Till now

→ Z test, t test, Chi square test

Agenda

→ F test, ANOVA, Practical implementt

* Chi-square test (Practice Question H/w)

Q. A school principal would like to know which days of the week students are most likely to be absent. The principal expects the student will be absent equally during the 5-day school week. The principal selects a random sample of 100 teachers asking them which of day of the week they had the highest no of student absences.

	Monday	Tuesday	Wed	Thur	Friday
Observed	23	16	14	19	28
Expected	20	20	20	20	20

Test the hypothesis with 95% confidence based on above table.

→

* Accept the H_0.

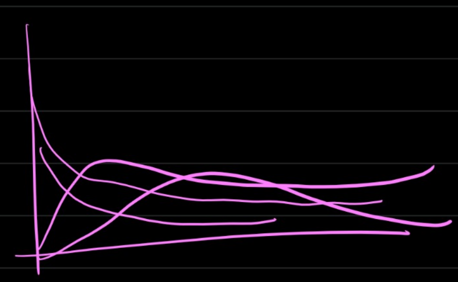
F-test

- follow F-distn
- Comparing variances.
- $X = \dfrac{S_1/d_1}{S_2/d_2}$

where S_1^2 is var of S_1, S_2^2 is var of S_2, d_1 & d_2 are degree of freedom for both sample

$$F_{statistics} = \frac{S_1^2}{S_2^2}$$

- right skewed
- Non-neg.
- it also depend on dof



F_{test} (Variance ratio test)

Q. The following data is about the number of bulbs produced daily by two workers A and B.

A	B
40	39
30	38
38	41
41	33
38	32
35	39
	40
	34

$\alpha = 0.05$, Can we consider based on data that worker B is more stable and efficient?

* Why not mean can be used for test? → mean is same for the both the sample. So we will compare variance.

⇒ ① $H_0 : S_1^2 = S_2^2$ $\quad H_A : S_1^2 \neq S_2^2$

② Ftest, one tail test, $\alpha = 0.05$

③ $F_{statistics} = \dfrac{S_1^2}{S_2^2}$

Worker A

x_1	$\bar{x}_1$	$(x_i - \bar{x}_1)^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
$\bar{x}_1 = 37$		$\sum(x_i - \bar{x}_1)^2 = 80$

$S_1^2 = \dfrac{80}{n-1} = \dfrac{80}{6-1} = \dfrac{80}{5} = 16$

Worker B

X_2	$\bar{X}_2$	$(X_2 - \bar{X}_2)^2$
39	37	4
38	37	1
41	37	16
33	37	16
32	37	25
39	37	4
40	37	9
34	37	9
$\bar{X}_2 = 37$		$\sum(x_2 - \bar{x}_2) = 84$

$S_2^2 = \dfrac{84}{n-1} = \dfrac{84}{8-1} = \dfrac{84}{7} = 12$

$F_{statistics} = \dfrac{16}{12} = 1.33$

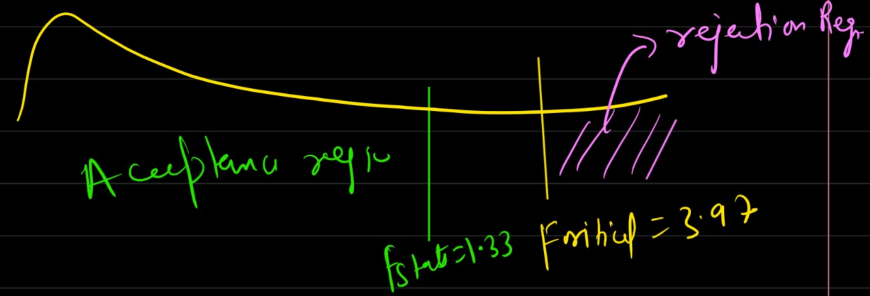
④ $F_{critical}$ $\alpha = 0.05$, $dof_1 = 5$, $dof_2 = 7$

$F_{cri} = 3.97$

⑤

1.33 < 3.97

We fail to reject the H_0



(Worker B is ~~not~~ Stable | effective as compared to A)

* Why can not we compare directly the absolute variance?

A	40	30	38	41	38	35		
B	39	38	41	33	32	39	40	34

→ Reason-1 No of days/hours are different

Reason-2 Sub data of both the sample can produce same variance. So we want to be statistical confident the even a slight change in A or B working hours will make no difference in variances.

Q. radius of Tomatoes from field A and field B are given :-

Group A : H_0 : 16 17 25 26 32 34 38 40 42

Group B : H_A : 14 16 24 28 32 35 37 42 43 45 47

Do you think if VAR A = VAR B is statistically significant.

