- Head first statistics Dawn Griffiths
- Introductory statisics Neil A Weiss

Chap 1 → Introductory statistics

- Collecting data → summarize → vizualization → analysis & making inferences

Descriptive statistics v/s Inferential statistics

- Abt summarizing data set & visualization
- Predection of some data in data set

Ex Flower classification → (Inferential statistics)

- distinguish rose & lotus 50 roses & 50 lotus
- we have patel length & width for each recorded
- we now have a new flower given patel length = a units & patel width = b units, determine species of flower

- We find object with nearest data and this flower belong to this category.

optimize

- find centroid of data for features & find nearest catego

Mean Median Mode

- Just compare it with centroid of diff. category & interprete.

Descriptive Statistics →

- Measures of central tendency → mean, median, mode
- Measure of varience → range, IQR, percentile.
- Data visualization → piechart, Box plot, Bar charts, Stem & leaf diagram, surface plots. → use matplotlib library in python

- Observational study → based on observation only
- Experimental study → based on experiment done on some data grp

Inferential Statistics →

- regression - relationship b/w dependent & independent variable
- naive Bayes classification

$$y = w_0 + w_1 x + w_2 x^2$$

↳ Polynomial regression of order 2

Ex. Need to predict age given value of Blood suger, weight, height
∴ we need to establish a function for that we make assumption known as hypothesis

let Age = $w_0 + w_1 \cdot \underset{x_1}{BS} + w_2 \cdot \underset{x_2}{Weight} + w_3 \cdot \underset{x_3}{height}$.

or y

let $y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 \cdot x_1^2 + w_5 x_2^2 + w_6 x_3^2 + w_7 x_1 x_2 + w_8 x_2 x_3 + w_9 x_1 x_3$

- optimal order = 1 more than number of turns

- Sample should be with no bais & ideal group should be selected from population
- sampling technique → random sampling, systemetic random sampling, cluster based sampling.

- Discrete variable → table (Prob. dist.)
- Random variable → probo distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Tally mark :-

Ex 1 1 2 2 2 3 3 3 4 4 4 5 5 1 1 2 2 3 5 6

→ One pass

1	||||
2	𝍸 |
3	||||
4	|||
5	|||
6	|

- Write a prog to compute mean & standard deviation (σ) ?

- We just need 1 pass.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} = \frac{\sum (x_i^2 - 2x_i\mu + \mu^2)}{n}$$

$$= \sum_{i=1}^{N} \frac{x_i^2}{N} - 2\mu \sum_{i=1}^{N} \frac{x_i}{N} + \frac{N\mu^2}{N}$$

$$= \sum_{i=1}^{N} \frac{x_i^2}{N} - 2\mu^2 + \mu^2$$

$$= \sum_{i=1}^{N} \frac{x_i^2}{N} - \mu^2$$

$$= \sum_{i=1}^{N} \frac{x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2$$

→ Computation formula for standard deviation

- If we use above methode upon adding data our previous computation are not lost

- An algorithm is said to be incremental when u can update the model on availibility of additional data.

Design experiment →

- Experimental units
- treatment ⟶ 8
- factor ⟶ Irr. regin, polymer
- level of factor ⟶ Irr. regin ⇒ 4, polymer ⇒ 2
- response variable ⟶ wt. gain

Ex Impact of irregation regin & use of a particular polymer? for max weight gain of cactus plant

irregation regin → none low moderate high

polymer → yes no

	N	L	M	H
Y	T1	T2	T3	T4
N	T5	T6	T7	T8

Q. A ML algorithm named SPM has 2 free parameters named regularization parameter and kernal width parameter, σ. In orderto find optimal values of parameter we perform a grid search

$C \in \{2^{-18}, 2^{-16}, \dots 2^{0} \dots 2^{50}\}$ = 35 terms

$\sigma \in \{2^{-18}, 2^{-16}, \dots 2^{0} \dots 2^{20}\}$ = 20 terms

for which accuracy is maximum

factors → C, σ

level of factor → C = 35, σ = 20

Response var → Accuracy

treatment → 35 X 20

Q. An exp. was conducted to study the impact of follic acid on birth defect, to perform this study a grp of 200 women considered, 100 women were given 10mg follic acid tablet & the rest of women were given trace element find out the factors, number of treatment experimental variable & response variable!

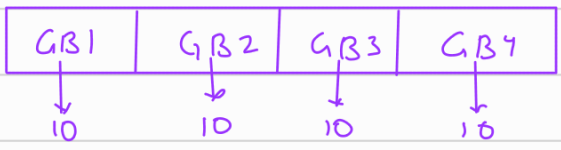
Exp. unit ⇒ subject (Biology) 200
Res. var ⇒ Child birth
factor ⇒ Acid
level ⇒ 2
Treatment ⇒ 2

Sampling → with replacement
Sampling → Without replacement

How to distribute experimental unit among treatments →

- Complete randomized design
- Randomized block design

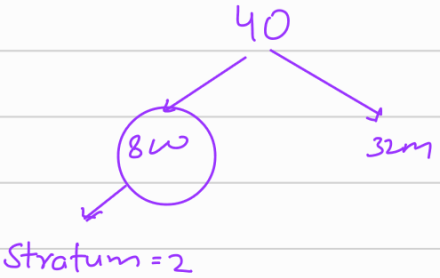
Q. Driving distance of golf ball - 4 Brands golf balls
their are 40 golfers 8 are women & 32 are men



8+2 10M 10M 10M

GB2 driving dist Less
As sample is not good.

- Sample should be representative of population
- For good same their shd be 8M + 2W



Sampling

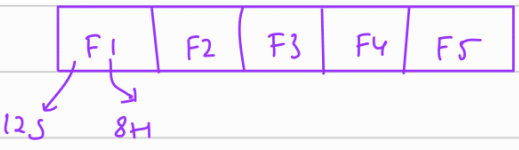
- SRS (simple random sampling)
  - without replacement
  - with replacement
    - same no more than once
- Cluster based
  - Packets instead of individual
- Systematic random sampling.
  - Sample size = ss
  - population size = ps
  - $m = \frac{ss}{ps}$
  - Generate random no k from 1 to m
  - Now my samples are k, k+m, k+2m, ...
- Stratified random sampling.
  - Population of subconcepts
  - 4 types of colors in trophy

- Cross validation used stratified random sampling
-

100 Emails → 80 training, 20 testing
- 60 S
- 40 H

Then how??

data of 5 equal parts & maintain ratio



	Train	Test
T1:	F1	F2 F3 F4 F5
T2:	F2	F1 F3 F4 F5
T3:	F3	F1 F2 F4 F5
T4:	F4	F1 F2 F3 F5
T5:	F5	F1 F2 F3 F4

m-classes

C_1	n_1
C_2	n_2
C_3	n_3
⋮	⋮
C_m	n_m

Find unique no. of experimental set up as per using k-fold cross validation.

$\underbrace{x_1\ x_2\ x_3\ x_4}_{R}\quad \underbrace{y_1\ y_2\ y_3\ y_4}_{L}$ = 8 instances

2-fold cross validation
↳ divide in 2 equal part

4 instances	4 instances
fold 1	fold 2

maintaing ratio.

