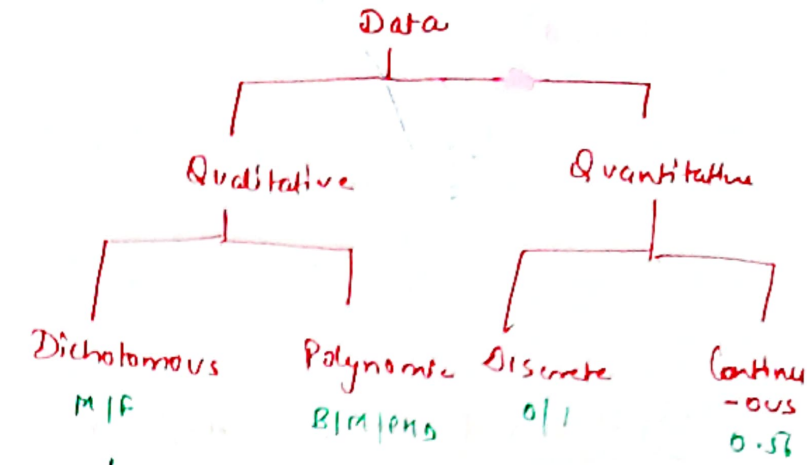
EDA (Exploratory Data Analytics)

- ✓ Correlation Matrix, heatmap
- ✓ Box Plot
- ✓ Histogram
- ✓ Kernel Density Estimate
- ✓ Central Tendency: Mean, Median, Mode
- ✓ Variance:- Range, std Deviation, IQR ($Q_3 - Q_1$)



Correlation Matrix

[Correlation matrix is useful to find highly correlated variables, these variables should be removed during the flow of analysis.]

↳ Multicollinearity Problem] → when one variable is predictive enough of other variable misleading ML models.

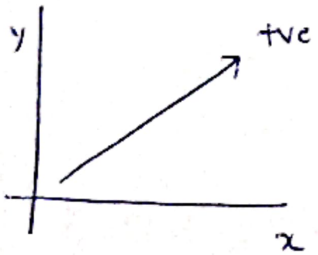
- delete one
- Dimensionality Redn
  - ↳ PCA

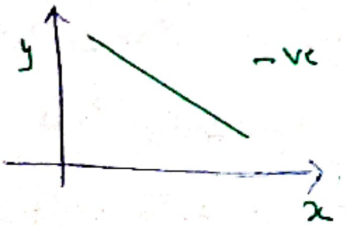
Correlation - Two variables

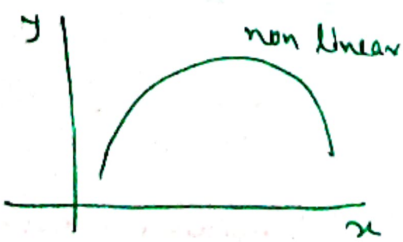
Scatter plot can tell us whether there is a linear correlation b/w two variables

+ve linear, -ve linear

non linear, no correlation







Correlation Coeff - strength + direction of correlation.

$(-1, 1)$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Significance of correlation coefficient

α - level of significance ≈ 0.05

n → number of pairs of data in sample.

if $|r| >$ critical value at α & n in Pearson Correlation Coefficient table

∴ Correlation significant

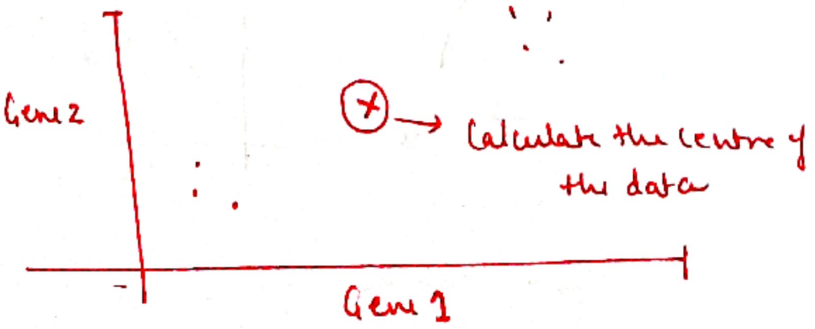
Correlation ≠ Causation.

