DATE : 2/9/22
PAGE NO. :
Friday

Normal Distribution

(1SD) (2SD) (3SD)

- 68% - 95% - 99.7% rule (use z-table)

z → score away from mean in terms of SD

z = 0 → on mean → divides population ~~[illegible]~~ in half.

- Dist" of height with mean at 67in & SD 3in. what % are b/w 63 & 67.

$$z = \frac{x-\mu}{\sigma} = \frac{x-67}{3} \quad \therefore \frac{63-67}{3} = \frac{-4}{3} = -1.33$$

& (0)

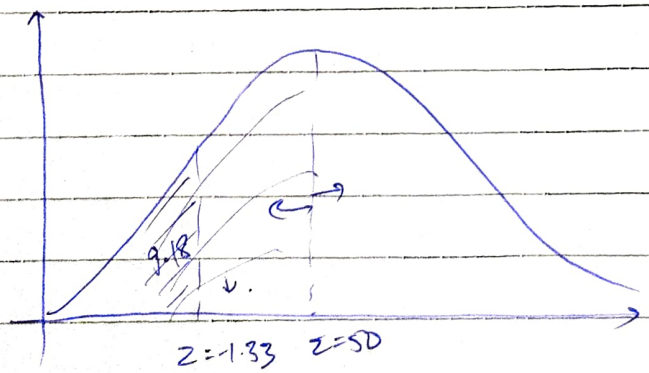
→ 0.03

↓

-1.3 → 0.0918 → ×100 = 9.18% to the left of 63. (less than 63)

0.0 → 50%

∴ 50 − 9.18 = 40.82 ← answer



9.18 90.82

100.00
− 9.18
90.82

DATE: / /
PAGE NO.:

Q: What is the 40% percentile of heights?
→ 40% ppl following you!.

Find 0.4000 in table ← that z-score is answer.

- Monthly rents in neighbourhood
  Mean = $900
  SD = $600 →
→ No bound on monthly rent
→ Is it a normal distr? → No. (Reason)
→ SD is very large, order of mean
→ ∴ Dispersion of data is huge.
→ Normal distr! atleast 3 SD to one side of mean should be in range
→ ∴ For this, t-table / stats is used.
→ In these cases, we can use chebyshev inequality for bound
→ Good bound is better than a very weak approximation

- Binomial approximation with large values of n
  - approxed by a Normal distr
- Mean = np, Variance = np(1-p), draw Normal curve

DATE : / /
PAGE NO. :

Central Limit Theorem

- Aggregate (sum) of large no. of independent random vars, leads to a ~~no~~ random var. that is approximately normal.

- If sample size > 30, whatever be the dist^n, sample mean distr will be a normal dist^n.

Sample mean dist^n: Take the same-size sample, randomly, many times, and plot the distribution of all the means in each sample-set.

- As sample size ↑↑, we get closer to normal mean
- sample size = 1 → Mean = Actual mean.
- Mean of sample mean dist^n = mean of actual popul.

What abt their SDs?

→ Not the same

→ SD of sample < SD of popul.

Reason: In samples, we are evening out the values closely.

$$\sigma_{\text{sample mean dist}^n} = \frac{\sigma}{\sqrt{n}}$$

σ ← SD of actual dist^n.

n ← sample size

↓

(Standard Error)

DATE: 5/9/22
PAGE NO.:

Inferential statistics

Population: collection of same units
Parameter: A number we're interested in about popul.
Sample: Subset of populat.
Estimate: guess for parameter, calculated from sample
- Estimate is good only if sample is good
- Sample should avoid Bias

~~Bias~~

- To avoid bias: cover larger area, select elements for sample randomly.
- sample should be representative of all complexiti of the population

Selection Bias: - literary Digest used Phone Book.
- Phones were a luxury at that time
- : Biased towards ppl with phone

Non-response Bias:
- only 2.4 mn responded out of 10 mn.
- Losing almost 76% of ppl

- Sample should not be a sample of convenience
- Should draw uniformly at random with or w/o replacement from population.
- Simple Random Sample:
  - Random sample
  - all units equally likely
- Random Sample
  - all units need not be equally likely

DATE : ~~5/9/22~~
PAGE NO. : ~~Monday~~

Q: A population has 4 ppl A, B, C, D.
A is chosen in sample & other 3 are chosen only if heads on coin toss ⟶ Biased towards A.

