Statistics

31.08.24

- Raw data
- Arrays

Averages:

Arithmetic mean (A.M.) : $f_1, f_2, f_3 \ldots$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{\Sigma f X}{\Sigma f}$$ long method

$$d = X - A$$

$$\bar{X} = A + \frac{\Sigma f d}{\Sigma f}$$ → short method

Coding method or step derivation method:

$$\bar{X} = A + \frac{\Sigma f u}{\Sigma f} i \text{, where } u = \frac{X - A}{i}$$

* Find the arithmetic mean (A.M.) for the following distribution by step derivation method.

class frequency	frequency	Mid value X	$u = \frac{X-A}{i}$	fu
0 - 10	7	5	-2	-14
10 - 20	8	15	-1	-8
20 - 30	20	25	0	0
30 - 40	10	35	1	10
40 - 50	5	45	2	10
	50			-2

i = size of the class interval = difference b/w class boundaries
= 10

A = the assumed mean

$$\bar{X} = 25 + \frac{-2}{50} \times 10$$

$$= 25 - \frac{2}{5}$$

$$= 24.6$$

0 - 9
10 - 19

এমন থাকলে graph plot এ ঝামেলা হবে class boundaries column add করা লাগবে।

class boundaries	class frequency	frequency	Mid value X	$u = \frac{X-A}{i}$	fu
-0.5 - 9.5	0 - 9	7	4.5	-2	-14
9.5 - 19.5	10 - 19	8	14.5	-1	-8
19.5 - 29.5	20 - 29	20	24.5 (A)	0	0
29.5 - 39.5	30 - 39	10	34.5	1	10
39.5 - 49.5	40 - 49	5	44.5	2	10
		50			-2

Median:

5 7 (8) 12 15

5 7 [8 12] 15 18

$\frac{8+12}{2}$

For grouped data:

$$\text{Median} = L + \left(\frac{\frac{N}{2} - c}{f}\right) i$$

where, L is the lower class boundary of the median class

N is the total frequency

c is the sum of the frequencies of all classes lower than the median class

i is the size of the median class

Mode:

series of data র মধ্যে যেটার frequency বেশি

2, 2, 5, 6, 6, 6, 9, 12, 15 → 6

2, 5, 7, 12 → mode নাই
(at least দুইবার থাকা লাগবে)

2, 2, 5, 5, 7, 7, 12, 12 → 2, 5, 7, 12
(more than 1 বার আছে)

07/09/24

$$\text{Median} = L_1 + \left(\frac{N/2 - F}{f}\right) i$$

where L_1 is the lower class boundary of the median class
N is the total frequency of the data
F is the sum of the frequencies of all classes lower than the median class.
f is the frequency of the median class.
and i is the width of the median class.

Find the median value from the following data.

class interval :	0-5	5-10	10-15 ← median group	15-20	20-25	25-30	30-35	35-40	40-45
frequency :	29	195	241	117	52	10	6	3	2

Total = 655

$$\frac{N}{2} = \frac{655}{2} = 327.5$$

10-15 এ থাকে।

$L_1 = 10$
$F = 19 + 195 = 224$
$f = 241$
$i = 5$

$$\text{Median} = L_1 + \left(\frac{N/2 - F}{f}\right) i$$

$$= 10 + \left(\frac{655/2 - 224}{241}\right) \times 5$$

$$= 12.147$$

↳ 10-15 এর মধ্যে
(∴ correct)

For grouped data :

$$\text{Mode} = L_1 + \left(\frac{f - f_1}{2f - f_1 - f_2}\right) i$$

where, L_1 is the lower class boundary of the modal class
f is the frequency of the modal class
f_1 is the frequency before the modal class frequency.
f_2 is the frequency after the modal class frequency.

Find the mode from the following data:

class interval :	0-6	6-12	13-18	18-24 (modal group)	24-30	30-36	36-42
frequency	6	11	25	35	18	12	6

$L_1 = 18,\ f = 35,\ f_1 = 25,\ f_2 = 18,\ i = 6$

$$\text{Mode} = L_1 + \left(\frac{f - f_1}{2f - f_1 - f_2}\right) i$$

$$= 18 + \frac{35 - 25}{2 \times 35 - 25 - 18} \times 6$$

$$= 20.222$$

২টার frequency 35 থাকলে how? — next class

The mean of 200 items was 50. Later on it was discovered that 2 items were misread as 92 and 8 instead of 192 and 88. Find the correct mean.

