Life Cycle of Data Science Project

Requirment Gathering → **Data analytics team** ←PM ←BA —Data Source→ **Big Data team** → **Life cycle of DS project**

- Requirment Gathering ↑ Product / Project manager (PM); Business analyst (BA)
- Data analytics team ↳ Data analyst, Data scientist, Big Data engineer, Cloud Engineer (Decide to get data and from where or sources)
- Big Data team ↳ Big Data engineer
  - Sources:
    - Internal
    - Third party
    - web Scrapping.
- Life cycle of DS project
  1. EDA (Exploratory Data analysis)
  2. FE (Feature extraction)
  3. FS (Feature selection)
  4. Model Training
  5. Hyperparameter tunning
  6. Deployment

---

* **Statistics:** is the science of collecting, organizing and analyzing the data.
* **Data:** facts or pieces of information

Statistics

- **Descriptive** [EDA+FE]
  1. It consists of organizing and summarizing data using all different kind of plots.
- **Inferential**
  It consists of collecting sample data and making conclusion about population data using some experiments. → Hypothesis Testing

  ⇒ Entire data is population data
  ⇒ Sample from population is sample data.

  Sample Data —conclusion→ Population Data

  ⇒ Relationships / comparisions

---

* **Sampling Techniques:** (select techique which will give good results/data)
  - Simple random sampling :- Every member of population (N) has equal chance of being selected for your sample (n)
  - Stratified sampling :- Strata → Layers → Clusters → Group.
  - Systematic sampling :- select every nth (1st, 2nd ...) individual out of population (N).
  - Convenience sampling :- only those who are intrested in Survey will only participate.

---

* **Variables:** a property that can take values.
  - Quantitative: measured nummerically (mathematical operations)
  - Qualitative / categorical: categories based on characterstics
  - Discriptive: whole numbers, many repetations.
  - Continous: Fractions/decimals,

Histogram

To create histogram do :

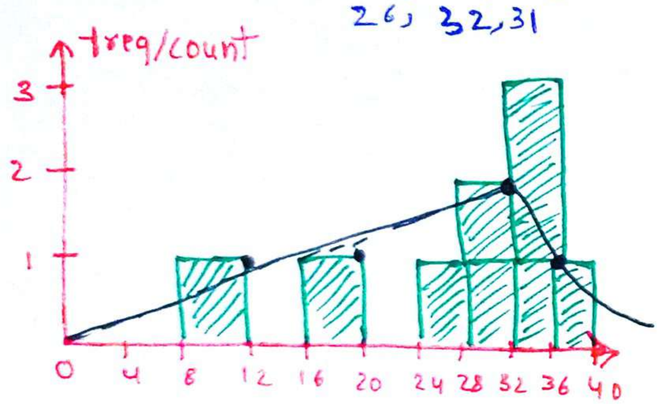
① Sort the values (asec)

② Create bins (groups)

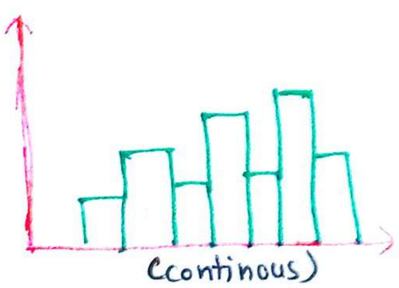
No. of Bins = My choice

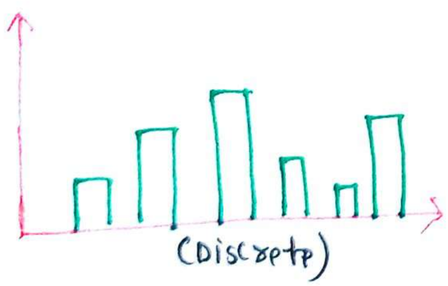
$$\text{No. of bins} = \frac{\text{Last value (Max)}}{\text{First value (Min) or No. of Bins}}$$

Ex→ Values = [10, 20, 25, 30, 35, 40]
26, 32, 31



- Plotting Continous values as well as discrete.





PDF: Probability density function

PMF: Probability Mass function

- PDF use to smooth discrete histogram and PMF used to smoothen continous histogram.

Measure of Central Tendency (MCT)

MCT is single value that attempts to describe a set of data identifying the central tendancy.

① Mean/Average:

$$\text{Population mean } (\mu) = \sum_{i=1}^{N} \left[\frac{x_i}{N}\right]$$

N >> n

$$\text{Sample mean } (\bar{x}) = \sum_{i=1}^{n} \left[\frac{x_i}{n}\right]$$

But, $\bar{x} > \mu$ or $\mu > \bar{x}$ depending on sampling technique.

➡ In practical, we use mean to replace NAN values [Calthough not efficient) due to outliers]

② Median:

To calculate it do
↳ Sort the values
↳ Find central number
↳ if no. of values even, find average of central numbers
↳ if no. of values odd, find the central number

→ Central number is value in mid of the all values.

→ If no outliers use mean, if outliers use median to replace the NAN values.

③ Mode:

most frequent occuring element

→ Used to replace the NAN values in categorical data within the most frequent value.

④ Measure of Dispersion:

4.1 Variance (σ^2)
- spred of data

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

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

Variance ∝ Spread of data

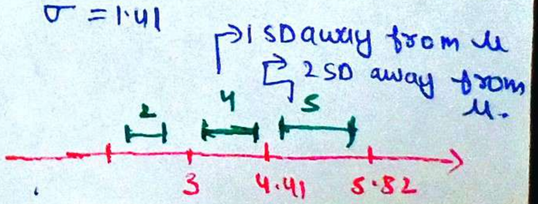
Less variance, less spread

More variance, more spread

4.2 Standard Deviation ($\sqrt{\sigma^2}$) = σ
(SD)

- Tells how many standard deviation a value is away from mean.

