Student's t - distribution / t distribution

Z test → Z score

→ σ (population standard deviation) is already given

→ Sample Size $\geq$ 30

But in majority of the case σ (pop std) will not be known.

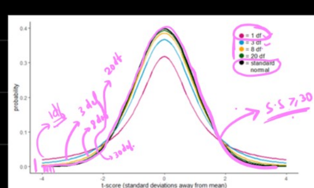
Then what analysis you wid do?

* Whenever S.S < 30, and σ (pop std) not given then use t - test.

$\Downarrow$

t distribution.

S.S < 30 ⟶ t distribution.



df → degree of freedom

S.S < 30

df = S.S − 1

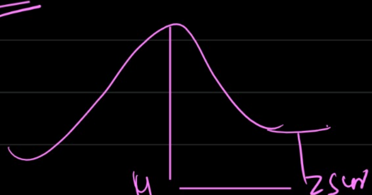
S.S = 28

df = 28 − 1 = 27

$$Z \text{ statistics} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$\Downarrow$

Z score



t distribution

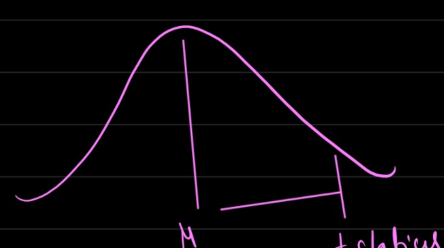
$\Downarrow$

$$t \text{ statistics} = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

$\Downarrow$

t score

S: Sample's Standard deviation.



$\Downarrow$

t table ⟵ t score

$dof = n-1$

Unconstrained

2 8 10 10 20 = 20

(avg)

Constrained

$5-1=4$

$(n-1)$

$$\frac{2+8+10+10+x}{5} = 20$$

$40+x = 20 \times 5$

$x = 100-40$

$= 60$

12 8 10

avg.