Training	Test
fold 2	fold 1
fold 1	fold 2

- I/P in machine learning is data
- In regression target var. contineous & in classification target is discrete

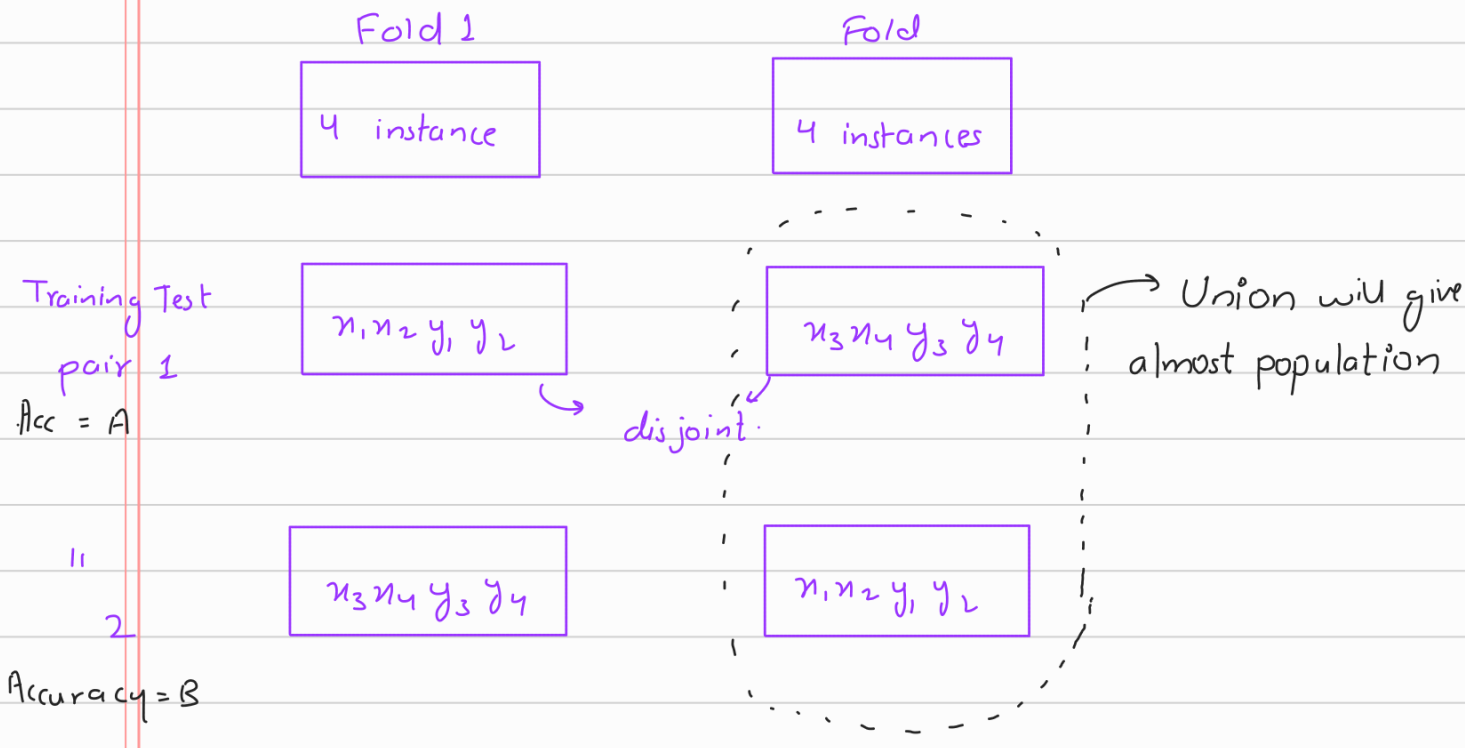
Training	
T	n
S	5'
T	6'
T	7'
S	4'

Test		
P	T	n
S	S	5'1"
S	S	5'2"
T	T	6'3"
T	T	6'7"

S* 4.5' T* 6.5'

S* 5'1.5" T* 6'5"

- Here test instance is easy
- To get better dct we use k (2) fold cross validation



$$\text{Net Accuracy} = \frac{A+B}{2}$$

$$\text{Unique} = \frac{{}^4C_2 \times {}^4C_2}{2}$$

Ass. →

$x_1 \text{ ---- } x_{10} \Rightarrow$ Class 1

$y_1 \text{ ------ } y_{10} \Rightarrow$ Class 2

find out unique experimental setups using 2&5-flod cross validation prove that no. of unique experimental setups for k-fold cross validation where the $C_1, C_2 \text{ --- } C_m$ are the m classes with $n_1, n_2, \text{ --- } n_m$ instances respectively is

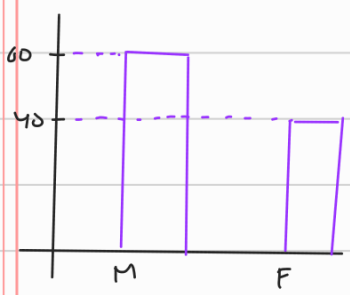
$$\frac{\prod_{i=1}^{m} {}^{n_i}C_{n_i/k}}{k}$$

	Not suitable
Mean	In presense of outliers (Teacher ↑ avg age of class)
Medians	(Parent with kid in swimming class)
Mode	(5 5 36 36)

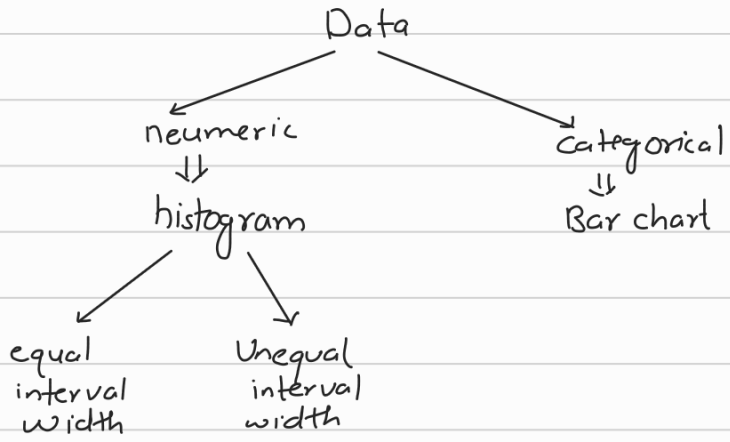
mode = $\frac{36+5}{2}$ = 20.5

Distribution →

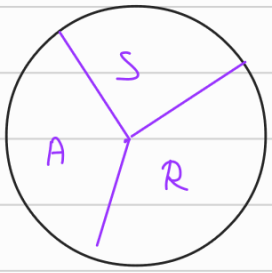
- bar-chart →



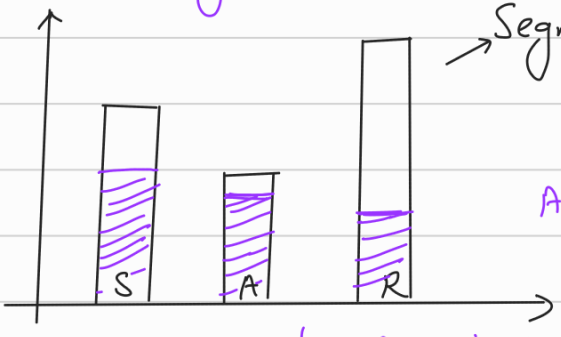
Height of bar ∝ Freq.