→ Step-1 H_0 : VAR A = VAR B , H_A : Var A ≠ Var B

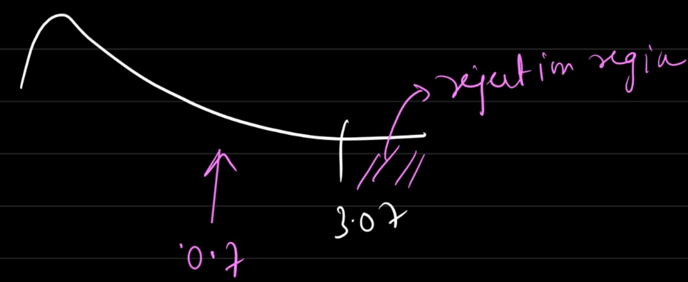
Step 2 $\alpha = 0.05$, F test

Step 3 $F_{\text{statistics}} = \dfrac{\text{VAR A}}{\text{VAR B}} = \dfrac{91.75}{129.8} = 0.7068$

Step 4 F_{critical} $df_1 = 9-1 = 8$, $df_2 = 11-1 = 10$

F_{critic} $\alpha = 0.05$, $df_1 = 8$, $df_2 = 10$

$F_{crit} = 3.07$



⑤ fail to reject the H_0

* ANOVA (Analysis of Variance)

Defn → ANOVA is a statistical method used to compare the means of 2 or more group.

* Factors / variables / criteria

factors / variables / Criteria → Medicine

0mg 10mg 100mg (Dosage) → levels

factors / variab / criteria ← Music

High sound, medium sound, low sound. → levels.

factor / variable / criteria ← Payment mode

GPay, PhonePe, IMPS, Paytm, NEFT, Cash → levels

* ## Assumptions of ANOVA

→ The population from which samples are drawn should be normally distribution.

→ The sample should be independen of each other (One way ANOVA)

→ Absence of outliers. → The sample should be random.

→ Homogeneity of variance → Homogeneity means that the variance among the group should be approx equal.

$$S_1 \quad S_2 \quad S_3$$

$$S_1^2 \approx S_2^2 \approx S_3^2$$

* ## Types of ANOVA

① **One way ANOVA** → One factor with atleast two levels and levels are independent.

eg. Medicine → (factor)

0mg 10mg 100mg (level)

eg. Music (factor)

high medium low] levels

② **Repeated measures ANOVA** → one factor with atleast two levels, but levels are dependent.

eg. No of hours studied

Day 1	Day 2	Day 3	Day 4
10 hours	5 hours	4	10

eg. Gym

Day 1	Day 2	Day 3	Day 4
6 hrs.	0	0.5	1

③ Factorial ANOVA :- Two or more factors (each of which with atleast 2 levels). Levels should be independent/dependent/both

↳ Two way ANOVA (deals with two factors)

Medicine (factors)

Gender (factor)	level ↓	0 mg	10 mg	100 mg (3 levels)
	male	2	3	3
		3	4	2
		4	5	1
		5	6	5
	female	1	1	4
		2	2	3
		1	3	2
		2	1	1

Medicine

Agegroup		0 mg	10 mg	100 mg
	Child			
	&			
	Adult			

music (Ist factor)

3rd	2nd factor	high	medium	low
male	1x	10	6	2
Female		9	5	3
male	x	8	1	1
femal		7	2	5
male	x1	5	3	6
faemal		4	4	4

ANOVA Test (partioning of variance)

$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ ---- } \mu_K$

H_A : Atleast one of the sample mean is not equal.

$\mu_1 = \mu_2 = \mu_3 = \text{ - - - } \neq \mu_K$

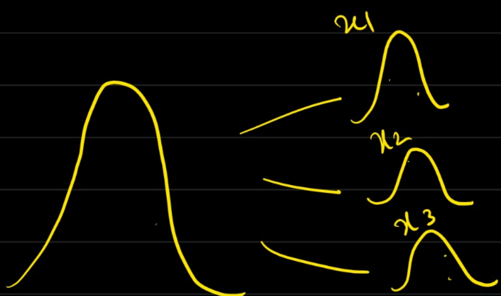
test statistics $= \dfrac{\text{Variance b/w samples}}{\text{variance within samples}}$