$2-1=1$

3 people

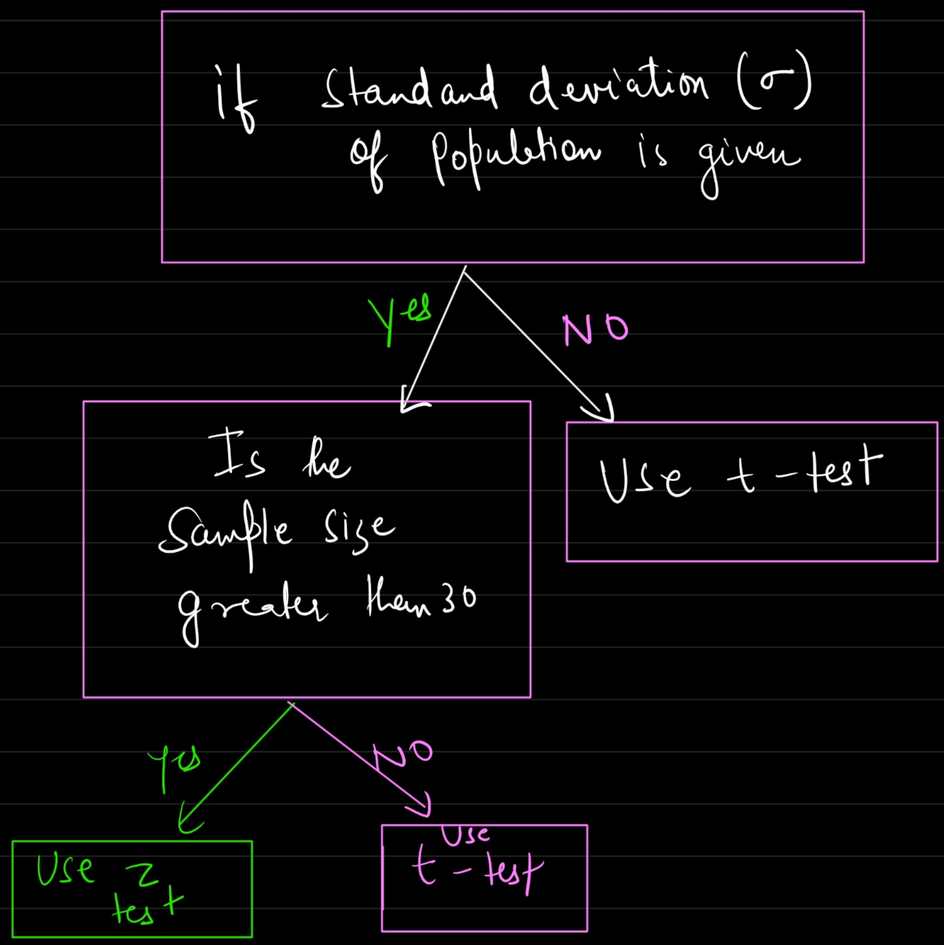
$\boxed{X}$ $\boxed{\ }$ $\boxed{X}$

degree of freedom

$3-1=2$

$(n-1)$

T test Vs Z test.



* If Population Standard deviation is given and Sample size is greater then 30, then use Z test, otherwise t-test.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Confidence interval and Margin of error

* Example → what is the potential score?

→ ~~80 marks~~

75 − 85 → interval

80 ± 5

Analysis → Sample → Avg ht of Employees of a company → Population

Sample → $\bar{X}$ → Point estimate

* estimate → An estimate of a population parameter is an approximation depending solely on sample information.

* A point estimate is a single no.

eg. 180 cm (ht)

* Confidence interval is an interval.

eg. Point Estimate ± Error

180 ± 20

Confidence Interval = 160 — 200

* Point estimate is located exactly in the middle of confidence interval

160 — 180 → Point estimate — 200

Confidence interval

Confidence Interval (CI) = Point Estimate ± Margin of error.

eg: People visiting this restaurent spends 1000 Rs. on an avg.

↑ point estimate.

↳ It is much safer to say that people spend 800 – 1200 Rs

↑ confidence interval

↓ more accurate representation of reality

⇓

Why? → You can not be 100% confident unless you go through entire population.

$$C.I = \text{point estimate} \pm \text{margin of error.}$$

Z test (two tail test) ⟶ $CI = \bar{x} \pm Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

Z test (one tail test) ⟶ $CI = \bar{x} \pm \boxed{Z_{\alpha}\dfrac{\sigma}{\sqrt{n}}}$

margin of error.

t-test $= \bar{x} \pm t_{\alpha/2}\dfrac{s}{\sqrt{n}}$

↳ s – Sample's standard dev.

where σ is population's standard deviation

n – sample size

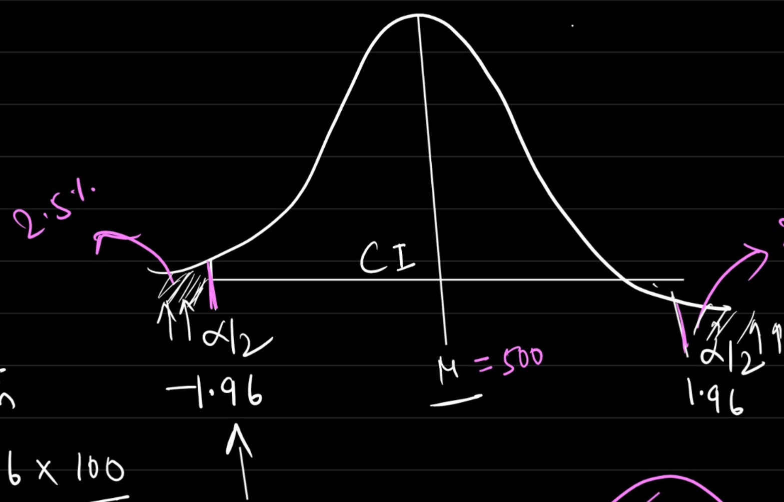
$\bar{x}$ – sample mean.

$Z_{\alpha/2}$ → Z score corresponding to given $\alpha/2$

α – level of significance

* CI → what value the sample statistics will take can be known through confidence interval.

