Statistics Assignment 2

Submitted by: Amit Kumar

Problem Statement 1:

In each of the following situations, state whether it is a correctly stated hypothesis

testing problem and why?

1. $H0$: $\mu = 25$, $H1$: $\mu \neq 25$ - correct , equal to and not equal to.

2. $H0$: $\sigma > 10$, $H1$: $\sigma = 10$ – correct.

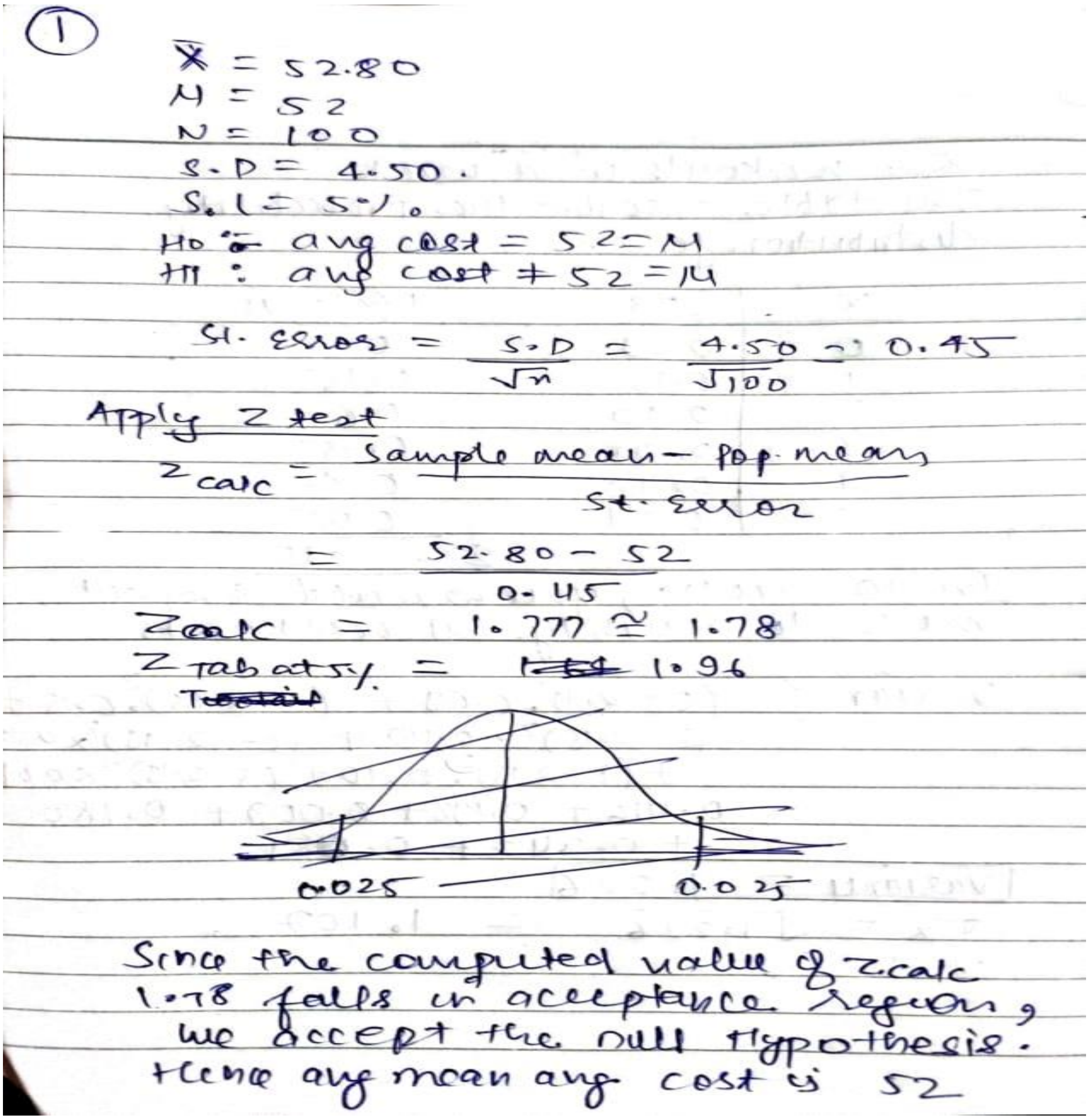
3. $H0$: $x = 50$, $H1$: $x \neq 50$ -- correct, equal to and not equal to.

4. $H0$: $p = 0.1$, $H1$: $p = 0.5$ -- incorrect, values are different for assumptions

5. $H0$: $s = 30$, $H1$: $s > 30$ ---correct

Problem Statement 2:

The college bookstore tells prospective students that the average cost of its textbooks is Rs. 52 with a standard deviation of Rs. 4.50. A group of smart statistics students thinks that the average cost is higher. To test the bookstore's claim against their alternative, the students will select a random sample of size 100. Assume that the mean from their random sample is Rs. 52.80. Perform a hypothesis test at the 5% level of significance and state your



①

$\bar{X} = 52.80$

$\mu = 52$

$N = 100$

S.D = 4.50.

S.L = 5%.