Ex→ {1, 2, 3, 4, 5}, $\mu = 3$, $\sigma^2 = 2$, $\sigma = 1.41$



4.3 Percentiles and Quartiles

$$\text{Percentage} = \frac{\text{sum of all}}{\text{Total number}} \times 100$$

- A percentile is a value below which a certain percentage of observation lies.

  ex→ 99 percentile means a person got better marks than 99% people.

$$\text{Percentile rank of } x = \frac{\text{No. of value below } x}{\text{Total values}}$$

- It may tell us about the distribution of data, like how much data in how much percentile.
- Two value can have same percentile rank.

$$\text{value index} = \frac{\text{Percentile}}{100} \times (\text{Total values} + 1)$$

* 5 Number theory

[A] Minimum

[B] First Quartile (25 percentile) (Q_1)

[C] Median

[D] Third Quartile (75 percentile) (Q_3)

[E] Maximum

→ Used to remove outliers

→ Used to draw boxplot to remove outliers

→ To remove outliers we use fences.

→ we keep the values between the lower and higher fence and remove others.

$$\text{Lower fence (LF)} = Q_1 - 1.5[IQR]$$

$$\text{Higher fence (HF)} = Q_3 + 1.5[IQR]$$

where,

IQR is Inter Quartile range,

$$IQR = Q_3 - Q_1$$

→ To treat outliers make fence

→ To visualize outliers use Boxplot.

→ Drawing Boxplot,

Ex- {1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

$$Q_1 \text{ (ind)} = \frac{25}{100} \times (20+1) = 5.25 \Rightarrow Q_1 \text{ val} = \frac{3+3}{2} = 3$$

$$Q_3 \text{ (ind)} = \frac{75}{100} \times (20+1) = 15.75 \Rightarrow Q_3 \text{ val} = \frac{8+7}{2} = 7.5$$

So,

$LF = 3 - 1.5 \times 3 = -3.65$.

$HF = 3 + 1.5 \times 7.5 = 14.25$

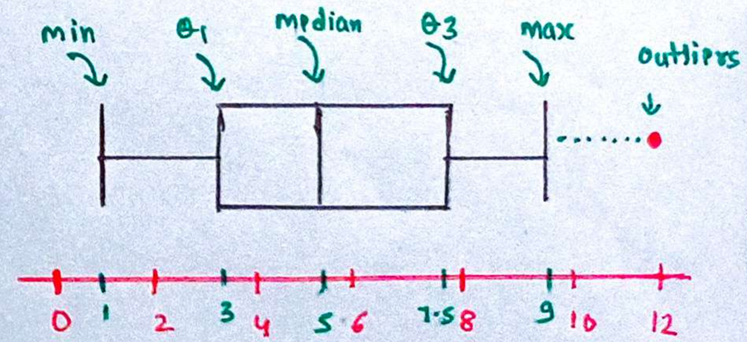
Our values must be within, [-3.65 to 14.25]

So, 27 is only outlier in our dataset.

Now,

minimum = 1, $Q_1 = 3$, median = 5

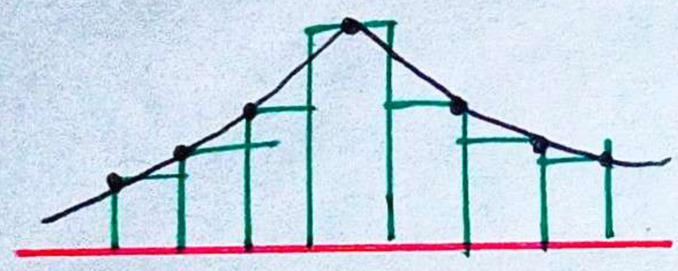
$Q_3 = 7.5$, maximum (without outliers) = 9.



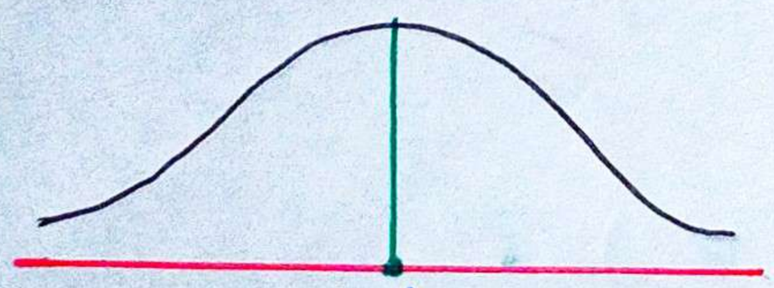
* Normal / Gaussian Distribution

- type of continous probability distribution for real valued random variable.
- Bell shaped curve
- Assumed that during any experiment measured values will follow this distribution with equal number of values before and after a 'central tendency' (mainly median).

[KDE = Kernel Density Estimator]



apply KDE to smoothen the histogram to obtain bell curve.



central tendency [CT] (mean)

→ Total area under curve is '1' or 100% as it is probability distribution.