- Pie-chart →



How to plot a add on data
- Using barcharts



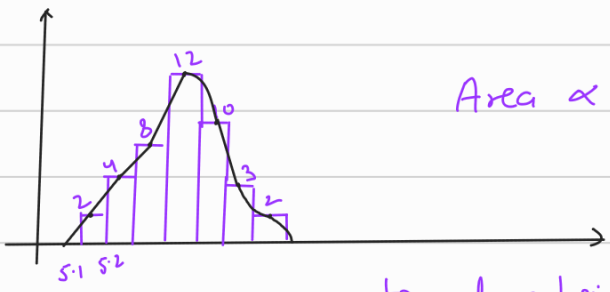
Add. data by coloring.

↳ Gaps in b/w barchart

- Histogram →

	freq
5 → 5.1	2
5.1 → 5.2	4
5.2 → 5.3	8

6.1 → 6.2	8
6.2 → 6.3	3



Area ∝ Freq.

- No gap b/w bar

- to reduce height we can use fraction
- join centre of histogram

Histogram →

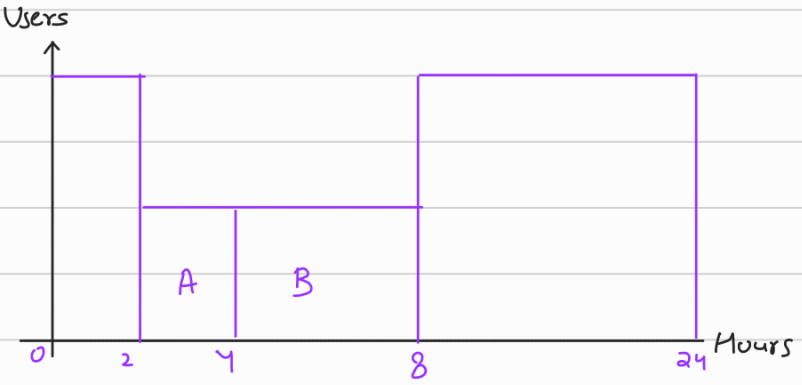
3000 Customers

0 → 2	1000
2 → 4	500
4 → 8	500
8 → 24	1000

lower limit, upper limit

[8, 24)

Bar chart →



B > A ?? NO Barchart drawback

(For Nominal or Categorical attribute)

Histogram ↓

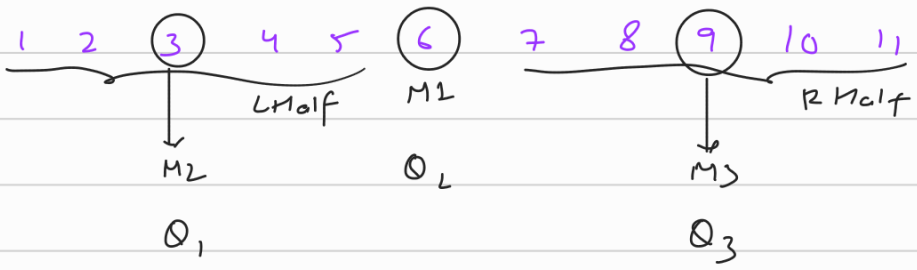
Class	freq	Class width	height
0 → 2	1000	2	500
2 → 4	500	2	250
4 → 8	500	4	125
8 → 24	1000	16	62..

Box plot →

Central tendency	Variance
→ Mean (fail in outliner)	→ trimmed range (Range - Outliner)
→ Median (fail 2 major class) (kid + parent)	→ range (It contaiant outliner)
→ Mode	→ IQR (for median) - Interqurtile range ⇒ 25% (Q_1), 50% (Q_2), 75% (Q_3); $IQR = Q_3 - Q_1$
	→ Varience

- both collectively ct & var. provide much more info about data

IQR →



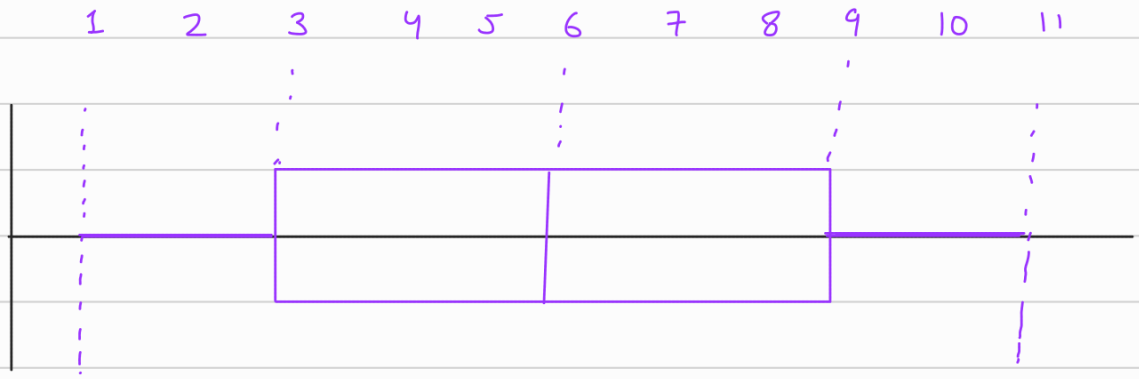
25% data lie below Q_1

lower bound = 1

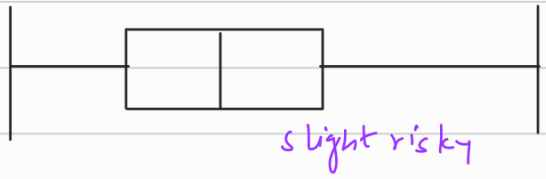
Upper bound = 11

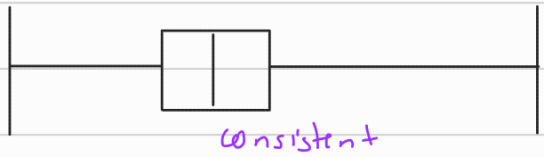
$Q_1\ Q_2\ Q_3$ = 3, 6, 9

Median = 6

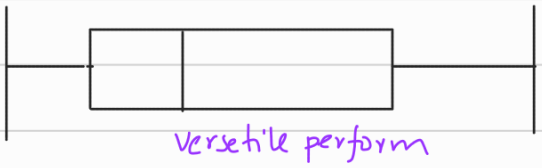


Q.





Whom to select??



Q Compare the profit of 5 companies for the month Jan-Apr 2023.

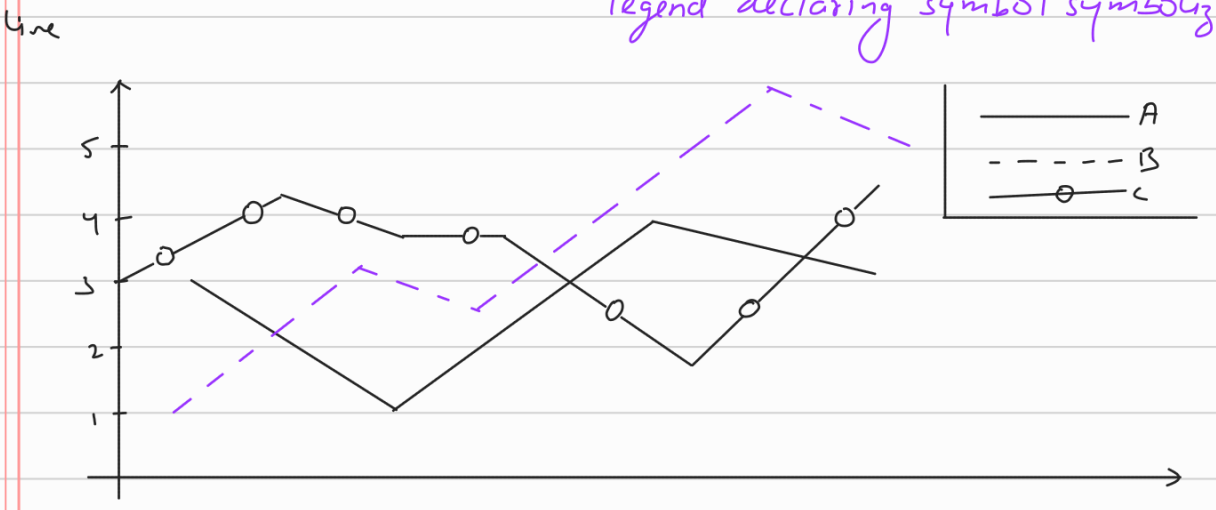
- What to use

(best)

