Axioms of Probability:

i. $0 \leq P(A) \leq 1$

ii. $P(S) = 1$

iii. $P(A \cup B) = P(A) + P(B)$ if $(A \cap B) = \phi$

Prove that: $P(\phi) = 0$

$\phi \leq S$

$\therefore S \cup \phi = S$

$P(S) = P(S \cup \phi)$

$\Rightarrow S \cap \phi = \phi$

$= P(S) + P(\phi) = 1$

$1 + P(\phi) = 1$

$\therefore P(\phi) = 0.$

Equiprobability:

Equiprobability is a property for a collection of events that each have the same probability of occuring.

If there are n events under consideration, the probability of each occuring is $1/n$.

Example: Consider 4 Red balls 5 white balls

$P(S) = P(R_1 \cup R_2 \cup R_3 \cup R_4 \cup W_1 \cup W_2 \cup W_3 \cup W_4 \cup W_5) = 1$

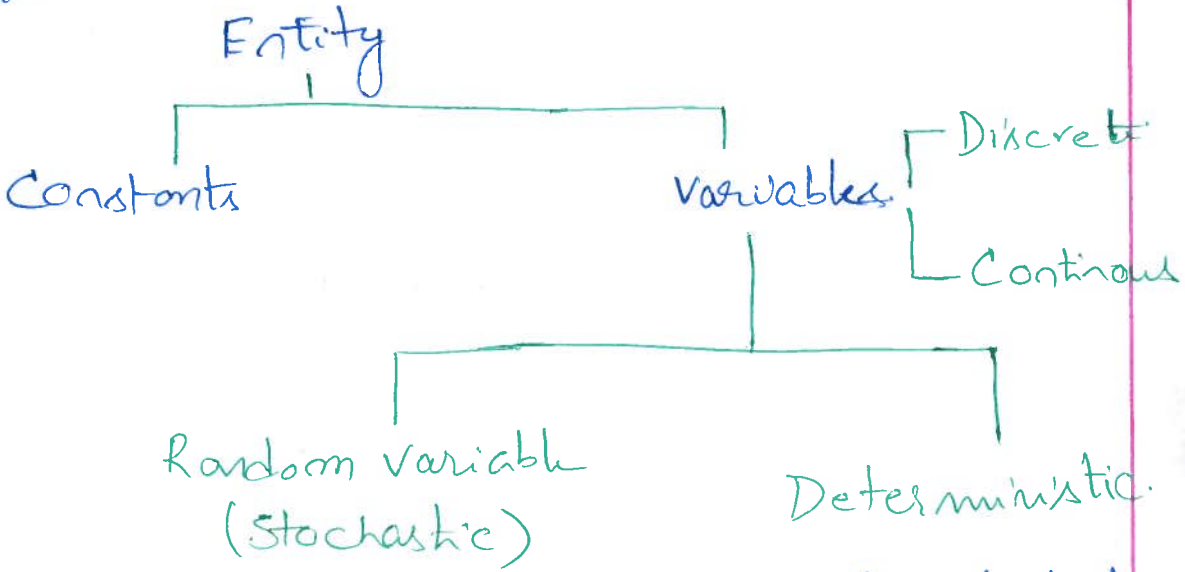
$P(R_1) + P(R_2) + P(R_3) + \cdots + P(W_4) + P(W_5) = 1$

$\frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = 1$

Relative frequency presumes probability is a rational number i.e p/q".

∴ Probability of picking a Red ball in above example is $4/9$.

Types of variables:



Deterministic: whose values can be predicted exactly.

Stochastic: whose values cannot be predicted exactly.

Stocastic variable is a variable whose value cannot be predicted but can be estimated with an associate probability. This is also known as Random variable.

Continous variable eg: Time taken tube light to stop walking.

Discrete variable eg: Number of tube lights in a hall.

Random variable X

$$P(a \le x \le b) = \int_a^b f(x)\,dx$$

It is a function when integrated with any two limits gives probability of * which lies between a and b.

Probability Density Function :Pdf: A function of a Continous random variable, whose integral gives the probability that the value of the variable lies with in the same interval.

