for dof = 29 the value in table t is 2.045

if t test is less than −2.045 or greated than +2.045, reject the null hypothesis.

(6) calculate test statistics

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{30}} = \frac{40}{3.65} = 10.96$$

$t = 10.96$

Since

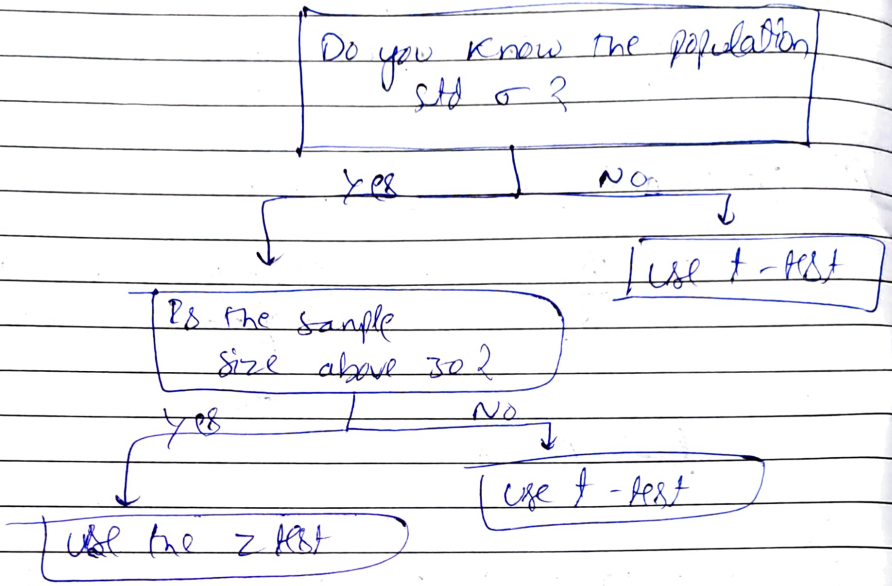
$t = 10.96 > 2.045$ { Reject the null hypothesis }

Conclusion : Medication used has affected the intelligence

Medication has increased the intelligence

Date
Page

When to use T-test vs z-test?



Type 1 and Type 2 errors

Reality: Null hypothesis true or Null hypothesis is false

Decision: Null hypothesis is true or Null hypothesis is false

Outcome 1: we reject the Null hypothesis when in reality it is false → Good.

outcome 2: we reject the Null hypothesis when in reality it is true → Type 1 error

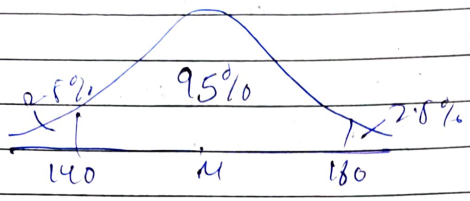
outcome 3: we retain the Null hypothesis when in reality it is false → Type 2 error

outcome 4: we retain the Null hypothesis when in reality it is true → Good.

Confidence Interval and Margin of Error

$\mu = 160$

C.I = 95%



140 ⟷ 160

point estimate : A value of any statistics that estimate the value of an unknown population parameter is called point estimate.

$\bar{x} \longrightarrow \mu$

$\bar{x} = 2.95 \qquad \mu = 3$

Confidence Interval

we construct a confidence interval to help estimate that the actual value of the unknown population mean is

point estimate ± Margin of error

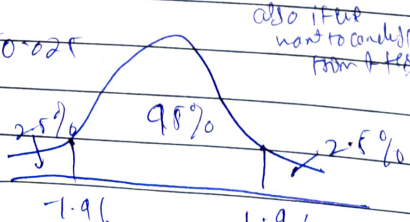
z test

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad \alpha = 0.05$$

Here we can use t test under also if we want to conclude from t test

we can use t test if t test is given

$z_{0.05/2} \Rightarrow z_{0.025}$

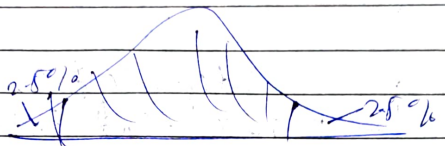


Date
Page

① On the verbal section of CAT Exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. construct 95% C.I about the mean?

