Hypothesis :- A premise or claim that we want to test (or) experiment

Null hypothesis: (in statistics Null hypothesis means default hypothesis)

↓

(H_0) → currently accepted value for a parameter

Alternative hypothesis : (H_a) → also called Research hypothesis
↳ Involves the claim to be tested

eg:

It is believed that a candy machine makes chocolate bars that are on average 5g. A worker claims that the machine after maintenance no longer makes 5g bars.

Write H_0 and H_a?

(default assumption) $H_0 : \mu = 5g$

$H_a : \mu \neq 5g$

> H_0 & H_a are mathematical opposites

Possible outcomes of this test:

- Reject Null hypothesis H_0
- Fail to Reject Null hypothesis H_0

2016

01 TUESDAY • MARCH

WK10 • 061-305

MAR - 2016

M T W T F
7 1 2 3
14 8 9 10
21 15 16 17
28 22 23 24 25
29 30 31

Test statistic - calculate from sample data used to decide

Ex:) Take sample 50 bars
- Get Average val
- calculate test statistic

Statistically significant - where do we draw the line to make a decision.

from 3 Random samples we are getting Averages as below

1) Avg = 5.21 → more (or) less equal to 5 so we cant Reject H_0.
2) Avg = 5.79 → little far away from 5 so we could give a thought of Rejecting H_0
3) Avg = 7.21 ✓ much Far away From 5 so we can definetly Reject H_0

Level of Confidence: C - 95%, 99%.

How confident are we in our decision.

Level of Significance $\alpha = 1 - C$

So Level of Confidence $C = 95\%$

$C = 0.95$

$\Rightarrow \alpha = 1 - C$

$= 1 - 0.95$

$\boxed{\alpha = 0.05}$

2016

Basic Statistics

1> Descriptive statistics 2> Inferential statistics

1) Descriptive statistics

↳ methods of summarizing the information we have collected for an analysis. we can summarize information by means of graphs, such as a pie chart, bar chart or numbers such as means, percentage, correlation coefficient.

2) Inferential statistics:

↳ It is about drawing conclusions about a population on the basis of only a limited number of cases.

Data and visualization:

e.g.: football

⇒ goals, winners, stopped penalties

variables ⇒ characteristics of something or someone

cases ⇒ something or someone

2016

Player themselves are

Cases { player 1, player 2, player 3 ... }

Characteristics & [variables] { body weight ... ..., age ... ..., color ... ..., goals ... ..., matches ... ... }

* A case can be individual player or Team
* & variable means features (or) characteristics of those cases (player or team)

* variable need to vary (variables should not be same, e.g) country of the team, here the case is particular Team. For the particular team has many players and all these players are from same country only, so we can treat the "country" Feature/characteristics as variable, that feature is called "Constant".

* Levels of measurement:

diff kinds of variable representing strongly diverging characteristics. for this reason it's of essential important to distinguish diff levels of measurement

2016

nominal level : made up of various categories that differ from each other. There is no order, however. (means no ranking order)

=) It's not possible to argue one that one category is better or worse, or more, or less than another.

E.g 1 : nationality of the football players.
[Spanish, French, Italy, German, India, USA]

There is no order ranking order but differ from each other

E.g 2 : Gender of the football players, city the football teams come from.

ordinal level : There is not only the difference b/w the categories of the variable, there is also an order.

E.g 1 : winners of the competition. which team is no 1, no 2, no 3. [1, 2, 3].

It's differ from each other and has order. but we dont know anything about the different b/w categories means how much the no 1 was better than no 2.

Categorical variables :

both nominal and ordinal levels are called Categorical variables.

2016

Interval level : different categories and has ranking order + also has similar intervals b/w the categories.

e.g 1: Age of the football player

[player 1 ⇒ 16 , player 2 ⇒ 18,
player 3 ⇒ 12 , player 4 ⇒ 10]

Same Intervals { 16, 18 } differs and we can say player 2 is older than player 1
{ 10, 12 } " " " "

ratio level : It is similar to Interval levels but In addition a meaningful zero point.

e.g 1 : player's body height measured in cms.

A height of 0 cm means there is no height at all.

Note: Age cannot be 0 so there is no zero point. So therefore Age is an Interval variable.

06 SUNDAY Quantitative variables:

both Interval and ratio levels are called quantitative variables. because the categories are represented by numerical variables. (or) values

2016

Quantitative variables

Discrete	Continuous
* set of separate numbers e.g: no of goals scored by Football players. can be 1, 2, 3 ... but we cant saye 1.23, 2.4 ...	* infinite region of values. A variable is continuous if the possible values of the variables form an interval. e.g: height of the players can be 170 cm, 170.25 cm, 170.0245 cm. * we dont have separate numbers, but an infinite region of values.

statistical methods

		Difference	Ranking orders	similar Intervals	meaningful zero point
Categorical	nominal	+	-	-	-
	ordinal	+	+	-	-
Quantitative	interval	+	+	+	-
	ratio	+	+	+	+

→ Discrete – set of separate numbers

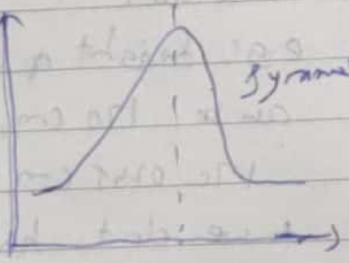
→ continuous – infinite region of values

2016

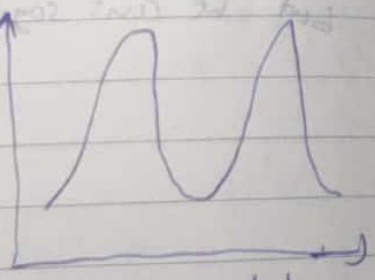
Frequency Table:
shows the how the values are distributed over the cases.

Graphs:
categorical ⇒ pie, bar chart
quantitative ⇒ histograms (continus variable)

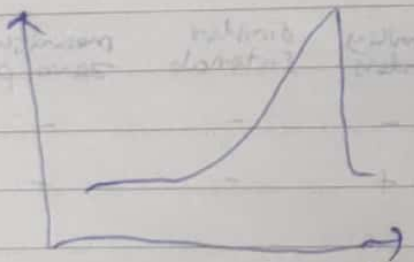
shape of the Histogram



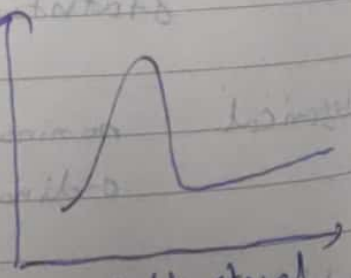
Bell shape (unimodal)



Bimodal



Right skewed



Left skewed (Annual Income of Football Players)

measures of central tendancy:

→ describe the center of a distribution

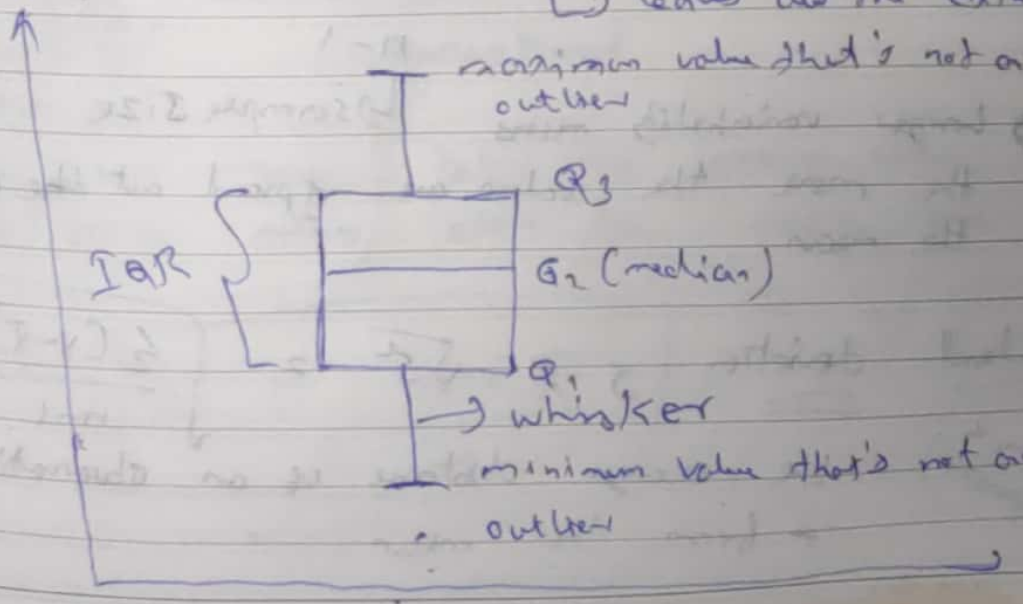
→ mode, median, mean

→ graph that nicely presents the variability of a distribution is the "box plot".