Note: $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Problem:
Find the value of a where $f(x)$ is pdf.

$$f(x) = ax \qquad 0 \le x \le 1$$
$$= a \qquad 1 \le x \le 2$$
$$= -ax + 3a \qquad 2 \le x \le 3$$
$$= 0 \qquad \text{else where}$$

Sol:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_0^1 f(x)\,dx + \int_1^2 f(x)\,dx + \int_2^3 f(x)\,dx + \int_3^{\infty} f(x)\,dx$$

$$= 0 + \int_0^1 ax\,dx + \int_1^2 a\,dx + \int_2^3 (-ax+3a)\,dx + 0 = 1$$

apply integral we know

$$\int x^n\,dx = \frac{x^{n+1}}{n+1}.$$

$$= 0 + \frac{ax^2}{2}\Big|_0^1 + ax\Big|_1^2 + \frac{-ax^2}{2} + 3ax\Big|_2^3 + 0 = 1$$

$$= 0 + a\cdot\frac{1}{2} + (2a - a) + \left(-\frac{9a}{2} + \frac{4a}{2} + 9a - 6a\right) = 1$$

$$= \frac{a}{2} + a - \frac{9a}{2} + 2a + 9a - 6a = 1$$

$$= \frac{a}{2} + 2a - 9a + 2a + 9a - 6a = 1$$

$$= \frac{a + 4a + 6a - 5a}{2} = 1$$

$$= \frac{a - 2a}{2} \neq 1 \qquad \Rightarrow \frac{4a}{2} = 1$$

$$\Rightarrow a = 1/2$$

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

$$a = 1/2$$

Data Generating function: DGF:

We observe data at specific intervals of time to come to a business decision with certain amount of approximation.

Moment:

If $f(x)$ is a valid density function then $E[X] = \mu = \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx$ is called First Moment. It is also called Expectation of, Expected value of X.

Second Moment: If mean $\mu = 0$, the second

Moment is called variance.

$$V[x] = \int_{-\infty}^{\infty} x^r f(x)\, dx$$

Property of Moment

i) $E[cx] = c\,E[x]$

ii) $E[c] = c$

iii) $E[x+y] = E[x] + E[y]$

Bivariate distribution:

density function with two variables

Ex: temperature in Banglor

$f(x, y)$

$$P(a \le x \le b;\ c \le y \le d) = \int_a^b \int_c^d f(x, y)\, dx\, dy.$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

* Bivariate data essentially focuses on cause and effect relationship.

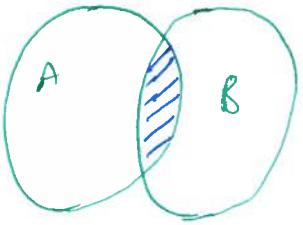
Ex If there is a rain in Bagmadala there is a flow in KRS, Dam, Mysore.

* Data is said to be univariate, If there is only one variable

* Multivariate data can have more than one variable

Ex: Meteorology w.r.t weather forcast in India considers 64 parameters.

Conditional Probability:



A and B are not mutually exclusive.

The probability of an event A, given that another event B has already occured.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{if and only if } P(A) \neq 0.$$

Note: It is not A divided by B. It is A pipeline B.

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

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

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

$$\therefore P(B|A) \cdot P(A) = P(A|B) \cdot P(B)$$

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

This is called Baye's Theorem.

Data Science:

* Population is all possible values
* Sample is part of population.

$$\bar{x}_{n_1, n_2, \ldots, n_{10}} = \sum_{i=1}^{n} \frac{n_i}{n}$$

"It is Unbiased estimate of population mean"

* Data generation function DGF for a Continous Random variable is measuring the value of Random variable at specific intervals of time and find a pattern in the observation.

* μ & σ for population, approximation for DGF is density function.