Soln:

$$\text{mean} = 50$$

$$\text{sum} = 200 \times 50$$

$$\text{correct sum} = 200 \times 50 - 92 - 8 + 192 + 88$$

$$= 10180$$

$$\therefore \text{correct mean} = \frac{10180}{200} = 50.9$$

group data র মধ্যে যদি কোনোটার frequency missing থাকে তাহলে কীভাবে?

From the following data find the missing frequency.

class interval : 40-43 43-46 46-49 49-52 52-55

frequency : 31 58 60 ? 27

It is given that the mean of the frequency is 47.2.

Soln:

class interval	f	mid value	u	fu	d	fd
40-43	31	41.5	-2	-62	-6	-186
43-46	58	44.5	-1	-58	-3	-174
46-49	60	47.5 (A)	0	0	0	0
49-52	x	50.5	1	x	3	3x
52-55	27	53.5	2	54	6	162
	176+x			-66+x		-198+3x

$$\bar{X} = A + \frac{\Sigma f u}{\Sigma f} i$$

$$47.2 = 47.5 + \frac{-66 + x}{176 + x} \times 3$$

$$\rightarrow x = 44$$

or

$$\bar{X} = A + \frac{\Sigma f d}{\Sigma f}$$

$$\rightarrow 47.2 = 47.5 + \frac{-198 + 3x}{176 + x}$$

$$\therefore x = 44$$

mean of the frequency 47.2 $\rightarrow$

$$\frac{176 + x}{5} = 47.2$$

$$\rightarrow x = 60$$

11.09.24

Standard deviation is defined as the square root of the mean square of the deviation from the A.M.

$$S.D = \sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}}, \quad \sigma^2 = \text{variance}$$

Standard deviation zero হলে কি বোঝায়?

Calculate the S.D. for the following data:

class interval	f	x	$d = x - A$ (A=6)	fd	fd^2
0–4	4	2	−4	−16	64
4–8	8	6 (A)	0	0	0
8–12	2	10	4	8	32
12–16	1	14	8	8	64
	$\Sigma f = 15$			$\Sigma fd = 0$	$\Sigma fd^2 = 160$

$$S.D = \sqrt{\frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2} = 3.266$$

যে formula দিয়েই করি same আসবে।

Following are the marks obtained by two students X and Y in 10 tests of 100 marks each:

marks obtained by X :	44	80	76	48	52	72	68	56	60	54
marks obtained by Y :	48	75	54	60	63	69	72	51	57	66

If the consistency of performance is the criterion for awarding a prize, who should get the prize?

$\bar{x} = \frac{610}{10} = 61$ $\quad \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}} = 11.7$ $\quad \frac{\sigma}{\bar{x}} \times 100 = 19.18$

$\bar{y} = \frac{615}{10} = 61.5$ $\quad \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}} = 8.63$ $\quad \frac{\sigma}{\bar{x}} \times 100 = 14.03$ (Winner)

coefficient of variation (C.V) $= \frac{\sigma}{\bar{x}} \times 100$

যার C.V. কম → তার performance ভালো।

Ref:

R.S. Pillai
V. Bagavathi
} statistics

Moments:

The rth moment of a variable x about the mean $\bar{x}$ is usually denoted by μ_r and is given by

$$\mu_r = \frac{1}{N} \Sigma f_i (x - \bar{x})^n \text{, where } \Sigma f_i = N$$

