Correlation, Regretion and Sampling Distribution

Correlation (Introduction): For study the characteristic of only one variable like marks, weights, heights, prices, ages, sales etc.

* This type of analysis is called univariate analysis.
* If there exists some relationship between two variables then the statistical analysis of such data is called bivariate analysis.

* Correlation refers to the relationship of two or more variable. There exists a relationship between the height of th a father and a son. The study of the relation is called Correlation. It measures the closeness of the relationship between the variables.

definition: Correlation is a statistical analysis which measures and analyze the degree or extent to which two variables fluctuate with reference to each other.

* The Correlation express the relationship or interdependent of two sets of variables upon each other. one variable may be called the subject (independent) and the other relative (dependent).
* A distribution involving two variables is known as bivariate distribution. If these two variable vary such that change in one variable effects the change in other variable and the variables are said to be Correlated

26-10-2023

Ex: There exist some relation between height and weight of person.

2) price of committee and its demand.

Note:

(i) The degree of relationship between the variable under consideration is measured through the correlation analysis.

(ii) The measures of correlation is called as Coefficient of Correlation or Correlation index.

Types of Correlation:

* Correlation are classified into many types.

(i) positive and negative

(ii) simple and multiple

(iii) Total and partial

(i) positive and negative: If two variables tend to move together in the same direction that is an increase in the value of one variable is accompanied by an increase in another variable or vice versa is called a "positive Correlation".

* If two variables tends to move together in opposite direction so that an increase or decrease in the value

26-10-2023

of one variable by ~~a degrees~~ is accompanied by a degrees of another variable is called "negative correlation".

(ii) simple and multiple: When we study only two variables i.e. so relationship is called "simple".

* when we study more than two variables simultaneously that relationship is called "multiple".

(iii) Total and partial: The study of two variables excluding some other variables is known as partial.

* In total Correlation all the facts are ~~considered~~ taken into account.

linear and Non-linear Correlation: If the ration of change between two variables is uniform then there will be a linear correlation between them.

* The amount of change in one variable does not bear a constant ration of the amount of change in other variable is known as Non-linear Corelation.

Methods of studying Correlation: There are two methods.

1) Graphical method

2) Mathematical method.

Mathematical methods: There are two types

(i) Karl person coefficient of Correlation

(ii) Spearman's rank Correlation

(i) **Coefficient of Correlation**: This method is used for measuring the magnitude of linear relationship between two variables. It is denoted by 'r' and it is defined as

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{N}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{N}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{N}\right\}}}$$

where
$$\left\{\begin{array}{ll} x = x - \bar{x}, & \bar{x} = \frac{\Sigma x}{N} \\ y = y - \bar{y}, & \bar{y} = \frac{\Sigma y}{N} \end{array}\right.$$
deviation

* r lies between ± 1

Problems:

1. Calculate the Coefficient of Correlation for the following data

~~x = 12, 9, 8~~

x	12	9	8	10	11	13	7
y	14	8	6	9	11	12	3

Sol:

x	y	x^2	y^2	xy
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\Sigma x = 70$	$\Sigma y = 63$	$\Sigma x^2 = 728$	$\Sigma y^2 = 651$	$\Sigma xy = 676$

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{N}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{N}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{N}\right\}}}$$

$$r = \frac{676 - \left(\frac{70 \times 63}{7}\right)}{\sqrt{\left\{728 - \frac{(70)^2}{7}\right\}\left\{651 - \frac{(63)^2}{7}\right\}}}$$

$$r = 0.9485$$

30-10-2023

2, Find the Coefficient of correlation between heights and weights given below

Height (in inches) x	57	59	62	63	64	65	55	58	57
weight y	113	117	126	126	130	129	111	116	112

Sol:

x	$x = X - \bar{X}$	y	$Y = y - \bar{Y}$	x^2	Y^2	xY
57	-3	113	~~113~~ -7	9	49	21
59	-1	117	~~117~~ -3	1	9	3
62	2	126	~~126~~ 6	4	36	12
63	3	126	~~126~~ 6	9	36	18
64	4	130	~~130~~ 10	16	100	40
65	5	129	~~129~~ 9	25	81	45
55	-5	111	~~111~~ -9	25	81	45
58	-2	116	~~116~~ -4	4	16	8
57	-3	112	112 -8	9	64	24
$\Sigma X = 540$	$\Sigma x = 0$	$\Sigma y = 1080$	$\Sigma Y = 0$	$\Sigma x^2 = 102$	$\Sigma Y^2 = 472$	$\Sigma xY = 216$

$\bar{X} = \frac{\Sigma x}{N} = \frac{540}{9} = 60$

$\bar{Y} = \frac{\Sigma y}{N} = \frac{1080}{9} = 120$

$$r = \frac{\Sigma xY - \frac{\Sigma x \Sigma Y}{N}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{N}\right\}\left\{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right\}}}$$

30-10-2023

$$r = \frac{216 - 0}{\sqrt{\{102 - 0\}\{472 - 0\}}}$$

$r = 0.98$

3, Calculate Coefficient of Correlation for the following data

x	28	41	40	38	35	33	40	32	36	33
y	23	34	33	34	30	26	28	31	36	38

Sol:

X	$x = X - \bar{X}$	y	$Y = y - \bar{Y}$	x^2	Y^2	xY
28	-8	23	-8	64	64	64
41	5	34	3	25	9	15
40	4	33	2	16	4	8
38	2	34	3	4	9	6
35	-1	30	-1	1	1	1
33	-3	26	-5	9	25	15
40	+4	28	-3	16	9	-12
32	-4	31	0	16	0	0
36	0	36	5	0	25	0
33	-3	38	7	9	49	-21
$\Sigma X = 356$	$\Sigma x = -4$	$\Sigma y = 313$	$\Sigma Y = 3$	$\Sigma x^2 = 160$	$\Sigma Y^2 = 195$	$\Sigma xY = 76$

$\bar{X} = \frac{\Sigma x}{N} = \frac{356}{10} = 35.6 \cong 36$

$\bar{Y} = \frac{\Sigma y}{N} = \frac{313}{10} = 31.3 \cong 31$

31-10-2019

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left\{\sum X^2 - \frac{(\sum X)^2}{N}\right\}\left\{\sum Y^2 - \frac{(\sum Y)^2}{N}\right\}}}$$

$$r = \frac{76 - \frac{(-4)(3)}{10}}{\sqrt{\left\{160 - \frac{(\cancel{356}\,4)^2}{10}\right\}\left\{195 - \frac{(3)^2}{10}\right\}}}$$

$$r = 0.44$$

Rank correlation (or) non-repeated ranks: This method is based on rank and is used in dealing with qualitative characteristics such as intelligence, beauty, morality etc.

* It is based on the ranks given to the observation.
* Rank correlation is applicable only to the individual observation.
* It is denoted as ρ and it is defined as

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)}$$

~~where~~ where D = sum of squares of difference between two ranks.

N = number of paired observation

31-10-2019

problem:

1, The following are the ranks obtained by 10 students in 2 subjects

statistical values (S)	1	2	3	4	5	6	7	8	9	10
mathematical (M) values	2	4	1	5	3	9	7	10	6	8

sol:

x	y	$D = x-y$	D^2
1	2	−1	1
2	4	−2	4
3	1	2	4
4	5	−1	1
5	3	2	4
6	9	−3	9
7	7	0	0
8	10	−2	4
9	6	3	9
10	8	2	4
			$\sum D^2 = 40$

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)}$$

$$\rho = 1 - \frac{6 \times 40}{10(10^2-1)}$$

$$\rho = 0.75$$

31-10-2019

2, A random sample of 5 college students are selected and their grades in mathematics and statistic values are found to be

M	85	60	73	40	90
S	93	75	65	50	80

Sol:

x	X	y	Y	D = X − Y	D^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	−1	1
40	5	50	5	0	0
90	1	80	2	−1	1
					$\Sigma D^2 = 4$

$$\rho = 1 - \frac{6 \Sigma D^2}{N(N^2-1)}$$

$$\rho = 1 - \frac{6(4)}{5(5^2-1)}$$

$$\rho = 0.8$$

31-10-2019

Equal or Repeated ranks: If any two or more persons are bracketed equally in any classification or if there is more than one item with the same value in the series then we will apply repeated rank correlation.

* The common rank is the average of the ranks which these items would have assumed if they were different from each other and the next item will get the rank next to ranks already assumed and it is defined as

$$\rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \cdots\right]}{N(N^2-1)}$$

where m = number of items repeated

Problem:

1, From the following data calculate the rank correlation coefficient after making adjustment for tied ranks.

x	48	33	40	9	16	16	65	24	16	57
y	13	13	24	6	15	4	20	9	6	19

01-11-2023

Sol:

X	x	y	Y	D = X − Y	D^2
48	8	13	5.5	2.5	6.25
33	6	13	5.5	0.5	0.25
40	7	24	10	−3	9
9	1	6	2.5	−1.5	2.25
16	3	15	7	−4	16
16	3	4	1	2	4
65	10	20	9	1	1
24	5	9	4	1	1
16	3	6	2.5	0.5	0.25
57	9	19	8	1	1
					$\Sigma D^2 = 41$

Here $m = 3, 2, 2$

$$\rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m)\right]}{N(N^2-1)}$$

$$\rho = 1 - \frac{6\left[41 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right]}{10(10^2-1)}$$

$$\rho = 0.73$$

01-11-2023

2) A sample of 12 fathers and their elder sons gave the following data

X	65	63	67	64	68	62	70	66	68	67	69	71
y	68	66	68	65	69	66	68	65	71	67	68	70

Sol:

X	x	y	Y	D = X − Y	D^2
65	4	688	7.5	−3.5	12.25
63	2	664	3.5	−1.5	2.25
67	7	682	7.5	−0.5	0.25
64	3	65	1.5	1.5	2.25
68	~~4~~ 57.5	69	10	−2.5	6.25
62	1	663	3.5	−2.5	6.25
70	10	689	7.5	2.5	6.25
66	5	65	1.5	3.5	12.25
68	~~4~~ 575	71	12	−4.5	20.25
67	6	67	5	1	1
69	9	686	7.5	1.5	2.25
71	11	70	11	0	0
					$\Sigma D^2 = 71.5$

Here $m =$

$$\rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m)\right]}{N(N^2-1)}$$

Regression (2M) The study of Correlation measures the direction and the strength of relationship between two variables.