summary
↓
summarizing a distribution
graphs ↓ center
↓
mode → categorical
quantitative { median → influential outlier/skewed
mean → ($\bar{x} = \frac{\Sigma x}{n}$)

Information about the variability of the data:

↳ range ⇒ $Q_2 - Q_1$ (last value − first value)
↳ interquantile range ⇒ IQR = $Q_3 - Q_1$
↳ leaves out the extreme values

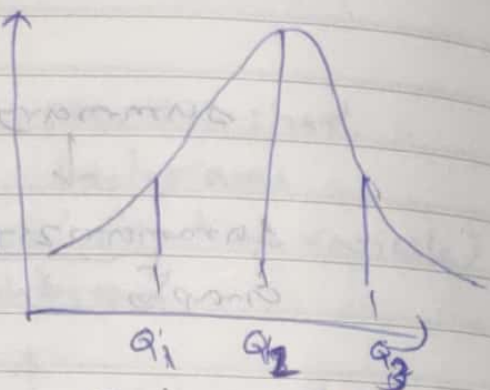


center of distribution + variability of distribution
= more complete picture

values ordered from low to high

0, 5.6, (8.7) Q1, 14.1, 14.1,

(15) Q2 (median), 17.2, 19.2, (19.3) Q3, 24.1, 27.9



outliers:
= value lower than Q1 − 1.5(IQR)
(or)
higher than Q3 + 1.5(IQR)

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

↳ sample size

↳ Larger variability means the more the values are spread out around the mean

Standard deviation $s = \sqrt{s^2} = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$

(most often used measure of dispersion)

= the average distance of an observation from the mean

2016

Z-scores

Sometimes we want to know if a specific observation is common or exceptional. To answer that question, they express a score in terms of the number of standard deviations it is removed from the mean. This no is called Z-score.

if we recode orginal scores into z-scores we say that we standardize the variable

recode orginal scores into z-scores

↓

Standardization

↓

replace the orginal scores by standard deviation (s) from the mean ($\bar{x}$)

+ easy to see whether a specific score is relatively common or exceptional

mean $\bar{x}$ = balance point

Z-scores

(-) scores means below the mean value
(+) scores means above the mean value

$$z = \frac{x - \bar{x}}{s}$$

2016

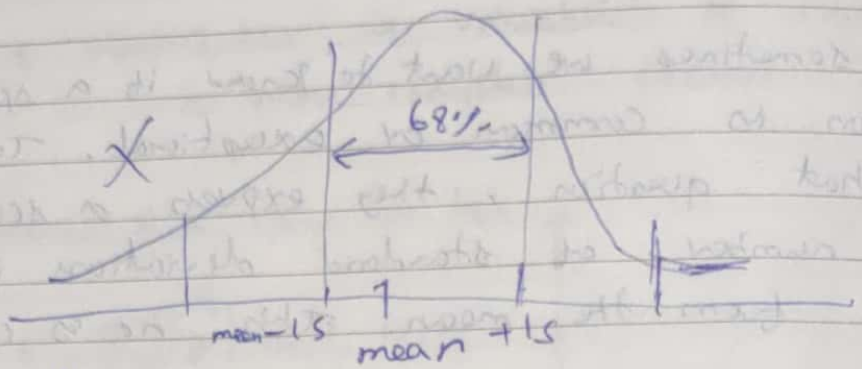
12 SATURDAY • MARCH

WK11 • 072-294

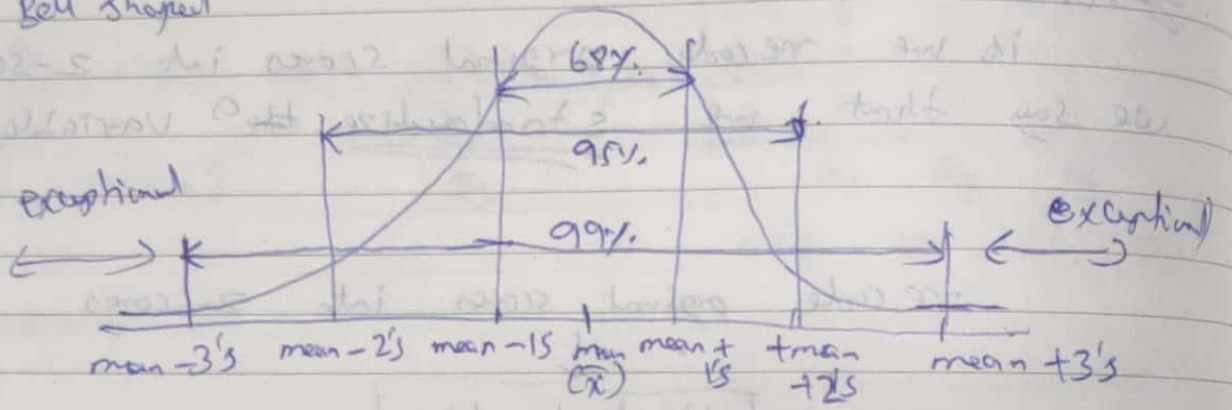
MAR - 2016

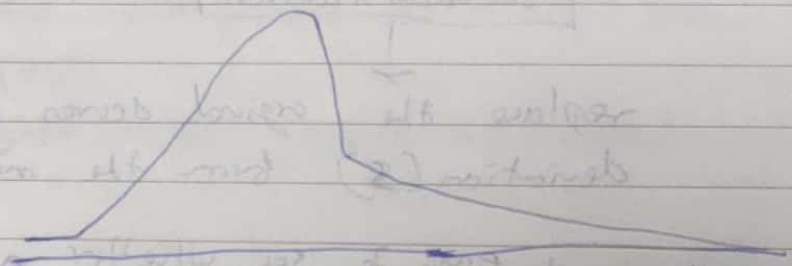
M T W T F S S
1 2 3 4 5 6
7 8 9 10 11 12 13
14 15 16 17 18 19 20
21 22 23 24 25 26 27
28 29 30 31

Rule:



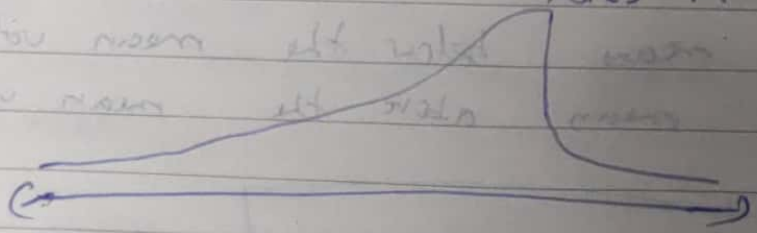
Bell shaped





Large positive z-scores are more common because more extreme values lie's in Right side.

13 SUNDAY



Large negative z scores are more common b'cz more extreme values lay in left side

2016

& 2 scores are more useful to compare different distributions.

Example:

city with 8 high schools, average grade for Chemistry has b/w 0 to 10

School name	Average grade Chemistry
School 1	7.4
" 2	7.9
" 3	4.1
" 4	8.1
" 5	6.2
" 6	7.1
" 7	7.4
" 8	6.7

1) What does the distribution of the variable "average grade for Chemistry" looks like?

