STATISTICS

18/04/2021

→ eg- Budget

Data: A number in either ascending or descending order.

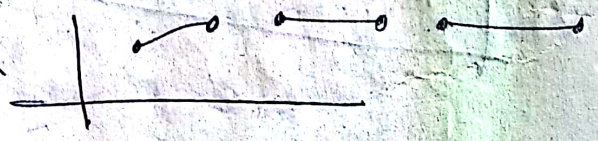
Class interval: lower value, upper value

Def: Statistics is defined as a branch of science, dealing with collection of data, organising, summarising, presenting and analysing data and drawing a valid conclusion. And thereafter making reasonable decisions on the basis of such analysis.

eg→ Analysing the datas we can conclude that weather rain will occur or not.

Discrete data:

eg There is a break



Continuous data:

eg

Continuously	0-10	5
	10-20	8
	20-30	4

CURVE FITTING:

The general problem of finding equations of approximating curves that closely fit the given data is called curve fitting.

In other words, curve-fitting is the representation of relationship b/w the variables two by means of an algebraic expression.

→ This type of suggestion first given by Scatter.

→ And the diagram is known as scattered diagram

→ Calculation of scattered diagram by Li-square method.

* The curve passes through maximum no. of points from scattered diagram, more than one curve may be seen to be appearing to the given set of data and this may largely be due to human judgement.

* To achieve this method of Li-square is used

* Let $P(x_i, y_i)$ be one of the points, of the plotted diagram

* from P draw $\perp^r$ PM on the X axis

* Let it meet the curve at M'

* The distance $PM' = PM - MM'$

★ PN is called the error of the estimate or residual or deviation.

★ This may be positive or zero or negative, depending upon whether the point is above, on or below the curve.

★ Similar residuals may be obtained for (n-1) remaining observations.

★ Let the residuals be denoted by $e_1, e_2, e_3, \ldots, e_n$.

★ Sum of the squares of the residuals, i.e. $e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2$

$$\sum e_i^2$$

★ $\sum e_i^2$ is called the sum of the squares of the residuals.

★ The line for which $\sum e_i^2$ is the least is called the line of best fit.

* This method is known as Legendre's method of least squares.

The lines of best fit :

Let $y = a + bx$ be the lines of best fit.

Let $(x_1, y_1) \ldots (x_n, y_n)$ be n points through which it passes.

If these lines will be lines of best fit then

$$\sum_{i=1} e_i^2 = E = \sum (y_i - a - bx_i)^2$$

* E to be maxima or minima :

By using principle of least square :

Partial derivative of E wrt a and b must vanish separately

$$\frac{\partial E}{\partial a} = -2 \sum (y_i - a - bx_i) = 0$$

$$\frac{\partial E}{\partial b} = -2 \sum (x_i)(y_i - a - bx_i) = 0$$

$$\sum y_i - a - bx_i = 0$$

$$\sum x_i (y_i - a - bx_i) = 0$$

$$\Rightarrow \boxed{\begin{aligned} \sum y_i &= na + b \sum x_i \\ \sum x_i y_i &= a \sum x_i + b \sum x_i^2 \end{aligned}}$$

Q1

$y = ax + b$

$\Sigma y = a\Sigma x + nb$

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

Q2

$y = a + bx + cx^2$

$E = \Sigma (y_i - a - bx - cx^2)^2$

$\frac{\partial E}{\partial a} = -2\Sigma (y_i - a - bx - cx^2) = 0$

$\frac{\partial E}{\partial b} = -2\Sigma x (y_i - a - bx - cx^2) = 0$

$\frac{\partial E}{\partial c} = -2\Sigma x^2 (y_i - a - bx - cx^2) = 0$

$\Sigma y = na + b\Sigma x + c\Sigma x^2$

$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$

$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$

Q3 $y = ax + bx + c$

$\Sigma y = an + b\Sigma x + nc$

$\Sigma xy = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^2$

$\Sigma x^2 y = a\Sigma x^3 + b\Sigma x^4 + c\Sigma x^3$

Fit a straight line $y = a + bx$ by method of least squares from the following data.