* In Correlation we can estimate the value of other variable when the value of one variable is given.
* But in regression we can estimate the value of one variable with the value of other variable which is known.
* The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called Regression.

Methods of studying Regression:

(i) 1) Graphical method

2) Algebraic method

Regression line: A regression line is a straight line fitted to the data by the method of least squares.

* There are always two regression lines constructed for the relationship between two variables x and y.

Regression equation for a straight line equation of y on x:

$$y = a + bx$$

The normal equation

$$\Sigma y = Na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Regression equation for a straight line equation of x on y:

$$x = a + by$$

The normal equation

$$\Sigma x = Na + b\Sigma y$$

$$\Sigma xy = a\Sigma x + b\Sigma y^2$$

Deviation taken from arithmetic mean:

1) Regression equation of x on y

* The equation is $x - \bar{x} = r\dfrac{\sigma x}{\sigma y}(y - \bar{y})$

where $r\dfrac{\sigma x}{\sigma y} = b_{xy} = \dfrac{\Sigma xy}{\Sigma y^2}$

$x = x - \bar{x}$ $\quad \bar{x} = \dfrac{\Sigma x}{N}$, $\bar{y} = \dfrac{\Sigma y}{N}$

$y = y - \bar{y}$

2) Regression equation of y on x

$$y - \bar{y} = r\frac{\sigma y}{\sigma x}(x - \bar{x})$$

01-11-2023

where $r\frac{\sigma_y}{\sigma_x} = b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$

$x = X - \bar{X}$ $\qquad \bar{X} = \frac{\Sigma X}{N}, \ \bar{Y} = \frac{\Sigma Y}{N}$

$y = Y - \bar{Y}$

Problems:

1, Determine the equation of a straight line which fits best data

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

Sol:

X	Y	X^2	Y^2	XY
10	10	100	100	100
12	22	144	484	264
13	24	169	576	312
16	27	256	729	~~384~~ 432
17	29	289	841	493
20	33	400	1089	660
25	37	625	1369	925
$\Sigma X = 113$	$\Sigma Y = 182$	$\Sigma X^2 = 1983$	$\Sigma Y^2 = 5188$	$\Sigma XY = 3186$

The straight line equation for Y on X is

$$Y = a + bX \quad \text{—— (1)}$$

The normal equations are

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

$$182 = 7a + 113b$$

$$3186 = 113a + 1983b$$

$$a = 0.79, \quad b = 1.56$$

From (1)

$$\boxed{Y = 0.79 + 1.56X}$$

2, A panel of 2 judges P and Q graded 7 dramatic performance by independently awarding marks as follows

Marks of P	46	42	44	40	43	41	45
Marks of Q	40	38	36	35	39	37	41

The 8th performance which judge Q would not attend was awarded 37 marks by judge P. If judge Q has also been present how many marks would be expected who have been awarded by him for the 8th performance.

04-11-2023

Sol:

x	$X = x - \bar{x}$	y	$Y = y - \bar{Y}$	XY	X^2	Y^2
46	3	40	2	6	9	4
42	-1	38	0	0	1	0
44	1	36	-2	-2	1	4
40	-3	35	-3	9	9	9
43	0	39	1	0	0	1
41	-2	37	-1	2	4	1
45	2	41	3	6	4	9
$\Sigma x = 301$	$\Sigma X = 0$	$\Sigma y = 266$	$\Sigma Y = 0$	$\Sigma XY = 21$	$\Sigma X^2 = 28$	$\Sigma Y^2 = 28$

$$\bar{x} = \frac{\Sigma x}{N} \qquad \bar{Y} = \frac{\Sigma y}{N}$$

$$\bar{x} = \frac{301}{7} = 43 \qquad \bar{Y} = \frac{266}{7} = 38$$

Regression equation of Y on x.

$$Y - \bar{Y} = r \frac{\sigma y}{\sigma x} (x - \bar{x}) \text{ —— } (1)$$

$$r \frac{\sigma y}{\sigma x} = \frac{\Sigma XY}{\Sigma X^2} = \frac{21}{28} = 0.75$$

From (1)

$$Y - 38 = 0.75 (x - 43)$$

$$Y = 0.75x - 32.25 + 38$$

$$Y = 0.75x + 5.75$$

When $x = 37 \Rightarrow Y = 0.75(37) + 5.75$

$Y = 33.5$

Deviation taken from assumed frequency if the actual mean is a fraction this method is used.

