Welcome Everyone

Agenda :-

1. ANOVA
2. EDA & Feature Engineering
3. CRISP -DM
4. DS libraries → numpy, pandas, matplotlib
5. Practical of statistics
6. Few projects and EDA parts

Hypotheses testing

0. write everything given. Finalize the test
1. H_0 & H_1
2. Decision boundary, CI, α
3. Test statistics, p-value or direct, value
4. Decision & Conclusion

ANOVA ⇒ analysis of variance

anova is a statistical test used to compare mean of 2 or more groups. & see if atleast one of them is significantly different from the other.

2 main important things to understand

1. Factor
2. Level

G

M
F
M
F

G with 2 levels

Gender

M
F
O
O
M
F

Medicine

Dolo ⇐ factor

500
650
900

} ⇒ levels

Say you are given all of these dosages as a part of test group & you are asked to rate out of 10, how much was each dosage effective?

Medicine → Factor

500	650	900	→ levels
7	9	5	
8	3	6	
7	9	8	
8	3	5	
3	1	9	

{Gender} Factor / Variable

{M, F, O} Levels / Sub-categories

Assumption in Anova

1. Normality of means of sample distribution
   ⇒ means of sample dist. should be normally distributed
2. Absence of Outliers ⇒ outliers needs to be removed
3. Homogenity of variance of samples
   ⇒ sample should have same variance
4. Population variance in diff levels of each independent factor/variable are equal.
5. Samples are random & independent.

Types of ANNOVA

1. One way Annova : When we have one factor with atleast 3 levels & all the levels are independent.

   Gender ⇒ M F Others

   Data ⇒ 600 500 900

2. Repeated Measure Anova :- We have one factor with atleast 3 levels but levels are dependent.

   Sleep

Day 1	Day 2	Day 3
18	6	6

   There will be some relationship b/w levels.

3. Factorial Anova :- 2 or more factors & each of which have atleast 2 levels. Levels can be dep. or indep.

		Medicine		
		500	650	900
Gender	M	3	5	4
		3	6	8
		8	7	1
	F	2	6	8
		2	6	9
		3	7	10

← random variable

One way anova [F-test]

⇒ used for inferential statistics

⇒ Comparing mean of more than 2 groups

Hypotheses in anova

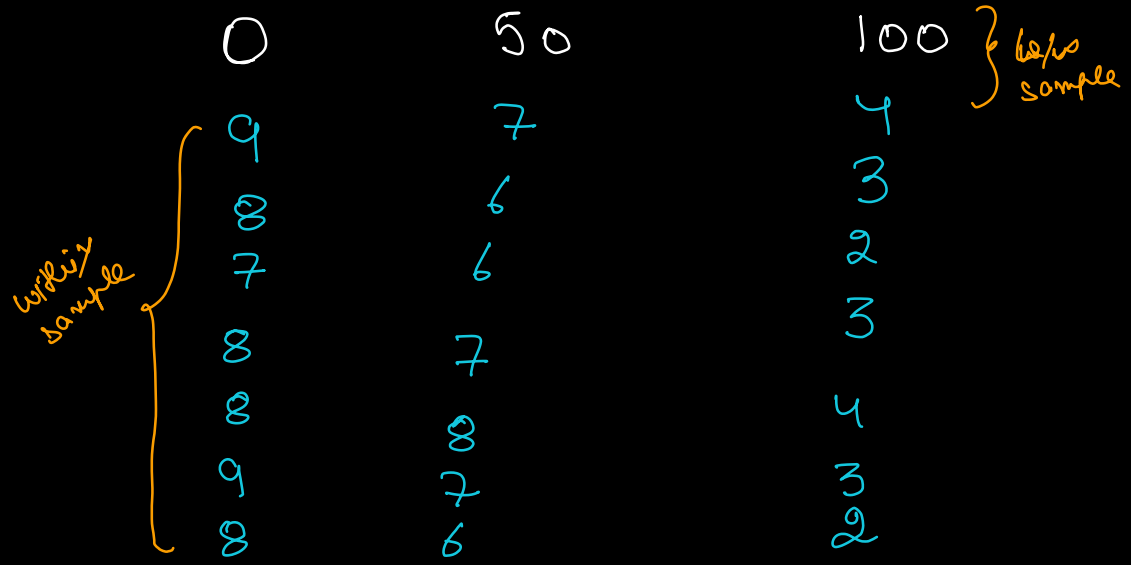
$H_0 \Rightarrow \mu_1 = \mu_2 = \mu_3 \cdots \mu_n$