$r = 1 \rightarrow$ first moment about the mean

$\rightarrow \mu_1$ will be always zero

x	$x - \bar{x}$
2	-1
3	0
4	1

$\bar{x} = 3$

The rth moment of a variable x about any point a is defined by

$$\mu_r' = \frac{1}{N} \sum f_i (x_i - a)^r$$

Relation between moments about mean and moment about any point a

$$\mu_1 = 0 = \mu_1' - \mu_1'$$

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

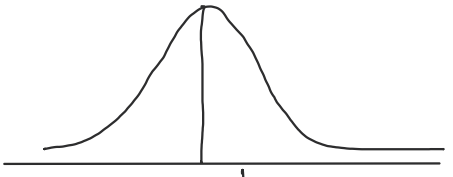
14.09.24

Skewness is the lack of symmetry. The measures of asymmetry are usually called measures of skewness

$$\text{coefficients of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{s.d.}}$$

symmetric

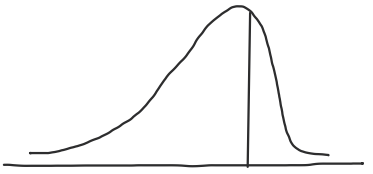
$(\beta_1 = 0)$



+ve skewness

or right hand skewness

$(\beta_1 > 0)$



(-ve) skewness

or left hand skewness

$(\beta_1 < 0)$

Measures of skewness: $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$

$\mu_4 \rightarrow$ 4th moment about the mean

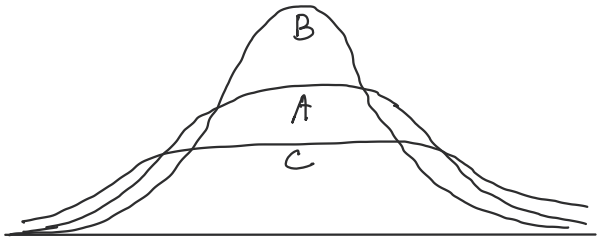
Kurtosis: It measures the degree of peakedness of a distribution and is given by

$\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 3$, the curve is normal or mesokurtic distribution

$\mu_2 \rightarrow$ 2nd moment about the mean

If $\beta_2 > 3$, the distribution is leptokurtic

and if $\beta_2 < 3$, the distribution is platykurtic distribution.



A – meso kurtic

B – lepto kurtic

C – platy kurtic

⊛ In a certain distribution the first four moments about the value 5 are 2, 20, 40 and 50. Calculate β_1 and β_2 and comment about any value a on the shape or nature of the distribution. (μ')

Solution: Here, $\mu_1' = 2$, $\mu_2' = 20$, $\mu_3' = 40$ and $\mu_4' = 50$

$$\mu_1 = \mu_1' - \mu_1' = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 16$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 = -64$$

$$\mu_4 = 162$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 1 \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = 0.6328$$

→ (+ve) distribution ($\beta_1 > 0$)

→ platykurtic ($\beta_2 < 3$)

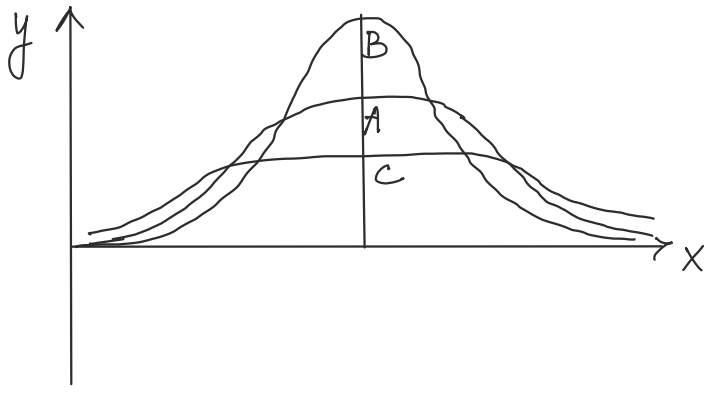
18.09.24

Kurtosis: The degree of Kurtosis of a distribution is measured relative to the peakedness or a normal

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3,$$ the curve is normal or mesokurtic distribution

μ_2 → 2nd moment about the mean

If $\beta_2 > 3$, the distribution is leptokurtic
and if $\beta_2 < 3$, the distribution is platykurtic distribution.



A — mesokurtic
B — lepto kurtic
C — platy kurtic

* Calculate the first four moments from the following data and find out β_1 and β_2 and also comment on your result.