Q. In an exam the standard deviation of marks is 100. A sample of 36 students has a mean of 500 marks.
Construct a 95% Confidence Interval about the mean?

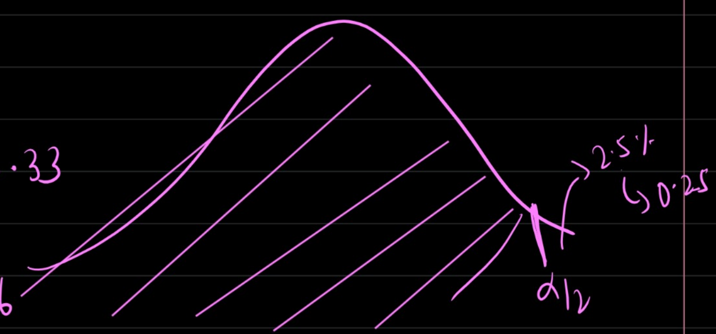


$$C.I = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$500 \pm 1.96 \times \frac{100}{\sqrt{36}}$$

$$\text{Lower CI} = 500 - 1.96 \times \frac{100}{\sqrt{36}} = 467.33$$

$$\text{Upper CI} = 500 + 1.96 \times \frac{100}{\sqrt{36}} = 532.6$$



$\Rightarrow 1 - 0.25$

$\Rightarrow 0.9750$

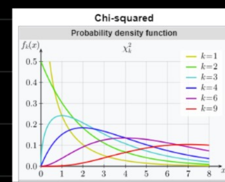
Z table → Corresponding to 0.9750 what is Z score.

→ The Z score Corresponding to 0.9750 is 1.96

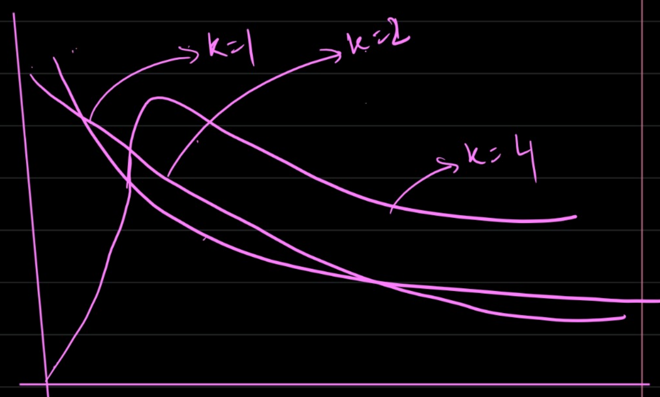
→ I am 95% Confident that the mean score in the exam lies between 467.3 and 532.6

Chi-Square distribution and Chi-square test

* The chi-square distribution is a probability distribution that describes the distribution of a sum of squares of K random variables.



→ degree of freedom (k) = n-1
→ Chi-square distribution shape is determined by "k"
→ non-negative distribution
→ right skewed distribution



* Chi-square test (χ^2 test)
  → chi-square distribution.
  → Goodness of fit test — Used to compare observed and expected categorical data.
  → Test of independence — To determine the relationship b/w two categorical variable.
  → It test the claims about population proportions.
  categorical
  → Non-parametric test — No assumption about the population.

Types of car	(Expected) Theory	(Observed) Sample
Sports car	1/3	22
SUV	1/3	17
Sedan	1/3	59

→ observed categorical distribution
⇓
Using this Observed Sample dist^n, you have to verify it

theoretical categorical distribution

the theoretical distribution is true or not

$\longrightarrow$ goodness of fit

ex. In a class of 75 students, 11 are left handed.

theory $\rightarrow$ 12% of the people are left handed.

12% of 75

$\frac{12}{100} \times 75 = 9$

	O	E
left handed	11	9
right handed	64	66

$$\chi^2_{\text{statistics}} = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$\underbrace{\frac{(11-9)^2}{9}}_{\text{left}} + \underbrace{\frac{(64-66)^2}{66}}_{\text{right hand}}$$

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

Probability and Baye's theorem.