x	0	1	2	3	4
y	1	3	2	5	4
xy	0	3	4	15	16
x^2	0	1	4	9	16

$\Sigma x = 10$, $\Sigma y = 15$, $\Sigma xy = 38$, $\Sigma x^2 = 30$

$y = a + bx$

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

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

x	y	x^2	xy
0	1	0	0
1	3	1	3
2	2	4	4
3	5	9	15
4	4	16	16
$\Sigma x = 10$	$\Sigma y = 15$	$\Sigma x^2 = 30$	$\Sigma xy = 38$

$15 = 5a + 10b$ ... (1)

$38 = 10a + 30b$ ... (2)

Solving we get $a = 1.6$, $b = 0.3$

$y = 1.6 + 0.3x$

x	y	xy	x^2	x^3	x^2y	x^4
0	1	0	0	0	0	0
1	3	3	1	1	3	1
3	2	6	9	27	18	81
6	5	30	36	216	180	1296
8	4	32	64	512	256	4096
18	15	71	110	756	457	5474

$y = a + bx + cx^2$

$\Sigma y = na + b\Sigma x + c\Sigma x^2$

$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$

$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$

$15 = 5a + 18b + c\times 110$

$71 = a\,18 + b\,110 + c\,756$

$457 = 110a + 756b + 5474c$

$a = 1.34$

$b = 734.2 \times 10^{-3}$

$c = -44.92 \times 10^{-3}$

$a = 1.34$

$b = 0.734$

$c = -0.04$

$\Rightarrow y = 1.34 + 0.734x - 0.04x^2$

FITTING OF OTHER CURVES

$y = ax^b$

$\log_{10} y = \log_{10} a + b \log_{10} x$

$Y = A + BX$

$Y = \log_{10} y$

$A = \log_{10} a$

$B = b \log_{10}$

$y = a e^{bx}$

$\log_{10} y = \log_{10} a + bx \log_{10} e$

$Y = A + BX$

$B = b \log_{10} e$

combined gas equn:

$PV^r = \text{const.}$

$xy^n = b$

$\log_{10} x + n \log_{10} y = \log_{10} b$

$\log_{10} y = -\frac{1}{a}\log_{10} x + \frac{1}{a}\log_{10} A$

$X = \log_{10} x$

$Y = \log_{10} y$

$A = \frac{1}{a}\log_{10} A$

$B = -\frac{1}{a}$

Eg:2

Find the curve of best fit of the type $y = ax^b$ by method of least squares for the following data

Data given:

x	1	2	3	4
y	10	15	20	25

$Y = \log y$

$A = \log a$

$B = b \log e$

$\log_{10}$

$x = X$	$y = Y$	$Y = \log_{10} y$	y	x^2	xy
1	10	1	1	1	10
5	15	1.17	25	25	75
7	12	1.07	49	49	84
9	15	1.17	81	81	135
12	21	1.32	144	144	252
$\Sigma x = 34$	$\Sigma y = 73$	Σ 5.73	300		556

$Y = \log_{10} y$

$y = A + BX$

$\Sigma y = nA + B\Sigma x$

$\Sigma xy = A\Sigma x + B\Sigma x^2$

$5.73 = 5A + 34B$

$556 = 34A + 300B$

$A = 8.70$

$B = 0.867$

$\log_a b$

	5.0
	12.8
	17.55
	29.72
	64.96
1	
5.85	
7.49	
10.53	
15.84	
40.71	

Co-relation and Regression:

→ Co-relation is the co-variation of 2 independent magnitudes. If 2 variables x and y are related in such a way that increase or decrease in one of them corresponds to increase or decrease in the other.

→ In other words we say that the variables are positively co-related, also if increases or decrease in one of them corresponds to decrease or increase in other, then the variables are said to be negatively co-related.

