Statistics

DESCRIPTIVE

It consists of organizing and summarizing of data

1. Measure of Central Tendency
   - i) Mean
   - ii) Median
   - iii) Mode
2. Measure of Dispersion
   - i) Variance
   - ii) Standard Deviation

INFERENTIAL

Sample data ← Collect Data (Z test, t test . . . { Using Some Experiments }

↓

→ Conclusions or Inferences

↓

Other data

↓

{ Population data }

Eg :- There is a College A → 1000 students → Avg height

↓ population data.

Class Statistics

Sample ← { 180cm, 170cm, 162cm, 150cm, 160cm }

Descriptive — Mean, Media, — Average = 165 ✓

Inferential :- Conclusion, Inferences }

Random Variables

$X \longrightarrow$ function $\longrightarrow$ Values

$\begin{cases} y = 5x + 2 \\ y = 5 \cdot 1 + 2 = 7 \end{cases}$ $x=1, x=2, x=3$

$\Downarrow$

Process or Experiments

$X = \begin{cases} 0 & H \checkmark \\ 1 & T \checkmark \end{cases}$ Tossing a Coin

$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$ Rolling a fair dice

Eg: Height of the people attending the event tomorrow

Eg: 150cm, 160cm, 160.1cm, - - - -

Random Variables

Discrete Random

Eg: Tossing a Coin
Rolling a dice

Continuous Random Variable

Eg: Tomorrow how many inches it is going to rain

[0, 1.1, 5.5, 10.5, 10.75]

Sample Variance

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

n = Sample data

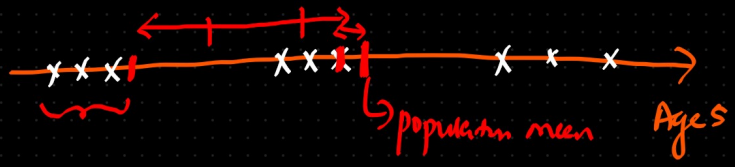
Sample Mean
Population Mean

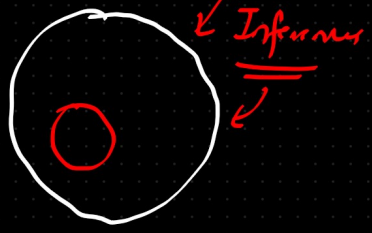
↑↑↑↑ {Imp interview question}

→ n-1 → ??