Probability → Share of Success | Total No of Possible outcomes.

ex. if you toss a coin, what is the probability that you will get a head?

→ $P(H) = \frac{1}{2} = 0.5$

ex. dice is rolled. What is probability that the outcome is an even no?

→ 1, 2, 3, 4, 5, 6

$P(\text{Even no}) = 3/6$

* Probability rules :-

① For any event A → $0 \leq P(A) \leq 1$

② The sum of all probabilities of all possible outcomes is 1.

Rule of subtraction.

$P(H)$ or $P(T)$

$P(H) + P(T) = 1$

$P(T) = 1 - P(H)$

③ Complement rule

$P(\text{not } A) = 1 - P(A)$

④ General addition rule

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

⑤ The multiplication rule

Events → Independent events, Dependent events

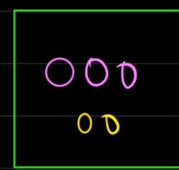
* Independent events

→ Tossing a coin

→ Throwing a dice

$Pr(1) = 1/6$

* Dependent Event



Scen-1 $P(Y) = \frac{2}{5}$ $\xrightarrow{\text{one yellow ball is taken out}}$ $P(Y) = \frac{1}{4}$

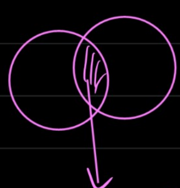
Scen-2 $P(R) = \frac{3}{5}$ $\xrightarrow{\text{one yellow ball is taken out}}$ $P(R) = 3/4$

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

$$3/5 \times 3/4$$

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

$P(B/A)$ = probability of event B when A has already occurred



A and B
or
B and A } same

$$P(A \text{ and } B) = P(A) * P(B/A) \quad \text{— a}$$

$$\parallel$$

$$P(B \text{ and } A) = P(B) * P(A/B) \quad \text{— b}$$

equating eqn ⓐ & ⓑ

$$P(A) * P(B/A) = P(B) * P(A/B)$$

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

Baye's theorem.

$$\Downarrow$$

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

$P(A/B)$ — Prob of event A given B has already occurred

$P(B/A)$ = Prob of event B, given A has already occurred

$P(A), P(B) \rightarrow$ Independent probability of A and B

Q. 10% of patients in a clinic have liver diseases. Five percent of the clinical patients are alcoholics. Among these patients diagnosed with liver disease 7% are alcoholics.

What is prob of patients having liver disease given that he is an alcoholics?

$\rightarrow$

$P(A)$ = Prob of having liver disease $= 0.10$

$P(B)$ = Prob of alcoholism $= 0.05$

$P(B|A) = 0.07$

$$P(A|B) = \frac{P(B/A) \cdot P(A)}{P(B)} = \frac{0.07 \times 0.10}{0.05} = 0.14 = 14\%$$

Use of Baye's theorem

(Bayesian statistics) $\rightarrow$ data analysis & Parameter estimation based on baye's theorem

$\rightarrow$ Naive Bayes ML model.

x_1 # room	x_2 Area	x_3 loc	y Price of House
—	—	—	—
—	—	—	—
—	—	—	(?)

$$P(\overset{A}{y}/\overset{B}{x_1, x_2 x_3}) = \frac{\overset{P(A)}{P(y)} \cdot \overset{P(B|A)}{p(x_1 x_2 x_3/y)}}{\underset{P(B)}{P(x_1 x_2 x_3)}}$$

$\rightarrow$ Baye's theorem.

Chi-square test | goodness of fit test

Q. 12% of people are left handed. To verify this theory, you took a sample of 75 students, 11 are left handed.

→ Level of Significance 5%.