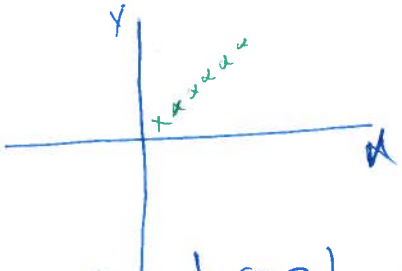
* $\bar{x}$ is a random variable because it varies from sampl to sample 2 to sample n, and it has its own distribution.

$$\text{Variance } s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1}$$

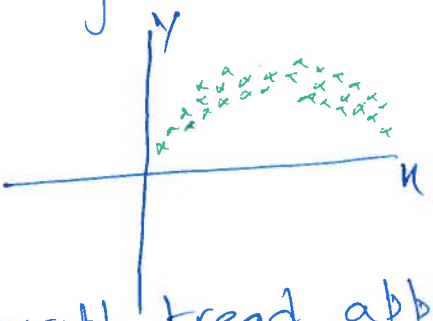
$\bar{x}$ is estimated value.

* Variance of population is σ^2

Scatter Plot:

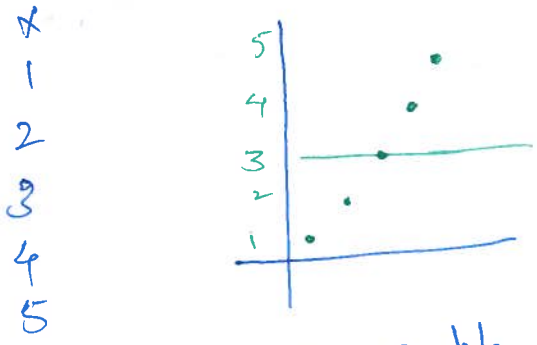


In the above graph trend appears to be linear.

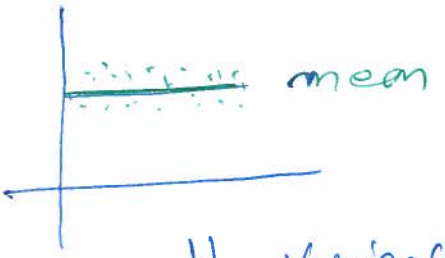


In the above graph trend appears to be non-linear.

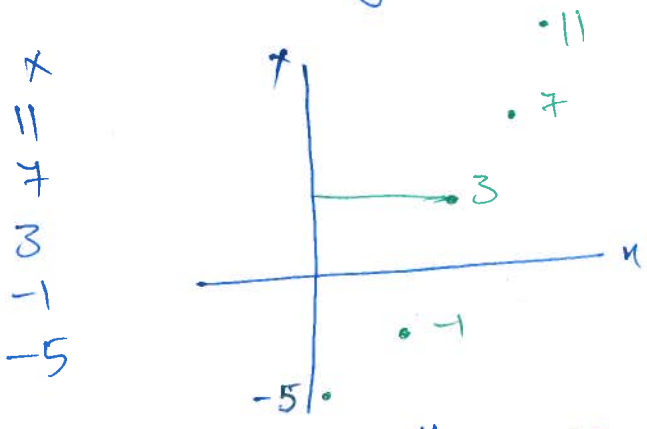
In the above graph there is no trend



In the above graph mean is 3 and range is 1 to 5.



In the above graph, Variance is distribution of data around mean.



In the above graph mean is 3; because of Data is scattered around mean and far away too Ex -5, 11 are away from mean 3. Hence Variance is high.

c'

$$s_x^2 = \text{Var}(x) = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

$$\text{covariance}(x, y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n-1}$$

Correlation Coefficient:

$$r = \frac{\text{cov}(x, y)}{s_x s_y} \qquad -1 \le r \le +1$$

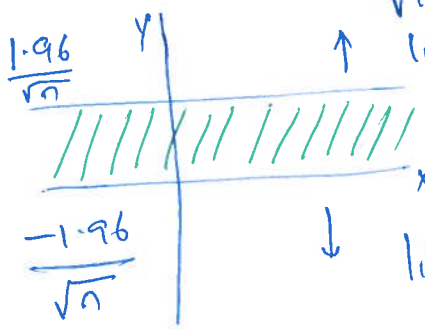
r indicates linear relationship of x and y

where $r = 1$ indicates positive relationship

$r = 0$ indicates

$r = -1$ indicates negative relationship.

Threshold value $= \frac{1.96}{\sqrt{n}}$



↑ linear relationship is "high"

↓ linear relationship is "low".