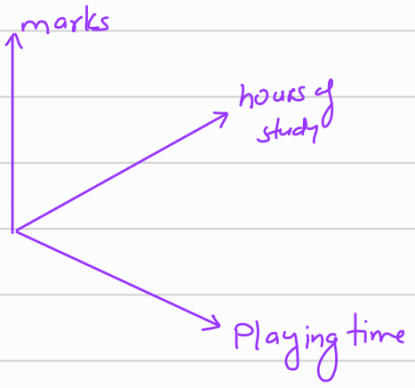
Piechart — line Chart → (diff line similar to below one)

Barchart — Scatter blots → (diff symbol for diff people)

Just plot point on x-y & we create a box on right corner called legend declaring symbol symbolization.



Surface plot 3D →



Normal distribution →

- Can be used to summarize numerical data.

↗ Parametric study (Assume data follow a particular prob. density function)
↘ Non-parametric

data ↗ Categorical/Nominal ⇒ prob. distribution table (coin HT table)
data ↘ Numeric ⇒ prob. density function

Temp-play →

Yes	No
Mild	Mild
Mild	
Mild	Mild
Mild	
Mild	hot
Hot	

summarize →

	Hot	Mild	Cool
yes (9)	2/9	4/9	3/9
No (5)	3/5	2/5	0/5

→ likelihood

Hot	hot
cool	
cool	hot

$P\left(\frac{\text{Hot}}{\text{Yes}}\right)$??

Numeric

Temp - Play	
Yes	No
31	41
32	42
33	43
⋮	44
39	45

Assume data follow normal or gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

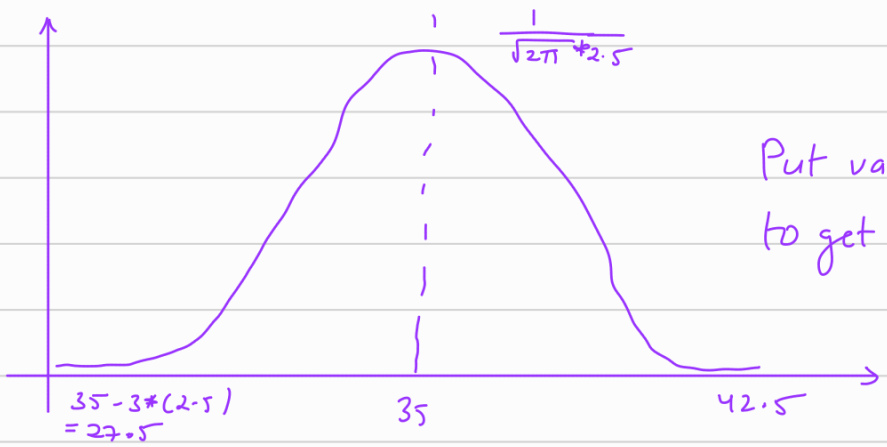
μ, σ = ??

$\Rightarrow$ temp = 36.5 play = ??

$\mu = 35$, $\sigma = \sqrt{\frac{60}{9}} = \frac{7.745}{3} = 2.5$

X = random var.

x = value of random var.



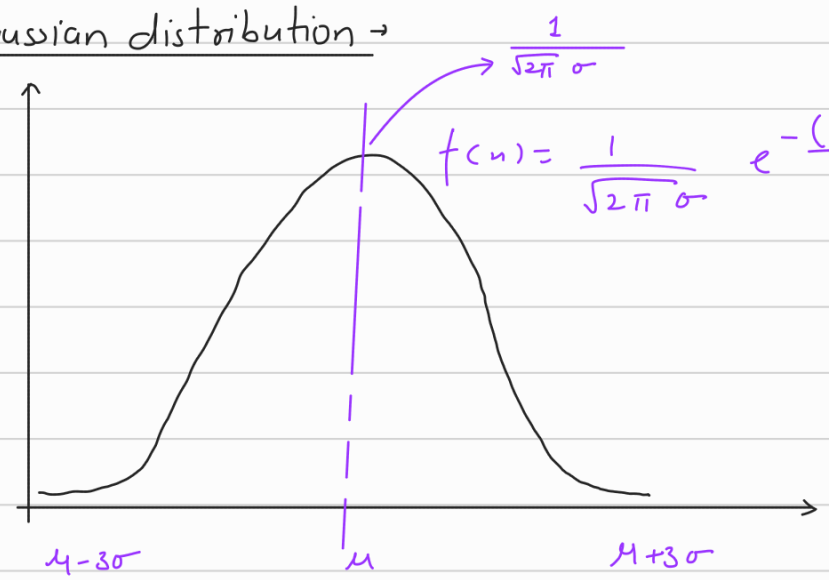
Put value in function to get ans.

Q. Compute the likelyhood of play = yes & play = NO when temprature is 40°c assume that the data follows normal/gaussian pdf

→ P (X < 40 & Play = Yes) = Area under curve till x = 40

- To solve Area under curve we convert this distribution into standard normal distribution ($\mu = 0$, $\sigma = 1$)

Gaussian distribution →



Emperical rule → (65-95-99.7 rule)

65% of entire population lie in range of $\mu-\sigma$ to $\mu+\sigma$
95% of entire population lie in range of $\mu-2\sigma$ to $\mu+2\sigma$
99.7% of entire population lie in range of $\mu-3\sigma$ to $\mu+3\sigma$

Standard normal distribution → (Z-Table)

Mean = 0 , Standard deviation = 1

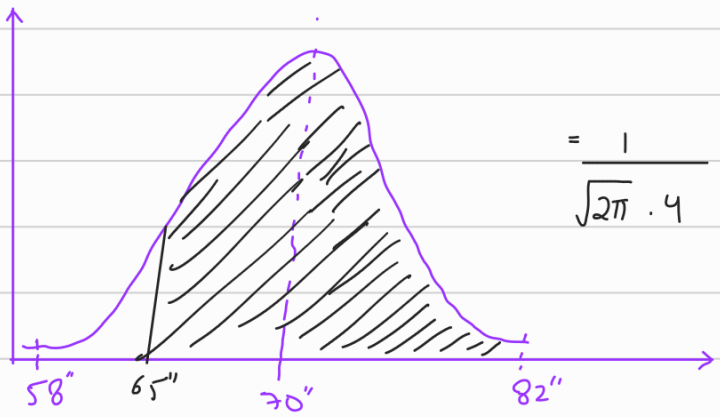
$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}} \xrightarrow[\sigma=1]{\mu=0} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

→ eqn of standard normal distribution

Q. She want to marry a man who is taller than her. The dis. of the height of man is given by $N(70'', 16'')$ Given girl height = 65".