এই case-এ $\bar{x} = 4$ হয়; so μ_r', μ_r বলা যাবে।

$A = 4$

x	f	$f(x)$	$d = x - A$	$f(x-4)$	$f(x-4)^2$	$f(x-4)^3$	$f(x-4)^4$
0	5		-4	-20	80	-320	1280
1	10		-3	-30	90	-270	810
2	15		-2	-30	60	-120	240
3	20		-1	-20	20	-20	20
4	25		0	0	0	0	0
5	20		1	20	20	20	20
6	15		2	30	60	120	240
7	10		3	30	80	270	810
8	5		4	20	90	320	1280
	125 = N = ΣF	$\Sigma + x$ = 500		$\Sigma f d = 0$	$\Sigma f d^2$ = 500	$\Sigma f d^3$ = 0	$\Sigma f d^4$ = 4700

$\mu_1 = \frac{\Sigma f d}{N} = 0$ $\quad \mu_2 = \frac{\Sigma f d^2}{N} = \frac{500}{125} = 4$ $\quad \mu_3 = \frac{\Sigma f d^3}{N} = 0$ $\quad \mu_4 = \frac{\Sigma f d^4}{N} = 37.6$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \quad \text{(symmetric distribution)}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 2.37 < 3 \quad \text{(platy kurtic)}$$

Quartiles, Deciles and Percentiles:

(3 Q) (9 D) (99 P)

First quartile, $Q_1 = L_1 + \frac{N/4 - c.f.}{f_{Q_1}} \times c$

↳ frequency of the first quartile class.

$$Q_2 = L_2 + \frac{2N/4 - c.f.}{f_{Q_2}} \times c$$

L_2: lower class boundary of the second quartile class

$$Q_3 = L_3 + \frac{3N/4 - c.f.}{f_{Q_3}} \times c$$

2nd quartile = 5 th decile = 50th percentile

6th decile, $D_6 = L_6 + \frac{6N/10 - c.f.}{f_{D_6}} \times c$

L_6: lower class boundary of the 6th decile class.

↳ frequency of the 6th decile class

nth percentile এ Nn

$$65\text{th percentile} = L_{65} + \frac{65N/100 - c.f.}{f_{65}} \times c$$

$c \rightarrow$ interval

$c.f. \rightarrow$ ঐ class এর আগের class পর্যন্ত frequency র sum.

21.09.24

Probability : If an event A can happen in m ways, and failed to happen in n ways, all this ways being equally likely to occur then the probability of happening A is

$$= \frac{\text{number of favourable cases}}{\text{Total number of mutually exclusive and equally likely cases}} = \frac{m}{m+n} = p$$

and that the probability of not happening $= \frac{n}{m+n} = q$

$$p + q = 1$$

mutually exclusive — H আসলে T আসবে না and vice versa.

Dice : (more than 1 Die)

Die — একটা ছক্কা

Find the probability of throwing 9 with two dice.

sum of the upper face of the dice.

Sum = 9

(5,4)
(4,5)
(6,3)
(3,6)

$\therefore$ probability $= \frac{4}{36}$

$= \frac{1}{9}$

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

Expected Value: If $p_1, p_2, p_3 \ldots, p_n$ are probabilities of the events $x_1, x_2 \ldots, x_n$ respectively then expected value:

$$E(x) = p_1 x_1 + p_2 x_2 + \ldots + p_n x_n = \sum_{i=1}^{n} p_i x_i$$

Conditional Probability: The probability of happening an event A, such that event B has already happened, is called the probability of happening A on the condition that B has already happened. It is usually denoted by $P(A|B)$.

If two events are mutually exclusive, then the probability of the occurrance of either A or B is the sum of the probabilities of A and B.

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

Ex: A bag contains 3 white, 2 black and 5 red balls. What is the probability of getting a white or red ball at random in a single draw.

Solution: The probability of getting a white ball is $= \frac{3}{10}$

The probability of getting a red ball is $= \frac{5}{10}$

∴ The probability of getting a white or red ball is $= \frac{3}{10} + \frac{5}{10}$

$= \frac{8}{10}$

$= \frac{4}{5}$

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

When events are not mutually exclusive:

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

২ জনকে একটা problem দেওয়া হলো।

X এর solve করার probability $\frac{3}{4}$

Y " " " " $\frac{2}{3}$

problem টা solve হওয়ার probability = ?