* Regression equation of y on x.

$$y - \bar{y} = \frac{r\sigma y}{\sigma x}(x - \bar{x})$$

where $x - \bar{y} = r\frac{\sigma y}{\sigma x}(x - \bar{x})$

Problems:

1) price index of cotton and wool are given below for the 12 months of a year obtained the regression equation of lines between the two index

x	78	77	85	88	87	82	81	77	76	83	97	93
y	84	82	82	85	89	90	88	92	83	89	98	99

Sol:

x	$dx = x - A$ (A = 84)	y	$dy = y - A$ (A = 88)	dx^2	dy^2	$dxdy$
78	-6	84	-4	36	16	24
77	-7	82	-6	49	36	42
85	+1	8.2	-6	1	36	-6
88	4	85	-3	16	9	-12
87	3	89	1	9	1	6
82	-2	9.0	2	4	4	
81	-3	8.8	0	9	0	
77	-7	92	4	49	16	
76	-8	83	-5	64	25	
83	-1	8.9	-1	1	1	
97	13	98	10	169	100	
93	9	99	11	81	121	
Σx =						

Regression equation of x on Y.

$$x - \bar{x} = \frac{r\sigma x}{\sigma y}(y - \bar{y}) \quad \text{—— (1)}$$

where $$\frac{r\sigma x}{\sigma y} = \frac{\Sigma dx\,dy - \frac{\Sigma dx\,\Sigma dy}{N}}{\Sigma dy^2 - \frac{\Sigma (dy)^2}{N}}$$

$$\frac{r\sigma x}{\sigma y} =$$

Sampling distribution:

Population: It is the aggregate or totality or statistical data forming a subject of investigation

Ex: Heights of India or Nationalized banks in India

Note: The number of observation in the population is defined to be the size of the population.

* It is denoted by 'N'.
* It may be finite or infinite.

Sampling: Most of the time study of entire population may not be possible to carry out and hence a part alone is selected from the given population

* A portion of the population which is examine with a view to determine a population characteristic is called sample.
* A sample is a subset of population and number of objects in the sample is called size of the sample and denoted by n

Ex: Cars produced in India is the population and NANO cars comes under sample.

Classification of samples:

* The samples are classified into two ways.

1) large sample 2) small sample

Large sample: If the size of the sample i.e if $n >= 30$ the sample is said to be large sample

Small sample: If the size of the sample i.e if $n <= 30$ the sample is said to be small sample

Note:

* The number of samples with replacement (infinite is N^n)
* The number of samples without replacement (finite) is ^NC_n

Simple mean: If $x_1, x_2, \ldots x_n$ represent a random sample of size n then the simple mean is defined as $\bar{x} = \frac{\Sigma x}{n}$

Simple variance: If $x_1, x_2 \ldots x_n$ represents a random sample of size n then the simple variance is defined as

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

Standard error: The sampling distribution of a statistic is known as it's standard error and it is denoted by S.E.

$$S.E = \frac{\sigma}{\sqrt{n}}$$

Central limit theorem: If $\bar{x}$ is the mean of the sample and 'n' is the size of the sample drawn from a population mean with a mean 'μ' under standard deviation σ, then the standardised simple mean is

01, 07-11-2023

defined as $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Correction factor (C.F)

$$CF = \frac{N-n}{N-1}$$

Problem

1, What is the value of Correction factor if $n = 5$, $N = 200$

Sol:

$$Cf = \frac{N-n}{N-1}$$

$$= \frac{200-5}{200-1} = \frac{195}{199}$$

$$\boxed{CF = 0.979}$$

2, How many different samples of size 2 can be chosen from a finite population of size 25.

Sol: $N = 25$, $n = 2$.

Number of samples $N_{C_n} = {}^{25}C_2$

$= 300$ ways

1, A population consists of 5 numbers, 2, 3, 6, 8 and 11 consider all possible samples of size 2 which can be drawn with replacement from the population. Find

(i) Mean of the population

(ii) Standard deviation of the population

(iii) Mean of the sampling distribution of means

(iv) S.D of the sampling distribution

Sol: Given population are 2, 3, 6; 8, 11, $N = 5$, $n = 2$

(i) $\mu = \frac{2+3+6+8+11}{5}$

$= 6$

(ii) $\sigma^2 = \sum_{i=1}^{5} \frac{(x-\bar{x})^2}{N}$

$$= \frac{(2-6)^2 + (3-6)^2 + (8-6)^2 + (6-6)^2 + (1 \cdot 6)^2}{5}$$

$= 10.8$

(ii) $\sigma = \sqrt{10.8} = 3.28$

(iii) The number of samples with replacement is

$N^n = 5^2 = 25$ ways

(2, 2) (2, 3) (2, 6) (2, 8) (2, 11) (3, 2) (3, 3) (3, 6) (3, 8)
(3, 11) (6, 2) (6, 3) (6, 6) (6, 8) (6, 11) (8, 2) (8, 3) (8, 6)
(8, 8) (8, 11) (11, 2) (11, 3) (11, 6) (11, 8) (11, 11)

The mean of sample are

		4	5	6.5	2.5	
2	2.5	4	5	6.5		
3	4.5	5.5	7	4	4.5	
6	7	8.5	5	5.5	7	
8	9.5	6.5	7	8.5	9.5	11

The mean of the standard deviation of mean is

$$\bar{x} = \frac{2 + 2.5 + 4 + \dots + 11}{25}$$

$$= 6$$

(iv) $\sigma^2 = \dfrac{\sum_{i=1}^{5} (x-\bar{x})^2}{n}$

$$= \frac{(2-6)^2 + (2.5-6)^2 + \dots + (11-6)^2}{25}$$

$\sigma^2 = 5.5 \Rightarrow \sigma = \sqrt{5.5}$

$\sigma = 2.4$

2, A population consists of 5, 10, 14, 18, 13, 24. Consider the possible samples of size 2 which can be drawn without replacement. Find

(i) Mean

(ii) S.D of population

(iii) Mean of the sampling distribution of mean

(iv) S.D of the sampling distribution.