$\bar{x} = 520$ $\sigma = 100$ $n = 25$ C.I = 0.95

$\alpha = 0.05$

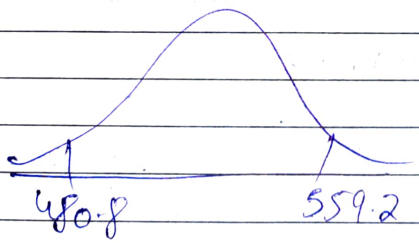


$1 - 0.25 = 0.9750$

$\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $\pm 1.96 \Rightarrow$ z table

Lower CI $= 520 - (1.96) * \frac{100}{\sqrt{25}} = 480.8$

Higher CI $= 520 + (1.96) * \frac{100}{\sqrt{25}} = 559.2$



I am 95% confident that the mean CAT score lies b/w 480.8 and 559.2.

Date
Page

Bayes statistics (Bayes theorem)

Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes theorem.

Bayes theorem

Probability → independent events
→ dependent events

(1) Independent events

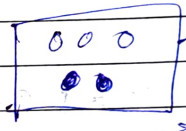
e.g. Rolling a die
{1, 2, 3, 4, 5, 6}

$P(1) = \frac{1}{6}$ $P(2) = \frac{1}{6}$ ...

Tossing a coin

$P(H) = 0.5$ $P(T) = 0.5$

(2) Dependent event



$P(R) = \frac{2}{5}$

yellow

$P(Y) = \frac{3}{4}$

$P(R \text{ and } Y) = P(R) \times P(Y/R)$

conditional probability

$\Rightarrow \frac{2}{5} \times \frac{3}{4} = \frac{6}{20}$

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

$P(A) \times P(B/A) = P(B) \times P(A/B)$

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

→ Bayes's thm

$$Pr(A|B) = \frac{Pr(A) \times Pr(B|A)}{Pr(B)}$$

A, B = events

already occurred

$Pr(A|B)$ = probability of A given B is true

$Pr(B|A)$ = " " B " A is true

$Pr(A)$, $Pr(B)$ = Independent probability of A and B

DATASET :

Size of House	No. of Rooms	location	Price
	← Independent		← o/p/dependent
x_1	x_2	x_3	y

$$Pr(y|x_1, x_2, x_3) = \frac{Pr(y) \times Pr(x_1, x_2, x_3|y)}{Pr(x_1, x_2, x_3)}$$

↓

Bayes Theorem

Date
Page

CHI SQUARE TEST

The chi square test for Goodness of fit test claims about population proportions (categorical)

It is a non parametric test that is performed on categorical (ordinal and nominal) data.

There is a population of Make who likes different color bikes

	Theory	Sample
Yellow Bike	1/3	22
Red Bike	1/3	17
Orange Bike	1/3	59

(*) Goodness of fit test

Does sample support the theory

Observed categorical distribution

Theory categorical distribution

Goodness of fit data

In a school class of 75 students, 11 are left handed. Does this class fit the theory that 10% of people are left handed

	O	E → expected value based on the theory	
Left handed	11	9	2
Right handed	64	66	
Total	75	75	

Now we need to check the observed value fit for the hypothesis

Chi Square for Goodness of Fit

In 2010 census of the city, the weight of the individuals in a small city were found to be the following

<50kg	50-75	>75
20%	30%	50%

In a 2020, weight of $n=500$ individuals were sampled. Below are the results

<50	50-75	>75
140	180	200

Using $\alpha = 0.05$, would you conclude the population different of weights has changed in the last 100 ye