n

{Unbiased Estimation}





{ n-2 , n-3, n-4 }

Experimentation {n-1} ✓

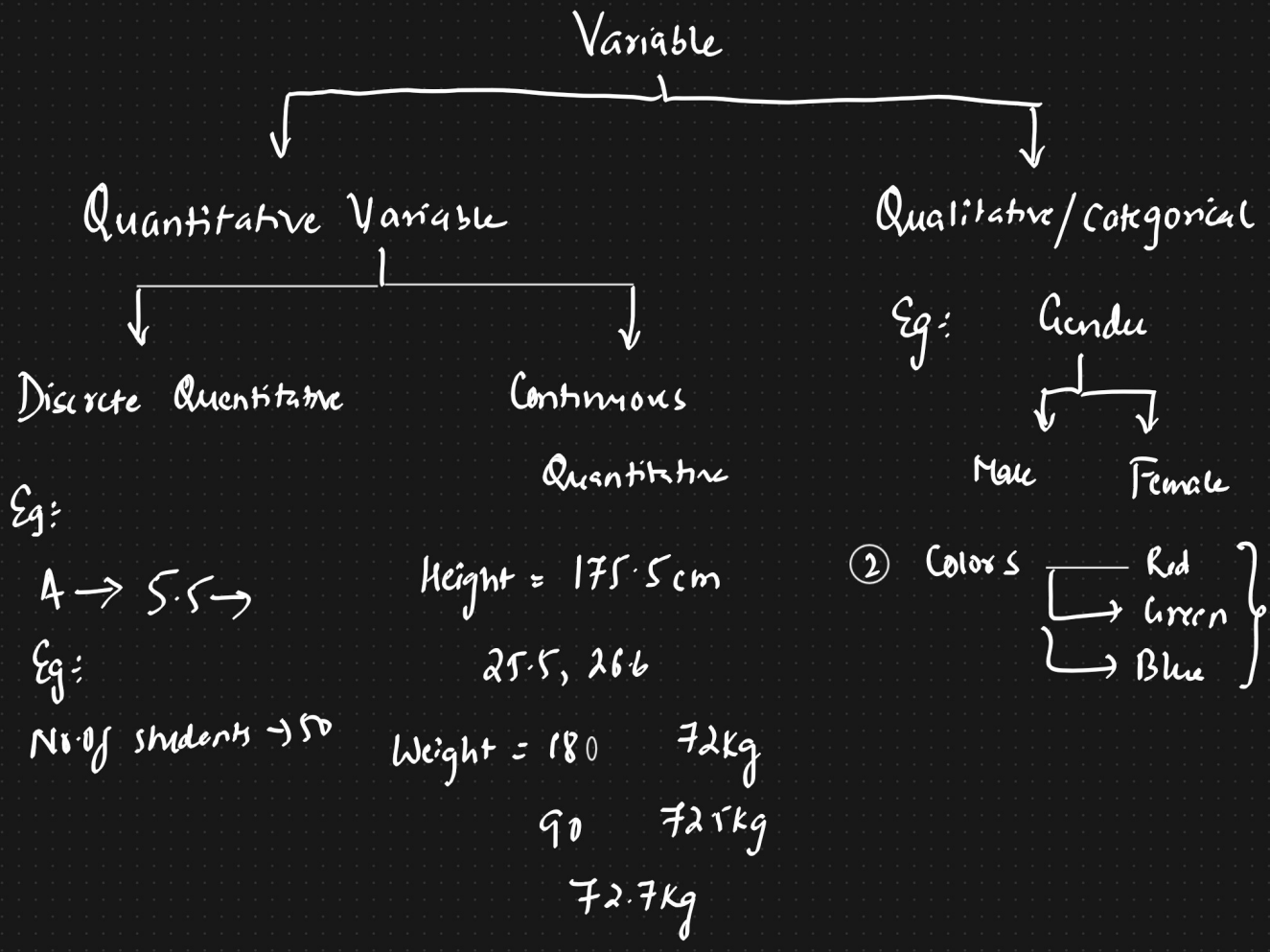
Variable

Defn : Variable is a property that can take up any value

{Age = 25} Gender = Male Height = 7.2

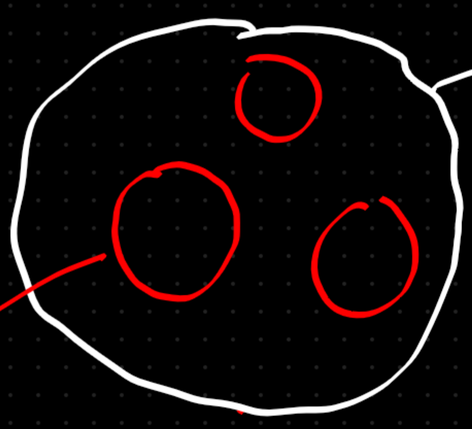
Ages = [12, 20, 25, 36, 70] ✗ Variable

Different types of Variable.



Population And Sample

100kg
70kg



100K → Population ↓
↳ N

① Exit Polls

10K
↳ Sample → n

Collect all the weights of all the people

Measure of Central Tendency

① Mean

② Median

③ Mode

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad (N \checkmark)$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (n \checkmark)$$

