Page 1

lec 1

Week - 1.

Types of data.

Data.
- Categorical (define categories or groups)
- Numerical
  - Discrete → counted items
  - Continuous. → time, pressure

→ Benefit of visual representation of data:

→ Create picture of data: makes abstract data more concrete and understandable.

→ Make cognitive processing easier

→ content view at glance: giving users comprehensive understanding of data

→ direct attention toward content:

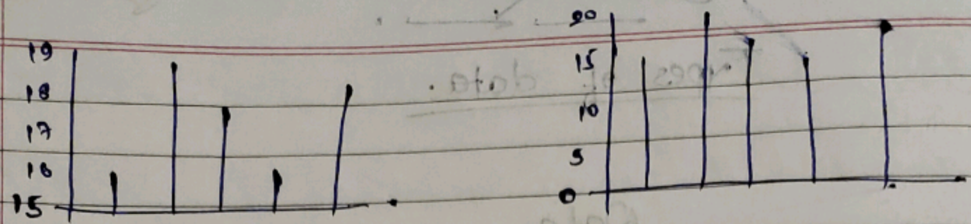
→ describe, explore, and summarize

FOUR UMBRELLA PRINCIPLE of effective visualizat^n.

1) KNOW PURPOSE :- having purpose statement for every table or graph you creat

• purpose is not necessary a message

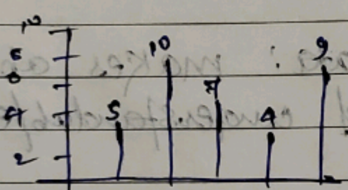
2) ENSURE INTEGRITY :- accuracy & consistency



same data but different representat^n says different informat^n.

3) MAXIMIZE DATA INK ; MINIMIZE NON-DATA INK :- use less amount of ink to convey most amount of info.

4) SHOW YOUR DATA : ANNOTED :- don't hide data, show it, use annotat^n.



EXECUTING YOUR INFO display : a 3-step PROCESS :

1) DEFINING MESSAGE :-
- what I'm trying to communicate
- What's message
- How do i make msg clear at glance

2) CHOOSING FORM :-
- what should I use : text, table graph chart, diagram

3) CREATING DESIGN :- for good design people pay for it, presentat^n matters.

1) DEFINING MESSAGE:

→ what you want your audience to understand and remember. It could be data trends, relationship b/w various variables or summary.

2) CHOOSING FORM – TABLE or CHART ??

1) Choose a Table :-

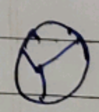
- → display complete dataset
- → present wide range of dataset that is difficult to scale using graph.
- → focus on specific item
- → explain how numbers are derived – how results were calculated.

2). CHOOSE a chart :-

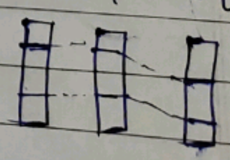
- → compare a slice of information
- → show changes over time
- → show "causes of changes"
- → show patterns of data distribution

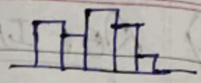
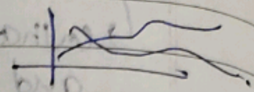
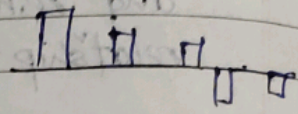
CHOOSE Appropriate Graph:

- components of one item → Pie chart



- Components of multiple items → 100% column / stacked column chart



- Item comporison → BAR CHART 
- Change over time → column/line chart 
- Frequency, distribut? → Histogram 
- correlation → Paired bar, scatter dot 

3) CREATING DESIGN :

- avoid 3-d effect
- avoid legends; consider using labels
- avoid constrasting borders around objects
- use annotations to highlight keys

VISUALIZATION ON DASHBOARD:

Dashboard : A visual display of most imp. info

Should on Dashboard	Should not.
→ Real time data.	→ Complex visual
→ visual summary	→ excessive detail ↳ use hyper links.
→ filter/drill down optn	
→ comparative data :	→ Redundant matrics

PA-1.

1) to show complete data - use tabular form

2) to show pattern - use graph rather table

6) One item proportional to total → pie chart

Multiple item components → multiple item comparis^n

comparison b/w items → Bar chart.

Trend over time → line chart

Histogram frequency of items → ~~Bar chart compare b/w items~~

outliers identificat^n → Box plot

correlat^n representat^n → scatter plot

Bar chart → compare b/w items.

13)

Week-2

→ The data values themselves are used directly in simulation. This is called "trace-driven simulation".