→ The numerical of co-relation b/w 2 variables x and y is known as Pearson's co-efficient of co-relation, usually denoted by γ.

$$\gamma = \frac{\sum (x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$$

$$\sigma_x = \frac{\sum (x-\bar{x})^2}{n}$$

$$= \frac{\sum X^2}{n} \qquad X = x-\bar{x}$$

$$\sigma_y = \frac{\sum (y-\bar{y})^2}{n} \qquad Y = y-\bar{y}$$

$$= \frac{\sum Y^2}{n}$$

$$r = \frac{\sum fxy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}, \quad r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

→ The co-efficient of co-relation numerically doesn't exceed unity.

$$-1 \leq r \leq 1$$

→ if $r = \pm 1$, we say that x and y are perfectly co-related.

→ if $r = 0$, we say that x and y are non-co-related.

$$\rightarrow r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

Problem

Regression

→ It is an estimation of one independent variable in terms of other.

→ If x and y are co-related the best fitting straight line in the least square sense gives a reasonable good relation b/w x and y

→ The best fitting straight line of the form $y = ax + b$.

→ (x being the independent variables) is called the regression

of y on x.

→ again if x = a + by (y be the independent variable) is called regression of x on y.

→ Derivation of equ of the regression lines:

Regression line of y on x

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

Regression of x on y

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

Problem 1) Compute co-efficient of co-relation and the equ of the lines of regression from the following data.

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

We shall prepare the following columns:

Scanned by CamScanner

$$r = \frac{9 + 9 - 4/7}{2 \times 2 \times 2}$$

$r = 0.928$

$= 0.93$

Regression of y on x:

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

$= (y - 11) = 0.93\,(x - 4)$

$= y - 11 = 0.93x - 3.72$

$= 0.93x - y + 7.28$

Regression of x on y:

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

$= (x - 4) = 0.93\,(y - 11)$

$= x - 4 = 0.93y - 10.23$

$= x - 0.93y + 6.23.$

Q. Given:

$8x - 10y + 66 = 0$

$40x - 18y = 214$

are the 2 regression lines, find the means of x and y and the co-relation co-efficient.

Find σ_y? if $\sigma_x = 3$

Ans: Suppose 2 regression lines passing through the means, i.e they must satisfy the eqn.

$8\bar{x} - 10\bar{y} + 66 = 0$

$40\bar{x} - 18\bar{y} = 214$

$\bar{x} = 13, \bar{y} = 17$

$8x - 10y + 66 = 0$

$10y = 8x + 66$

$y = 0.8x + 6.6$ — y on x

$40x - 18y = 214$

$x = \frac{214}{40} + \frac{18y}{40}$ — x on y

$= 0.45y + 5.35$

$r\frac{\sigma_y}{\sigma_x} = 0.8$, $r\frac{\sigma_x}{\sigma_y} = 0.45$.

multiplying the two eqns,

$r^2 = 0.6$ ← As r^2 is always +ve.

While those of samples are denoted by $\bar{x}$ and s^2 for means & variance

Standard error: (S.E)

The standard deviation of the sampling distribution of a statistic is known as the standard error.

→ It plays an important role in the theory of large sample and it forms on the basis of testing of hypotheses.

→ for any static 't' (for large samples)

$$Z = \frac{t - E(t)}{S.E(t)}$$

is normally distributed with mean $(\bar{x}) = 0$ and $S^2 = 1$

For large samples

n → Sample size

σ^2 → population variance

s^2 → sample variance

P → population proportion

Q = 1 - P

n_1 and n_2 are sizes of two independent random samples.

Sl.No	Statistic	S.E
1	$\bar{x}$	$\sigma/\sqrt{n}$
2	s	$\sqrt{\sigma^2/2n} = \sigma/\sqrt{2n}$
3	Difference of 2 sample means i.e $\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
4	Difference of 2 sample standard dev i.e $s_1 - s_2$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
5	Difference of 2 sample proportions i.e $P_1 - P_2$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$
6	Observed sample proportions i.e p	$\sqrt{\frac{PQ}{n}}$

Test of significance :

Test of significance occurs in 2 conditions :

1) The deviation b/w the observed sample statistics and the hypothetical parameter.
2) The deviation b/w 2 samples statistics is significant due to the fluctuation of sampling

Testing of hypothesis :

Hypothesis is a definite statement ab the population parameter, for applyi test of significance, called null hyp

denoted by H_0.