Q: In a pop. avg age = 37, SD = 15.
sample size = 200 taken from pop
Avg. age for sample = ?

Sol^n: Expected Age = 37

$$\text{Standard Error} = \frac{15}{\sqrt{200}} = 1.06$$

$$\frac{SD(\sigma)}{\sqrt{n}}$$

Q: Sample size = 100, Avg = 38, SD = 14.
Avg age of population ?

Sol^n: Estimated age = 38

$$\text{Std. error} = \frac{SD}{\sqrt{n}} = 1.4.$$

- SD is for sample here
- We want SD of population
∴ contradiction!

But it is correct coz of assumption:

Bootstrap: A large sample size is like a population, then SD of the sample is an approximation to the SD of population.

7.42
8.7 8.22 SP 7.1 7.83

Confidence Intervals

DATE: / /
PAGE NO.:

Q: Sample of 625 households,
avg. income = $63000,
SD = $40,000
Find an approximate 95% confidence interval for the avg income of households.

Sol^n: Std error $= \frac{40000}{\sqrt{625}} = 1600$ (previous method)

For 95% accuracy : need to see (2) SD away from mean

$\therefore 63000 \pm 2 \times \frac{40000}{\sqrt{625}}$

$= 63000 \pm 2 \times 1600$

$= 59800$ to 66200 (confidence interval)

$\therefore$ 95% of time, ~~it will fall into this range, if a house is picked at random.~~ mean will fall into this range

Given dist^r is not normal, but how did we use normal dist^n approximation for a normal dist^r

- Sample mean dist^n by central mean th. says that it is a normal dist^r.

On Confidence Intervals

DATE : / /
PAGE NO. :

- Good confidence Intervals
  - on which, decisions can be made
- Bad : when claim does not match the actual position of mean.

Testing Hypothesis

- Toss coin 15 times → 10 heads
  → 5 tails

- Coin is fair : H_0 → null hypothesis
- Coin is unfair : biased towards 'head' - H_A

H_A : Alternate Hypothesis

This is a 1 side test (Don't see if biased towards tails)

- Assume H_0 is true & calculate chance of H_A, i.e. above obs. of $\frac{10 \text{ heads}}{15 \text{ toss}}$ = p-value

Method

If p-value < 5% (small) → choose H_A
We call it statistically significant
O/W (P-value > 5%), no change in our assumption that H_0 is true.

- P(10 heads / 15 tosses) < 5% → means prob was less than 5%, but it still happened, that means coin is unfair.

Q: Red-flowering = 25% in each plant
Data: 400 plants, 88 are red-flowering
Is data in sync with theory?

Sol^n: . $H_0 = p = 0.25$, $H_A = p < 0.25$

. P value: binomial with $p = 0.25$ & $k \leq 88$

$\Rightarrow$ P = 9.08%

of 400 plants, each → red (0.25)
→ not red (0.75)
$k \leq 88$

since p (9.08) > 5% → hence theory (25%) is fine.

This is called exact binomial test.

8: 25% of chance of red flowering
Data: 400 plants, 88 → red

DATE: 7/9/22
PAGE NO.:
Wednes

Ans:
- $z = \frac{88.5 - 100}{8.66} = -1.328$
- Get area under curve
- $P = 9.21\% > 5\%$ → Theory looks good
- This is one-sample z-test

8: 16000 Jurors → 26% of them are black
Data: 100 men on Jury → 8 are black.
Is data % in sync with theory %?

Sol^n:
- Selection of 100 men is like a trial conducted 100 times with selecting black as 1 & white as 0.

