Name ⇒ Dheeraj Chaudhary
Roll ⇒ 17BC5009

ASSIGNMENT ⇒ Hypothesis testing

Testing of Hypothesis

* ⇒ Z-test (Large sample test) $n \geq 30$

One Sample Test : let $x_1, x_2, x_3 \ldots x_n$ be random obs. of sample 'x' is drawn from a normal population with mean 'μ' and variance is σ^2, i.e, $N(\mu, \sigma^2)$, resp.

Null hypoth. $\mu = \mu_0$

Alternate hypo. : $\mu \neq \mu_0$
$\mu > \mu_0$
$\mu < \mu_0$

l.o.s, $\alpha = 1\%$ or 5% or 10%

Test statistic : let $X \sim N(\mu, \sigma^2)$, z-statistic defined as

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$, Pop^n var σ^2 is known

and $$Z = \frac{\bar{X} - \mu}{(S/\sqrt{n})} \sim N(0,1)$$ under H_0 if pop^n var is unknown.

Conclusion :- If Z-calculated value is less than Z-tabulated value Z-cal < Z-tab than accept H_0 otherwise reject H_0.

Problem ⇒ 1 ⇒ A sample of 900 members is found to have a mean 3.5 cm. Can it be reasonably regarded as sample sample from a large popl^n whose mean is 3.8 and std. dev. is 2.4 cm? $\alpha = 5\%$.

1) $\mu = 3.38$ and $\bar{x} = 3.5$ $n = 900$

$\sigma = 2.4$

Now, $H_0 : \mu = \mu_0$

$\mu = 3.38$

$H_1 : \mu \neq \mu_0$ & $\alpha = 5\%$ (l.o.s)

$Z_{cal} = 1.5$ & $Z_{tab} = 1.96$

(since $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$)

$\Rightarrow \dfrac{3.5 - 3.38}{2.4/30} = \dfrac{3}{2} = 1.5$

Since $Z_{cal} < Z_{tab} \Rightarrow$ accept H_0:

Given sample is drawn from the normal popl^n.

Two Sample Z-test :- let x be a random sample $x_1, x_2 .. x_n$ drawn from a normal pol^n i.e, $N(\mu_1, \sigma_1^2)$ & y be another random sample drawn from other pol^n i.e, $N(\mu_2, \sigma_2^2)$

∴ Null hypothesis :- $H_0 : \mu_1 = \mu_2$

there is no significant diff b/w means of populations.

∴ H_1 or H_A : $\mu_1 \neq \mu_2$

$\mu_1 > \mu_2$

$\mu_1 < \mu_2$

there is significance diff b/w means of popul^n

Test statistic:

Case → I → σ_1^2 & σ_2^2 is known

$$z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0.$$

where $\bar{X}$ & $\bar{Y}$ are sample mean & σ_1^2 & σ_2^2 are pop^n variance and n_1 & n_2 are sample sizes

case-II → σ_1^2 & σ_2^2 are unknown.

$$z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$ when S_1^2 & S_2^2 are Sample var.

Problem-2 → Two samples are drawn from two large population give the following result.

	mean	Std.	sample size
Sample I:	250	40	400
Sample II:	220	55	400

To examine whether there's any significant diff b/w means? Z_{tab} at 1% is 2.580.

Sol→ H_0: $\mu_1 = \mu_2$ no significant diff

H_1: $\mu_1 \neq \mu_2$

$$z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(250 - 220)}{\sqrt{\frac{(40)^2}{400} + \frac{(55)^2}{400}}} = \frac{20 \times 30}{\sqrt{1600 + 3025}}$$

$$= \frac{600}{68} = 8.822$$

$\therefore$ $|Z_{cal}| > |Z_{tab}|$ ⇒ reject H_0.

There is significant diff b/w mean.

Small Sample Test ($n \le 30$)

One Sample Test

→ Student t-test :- Let X be a random variable i.e, $X_1, X_2, \dots X_n$ be a random sample drawn from a normal population with mean 'μ' and variance σ^2 $N(\mu, \sigma^2)$. Let us consider $(\bar{x}, s^2)$ be the sample mean & sample variance.

- H_0: $\mu = \mu_0$ ie, the sample is drawn from normal popl^n.
- H_A: $\left.\begin{array}{l}\mu \neq \mu_0 \\ \mu > \mu_0 \\ \mu < \mu_0\end{array}\right\}$ the sample are not drawn from the normal popl^n.

Test statistic : $t = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n-1}} \sim N(0,1)$

where, n = sample size, μ_0 : pop^n mean

$\bar{x}$: sample mean, S : std. deviation

- Decision rule if $t_{cal} < t_{tab}$ with $(n-1)$ d.f. (degrees of freedom) at α% l.o.s. then accept H_0. otherwise reject H_0.

Problem → A manager has claimed that the average age of employee his company is 30.0. A random sample of 100 employee holding different pos^n give the following age list.