Given population are 5, 10, 14, 18, 13, 24, $N = 6$, $n = 2$

$$\mu = \frac{5 + 10 + 14 + 18 + 13 + 24}{6}$$

$\mu = 14$

(ii) $\sigma^2 = \sum_{i=1}^{6} \dfrac{(x-\bar{x})^2}{N}$

$$= \frac{(5-14)^2 + (10-14)^2 + (14-14)^2 + (18-14)^2 + (13-14)^2 + (24-14)^2}{6}$$

$\sigma^2 = 35.6$

$\sigma = \sqrt{35.6} = 5.97$

(iii) The number of samples without replacement is

$N_{C_n} = {}^6C_2 = 15$ ways

(5, 10) (5, 14) (5, 18) (5, 13) (5, 24) (10, 14) (10, 18)

(10, 13) (10, 24) (14, 18) (14, 13) (14, 24) (18, 13) (18, 24)

(13, 24)

The mean of sample are

7.5 9.5 11.5 9 14.5 12 14 11.5 17 16 13.5

18 15.5 21 18.5

~~$\bar{x} = 7$~~

The mean of the standard deviation of mean is

$$\bar{x} = \frac{7 + 7.5 + 11.5 + 9 + \dots + 18.5}{15} = 14$$

07-11-2013

(iv) $\sigma^2 = \dfrac{\sum_{i=1}^{6} (x - \bar{x})^2}{n}$

$\sigma^2 = \dfrac{(7.5-14)^2 + (9.5-14)^2 + (11.5-14)^2 + \ldots\ldots (18.5-14)^2}{15}$

$\sigma^2 = 14.26 \Rightarrow \sigma = \sqrt{14.26}$

$\sigma = 3.77$

3, The variance of a population is 2. The size of the sample collected from the population is 169. what is the ~~S.D~~ S.E?

Sol: $\sigma^2 = 2 \Rightarrow \sigma = \sqrt{2} = 1.414$

$n = 169$

$S.E = \dfrac{\sigma}{\sqrt{n}}$

$= \dfrac{\sqrt{2}}{\sqrt{169}} = 0.108$

08-11-2013

4, A random sample of size 100 is taken from an infinite population having the mean $\mu = 76$ and the variance $\sigma^2 = 256$. what is the probability that $\bar{x}$ will be between 75 and 78?

Sol: $\mu = 76,\ \sigma^2 = 256,\ \sigma = 16,\ n = 100,\ \bar{x}_1 = 75,\ \bar{x}_2 = 78$

we know that

$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

when $\bar{x}_1 = 75$

$z_1 = \dfrac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}}$

$z_1 = \dfrac{75 - 76}{16/\sqrt{100}} = \dfrac{-1}{16/10}$

$z_1 = -0.625 < 0$

when $\bar{x}_2 = 78$

$z_2 = 1.25 > 0$

$P(75 < \bar{x} < 78) = |A(z_2) + A(z_1)|$

$= |A(1.25) + A(-0.625)|$

$= |A(1.25) + A(0.625)|$

$= |A(1.2 + 0.05) + A(0.6 + 0.025)|$

$= |0.3944 + 0.2324| = 0.6268$

5, A random sample of size 64 taken from a normal population with a mean $\mu = 51.4$ and $\sigma = 6.8$. What is the probability that the mean of the sample will be

(i) Exceed 52.9

(ii) Falls between 50.5 and 52.3

(iii) less than 50.6.

Sol: $n = 64$, $\mu = 51.4$, $\sigma = 6.8$,

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

$z_1 = \dfrac{52.9 - 51.4}{6.8/\sqrt{64}} = 1.76 > 0$

$P(z > z_1) = 0.5 - A(z_1)$

$= 0.5 - A(1.76)$

$= 0.5 - A(1.7 + 0.06)$

$= 0.5 - 0.4608$

$= 0.0392$

(ii) $z_1 = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

when $x_1 = 50.5$

$z_1 = \dfrac{50.5 - 51.4}{6.8/\sqrt{64}}$

$z_1 = -1.05 < 0$

$z_2 =$

when $x_2 = 52.3$

$z_2 = \dfrac{52.3 - 51.4}{6.8/\sqrt{64}}$

$z_2 = 1.05 > 0$

$P(50.5 < z < 52.3) = |A(z_2) + A(z_1)|$

$= |A(1.05) + A(-1.05)|$

$= |0.3531 + 0.3531|$

$= 0.7062$

(iii) $z_1 = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$= \dfrac{50.6 - 51.4}{6.8/8}$

$= -0.94$

$= 0.5 - A(z_1)$

$= |0.5 - 0.3264|$

$= 0.1736$

6, What is the effect on standard error if sample is taken from infinite population of sample size is increased from 400 to 900

Sol:

$n_1 = 400 \quad , \quad n_2 = 900$

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$SE_1 = \frac{\sigma}{\sqrt{n_1}} = \frac{\sigma}{\sqrt{400}} = \frac{\sigma}{20}$$

$$SE_2 = \frac{\sigma}{\sqrt{n_2}} = \frac{\sigma}{\sqrt{900}} = \frac{\sigma}{30}$$

$$SE_1 = \frac{3}{2} SE_2 .$$

Estimation :

Estimate : To find an unknown population parameter or judgement or a statement is made which is an estimate.