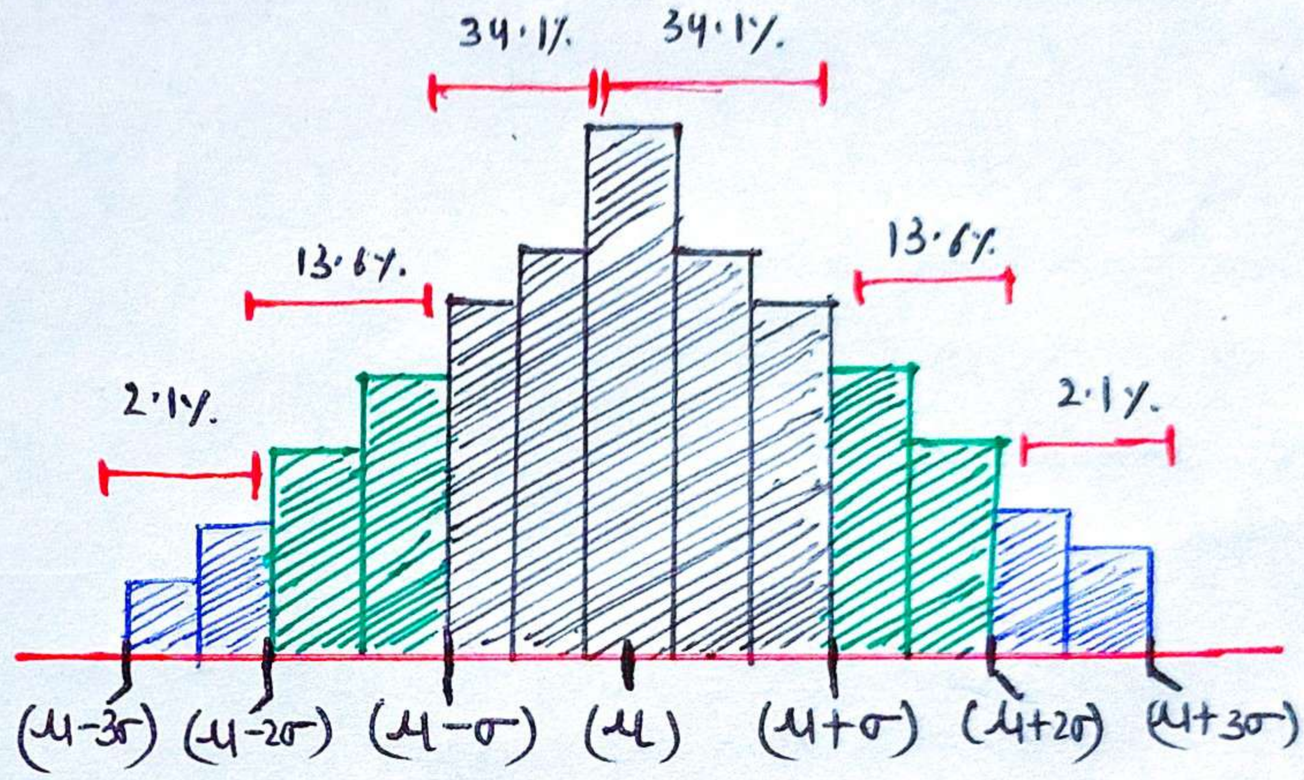
→ 50% data lie on each side of CT.

→ By exp. it is found most important data like age, height, many more follow this distribution.

65 - 95 - 99.7 %

Empircal formula of Gaussian Distribution (rule)

μ = population mean σ = standard deviation.



By experimentation it is determined that for any Gaussian Distribution,

→ 68% data lies within the range of (34.1 + 34.1) first standard deviations i.e, from $(\mu-\sigma)$ to $(\mu+\sigma)$

→ 95% data lies within range of the (34.1 + 13.6 + 34.1 + 13.6) second standard deviations i.e, from $(\mu-2\sigma)$ to $(\mu+2\sigma)$

→ 99.7% data lies within the range of (2.1 + 13.6 + 34.1 + 34.1 + 13.6 + 2.1) third standard deviations i.e from $(\mu-3\sigma)$ to $(\mu+3\sigma)$

* Central Limit theorem

- It states that distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the distribution type of population.
- Sample size must be atleast 30.

→ If we choose 'm' samples using any sampling technique from a given population with '$\bar{x}_m$' mean and plot these means then it will yield a normal distribution regardless of the nature of distribution of population.

→ Larger than sample size, closer the distribution towards normal distribution.

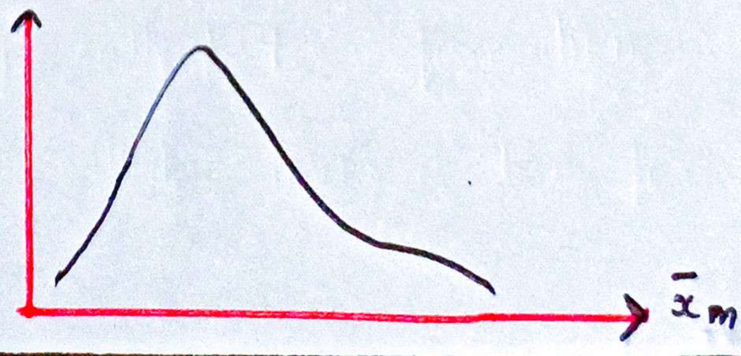
Say, we have few samples. $(S_1, S_2, \dots S_m)$ with each having some random elements giving means $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m)$. Then these means are normally distributed.

$S_1 \rightarrow \{x_1, x_2, \dots x_m\} \rightarrow \bar{x}_1$

$S_2 \rightarrow \{x_4, x_6, x_9, \dots x_{11}\} \rightarrow \bar{x}_2$

$\vdots$

$S_m \rightarrow \{x_3, x_5, x_7, \dots, x_{m-1}\} \rightarrow \bar{x}_m$



* Covariance

- Covariance measure the direction of relationship between two variables (features)
- A positive covariance means both variables tend to be high or low at the same time (directly proportional)
- A negative covariance means that when one variable is high, the other one tends to low. (inversly proportional.
- A '0' covariance means no relation.
- It is used to measure the distribution of the data points.

→ Variance is also used to measure the distribution of data points, but just for one variable, while covariance is used for multiple variables.