Ho : avg cost = 52 = μ

H1 : avg cost ≠ 52 = μ

$$St.\ Error = \frac{S.D}{\sqrt{n}} = \frac{4.50}{\sqrt{100}} = 0.45$$

Apply Z test

$$Z_{calc} = \frac{\text{sample mean} - \text{pop. mean}}{\text{St. Error}}$$

$$= \frac{52.80 - 52}{0.45}$$

$Z_{calc} = 1.777 \cong 1.78$

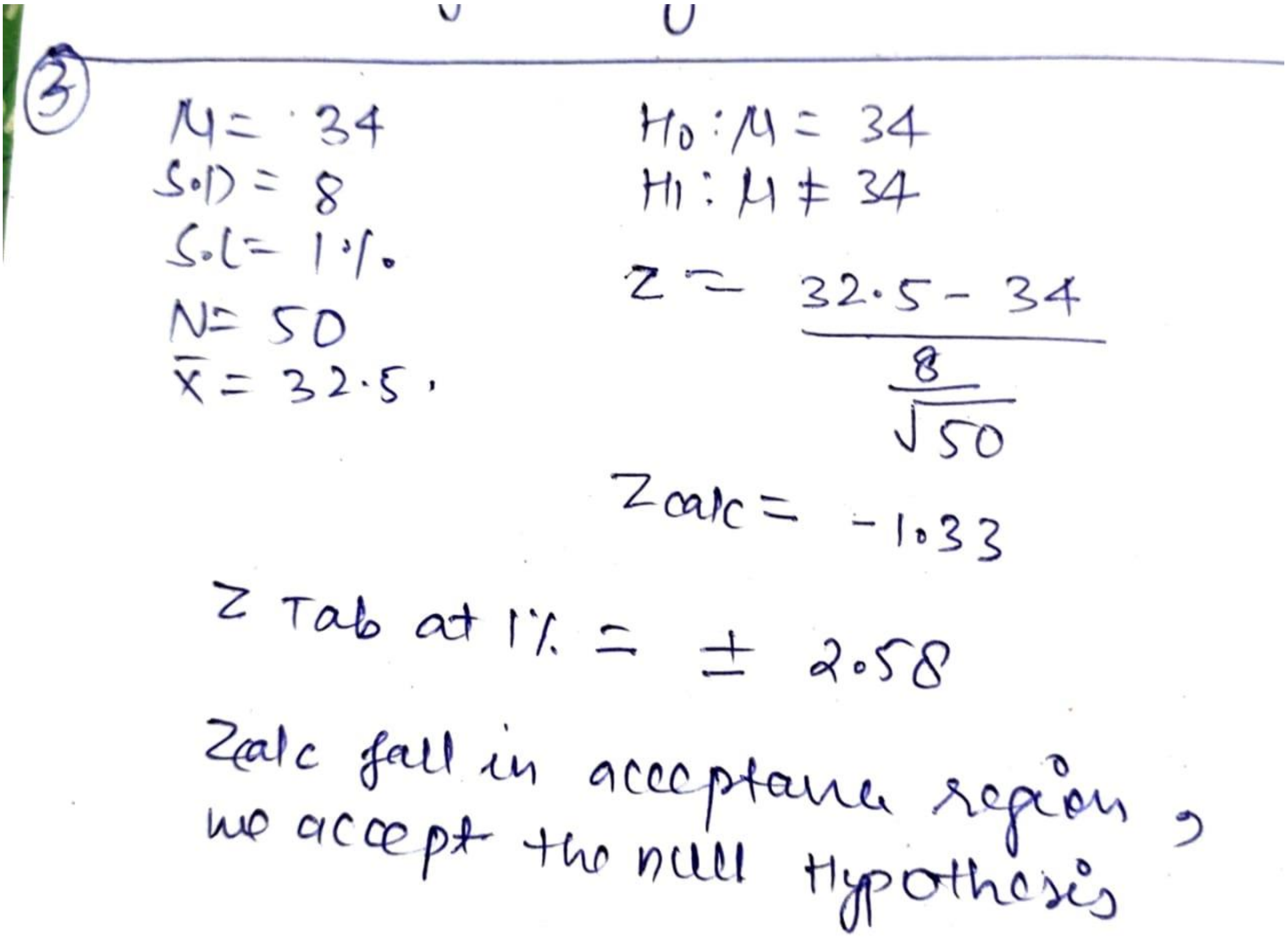
Z Tab at 5% = 1.96

0.025 0.025

Since the computed value of Zcalc 1.78 falls in acceptance region, we accept the null hypothesis. Hence avg mean avg cost is 52

Problem Statement 3:

A certain chemical pollutant in the Genesee River has been constant for several years with mean μ = 34 ppm (parts per million) and standard deviation σ = 8 ppm. A group of factory representatives whose companies discharge liquids into the river is now claiming that they have lowered the average with improved filtration devices. A group of environmentalists will test to see if this is true at the 1% level of significance. Assume \ that their sample of size 50 gives a mean of 32.5 ppm. Perform a hypothesis test at the 1% level of significance and state your decision.



(3)