μ σ^2

$\sigma = 4$



$$= \frac{1}{\sqrt{2\pi}\cdot 4}\, e^{-(x-70)^2/2(16)}$$

↓ integrate in range

or

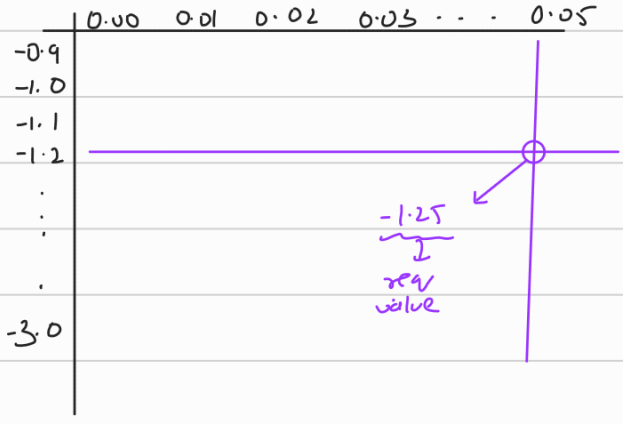
transform it to standard normal distribution

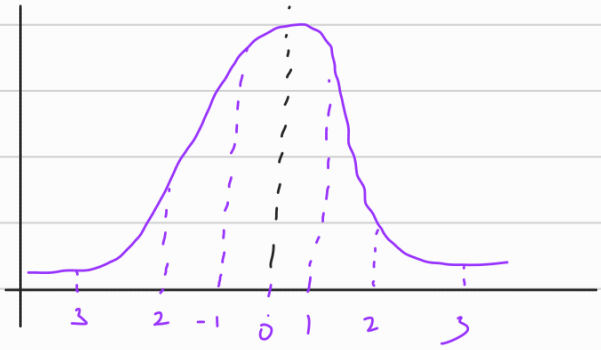
x	$x-\mu$
1	-4.5
2	-3.5
3	-2.5
4	-1.5
5	.5
6	.5
7	1.5
8	2.5
9	3.5
10	4.5
$\mu = 5.5$	$\mu = 0$

to convert σ to 1 divide $x-\mu/\sigma$

$$z = \frac{65-70}{4}$$

- z-table give area on left hand side of curve



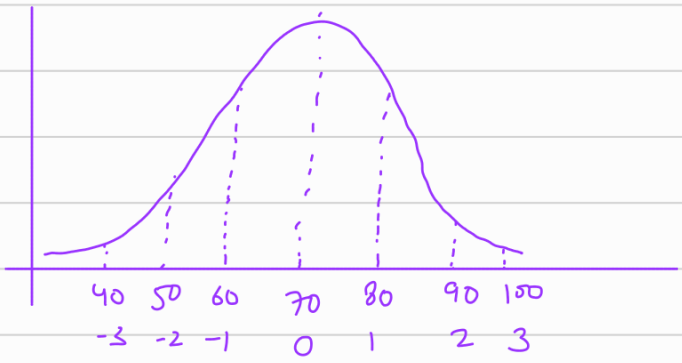


- This will give are upto this portion
- We need area on right ∴ 1 - Area of left.

Q. The marks of students of CSE follow gaussian distribution with mean 70 marks & variance of 100 marks find out the following prob.
$P(X \geq 80)$, $P(60 \leq X \leq 80)$, $P(X \leq 50)$

- $\mu = 70$, $\sigma = 10$

(i)



$P(X \geq 80) = 1 - (z \text{ at } 1)$

$P(60 \leq X \leq 80) = (z \text{ at } 1) - (z \text{ at } -1)$

$P(x \leq 50) = (z \text{ at } -2)$

Binomial distribution →

$P(X = r) = {}^{n}C_{r}\, p^{r} q^{n-r}$

$P(X = 2) = {}^{n}C_{4} (1/2)^{2} (\tfrac{1}{2})^{2}$

$n = 4$

coin toss

Head 1, 2, 3, 4 first

Distribution to learn
- Geometric
- Normal or gaussian
- Poisson
- Binomial

Q. There are 100 questions with 4 possible ans A, B, C, D in a test all the options are equally likely to be correct, each question carry 1 mark, find out the probability that the no. of ans successfully marks < 70.

→ Binomial is over calculative ∴ we will transform it to other distribution

Expectation & varience of random variable:-

Ex- lottery system → 3-window & D1 symbol → $, C, L, other

0.1 0.2 0.2 0.5

same for D2 & D2, say if we get $ $ $ we win 20 Rs, $ $ C = 15 Rs (in any seq.), CCC = 10 Rs, LLL = 5 Rs

we need to find expected amount we will win, entry fee: 1 Rs

we need to make prob dist. table

X	Prob. $(X=x)$	
19	$0.1 * 0.1 * 0.1$	$= 0.001$
14	$0.1 * 0.1 * 0.2 * 3$	$= 0.006$
9	$0.2 * 0.2 * 0.2$	$= 0.008$
4	$0.2 * 0.2 * 0.2$	$= 0.008$
-1	$1 - AU$	$= 0.977$

$$\text{Expectation } [x] = \sum_x X P(X=x)$$

$$= 19 * 0.001 + 14 * 0.006 + 9 * 0.008 + 4 * 0.008 - 1 \times 0.977$$

$$E[x] = -0.77$$

$$\text{Varience } [x] = E[(x-\mu)^2]$$

$$\because \quad E[f(x)] = \sum_x f(x) \cdot P(X=x)$$

$$= \sum_x (x-\mu)^2 P(X=x)$$

$$= \sum_x (x^2 + \mu^2 - 2\mu x) P(X=x)$$

$$= \sum_x x^2 P(X=x) + \mu^2 \sum_x P(X=x) - 2\mu \sum x P(X=x)$$

(μ^2 and $2\mu^2$)

$$= E(X^2) - E^2(X)$$

X	Prob. $(X=x)$		$X-\mu$	$(x-\mu)^2$
19	$0.1 * 0.1 * 0.1$	$= 0.001$	19.77	390.85
14	$0.1 * 0.1 * 0.2 * 3$	$= 0.006$	14.77	218.15

9	$0.2 * 0.2 * 0.2 = 0.008$	9.77	95.45
4	$0.2 * 0.2 * 0.2 = 0.008$	4.77	22.75
-1	$1 - AU = 0.977$	-0.23	0.05

$$var = 0.001 * 390.85 + 0.006 * 218.15 + 0.008 * 95.45 + 0.008 * 22.75 + 0.977 * 0.05$$

$$= 2.69$$

$$\sigma = \sqrt{2.69} = 1.64$$

- In general we lose 0.77 vs by gussian dist.

$\therefore$ Our win range lie in $[-0.77 - 3 * 1.64, -0.77 + 3 * 1.64]$

- Price money for the lottery is made 5 times & fee is doubled.