As variance,

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

$$\Rightarrow \sigma^2 = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

$$\Rightarrow Cov(X,X) = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

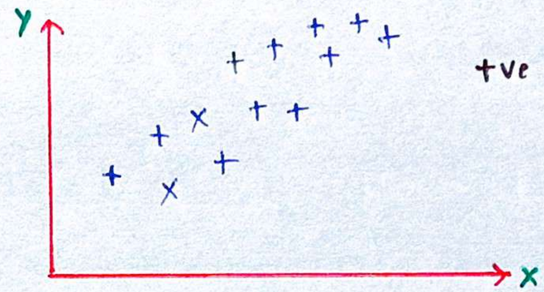
$$\therefore \boxed{Cov(X,X) = Var(X)}$$

And formula for covariance is,

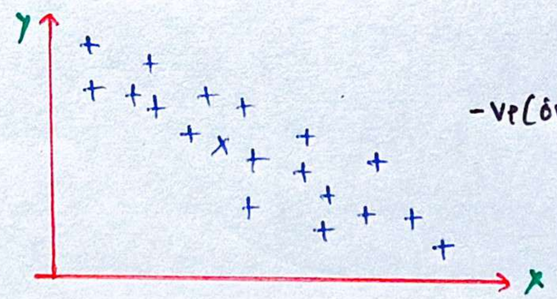
$$\boxed{Cov(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}$$

→ Variance measures the spread of data along single axis, while covariance measures directional relationship.

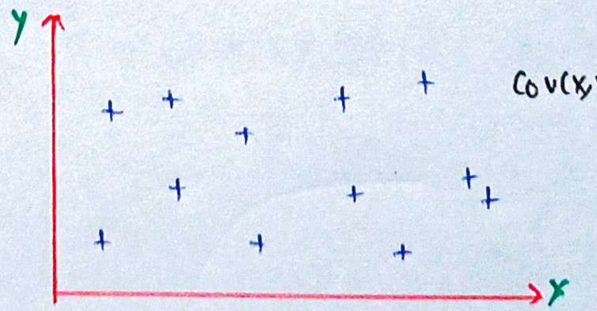
- Few graphical representation of Cov(X,Y),



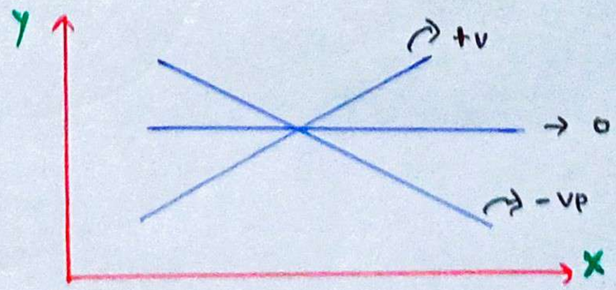
+ve Cov(X,Y)



-ve(Cov(X,Y)



Cov(X,Y) = 0



ex- We have two features with certain values.

X	Y
12	40
13	45
15	48
17	60
18	62

$\bar{x} = 15$ $\bar{y} = 51$ (means)

then,

$$Cov(x,y) = [(12-15)(40-51) + (13-15)(45-51) + (15-15)(48-51) + (17-15)(60-51) + (18-15)(62-51)] / 5-1$$

$$= [(-3 \times -11) + (-2 \times -6) + (0 \times -3) + (2 \times 11) + (3 \times 11)] / 4$$

$$= [33 + 12 + 0 + 22 + 33] / 4$$

$$= 25$$

∴ Cov(x,y) = 25, is +ve so their relationship is directly proportional.

→ It is used in feature selection

→ One drawback of covariance is that, it only tells direction of relation and not the strength.

So we cant compare features relation with each other and say it has stronger relation compared to another one as Cov has no scale and its value can go as high and low.

* Correlation

- Correlation measures the degree to which two variables move in relation to each other.
- Correlation shows the direction of relationship as well as the strength of it. and represent it nummerically as correlation coefficent.
- Correlation coefficents value range in -1 to $+1$.
- Positive correlation coefficent means that when one variable is increasing other tends to follow the same.

  Closer the coefficent to 1, stronger the directly proportionality relation.
- A negative correlation coefficient means that both variable tends to follow inversly proportional relation.

  Closer the coefficient to -1, stronger the inversly proportional relationship.
- It solves the major drawback of covariance by limiting the value of correlation coefficient within a range of -1 to 1, which help in comparing strengt of relationships in for accurate way.
- Their are different type of correlation coefficients available.

* Pearsons product-moment correlation coefficient

- We tweak the formula of covariance to limit its range between -1 to 1.

$$\boxed{\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}, \quad \sigma_x \sigma_y > 0}$$

- if $\rho(x,y) = 1$, perfect direct increasing or linear relationship.
- if $\rho(x,y) = -1$, perfect inverse linear relationship.
- all other cases within $(-1, 1)$
- if $\rho(x,y) = 0$, no relation.

* Spearmens Rank Correlation.

- We tweak the formula of pearsons coefficient to get better and more precise values.

$$\boxed{\text{Spearmen rank} = \frac{Cov[R(x), R(y)]}{\sigma_{R(x)} \sigma_{R(y)}}}$$

- to get the Rank of a variable $(R(x))$, we simply allot a value within $(1, \infty)$ starting from smallest value to the largest.
- if two values are same they will get same rank.

ex -

X	Y		R(x)	R(y)
10	20		4	1
8	30	then	3	2
7	40		2	3
6	50		1	4
6	70		1	5

then use formula accordingly.

* Permutation

- It is a technique that determines the number of possible arrangments in set when the order of arrangments matters.
- Permutation is set of arrangment of its members into sequence. And if the members already ordered then reordering them.

$$\boxed{{}^{n}P_{r} = \frac{n!}{(n-r)!}}$$

n = total no. of elements in set

r = no. of selected elemnts arranged in order

* Relation of permutation and combination

$$\boxed{\text{Theorem: } {}^{n}P_{r} = {}^{n}C_{r} \cdot r!}$$

$$ {}^{n}C_{r} \cdot r! = \frac{n!}{r!(n-r)!} \cdot r!$$

$$= \frac{n!}{(n-r)!}$$

$$= {}^{n}P_{r}$$

→ It means that for each combination of ${}^{n}C_{r}$ we have $r!$ permutations because r objects in every combination can be rearranged in $r!$ ways.