2010 expected	<50kg	50-75	>75
	20%	30%	50%

2020 n=500 observed	<50	50-75	>75
	140	180	200

expected	<50	50-75	>75
	0.2×500 = 100	0.3×500 = 150	0.5×500 = 250

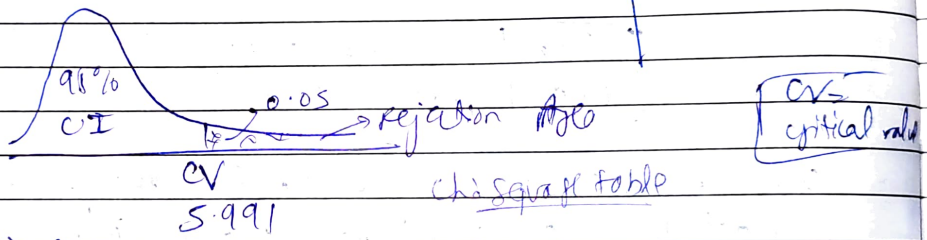
Null Hypothesis: H_0: The data meets the expectation
Alternate Hyp: H_1: The data does not meet the expectation

④ $\alpha = 0.05$ CI = 95%

⑧ Degree of freedom

$df = K - 1 = 3 - 1 = 2$

⑨ Decision boundary



CV = critical value

Chi square table

notation for chi square

If χ^2 is greater than 5.99, Reject H_0.
else
we fail to reject the null hypothesis

⑤ calculate chi square test statistics

$$\chi^2 = \frac{\Sigma (O-E)^2}{E}$$

2020

n=506	< 50	50–75	>75
observed	140	160	206

expected	< 50	50–75	>75
	=100	=150	=250

$$\Rightarrow \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$\Rightarrow \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$\Rightarrow 16 + 0.66 + 10$$

$$\Rightarrow 26.66$$

$$\chi^2 = 26.66$$

If χ^2 is greater than 5.99, reject H_0
else
we fail to reject the null hypothesis

$26.66 > 5.99$ Reject H_0

The weight of 2020 population are different than those expected in the 2010 population

Specifically used for hypothesis testing to analyse or compare the ratios of variances between two different groups

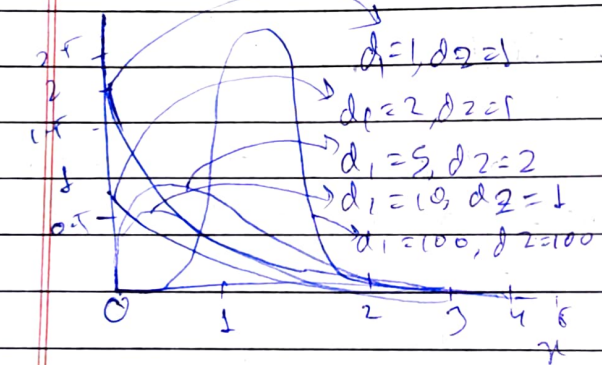
F distribution

In probability theory and statistics, the F-distribution or F-ratio, also known as Snedecor's F dist. or the Fisher–Snedecor dist. (after Ronald Fisher and George W. Snedecor) is a continuous probability distribution that arises frequently as the null dist. of a test statistic, most notably in the analysis of variance (ANOVA) and other F-tests.

used in ANOVA

Fisher–Snedecor prob. density fun^n

using F-test we compare two groups, or more groups so two d.o.f



parameters : $d_1, d_2 > 0 \Rightarrow$ Def

support parameter $x \in (0, +\infty)$

$$Pdf = f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

samples of two groups of data

Beta function

$$B(m,n) = \frac{(m-1)!(n-1)!}{(m+n-1)!}$$

$$\Rightarrow \frac{m+n}{mn} \Big/ \binom{m+n}{m}$$

F distribution with d_1 and d_2 degree of freedom distribution of of a Random variable $\Rightarrow$

Chi-square is Right skewed dist.