$H_1 \Rightarrow$ atleast one is not equal

Test statistic

$$F = \frac{\text{Variance between sample}}{\text{Variance within sample}}$$

Q :- Researchers want to test a new medicine. They split participants into 3 levels (dosage) & asked them to rate the effect of medication from 1-10



Are there any diff. b/w 3 dosages using $\alpha = 0.05$?

① $H_0 \Rightarrow \mu_0 = \mu_{50} = \mu_{100}$

$H_1 \Rightarrow$ mean are not equal

② $\alpha = 0.05$ C.I. 95%

③ In f test we calculate d.o.f.

$n = 7$ [total data points in a sample in a level]

$N = 21$ [all data points collected]

$a = 3$ [no. of categories]

In ANOVA, we use sample data to make inference about population parameters.

Hypothesis test in ANOVA compares the population mean, but by using sample means

$df_{B/w} = a - 1 = 2$

$df_{within} = N - a = 18$

$df_{total} = N - 1 = 20$

$\alpha = 0.05$

f-value [$df_{B/w}$, df_{within}]

[2, 18]

Critical Values of the F-Distribution $\alpha = 0.05$

Denom. d.f.	Numerator Degrees of Freedom 1	2	3	4	5	6	
1	161.448	199.500	215.707	224.583	230.162	233.986	236.7
2	18.513	19.000	19.164	19.247	19.296	19.330	19.3
3	10.128	9.552	9.277	9.117	9.013	8.941	8.8
4	7.709	6.944	6.591	6.388	6.256	6.163	6.0
5	6.608	5.786	5.409	5.192	5.050	4.950	4.8
6	5.987	5.143	4.757	4.534	4.387	4.284	4.2
7	5.591	4.737	4.347	4.120	3.972	3.866	3.7
8	5.318	4.459	4.066	3.838	3.687	3.581	3.5
9	5.117	4.256	3.863	3.633	3.482	3.374	3.2
10	4.965	4.103	3.708	3.478	3.326	3.217	3.1
11	4.844	3.982	3.587	3.357	3.204	3.095	3.0
12	4.747	3.885	3.490	3.259	3.106	2.996	2.9
13	4.667	3.806	3.411	3.179	3.025	2.915	2.8
14	4.600	3.739	3.344	3.112	2.958	2.848	2.7
15	4.543	3.682	3.287	3.056	2.901	2.790	2.7
16	4.494	3.634	3.239	3.007	2.852	2.741	2.6
17	4.451	3.592	3.197	2.965	2.810	2.699	2.6
18	4.414	3.555	3.160	2.928	2.773	2.661	2.5
19	4.381	3.522	3.127	2.895	2.740	2.628	2.5
20	4.351	3.493	3.098	2.866	2.711	2.599	2.5
21	4.325	3.467	3.072	2.840	2.685	2.573	2.4
22	4.301	3.443	3.049	2.817	2.661	2.549	2.4
23	4.279	3.422	3.028	2.796	2.640	2.528	2.4
24	4.260	3.403	3.009	2.776	2.621	2.508	2.4
25	4.242	3.385	2.991	2.759	2.603	2.490	2.4
26	4.225	3.369	2.975	2.743	2.587	2.474	2.3
27	4.210	3.354	2.960	2.728	2.572	2.459	2.3
28	4.196	3.340	2.947	2.714	2.558	2.445	2.3
29	4.183	3.328	2.934	2.701	2.545	2.432	2.3
30	4.171	3.316	2.922	2.690	2.534	2.421	2.3