$\mu = 34$

$S.D = 8$

$S.L = 1\%$

$N = 50$

$\bar{x} = 32.5$

$H_0 : \mu = 34$

$H_1 : \mu \neq 34$

$$z = \frac{32.5 - 34}{\frac{8}{\sqrt{50}}}$$

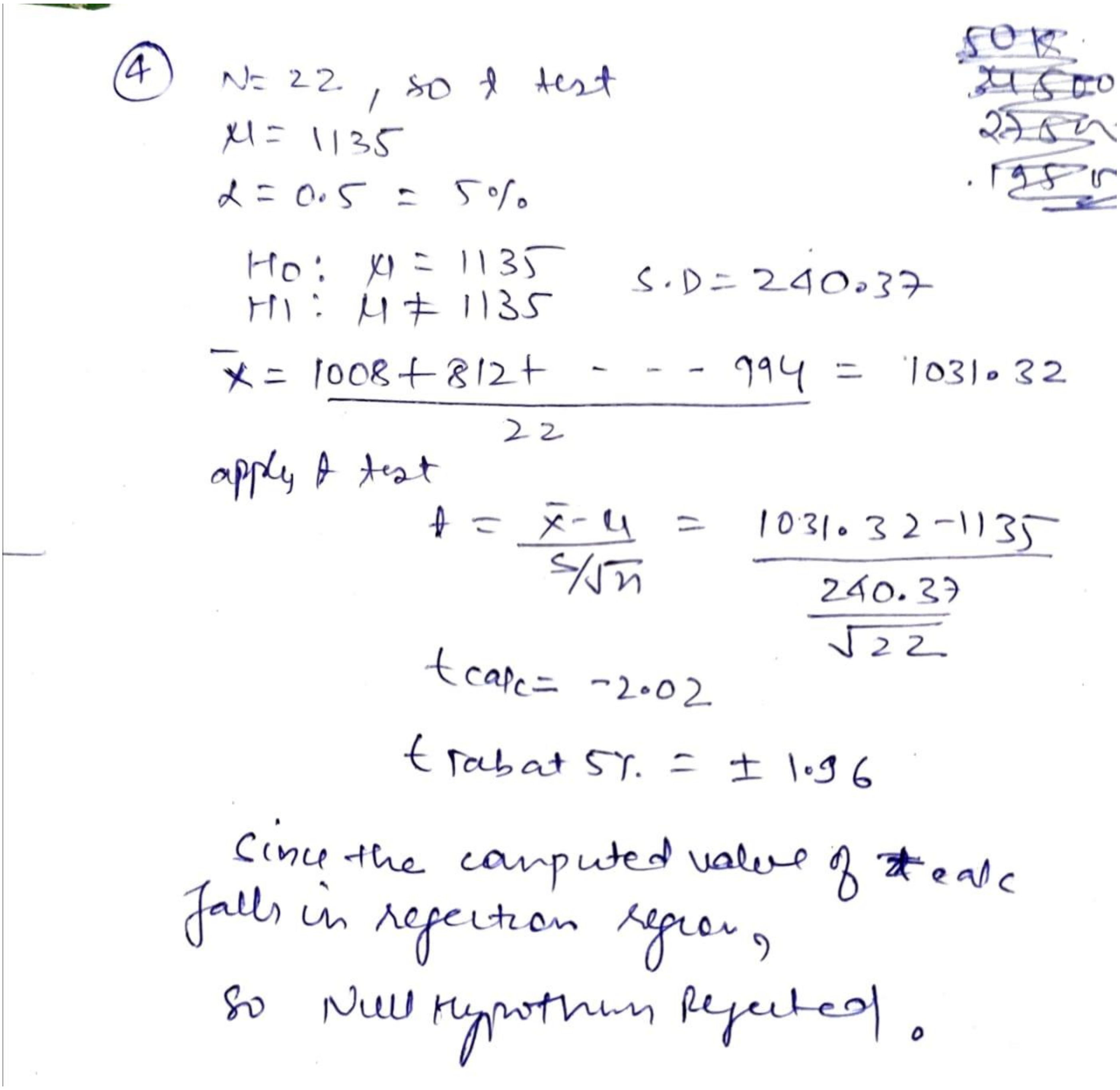
$Z_{calc} = -1.33$

Z Tab at 1% = ± 2.58

Zcalc fall in acceptanc region, we accept the null Hypothesis

Problem Statement 4:

Based on population figures and other general information on the U.S. population, suppose it has been estimated that, on average, a family of four in the U.S. spends about $1135 annually on dental expenditures. Suppose further that a regional dental association wants to test to determine if this figure is accurate for their area of country. To test this, 22 families of 4 are randomly selected from the population in that area of the country and a log is kept of the family's dental expenditure for one year. The resulting data are given below. Assuming, that dental expenditure is normally distributed in the population, use the data and an alpha of 0.5 to test the dental association's hypothesis.
1008, 812, 1117, 1323, 1308, 1415, 831, 1021, 1287, 851, 930, 730, 699, 872, 913, 944, 954, 987, 1695, 995, 1003, 994



~~50K~~
~~24500~~
~~25525~~
~~.1955~~

(4) N = 22, so t test

$\mu = 1135$

$\alpha = 0.5 = 5\%$

$H_0: \mu = 1135$ S.D = 240.37

$H_1: \mu \neq 1135$

$$\bar{x} = \frac{1008 + 812 + \text{- - -} 994}{22} = 1031.32$$

apply t test

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{1031.32 - 1135}{\frac{240.37}{\sqrt{22}}}$$