Ages = {1, 3, 4, 5} → Distribution

$$\mu = \frac{1+3+4+5}{4} = \frac{13}{4} = 3.25$$

$\mu \downarrow \quad \bar{x}$

② Median

Ages = {1, 3, 4, 5, 100} {Outlier} ← Big Number

$$\mu = \frac{1+3+4+5+100}{5} = \frac{113}{5} = \boxed{22.6}$$

3.25

Ages = {4, 3, 1, 5, 100} → odd

↳ Sort the numbers {1, 3, [4, 5], 100, 200} → Even

↳ Median → 4

$$\frac{4+5}{2} = 4.5$$

Good

Measure of Dispersion

① Variance

② Standard Deviation

Age1 = {2, 2, 4, 4} → Distr 1

$\mu = \frac{2+2+4+4}{4}$

$= 3$

{Spread is less}

{Spread is more}

Age 2 = {1, 1, 5, 5} → Distr 2

$\mu = \frac{1+1+5+5}{4}$

$= 3$

① Variance

Population Data {N} → Population Size

Sample Data {n} → Sample Mean ($\bar{x}$)

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

x_i ↑

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

Imp Interview

Age 1 = {2, 2, 4, 4}

$\mu = \frac{2+2+4+4}{4} = 3$

x_i	μ	$(x_i-\mu)^2$
2	3	1
2	3	1
4	3	1
4	3	1

$N = 4$

$\sigma^2 = 4/4 = 1$

Age 2 = {1, 1, 5, 5}

$\mu = \frac{1+1+5+5}{4} = 3$

x_i	μ	$(x_i-\mu)^2$
1	3	4
1	3	4
5	3	4
5	3	4

Σ $16/4 = 4$

$N = 4$

Revising All we have learnt.

Population

$$\mu = \frac{\sum_{i=1}^{N} (x_i)}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

How far a Data point is away from the mean.

$\sqrt{\sigma^2} \rightarrow$ Population Standard deviation

$$\sigma = \sqrt{\text{Population Variance}}$$

Sample

$$\bar{x} = \frac{\sum_{i=1}^{n} (x_i)}{n}$$

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

$$s = \sqrt{\text{Sample Variance}}$$

↓

Sample Standard Deviation

Covariance And Correlation

Covariance and correlation are two statistical measures used to determine the relationship between two variables. Both are used to understand how changes in one variable are associated with changes in another variable.

Covariance

Definition: Covariance is a measure of how much two random variables change together. If the variables tend to increase and decrease together, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative.

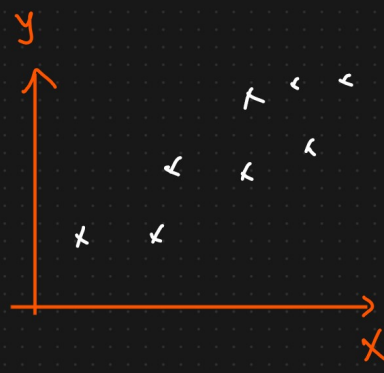
[Quantify the Relationship between X and Y]

X	Y
2	3
4	5
6	7
8	9

X↑	Y↑
X↓	Y↑
X↑	Y↓
X↓	Y↓

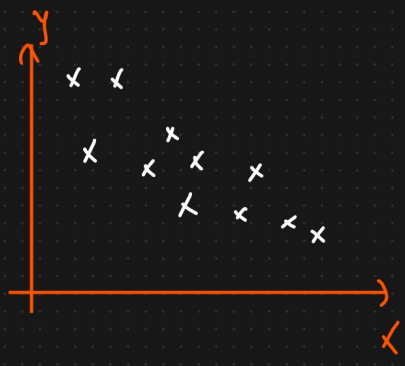
Dataset

↓ ↑ Size of House	Price ↑ ↓
1200	45 lakhs
1300	50 lakh
1500	75 lakh



X↑	Y↑
X↓	Y↓

⇒ +ve Covariance ⇒ +ve value



X↓	Y↑
X↑	Y↓

X	Y
7	10
6	12
5	14
4	16

⇒ -ve Covariance ⇒ -ve value

Covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\Rightarrow Cov(x,x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

$$\boxed{Cov(x,x) = Var(x)}$$

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

$x_i \rightarrow$ Datapoint of random variable x

$\bar{x} \rightarrow$ Sample mean of x

$y_i \rightarrow$ Datapoints of random variable y

$\bar{y} \rightarrow$ Sample mean of y

Students

Hour studied (X)	Exam Score (Y)
2	50
3	60
4	70
5	80
6	90

$x\uparrow \; y\uparrow$ $x\downarrow \; y\downarrow$ $\Rightarrow$ +ve covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

① $\bar{x} = \frac{2+3+4+5+6}{5} = 4$ //

② $\bar{y} = \frac{50+60+70+80+90}{5} = 70$ //

$$Cov(x,y) = \frac{(2-4)(50-70)+(3-4)(60-70)+(4-4)(70-70)+(5-4)(80-70)+(6-4)(90-70)}{4}$$

$Cov(x,y) = 20$ //.