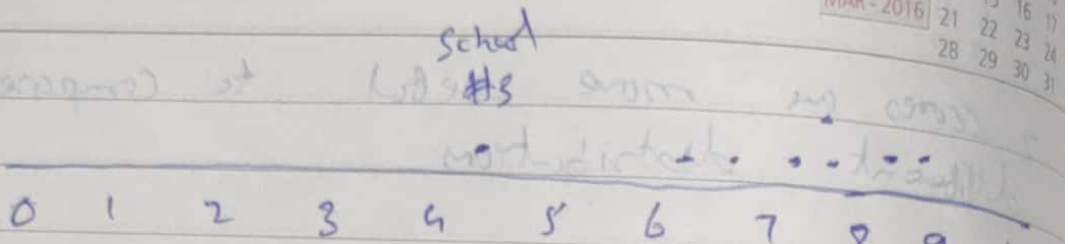
2) what is the center of the distribution?
3) The variability of the distribution?
4) Construct a box plot
5) what is the z-score of School #3?

15 TUESDAY • MARCH

WK12 • 075-291

MAR - 2016

1)



2) mode => 4.1, 6.2, 6.7, 7.1, 7.4, 7.4, 7.9, 8.1

=> 7.4 occurs twice

median => $\frac{(7.1+7.4)}{2} = 7.25$

mean => $\frac{\Sigma x}{n} = \frac{54.9}{8} = 6.86$

3)

Range => highest value - Lowest value

$8.1 - 4.1$

$= 4$

IQR => 4.1, 6.2, 6.7, 7.1 | 7.4, 7.4, 7.9, 8.1

Q_1 median Q_2 Q_3

$\frac{6.2+6.7}{2}$ $\frac{7.4+7.9}{2}$

$= 6.45$ $= 7.65$

$= Q_3 - Q_1 = 7.65 - 6.45$

$IQR = 1.2$

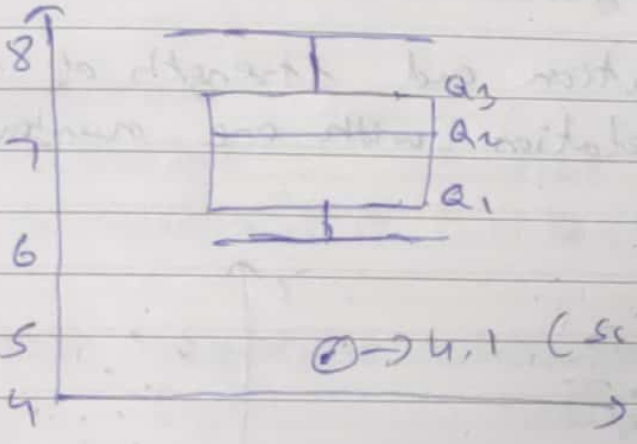
Standard deviation $(s) = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

$s = 1.27$

2016

4) $Q_1 = 6.45$, Q_2 = median = 7.25, $Q_3 = 7.65$,
IQR = 1.21

outlier = value lower than $Q_1 - 1.5(IQR)$
or higher than $Q_3 + 1.5(IQR)$



5) $z = \frac{x - \bar{x}}{s} = \frac{4.1 - 6.86}{1.27} = -2.17$

school #3
+
average of 4.1
+

lies 2.17 standard deviation below the mean and can therefore be conceived of as a rather exceptional value.

17 THURSDAY • MARCH

MAR - 2016

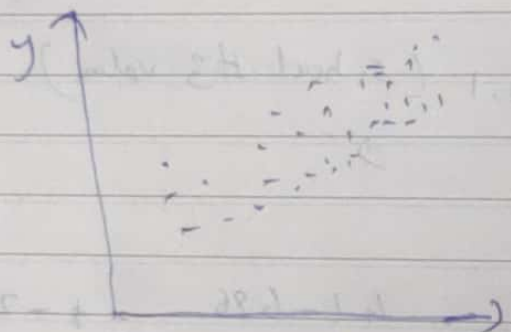
WK12 • 077-289

Contingency table → nominal variables
ordinal "

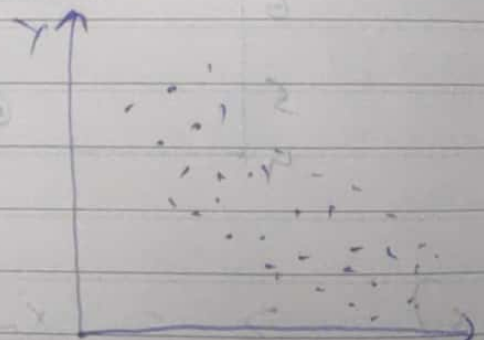
scatterplot → quantitative variables

pearson's R (correlation)

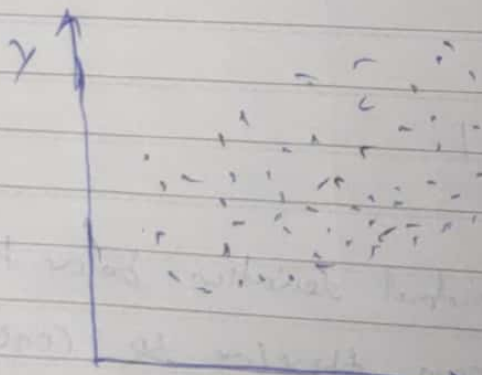
↳ direction and strength of linear correlation with one number.



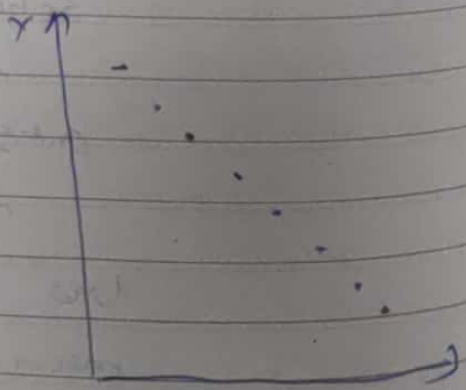
strong positive linear relationship



strong negative linear relationship

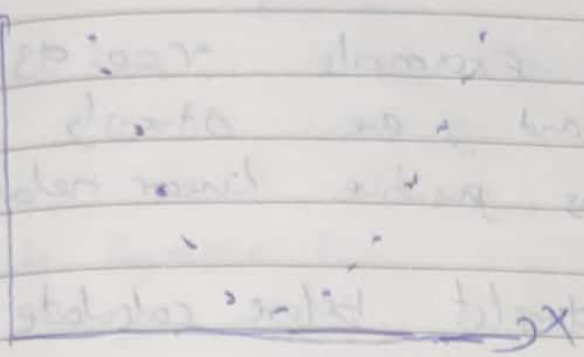


weak positive linear relationship



perfect negative linear relationship

2016



Curvilinear relationship

Pearson's R ⇒ how strong or weak correlation.

↳ direction
+ = positive
− = negative

↓ Strength
−1 = perfect negative
+1 = perfect positive

e.g.: body weight vs chocolate consumption.
(y) (x)
(dependent variable) ↳ (Independent variable)

Person	X	Y	Σx	Σy	$\Sigma x * \Sigma y$
P1	50	50	−1.01	−1.15	1.17
P2	100	70	−0.56	−0.07	0.04
P3	200	70	0.34	−0.07	−0.02
P4	300	95	1.24	1.29	1.60

Pearson's R ⇒ $r = \frac{\Sigma \Sigma x \Sigma y}{n-1}$

$r = 2.78/3 = 0.93$

2016

So for this example $r = 0.93$ says that x and y are strongly correlated i.e. strong positive linear relationship