* Combination

- It is a techique that determines the number of possible arrangments in the collection of items where order of selection doesn't matters.
- Combination is selection of items from a set that has distinct members, such that order of selection doesnt matter.

$$\boxed{{}^{n}C_{r} = \frac{n!}{r!\,(n-r)!}}$$

n = total no. of elements in set

r = no. of selected elements.

$$\boxed{\text{Theorem: } {}^{n}C_{r} + {}^{n}C_{r-1} = {}^{n+1}C_{r}}$$

$$ {}^{n}C_{r} + {}^{n}C_{r-1} = \frac{n!}{r!(n-r)!} + \frac{n!}{(r-1)!(n-r+1)!}$$

$$= \frac{n!}{r(r-1)!(n-r)!} + \frac{n!}{(r-1)!(n-r+1)!(n-r)!}$$

$$= \frac{n!}{(r-1)!(n-r)!}\left[\frac{1}{r} + \frac{1}{(n-r+1)}\right]$$

$$= \frac{(n+1)!}{r!(n+1-r)!}$$

$$= {}^{n+1}C_{r}$$

* Permutation vs Combination

→ **Permutation:**

- the different ways of arranging a set of objects into a sequential order.
- order matters.
- denotes arrangment of objects
- multiple permutations can be derived from single combination.
- simply these are ordered elements.

→ **Combination:**

- one of the several way of choosing object from large set of objects, without considering the order.
- order doesn't matters.
- denotes selection of objects.
- single combination can be derived from single permutation.
- simply defined as unordered sets.

→ If order matters use permutation, else combination.

→ Ex, Let elements be X, Y, Z. (for two)

- Possible permutations are,
  XY, YX, YZ, ZY, XZ, ZX
- Possible combinations are,
  XY, YZ, ZX.

- Experiment are the uncertain situations which could have multiple outcomes.
- Outcome is the result of single trial.
- Event is one or more outcome from an experiment.
- Probability is a measure of likelihood of an event.
- Mutually exclusive events:
  - → Two events are mutually exclusive if they can't occur at the same time.
  - → ex- Tossing of coin.
- Non mutual exclusive events:
  - → Two events are non mutual exclusive if they can occur at same time.
  - → ex- picking heart and King from deck.
- Dependent events:
  - → Two events are dependent if they affect occurence of each other.
  - → ex- If bag has 5R, 5B balls then, picking a red ball has probability of $P(R) = 5/10$ now if again we pick red probability is $P(R) = 5/9$.
- Independent events:
  - → Two events are independent if they dont affect occurence of each other.
  - → ex- In prev example if we put ball back in bag after picking probability remain same.

* Addition Law of probability.

- For mutually exclusive events:

$$\boxed{P(A \text{ or } B) = P(A) + P(B)}$$

- For non-mutual exclusive events:

$$\boxed{P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)}$$

* Multiplication Law of probability

- For Dependent events:

Conditional probability

$$\boxed{P(A \text{ and } B) = P(A) * P(B/A)}$$

- For Independent events:

$$\boxed{P(A \text{ and } B) = P(A) * P(B)}$$

* Naive Bayes theorem:

- Its basically conditional probability for mutually exclusive events.
- If $A_1, A_2, \ldots A_n$ are mutually exclusive events with $P(A_i) \neq 0$ $(i = 0, 1, \ldots n)$ of random experiment then for an arbitary event B of the sample space of above experiment, $P(B) > 0$.

$$\boxed{P(A_i/B) = \frac{P(A_i)\, P(B/A_i)}{\sum_{i=1}^{n} P(A_i)\, P(B/A_i)}}$$

* Odds in favour and against event

m = number of favourable events (ways)
n = number of non-favourable events

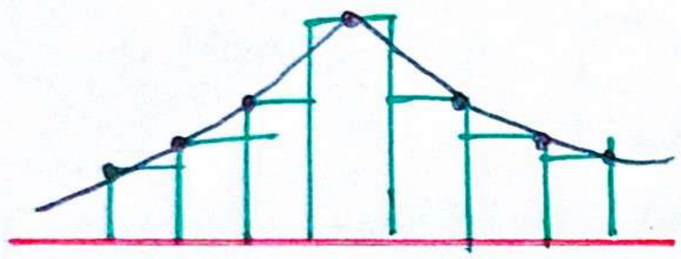
$$\boxed{\text{Odds}_{\text{Favour}} = m/n}$$

$$\boxed{\text{Odds}_{\text{Against}} = n/m}$$

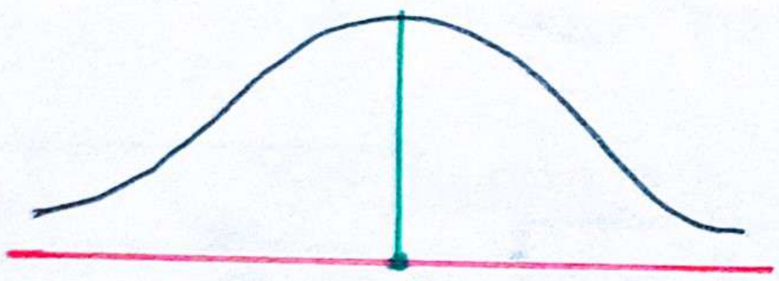
* Normal / Gaussian Distribution

- type of continous probability distribution for real valued random variable.
- Bell shaped curve
- Assumed that during any experiment measured values will follow this distribution with equal number of values before and after a 'central tendency' (mainly median).

[KDE = Kernel Density Estimator]



apply KDE to smoothen the histogram to obtain bell curve.



central tendency [CT] (mean)

→ Total area under curve is '1' or 100% as it is probability distribution.