Alternative hypothesis:

Any hypothesis complementary to null hypothesis is called alternate hypothesis, denoted by H_1.

We want to test null hypothesis H_0 with mean μ, under the specified condition

$H_0 : \mu = \mu_0$

Alternative hypothesis will be as follows

$H_1 : \mu \neq \mu_0$ $\begin{cases} \mu > \mu_0 \\ \text{or} \\ \mu < \mu_0 \end{cases}$

Two tailed alternative hypothesis

if $H_1 : \mu > \mu_0$

Right tailed alternative hypothesis or single tailed

if $H_1 : \mu < \mu_0$

Left tailed alternative hypothesis or single tailed.

Alternative hypothesis tells us wether the test is two tailed test or one tailed test.

Critical region:

A region corresponding to a statistics 't' in the sample space 'S' which amounts to rejection of the null hypothesis H_0, is called as critical region.

→ The region of the sample space which amounts to exceptance of H_0, is called exceptance region.

→ Critical region is also known as region of rejection.

$P(|Z| > Z_\alpha) = \alpha$

$P(Z > Z_\alpha) = \alpha$

$P(Z < -Z_\alpha) = \alpha$

level of significance:

	1% (0.01)	5% (0.05)	10% (0
Two-tailed test	$\|Z_\alpha\| = 2.58$	$\|Z_\alpha\| = 1.96$	$\|Z_\alpha\| =$
Right-tailed	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha =$
Left-tailed	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha =$

Working procedure for testing of statistical hypothesis.

Step 1) Null hypothesis:

Set up H_0.

Step 2) Alternative hypothesis

Set up H_1 so that we could decide wether we should use one tailed system or two tailed system.

Step 3) Level of significance

Select the appropriate level of significance in advance depending on the reliability of the estimate

Step 4) Test of statistics

Compute the test static:

$$z = \frac{t - E(t)}{S.E(t)}$$

Under the null hypothesis.

Step 5) Conclusion

Compare the compute value of z with critical values z of at level of significance of

If $|Z| > Z_\alpha$ we reject H_0 and conclude that there is a significant difference.

If $|Z| < Z_\alpha$ we accept H_0 and conclude that there is no significant difference.

Test of significance for large sampling:

⇒ $(n > 30)$ then large sample

⇒ Classified as:

1) Testing of significance for single proportion.
2) Testing of significance for different proportion.
3) Testing of significance for single mean.
4) Testing of significance for different mean.

1) For single proportion:

Let X be the no. of successes in 'n' independent trials with constant probability 'p' of success of each trial

$E(X) = np$

$V(X) = npq$

$q = 1 - p$ probability of failure

$p = X/n$ called observed proportion of success.

$E(p) = E(X/n)$

$= \frac{1}{n} E(X) = \frac{np}{n} = p$

$V(p) = V(X/n)$

$= \frac{1}{n^2} V(X) = \frac{pqn}{n^2} = \frac{pq}{n}$

$S.E(p) = \sqrt{\frac{pq}{n}}$

$Z = \frac{p - E(p)}{S.E(p)}$

$= \frac{p - p}{\sqrt{pq/n}}$

Z is called the test static

Note:

1) The probable limit for a the observed proportion of success are

$P = Z_\alpha \sqrt{pq/n}$

$\therefore Z_\alpha$ be the level of significance α.

2) If small 'p' is not known, the limits for the proportion in the population are

$$p \pm 3\sqrt{\frac{pq}{n}}$$

$$q = 1 - p$$

$$p \pm 2\sqrt{\frac{pq}{n}}$$

$$q = 1 - p$$

3) If α is not given so we can take safely 3σ limits.

4) Confidence limits for the observed proportion

$$p \pm 3\sqrt{\frac{pq}{n}}$$

5) Confidence limits for the population proportion

$$p = p \pm \sqrt{\frac{pq}{n}}$$

eg) A coin was tossed 400 times and the head turned up 216 times test the hypothesis that the coin is un-biased.