Dimensionality Red^n

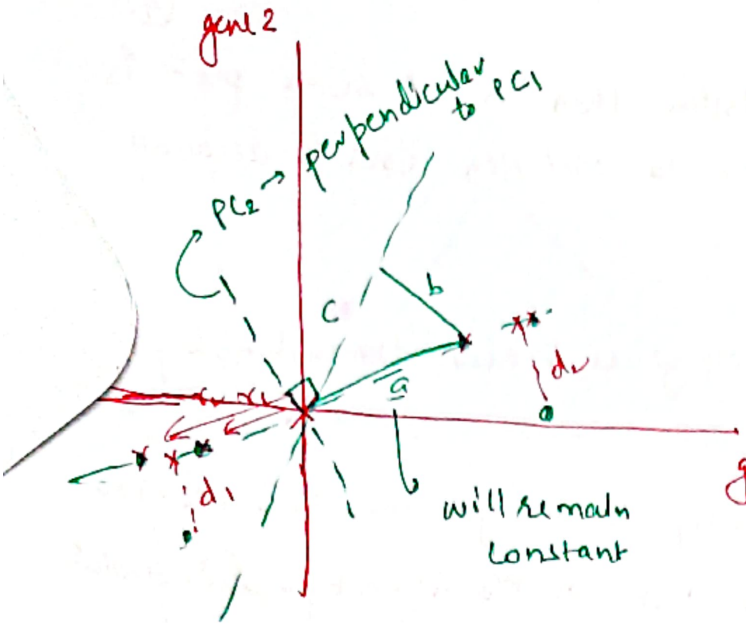
PCA: Principle Component Analysis

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

Lets say this is the data



Lets find the centre of the data for the first two genes and then shift the graph such that centre comes to origin



3

Now try fit a line to fit the data as well as it can.

* we find the line which minimize the distances from the data points d_1, d_2

* Or maximize the distance of projections from origin $f_1, f_2, \ldots$

PCA uses this.

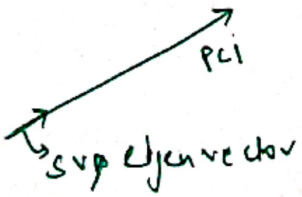
$a^2 = b^2 + c^2$

if b ↑ c ↓

if c ↑ b ↓

* first best fit line is called PC1, Principle component 1.
  - ↳ linear combination of gene 1 and (2)
  - linear combination of variables

* when we do PCA with SVD everything is scaled to 1, especially the hypotenuse, this unit vector along PCA is called a singular vector or eigenvector for PC1.



eigen value = sum of distance b/w projected point & the origin

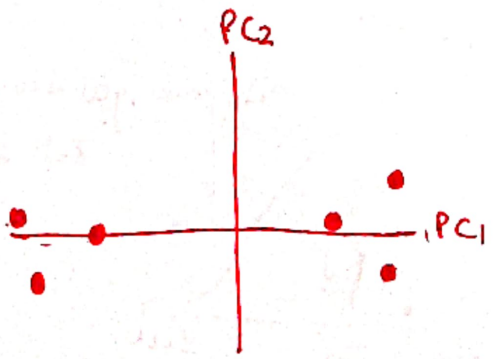
eigen value = $d_1^2 + d_2^2 + \ldots$ (projected distance from origin)

Variation = $\frac{d_1^2 + d_2^2 + \ldots}{n-1}$

↓

we select a line with largest variations as PC1

PCA plot



* PC_1 accounts for more variation than PC_2. A scree plot is a graphical representation for the variation each PC accounts for.

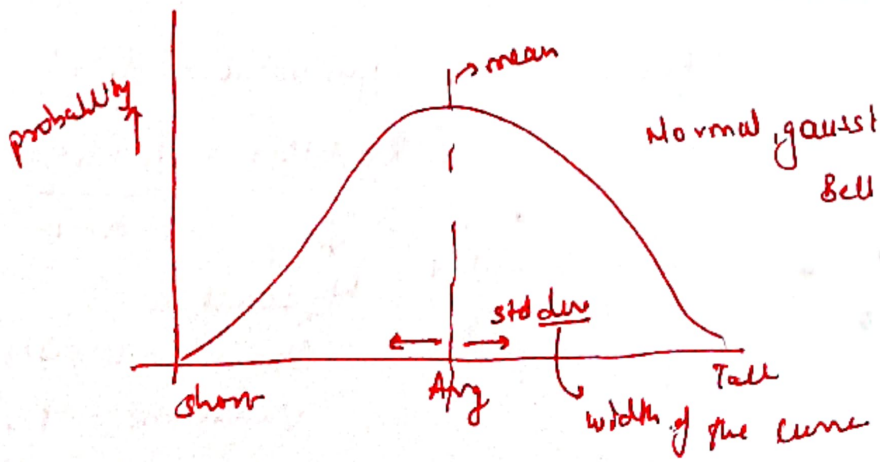
for 3D $PC_1, PC_2, -- PC_3$ → It gives a new dimensional plane.