Soln: $P(X \text{ or } Y) = P(X) + P(Y) - P(X \cap Y)$

$= \frac{3}{4} + \frac{2}{3} - \frac{3}{4} \times \frac{2}{3}$

$= \frac{11}{12}$

19.10.24

Bayes Theorem : (inverse probability)

The applications of the result of the probability theory involves estimating unknown probabilities and making decisions on the basis of the new sample information. This concept is reffered to as Bayes theorem.

1. prior probability
2. posterior probability

- Probabilities assigned on the basis of the personal experience, before observing the outcomes of the experiment are called prior probabilites.

e.g. bulb production ৳
defective percentage আগের experience থেকে বের করা → prior probability

- When the probabilities are revised with the use of Bayes rule, they are called posterior probabilities.

✳ A company has two plants to manufacture scooters.

plant - 1 —— manufactures 80%
plant - II —— manufactures 20%

plant I maintained 85 out of 100 scooters standard quality.
plant II " 65 " " 100 " " "

additional information → Baye's theorem

What is the probability that the scooter selected at random came from plant I if it is known that the scooter is of standard quality?
Similarly for plant - II.

Soln: Let A_1 be the event that of drawing a scooter produced by plant I and A_2 " " " " " " " " " " " II.

B be the event of drawing a standard quality scooter produced by either plant I or plant II.

Then from the first information, $P(A_1) = 80\% = 0.8$

and $P(A_2) = 20\% = 0.2$

From the additional information, $P(B/A_1) = \frac{85}{100} = 85\%$

$P(B/A_2) = 65\%$

The required values are computed in the following table:

Event ①	Prior probability ②	Conditional probability ③	Joint probability ④ ⇒ (2×3)	Posterior Probability (Revised) ⇒ {④ ÷ P(B)}
A_1	0.80	0.85	0.68	$\frac{0.68}{0.81} = \frac{68}{81}$
A_2	0.20	0.25	0.13	$\frac{0.13}{0.81} = \frac{13}{81}$
	sum = 1		sum = 0.81	sum = 1

From the first information we may say that the standard scooter is drawn from plant I since

$P(A_1) = 80\%$ which is greater than $P(A_2) = 20\%$

From the additional information, we may conclude that the standard quality of scooter is more likely drawn by the output given by plant I.

23.10.24

Population: The group of individuals under study is called population or universe.

Sampling: A part selected from the population is called a sample. The process of selection of a sample is called sampling.

A random sample is one in which each member of the population has an equal chance of being included it. There are $^{N}C_{n}$ different samples of size n that can be picked up from a population of size n.

Parameters: functions of population.

— mean (μ), standard deviation (σ) which are the statistical constants of the population are called parameters.

The mean $\bar{x}$, standard deviation (s) of a sample are called statistic.

Statistic → functions of samples

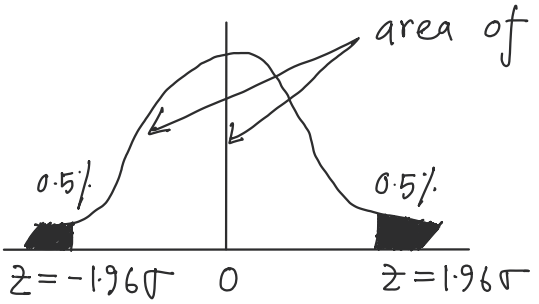
Standard error : (S.E) is the standard deviation of sampling distribution. For assessing the difference between the expected value and observed value, standard error is used.

Testing a hypothesis : on the basis of the sample information, we make certain decisions about the population. In taking such decisions, we make certain assumptions. These assumptions are known as statistical hypothesis.

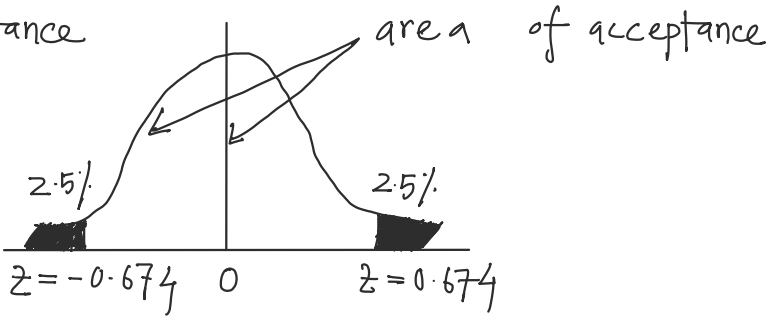
5% level of significance : 5 টা খারাপ 95 টা ভালো।

Level of significance :

There are two critical regions which cover 5% and 1% areas of the normal curve. This shaded portions are the critical regions.