$\Rightarrow$ The positive covariance indicates the no. of hours studied increased the exam score also.

X	Y
7	10
6	12
5	14

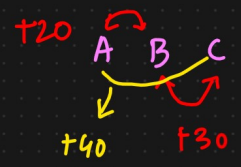
$\Rightarrow$ -Ve //.

$x\uparrow \; y\downarrow$ $x\downarrow \; y\uparrow$

0.96 Cov(A, B) 0.88 Cov(B, C)

-200 +100 20 = Cov(A, B)

-300 +300 30 = Cov(B, C)



Advantages

-1 to 1

Disadvantage

① Quantify the Relationship between X and Y

① Covariance does not have a specific limit value.

$$Cov(X,Y) \Rightarrow -\infty \text{ to } \infty$$

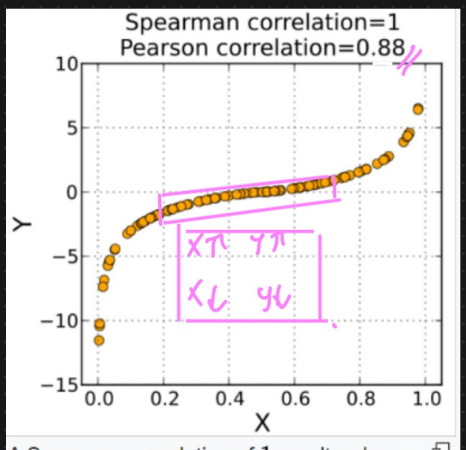
② Correlation
- → Pearson Correlation Coefficient
- → Spearman Rank Correlation

① Pearson Correlation Coefficient $\Rightarrow [-1 \text{ to } 1]$

$$\rho_{x,y} = \frac{Cov(x,Y)}{\sigma_x \cdot \sigma_y} = \frac{20}{\sigma_x \cdot \sigma_y} \Rightarrow 0 \text{ to } 1$$

(Y) The more the value towards +1 the more +ve correlated X & Y is.

(K) The more the value towards -1 the more -ve correlated it is (X,Y)

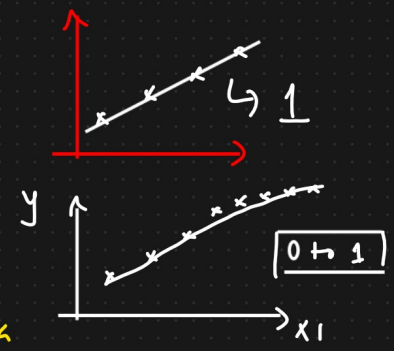


A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater x values than that of a given data point will have greater y values as well. In contrast, this does not give a perfect Pearson

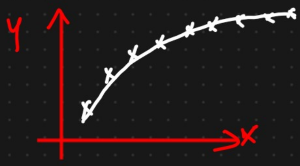
⇒ X↑ Y↑
⇒ X↓ Y↓

Pearson Correlation = 0.88

② Spearman Rank Correlation



Pearson Correlation for non linear data



x	y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5
0	7	1	4

$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

Feature Selection

Size of House ↑

No. of Room ↑ (+ve correlated)

Location ↑

~~No. of People staying in the house~~ (≈ 0) → No → Price

Haunted ↔ Price (−ve correlation)