t calc = -2.02

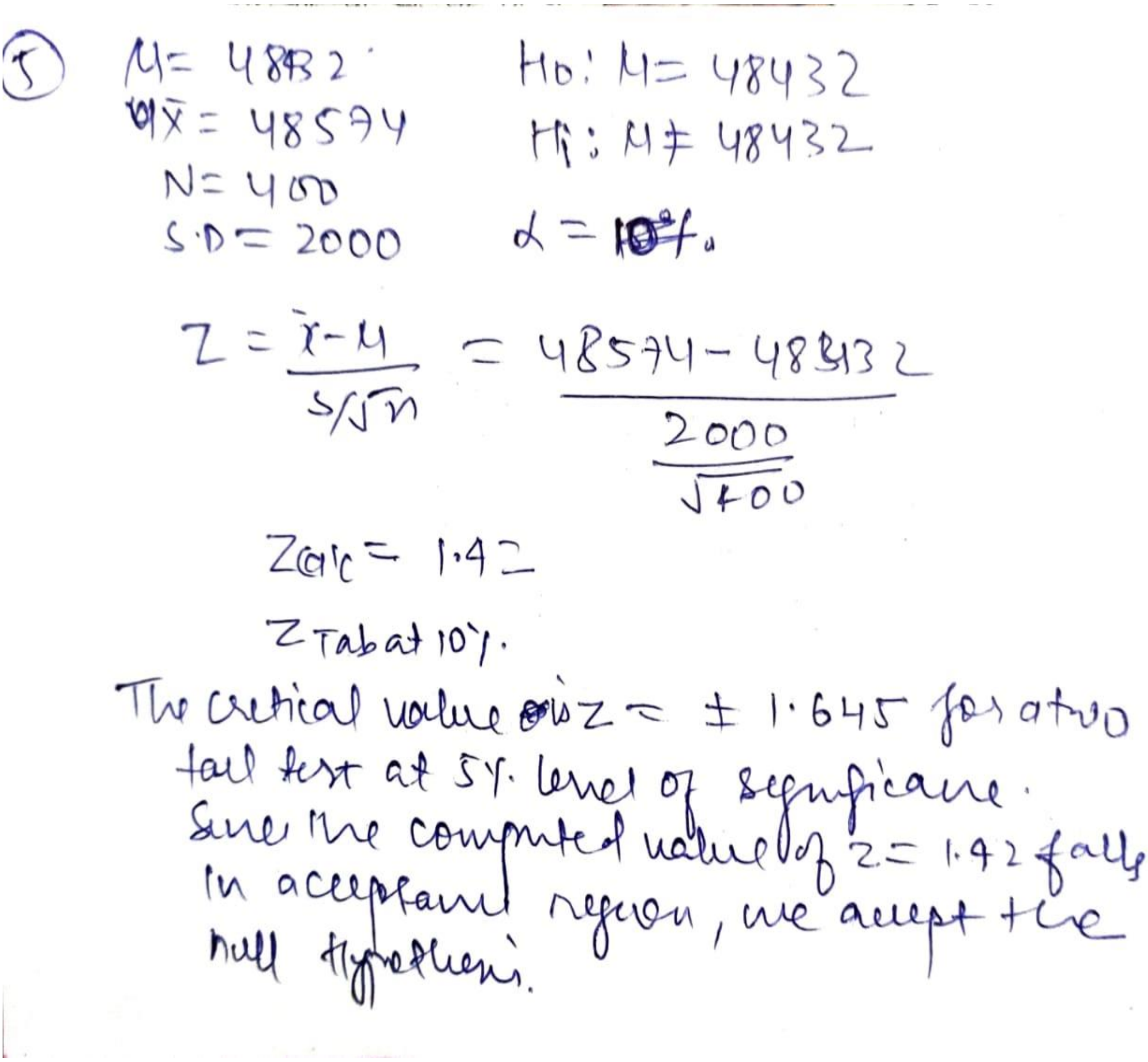
t tab at 5% = ± 1.96

Since the computed value of ~~t~~ calc falls in rejection region,

So Null Hypothesis Rejected.

Problem Statement 5:

In a report prepared by the Economic Research Department of a major bank the Department manager maintains that the average annual family income on Metropolis is $48,432. What do you conclude about the validity of the report if a random sample of 400 families shows and average income of $48,574 with a standard deviation of 2000?



(5) $\mu = 48432$ H0: $\mu = 48432$

$\bar{x} = 48574$ H1: $\mu \neq 48432$

$N = 400$

$S.D = 2000$ $\alpha = 10\%$

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{48574 - 48432}{\frac{2000}{\sqrt{400}}}$$