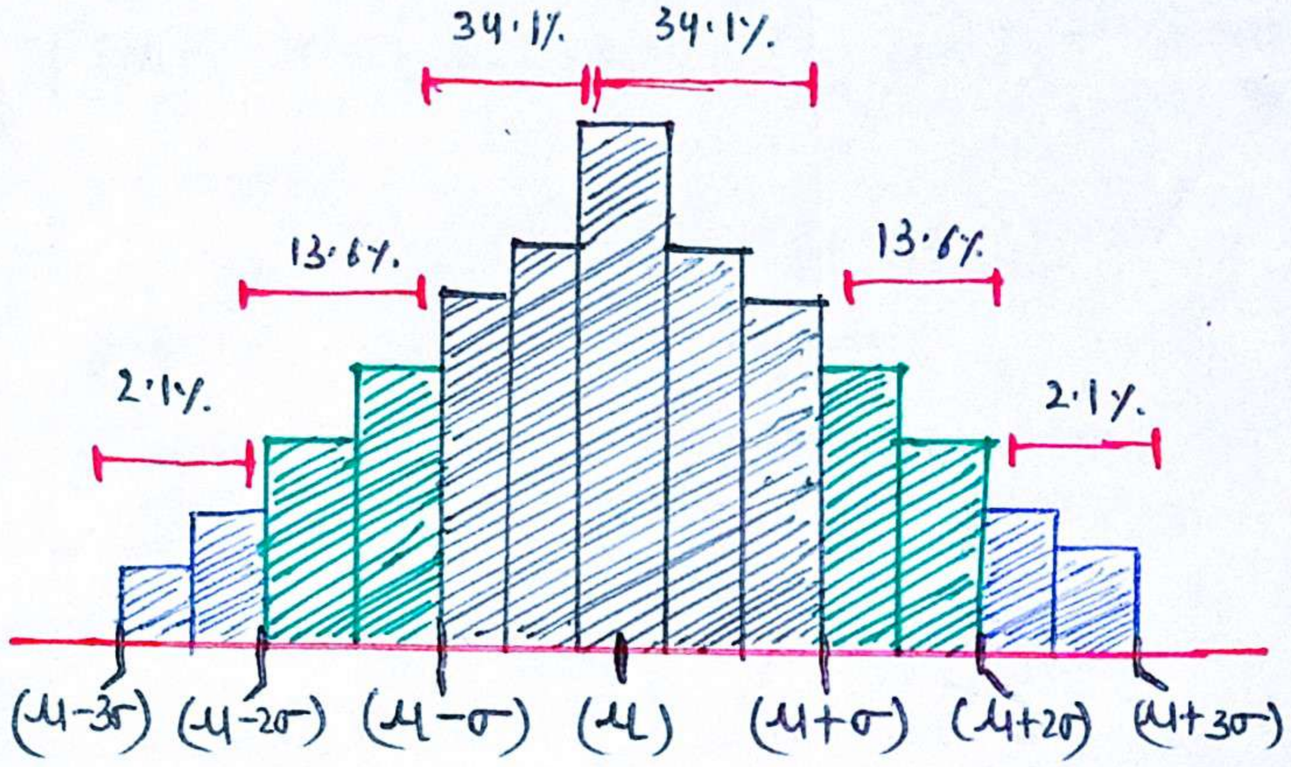
→ 50% data lie on each side of CT.

→ By exp. it is found most important data like age, height, many more follow this distribution.

$\boxed{65 - 95 - 99.7\%}$

Empircal formula of Gaussian Distribution (rule)

μ = population mean σ = standard deviation.



By experimentation it is determined that for any Gaussian Distribution,

→ 68% data lies within the range of first standard deviations i.e, from $(\mu-\sigma)$ to $(\mu+\sigma)$
$(34.1+34.1)$

→ 95% data lies within range of the second standard deviations i.e, from $(\mu-2\sigma)$ to $(\mu+2\sigma)$
$(34.1+13.6+34.1+13.6)$

→ 99.7% data lies within the range of third standard deviations i.e from $(\mu-3\sigma)$ to $(\mu+3\sigma)$
$(2.1+13.6+34.1+34.1+13.6+2.1)$

* Standard Normal Distribution

Let 'x' be a random variable which belongs to a Gaussian distribution having 'μ' mean and 'σ' standard deviation.

Then we can transform this 'x' to 'y' which will belong to Standard Normal Distribution having $\mu = 0$ and $\sigma = 1$, this process is called standardization and done by calculating Z-score.

$x \in$ Gaussian (μ, σ)

$y \in$ Normal standard

Transformation (standardization) using Z-score.

→ x to y called transformation.

→ y to x called inverse transformation.

$$\boxed{Z \text{ score} = \frac{x_i - \mu}{\sigma/\sqrt{n}}}$$

here,

$$\boxed{\sigma/\sqrt{n} = \text{Standard error}}$$

where n is sample size.

as in case of Gaussian to Normal standard transformation, we transform each value or features so 'n' tends to be 1, as area under probability distribution as 1.

So, Z-score formula for this specific transformation case is $(n=1)$,

$$\boxed{Z\text{-score} = \frac{x_i - \mu}{\sigma}}$$

Ex → $x = \{1, 2, 3, 4, 5\}$, $\mu = 3$, $\sigma = 1.414$

Zscores,

for 1 its -1.414 as $(1-3/1.414)$

Similarly for 2 its -0.707,

for 3 its 0, for 4 its 0.707, for 5 its 1.414

$\therefore$ $y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

Comparing x and y we can conclude that after standardization scope has decreased leading to faster calculation and easy visualization of data.

→ This standardization helps us build models efficiently as it decreases the scope for those datasets where each features range vary widely.

ex-

Age (yr)	Height (cm)	Weight (Kg)
20	180	72
22	190	94
24	160	82

Here all these features have huge scope, so we should standardize it first.