Estimator : The method or rule to determine an unknown population parameter is called estimator.

* The estimate can be done in 2 ways

1) Point estimation
2) Interval estimation

Maximum error of estimate : The maximum error of estimate is $E_{max} = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Sample size (when mean is given):

$$n = \left[\frac{Z_{\alpha/2}\,\sigma}{E_{max}}\right]^2$$

Sample size (when proportion is given):

$$n = \left[\frac{Z_{\alpha/2}}{E_{max}}\right]^2 PQ$$

where P = success of the proportion
Q = Failure of the proportion

* Maximum error $E_{max} = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

Confidence interval estimate of parameter:

* In an interval estimation of the population parameter (θ), if we can find two quantities t_1 and t_2 based on a sample observation drawn from the population such that the unknown parameter θ is included in the interval $[t_1, t_2]$ it in a specified percentage of cases then this interval is called a confidence interval for the parameter θ.

Confidence limits:

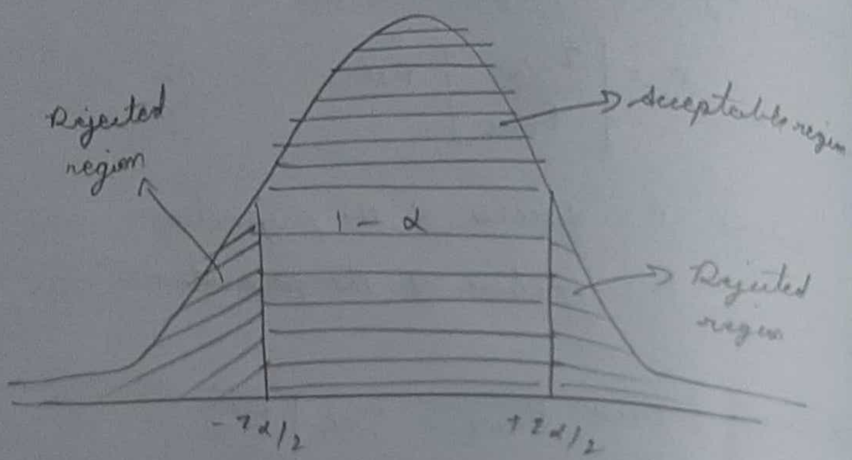
1) 95%. Confidence limits are 1.96 i.e., $Z_{\alpha/2} = 1.96$

2) 99%. Confidence limits are 2.58 i.e., $Z_{\alpha/2} = 2.58$

3) 98%. Confidence limits are 2.33 i.e, $Z_{\alpha/2} = 2.33$

4) 90%. Confidence limits are 1.64 i.e, $Z_{\alpha/2} = 1.64$

$$E_{max} = Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$



Confidence Intervals:

$$C.\ Interval = \left(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Problems:

1) In a study of automobile insurance a random sample of 80 bodies body repair cost had a mean of rupees 472.36 And a standard deviation of rupees 62.35. If $\bar{x}$ is used as the point estimate to the true average repair cost with what confidence we can assert that the maximum error does not exceed rupees 10.

Sol:

$\bar{x} = 472.36$, $\sigma = 62.35$, $n = 80$, $E_{max} = 10$.

Confidence interval (C.I) = ?

$$E_{max} = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$10 = Z_{\alpha/2}\frac{62.35}{\sqrt{80}}$$

$$Z_{\alpha/2} = \cancel{62}\ \frac{\sqrt{80}\times 10}{62.35}$$

$$Z_{\alpha/2} = 1.43$$

$$\frac{\alpha}{2} = 0.4236 \quad [\because \text{From the normal distribution table}]$$

$$\alpha = 0.8472$$

∴ The C.L Confidence levels C.L [i.e. 1 − α] = 84.72%.

2, What is the size of the smallest sample required to estimate an unknown proportion within a maximum error 0.06 with atleast 95% Confidence.

Sol:

$E_{max} = 0.06$

$n = ?$

$Z_{\alpha/2} = 1.96$ [∵ 95%]

$P = \frac{1}{2}$ [∵ If proportion is not given then it is $\frac{1}{2}$]

$Q = \frac{1}{2}$

$$n = \left[\frac{Z_{\alpha/2}}{E_{max}}\right]^2 PQ$$

$$n = \left[\frac{1.96}{0.06}\right]^2 \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$$

$n = 266.77$

$n = 267$ [only sample size should be a whole number]

3, Assuming that $\sigma = 20$, how large a random sample be taken to assert with the probability 0.95 that the sample mean will not differ from the true mean by 3 point max error.