$\chi = \frac{S_1/d_1}{S_2/d_2}$

S_1 & S_2 ⇒ Independent random variables with chi square dist.

d_1 & d_2 ⇒ Respective degrees of freedom

f-test [Variance Ratio test]

f-test [variance Ratio test]

(1) The following data shows the no. of bulbs produced daily for some days by 2 workers A and B

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

Can we conclude based on the data worker B is more stable and efficient

$\alpha = 0.05$

samples of two groups of data

⇒ variance

(1) Null Hypothesis $H_0: \sigma_1^2 = \sigma_2^2$

$H_1: \sigma_1^2 \neq \sigma_2^2$

(2) calculation of variance :

$$S_1^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Worker A			Worker B		
X_1	$\bar{X}_1$	$(X_1 - \bar{X}_1)^2$	X_2	$\bar{X}_2$	$(X_2 - \bar{X}_2)^2$
40	37	9	39	37	4
30	37	49	38	37	1
38	37	1	41	37	16
41	37	36	33	37	16
38	37	1	32	37	25
35	37	4	39	37	4
			40	37	9
			34	37	9
$\bar{X}_1 = 37$		$\Sigma(x_1 - \bar{x}_1)^2 = 80$	$\bar{X}_2 = 37$		$\Sigma(x_2 - \bar{x}_2)^2$
					$\Rightarrow 84$

$$S_1^2 = \frac{80}{n-1} = \frac{80}{5} = 13$$

$$S_2^2 = \frac{84}{8-1}$$

$$\Rightarrow \frac{84}{7} = 12$$

Calculation of variance Ratio Test (F test)

$$F = \frac{S_1^2}{S_2^2} = \frac{13}{12} = 1.33$$

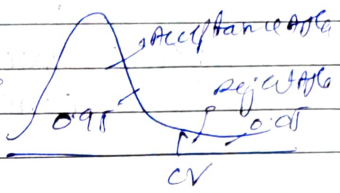
F test also follows right skewed dist.

Date
Page

(x) Decision Rule;

$df_1 = 6-1 = 5$
$df_2 = 8-1 = 7$
$\alpha = 0.05$

F table @



3.9715

Decision
F test is greater than 3.9715, Reject the null hypothesis

1.3 < 3.9715, we fail to Reject the null hypothesis

Conclusion;

(xi) worker B is not more stable or efficient when compared to worker A

Analysis of variance (ANOVA)

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

ANOVA — independent / dependent

(1) Factors (variable)
(2) levels

} components

Eg: factor = medicines (Factor)

levels = 5mg, 10mg, 20mg [Dosage]

Mode of payment → (factor)

GPAY, PHONEPE, IMPS, NEFT [levels]

Date

Assumptions of ANOVA

CLT (Central Limit Theorem)

(1) Normality of sampling Distribution of Mean
The distribution of sample mean is normally distributed
(Gaussian dist.)

(2) Absence of outliers
outlying score need to be removed from the dataset. [outliers need to be removed?]

(3) Homogenity of variance
population variance in different levels of each independent variable are equal.
$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$ [one variable do not impact the other variables]

(4) Samples are independent and random
[randomly selected].

Types of ANOVA (3 Types)

① one way ANOVA : one factor with atleast 2 levels, these levels are independent.

eg: Doctor wants to pretest a new medication to decrease headache
They split the participants in 3 conditions (10mg, 20mg, 30mg)
Doctor ask the participants to rate the headache (1-10)

Medication → factor

10mg	20mg	30mg	→ levels are independent.
5	7	2	
3	4	6	
—	—	—	
—	—	—	

② Repeated Measures ANOVA : one factor atleast 2 levels, levels are dependent.