→ Their is another way to scale random variables which is called Normalization.

→ In case of standardization range of scaled values determined by the formula outcome.

→ In case of Normalization range of scaled variable can be determined in advance.

One method to do Normalization is min-max normalization which gives the scaled variables withing range of 0 to 1.

And is used mostly in deep learning.

Another example is that pixels of image are kept within range of 0-255.

→ Normalization and Standardization are part of procedure called Feature scaling.

→ Basically in normalization we are shifting our mean close to Z-score, same with standardization.

→ Area under the curve can be calculated by Z-score and Z-table.

* Log Normal Distribution :

→ It is right skewed continous probability distribution, meaning it has long tail towards right.

→ It is continous probability distribution of a random variable whose logarithm is normally distributed.

→ Also Known as 'Galtons distribution'

- For any random variabe 'X' can be transformed to log normal distribution by calculating its natural log.
- Similarly log normal distribution can be inverse transformed to original form by calculating exponent of the data points of log normal distribution.

$x \in$ Any Distribution (μ, σ)

Transformation

$y \in$ Log Normal Distribution (μ, σ^2)

$$\rightarrow \text{PDF} = \frac{1}{x\sigma\sqrt{2\pi}} e^{\left(\frac{-[\ln(x)-\mu]^2}{2\sigma^2}\right)}$$

→ x to y called transformation.

→ y to x called inverse transformation.

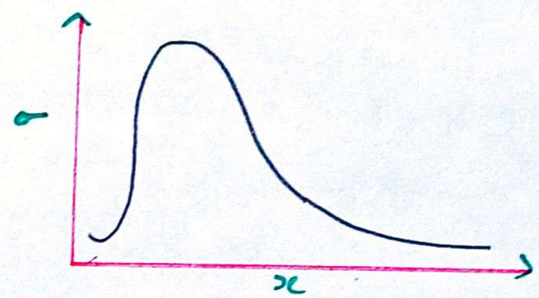
→ Transformation: If the random variable x of any distribution converted to Normal log distribution, then value of the x on new distribution is y.

$$\boxed{y = \ln(x)}$$

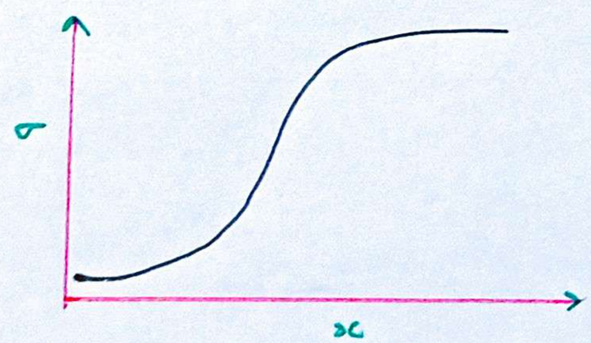
→ Inverse Transformation: If the random variable y of log normal distribution converted back to original distribution, then value of y on new distribution is x.

$$\boxed{x = \exp(y)} \quad \text{[antilog]}$$

→ Log normal Distribution (PDF)



→ Log Normal Distribution (CDF)



→ For log normal distribution,

$$\boxed{\text{Mean } (\mu) = \exp\left(\mu + \frac{\sigma^2}{2}\right)}$$

$$\boxed{\text{Median} = \exp(\mu)}$$

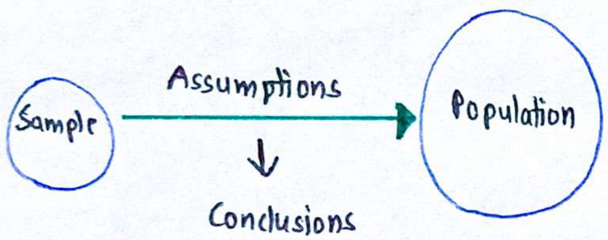
$$\boxed{\text{Mode} = \exp(\mu - \sigma^2)}$$

$$\boxed{\text{Variance} = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)}$$

* Hypothesis testing

- It is an act in statistics where we test an assumption regarding a population parameter.
- We test an hypothesis by measuring and examining a random population sample.
- It is done to check plausability of null hypothesis

→ We take sample of a population and make certain assumptions about population based on sample called hypothesis and then try to conclude about the hypothesis.



- Conclusion can be that Null hypothesis accepted or we failed to reject Null hypothesis within Confidence Interval.
- Conclusion can also be that Alternate hypothesis accepted or Null hypothesis rejected within Confidence Interval.

⊙ Test stastic (T) is dependent on test and is basically calculated value of critical value (c).

⊙ Null hypothesis is stastical theory that suggests their is no statistical significane exists between the sample and population.
It is basically the default assumption which is made about population based on facts.

⊙ Alternate hypothesis is stastical theory that suggests their is statistical significance between the sample and population.
Basically its just opposite of what we assumed in null hypothesis.

→ Null hypothesis represented by H_0 and alternate hypothesis by H_1 or H_a.

⊙ Level of significance (α) is fixed probability of rejecting a True Null Hypothesis.

$$\boxed{\alpha = 1 - CI}$$

ex- if $\alpha = 5\%$ that means we are okay to take 5% risk and conclude their exists a difference when their is no actual difference.