→ "FIT" a theoritical distribution to the data (and check whether that "fit" is good!)

→ using data, we build our own distributions.
↳ this is Empirical distribution.

→ for SYMMETRIC DISTRIBUTION : mean, median, mode should be equal.
→ if these values are sufficiently closer to each other we can think of symmetric distribution.

→ Cofficient of variation (CV) : ~~ratio~~ $\frac{\text{std. dev}}{\text{mean}} = \frac{\sigma}{\mu}$

→ the CV = 1 for exponential distribut^n

Low CV : less variability in relat^n to mean.
→ means data points are close to av. value.

High CV : greater variability related to mean.
→ means data points are more spread out.

→ Lexis ratio (L) : same as CV for discrete distribut^n.

$$L = \frac{\text{observed variance}}{\text{expected variance}}$$

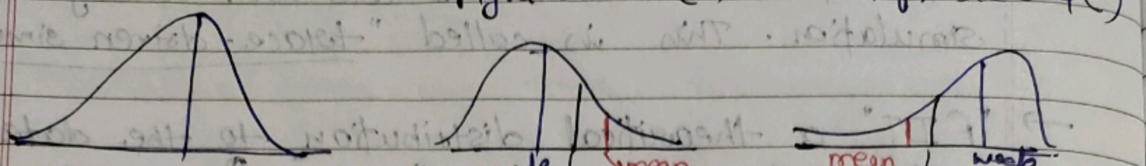
L < 1 : data is more uniform than expected

$$\frac{n}{(n-1)(n-2)} \sum \frac{(x_i - \bar{x})^3}{s^3}$$

→ Skewness (v) : measure of symmetry of distribution.

for normal dist. v = 0
for v > 0. right skewed (exponent^n dist. v = 2)
for v < 0 left skewed

right skewed (+) | left skewed (−)



skewness = 0
mean = median = mode

mean > median > mode

mean < median < mode

lec 3.

Goodness-of-fit.

↳ can be checked by several methods:

i) frequency composition (a bit technical)
ii) probability plots (visual tool)
iii) Goodness-of-fit tests. (statistical test of goodness)

Probability plots:

1) Q-Q plot: Quantile-quantile plot

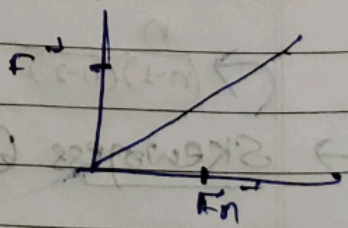
→ Graph of q_i – quantile of fitted (model) distribut^n versus q_i – quantile of sample distribut^n.

$$x^{M}_{q_i} = \hat{F}^{-1}(q_i)$$

$$x^{S}_{q_i} = \tilde{F}_n^{-1}(q_i) = x_i$$

If this and this is equal then Q-Q plot will be approximately linear with intercept 0 & slope 1

→ valid for descrete dataset only.



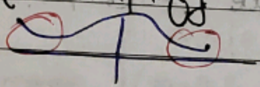
2) P-P plot : prob-prob. plot.

→ A graph of model probability $\hat{F}(x_{(i)})$ against the sample probability $\tilde{F}_n(x_i) = q$

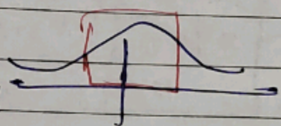
→ valid for both continuous & discrete dataset.

→ if $\hat{F}(x)$ is correct distribution that is fitted, for a large sample size, then $\hat{F}(x)$ & $F_n(x)$ will be close together and P-P plot will be approx. linear with intercept 0 & slope 1.

Q-Q plot amplify difference b/w tails of model distribut^n.



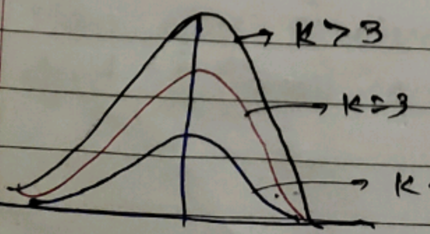
P-P plot amplify difference at middle portion



KURTOSIS (K): describes the shape of distribution's tail

→ overall focusing on presence of outliers and "tailedness" of distribution.

- K = 3 (Mesokurtic) : Normal distribut^n like behaviour with moderate tails.
- K > 3 (Leptokurtic) : More data in the tails, indicating higher likelihood of extreme outliers. (.fatter tail .sharper peak)
- K < 3 (platykurtic) : less data in the tails, indicating fewer outliers & flatter peak. (.thin tail .flat peak)



Week-2 PA.