1% level of significance



5% level of significance

small sample — sample size ≤ 30
↳ tree distribution / tree test

large sample — z-test

10 জন people height average 65 inch হবে কি না ?

↓
hypothesis

5% level of significance দিয়ে বের করতে হবে।

assume $t = 0.05$ ⟶ 3.5

↓
table value

2.61
calculated value

here,
calculated value < table value
therefore hypothesis is accepted.

26.10.24

Small space :

To calculate a significance of sample mean at 5% level of significance

calculate $t = \frac{\bar{x} - \mu}{s}\sqrt{n}$ and compare it to the value of t with $(n-1)$ degrees of freedom at 5% level obtained from the table. Let this tabulated value of t be t_1.

If $t < t_1$, then we accept the hypothesis, i.e we say that the sample is drawn from the population.

If $t > t_1$, we compare it with the tabulated value of t at 1% level of significance for $(n-1)$ degrees of freedom.

Denote it by t_2, if $t_1 < t < t_2$, then we say that the value of t is significant.

If $t > t_1$, we reject the hypothesis and the sample is not drawn from the population.

Sample: Ten individuals are chosen at random from a population and their heights are found to be in inches 63, 63, 64, 65, 66, 69, 69, 70, 71. Discuss the suggestion that the mean height of universe is 65. It is given that for 9 degrees of freedom t at 5% level of significance = 2.262.

Solution: Let us take the hypothesis that the mean height of universe is 65 inches.

Applying t-test: $t = \frac{\bar{x} - \mu}{s}\sqrt{n}$

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} = \sqrt{\frac{88}{9}} = 3.13$$

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{67 - 65}{3.13}\sqrt{10}$$

$$= 2.02$$

x
63
63
⋮
$\Sigma = 670$

Since the calculated value is less than the table value, so the hypothesis is accepted.
Therefore, the mean height of universe is 65 inches.

The mean life time of sample of 100 flurescent light bulbs produced by a company is computed to be 1570 hours with a standard deviation of 120 hours. The company claims that the average life of bulbs produced by it is 1600 hours. Using the level of significance of 0.05, is the claim acceptable?

Solution: Let us take the hypothesis that the claim is acceptable of

$t = \frac{\bar{x} - \mu}{s}\sqrt{n}$, here $\bar{x} = 1570$, $s = 120$, $n = 100$, $\mu = 1600$

$t = 2.5$ (calculated value)

But at 5% level of significance $t_{0.05} = 1.96$

since calculated value > table value so the claim is not acceptable so it is rejected

02.11.24

Certain pesticide is packed into bags by a machine. A random sample of 10 bags is drawn and their contents are found to weigh (in kg) as follows:

50, 49, 52, 44, 45, 48, 46, 45, 49, 45.

Test if the average packing can be taken to be 50 kg.

Solution: Null hypothesis:

H_0, $\mu = 50$ kgs in the average packing is 50 kgs.

Alternative hypothesis:

H_1 : $\mu \neq 50$ kgs (Two-tailed)

Level of significance:

let, $\alpha = 0.05$

Calculation of sample mean and S.D.

x	$d = x - 48$	d^2
50	2	4
49	1	1
52	4	16
⋮	⋮	⋮
	$= \Sigma d = -7$	$= \Sigma d^2 = 69$

assumed mean দিয়ে করা।
—actual mean দিয়েও করা যাবে।

$$\bar{x} = A + \frac{\Sigma d}{n} = 48 + \frac{-7}{10} = 47.3$$

$$s^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = 7.12$$

Calculation of statistic :

Under H_0, the test statistic is :

$$t_0 = \left|\frac{\bar{x} - \mu}{\sqrt{s^2/\mu}}\right| = \left|\frac{47.3 - 50.0}{\sqrt{7.12/10}}\right| = 3.2$$

(calculated value)

Expected value:

$$t_e = \left|\frac{\bar{x} - \mu}{\sqrt{s^2/\mu}}\right|$$ follows t distribution with $(10-1)$ d.f.