$\mu = 100 * 0.26 = 2$

$\sigma = \sqrt{100 * 0.26 * 0.74} = 4.39$ $(\sigma = \sqrt{npq})$

Calculate z.

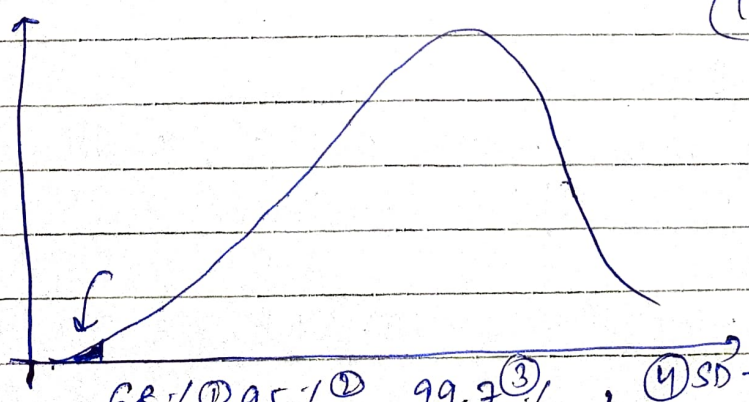
$x = 8$, how far is it from 26. in terms of SD.

$$z = \frac{8.5 - 26}{4.39} = -3.99 \approx -4$$

∴ z is almost 4 std error away from the mean.

→ Extreme cond^n

∴ $\leq 5\%$ → Not in sync with given (16000 - 26%) hypothesis



Significance value: level of tolerance for error.

Here, by default, we consider significance value = 5%.

$\therefore p < 5\%$ → alternate hypothesis considered.

W.R.T. the given sample, & the significance val, we conclude that theory is unfair.

Q: Coin tossed 20 times - H_0 : $p = 0.5$ & H_A : $p = 0.8$
If no. of head is more than 14, then select H_A, o/w H_0

Sol: Possible Scenario

Reality \ Observation	$p = 0.5$	$p = 0.8$
$p = 0.5$	correct	Type-1 error Prob = 5.8%
$p = 0.8$	Type II error	correct Prob = 91.3%

Type 1: Significant level

Type 2: ~~Power~~.

Calculate Type I error

Binomial, $n = 20$, $p = 0.5$, $K \geq 14$.

Prob = 5.8% → significance level

Type II

Reality

Binomial, $n = 20$, $p = 0.8$, $K \geq 14$

Prob = 91.3% → Power

$\therefore$ Type II error = 100 - 91.3% = 8.7%

DATE : / /
PAGE NO. :

- Significance & Power are inter-related
- Want to minimize error
  If error ↓, power ↓ & vice-versa.

Neyman-Pearson Lemma

Design the test such that we maximize power, given the fixed significance level.

— See slides —

Eg: Obs: 400 coin toss, 225 heads seen
H_0 : coin is fair, $p = 0.5$
Under H_0, expect 200 heads with std error of 10 ($\sqrt{np(1-p)}$)

1st alternative : coin is biased towards → $p > 0.5$
2nd : coin is not fair → $p \neq 0.5$
3rd : coin is biased to tail → $p < 0.5$

Find P-value of all above cases.

(1)
— chance of getting 225 heads or more, assuming coin is fair.

$$z = \frac{225 - 200}{10} = 2.5$$ ← significant

$\therefore P = 0.71\% < 5\%$

$\therefore$ can say biased towards head.

This is 1-tailed test

DATE: / /
PAGE NO.:

(2) no. of heads as ≥25 more or less the expected value of 200, assuming fair

$z = \pm 2.5$, $p = 2 * 0.71\% = 1.42\%$,

$p < 5\% \rightarrow$ choose 2nd alt.

called - 2-tailed test.

(3) chance of getting heads = 225 or fewer, assuming coin is fair

$$z = \frac{226 - 200}{10} = 2.6$$

$p = 99.4\% > 5\%$, so choose H_0

→ This is another 1-tailed test.

---

z-test

Q: Avg. height of Indian men = 175 cm
Data: simple random sample of 100 studs taken from univer. in India with 10k(s) stu—.