Sol:

$n = ?$, $\sigma = 20$, $E_{max} = 3$, $Z_{\alpha/2} = 1.96$

$$n = \left[\frac{Z_{\alpha/2}\,\sigma}{E_{max}}\right]^2$$

$$n = \left[\frac{1.96 \times 20}{3}\right]^2 = 170.73$$

$n = 171$

4, A random sample of size 100 has a standard deviation 5. What can you say about the maximum errors with 95% Confidence.

Sol:

$E_{max} = ?$, $Z_{\alpha/2} = 1.96$ [∵ 95%]

$n = 100$, $\sigma = 5$

$$E_{max} = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$= 1.96 \times \frac{5}{\sqrt{100}}$$

$$= 0.98$$

5, A random sample of size 81 was taken whose variance is 20.25 and the mean is 32. Construct 98% Confidence interval.

Sol:

$n = 81$, $\bar{x} = 32$, $\sigma^2 = 20.25 \Rightarrow \sigma = 4.5$

$Z_{\alpha/2} = 2.33$

$$C.I = \left(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$= \left(32 - \left(2.33 \times \frac{4.5}{\sqrt{81}}\right),\ 32 + \left(2.33 \times \frac{4.5}{\sqrt{81}}\right)\right)$$

$$= [30.85,\ 33.13]$$

11-11-2019

6. Find the mean values of x and y and Correlation Coefficient of Correlation from the following regression equation

$2y - x = 50, \quad 3y - 2x = 10$

?

Sol: Given regression lines of y on x are

$2y - x = 50 \quad \text{—} (1), \qquad 3y - 2x = 10 \quad \text{—} (2)$

$x = 130, \quad y = 90$

$\bar{x} = 130, \quad \bar{y} = 90$

Rewrite the equation (1) and (2)

From (1) $\Rightarrow \quad y = \frac{x}{2} + 25$

From (2) $\Rightarrow \quad x = \frac{3}{2} y - 5$

$r \frac{\sigma y}{\sigma x} = \frac{1}{2}$

$r \frac{\sigma x}{\sigma y} = \frac{3}{2}$

$r^2 = \frac{3}{4}$

$r = 0.86$

Hypothesis: There are many problem in which rather than estimating the value of a parameter we need to decide whether to accept or reject a statement about the parameter

* This statement is called hypothesis and the decision making process about the hypothesis is called testing of hypothesis

* A drug chemist is to decide whether a newly drug is really effective in curing a disease.

* A quality control manager is to determine whether the process is working properly.

* There are two types of hypothesis

1) Null hypothesis (H_0): A null hypothesis is the hypothesis which assist that there is no significance to difference between the statistics and the population parameter and whatever observed, difference is there merely due to fluctuation in a sampling from the sample population

2) Alternative hypothesis (H_1): Any hypothesis which contradicts the null hypothesis is called Alternative Hypothesis.

* It is denoted by H_1 and it is defined as

a) $H_1 : \mu \neq \mu_0$

b) $H_1 : \mu > \mu_0$ — Right tailed

c) $H_1 : \mu < \mu_0$ — left tailed

} one-tailed

14-11-2019

Level of significance: The level of significance is denoted by α is the confidence with which we reject or accept the null hypothesis (H_0).

* The level of significance is generally specified by some certain levels.

$\alpha = 5\%$ (95% confidence)

$\alpha = 10\%$ (90% confidence)

$\alpha = 1\%$ (99% confidence)

Note: If the level of significance is not mentioned then by default it is considered as 5%.

Errors of sampling:

The object in sampling theory is to draw valid inference about the population parameter on the basis of the sample result.

* In a practice we decide to accept or reject after examining a sample.

* There are two types of errors:

1) **Type I error**: Reject H_0 when it is true i.e. if the null hypothesis H_0 is true but it is rejected by the test procedure then the error made is called Type I error.

2) **Type II error**: Accept H_0 when it is wrong i.e. accept H_0 when H_1 is true if the H_0 is false but it is accepted by the test procedures, then the error committed is called Type II error.

Critical values:

	Level of significance		
	1% (0.01)	5% (0.05)	10% (0.1)
Two tailed Test	$\lvert Z_2\rvert = 2.58$	$Z_2 = 1.96$	$Z_2 = 1.64$
Right tailed Test	$Z_2 = 2.33$	$Z_2 = 1.64$	$Z_2 = 1.28$
Left tailed Test	$Z_2 = -2.33$	$Z_2 = -1.64$	$Z_2 = -1.28$

Procedure for testing of Hypothesis:

Step-1: Null hypothesis (H_0): Define or setup a null hypothesis (H_0) taking into consideration, the nature of the problem and the data involved.

Step 2: Alternative hypothesis (H_1): setup the alternative hypothesis so that we could decide whether we should use one tailed or two-tailed test.

step-3: level of significance (α): Select the appropriate level of significance (α) usually we chose 5%. level of significance.

step-4: Test of statistics (z test): Compute the test of statistics under the null of hypothesis

step-5: Conclusion :i) If $|z| < z_\alpha$, H_0 is accepted

(ii) If $|z| > z_\alpha$, H_0 is rejected

$|z|$ = Calculated value

$|z_\alpha|$ = Tabular value.