$= 2.262$ (table থেকে নেওয়া)

(table value)

calculated value > table value

→ hypothesis rejected.

Inference: Since $t_o > t_e$, H_o is rejected at 5% level of significance and we conclude that the average packing can not be taken to be 50 kgs.

defn:
- level of significance
- test of "
- degree of freedom

T test sample size < 30

One-sample z-test:

A one sample z-test is used to check if there is a difference between the sample mean and the population mean when the population standard deviation is known.

06.11.2024

Z-test:

(i) Z test is a statistical test that is conducted on normally distributed data to check if there is a difference in means of two data sets.

(ii) The sample size should be greater than 30 and the population variance must be known to perform a Z-test.

(iii) the one sample Z-test checks if there is a difference in the sample and population mean.

(iv) The two sample Z-test checks if the means of two different groups are equal.

One sample Z-test:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$, where $\bar{X}$ is the sample mean
μ is the population mean
σ is the population standard deviation and n is the sample size.

The algorithm to set a one sample Z-test based on the Z-test statistic is given as follows:

Left tailed test:

Null hypothesis : $H_0 : \mu = \mu_0$

Alternative " $H_1 : \mu < \mu_0$

Decision criteria : If the z - statistic $<$ z - critical value, then reject the null hypothesis.

Right tailed test:

Null Hypothesis: $H_0: \mu = \mu_0$

Alternative " : $H_1: \mu > \mu_0$

Decision Criteria : If the z - statistics $>$ z critical value then reject the null hypothesis.

Two tailed test:

Null Hypothesis: $H_0: \mu = \mu_0$

Alternative " : $H_1: \mu \neq \mu_0$

Decision Criteria : If the z - statistics $>$ z critical value then reject the null hypothesis.

Two sample z - test:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

where $\overline{x_1}$, μ_1, σ_1^2 are the sample mean, population mean and population variance respectively for the first sample and

$\overline{x_2}$, μ_2 and σ_2^2 are the sample mean, population mean and population variance respectively for the second sample

Z-test for proportions:

A z-test for proportion is used to check the difference in proportions. A z-test can either be used for one proportion or two proportions.

One proportion z-test is used where there are two groups and compares the value of an observed proportion to a theoretical one. The z-test statistic for a one proportion z-test is as follows

$$z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$, Here, p is the observed value of the proportion, P_0 is the theoretical proportion value and n is the sample size.

Two proportion z-test:

A two proportion z-test is conducted on two proportions to check if they are same or not.

09.11.24

Chi-square test : χ^2-test

Chi-square test is applied in statistics to test the goodness of fit to verify the distribution of observed data with assumed theoretical distribution.

If there is no difference between the actual and expected frequencies χ^2 is zero.

χ^2-test of goodness of fit:

Through the the test we can find out the deviations between the observed values and the expected values.

The χ^2-test may be defined as,

$$\chi^2 = \sum \frac{(O_i - E_i)}{E_i}$$, where O_i is the observed frequencies and E_i is the expected frequencies

For a contingency table, i.e. if it is 2×2 table, then the degrees of freedom is

$$\nu = (c-1)(r-1) = (2-1)(2-1) = 1$$ degree of frequency.

Example: A dice is tossed 120 times with the following results

No. turned up :	1	2	3	4	5	6	Total
Frequency :	30	25	18	10	22	15	= 120

Test the hypothesis that the dice is unbiased.

Solution: Let us take the hypothesis that the dice is unbiased one.

The expected frequency is = $120 \times \frac{1}{6} = 20$

Applying the χ^2 – test :

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
30	20	10	100	5
25	20	5	25	1.25
18	20	−2	4	
10	20	−10	100	
22	20	2	4	
15	20	−5	25	
				12.90

$$= \sum \frac{(O_i - E_i)^2}{E_i}$$

Here, d.f. = 6 − 1 = 5

The table value for 5 degrees of freedom at 5% level is 11.07.

since calculated value > table value

$\therefore$ hypothesis is rejected.

