Stdev.p(range)

Sample – Variance/Std Dev

Mean,

$$\bar{X} = \frac{\sum x_i}{n}$$

Variance,

$$s^2 = \frac{\sum (x_i - \bar{X})^2}{n-1}$$

Standard deviation,

$$s = \sqrt{s^2}$$

For sample, variance is divided by n-1
Excel Formula:
Sample variance, =Var.s(range)
Sample standard deviation, =Stdev.s(range)

Estimating Population Parameters

Population mean

$$= \textit{Sample mean} \pm \textit{Sampling Error}$$

$$= X' \pm \textit{Critical value} * \frac{s}{\sqrt{n}}$$

Population Proportion

$$= \textit{Sample proportion} \pm \textit{Sampling Error}$$

$$= p \pm \textit{Critical Error} * \sqrt{\frac{p(1-p)}{n}}$$

Where,
s – Std deviation of sample (divide by n-1)
p – proportion found in the sample
n – sample size
Critical Value – t table (mean), z table (proportion)

$$\textit{Standard Error of mean} = \frac{s}{\sqrt{n}}$$

$$\textit{Standard Error of proportion} = \sqrt{\frac{p(1-p)}{n}}$$

To estimate, population mean – use t table
To estimate proportion – use Z table

To use t table, degree of freedom, df=sample size – 1

Example:
The weights of 5 potatoes drawn randomly from a consignment are 136, 58, 79, 48 and 46 g.

						Size, n	Mean, X'	Stdev, S
Sample-1	136	58	79	48	46	5	73.4	37.4

Estimate average weight (mean) of potatoes in the consignment.

SE= S/√n	$t_{90\%}$ value	Estimate, Mean ± $t_{90\%}$ * SE
16.7	2.132	73.4 ± 35.6 or, 37.8 to 109

Estimates from 5 samples taken from same consignment

						Size, n	Mean, X'	Stdev, S	SE= S/√n	$t_{90\%}$ value
Sample-1	136	58	79	48	46	5	73.4	37.4	16.7	2.132
Sample-2	73	58	79	48	46	5	60.8	14.8	6.6	2.132
Sample-3	115	58	79	48	46	5	69.2	28.8	12.9	2.132
Sample-4	64	58	79	48	46	5	59.0	13.4	6.0	2.132
Sample-5	87	58	79	48	46	5	63.6	18.5	8.3	2.132

Estimate, Mean ± $t_{90\%}$ * SE
73.4 ± 35.6 or, 37.8 to 109
60.8 ± 14.1 or, 46.7 to 74.9
69.2 ± 27.4 or, 41.8 to 96.6
59 ± 12.8 or, 46.2 to 71.8
63.6 ± 17.6 or, 46 to 81.2

FORMULAS

Mean for Population:

$$\overline{X} = \frac{\sum x_i}{n}$$

Mean for Sample:

$$\mu = \frac{\sum x_i}{N}$$

Percentiles:

$$i = \left(\frac{p}{100}\right) n$$

if, i ≠ integer, xroundup(i)
else, Avg(xi, x(i+1))

Percentile Rank

$$\% \, rank = \left[\frac{(\#of \ values \ below \ x) + 0.5}{Total \ \# \ of \ values}\right] \times 100$$

Range

Range = Max Val – Min Val

Inter Quartile Range

IQR = Q3 – Q1

Variance for population,

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Variance for sample,

$$s^2 = \frac{\sum (x_i - \overline{X})^2}{n-1}$$

Z-Score

$$z = \frac{x - \overline{X}}{s}$$

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

Chebyshev's Theorem

$$(1 - 1/z^2)$$

Probability

Not mutually exclusive events,

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

For mutually exlusive events,

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

For not independent events,

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

For independent events,

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

$$\text{Probability} = \frac{No. of \ outcomes \ in \ which \ the \ event \ occurs}{Total \ no. of \ possible \ outcomes}$$

Bayes' Theorem

$$P(E_i|F) = \frac{P(E_i)P(F|E_i)}{P(E_1)P(F|E_1) + P(E_2)P(F|E_2) + \cdots + P(E_n)P(F|E_n)}$$

$$P(E_i|F) = \frac{P(E_i)P(F|E_i)}{P(E_1 \cap F) + P(E_2 \cap F) + \cdots + P(E_n \cap F)}$$

$$P(E_i|F) = \frac{P(E_i)P(F|E_i)}{P(F)}$$

Discrete Probability Distribution

$$0 \le f(x) \le 1$$

$$\sum f(x) = 1$$

$$E(X) = \mu = \sum Xf(X)$$

$$Var(X) = V(X) = \sigma^2 = \sum (X - \mu)^2 f(X)$$

$$E(\alpha X) = \alpha E(X)$$

$$V(\alpha X) = \alpha^2 V(X)$$

$$E(X \pm c) = E(X) \pm c$$

$$V(X \pm c) = V(X)$$

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$$

If X& Y are independent,

$$V(\alpha X + \beta Y) = \alpha^2 V(X) + \beta^2 V(Y)$$

Poisson Probability Distribution Π(μ)

I: The specified interval
X = the number of occurrences in an interval
f(x) = the probability of x occurrences in an interval
μ = mean number of occurrences in an interval

X ~ Π(μ)

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!} = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$\mu = \lambda = E(X) = V(X)$$

Binomial Probability Distribution B(n,p)

X ~ B(n, p)
p: Probability of Success
q: Probability of Failure (q=1-p)

$$\frac{n}{N} < 5\%$$

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{(n-x)} = C_x^n p^x q^{(n-x)}$$

$$= \frac{n!}{x!\,(n-x)!} p^x q^{(n-x)}$$

$$\mu = np$$

$$\sigma^2 = npq$$

$$\sigma = \sqrt{npq}$$

Uniform Probability Distribution U(a,b)

$$f(x) = \begin{cases} \frac{1}{b-a} & for \ a \le x \le b \\ 0 & else \ where \end{cases}$$

$$E(U) = \frac{a+b}{2}$$

$$V(U) = \frac{(b-a)^2}{12}$$

So, considering x2 > x2

$$P(x_1 < U(a,b) < x_2) = \frac{x_2 - x_1}{b-a}$$

Exponential Probability Distribution – exp(μ)

$$f(x) = \left(\frac{1}{\mu}\right) e^{-x/\mu}$$

$$P(X < x) = 1 - e^{-x/\mu}$$

Where, μ : Avg interval time
1/ μ : Arrival rate

Transforming N(m, s) to N(0, 1)

X ~ N (μ, σ)

$$\mathbf{z} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

Z ~ N(0, 1)

z is the number of standard deviations x is from μ.

Population Mean

$$\begin{aligned}\boldsymbol{Population\ mean} \\ &= \boldsymbol{Sample\ mean} \\ &\pm \boldsymbol{Sampling\ Error} \\ &= \boldsymbol{X'} \pm \boldsymbol{Critical\ value} * \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}\end{aligned}$$

Population Proportion

$$\begin{aligned}\boldsymbol{Population\ Proportion} \\ &= \boldsymbol{Sample\ proportion} \\ &\pm \boldsymbol{Sampling\ Error} \\ &= \boldsymbol{p}\ \pm \boldsymbol{Critical\ Error} * \sqrt{\frac{\boldsymbol{p(1-p)}}{\boldsymbol{n}}}\end{aligned}$$

Standard Error of mean

$$\boldsymbol{Standard\ Error\ of\ mean} = \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}$$

Standard Error of proportion

$$\boldsymbol{Standed\ Error\ of\ proportion} = \sqrt{\frac{\boldsymbol{p(1-p)}}{\boldsymbol{n}}}$$