Sol) H_0 = The coin is un-biased

$p = 1/2 = 0.5$

H_1 = the coin is not unbiased (biased)

$p \neq 1/2 = 0.5$

$n = 400$

X = no. of success = 216

$$p(X) = \frac{216}{400} = 0.54$$

P = population proportion = 0.5

$Q = 1 - P = 0.5$

Test statistic

$$|Z| = \frac{p - P}{\sqrt{PQ/n}}$$

$$= \left| \frac{0.5 - 0.54}{\sqrt{\frac{0.5 \times 0.5}{400}}} \right|$$

$$= 1.6$$

$|Z| < Z_\alpha$

We use two tailed test

conclusion:

since $|Z| = 1.6 < 1.96$

$|Z| < Z_\alpha$, Z_α is the significant value at 5% level of significance.

27) A certain cubical die has thrown 9000 times and 5 or 6 was obtained 3240 times. On the assumption of certain throwing do the data indicate an unbiased die.

$H_0 = 2/6 = 1/3$ i.e. P

$q = 2/3$

$p = \frac{3240}{9000}$

$p = 0.36$

$H_1 \neq 1/3 \neq P$

Two tailed test.

$$|Z| = \frac{|0.36 - 1/3|}{\sqrt{\frac{1/3 \times 2/3}{9000}}}$$

$$= 6.009$$

Hypothesis is rejected as

$|Z| > Z_\alpha$

3) A manufacture claims that only 4% of product supplied by him are defective. A random sample of 600 products contains 36 defectives. Test the claim of manufacturer.

$= 0.06 = 6\%$

$P = 0.06$, $Q = 0.94$.

$p = 0.04$.

$H_0 = 0.04$

$$|Z| = \frac{p - P}{\sqrt{PQ/n}}$$

$$= \frac{0.04 - 0.06}{\sqrt{\frac{0.06 \times 0.94}{600}}}$$

$$= 2.06$$

$|Z| > |Z_\alpha|$ so two tailed test.

So, manufacture's claim is rejected.

Test of difference b/w 2 proportions:

Consider two samples X_1 and X_2 of size n_1 and n_2 respectively taken from 2 different populations. To test the significance of the difference b/w 2 sample proportions p_1 and p_2. The test static

under the null hypothesis H_0, that there is no significant difference b/w 2 sample proportions.

$$|Z| = \frac{P_1 - P_2}{\sqrt{PQ\left(1/n_1 + 1/n_2\right)}}$$

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

$$Q = 1 - P$$

eg) A machine produced 16 defective articles in a batch of 500. After overally it produces 3 defectives in a batch of 100. Has the machine improved.

$n_1 = 500$, $n_2 = 100$.

$P_1 = \frac{16}{500} = 0.032$, $P_2 = \frac{3}{100} = 0.03$

null hypothesis H_0 = $\frac{16}{500}$ $\frac{3}{100}$

i.e $P_1 = P_2$

Alternative hypothesis $H_1 : P_1 > P_2$.

i.e right tailed

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = 0.031$$

$Q = 0.96$

$$|Z| = \frac{0.032 - 0.031}{\sqrt{0.031 \times 0.96 \left(\frac{1}{500} + \frac{1}{10}\right)}}$$

$$= 0.105$$

Conclusion :

$|Z| < 1.645$

level of significance less than 5%.

accepted - machine is not approved.

less than accepted

greater than rejected.

2% 500 articles from a factory are examined and found to be 2% defective, 800 similar articles from a same factory are found to have only 1.5% defective, can it reasonably concluded that the product of the first factory are superior to those of 2nd.

$n_1 = 500$, $n_2 = 800$

$p_1 = 0.02$ $p_2 = 0.015$

$p_1 \quad p_2$

$|Z| =$

$p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$

$= \frac{10 + 12}{1300}$

$= 0.016$

(Complete later)

Test for single mean:

i.e, to test wether difference b/w sample mean and population mean is significant or not.