Note: check scatterplot before calculate pearson's

e.g.: for curvilinear case

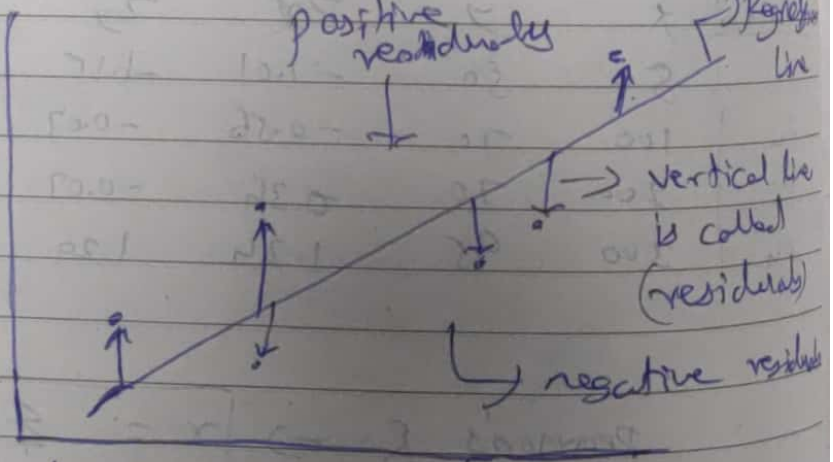
is $r = -0.15$ means not weak negative correlation it is weak linear correlation

$r = -0.15$ = weak linear correlation

Regression:

Find Regression line

line with the smallest sum of squared residuals



20 SUNDAY

OLS ⇒ (ordinary Least Square Analysis)

⇒ used to find the Regression line

2016

describes
Regression Line: linear Relationship between two variables.

Compute Regression line:

$$\hat{y} = a + bx$$

$$b = r\left(\frac{s_y}{s_x}\right) \quad \text{(standard deviation of } y\text{)} \qquad a = \bar{y} - b(\bar{x})$$

R squared (R^2)

a) tells you how much better a regression line predicts the value of a dependent variable than the mean of the variable

b) tells you how much of the variance in your dependent variable is explained by your independent variable.

e.g. $r^2 = 0.69$ means ⇒ prediction error is 69% smaller than when you use the mean
⇒ 69% of the variance in the values can be predicted by the previous values.

the higher the value means higher variance

Randomness:

Randomness it is not a property of a phenomenon.

humans → are not very good in assessing randomness.
→ over Interpretation of randomness (apophenia)
→ bad in constructing random data.

probability:-

→ probability is way to quantify randomness.
→ $0 \leq$ probability ≤ 1
→ Σ probability $= 1$
* Important terms:
'experiment', 'event', 'independent trails', 'relative frequency.
→ law of Large numbers
→ keep calm and carry on many trails.

sample space, events & tree diagrams:

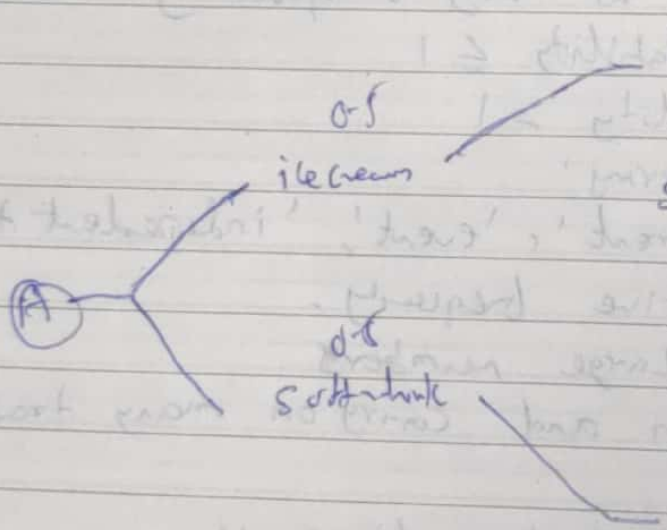
sample space → all possible outcomes for the experiment (or) collections of all possible outcomes for a random phenomenon

elementary (or) Combined outcomes (or) subset of sample space → events

2016

* note:
never know whether your assessment of the probability was correct only three trials.
↓
no accurate probability estimate

e.g: i) Box have 2 softdrinks, 1 icecream,
ii) There are 4 persons waiting to take 5th from box
what is the probability of 4th person's pick



Venn diagram:

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

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Note: joint means intersection

joint and marginal probabilities

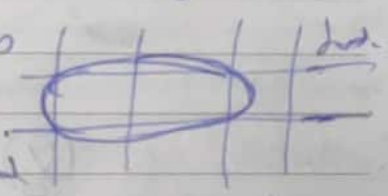
e.g.:

gender	rest	active	swim	total	
male	34	12	5	51	51
female	45	8	9	62	
Total	79	20	14	113	113

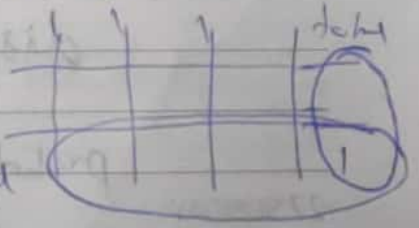
into percentage

	0.301	0.106	0.044	0.451
	0.398	0.071	0.080	0.549
	0.699	0.177	0.124	1

Joint probabilities:
→ probabilities for the intersection of certain outcomes of the variables



marginal probabilities:
→ probabilities for an outcome of each individual variable



Conditional probability : $P(A/B)$

the probability of an event, given that another event occurs.

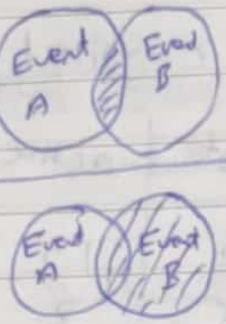
A given B = A/B

$P(A \text{ given } B) = P(A/B) = \frac{P(A \text{ and } B)}{P(B)}$

2016

$P(A/B) \leq P(B)$

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



Independence => knowing the outcome of one event does not influence your knowledge about the outcome for the other.

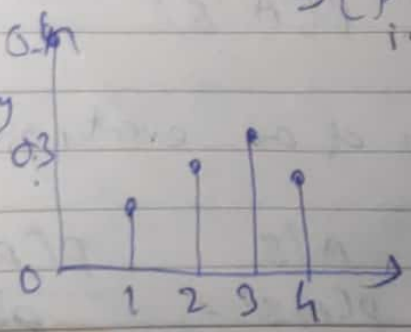
disjoint (dependant) => if one event occurs the other cannot occur. So the joint probability is zero.

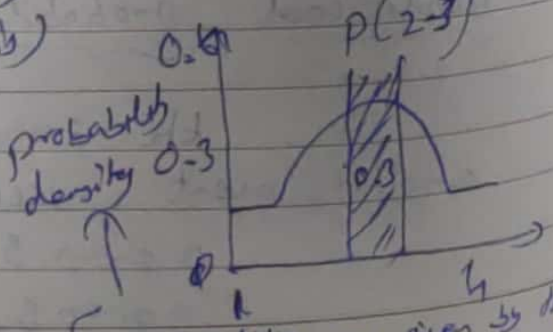
Random Variable → [variables whose possible values are numerical outcomes of a random phenomenon]

probability distribution

discrete (countable no of distinct values) → probability mass function → (probability in y axis)

continuous (infinite no of possible values) → probability density function





(probabilities are given by the surface area under the curve)

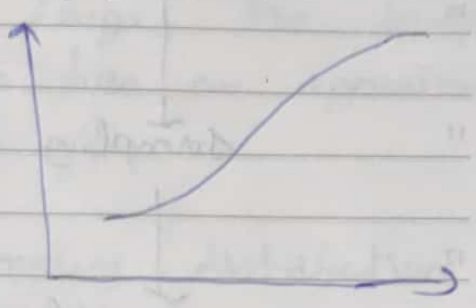
27 SUNDAY

2016

Cumulative probability:

a) discrete

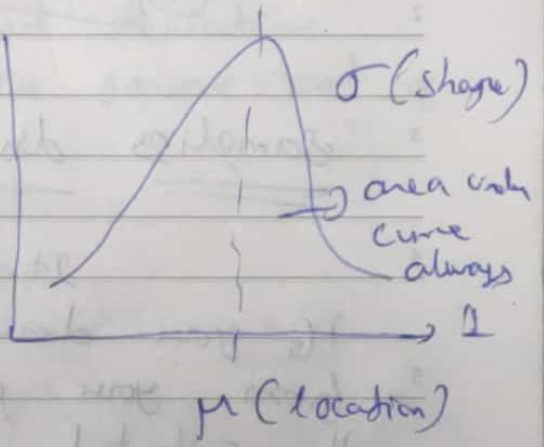
b) continuous



Normal distribution:

Equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(X-M)}{\sigma}\right)^2}$$



populations :- All data
sample - subset of population (all data)

sample —(for conclusion)→ populations

statistics — Inferential statistics → parameters

statistics:
roman characters
$\bar{x}$ = mean
s = standard deviation

parameters:
Greek letters
μ = mean
σ = standard deviation in population

2016

Simple Random Sample

population → all student

↓

sampling frame → [collection of all student names]

↓

sample data → 200 students [from the the above collection frame]

(X) Bigger is Better → Sample data

Sampling distribution of sample mean:

It is the distribution of that you get if you draw an infinite number of samples from your population and compute the mean of all the collected sample means.