Avg height of sample = 174 cm
Std. Dev = 5 cm

Q: Is avg height of students shorter than Indian men?

DATE : 9/9/22
PAGE NO. :

t-test

For P-value, find t & look at t-table.

df : sample size = 5
: see row of df = 4.

Now see, t = 1.79
it is b/w 1.533 & 2.132 (in that row)

why ÷ by n-1, not n.

Two-Sample Test

Eg⊙ Take 2 samples of egg, 1 before ↑ing temp^r, 1 after temp^r ↑se, compare their qualities

Method:

- H_0; $\mu_1 = \mu_2$ & H_A : $\mu_1 \neq \mu_2$
- Assume H_0 & calc. z.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

($\bar{x}_1$: mean of sample 1)

n_1: sample size of 1st sample

d_0: $\mu_1 - \mu_2$ in 'H_0'
here $d_0 = 0$.
& most cases

- Look at z-table, find P-value. If P < 5% ⇒ discard the null hypothesis
- Unpaired Test : different sample items, diff sample size.
  Paired Test : same sample items in both situations

DATE : / /
PAGE NO. :

If sample size too small, use t-test

- If variance of both samples are same

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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

$\rightarrow d_f = n_1 + n_2 - 2$

degree of freedom

- If variance diff

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$d_f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\quad\right)}$$

DATE : / /
PAGE NO. :

ANOVA (More than 2 samples)

Mean of all 4 samples = $\bar{\bar{y}}$

Mean of i^{th} sample = $\bar{y}_i$

$n \rightarrow$ sample size ~~(total)~~

$y_{i,j} \rightarrow$ individual value

$\bar{y}_i \rightarrow$ mean of that particular sample.

$N \rightarrow$ total items, ~~including~~

$= na.$

F-score = MSB / MSE

x^2 - chi-square Distribution

- By Aloke Dutta

DATE 12/9/22
PAGE NO.:

$\mu_x = 130$ (f1) $\quad \mu_y = 60$ (f2)

height = x	weight = Y	$x - \mu_x$	$y - \mu_y$
130	55	0	-5
135	60	5	0
125	65	-5	5
140	70	10	10
120	50	-10	-10

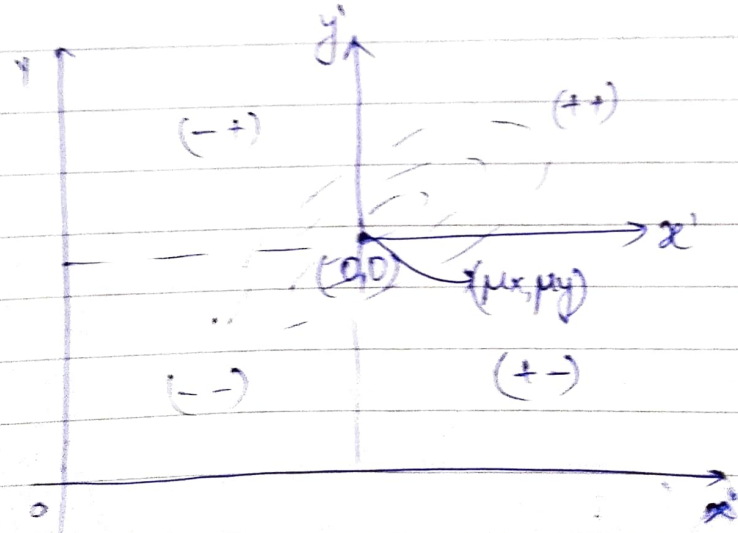
$\Rightarrow$

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} p_i x_i$$

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

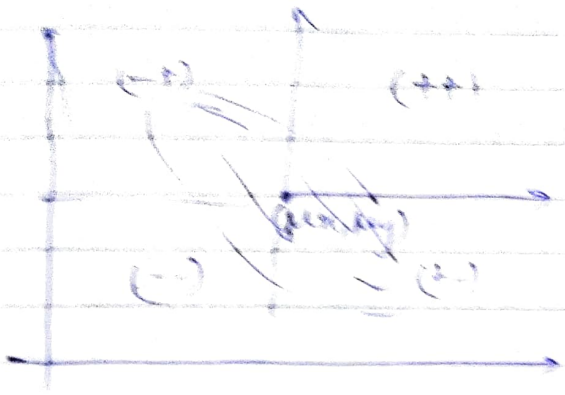
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)$$