$|Z|$

$X_1, X_2, X_3, \ldots X_n$ are sample size from large population

$X_1, X_2, \ldots X_N$ with mean μ and variance σ^2

The standard error of mean of random sample of size 'n' from of population with variance σ^2 is

$\sigma / \sqrt{n}$

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

σ = standard deviation

If σ is not known

$$|z| = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Note:

If the level of significance is α and z_α is the critical value then

$$-z_\alpha < |z| = \left| \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right| < z_\alpha$$

limit of the population mean μ are given by

$$\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

At 5% level of significance 95% confidence limit are:

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

At 1% level of significance 99% confidence limit are given by:

$$\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}}$$

These limits are called confidence limits.

1) A Normal population has a mean of 6.8 and standard deviation of 1.5. A sample of 400 members gave a mean of 6.75, is the difference significant.

$\mu = 6.8$
$\sigma = 1.5$
$n = 400$
$\bar{x} = 6.75$

Given:

$\mu = 6.8$, $\sigma = 1.5$, $n = 400$

$\bar{x} = 6.75$

$$|Z| = \left| \frac{6.75 - 6.80}{1.5/\sqrt{400}} \right|$$

$= 2/3 = 0.67$

H_0 : there is no significant difference b/w the two.

H_1 : there is significant difference b/w $\bar{x}$ and μ

$|Z| = 0.67$

Conclusion:

$|Z| < Z_\alpha = 1.96$ 5% level of significance.

Selected H_0 accepted.

Q) The mean weight obtained from a random sample of size 100 is 64 gms. The standard deviation of the weight of the distribution of the population is 3 gms. Test the statement that the mean weight of the population is 67 gms at 5% level of significance. Also setup 99% confidence limits of the mean weight of the population.

$n = 100$

$\mu = 67 \quad (5\%)$

$\bar{x} = 64$

$\sigma = 3$

H_0 = There is no significant difference b/w sample & population mean

$H_1 = \mu \neq 67$

$$Z = \frac{|64 - 67|}{3/\sqrt{10}}$$

$$|Z| > 1.96$$

Conclusion: H_0 is rejected.
i.e sample is not drawn from the population with $\mu = 67$.

To find confidence limit.

$$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

$= 64.776$ } Any one
63.224 } of the mean.

Test of significance for difference of means of 2 large samples.

Let $\bar{x}_1$ be the mean of sample size n_1, from a population with mean μ_1, and varriance σ_1^2.

Similarly $\bar{x}_2$ be the mean of sample size n_2, from a population mean μ_2 and varriance σ_2^2.

$$|Z| = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under null hypothesis samples are drawn from the sample.

population, where $\sigma_1 = \sigma_2$ and $q_1 = q_2$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{n_2 + n_1}{n_1 \cdot n_2}}}$$

Note: If σ_1 and σ_2 are not known

$$\sigma_1 \neq \sigma_2$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

c) σ is not known and

$$\sigma_1 = \sigma_2$$

$$\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

Ques) The average income of a person was of ₹210 with a standard deviation of ₹10 in sample of 100 people of a city, for another sample of 150 persons, the average income was ₹220 with standard deviation of ₹12. The standard deviations of income of the people of the city was ₹11. Test wether there is any signif-icant difference b/w the average incomes of the localities.

$n_1 = 100$
$\bar{x}_1 = 210$
$\sigma = 10$

$n_2 = 150$
$\bar{x}_2 = 220$
$\sigma^2 = 12$

$H_0 : \bar{x}_1 = \bar{x}_2$

$H_1 : \bar{x}_1 \neq \bar{x}_2$

$$|Z|_2 = \frac{210 - 220}{\sqrt{\frac{100}{100} + \frac{144}{150}}}$$

$$= 7.14$$

Conclusion: H_0 at 5% level of significance is Rejected.

Ques 2) For sample 1 $N_1 = 1000$

$\sigma_x = 49,000$

$(\sigma_x - \bar{x})^2 = 7,84,000$

For sample 2

$N_2 = 4500$

$\sigma_x = 70,500$

$(\sigma_x - \bar{x})^2 = 24,00,000$

Discuss the significance of the difference of sample mean.