Price ↑↑ (o/p) ↑↓ ⇅ %

Histograms

{ Kernel Density Estimation }

Ages

$X = \{23, 24, 25, 30, 34, 36, 40, 50, 60, 75, 80\}$

20-30 = 4

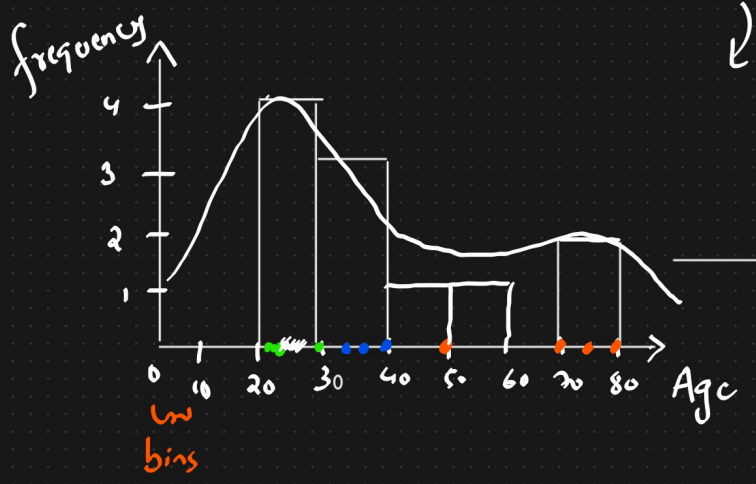
30-40 = 3

41-50 = 1

51-60 = 1

61-70 = 0

71-80 = 2



↳ Histograms

→ Smoothen the Histogram

Percentiles And Quartiles

Percentage : $\{1, 2, 3, 4, 5, 6\}$

No. of odd numbers = 3

Percentage of odd numbers in this group $= \frac{3}{6} \times 100 = 50\%$

Percentiles: A percentile is a value below which certain percentage of observations lie.

$\{2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 9, 9, 10\}$ $n = 9$ $\frac{3+4}{2} = 3.5$

Percentile of value $x = \frac{\#\text{ of values below } x}{n} \times 100$

$= \frac{11}{14} \times 100$

$= 78.57\%$ of value 9

Percentile Ranking

25% is 3.75

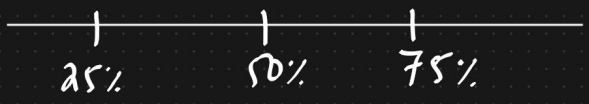
$\Rightarrow$ Value $= \frac{\text{Percentile}}{100} \times (n+1)$

$= \frac{25}{100} \times (15)$

$= 3.75 \approx 3.5$

② Quartiles

25% = 1st Quartile
50% = 2nd Quartile
75% = 3rd Quartile



5 Number Summary

1) Minimum ✓
2) 1st Quartile (25%)
3) Median
4) 3rd Quartile (75%)
5) Maximum

Eg: 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 → outlier

[lower fence ⟷ Higher fence]

lower fence = Q1 − 1.5(IQR)

Higher fence = Q3 + 1.5(IQR)

(25%)

$Q1 = \frac{\text{Percentile}}{100} \times (n+1) = \frac{25}{100} \times (20) = 5^{th}$ position = 3 ✓

(75%)

$Q3 = \frac{75}{100} \times (20) = 15^{th}$ position = 7 ✓

outlier

[−3 ⟷ 13]

IQR = Q3 − Q1 = 7 − 3 = 4 ✓

lower fence = Q1 − 1.5(IQR)

= 3 − 1.5(4)

= 3 − 6 = −3

Higher fence = Q3 + 1.5(IQR)

= 7 + 1.5(4)

= 13 //

Eg: 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

Box plot ✓✓

Minimum = 1

1st Quartile = 3

Median = 5

3rd Quartile = 7

Maximum = 9

Probability

1. Introduction ✓
2. Addition Rule (For mutually exclusive event)
3. Addition Rule (For non mutually exclusive event)
4. Multiplication Rule (Independent & Dependent Events)