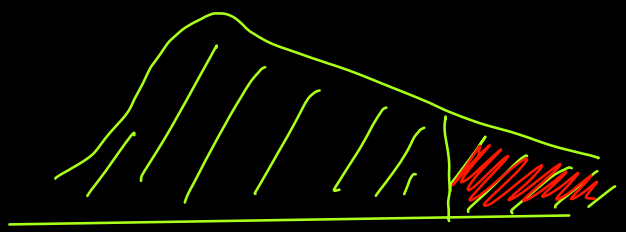
2 ← dof B/w

dof w.r.t

18

f-value = 3.555

If f-test statistic is greater than f-value we reject null value

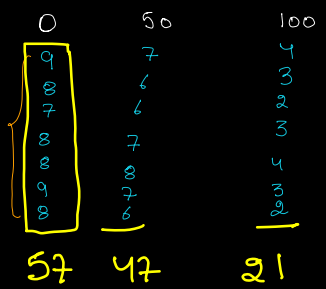


H.W.?

Why f-test graph is rightly skewed?

Calculate F-statistics

	SS	DoF	Mean Square SS/dof	F
Between	98.67	2	→ 49.33 ①	
Within	10.29	18	→ 0.54 ②	
Total	108.95	20		



$$SS_{Between} = \frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$$

a_i = all data point key level

T = total summation of all values

$$\frac{57^2 + 47^2 + 21^2}{7} \Rightarrow 842.71 \qquad n = 7$$

T = total summation of all values

$= (57 + 47 + 21)^2$ $\qquad N = 21$

$= 125$ $\Rightarrow 744.04$

$$SS_{Between} = \frac{57^2 + 47^2 + 21^2}{7} - \frac{125^2}{21}$$

$\Rightarrow 98.67$

$$SS_{within} = \Sigma y^2 - \frac{\Sigma (\Sigma a_i^2)^2}{n}$$

Σy^2 = all data point squared

$$= 853 - 842.71$$

$$= 10.29$$

$$F = \frac{\text{variance between sample}}{\text{variance within sample}} \Rightarrow \frac{①}{②}$$

$$\frac{49.33}{0.57} = \frac{86.54}{}$$

Critical Values of the F-Distribution $\alpha = 0.05$

Denom. d.f.	Numerator Degrees of Freedom 1	2	3	4	5	6	
1	161.448	199.500	215.707	224.583	230.162	233.986	236.
2	18.513	19.000	19.164	19.247	19.296	19.330	19.
3	10.128	9.552	9.277	9.117	9.013	8.941	8.
4	7.709	6.944	6.591	6.388	6.256	6.163	6.
5	6.608	5.786	5.409	5.192	5.050	4.950	4.
6	5.987	5.143	4.757	4.534	4.387	4.284	4.
7	5.591	4.737	4.347	4.120	3.972	3.866	3.
8	5.318	4.459	4.066	3.838	3.687	3.581	3.
9	5.117	4.256	3.863	3.633	3.482	3.374	3.
10	4.965	4.103	3.708	3.478	3.326	3.217	3.
11	4.844	3.982	3.587	3.357	3.204	3.095	3.
12	4.747	3.885	3.490	3.259	3.106	2.996	2.
13	4.667	3.806	3.411	3.179	3.025	2.915	2.
14	4.600	3.739	3.344	3.112	2.958	2.848	2.
15	4.543	3.682	3.287	3.056	2.901	2.790	2.
16	4.494	3.634	3.239	3.007	2.852	2.741	2.
17	4.451	3.592	3.197	2.965	2.810	2.699	2.
18	4.414	3.555	3.160	2.928	2.773	2.661	2.
19	4.381	3.522	3.127	2.895	2.740	2.628	2.
20	4.351	3.493	3.098	2.866	2.711	2.599	2.
21	4.325	3.467	3.072	2.840	2.685	2.573	2.
22	4.301	3.443	3.049	2.817	2.661	2.549	2.
23	4.279	3.422	3.028	2.796	2.640	2.528	2.
24	4.260	3.403	3.009	2.776	2.621	2.508	2.
25	4.242	3.385	2.991	2.759	2.603	2.490	2.
26	4.225	3.369	2.975	2.743	2.587	2.474	2.
27	4.210	3.354	2.960	2.728	2.572	2.459	2.
28	4.196	3.340	2.947	2.714	2.558	2.445	2.
29	4.183	3.328	2.934	2.701	2.545	2.432	2.
30	4.171	3.316	2.922	2.690	2.534	2.421	2.