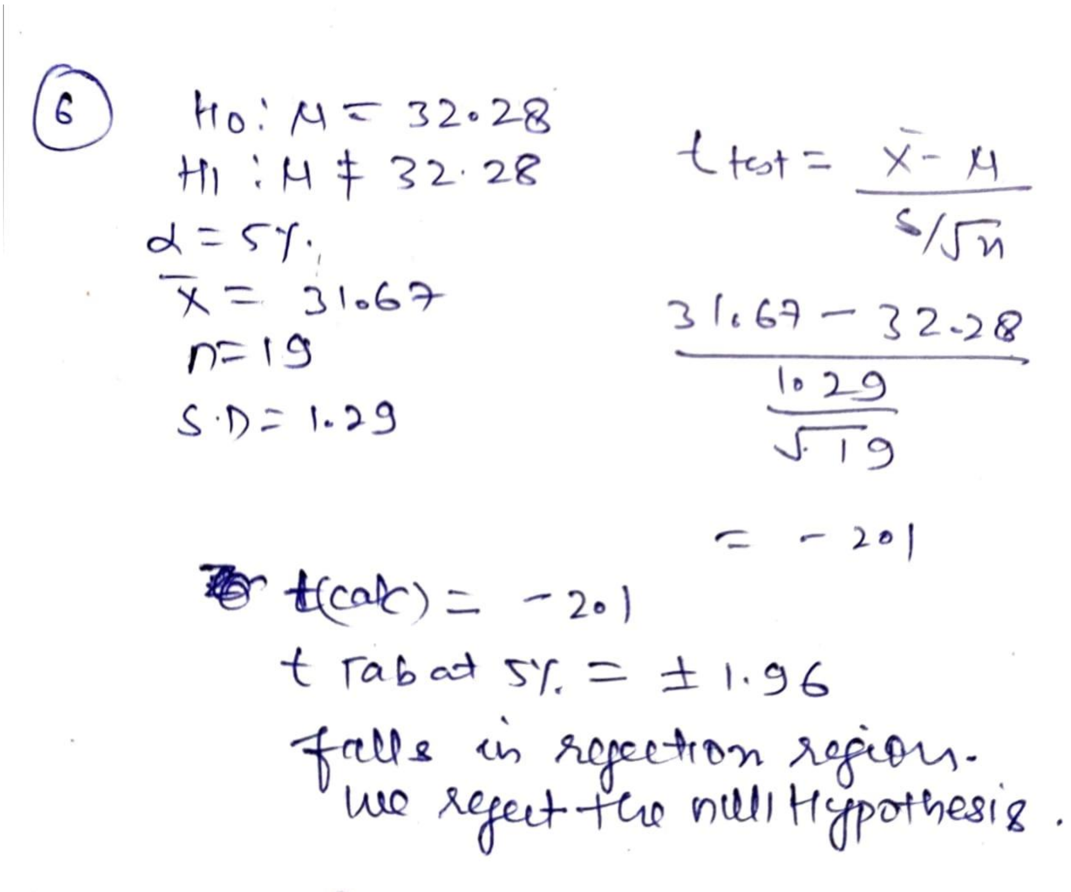
$Z_{calc} = 1.42$

Z Tab at 10%.

The critical value is z = ± 1.645 for a two tail test at 5% level of significance. Since the computed value of z = 1.42 falls in acceptance region, we accept the null hypothesis.

Problem Statement 6:

Suppose that in past years the average price per square foot for warehouses in the United States has been $32.28. A national real estate investor wants to determine whether that figure has changed now. The investor hires a researcher who randomly samples 19 warehouses that are for sale across the United States and finds that the mean price per square foot is $31.67, with a standard deviation of $1.29. assume that the prices of warehouse footage are normally distributed in population. If the researcher uses a 5% level of significance, what statistical conclusion can be reached? What are the hypotheses?



(6) $H_0: \mu = 32.28$

$H_1: \mu \neq 32.28$

$\alpha = 5\%$

$\bar{x} = 31.67$

$n = 19$

$S.D = 1.29$

$$t_{test} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$\frac{31.67 - 32.28}{\frac{1.29}{\sqrt{19}}}$$

$$= -2.01$$

$t(cal) = -2.01$

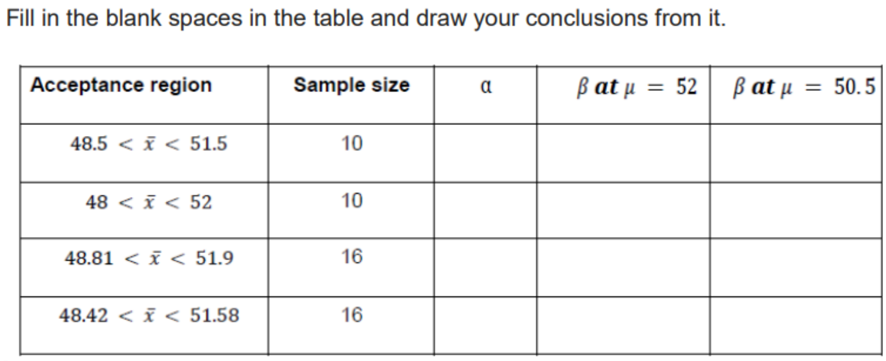
t Tab at 5% = ± 1.96

falls in rejection region.

we reject the null Hypothesis.

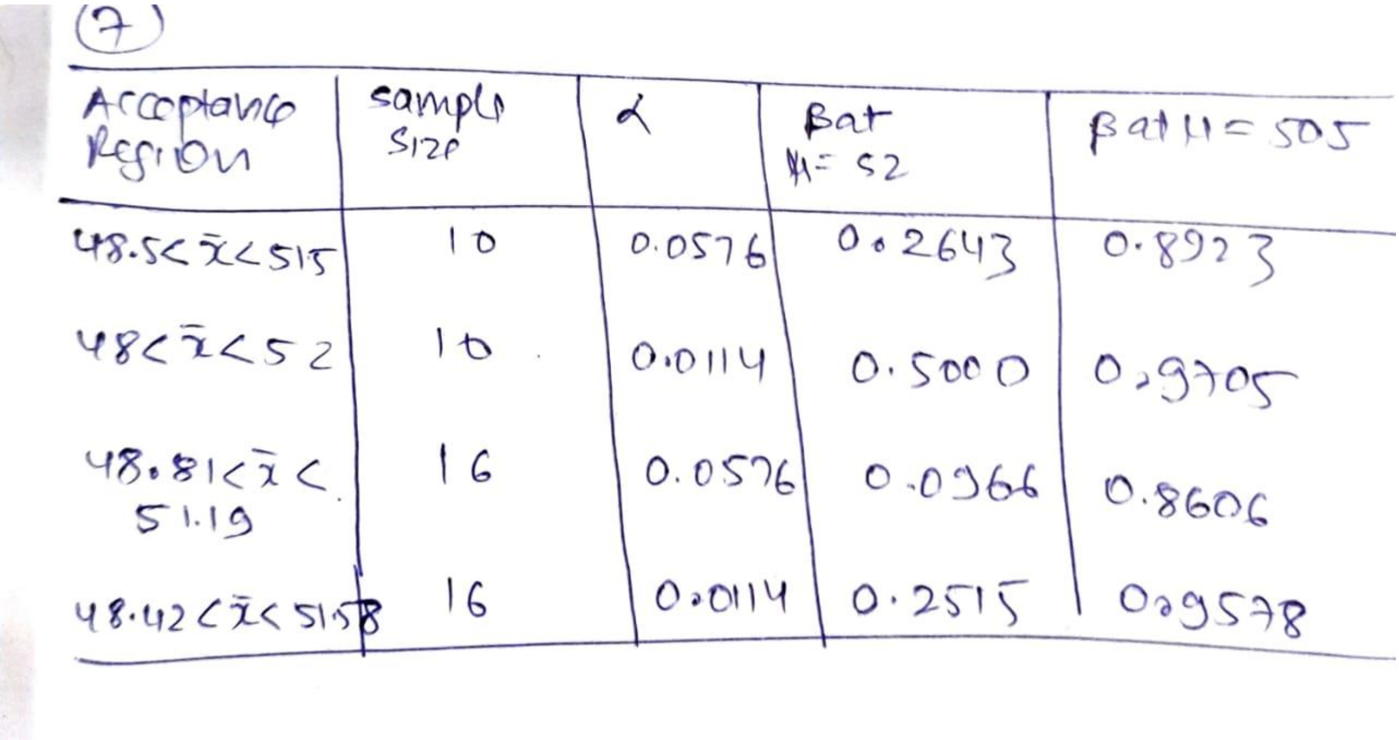
Problem Statement 7:

Fill in the blank spaces in the table and draw your conclusions from it.



Fill in the blank spaces in the table and draw your conclusions from it.

Acceptance region	Sample size	α	β **at** $\mu = 52$	β **at** $\mu = 50.5$
$48.5 < \bar{x} < 51.5$	10			
$48 < \bar{x} < 52$	10			
$48.81 < \bar{x} < 51.9$	16			
$48.42 < \bar{x} < 51.58$	16			



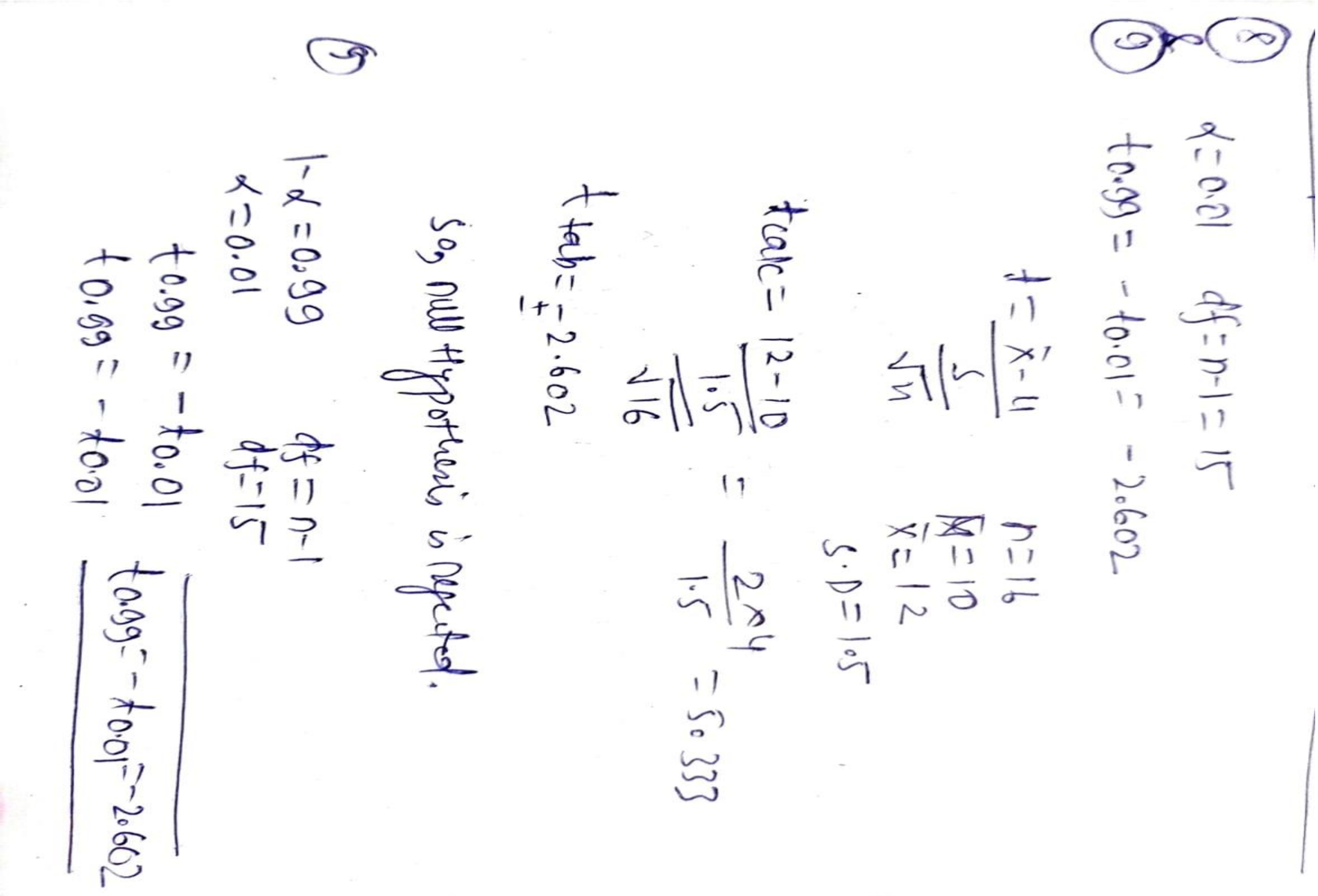
(7)

Acceptance Region	sample size	α	β at $\mu = 52$	β at $\mu = 50.5$
$48.5 < \bar{x} < 51.5$	10	0.0576	0.2643	0.8923
$48 < \bar{x} < 52$	10	0.0114	0.5000	0.9705
$48.81 < \bar{x} < 51.19$	16	0.0576	0.0966	0.8606
$48.42 < \bar{x} < 51.58$	16	0.0114	0.2515	0.9578

Problem Statement 8 and Problem Statement 9:

Find the t-score for a sample size of 16 taken from a population with mean 10 when the sample mean is 12 and the sample standard deviation is 1.5.