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$



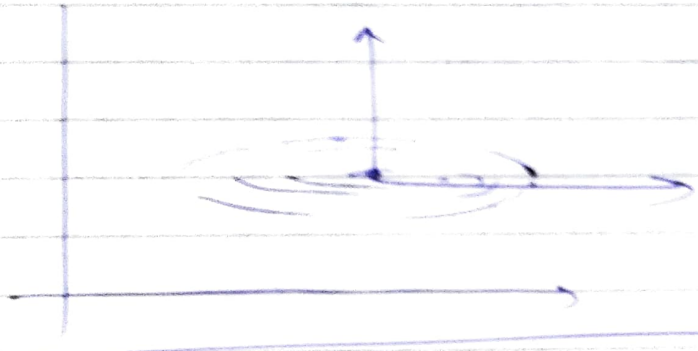
+ve values are larger (more values in 1st & 3rd quadrant)

$\Sigma x_i y_i > 0 \rightarrow cov(x,y) > 0$ ($x \& y \rightarrow$ directly prop)



← here, more ve values

$\Sigma x_i y_i < 0 \rightarrow cov(x,y) < 0$ ($x, y \rightarrow$ inverse prop)



$\Sigma x_i, y_i \approx 0$

$cov(x,y) \approx 0$

correlation $\rightarrow \rho(x,y) = \frac{cov(x,y)}{\sqrt{var(x)\, var(y)}}$

$\rho(x,y) \in (-1, 1)$

Let there are D attributes
f_1(height), f_2(weight), ... f_D( )

- Need to calculate rel^ b/w all pairs of attributes

$$\Sigma = \begin{matrix} f_1 \\ \vdots \\ f_D \end{matrix} \begin{bmatrix} & & f_1 \cdots f_D \\ & \bigcirc & \\ & & \end{bmatrix}_{D \times D}$$ ← No. of attributes

each cell = covariance of (f_i & f_j)

→ $\underbrace{cov(f_i, f_i)}_{\text{diagonal elements}} = var(f_i)$

→ $cov(f_i, f_j) = cov(f_j, f_i)$

Σ → Dispersion Matrix → symmetric wrt diagonal elements

→ Also called, covariance matrix.

\# Σ is real symmetric Positive Definite Matrix

Real: all elements are real values (never imaginary)

Symmetric: $cov(f_i, f_j) = cov(f_j, f_i)$
upper & lower are sym wrt diagonal

Positive Definite:
A is +ve definite if

$\vec{x}^T A \vec{x} > 0, (\forall\ x)$

DATE : / /
PAGE NO. :

Eg: $A = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}$ → Is it +ve definite

Let $\vec{x} = \begin{bmatrix} a \\ b \end{bmatrix}$ $\{a, b \in R\}$

$\vec{x}' = [a \quad b]$

$\therefore \bar{x}' A \bar{x} = [a \quad b] \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$

$= \begin{bmatrix} 2a-b \\ 3b-a \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$

$= [(2a-b) \quad (3b-a)] \times \begin{bmatrix} a \\ b \end{bmatrix}$

$\therefore 2a^2 - 2ab + 3b^2$ → D

$= 2\left(a^2 - \frac{2ab}{2} + \frac{b^2}{4}\right) + \frac{5}{2}b^2$

$= 2\left(a - b/2\right)^2 + \left(5/2\right)b^2 \geq 0$

→ $|\Sigma| > 0$ (Determinant of $\Sigma > 0$)

→ Σ^{-1} also be +ve definite matrix

→ All the eigen values of Σ are ≥ 0.