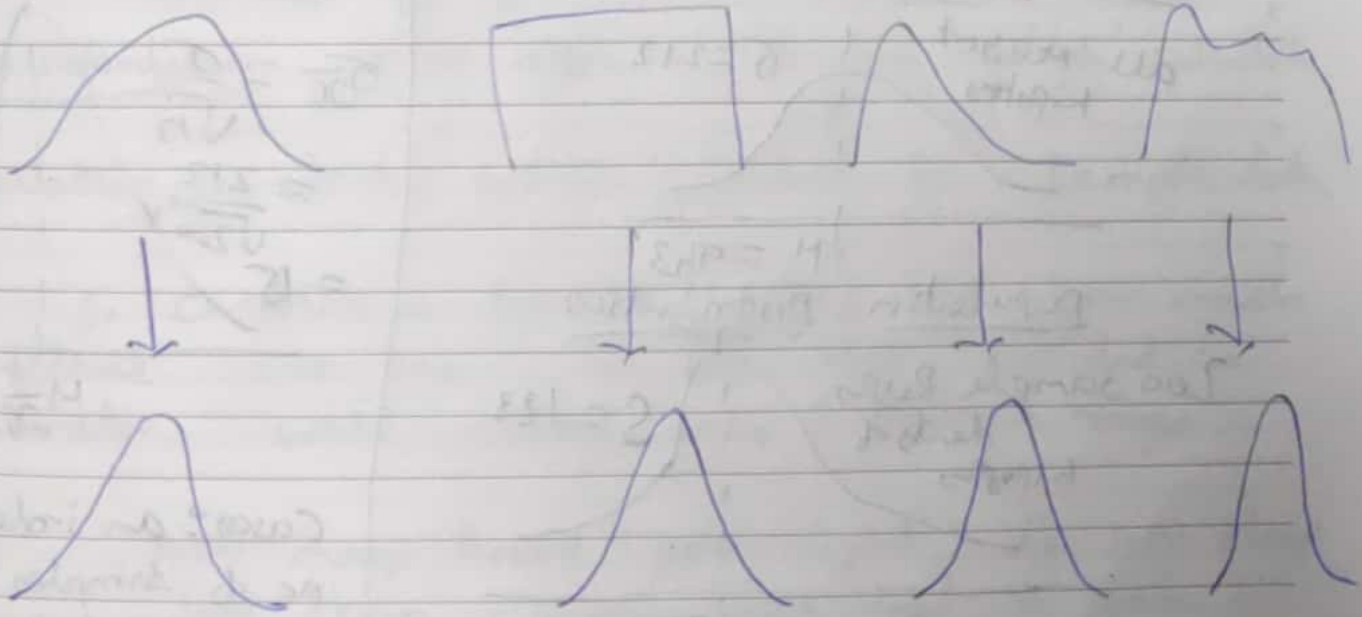
Central Limit theorem :-

* It says that, provided that the sample size is sufficiently large, the sampling distribution of the sample mean has an approximately normal distribution

* The mean of the sampling distribution equals the population mean, and ~~the~~

* Standard deviation ~~of~~ of the sampling distribution equals the standard deviation in the population divided by the square root of the sample size.

population distribution of mean



d → Sampling distribution of sample mean with sample size of $n = 30$

(No of samples)

2016

$$\mu_{\bar{x}} = \mu \text{ (population mean)}$$

where $\bar{x}$ is sample mean

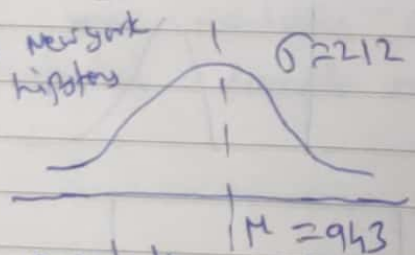
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where

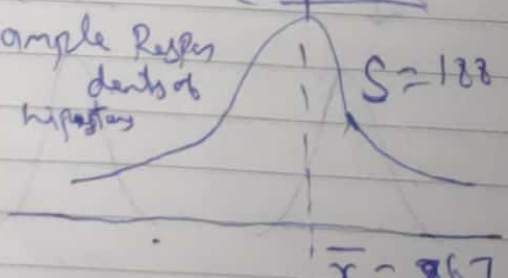
σ = standard deviation of population
$\bar{x}$ = " " " " sample mean
n = sample size
$\sigma_{\bar{x}}$ = stand deviation of sampling distribution

* The larger the variability in the population the larger the variability in the sample means

Three Distributions:



Population Distribution

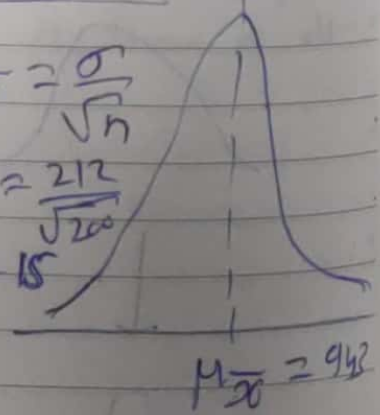


Data Distribution

(theoretical distribution)

Sampling Distribution

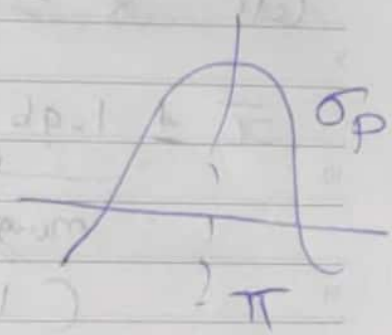
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{212}{\sqrt{200}} = 15$$



Cases: an indefinite no of samples of 200 respondents from the population of new york hipsters

Sampling distribution proportion:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$



where σ_p = Standard deviation of sampling distributions of sample proportion

p = sample proportions

Binary categorical variables

↓

Compute proportions

↓

π

Confidence Interval:

My understand: our objective is we need to calculate population parameter which is done by sample data.

So confidence level is nothing but how much confidence we are to say that population parameter will fall into certain range.

Ex: how many hours per night do you sleep less than before you had a baby?