→ null hypothesis H_0 : There is no significance difference b/w sample means.

$\bar{x_1} = \bar{x_2}$

$H_1 = \bar{x_1} \neq \bar{x_2}$

To calculate the

$S_1^2 = \frac{7,84,000}{1000} = 784$

$S_2^2 = \frac{24,00,000}{4500} = 1600$

$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = 49$

$\bar{x}_2 = 47$

$Z = \frac{49 - 47}{\sqrt{\frac{784^2}{1000} + \frac{1600^2}{1500}}}$

$= 0.0047$

Accepted.

Test of significance for difference of standard deviations:

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

σ_1 and σ_2 are Population Standard deviation.

When population, standard deviation are not known,

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}}$$

Q) Random samples are drawn from 2 countries gave the following data relating to the height of the adult males.

	Country A	Country B
mean height	67.42	67.25
S.D	2.58	2.50
number of sample	1000	200

1) Is the difference b/w the means is significant.

a) Is the difference b/w the σ significant.

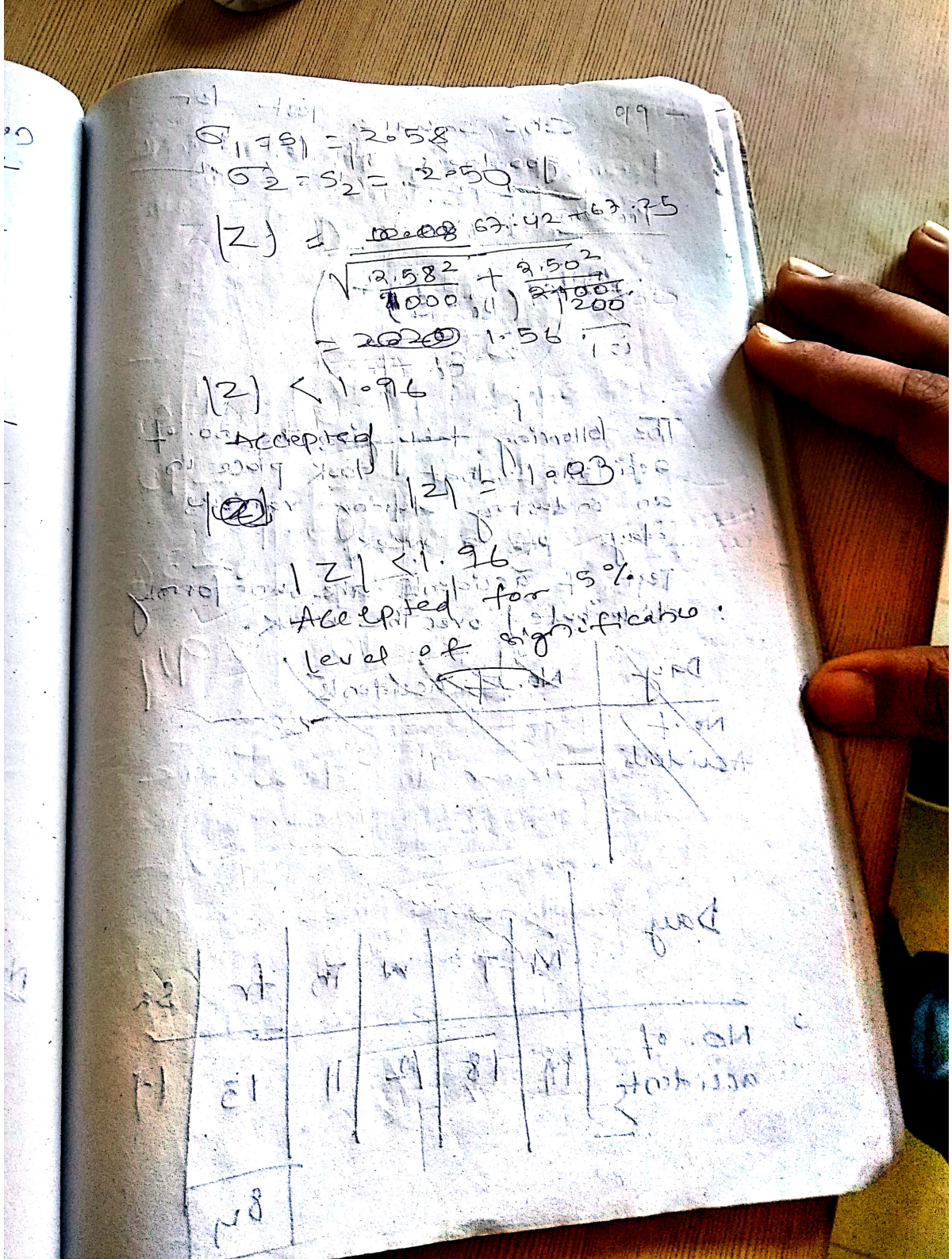
Sol?) $n_1 = 1000$, $n_2 = 1200$

$S_1 = 2.58$, $S_2 = 2.50$

$\bar{x}_1 = 67.42$, $\bar{x}_2 = 67.25$

$\sigma_1 = 2.58$, $\sigma_2 = 2.50$

Since sample size are large we can take



$\sigma_1^2 = 2.58^2$

$\sigma_2^2 = s_2^2 = 2.50^2$

$|Z| = \frac{67.42 - 63.75}{\sqrt{\frac{2.58^2}{1000} + \frac{2.50^2}{1200}}}$

$= 1.56$

$|Z| < 1.96$

Accepted

$|Z| = 1.93$

$|Z| < 1.96$

Accepted for 5% level of significance.

Chi square test for goodness of fit?

Problem

$$\sum_{i=1}^{n} \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

The following table gives the no. of accidents that took place in an industry during various days of a week. Test if accidents are uniformly distributed over the week.

Day						
No. of Accidents						

Day	M	T	W	Th	fr	Sa
