# Probability & Statistics

1) **Random variable** → A variable that can take any value from the sample space.

or

Can take an outcome as its value from the set of all the possible outcomes available.

2) **Sample space** → A set of all the possible outcomes of a random experiment.

3) **Outlier** → An extreme observation.

## Population & Sample size →

→ All the subjects for which an estimate need to be counted for.

ex: estimate the height of all the humans on earth.

$\mu$ → population mean.

→ It is often not feasible to estimate the statistic about all the subjects in the population.

**Sample** → A subset of a population that can be used to measure a statistic.

ex: Estimate the height of a few people in each state of a country.

$\bar{x}$ → Sample mean.

→ As the sample size increases, sample mean comes closer to population mean.

# Types of Events :-

1) **Mutually Exclusive :-** Events that can't happen at the same time.
   ex :- Sachin scoring a century & Sachin getting duck.

2) **Joint Events :-** Events that can happen together.
   ex :- Sachin scoring ½50 & India winning the match.

3) **Independent Events :-** Occurance of one event don't affect the occurance of other.
   ex :- outcome of sachin's performance & weather conditions.

4) **Exhaustive Events :-** Set of events that covers all the possible outcomes in the sample space.
   ex :- Sachin scoring ½50. & Sachin scoring < 50.

**Events & Set operations :-** 1) **Intersection :-** Both events happen. (Sachin scored ½50 & India won.)
   pd. merge(A, B, how='inner')

2) **union :-** either one of the events happen.
   pd.concat([eventA, eventB]).drop_duplicates()
   ex :- scored ½50 or India won.

3) **Compliment :-** Event don't happen.

**Probability Rules :-**

1) **Addition Rule :-** $P(A \cup B) = P(A) \cup P(B) - P(A \cap B)$

2) **Multiplication Rule :-** $P(A \cap B) = P(A|B)\,P(B)$

3) **Conditional Probability :-** $P(A|B) = \frac{P(A \cap B)}{P(B)}$

# Gaussian Distribution → (Normal Distribution)

1) Distribution → Very simple model of nature.

Given → 1) Type of distribution → One can easily
2) mean & variance model the probability/
↳ of a property frequency of occurrance
of each possible random variable of that property.

2) Why Gaussian distribution is important?

→ Most of the naturally occuring phenomena follow gaussian distribution.

ex → Height & weight of people.

3) Mathematical representation →

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

↳ quadratic power.
↳ exponential fn.

Simplified → standard normal
$\mu = 0, \sigma^2 = 1$

$$y = P(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{\left(-x^2/2\right)}$$

remove constants → Shape of fn don't change when constants are removed.

$$y = e^{-x^2} = \frac{1}{e^{x^2}}$$ → exponential with quadratic power.

Symmetricity

→ Due to quadratic exponentation $\frac{1}{e^{x^2}}$ always decreases.
if x changes unit → y reduces by 10 times.
x increases or decreases

ex → $P(x=1) = \frac{1}{e} = 0.3678$

$P(x=2) = \frac{1}{e^4} = 0.018$

$P(x=3) = \frac{1}{e^9} = 0.000123$

$\mu$ & $\sigma^2$ → Parameters of a gaussian dist.

↳ only these two are required to get the prob. density fn of a distribution.

→ Maximum probability is always at mean of the distribution.

→ Mean = = Mode.

CDF → Cumulative density fn.

Pdf is used when we want $P(X = x)$.

CDF is used we want $P(X \leq x)$ at a specific point or $P(X \geq x)$

↳ data points in a range.

→ relf of normal distribution is S shaped.

lower $\sigma^2$ → tighter shape.
higher $\sigma^2$ → loose shape.

Properties related to shape of a distribution →

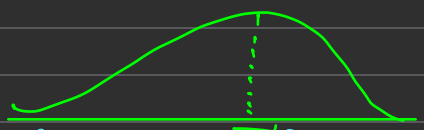
1) Symmetry → A distribution is symmetric if @ a point $\bar{x}$ about which the # of points left of $\bar{x}$ = = # of points right of $\bar{x}$.

mathematically → $f(x_0 - x_i) = = f(x_0 + x_i)$

→ Gaussian distribution is symmetric.

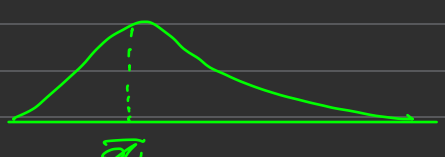
2) Skewness → Tailness of a distribution.

1) left (-ve) skewed → longer tail towards left side of mean.



2) right (+ve) skewed → longer tail towards right side of mean.

mathematically → $$\frac{1/n \sum (x_i - \bar{x})^3}{\left(1/(n-1) \sum (x_i - \bar{x})^2\right)^{3/2}}$$



3) **Kurtosis** → Another measure for fatness of a distribution.
→ used to detect outliers.

mathematically →

$$\text{Kurtosis}(x) = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum (x_i - \bar{x})^2\right)^2}$$

Gauss Kurtosis(x) = Kurtosis(x) − 3
(↓ relative measure. ↓ Kurtosis of gaussian distribution.

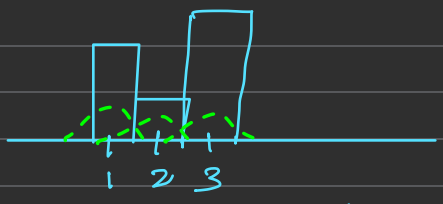
**Standard Normal** → $x_{im} = \frac{x_i - \mu}{\sigma}$ ∀ i ∈ (1 .. n)

Distribution where μ = 0 & σ = 1.
– Description of the distribution is clearer than normal distribution.

68 % of data within (−1 & 1) std. deviation.
95 % of data within (−2 & 2) std. deviation.
99 % of data within (−3 & 3) std. deviation.

**Kernel Density Estimate** → Convert histograms to probability density.

ex →




1) Draw a normal dist. around each point.
→ Distributions overlap for crowded points & are mutually exclusive for non-crowded points.

2) Kde(x) = Σ values of dist. overlapping at x.

value → ↑ higher probability → Points are crowded.
→ ↓ low probability → less crowded points.

# Sampling Distribution & Central Limit Theorem

Given → $X \to N(\mu, \sigma^2)$ population (& variance)

population mean

1) Collect $m$ samples of size $n \geq 30$ from the population.

$S_1, S_2 \ldots S_m$ total collection = $m \times n$ i.e. $30m$.

2) Get the mean of each sample.

$S_1 \to \bar{x}_1$, $S_2 \to \bar{x}_2 \ldots S_m \to \bar{x}_m$

3) Get the distribution of sample means →

$dist(\bar{x}_n)$ = Sampling distributions of sample means.

Central Limit Theorem → $dist(\bar{x}_n) \cong N(\mu, \sigma^2/n)$ $n \to \infty$.

The sampling distributions of sample means is normally distributed around the population mean with $\sigma^2/n$ variance.

Population (mean, variance) — can be computed by → Sampling distribution of sample means.

of samples from the population.

Quantile – Quantile Plot → (q-q plot)

why? – To check if a random variable is normally distributed or not.

How? Let $x$ be a random variable we want to check →

1) Take $n$ observations of $x$.
2) Divide the $n$ observations into percentiles. → $x_p$

1) Take another r.v $N(0,1)$, obtain 1000 observations.
2) Divide them into percentiles. → $y_p$

For each percentile $x_i$ in $x_p$ & $y_i$ in $y_p$
plot the pair $(x_i, y_i)$


If points lie on a straight line
$x$ ~ Normally distributed.

else
$x$ ~ Not Gaussian.