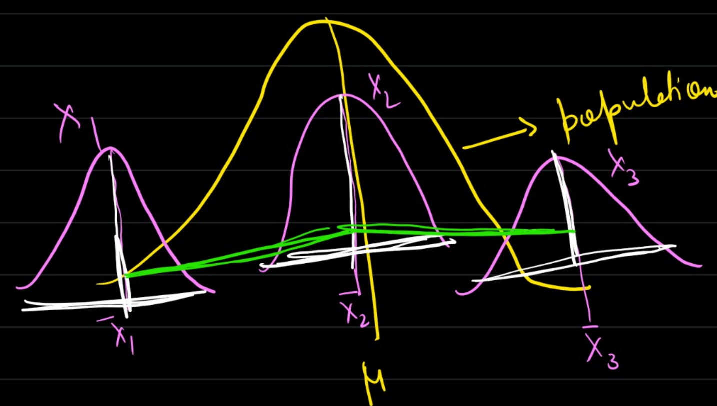
X

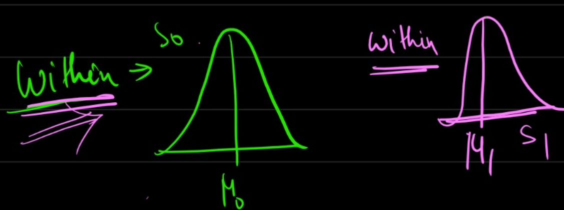
variance b/w sample (inter group)

x_1	x_2	x_3
1	5	9
2	6	10
3	7	11
4	8	12

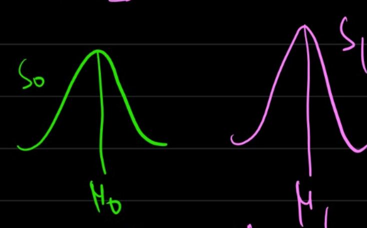
variance within sample (intra group)







between → two samples



$S_0 \approx \mu_1$

$S_1 \approx \mu_0$

One way ANOVA

Q. There are three dosage of a medicine given to three sample of patients. They rate headache is reduced (1–10).

Are there differences in the three condition? $\alpha = 0.05$

0mg	10mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

① $H_0 : \mu_{0mg} = \mu_{10mg} = \mu_{100mg}$.
H_A : not all are equal.

② $\alpha = 0.05$, one tail test (ANOVA follow F dist^n)

③ Calculate F statistics

$$F\text{statistics} = \frac{\text{Variance b/w Sample}}{\text{Variance within the Sample.}}$$

$N = 21$ (total no of element)

$n = 7$

$\frac{\text{Sum sq}}{\text{S.S}} = \text{mean sq}$

	Sum of squares (SS)	df	Mean Square (MS) (SS/df)
btw the sample			
within sample			
Total			

0mg	10mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

* Sum of Squares b/w the sample = SS_{btw} = $\frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$

- 0mg : $9+8+7+8+8+9+8 = 57 \Rightarrow a_1$
- 10mg : $7+6+6+7+8+7+6 = 47 \Rightarrow a_2$
- 100mg : $4+3+2+3+4+3+2 = 21 \Rightarrow a_3$

$T = a_1 + a_2 + a_3$

$$\frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N} \Rightarrow \frac{a_1^2 + a_2^2 + a_3^2}{n} - \frac{(a_1 + a_2 + a_3)^2}{N}$$

$$\Rightarrow \frac{57^2 + 47^2 + 21^2}{7} - \frac{(57+47+21)^2}{21}$$

$$\Rightarrow 98.67.$$

* Sum of Squares within the group

$$SS_{within} = \underline{\sum y^2} - \sum \frac{\left(\sum a_i\right)^2}{n}$$

0mg	10mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

$$\sum y^2 = 9^2+8^2+7^2 \dots 8^2 + 7^2 + 6^2 \dots 6^2 + 4^2 + 3^2 + \dots 2^2$$
$$= 853$$

$$SS_{within} = 853 - \left(\frac{57^2 + 47^2 + 21^2}{7}\right) = 10.29.$$

	Sum of squares (SS)	df	Mean Square (MS) (SS/df)
btw the sample	98.67	2	$98.67/2 = 49.34$
within sample	10.29	18	$10.29/18 = 0.57$
Total	108.96	total = 20	

dof b/w sample = # sample − 1

dof within sample = No of elements − Total no of sample

$= 21 - 3 \Rightarrow 18$

total Dof = N − 1
= 21 − 1
= 20

$$F_{stats} = \frac{MS \text{ b/w sample}}{MS \text{ with the sample}} \left(\frac{\text{Var b/w sample}}{\text{Var within sample}}\right) = \frac{49.34}{0.57} = 86.56$$

(4) Fcritical for $\alpha = 0.05$

df between $= 3-1 = 2$ ↙ df_1

df within $= 21 - 3 = 18$ $\quad df_2$

$F_{critical} = 3.55$

(5)

$F_{critical} = 3.55$ $\qquad$ $F_{stats} = 86.56$

$86.56 > 3.55$, Reject the H_0

H/W Q. levels of stress for employees of a company is given below :-

Sample-1	Sample 2	Sample 3
2	10	16
3	8	13
7	7	14
2	5	13
6	10	15
(Normal stress)	(anouncement of layoff)	(layoff is done)

Mean stress is different or not at three stages.

$\alpha = 0.05\%$

Ans - Reject the H_0

$F_{stats} = 4.5$ (Not sure)