Y	P(x)
98	0.001
73	0.006
48	0.008
23	0.008
-2	0.977

Original win - fee = X

original win = X + fee = (X+1)

$Y = 5(X+1) - 2$

$Y = 5X + 3$

↓

Linear dependency

- If $Y = aX + b$; $E(x)$ is known

then $E[y] = aE[x] + b$

$Var[y] = a^2 V[x]$

Derive → $E(x) = \sum_x f(x) \cdot P(X = x)$

$E(aX + b) = \sum_x (ax + b) \cdot P(X = x)$

$$= \sum_n a \cdot n \cdot P(X=n) + \sum_n b\, P(X=n)$$

$$= a \sum_n n\, P(X=n) + b$$

$$= a\, E[n] + b$$

$$\text{Var}(aX+b) = E[(an+b)^2] - (E[an+b])^2$$

$$= E[a^2n^2 + b^2 + 2abn] - (aE[n]+b)^2$$

$$= a^2E[X^2] + \cancel{2ab\,E[x] + b^2} - a^2E^2[x] - \cancel{2ab\,E[x] - b^2}$$

$$= a^2\left(E[x^2] - E^2[n]\right)$$

$$= a^2\, \text{Var}[x]$$

Note: $2X$ is diff from $X+X \Rightarrow 2X \Rightarrow$ price money doubled

$X+X$ is 2 lottery ticket purchase

$2X \rightarrow$ linear dependency

X	15	-2
$P(X=n)$	0.4	0.6
$Y = 2X$	30	-4
$P(Y=y)$	0.4	0.6

$X+X \rightarrow$ independent observation

X	$P(n)$
30	0.16
13	0.48
-4	0.36

both differ

- $E(X+X) = 2E(X)$ $\quad Var[X+X] = 2\,Var[X]$
- $E(2X) = 2E(X)$ $\quad Var[2X] = 4\,Var[X]$

Derive:

$E(aX+bY) = aE[X] + bE[Y]$

$Var[aX+bY] = a^2 Var[X] + b^2 Var[Y]$

$E[aX-bY] = aE[X] - bE[Y]$

$Var[aX-bY] = a^2 Var[X] + b^2 Var[Y]$

} For independent var

- Chap. 5 Head first statistics

Expectation & Varience of geometric distribution:

P = Prob. win r = trails r = n to success ?! q = lose prob

$E(X) = \sum X P(X=n)$

$= \sum_{r=1}^{\infty} r\,P(X=r)$

$E(x) = 1.P + 2.qP + 3q^2P + 4q^3P + \cdots$

$qE(x) = 1.q.P + 2q^2P + 3q^3P + \cdots$

$(1-q)E(x) = P + Pq + Pq^2 + \cdots Pq^5 + \cdots$

$PE(x) = P[1 + q + q^2 + \cdots q^{\infty}]$

$$\boxed{E(x) = \frac{1}{1-q} = \frac{1}{P}}$$

$Var[x] = E(x^2) - (E[x])^2$

$= E[x^2] - \frac{1}{P^2}$

$= E[x(x-1)] + E[x] - (E[x])^2$

$E[x(x-1)] = \sum_{k=1}^{\infty} k(k-1)P(X=k)$

$= \sum_{k=1}^{\infty} k(k-1)\,P.q^{k-1}$

$= P\sum_{k=1}^{\infty}(k-1)k\,q^{k-1} = P.\frac{d}{dq}\left(\sum_{k=1}^{\infty}(k-1)q^k\right)$

$= P\frac{d}{dq}\left(q^2\sum_{k=1}^{\infty}(k-1)q^{k-2}\right) = P\frac{d}{dq}\left(q^2\sum_{k=2}^{\infty}(k-1)q^{k-2}\right)$

$= P\frac{d}{dq}\left(q^2\frac{d}{dq}\left(\sum_{k=2}^{\infty}q^{k-1}\right)\right) = P\frac{d}{dq}\left(q^2\frac{d}{dq}\left(\sum_{k=1}^{\infty}q^k\right)\right)$

$= P\frac{d}{dq}\left(q^2\frac{d}{dq}\left(\frac{1}{1-q}-1\right)\right) = P\frac{d}{dq}\left(q^2\left(\frac{-1}{(1-q)^2}\right)\right)$

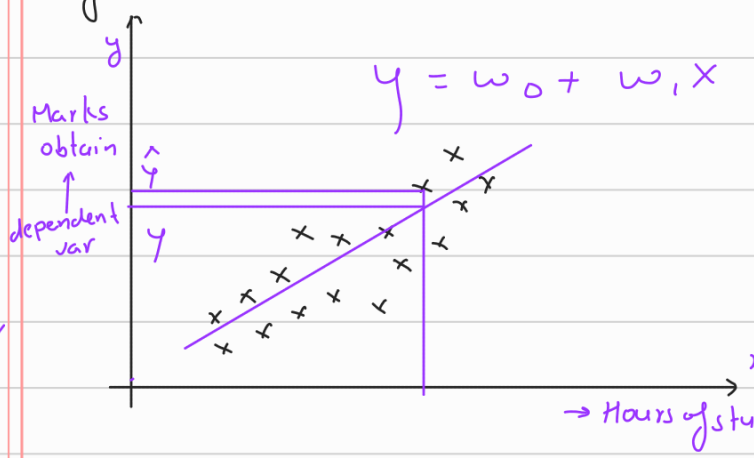
$= P\frac{d}{dq}\left(\frac{-q^2}{(1-q)^2}\right) = P.\frac{2q}{(q-1)^3} = P.\frac{2q}{P^3} = \frac{2q}{P^2}$

$\therefore Var[x] = \frac{2(1-P)}{P^2} + \frac{1}{P} - \frac{1}{P^2}$

$= 1-P/P^2$

$= q/P^2$

Regression →



$y = w_0 + w_1 x$

Machine learning
- Supervised → Regression
- Unsupervised

$Y = w_1 X + w_0$ (Assume linear relation)

let Assume sleep also affect then

$Y = w_1 X_1 + w_2 X_2 + w_0$ (Multiple linear relation)

polynomial reg → let 1 ind. var but 2nd order relationship

$Y = w_2 x^2 + w_1 x + w_0$ (reg with 2nd order poly)

let 2 ind. var →

$Y = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_1^2 + w_4 X_2^2 + w_5 X_1 X_2$

$\hat{Y}$	X	let $y = x$	Residual error	sq	$E = \frac{1}{2}\Sigma$ sq	let $y = 2x$	RE	sq
2	1	1	1	sq →	1	2	0	0
4	2	2	2	sq →	4	4	0	0
6	3	3	3	sq →	9	6	0	0
8	4	4	4	sq →	16	8	0	0

$E = 15$ → this need to min.

$\therefore E = 0$ → min

- In regression we minimize least squares errors

$$E = \frac{1}{2}\sum_{i=1}^{N} r_i^2 = \frac{1}{2}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

$$= \frac{1}{2}\sum_{i=1}^{N}(\hat{y}_i - w_0 - w_0 x_i)^2$$

↓ minimize

$$\frac{\partial E}{\partial w_0} = \frac{1}{2} \times 2 \sum_{i=1}^{N}(\hat{y}_i - w_0 - w_1 x_i)(-1) \Rightarrow 0$$

$$= \sum_{i=1}^{N} \hat{y}_i - w_0 N - w_1 \sum_{i=1}^{N} x_i = 0$$

$$= w_0 N + w_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \hat{y}_i \quad - (1)$$

$$\frac{\partial E}{\partial w_1} = \frac{1}{2} \times 2 \sum_{i=1}^{N}(\hat{y}_i - w_0 - w_1 x_i)(-x_i) \Rightarrow 0$$

$$= \sum_{i=1}^{N} \hat{y}_i x_i - w_0 \Sigma x_i - w_1 \sum_{i=1}^{N} x_i^2 = 0 \quad - (2)$$

Use 1, 2 & data, substitute it ⇒

y	x	x^2
2	1	1
4	2	4
6	3	9
8	4	16
20	10	30

(1) ⇒ $w_0(4) + w_1(10) = 20$

(2) ⇒ $w_0(10) + w_1(30) = 60$

$w_0 = 0 \quad w_1 = 2 \Rightarrow$ solution optimal

Ex-

$\hat{y}$	x
4	1
8	2
13	3
32	4

Fit in the following poly.