Now, how this reduces dimensionality, as we get the equal number of PCs as the dimensions. If the first two PCs account for substantial variation, we ignore other dimensions and take first two PCs regardless.

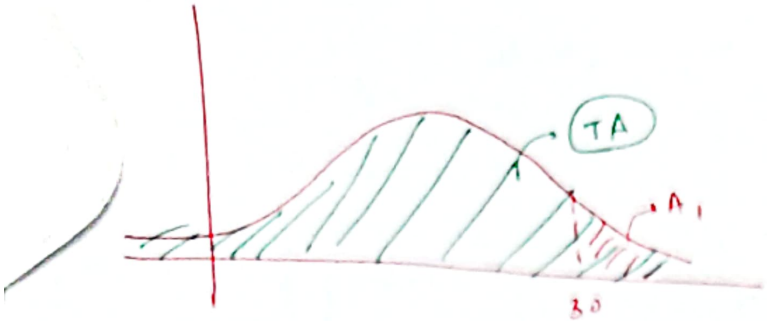
Histogram : we need to decide a good size of bin width to get a clear picture

Curve over a histogram is a distribution.

Curve with mean and std deviations.

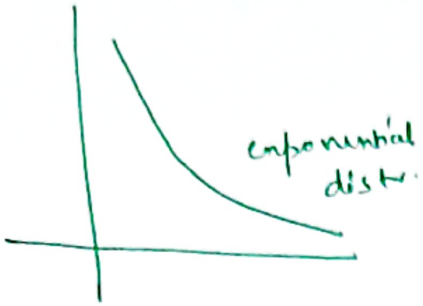


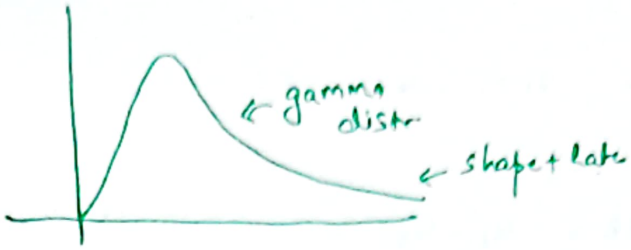
Normal, gaussian distribution
Bell shaped curve



Probability for values > 30

$\frac{A1}{\text{Total Area}}$





We use data to generate these curves and its very important to determine how much confidence we have in these estimates. Statisticians use p-values & confidence intervals for that.

mean = $\frac{\Sigma x}{n}$ → this mean will be an estimate of population not exactly the population mean because its not always the case tht we have enough data.

$(\bar{x})$ ← estimated mean

μ → population mean

$$\text{std dev} = \sqrt{\frac{\Sigma (x-\mu)^2}{n}}$$

$$\text{Population variance} = \sqrt{\frac{\Sigma (x-\mu)^2}{n}}$$

estimated population variance = $\frac{\Sigma (x-\bar{x})^2}{(n-1)}$

$$\text{estimated std dev} = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$

↳ so that we do not underestimate the variance

Central Tendency Measures - mean, median, mode

Box plot | Box & Whiskers Plot

Q_1 - 25% below, 75% above.
Q_2 - median (50% above, 50% below) - Not affected by the outliers.
Q_3 - 75% below 25% above

$IQR = Q_3 - Q_1$

$Q_1 - 1.5(IQR)$ → outlier

$Q_3 + 1.5(IQR)$ → outlier

Hypothesis Testing

If we generate a hypothesis we can clearly test the hypothesis by repeating the experiments for long, but we do not always have this option of repeating experiments multiple times. Hence we need a quick method to check the hypothesis.

H_0 / null hypothesis - negative hypothesis. - We assume a basic null hypothesis in order to avoid the large number of possible hypothesis we can have with us. We reject the null hypothesis then. or fail to reject it.