	Observed	Expected
left	11	9
Right	64	66

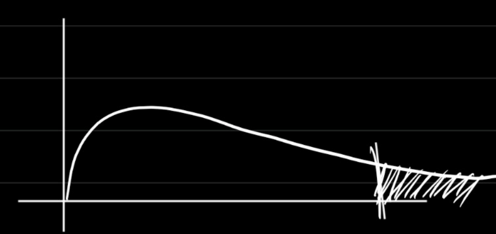
12% of 75

$\frac{12}{100} \times 75 = 9$

Step-1. $H_0 : \mu = 12\%$, $H_A : \mu_A \neq 12\%$

Step-2. $\alpha = 5\%$

Step-3 → $\chi^2 \text{statistics} = \sum \frac{(O-E)^2}{E}$



$$\chi^2_{\text{statistics}} = \frac{(11-9)^2}{9} + \frac{(64-66)^2}{66} = \frac{2^2}{9} + \frac{2^2}{66} = 0.505$$

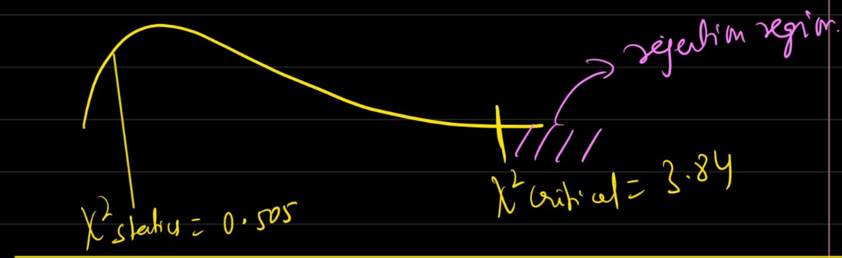
Step-4

χ^2_{critical} for $\alpha = 0.05$

dof = No of groups − 1
= 2 − 1
= 1

$\chi^2_{\text{critical}} = 3.84$

Step-5



if $\chi^2_{\text{statistics}} > \chi^2_{\text{critical}}$
reject the H_0

* $0.505 < 3.84$
  we fail to reject H_0.

* 12% of people are left handed with 95% Confidence.

Q. In 2010 census of the city, the weight of People in a city were found to be following :-

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

In 2020, weight of 500 people were sampled.

<50	50-75	>75
140	160	200

Using $\alpha = 0.05$, Can you conclude the population difference of weight has changed in last 10 years or not?

→ 2010 →

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

2020 $n = 500$ observed =

<50	50-75	>75
140	160	200

Expectation =

<50	50-75	>75
$0.2 \times 500 = 100$	$0.3 \times 500 = 150$	$0.5 \times 500 = 250$

Step-1 → H_0 : The data is as per expectation.
H_A : The data is not as per expectation

Step-2 - $\alpha = 0.05$.

Step-3 - χ^2 statistics $= \sum \frac{(O-E)^2}{E} = \frac{(140-100)^2}{100} + \frac{(160-150)^2}{160} + \frac{(200-250)^2}{250}$

$$= \frac{1600}{100} + \frac{100}{160} + \frac{2500}{2500}$$

$$= 16 + 0.66 + 10 = 26.66$$

Step-4 χ^2 critical $\alpha = 0.05$,

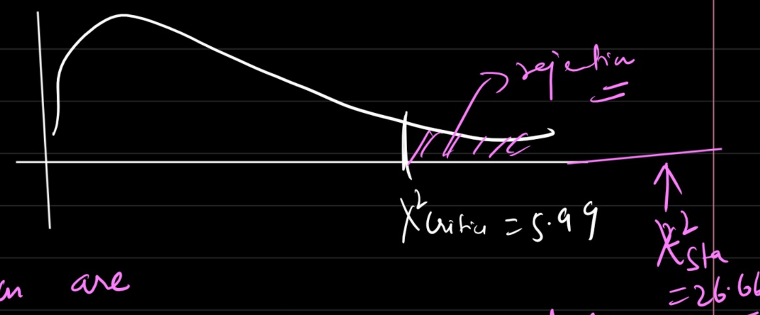
$dof = 3-1 = 2$

$\chi^2_{crit\, 0.05,\, dof\, 2} = 5.99$

Step 5 if $\chi^2_{stats} > \chi^2_{critical}$ — Reject the H_0

$26.66 > 5.99$

reject H_0.



Conclusion — The weight of 2020 population are different those expected in 2010 population.