f value from table
$\Rightarrow 3.55$

$86.54 > 3.55$

$\Rightarrow$ reject H_0

Conclusion :-

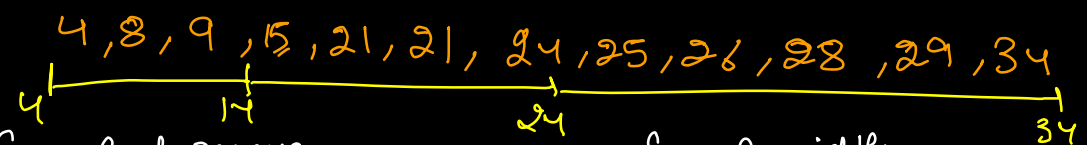
From f test value we reject null hypothesis, so there are difference b/w 3 condn (0, 50, 100) with 95% Interval

EDA :-

Data Scientist Plan :-

1. Load the data into the data mining environment.
2. Understand the data using graphical and non-graphical techniques
3. Document Insights
4. Pre process the data to suit your model
5. Divide into train & test
6. Quickly build the model with small data set and manually verify the error on test set for improving.

Changing numerical to categorical



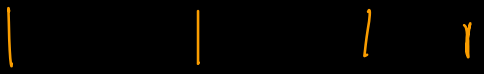
Equal frequency

Bin1 : 4,8,9,15

Bin2 = 21,21,24,25

Bin3 = 26,28,29,34

Equal width



$$\frac{34-4}{3} \quad \frac{(\text{max}-\text{min})}{(\text{no. of bins})}$$

$\Rightarrow 10$

$\Rightarrow \boxed{4+10}$

[4,14]

[14,24]

Categorical to numerical

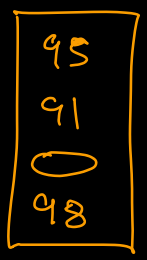
M $\longrightarrow$ 1

F $\longrightarrow$ 0

Missing Value

Why missing value :-

a) human error
b) Refuse to answer / fill form
c) Optional box
d) Technical reasons



Should we ignore or impute ?

⇒ Understand why missing value is present
⇒ Plot the graphs

↓↓

Delete the observ. or var.