H_0 → Drug A is not different from Drug B

↳ p value is a number b/w 0-1 which talks about the confidence in H_0. Closer the pvalue to 0 more confidence that drug A is ~~not~~ different from drug B.

P-values / one sided - rarely used.
\ two sided

Example 1:- Null hypothesis: My coin is not different from other coins. It is not special. It does not always gives a head.

↓

Let's validate is hypothesis using P-value.

We toss the coin twice

Outcome	Probability
2 heads	0.25
1h 1t	0.25
1t 1h	0.25
2 tails	0.25

P-value = Probability of getting something as

0.25 + 0.25 + 0

0.25 + 0.25 ← Probability of getting a result in our Observation

0 → Probability of getting something more extreme

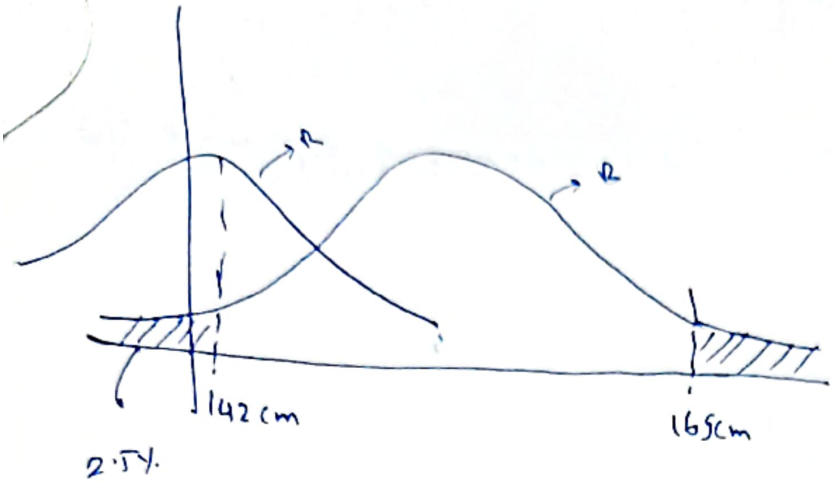
(as only two extreme cases are 2 head/2 tail)

$$P\text{-value} = 0.5$$

Hypothesis - If I get 2 heads in a row, my coin is no different from the normal coin.

We only reject this hypothesis if P value < 0.05

Now its not easy to calculate probabilities for other cases like p values hence we use some distribution of probability to calculate those



Now let say the height is 142cm. Is this height comes from blue distribution or from red distribution.

Ho → It does not come from Red distribution

~~Let's cal~~

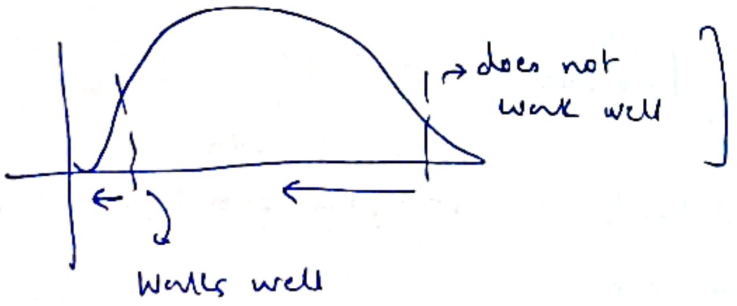
p-value = 0.0025 + 0.0025 = 0.05

↳ p value of the point that this height comes from Blue distribution is 0.05 → Now this is at border line of significance hence we say that this may or may not come from blue distribution.

Now this is ~~Facts~~ Inconclusive.

* One sides p value

We calculate the area only in the direction we want to see the change



Caution don't use this.

- ✓ Confidence Interval
- ✓ ANOVA + F test
- ✓ Chi-Square test
- ✓ Z test
- ✓ T test
- ✓ Cramer's V
- ✓ Point Biserial Correlation

- Probability
  - Bayes
  - Naive Bayes
  - Gaussian Naive Bayes

* Categorical and Categorical variables relations. → Chi square test

goodness of fit

Day	M	T	W	Th	F	S
Expected %	10	10	15	20	30	15
Observed	30	14	34	45	57	20
Expected	20	20	30	40	60	30

Expected amount of customer is told by the shop owner and observed is the actual number of customers visiting the restraunt we need to see how good is our hypothesis with regards to the observed customers.