① **Probability** : It is about determining the likelihood of an event

Eg : Toss a Coin $\{H, T\}$

$$Pr(H) = \frac{1}{2} = 50\%$$

$$Pr(T) = \frac{1}{2} = 50\%$$

Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$$Pr(x=1) = \frac{1}{6}$$

Mutual Exclusive Event

Two events are Mutual exclusive if they cannot occur at the same time

Eg : Tossing a Coin

$Pr(H) = \frac{1}{2}$ $Pr(T) = 1/2$

$Pr(H \text{ or } T) = Pr(H) + Pr(T)$ {Additive Rule for mutual Exclusive Event}

$= \frac{1}{2} + \frac{1}{2} = 1$

Eg: Rolling a dice $\{1,2,3,4,5,6\}$

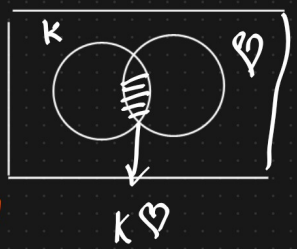
$Pr(1 \text{ or } 5) = Pr(1) + Pr(5)$

$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$ //

*) Non Mutual Exclusive Events

Eg: Taking a card from the deck

$Pr(K \text{ or } \heartsuit) = Pr(K) + Pr(\heartsuit) - P(K \text{ and } \heartsuit)$



$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$

↳ Non Mutual Exclusive Event

$= \frac{17}{52} - \frac{1}{52}$

$Pr(K \text{ or } \heartsuit) = \frac{16}{52}$ //

Probability

1. Introduction ✓
2. Addition Rule (For mutually exclusive event)
3. Addition Rule (For non mutually exclusive event)
4. Multiplication Rule (Independent & Dependent Events)

① **Probability** : It is about determining the likelihood of an event

Eg : Toss a Coin $\{H, T\}$

$$Pr(H) = \frac{1}{2} = 50\%$$

$$Pr(T) = \frac{1}{2} = 50\%$$

Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$$Pr(x=1) = \frac{1}{6}$$

Mutual Exclusive Event

Two events are Mutual exclusive if they cannot occur at the same time

Eg : Tossing a Coin

$Pr(H) = \frac{1}{2}$ $Pr(T) = 1/2$

$Pr(H \text{ or } T) = Pr(H) + Pr(T)$ {Additive Rule for mutual Exclusive Event}

$= \frac{1}{2} + \frac{1}{2} = 1$

Eg: Rolling a dice $\{1,2,3,4,5,6\}$

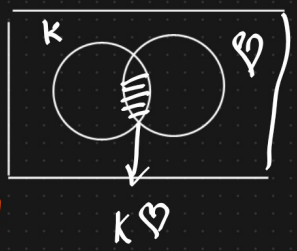
$Pr(1 \text{ or } 5) = Pr(1) + Pr(5)$

$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$ //

*) Non Mutual Exclusive Events

Eg: Taking a card from the deck

$Pr(K \text{ or } ♡) = Pr(K) + Pr(♡) - P(K \text{ and } ♡)$



$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$

Non Mutual Exclusive Event

$= \frac{17}{52} - \frac{1}{52}$

$Pr(K \text{ or } ♡) = \frac{16}{52}$ //

Multiplication Rule {Independent And Dependent Events}

2 events are Independent if they do not affect one another

Eg: Tossing a Coin {H and then Tails}

$Pr(H) = \frac{1}{2}$ $Pr(T) = \frac{1}{2}$

Eg: Rolling a dice

$Pr(1) = \frac{1}{6}$ $Pr(2) = \frac{1}{6}$

Dependent Events

2 Events are Dependent if they effect each other

Eg: Take a King from the deck and then the Queen Card from the deck

52 → K

$Pr(K) = \frac{4}{52}$ $\quad Pr(Q) = \frac{4}{51}$

Multiplication Rule

① Independent Event {Tossing a Coin}

$Pr(H \text{ and } T) = Pr(H) * Pr(T)$

$= \frac{1}{2} * \frac{1}{2}$

$= 1/4$

② Dependent Event → Conditional Probability

$Pr(K \text{ and } Q) = P(K) * Pr(Q/K)$

$= \frac{4}{52} * \frac{4}{51}$

$=$