(i) $y = w_0 + w_1 x$

(ii) $y = w_0 + w_1 x + w_2 x^2$

(i) ① $\Rightarrow$ $w_0 4 + w_1 10 = 57$

② $\Rightarrow$ $187 = w_0 10 + w_1 30$

$19 w_0 + 30 w_1 = 171$

$10 w_0 + 30 w_1 = 187$

$2 w_0 = -16$

$w_0 = -8 \Rightarrow w_1 = 57 - 4(-8)$

$= 57 + 32 = 89$

(ii) Solve diff.

Order = curve + 1 (Change) → n U

Q

$\hat{y}$	x	$\hat{y} - y$
2	1	$2 - w$
5	2	$5 - 2w$
5	3	$5 - 3w$
7	4	$7 - 4w$
11	5	$11 - 5w$

$y = wx$ → Given

$$\frac{1}{2}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 = \frac{1}{2}\sum_{i=1}^{N}(\hat{y}_i - wx)^2$$

$$\frac{dE}{dw} = \frac{1}{2} \times \sum_{i=1}^{N} 2(\hat{y}_i - w x_i)(-x_i) = 0$$

$$= \sum_{i=1}^{N}(\hat{y}_i - w x_i) x_i = 0$$

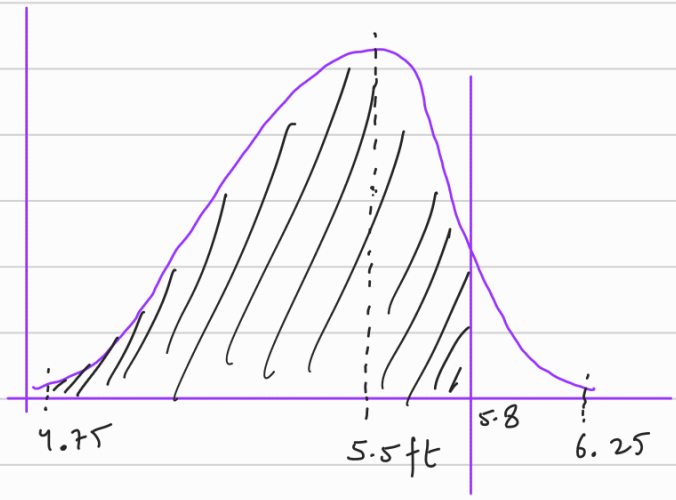
$$= \sum_{i=1}^{N} y_i x_i - w \sum_{i=1}^{N} x_i^2 = 0$$

$= 110 - 55w = 0$

$\Rightarrow w = 2$

$\therefore y = 2x$

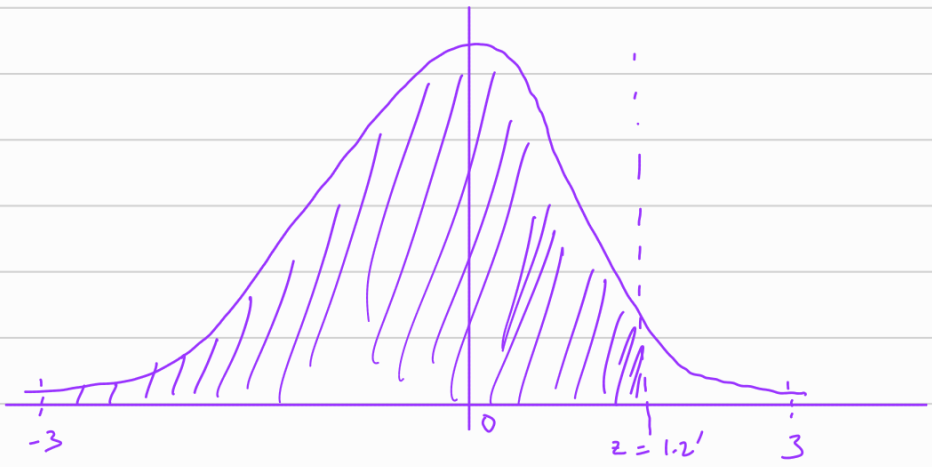
Q. The height of students of class follow normal distribution with avg height of 5.5ft and varience of 1/16 ft find out prob that student will have height less then 5.8 ft?



$\sigma^2 = 1/16$

$\sigma = 1/4$

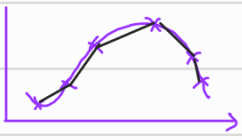
Standard. $z = \frac{x-\mu}{\sigma}$

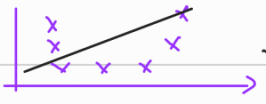


$z = \frac{5.8' - 5.5'}{0.25} = 1.2'$

Ans = z-table (1.2') =

SEE IN PYTHON

Curve fitting → Overfitting ⇒  taken deg. > req. deg.

Curve fitting → Underfitting.  req. poly degree > taken poly

- To resolve both we use regulaization

Regression with regulation →

- Here we modify error function to incorporate regularization term :-

We need to minimise ⇒ $\frac{1}{2}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + C \cdot \frac{1}{2} w^T w$

$C \in [2^{-20}, 2^{-18} \ldots\ldots, 2^{50}]$ ← Regularization parameter

$[w_1\ w_2]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ ← regulaization

$= \frac{1}{2}(w_1^2 + w_2^2)$

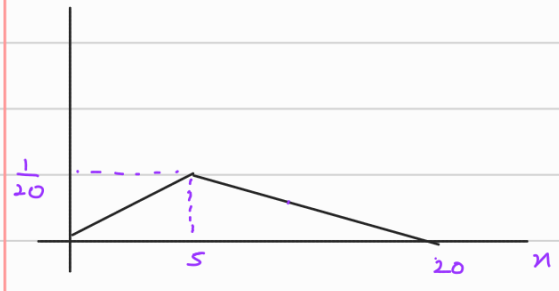
- In regulaization we need to tune C & d (chk all poss. for best result)

⇒ KEEL DATASET Repo. → regression prob.

- Find 3-4 reg. prob. & use suitable tool box & solve these problems find optimal value of degree of polynomial & regularization polynomial

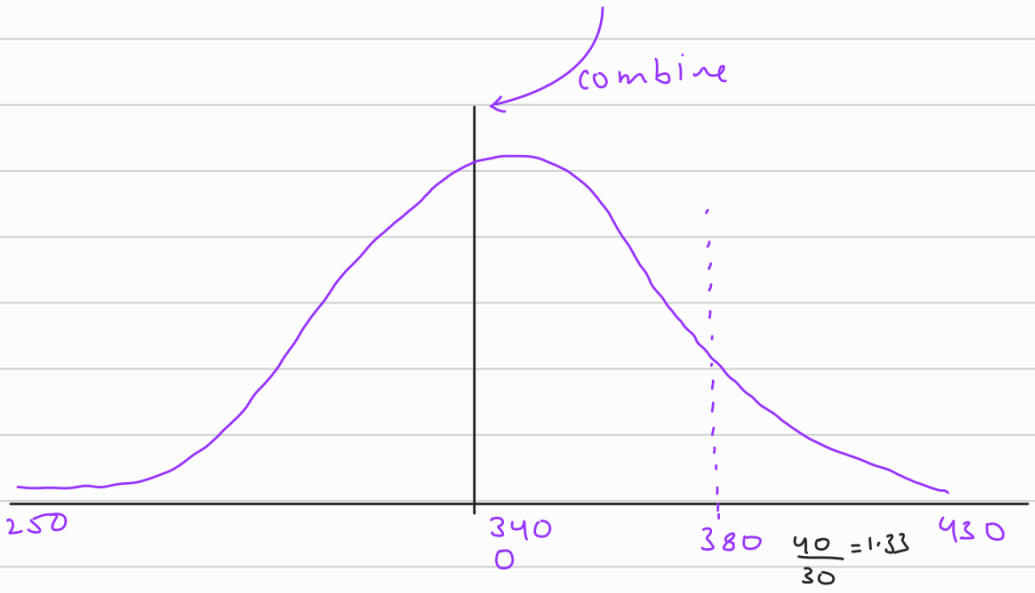
Chap-9 Beyond normal distr.→

Q.