1) $CV = \frac{\sigma}{\mu} = \frac{3.2}{10.7} = 0.299$

5) Binomial → mean = np var = npq
Geometric → Mean = $\frac{1}{p}$ var = $\frac{q}{p^2}$

poisson → mean = var = λ

7) Calculated value (19.70) is greater than tabulated value (8.25). This means calculated chi-square statistic exceeds critical value.

→ in hypothesis test, if calculated value exceeds the tabulated (critical value) we reject null hypothesis (H_0)

→ since calculated value > critical
p-value < 0.05

Week - 3

Total 1200 candidate (960 male, 240 female)
out of 1200, 324 were given offer letter.

~~offers made~~	Male	female	Total
~~Not~~ offered	288	36	324
~~Total~~ Not offered	672	204	876
Total	960	240	1200

$P(M \cap A) = 288/1200 = 0.24$

$P(M \cap A^c) = 672/1200 = 0.56$

$P(F \cap A) = 36/1200 = 0.03$

$P(F \cap A^c) = 204/1200 = 0.17$

M: male
F: female
A: admission offered
A^c: not offered.
$P(A) + P(A^c) = 1$

$C = \frac{J}{M}$

		Male	Female	Total
A	offer made	0.24	0.03	0.27
A^c	Not offered	0.56	0.17	0.73
	Total	0.8	0.2	1.0

Marginal

Joint probability.

conditional prob:

$P(A/M) = ? = \frac{P(A \cap M)}{P(M)} = \frac{0.24}{0.8}$

$= 288/960$

$$\boxed{C = \frac{J}{M}}$$

C = conditional prob
J = joint "
M = marginal "

Q. 40% of emails are spam
the word "lottery" appears 30% of spam ~~email~~ email
the word " " 5% of non-spam email
Now you received a mail contain word "lottery"
find P(mail is spam)?

A = spam mail   B = mail contain "lottery"

A' = not spam   B' = ... not contain ..

$P(A/B) = ?$

$$\boxed{P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}}$$

~~P(A) = 0.4~~
~~P(A') = 0.6~~
~~P(B) = 0.3~~
~~P(B') = 0.7~~

$\cancel{P(B/A) = \frac{P(A \cap B)}{P(A)}}$

$P(A) = 0.4$

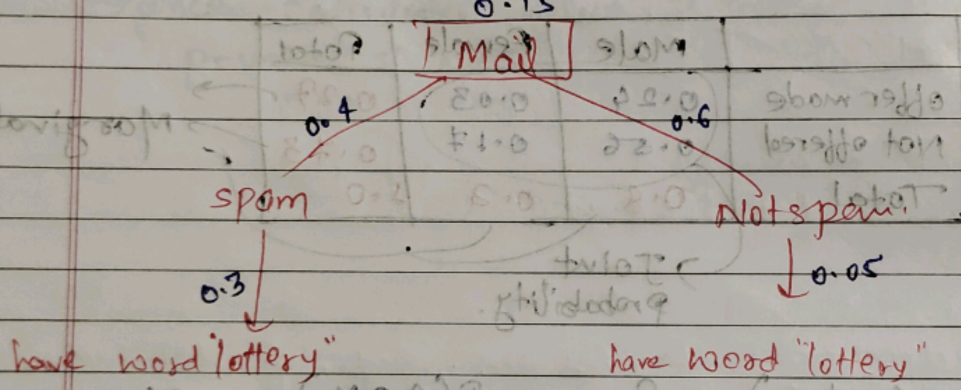
$P(B/A) = 0.3$

$P(B/A') = 0.05$

$P(A') = 0.6$

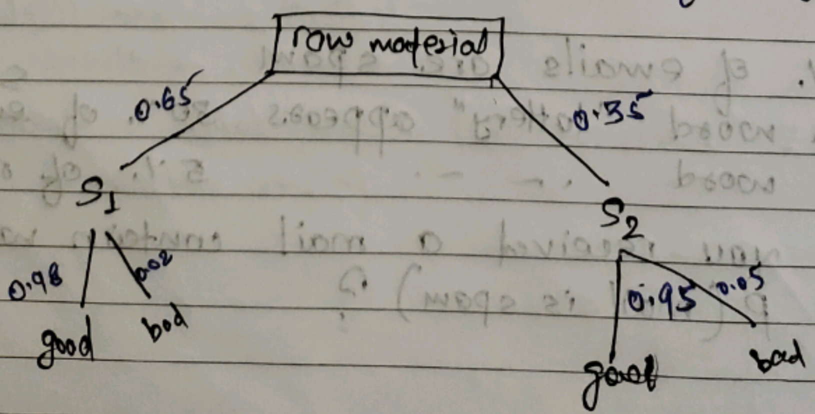
$P(B) = 0.3 \times 0.4 + 0.05 \times 0.6 = 0.15$

$$P(A/B) = \frac{0.3 \times 0.4}{0.15} = 0.8$$



Baye's rule

→ Baye's rule is used to calculate posterior probability if we have initial belief (probablity) and the additional sample information.