H_0 : Owners distribution is correct

H_1 : Owners distribution is not correct

Significance level = 0.05 (5%)

chi square statistics $(\chi^2) = \sum \frac{(O-E)^2}{E} = 11.44$

Degree of freedom = 5 (there are six values in datapoint)
6-1

Critical Chi-square value from the chart = 11.07 $= (\chi_c^2)$

$11.44 > 11.07$ hence null hypothesis is rejected
i.e Expected vs are less likely

What is chi-square distribution?

Normal distribution $(0, 1)$
mean → σ^2 (var)

$Q_1 = \chi_1^2$ → Sample random values from this normal distribution and square them

Chi-square distr
DOF = 1

$Q_2 = \chi_1^2 + \chi_2^2$ → take two samples of values from this normal distribution and square them & uptil to σ / uptil to 2σ

chi-square distribution
DOF = 2

$Q_3 = \sigma_1 2\sigma_1 3\sigma$

and like this the distribution are there.

Categorical & Categorical Variable relations

	Herb1	Herb2	Placebo	Total
Sick	20 25.3	30 29.4	30 25.3	80 21%
Not Sick	100 94.7	110 110.6	90 94.7	300 79%
Total	120	140	120	380

These are observed value for herb1, herb2 and Placebo impact on person being sick & not sick.

Assume

H_0 = Herbs do nothing → If herbs do nothing then the % of sick/total and not sick/total will remain the same irrespective of herb. Hence $80/350 = 21\%$

H_1 = Herbs have an impact

$\chi^2 = \sum \frac{(O-E)^2}{E} = 2.53$ $\quad \alpha = 0.1\%$

DOF = $(C-1)(R-1) = (3-1)(3-1) = 2$

(C = number of columns, R = number of rows)

~~χ_c^2~~ → $\chi^2 < \chi_c^2$ → 4.60

* Hence we does not reject the null hypothesis.

ANOVA (Analysis of Variance)

	m		
	1	2	3
	3	5	5
	2	3	6
n	1	4	7
$\bar{\bar{x}}$	2	4	6

Sum of Squares = $\frac{(3-4)^2 + (2-4)^2 + (1-4)^2 + \cdots}{\cancel{2}}$

$= 30$

grand mean → mean of all of stuff

$\bar{\bar{x}} = \frac{3+2+1+5+3+4+5+6+7}{9} = \frac{36}{9} = 4$

Variance = ~~SST / (m-1)(c-1)~~ SST / dof

$dof = m.n - 1$

$= \frac{30}{8}$ * = ~~7.5~~ 3.75

SSW (Sum of squares within) $= (3-2)^2 + (2-2)^2 + (1-2)^2 + (5-4)^2 + (4-4)^2 + (3-4)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2$

$= 6$

∴, 6 out of total 30 variation comes from within these groups.

$m \cdot (n-1)$ d.f → 6 dof.

SSB (variation b/w samples) = $(2-4)^2 + (2-4)^2 + (2-4)^2$ or $3(2-4)^2$ (column mean, total mean)

$+ 3(4-4)^2 + 3(4-6)^2$

$= 12 + 0 + 12 = 24.$

24/30 variation b/w samples.

dof = $m - 1 = 2$

∴ Inferencing from ANOVA is regarding the actual difference in the population means or distributions or just a difference in the means.

H_0 → means are the same or food doesn't make a difference

H_1 → Food does make a difference.

Assume H_0:

F-statistic = $\left[\left(\frac{SSB}{m-1}\right) \Big/ \left(\frac{SSW}{m(n-1)}\right)\right]$

chi sqm distri 2dof

chi sqm distri 6 dof

If this number is big means there will be a differ b/w the population mean of the sample.

harder to reject null hypothesis

if this is small difficult to reject null hypothesis.

= 12

Calculate critical f value = 3.46

Significance level = ~~0.01~~ 0.1

f value > critical f value

reject null hypothesis

probably a differ in population means.

T-test - Applied to one numerical variable.

↓

check mean of the sample reflects good with population

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$ ← std dev of sample

t statistics help us get in confidence intervals where the mean of various samples of populations will fall with this range. This number improves highly with t statistics and z statistics fails to improve the results on this number.