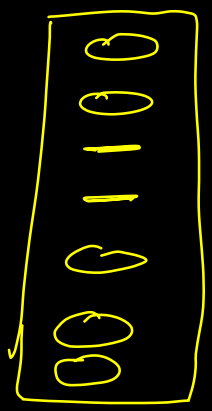
1] Drop obs 2] Drop var

1] Consider imputing values in var if missing value <10%

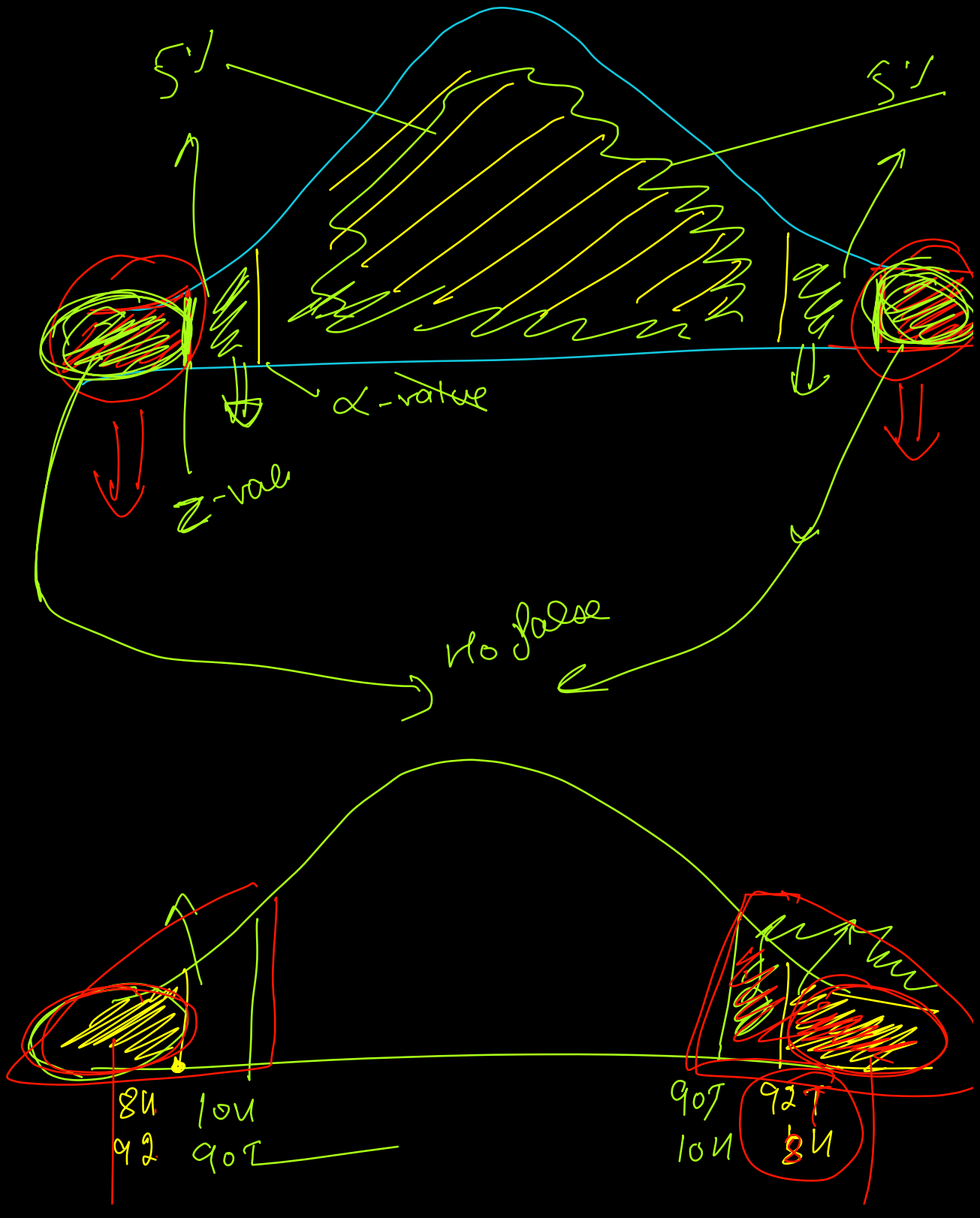
2] If more than 50%, drop var

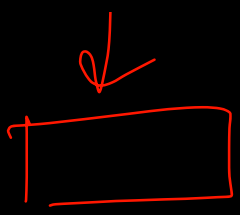
3] If we have huge data & say 10-20% are missing, drop observation

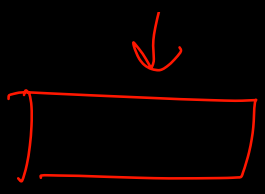
Marks } ⇒ dropping the whole variable



} ⇒ dropping observation


S.I
S.I
α-value
Z-val
No false
84
92
10M
90T
90T
92T
10M
8M





41 ∈ [0 , 100) ⇒ 41 ∈ [50 , 100)

$$\frac{100-0}{2} = 50$$

[0 , 50)

[50 , 100)

0 ≤ 41 < 49

50 ≤ 41 < 100