2016

02 SATURDAY • APRIL

WK14 • 093-273

APR - 2016

M T W T F S S
4 5 6 7 1 2 3
11 12 13 14 8 9 10
18 19 20 21 15 16 17
25 26 27 28 29 30 24

lets $\bar{x} = 2.6$, $\sigma = 1.1$ [Value from 60 Readings]

$\bar{x} \pm 1.96\,\sigma_{\bar{x}}$ where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

CI for mean with the known population std. dev

margin of error

$(1.96 * 0.142 = 0.28)$

So

2.6 ± 0.28

$\Rightarrow$ 95% Confidence Interval

$(22.32, 2.88)$

$\bar{x} \pm Z_{95\%}\,\sigma_{\bar{x}}$

where $Z_{95\%} \rightarrow$ Z score of 95% confidence

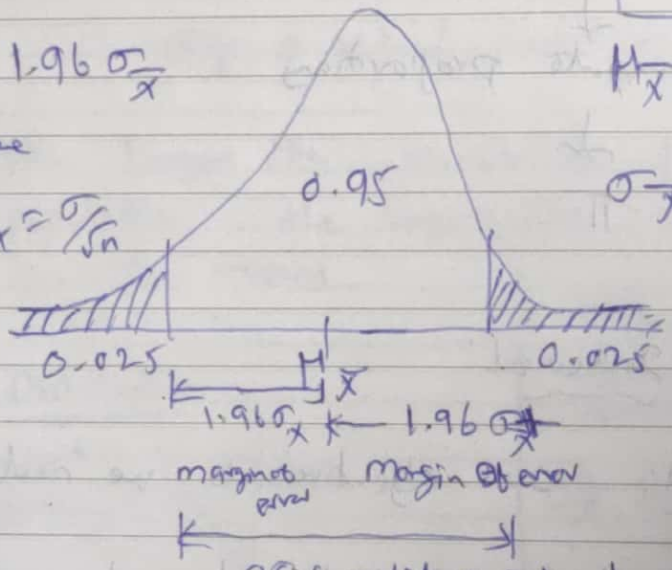
$\bar{x} \pm 1.96\,\sigma_{\bar{x}}$

where

$\sigma_x = \sigma/\sqrt{n}$

$\mu_{\bar{x}} = \mu$

$\sigma_{\bar{x}} = \sigma/\sqrt{n}$



03 SUNDAY

Confidence Interval

∞ (samples)

Compute the confidence intervals

in 95% of the samples the population value will fall within the confidence interval

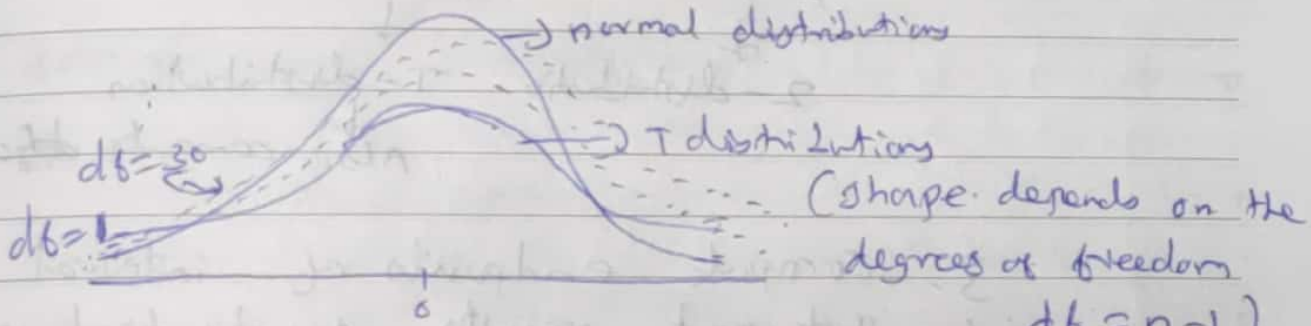
2016

CI for mean with unknown standard deviation:

Assumption's:

1. randomization.
2. approximately normal distribution

the worry of extreme outliers



(shape depends on the degrees of freedom df = n-1)

eg:

how many hours per night do you sleep less than before you had a baby?

$\bar{x} = 2.6 \quad s = 0.9$

$\bar{x} \pm t_{95\%} (se)$

where $t_{95\%}$ => t score at 95% confidence

$se = \frac{s}{\sqrt{n}}$

//ly

CI for proportion: $p \pm Z_{95\%} (se)$

where $se = \sqrt{\frac{p(1-p)}{n}}$

p = population mean

2016

Constructing a confidence interval

1. which confidence level?
2. proportion (or) mean?

proportion → Z-distribution
mean → T-distribution
↓
Also compute df (df = n-1)

3. Compute endpoints of interval
4. Interpret results substantively

Choosing the sample size:

Sample size
Mean
↓
1. magnitude of desired margin of error
(smaller margin of error → Larger sample size)
2. confidence level
(larger confidence level → larger sample size)
3. variability
(larger the std deviation → larger your sample size)

$n = \frac{\sigma^2 z^2}{m^2}$ where σ ⇒ std dev of population
z ⇒ z score respective to Confidence Interval
m ⇒ margin of error you want to maintain

sample size
[proportions]

$$n = \frac{p(1-p)z^2}{m^2}$$

sample size computations
↓
ideal world
↓
super large sample

real world → we don't have enough money to compute super large sample so we go for
↓
sample size computations

* Significance Test:

e.g.: → ## (dealing with proportions)

1. More than half of all certified divers in America have more than 35 hours of diving experience.

→ μ (dealing with mean)

2. Mean no of hours of diving experience of all certified divers in America is more than 35?

2016

07 THURSDAY • APRIL

WK15 • 098-268

APR - 2016

M T W T F S S
4 5 6 7 1 2 3
11 12 13 14 8 9 10
18 19 20 21 15 16 17
25 26 27 28 22 23 24
29 30

Example 1 — How many hours of experience

$p(>35 \text{ hours}) = 0.57$ — $n = 500$

proportion → (1) proportion or mean ← mean

$H_0: \pi = \pi_0 = 0.5$ — (2) Formulate hypothesis

$H_a: \pi \neq \pi_0$

$\pi > \pi_0$ → Right Tailed Test

$\pi < \pi_0$ → Left Tailed Test

Randomization

$n\pi \geq 15$ and

$n(1-\pi) \geq 15$

$\alpha = 0.05$ — (4) determine α

$$z = \frac{p - \pi_0}{se_0}$$ — (5) compute statistic

$$se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

$$z = \frac{(0.57 - 0.5)}{\sqrt{\frac{0.5(0.5)}{500}}}$$

$z = 3.13$

Example 2

$\bar{x} = 35.5$ $S = 8$

$H_0: \mu = \mu_0 = 35$

$H_a: \mu \neq \mu_0$

$\mu > \mu_0$ ← Right Tailed Test

$\mu < \mu_0$ ← Left Tailed Test

Randomization

Population distribution approximately normal (if one tailed test with small n)

$\alpha = 0.05$

$$t = \frac{\bar{x} - \mu_0}{se}$$

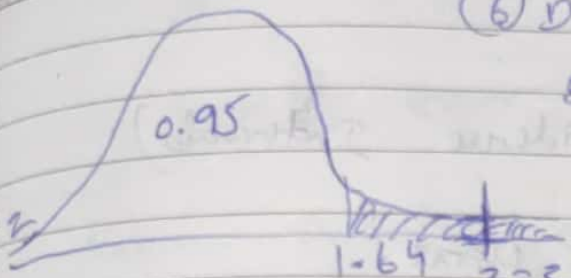
$se = S/\sqrt{n}$

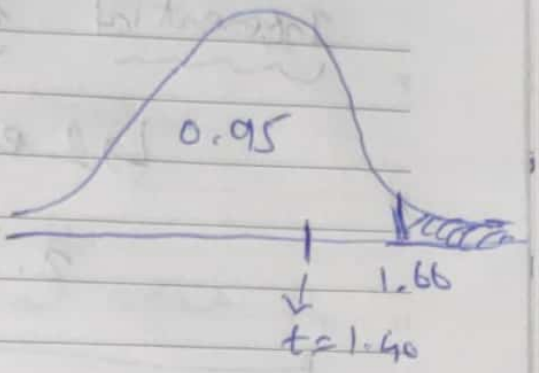
$$t = \frac{35.5 - 35}{\frac{8}{\sqrt{500}}}$$

$t = 1.40$

2016

(6) Draw Sampling Distribution





(7) Find location of ~~test~~ statistic.