Conditions for t-statistics
- Random
- normal — Central limit theorem — $n \geq 30$ – Normal distribution
- Independent observation

T statistics tells you about the plausibility of ~~one~~ a value in a distribution considering sample mean.

Foreg. if t stats give a range of -95, 105 and we expect the average thickness of 110 then our sample and population might not meet the standards.

Cramer's V → two categorical variables.

$$\text{Cramer's V} = \text{sqrt}\left(\chi^2 / [n(q-1)]\right)$$

$q = \min(\text{rows}, \text{columns})$

- 0 : not associated (values)
- 1 : perfectly associated
- 0.4 : weakly associated
- 0.7 : strongly associated

Point Biserial Correlation (Continuous + Categorical)

Continuous variable — Y

Binary variable — X

Z-score + Z test

of std dev away from pop^n mean for a particular datapoint.

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

Length of wings of turtles	Z-score
2	-0.59
2	-0.59
3	0
2	-0.59
5	1.3
1	-1.3
6	1.77

Why? → how usual or unusual a certain data point is

→ Outliers.

$\mu = 3$ $\sigma =$ 1.59

Is there a diff b/w dist A & dist B

Support new law	District A	District B	Total
Yes	58	52	110
No	42	48	90
Total	100	100	200

$\alpha = 0.05$

- Random
- Normal
- Independence

H_0 → There is no difference b/w Dist A & Dist B

H_1 → There is a difference b/w two districts

$$\sigma^2_{\hat{P}_A - \hat{P}_B} = \sigma^2_{\hat{P}_A} + \sigma^2_{\hat{P}_B}$$

$$\approx \frac{0.55(1-0.55)}{100} + \frac{0.55(1-0.55)}{100} \quad \left(\text{Considering } H_0 \text{ as true}\right)$$

100 → size of A; 100 → size of B

$$\sigma^2_{\hat{P}} = \frac{P(1-P)}{n}$$

→ In this case they are equal but may not be always.

$$\sigma_{\hat{P}_A - \hat{P}_B} \approx \sqrt{\frac{0.55(0.45)}{50}} \approx 0.07$$

$$\text{Z-score} = \frac{\hat{P}_A - \hat{P}_B}{\sigma_{\hat{P}_A - \hat{P}_B}} = \frac{0.06}{0.07} \approx 0.86$$

P (at z-score) → find using the table
value

if Pvalue > significance level fail to reject our null hypothesis. There is not enough evidence to suggest that myopia becomes more common over time

Confidence Interval

When we put a confidence interval of 95% i.e $\mu \pm 3\sigma$ this means that when we sample random data from the population true mean of the pop^n will be within this range of $\mu \pm 3\sigma$. 95% of the time.

Significance level

It talks about the comparision of conditional probability with regards to the significance level (α). if pvalue $< \alpha$ reject H_0 otherwise ~~else~~ do not reject H_0.

Bias vs Variance

error of fit ↑↑ → underfitting

error of fit ↓↓ → overfitting

Low Bias
High Variance

High Bias
High Variance

Best model

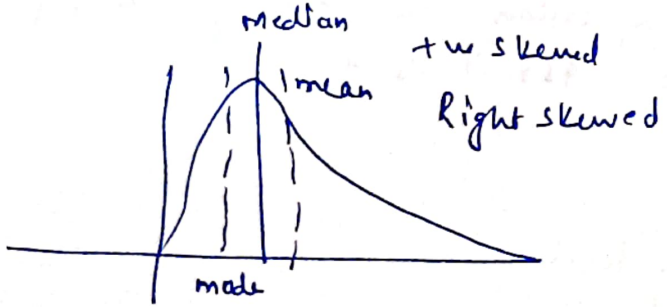
low Bias

low Variance

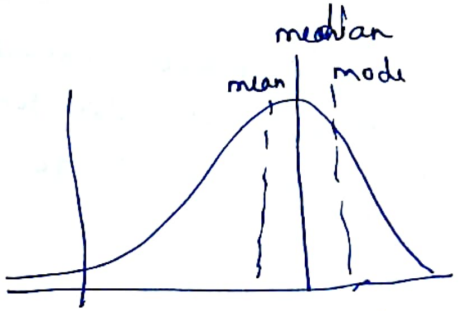
error of training data

error of test data

Data skewness



+ve skewed
Right skewed



-ve skewed.
left skewed