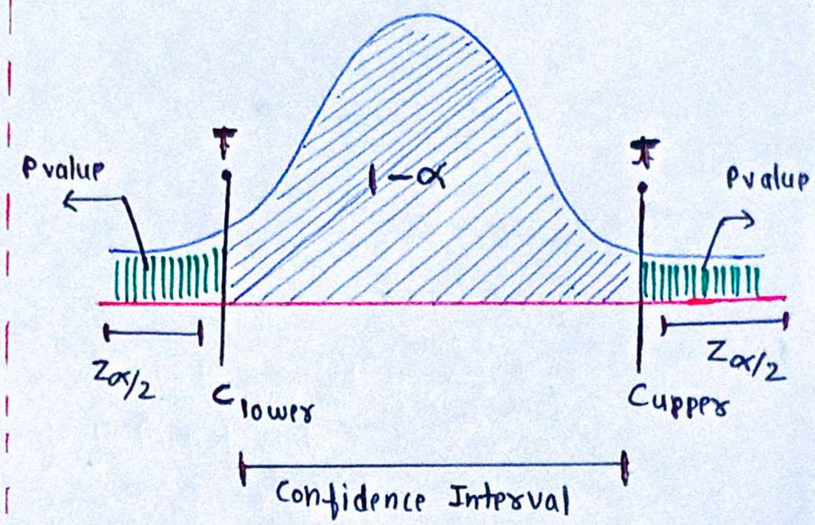
⊙ Critical value (C) is the value in distribution beyond which Null hypothesis is rejected.

⊙ Confidence Interval (CI) is the range of value in which Null hypothesis is accepted.
Critical value is the end point of CI.

⊙ P-value is proportion of samples that would be at extreme end of distribution and null hypothesis is rejected here.

P value is the region with minimum significance level.

ex- Consider Normal Distribution. with $\alpha = 5\% = 0.05$, $CI = 95\% = 0.95$



→ Procedure for testing

[A] Define Null and Alternate Hypothesis.
[B] Set α and CI.
[C] Compute test statistics (choose which suits data)
[D] Conclude:

- if $T \geq C_{upper}$ and $T < C_{lower}$, reject the hypothesis.
- if $T \geq C_{lower}$ and $T \leq C_{upper}$, fail to reject null hypothesis.

or

- if p value $> \alpha$, fail to reject null hypothesis.
- if p value $\leq \alpha$, reject null hypothesis.

→ Confusion matrix in hypothesis testing

Predicted values \ Actual values	H_0	H_a
Fail to reject H_0	Confidence $(1-\alpha)$ ●	Type II error (β) ●
Reject H_0	Type I error (α) ●	Power of Test $(1-\beta)$ ●

● wrong descission

● correct descission

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total predictions}}$$

- Confidence $(1-\alpha = CI)$ is probability of accepting True Null Hypothesis.
- Power of Test $(1-\beta)$ is probability of rejecting False Null Hypothesis.
- Type I error (α) occurs when we reject true Null Hypothesis.
- Type II error occurs when we accept False Null Hypothesis.

Confidence $\propto$ Power of Test

→ Point Estimate (PE)

- The value of any statistics that estimates the value of a parameter
- It is basically test statistic (t).

$$\boxed{PE \pm \text{Margin Error} = \text{Paramptr}}$$

$$\boxed{C_{lower} = PE - \text{Margin Error}}$$

$$\boxed{C_{upper} = PE + \text{Margin Error}}$$

$$\boxed{\text{Margin Error} = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}$$

(standard error)

$\frac{\sigma}{\sqrt{n}}$ is population standard deviation

- ex- Sample mean $(\bar{x})$ can be used to estimate population mean (μ).

$$\underbrace{\bar{x}}_{\text{statistic}} \xrightarrow{\text{estimate}} \underbrace{\mu}_{\text{parameter}}$$

- Based on distribution Hypothesis testing can be performed at various ends of distribution to find H_0 acceptance or rejection.

→ One Tail
  → Right Tailed
  → Left Tailed

→ Two tailed

→ Test statistics

Hypothesis Test	Test Statistics
Z - Test	Z - Score
T - Test	T - Score
F - Test	F - statistic
Chi-Square Test	Chi-square statistic.

- Z, T test are done on normally distributed data.
- Chi-square test done on chi-squared distributed data.
- F-Test done on f distributed data.
- Annova test is variation of F-test.

→ Z-Test

- Used when
  - → Sample size large ($n \geq 30$) or
  - → Population variance (σ^2) Known or we know standard deviation (σ)

$$\text{Margin Error (ME)} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$C_{lower} = PE - ME$$

$$C_{upper} = PE + ME$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- To perform the z-test
  - → Calculate ME, C_{lower}, C_{upper}, z
  - → Use z-table to get z values
  - → Take descision from stated rules of test statistic
  - → Make conclusion
- If population standard deviation not known and still want z-test we can use sample standard deviation (s).

  Formulas : σ will be replaced by s.
- Comparing two different groups,

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1/n_1)^2 + (\sigma_2/n_2)^2}}$$

→ T-Test

- Used when
  - → sample size small ($n < 30$) and
  - → Sample variance (s^2) Known or standard deviation (s) Known

$$\text{Margin Error (ME)} = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$C_{lower} = PE - ME$$

$$C_{upper} = PE + ME$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- To perform the T-test
  - → Calculate ME, C_{lower}, C_{upper}, t
  - → Use T-table to get t values
  - → Take descision from stated rules of test stats
  - → Make conclusion
- To refer T table we also need degree of fredom (df)

$$df = n - 1$$

- Comparing two different groups,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$df = n_1 + n_2 - 2$$

→ Chi-Square Test

- Used when data is chi-square distributed
- Used to check whether variables in population are independent or not
- This test checks the difference between observed (o) and expected (e) value which can be calculated using observed frequency (f_o) and expected frequency (f_e)
- ⊙ Its non parameter test performed on categorical data.

$$\text{number of categories (c)} = \text{Total categories}$$

$$df = c - 1$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_o - f_e)^2}{f_e}$$

- To perform chi-square test,
  - → calculate c, df, χ^2
  - → Get critical value (c_v) from χ^2 table using α and df
  - → If
    - $\chi^2 > c_v$, reject H_0
    - $\chi^2 < c_v$, fail to reject H_0