find P(good quality comes from S_1)

$$= \frac{0.65 \times 0.98}{0.65 \times 0.98 + 0.35 \times 0.95} = \frac{0.637}{0.9695} = 0.65$$

find P(bad quality comes from S_1)

$$= \frac{0.65 \times 0.02}{0.65 \times 0.02 + 0.35 \times 0.05} = \frac{130}{305} = 0.42$$

lec 3.

Chi-square distribution :

- Null hypothesis :-

  H_0 : the categorical variables are independent.

- Alternate hypothesis :-

  H_1 :- the categorical variables are not independent.

let f_o be the observed frequency (from sample)
let f_e be expected frequency

→ The expected frequency (f_e) for a cell equals the product of row & column total for that cell, divided by total sample size.

City	Brand A	Brand B	Brand C	Total
Mumbai	279 (261.4)	73 (90.7)	225 (244.9)	577
Chennai	165 (182.6)	47 (49.3)	191 (171.1)	403
Total	444	120	416	980

— observed frequency
— expected frequency

$261.4 = \frac{577 \times 444}{980}$

$182.6 = \frac{403 \times 444}{980}$

$49.3 = \frac{403 \times 120}{980}$

- Chi-squared test statistic:

$$\boxed{\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}}$$

$$\chi^2 = \frac{(279-261.4)^2}{261.4} + \frac{(73-70.7)^2}{70.7} + \cdots \text{ for total 6 time}$$

$$\chi^2_{cal} = 7.0$$ (this is Calculated test stastics value)

→ degree of freedom (df) = (row−1)(column−1)
$\boxed{K-p-1}$ = (2−1)(3−1) = 2

(bins) (parameters needs to calculate)

- df = 2, so at $\alpha = 0.05$ (95% confidence), tabular value of test statistic. = 5.99

$$\because \chi^2_{cal} = 7.0 > \chi^2_{test} = 5.99.$$

reject null hypothesis (H_0)

- df = 2, so at $\alpha = 0.01$ (99% confidence), tabular value of test statistic = 9.21

$$\because \chi^2_{cal} = 7.0 < \chi^2_{test} (9.21)$$

$\therefore$ ~~Can't~~ acept null hypothesis (H_0)

Page 4 | Date | Page 5

: for total 6 time

tics value

value

value of

PA-3

1) at $\alpha = 0.01$.

(I) → if p-value is greater than $\alpha = 0.01$: then accept Ho
→ if p-value is less than $\alpha = 0.01$: then reject Ho

(II) → if $\chi^2_{cal} > \chi^2_{test}$ reject Ho

→ if $\chi^2_{cal} < \chi^2_{test}$ accept Ho

(2)

	sec-A	sec-B	Total
Sci	288 (269.7)	60 (78.25)	348
Math	622 (640.25)	204 (185.7)	826
Total	910	264	1174

— observed freq.
— expected freq.

$269.7 = \frac{348 \times 910}{1174}$

$\frac{\text{total sci student}}{\text{total stud.}} = \frac{348}{1174} = 0.296$

(3) $P(\text{Math} / \text{sec-B}) = \frac{P(\text{math} \cap \text{sec-B})}{P(\text{sec-B})} = \frac{204}{264} = 0.77$

(4) $\chi^2 = \frac{(288-269.7)^2}{269.7} + \frac{(60-78.25)^2}{78.25} + \frac{(622-640.25)^2}{640.25} + \frac{(204-185.7)^2}{185.7}$