Age group	16-20	21-25	26-30	31-35	36-40
No. of employee	12	22	20	30	16

Calculate the A.M & Std. dev. of this list & test at 10% l.o.s.

Sol^n -) $\mu = 30.0$, $n = 100$, $\bar{x} = \frac{\sum f_i x_i}{N} = \frac{2880}{100} = 28.8$

C.I	f_i	x_i	$f_i x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
16-20	12	18	216	-10.8	116.64	1399.68
21-25	22	23	506	-5.8	33.64	740.08
26-30	20	28	560	-0.8	0.64	12.8
31-35	30	33	990	4.2	17.64	529.2
36-40	16	38	608	9.2	84.64	1354.24
			$\sum f_i x_i = 2880$			$\sum f_i(x_i - \bar{x})^2 = 4036$

H_0 : $\mu = \mu_0$ Sample is drawn from normal population $N(30, \sigma)$

H_1 : $\mu \neq \mu_0$
$\mu > \mu_0$
$\mu < \mu_0$

σ is unknown, Therefore,

$$S = \sqrt{\frac{1}{n-1}\left(\sum f_i(x_i - 28.8)^2\right)}$$

$$= \sqrt{\frac{1}{100-1} 4036} = \sqrt{40.76} = 6.384$$

$$z = \frac{\bar{X} - \mu}{S/\sqrt{n-1}} = \frac{28.8 - 30.0}{6.8/\sqrt{99}} = \frac{-1.2}{0.641} = -1.872$$

$\therefore$ $Z_{cal} = |-1.872| = 1.872$.

hence, $Z_{tab} < Z_{cal}$ Since $1.645 < 1.872$

$\therefore$ reject H_0. i.e, random sample is not drawn from the same popl^n.

Two Sample Student t-test

Let X be a random variable i.e, $x_1, x_2 \ldots x_n$ is random sample drawn from normal popl^n $N(\mu_1, \sigma_1^2)$.
Let us consider Y be a random variable i.e, $y_1, y_2 \ldots y_n$ be another random sample drawn from another normal popl^n $N(\mu_2, \sigma_2^2)$
Let us consider $\bar{x}$ & $\bar{y}$ are sample mean & s_1^2 & s_2^2 be sample var drawn from the normal pop^{ns}.

Ha or H_1 : $\mu_1 \neq \mu_2$
$\mu_1 > \mu_2$
$\mu_1 < \mu_2$

i.e, the two samples are not drawn from the same popl^n or not identical.

- Test statistic.

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{(n_1-1)} + \frac{s_2^2}{(n_2-1)}}} \sim N(0,1) \text{ under } H_0$$

Q⇒ General adult populn volunteer an avg. 4.2 hrs per week. A random sample of 20 female college student and 18 male college students indicated there result concerning the amount of time spent in volunteer increases / week. At 1% of los. is this sufficient evidence to conclude that a diff exist b/w the mean no. of volunteer hrs/week for male & female college student?

	male	female
Sample mean	2.5	3.8
Sample var.	2.2	3.5
Sample size.	18	20

Solⁿ :) H_0 : $\mu_1 = \mu_2$ there is no significant diff b/w the mean no. of volunteer.

H_A : $\mu_1 > \mu_2$

$\mu_1 < \mu_2$

$\mu_1 \neq \mu_2$

$$t = (\bar{x} - \bar{y}) \Big/ \sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}} = \frac{2.5 - 3.8}{\sqrt{\frac{2.2}{17} + \frac{3.5}{19}}} = \frac{-1.3}{\sqrt{0.129 + 0.184}}$$

$$= \frac{1.3}{\sqrt{0.313}} = \frac{1.3}{0.559} = 2.32$$

$t_{cal} = 2.32 < t_{tab} = 2.719$

accept H_0, i.e, no significant diff b/w mean calculated by male & female studen

Paired student t-test

Let X be a random variable and $x_1, x_2 \dots x_n$ be random sample drawn from normal popln with mean μ & std. dev. σ i.e, $N(\mu, \sigma^2)$

Null hypothesis (H_0) : $\mu_0 = 0$ there is no significant diff b/w before & after test

Alternate hypothesis (H_A): $\mu_0 = 0$

$\mu_0 > 0$

$\mu_0 < 0$

- Test statistic of paired student test is denoted by t_D & is defined as

$$t_D = \frac{\bar{D} - \mu_D}{S/\sqrt{n}} \sim N(0,1) \text{ under } H_0$$

where $D_i = x_1 - x_2$ $\quad \bar{D} = \frac{\sum D_i}{n}$

$$S_D = \sqrt{\frac{n(\sum D_i)^2 - (\sum D_i)^2}{n(n-1)}}$$

decision rule, $t_{cal} < t_{tab}$ with $(n-1)$ d.f. then accept H_0.

Q-) A physical education director claim by taking a special vitamin, a weightlifter can increases its strength after 2-weeks of regular training.
Compute the test at 5% l.o.s.