Find the t-score below which we can expect 99% of sample means will fall if samples of size 16 are taken from a normally distributed population.



(8) & (9)

$\alpha = 0.01 \quad df = n-1 = 15$

$t_{0.99} = -t_{0.01} = -2.602$

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

$n = 16$

$\mu = 10$

$\bar{x} = 12$

$S.D = 1.5$

$t_{calc} = \frac{12-10}{\frac{1.5}{\sqrt{16}}} = \frac{2 \times 4}{1.5} = 5.333$

$t_{tab} = \pm 2.602$

So, null Hypothesis is rejected.

(9)

$1 - \alpha = 0.99 \quad df = n-1$

$\alpha = 0.01 \quad df = 15$

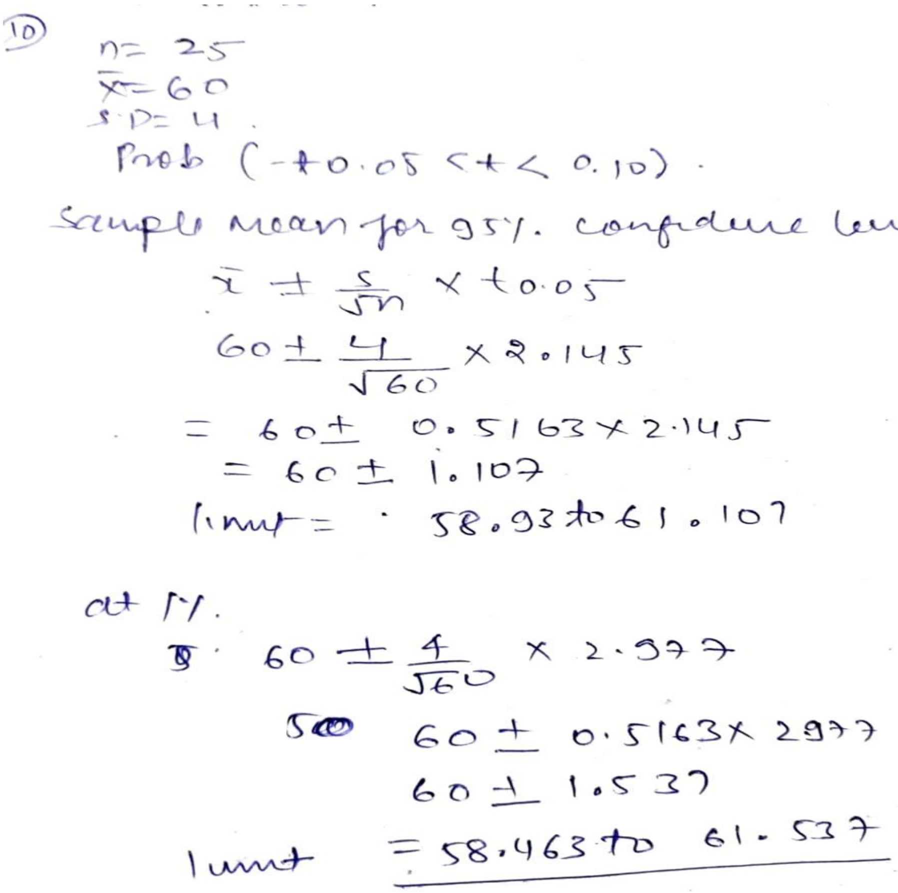
$t_{0.99} = -t_{0.01}$

$t_{0.99} = -t_{0.01}$

$\boxed{t_{0.99} = -t_{0.01} = -2.602}$

Problem Statement 10:

If a random sample of size 25 drawn from a normal population gives a mean of 60 and a standard deviation of 4, find the range of t-scores where we can expect to find the middle 95% of all sample means. Compute the probability that (−*t*0.05 <*t*<*t*0.10).



10)

$n = 25$

$\bar{x} = 60$

S.D $= 4$.

Prob $(-t_{0.05} < t < 0.10)$.

Sample mean for 95% confidence level

$$\bar{x} \pm \frac{s}{\sqrt{n}} \times t_{0.05}$$

$$60 \pm \frac{4}{\sqrt{60}} \times 2.145$$

$$= 60 \pm 0.5163 \times 2.145$$

$$= 60 \pm 1.107$$

limit $=$ 58.93 to 61.107

at 1%.

$$60 \pm \frac{4}{\sqrt{60}} \times 2.977$$

$$60 \pm 0.5163 \times 2977$$

$$60 \pm 1.537$$

limit $= \underline{58.463 \text{ to } 61.537}$

Problem Statement 11:

Two-tailed test for difference between two population means Is there evidence to conclude that the number of people travelling from Bangalore to Chennai is different from the number of people travelling from Bangalore to Hosur in a week, given the following:

Population 1: Bangalore to Chennai

n1 = 1200 x1 = 452

s1 = 212 Population

2: Bangalore to Hosur

n2 = 800 x2 = 523

s2 = 185

n>30 in both the cases hence we will apply z-test

$n1= 1200$

$\bar{x}_1 =452$

$\sigma_1=212$

□

$n2=800$

$\bar{x}_2 = 523$

$\sigma_2=185$

standard error= $\sqrt{(\frac{\sigma_1^2}{n_1}) + (\frac{\sigma_2^2}{n_1})}$

S.E= $\sqrt{(\frac{(212)^2}{1200} + \frac{(185)^2}{800})}$ = 8.96

z-test= $\frac{(\bar{x}_1 - \bar{x}_2)}{S.E}$

z-test= $\frac{(452 - 523)}{8.96}$ = -7.926 ....(1)