No. of accidents	14	18	12	11	15	14

Sy

H_0 = Accidents are uniformly distributed over the week.

Expected frequency of the accidents on each days of the week = 84

total days = 6

expected accidents = 14

observed frequency						
O_i	14	18	12	11	15	14
E_i	14	14	14	14	14	14

$(0)^2 + 16 + 4 + 9 + 1 + 0$

$\frac{30}{14} = \frac{15}{7} = 2.14$

χ^2

D.of $= 6-1 = 5$

$n = 6$
D.of. 5

$n = 6$

χ^2 at 5% = 11.49

Test from D.of table if below 5% significance then accepted else vice-versa.

Records taken of the no. of male and female birth in 800 families having 4 children

No. of male birth	No. of f. birth	No. of families
0	4	32
1	3	178
2	2	290
3	1	236
4	0	94

Test wether the data are consistent with the hypothesis that the binomial law holds and the chance of male birth is equal to that of female birth

i.e. $p = q = 1/2$

$H_0 = P = q = 1/2$

N = total frequency

N_γ = No. of families with γ male childrens.

$P(X) = \gamma$

$P(X=\gamma) = {}^{n}C_{\gamma}\, p^{\gamma} q^{n-\gamma}$

$\gamma = 0, 1, 2, 3, 4$

$N = 830$

$N(\gamma) = N \times P(X=\gamma)$

$N(0) = 830 \times {}^{4}C_0 \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^4 = 51.8$

$N(1) = 830 \times {}^{4}C_1 \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^3 = 207.5$

$N(2) = 830 \times {}^{4}C_2 \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^2 =$

$N(3) = 830 \times {}^{4}C_3 \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^1 =$

$N(4) = 830 \times {}^{4}C_4 \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^0 =$

51.8, 207.5, 311.25, 207.5, 51.8

$$\frac{(32-51.8)^2}{51.8} + \frac{(178-207.5)^2}{207.5} + \frac{(311.5)}{311.5}$$

$$+ \frac{(236-207.5)^2}{207.5} + \frac{(94-51.8)^2}{51.8}$$

$$= 7.5 + 4.19 + 1.48 + 3.91 + 34.39$$

$$= 51.459.$$

D.O.F = 9.49

Reject