Athletes	1	2	3	4	5	6	7	8
Before (X_1):	210	230	182	205	262	253	219	216
After (X_2):	219	236	179	204	270	250	222	216

Sol^n-) $D_i (X_1 - X_2) = -9 \quad -6 \quad 3 \quad 1 \quad -8 \quad 3 \quad -3 \quad 0$

$\because$ last D_i is 0. $\Rightarrow$ no effect $\Rightarrow$ reduce $n = 8$ to $n-1 = 7$.

$$\therefore \bar{D} = \frac{\sum D_i}{n} = \frac{-19}{7} = -2.714$$

and D_i^2 : 81 36 9 1 64 9 9

$$\Sigma D_i^2 = 81 + 36 + 9 + 1 + 64 + 9 + 9 = 209$$

$$\therefore S_D = \sqrt{\frac{n \times (\Sigma D_i^2) - (\Sigma D_i)^2}{n(n-1)}}$$

$$= \sqrt{\frac{7 \times 209 - (-19)^2}{7 \times 6}} \Rightarrow \sqrt{\frac{1102}{42}} = \sqrt{26.23}$$

$$= 5.1215$$

$$\therefore t_D = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} = \frac{(-2.714) - 0}{5.1215/\sqrt{7}} = -1.406$$

& $t_{5\%} = 2.447$ with d.f. $= 6 = (n-1)$

Since $t_{cal} < t_{tab} \Rightarrow$ accept H_0.

Chi-Square (χ^2) test

→ Test for goodness of fit (if taken from popⁿ)
→ test for independence (if data already exist)
→ Yate's correction χ^2-test (only for 2x2 table)

→ Test for goodness of fit

observed value O_i

expected value E_i

H_0: there is no association b/w observed & expected value $O_i = E_i$

H_A or H_1 There is an association b/w observed & expected $O_i \neq E_i$

Test statistics:

case 1 → let $x_1, x_2 \ldots x_n$ be random sample drawn from any population then chi-sq. test is denoted by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2 \underbrace{[(R-1)(C-1)]}_{d.f.}$$

case 2 → If data is 2x2 contigency table.

$$\chi^2 = \frac{[|ad - bc|]\times n}{[(a+b)(c+d)(a+c)(b+d)]} \sim \chi^2 \underbrace{(1)}_{df.}$$

decision rule,

if $\chi^2_{cal} < \chi^2_{tab}$ then accept Ho, otherwise reject Ho.

Q→ A reader read that fireman related death for people 1 to 18 years distributed as follows.

74%. works accidental.

16%. works home suicide

10%. works other reason. In this district, there were 68 accidental, 27 homes suicide, and 5 were other during last year. ~~At α = 10%.~~

Sol→ $O_i = 68$, $O_2 = 27$ & $O_3 = 5$

expected value, $E_1 = 74$, $E_2 = 16$, $E_3 = 10$

& d.f. = $(2-1)(3-1) = 2$

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
68	74	-6	36	36/74 = 0.48
27	16	11	121	121/16 = 7.56
5	10	5	25	25/10 = 2.5

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 0.486 + 7.562 + 2.5$$

$$= 10.548$$

$\because$ $\chi^2_{cal} > \chi^2_{tab} \Rightarrow$ reject H_0.

i.e, ~~observed~~ data doesn't belong ~~#~~ to the fireman related given data.

Chi - Square (Yate's Correction)

test for 2x2 contigency table ($O_i \le 5$)

Let $x_1, x_2 \dots x_n$ be a random sample drawn from any popⁿ. If observed frequency is less than 5 then Yates correction is defined as.

$$\chi^2 = \frac{\left[\,|ad - bc| - N/2\,\right]^2 \times N}{(a+d)(c+d)\ (a+c)(b+d)} \sim \chi^2_{(1)\ d.f.}$$

- Null hypothesis $H_0 : O_i - E_i = 0$

Alternate hypotheses $H_A : O_i - E_i \neq 0$

i.e, there's an association b/w observed and expected frequency.

$\chi^2_{cal} < \chi^2_{tab}$ with '1' d.f. then accept H_0,

otherwise reject H_0.

Ques → Data on smoking habit and presence of lung cancer in the study is given below.

$(\chi^2_{tab}) = 3.87$

Cancer	Yes	No	Total
Yes	6	1	7 → R_1
No	4	19	23 → R_2
Total	10 = C_1	20 = C_2	30 → N

$$\chi^2_{tab} = 3.87 \text{ (given)}$$

$$\chi^2_{cal} = \frac{[(6\times19 - 1\times4) - 30/2]^2 \times 30}{7\times23\times10\times20}$$

$$= \frac{[(114-4) - 30/2]^2 \times 3\cancel{0}}{3220\cancel{0}} \Rightarrow \frac{(110-15)^2 \times 3}{3220}$$

$$= \frac{95\times95\times3}{3220} = 8.408$$

Since, $\chi^2_{cal} < \chi^2_{tab} \Rightarrow$ accept Ho.