(8) Reject H_0?

yes
because z value falls in Rejection Area

No
because t value do not fall in rejection Area

(9) Interpret findings

more than half of all Certified drivers in America have more than 35 hours of duty experience

we cannot conclude that their mean no of duty exp is greater than 35 hours

09 SATURDAY • APRIL

WK15 • 100-266

APR - 2016

M	T	W	T	F	S	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

Inferential statistics:

1. estimation (confidence Intervals)
2. Significance Tests & Point

p-value in a two tailed significance test ≤ 0.05	= 95% confidence interval does NOT contain H_0 value
p-value in a two tailed significance test > 0.05	= 95% of confidence interval will contain the H_0 value

Type I and Type II error:

	set free	convicted (punished)
defendant innocent	✓	x
" guilty	x	✓

↓

Significance Test

	not rejected	rejected
H_0 True	✓	x (Type I error)
H_0 False	x (Type II error)	✓

10 SUNDAY

2016

Type I error : → False positive

Type II error : → False negative

if p-value $\leq \alpha$ → reject H_0
p-value $> \alpha$ → do not reject H_0

Chi-Squared Test:

$$x^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

how to choose degree of freedom for this distribution?

D.o.F = (No. of Row − 1) * (No. of collumn − 1).

	Fruit	Flowers	mixed	
Early Renaissance	11	5	1	17
late "	8	6	8	22
Baroque	3	10	12	25
	22	21	21	64

⇒ $(3-1)^*(3-1)^* = (2)^*(2)^* = 4$ D.o.F

if the p value is < 0.05 the two variables dependent

0 → dependent
1 → independent

2016

APRIL • WEDNESDAY 13

WK16 • 104-262

Machine learning: [Once written, they can keep on learning on its own and evolve as new data gets introduced with no human intervention required]

History of ML:

1642 - Mechanical Adder
1801 - First data storage of data
1847 - Boolean Logic
1890 - Mechanical system for statistical calculations
1950 - Turing Test
1952 - The perceptron
1967 - Pattern Recognition
1979 - Stanford cart
1981 - Explanation Based Learning
1990's - Machine learning Applications
2000's - Adaptive programming

Artificial Intelligence:

It is the study of how to train the computers so that computers can do things which at present human can do better.

Machine learning:

ML is said to learn from experience.

Artificial Intelligence	Machine Learning
* ↑ chance of success not accuracy	↑ accuracy, and doesn't care about success
* The goal is to simulate natural Intelligence to solve complex problem.	The goal is to learn from data on certain task to maximize the performance of machine on this task
* AI is decision making	ML allows system to learn new things from data.
* AI will go for finding the optimal solution	ML will go for only solution for that whether it is optimal or not.
* AI leads to Intelligence or wisdom	ML leads to Knowledge.

Limitations of ML:

* Performance
* Computation

Feb 9

machine learning process

collecting data → pre processing and preparing dat → Exploring data → choosing a model → Training a the model → Evaluating the model → Improving the performance of model.

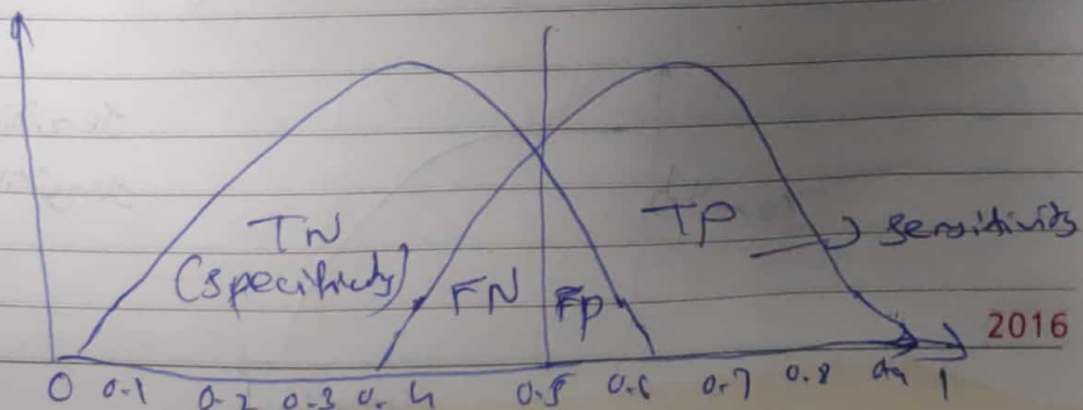
Confusion matrix:-

It's a technique used for summarizing the performance of the classification algorithm.

eg: to predict cancer Yes/No

n=165	Predicted: NO	Predicted: Yes	Total
Actual: NO	TN = 50	FP = 10	60
Actual yes	FN = 5	TP = 100	105
	55	110	165

↓ Type II error ↓ Type I error



2016

Sensitivity (or) Recall:

the proportion of patients that were identified correctly to have the disease. (i.e) True positive upon the total number of patients who actually have the disease is called Sensitivity or Recall.

$$\text{Recall} = \frac{TP}{TP + FN}$$

	Actual 1	Actual 0
predicted 1	TP	FP
predicted 0	FN	TN

II^ly

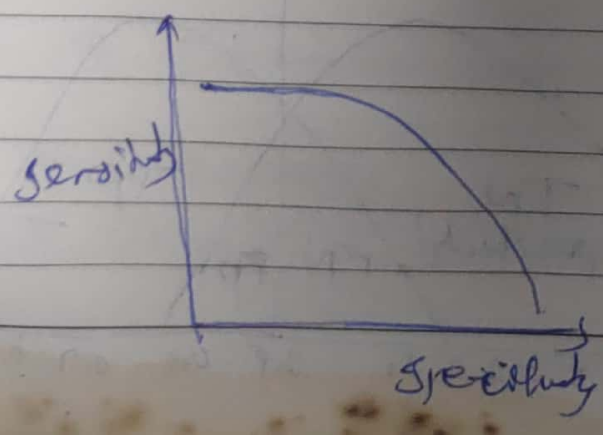
Specificity:

the proportion of Patients that were identified Correctly ~~not~~ not to have the disease i.e) True negative upon the total number of Patients who do not have the disease.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

	1	0
1	TP	FP
0	FN	TN

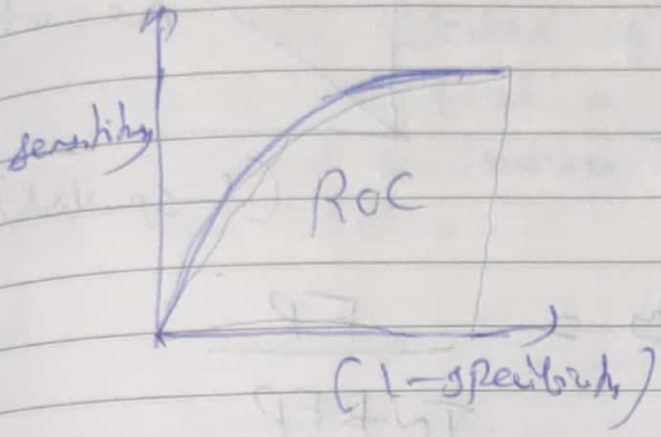
17 SUNDAY



Sensitivity ↑ specificity ↓
sensitivity ↓ specificity ↑

2016

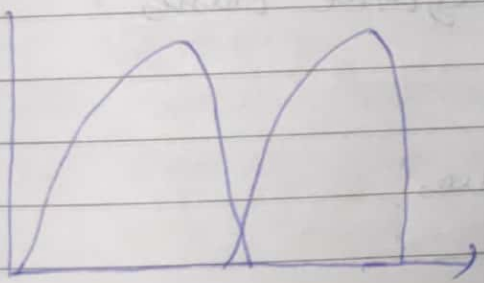
ROC curve: Receiver operating characteristics

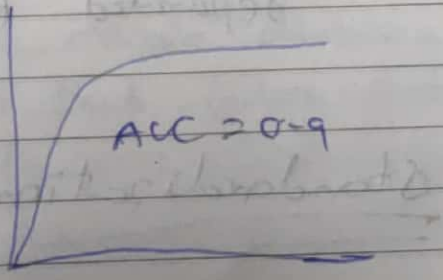


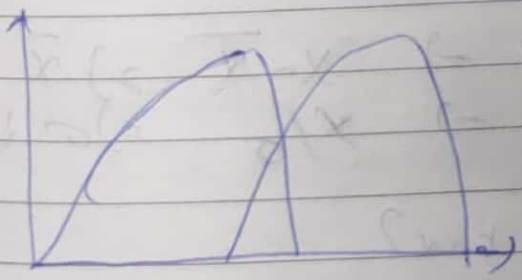
AUC → Area under the curve

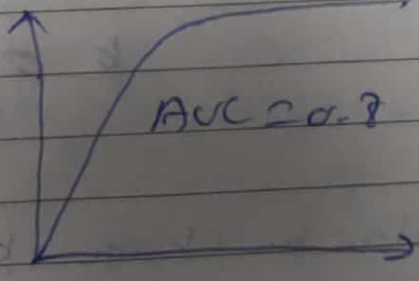
The AUC is the Area under the ROC curve. This score gives us a good idea of how well the model performances.