Test of significance for large sample (when sample mean is not given):

step-1: H_0

step-2: H_1

step-3: α

step-4: Test of statistics $z = \frac{x-\mu}{\sigma}$

step-5: Conclusion.

problems:

1) If a coin is tossed 960 times and returned head 183 times. Test the hypothesis that the coin is unbiased.

sol: Given, $n = 960$ ($n > 30$ large sample)

$x = 183$

$p = \frac{1}{2}$, $q = \frac{1}{2}$

$\mu = np = 960\left(\frac{1}{2}\right) = 480$

$\sigma = \sqrt{npq}$

$= \sqrt{960\left(\frac{1}{4}\right)}$

$= 15.49$

$\alpha = 5\%$

step I : Null Hypothesis H_0 : Coin is unbiased

step II : Alternative Hypothesis H_1 : Coin is biased

step III : level of significance - $\alpha = 5\%$ at $z_\alpha = 1.96$

step IV : Test of statistics $z = \frac{x-\mu}{\sigma} = \frac{180-480}{15.49}$

$= -19.17$

step V : Conclusion: $|z| > z_\alpha$ H_0 is Rejected

14-11-24

2, A dice is tossed 960 times and it falls with 5 upwards 184 times. Is the dice unbiased at the level of significance 1%.

Sol: Given, $n = 960$ ($n > 30$ large samples)

$x = 184$

$p = \frac{1}{6}$, $q = \frac{5}{6}$

$p + q = 1$

$\frac{1}{6} + q = 1$

$q = 1 - \frac{1}{6}$

$q = \frac{5}{6}$

$\mu = np = \overset{160}{\cancel{960}}\left(\frac{1}{\cancel{6}}\right) = 160$

$\sigma = \sqrt{npq} = \sqrt{\frac{\cancel{160} \times 5}{\cancel{6}}}$

$\sigma = \sqrt{\frac{\overset{80}{\cancel{160}} \times 5}{\cancel{6}_3}} = \sqrt{\frac{400}{3}}$

$\sigma = 11.54$

$\alpha = 1\%$

Step-I: H_0 : dice is unbiased

Step-II: H_1 : dice is biased

Step-III: $\alpha = 1\%$

Testing of Hypothesis for a single mean (large samples):

Step-I: Null hypothesis $H_0 : \mu = \mu_0$ (Two tailed)

Step-II: Alternative hypothesis $H_1 : \mu > \mu_0$ (Right tailed)

Step-III: level of significance (α) : 5%. 10%. 1%.

By default 5% is used when α is not given.

Step-IV: Testing of statistics $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

Step-V: Conclusion: 1) If $|z| < z_\alpha$, H_0 is accepted

2) If $|z| > z_\alpha$, H_0 is rejected

Problem:

1, According to the norms established for a mechanical aptitude test persons who are 18 years old have an average height of 73.2 with a standard deviation of 18.6. If 40 randomly selected person of that age average is 76.7 test the hypothesis as $\mu = 73.2$ and again its alternative hypothesis $\mu > 73.2$ at the level of significance 1%.

Sol: Given, $n = 40$, $\bar{x} = 76.7$, $\mu = 73.2$, $\alpha = 1\%$.

1) $H_0 : \mu = \mu_0 (= 76.7)$

2) $H_1 : \mu > \mu_0 (= 76.7)$ [Right tailed test]

3) $\alpha = 1\%$ i.e. $z_\alpha = 2.33$

18-11-2023

4) $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{76.7 - 73.2}{8.1 / \sqrt{40}} = 2.5$

5) $|Z| > Z_\alpha$

$\therefore$ H_0 is rejected

2) A sample of 64 students have a mean weight of 70 kgs. Can this be regarded as a sample mean from a population mean with weight 56 kgs and a standard deviation 28 kgs.

Sol: Given, $n = 64$, $\bar{x} = 70$, $\mu = 56$, $\alpha = 5\%$, $\sigma = 28$

1) $H_0 : \mu = \mu_0 = 70$

2) $H_1 : \mu \neq \mu_0$ [Two tailed test]

3) $\alpha = 5\%$ i.e. $Z_\alpha = 1.96$

4) $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = 4$

5) $|Z| > Z_\alpha$

$\therefore$ H_0 is rejected.

3) An ambulance service claims that it takes an average less than 10 min to reach its destination in emergency class. A sample of 36 calls has a mean of 11 mins and the variance of 16 mins. Test the level of significance at 5%.

Sol: Given, $n = 36$, $\bar{x} = 11$, $\mu = 10$, $\alpha = 5$, $\sigma^2 = 16$, $\sigma = 4$

1) $H_0 : \mu = \mu_0$

2) $H_1 : \mu < \mu_0$ [Left tailed test]

3) $\alpha = 5\%$ i.e. $Z_\alpha = -1.64$

4) $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{11 - 10}{4 / \sqrt{36}} = 1.5$

5) $|Z| < Z_\alpha$

$\therefore$ H_0 is ~~rejected~~ accepted.