=

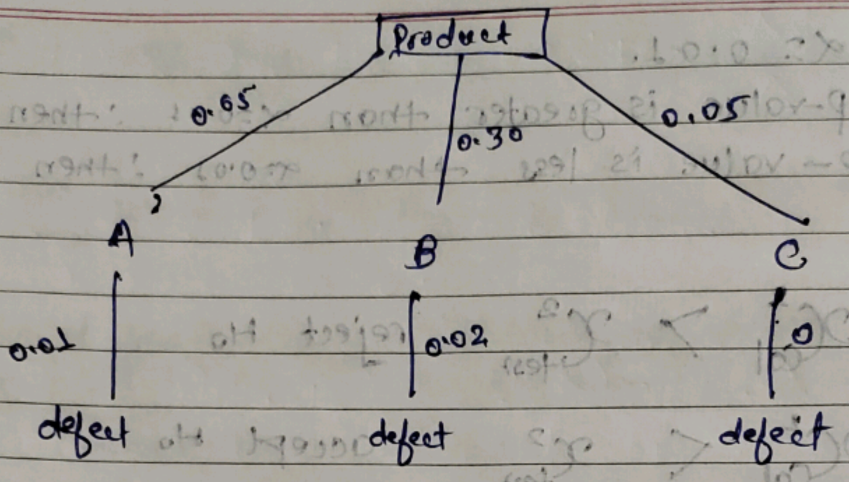
using excel p-value = 0.005
=(CHISQ.Test).

∴ p-value < α
so reject Ho

(5) ~~p-value is~~ $\alpha = 0.01$
still pvalue (0.05) < α
reject Ho

(6) $\alpha = 0.001$
p.value (0.05) > α
accept Ho

(7)

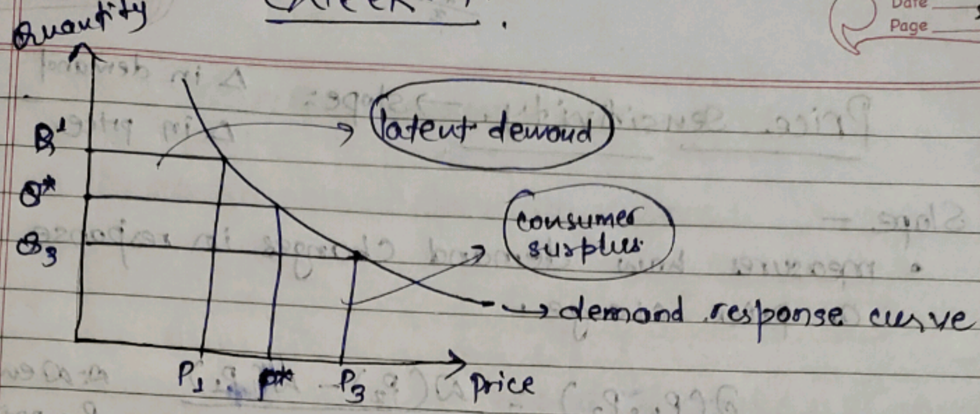


P(product come from B | it was defective) = ?

$$= \frac{0.30 \times 0.02}{0.65 \times 0.01 + 0.30 \times 0.02 + 0.05 \times 0}$$

$$= \frac{60}{65 + 60} = \frac{60}{125} = 0.48$$

Week-4.



Demand Response Curve.

→ a function that describes how demand for a product d(P) varies as a function of its price.

→ similar to the demand curve in economics, but for a single seller, in a single market.

Four properties:

1. Non-negative (both quantity & price)
2. continuous
3. Differentiable
4. Downward sloping

Latent Demand: a situation where there is a potential desire for a product, but market doesnot currently fulfill that demand. It might be because of product isn't available, price is high or consumer aren't yet aware of product.

Consumer Surplus: difference b/w what consumers are willing to pay for good and what they actually pay.

lec 3

Price sensitivity. → slope = $\frac{\Delta \text{ in demand}}{\Delta \text{ in price}}$

Slope –

- measures how demand changes in response to a price change.

$$\partial(P_1, P_2) = \frac{D(P_2) - D(P_1)}{P_2 - P_1}$$

D: Demand
P = price

① ELASTICITY: ratio of % change in demand to the percentage change in price.

$$\text{Elasticity } (E) = \frac{\text{\% change in quantity}}{\text{\% change in price}}$$

→ if $E > 1$, demand is elastic, means consumers are highly responsive to price change.
price increases demand decreases

→ $E < 1$, demand is Inelastic, mean even large price increase don't significantly reduce demand

→ $E = 0$, demand is perfect inelastic, demand doesn't change at all in response to price changes.

→ $E = \infty$, demand is perfect elastic, mean consumer will only buy at one price

→ Elasticity of 2 means 10% reduction in price will yield a 20% increase in sale.

→ $E = 2$:→ 1% increase in price leads to 2% decrease in demand.