Running → factor

Levels → are dependent.	Day 1	Day 2	Day 3
	8	5	4
	7	4	9
	—	—	—

Date
Page

(15) Factorial ANOVA of Two or more factors each of which with atleast 2 levels. Levels can be independent and dependent

Running → factor

Levels ↓	Day 1	Day 2	Day 3	levels
Gender ↓ factor: male	8	5	4	
female	9	4	3	
	2	4	6	
	7	8	3	

Date
Page

Hypothesis testing in ANOVA (partitioning of variance in the ANOVA)

Based on how many samples we have

Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 \cdots = \mu_K$

Alternate hypothesis H_1: Atleast one of the sample mean is not equal

$$\mu_1 \neq \mu_2 \neq \mu_3 \cdots \neq \mu_K$$

We cannot write like this

This means all the means are not equal
One mean is not equal to other
One mean is not equal to some other diff. mean

Test statistics

We use F-test

$$F = \frac{\text{variance between samples}}{\text{variance within samples}}$$

H_0: $\bar{X}_1 = \bar{X}_2 = \bar{X}_3$

H_1: Atleast one sample mean is not equal.

σ^2

variance between samples

variance within samples

X_1	X_2	X_3
1	6	5
2	7	6
4	5	3
5	2	2
3	1	4

$\bar{X}_1 = 3$ $\quad \bar{X}_2 = 19/5$ $\quad \bar{X}_3 = 4$

variance is spread

One way ANOVA
one factor with atleast 2 levels, levels are independent.

(1) Doctors want to test a new medication which reduces headache. They splits the participant into 3 condition [15mg, 30mg, 45mg]. later on the doctors ask the patient to rate the headache btw [1-10] Are there any differences btw the 3 conditions using alpha = 0.052

15mg	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

(2) Define Null and Alternate hypothesis ?

H_0: $\mu_{15} = \mu_{30} = \mu_{45}$
H_1: not all μ are equal

(2) Significance $\alpha = 0.05$ CI = 0.95

(3) Calculate Degree of freedom

$N = 21$ (Total sample size) $a = 3$ (no. of categories) $n = 7$ (sample size)

→ df blw samples

$df_{between} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$

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

df_1 df_2

(2, 18)

↓

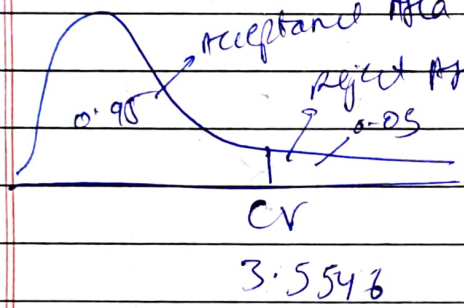
F table

$\alpha = 0.05$

↓

To find critical value

(4) Decision Boundary :



For $df_1 = 2$

$df_2 = 18$

The value in F table is 3.5546

Decision Rule

If F is greater than 3.5546, reject the null hypothesis

(5) calculate F test statistics

F = variance blw Samples / variance within Samples

	SS	df	MS	F
Between	98.87			
within	10.29			
Total	108.96			

(1) SS between $= \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}$

15mg = 4+8+7+8+8+9+8 = 57
30mg = 7+8+6+7+8+7+6 = 47
45mg = 4+3+2+3+4+3+2 = 21

$= \frac{57^2 + 47^2 + 21^2}{7} - \frac{[57^2 + 47^2 + 21^2]}{21}$

$= 98.67$

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

$\Sigma y^2 = 9^2 + 8^2 + 7^2 + 8^2 + 8^2 + \dots$

$= 853$

$\Rightarrow 853 - \frac{[57^2 + 47^2 + 21^2]}{7}$

$\Rightarrow$ 10.29

SS → Sum of Squares
df →
MS → Mean Square.

$MS = \frac{SS}{df}$

	SS	df	MS	F
Between	98.87	2	49.34	
within	10.29	18	0.54	
Total	108.96	20		

$F\ test = \frac{MS\ between}{MS\ within}$ F = variance B/w Samples / variance within Samples

$F = \frac{49.34}{0.54} = 86.56$

If F is greater than 3.5546, Reject the H_0

86.56 > 3.5546 Reject the H_0

∴ there are differences.