$A \underline{x} = \lambda \underline{x}$

$|A - \lambda I| = 0$

↳ gives values of (λ) → eigen values.

$A_{D \times D} \; x_{D \times 1} = \lambda_{1 \times 1} \; x_{D \times 1}$

(Scalar)

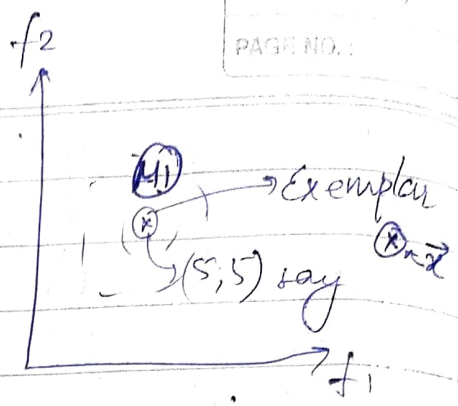
↓

D values of λ, D eigen values $(\lambda_1, \lambda_2 \ldots \lambda_D)$

corresponding eig. vectors $(v_1, v_2 \ldots v_D)$

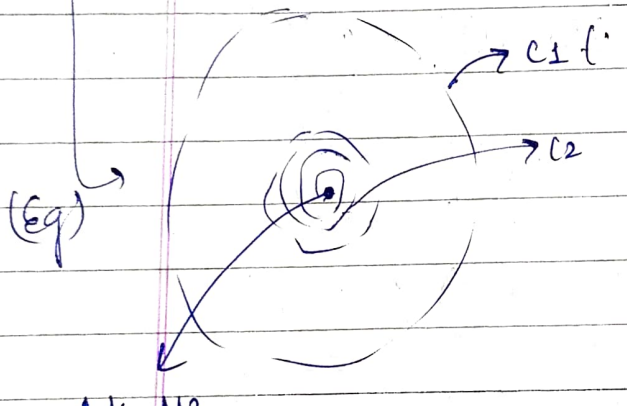
$d_2^2(\mu_1, \vec{x})$

If distance b/w $\vec{x}$ & μ_1 is very small



→ If μ_1 is not representing the class sufficiently, cultivate 2nd order stats of class,

eg: variance



$$d_2^2(\mu_1, \vec{x}) = (x-\mu_1)^T (x-\mu_1)$$

$$(c_1, \mu_1, \Sigma_1)\ d_M^2(\mu_1, \vec{x}) = (\underline{x}-\mu_1)^T (\Sigma_1)^{-1} (\underline{x}-\mu_1)$$

(Mahalanobis distance)

$$d_M^2(\mu_1, \mu_2) = (\mu_1-\mu_2)^T \Sigma^{-1} (\mu_1-\mu_2)$$

where c_1 c_2

$$\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$$

DATE : 13/9/22
PAGE NO. :
Wed.

f1	f2
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Cos difference

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$X \cdot Y = |X||Y| \cos\theta$. (Dot product)

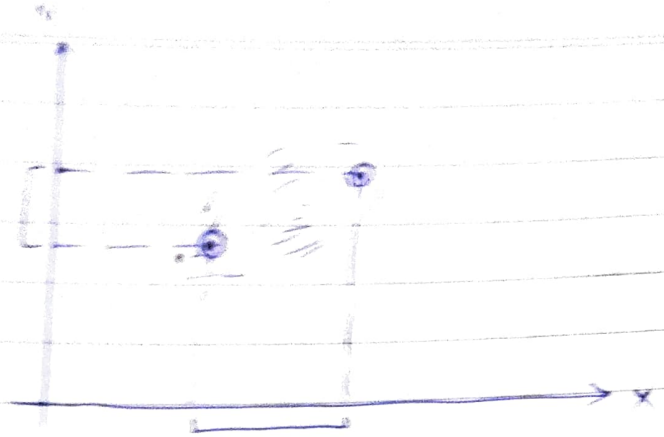
$$\therefore \cos\theta = \frac{X \cdot Y}{|X| \cdot |Y|}$$

$$= \frac{X^T \cdot Y}{\sqrt{\sum_{i=1}^{D} x_i^2} \sqrt{\sum_{i=1}^{D} y_i^2}}$$

Eigen vectors are perpendicular to each other.