→ $E = 0.9$ → 1% increase in price leads to 0.9% decrease in demand lec 3

LINEAR Response CURVE :

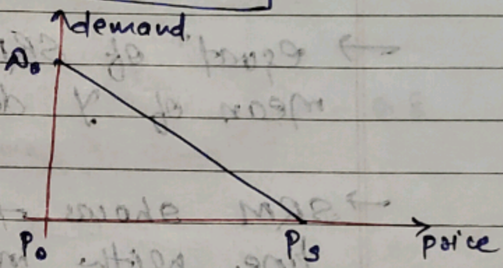
- simplest price Response curve :

$$\boxed{D(p) = D_0 - m \cdot p}$$

D_0 = demand at price = 0 (market size)
m = slope

→ the price at which demand = 0 is called Satiating price, $\boxed{P_s = \frac{D_0}{m}}$

→ elasticity of this curve is $\boxed{\varepsilon = \frac{mp}{D_0 - mp}}$

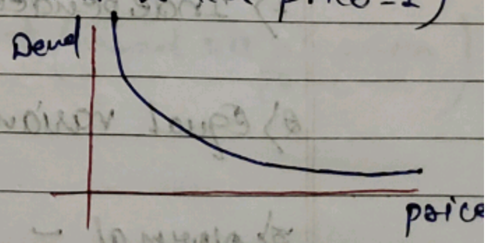


$\varepsilon = 0$ when $p = 0$ &
as $p \longrightarrow \frac{D_0}{m}$, $\varepsilon \longrightarrow \infty$

Constant elasticity Curve

→ after algebraic transition, the constant elasticity curve is given by : $\boxed{D = C \cdot p^{-\varepsilon}}$

C = constant (it is Demand when price = 1)



→ Revenue is $\boxed{R = p \cdot D}$

$$\boxed{p \cdot D = C \cdot p^{(1-\varepsilon)}}$$

lec-5

LINEAR DEMAND RESPONSE CURVE

- A SLR model can help to identify D_0 (Y-intercept) and m (slope)

- SLR tell us if the linear relationship is good fit for the data available from the market experiment.

SIMPLE REGRESSION MODEL (SRM)

→ equat^n of SRM describes how the conditional mean of Y depends on X.

→ SRM shows that these ~~means~~ MEANs lie on a line with intercept β_0 + slope β_1.

(β_0 ↳ D_0 (market size))

$$\boxed{\mu_{y|x} = E(Y|X=x) = \beta_0 + \beta_1 x}$$

→ deviation of responses around $\mu_{y|x}$ (predicted value) are called errors.

→ Error, is denoted by ϵ, and $E(\epsilon) = 0$

The SRM makes 3 assumptions about error term:

1) Independent – Errors are independent of each other.

2) Equal variance – all errors have same variance $Var(\epsilon) = \sigma_\epsilon^2$

3) Normal – the errors are normally distributed.

PA- 4

1) Consumer surplus = (Willing to pay) − (Actual price paid)

Product surplus = (Price received) − (min price producer would accept)

↳ e.g. = if a company willing to sell at ~~500~~/- but sells it for 800/- the producer surplus is 300.

2) by reducing price

3) $\varepsilon = \frac{\%\text{ chang in demand}}{\%\text{ change in price}}$ → $= \frac{(8-10)}{10} \times 100 = 20\%$.

$= \frac{(3.75 - 3.00)}{3.00} \times 100$ $\qquad \varepsilon = \frac{20}{25} = 0.8$

$= \frac{3}{12} \times 100 = 25\%$.

4) $\varepsilon = 0.9 = \frac{\%\text{ Change in demd}}{\%\text{ chge in price}}$ → $= \frac{(x - 1500)}{1500} \times 100$

$= \frac{x + 800}{15}$

$= \frac{(8000 - 10,000)}{8000} \times 100$

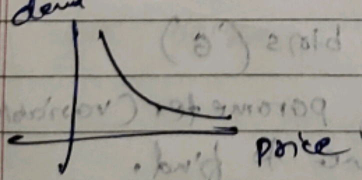
$= -20\%$.

− denote decrease

$0.9 = \frac{x - 1500}{15 \times 20}$

$x = 270 + 1500 = 1770$

5) demd.



→ No satiating point (demand = 0 at which price)

7) $E(Y \mid X = x) = \beta_0 + \beta_1 x$

β_1 → slope

β_0 ↳ (demnd at price 20)

9) linear co-relatⁿ b/w X & Y is symmetric