So, the dice is biased.

* A certain drug was administered to 456 males out of a total 720 in a certain locality, to test its efficiency against typhoid. The incidence of typhoid is shown below.

	Infection	No infection	Total
Administering the drug	144 (O_{11})	312 (O_{12})	456
Without administering the drug	192 (O_{21})	72 (O_{22})	264
Total	336	384	720

Find out the effectiveness of the drug against the disease (The table value of χ^2 for 1 d.f. at 5% level of significance is 3.84)

Expected frequency $\longrightarrow$ $\dfrac{\text{Row total} \times \text{column total}}{\text{Grand total}}$

degree of freedom, d.f. $= (r-1)(c-1) = 1$

2×2

$$O_{11} \rightarrow E_{11} = \frac{456 \times 336}{720} = 212.8$$

$$O_{12} \rightarrow E_{12} = \frac{456 \times 384}{720} = 243.2$$

$$O_{21} \rightarrow E_{21} = \frac{264 \times 336}{720} = 123.2$$

$$O_{22} \rightarrow E_{22} = \frac{264 \times 384}{720} = 140.8$$

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
144	212.8			
312	243.2			
192	123.2			
72	140.8			

16.11.24

How to split a stem and leaf plot?

- The split stem and leaf plot separates each stem into many stems based on its frequency. We place the smaller leaves on the first part of the split stem and the larger leaves on the subsequent stems.

For example, let us take 7 students whose marks in math test are as follows: 78, 82, 82, 90, 81, 72

Stem and leaf plot:

Stem	Leaf
7	28
8	122
9	04

Split stem and leaf plot:

Stem	Leaf
7	2
7	8
8	1
8	2
8	2
9	0
9	4

Ex: Richard is trying to read the plot given below. His teacher has given him stem and leaf plot worksheets.

(i) What is mode of the plot

(ii) " " mean of the plot

(iii) " " the range of the plot

Stem	leaf
1	24
2	158
3	246
5	0344
6	257
8	389
9	1

Soln of (i): leaf 4 occurs twice on the plot against stem 5.

Hence mode is 54.

(ii) the sum of the data values is:

$$12 + 14 + 21 + 25 + 28 + 32 + 34 + 36 + 50 + 53 + 54 + 54 + 62 + 65 + 67 + 83 + 88 + 89 + 91 = 958.$$

In order to find the mean, we have to divide the sum by the total number of the values

So, mean $= 958 \div 19$

$= 50.42$

(iii) Range = the highest value − the lowest value

$= 91 - 12$

$= 79$

20.11.24

Two proportion z-test:

A two proportion z-test is conducted on two proportions to check if they are the same or not. The test statistics formula is −

$$z = \frac{P_1 - P_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where, $p = \frac{x_1 + x_2}{n_1 + n_2}$

- P_1 is the proportion of the sample 1 with same size n_1 and x_1 is the number of trials.

- p_1 is the proportion of the sample II with same size n_2 and x_2 is the number of trials.

Ex: A company wants to improve the quality of products by reducing defects and monitoring the efficiency of assembly line A, there were 18 defects reported out of 200 samples, while in line B, 25 defects out of 600 samples were noted. Is there a difference in the procedures at a 0.05 alpha level?

Solution: This is an example of a two-tailed two proportion z-test.

Null hypothesis: H_0 : The two proportions are same.

Alternate hypothesis : H_1 : The two proportions are not same.

As this is a two tailed test the α level needs to be divided by 2 to get 0.025. Using this the critical value from the z-table is 1.96

Here, $n_1 = 200$, $n_2 = 600$

$$P_1 = \frac{18}{200} = 0.09$$

$$P_2 = \frac{25}{600} = 0.0416$$

$$P = \frac{18 + 25}{200 + 600} = 0.0537$$

Since calculated value > table value, i.e. 2.62 > 1.96, thus the null hypothesis is rejected and it is concluded that there is a significant difference between two lines.

- defⁿ
  - ↳ level of significance
  - → test of significance
  - → one tailed test
  - → two tailed "
  - → degrees of freedom