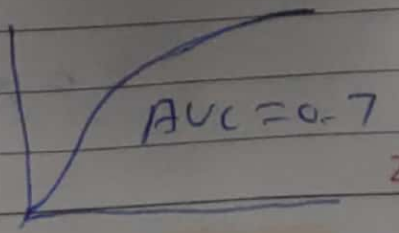
eg:



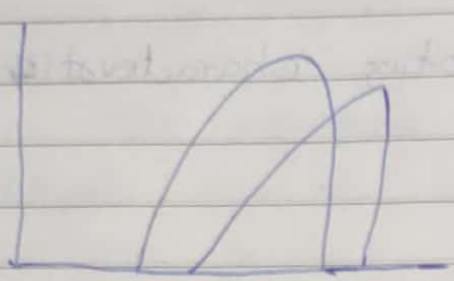








2016



AUC = 0.5

(1 - specificity)

1 - specificity = $\frac{FP}{TN + FP}$

Sensitivity → True positive Rate

(1 - specificity) → False positive Rate

" The AUC ROC indicates how well the probabilities from the positive classes are separated from the negative classes "

Standardization Techniques:

K = 1) Center → $X - \bar{X}$ 2) $\bar{X}$ is mean

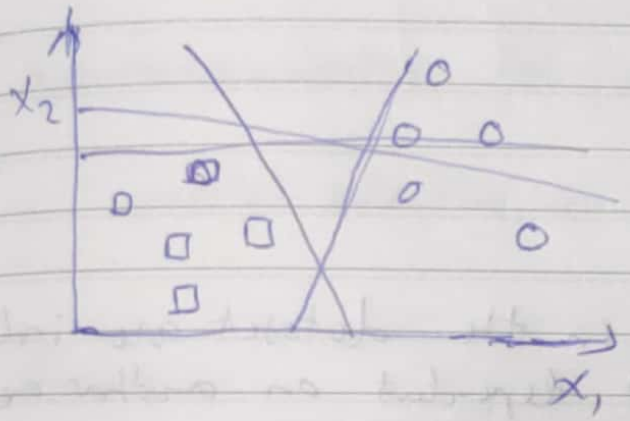
4) Scale → X/σ 2) σ is std. dev

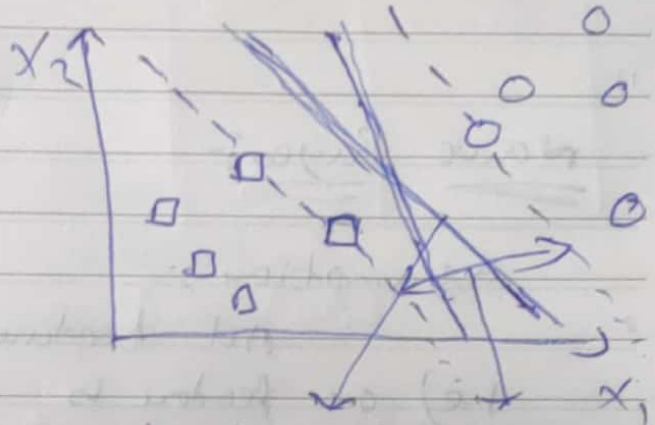
How to choose K in KNN?

$K = sqrt(N)$ where N → Total no of Records in dataset

SVM - Support vector machines:

objective: The objective of support vector machine algorithm is to find a hyperplane in an N dimensional space (N - # of features) that distinctly classifies the data points.

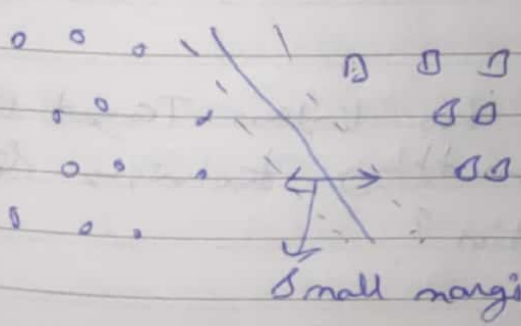


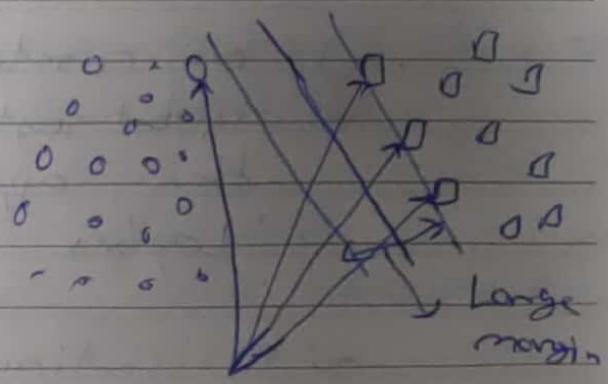


optimal hyperplane maximum margin.

maximum margin → maximum distance b/w the data points of both classes.

By having maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.





Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane.

Support vectors

2016

21 THURSDAY • APRIL

WK17 • 112-254

APR - 2016

in Logistic Regression we take linear function output and squash the value within the range of [0,1] using the sigmoid function.

In SVM we squashed the linear function output values to [-1, 1]

Naive Bayes:-

Assumption:

All features in the dataset are independent i.e) one feature is not dependent on another one.

$$P(C|x) = \frac{P(x|C)\,(P(C))}{P(x)}$$

$P(C|x)$ → posterior probability

$P(x|C)$ → likelihood

$P(C)$ → class prior probability

$P(x)$ → predictor prior probability

Tips to improve Naive Bayes model:

→ No correlation

→ If test data has unseen Target value then apply smoothing techniques like "Laplace Correction"

→ If Continuous variable do not have normal distribution, should use Transformation or diff methods to convert it in normal distribution.

2016

Application:

→ mostly used in Text classification.
→ Spam filtering / sentiment Analysis
→ Recommendation system.
→ multi class predictions

Hyperparameter for NB:

model selection ← Gaussian
multinomial
Bernoulli

Gaussian:-
It is used in classification and it assumes that features follow a normal distribution

multinomial:
It is used for discrete counts.
e.g: how often word occurs in the document (count)

Bernoulli:
If the feature vectors are binary (0's & 1's)
e.g: Text classification with "Bag of words" model where
'1' → word occurs in the document
'0' → word does not occur in the document

2016

NB example:

weather Vs play

weather	play
Sunny	No
overcast	Yes
Rainy	yes
Sunny	yes
.	.
.	.
.	.
n	n

Fregueny Table

weather	NO	Yes
overcast		4
Rainy	3	2
Sunny	2	3
Total	5	9

Likelihood Table

~~overcast~~		4	= 4/14 = 0.29
Rainy	3	2	= 5/14 = 0.36
Sunny	2	3	= 5/14 = 0.36
	= 5/14	= 9/14	
	= 0.36	= 0.64	

problem: players will play if weather is sunny. is this statement correct?

$$P(yes/sunny) = \frac{P(sunny/yes) * (P(yes))}{P(sunny)}$$

1) $P(sunny/yes) = 3/9 = 0.33$,

$P(sunny) = 5/14 = 0.36$

$P(yes) = 9/14 = 0.64$

$$P(\text{yes}/\text{sunny}) = \frac{0.33(0.64)}{0.36}$$

$$= 0.60$$

→ higher probability

So statement is correct.