Find prob $x > 10$

Q. Roller coster ride where X rep. weight of bride $X \sim N(150, 400)$ (μ, σ^2) & weight of groom is given by $Y \sim N(190, 500)$, they can ride if their combined weight is <=380



- When we sum up 2 distribution follow gaussian distribution it will also follow gaussian dist.

$E(x+y) = E(x) + E(y)$ $\quad x, y$ = independent

$Var[x,y] = Var[x] + E[y]$

$\therefore (x,y) \sim N(340, 900)$

$$Z = \frac{380-340}{30} = 1.33$$

$\therefore$ Area = Prob = 0.908

Q. In town male height ~ $N(71, 20.25)$, & for women $\sim N(64, 16)$ what is prob that man will be at least 5 inches taller than women

$X : N(71, 20.25)$ $\quad X - Y \geq 5$

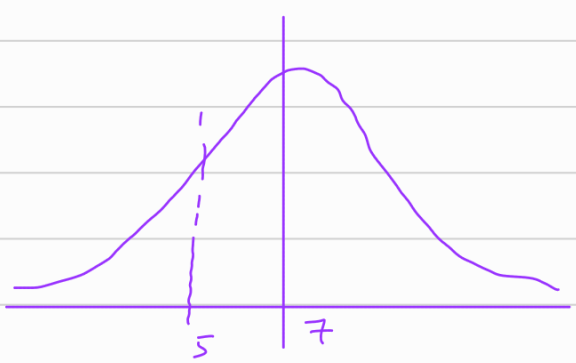
$Y : N(64, 16)$ $\quad x, y \in$ independent

$E(X-Y) = E(X) - E(Y) = E_N$

$V[X-Y] = V[X] + V[Y] = V_N$

$E_N = 7$, $V_N = 36.25$

$\sigma_N = 6.02$



$$P(<5) = z\left(\frac{5-7}{6.02}\right) = z\left(\frac{-2}{6.02}\right) = z(-0.33) = 0.37070$$

$$\therefore P(>=5) = 1 - 0.37070$$
$$\simeq 0.63$$

Till chap 9 Head first stat.

Binomial distribution →

- 'n' - Toss head success
- $P(X=r) = {}^nC_r\, p^r q^{n-r}$

Q. Ques. paper 50 Q → 4 possible ans (A, B, C, D) one of them is correct prob ≤30 correct?

$$p = \frac{1}{4}, \quad q = \frac{3}{4}, \quad n = 50$$

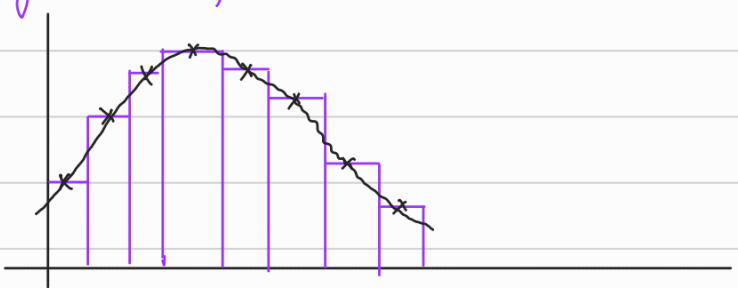
$$P(X \leq 30) = \sum_{n=0}^{30} {}^nC_r\, p^r . q^{n-r}$$

→ Calculative

∴ If $np \leq 10$ we plot a histogram for 'i' we plot bar from i-0.5 to i+0.5

→ but if $np > 10$ then this histogram tends to follow gaussian distribution.

- Then we normalize freq. of histogram by dividing freq/Total freq.
- Then we apply regression to found the curve



- let by req we get Pdf if $\int_{-\infty}^{\infty} pdf \cdot dx = p$ divide it by p $\therefore pdf = \frac{1}{p}(pdf)$

Else we can use parzen window estimator.

- let 3 student height 5', 5.5', 5.6' & i want to find overall pdf
- we will first find standard deviation 'σ'
- we will take diff function

$$Pdf_{Parzen} = \frac{1}{3}(GF('5', 'σ') + GF(5.5', σ) + GF(5.6', σ))$$

- The shape of window can be anything

$E[x] = np$

$Var[X] = npq$

X	0	1
P(x)	q	p

$$E[x]_{\text{single trail}} = \sum x P(x=n) = 0 \cdot q + 1 \cdot p = p$$

$E[x_1 + x_2 \cdots + x_n] = np$ ∵ trails are independent

$$E[x] = np$$

$Var[x_1 \cdots + x_n] = n\, Var[x]$

$$Var_{\text{single trail}}[x] = E[x^2] - (E[x])^2$$

$$= \sum x^2 P(x=n) - p^2$$

$$= 0 \cdot q + 1 \cdot p - p^2$$

$$= p - p^2 = p(1-p) = pq$$

$\therefore$ Var [x] = pq
for 1 trail

So. $\text{var}_n[x] = npq$

Q. Their are 20 Question the prob of getting 4 or less question correct (single MCQ ans 2 option)

$$P(X=4) = {}^{20}C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{16}$$

$$P(X=3) = {}^{20}C_3 \left(\frac{1}{2}\right)^{20}$$

$$P(X=2) = {}^{20}C_2 \left(\frac{1}{2}\right)^{20}$$

$$P(X=1) = {}^{20}C_1 \left(\frac{1}{2}\right)^{20}$$

$$P(X=0) = {}^{20}C_0 \left(\frac{1}{2}\right)^{20}$$

$$\Sigma = \frac{1}{2^{20}} \left[{}^{20}C_0 + {}^{20}C_1 + {}^{20}C_2 + {}^{20}C_3 + {}^{20}C_4 \right]$$

$$= \frac{1}{2^{20}} \left\{ 1 + 20 + 190 + 1140 + 4845 \right\}$$

$$= \frac{1}{2^{20}} \cdot (6196) = 0.0059$$

- Expectation is mean.

Transform

$\text{Exp}[x] = np = 20 \cdot \frac{1}{2} = 10 \geq 10$ $\therefore$ Gaussian distribution

$\mu = 10$, $\text{var} = npq = 5 \Rightarrow SD = \sqrt{5}$

$X = 4$ transform

$$Z = \frac{4-10}{\sqrt{5}} = -2.68$$

Area $P(X \leq 4) = 0.0037$ Wrong

we need to apply distribution correction

for i histogram goes till i+0.5

$\therefore$ $X = 4.5$ transform

$$Z = \frac{4.5-10}{\sqrt{5}} = \frac{-5.5}{\sqrt{5}} = -1.1 * \sqrt{5} = -2.46$$

Area = Area(Z) = 0.0069 $\simeq$ 0.0059 $\therefore$ Right