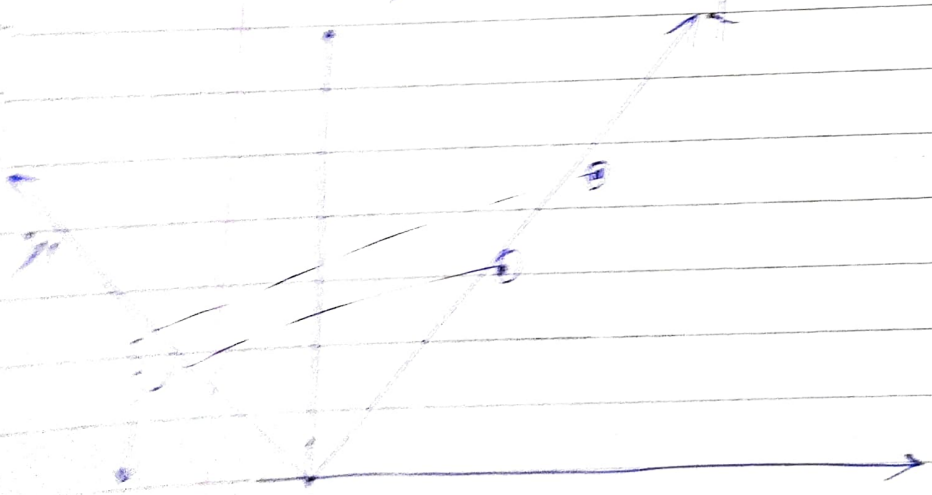
$\therefore v_i \cdot v_j = 0$.

- All dimensions are I.r coz effect of element of 1D should be zero on element of another dimension



- If variance is large wrt a dimension, it becomes easier to differentiate b/w the elements coz separadness ↑ ..

- We transform dataset into other dimensions to differentiate among patterns easily.



variance of these points in this dimension

DATE : / /
PAGE NO. :

1. Calculate Dispersion Matrix Σ of given dataset.

$X \rightarrow$ dataset $\rightarrow n \times D$ (n: entries, D: attributes), $X^T \rightarrow D \times n$

a) Subtract mean from each data in the data set. (μ)
(Making mean-subtracted data)
(Mean of new dataset = 0)

b) $\Sigma = X^T \cdot X$

c) Calculate eigen values and eigen vector of Σ

* trace (Σ) : sum of diagonal elements
: sum of variance wrt each attribute

trace (Σ) $= \sum_{\text{(sum)}} \lambda_i$, all $\lambda_i \geq 0$

• If arranged,

$\lambda_1 > \lambda_2 > \cdots > \lambda_D$
$v_1 \quad v_2 \quad \cdots \quad v_D$ ← corresponding eigen vectors

↳ all are $\perp^r$ to each other

$$\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i} > 99\%$$

∴ We can avoid 'd to D'
→ Not important in terms of variance

(Determ.-inant) $|\Sigma| = \lambda_1 \times \lambda_2 \times \lambda_3 - - \lambda_D$

$\lambda_1 \quad \lambda_2 \quad \lambda_d \quad \lambda_D$

$\updownarrow$

$v_1 (D \times 1) \quad v_2 \quad v_d - - v_D$

$\rightarrow d$

Transformation Matrix (W) contains all selected eigen vectors

$$W_{D \times d} = \begin{bmatrix} v_1 & v_2 - - & v_d \\ & & \end{bmatrix}_{D \times d}$$

$$x_1 = \begin{bmatrix} \\ \end{bmatrix}_{D \times 1}$$

$Y_1 = W^T \cdot X_1$
$(d \times 1)$
→ smaller dataset

- Variance (Y_1) > Variance (X_1)
  ↑ (Transformed)