According to question test is two tail hance $\frac{\alpha}{2}$ will be taken under consideration where α =0.05 => $\frac{\alpha}{2}$ =0.025

$z_{0.025}$ = -2.81

AS z-test < $z_{0.025}$

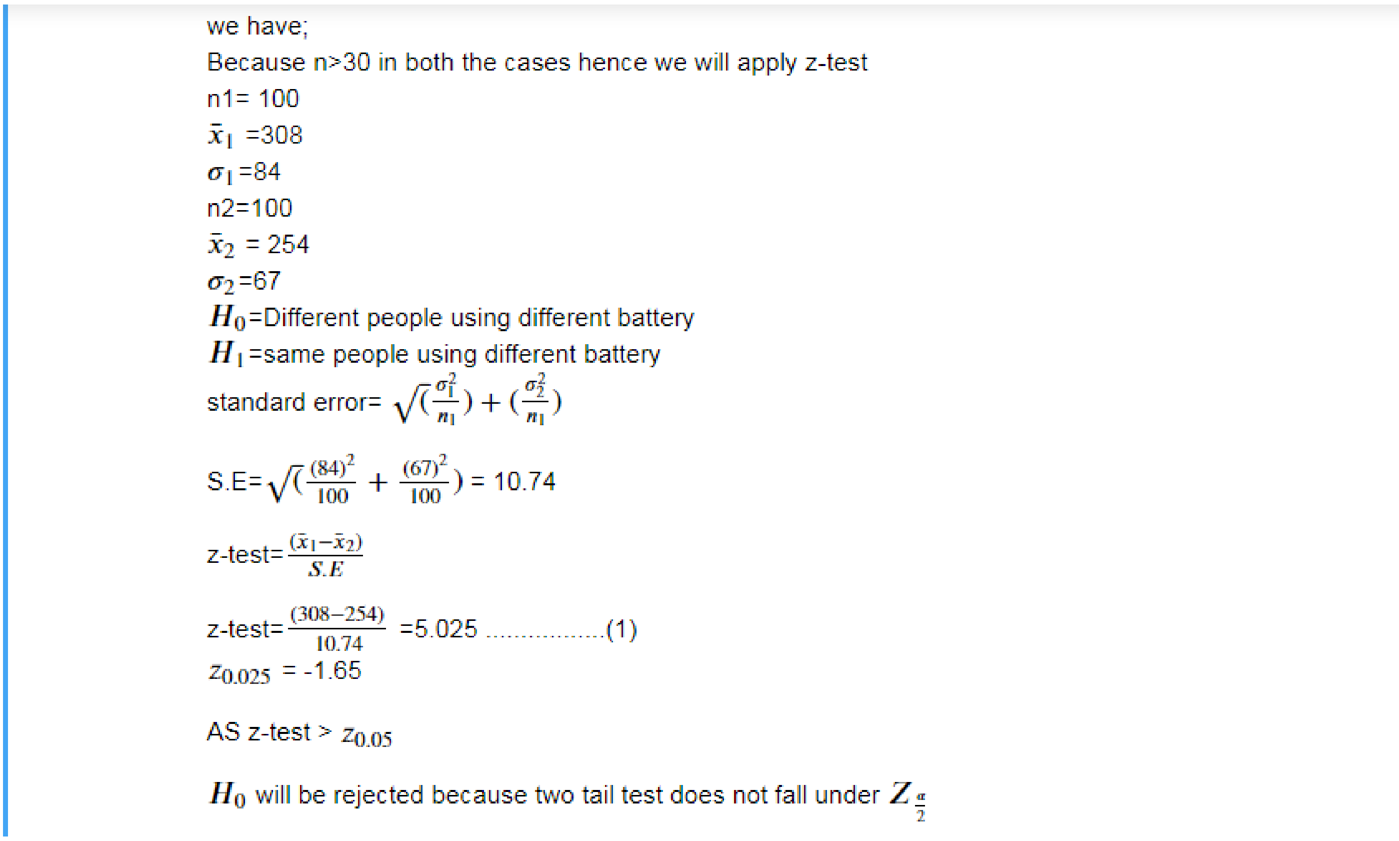
H_0 will be rejected

Problem Statement 12:

Is there evidence to conclude that the number of people preferring Duracell battery is different from the number of people preferring Energizer battery, given the following:

Population 1: Duracell n1 = 100 x1 = 308 s1 = 84

Population 2: Energizer n2 = 100 x2 = 254 s2 = 67



we have;

Because n>30 in both the cases hence we will apply z-test

n1= 100

$\bar{x}_1$ =308

σ_1 =84

n2=100

$\bar{x}_2$ = 254

σ_2 =67

H_0=Different people using different battery

H_1=same people using different battery

standard error= $\sqrt{(\frac{\sigma_1^2}{n_1}) + (\frac{\sigma_2^2}{n_1})}$

S.E= $\sqrt{(\frac{(84)^2}{100} + \frac{(67)^2}{100})}$ = 10.74

z-test= $\frac{(\bar{x}_1 - \bar{x}_2)}{S.E}$

z-test= $\frac{(308-254)}{10.74}$ =5.025 ....(1)

$z_{0.025}$ = -1.65

AS z-test > $z_{0.05}$

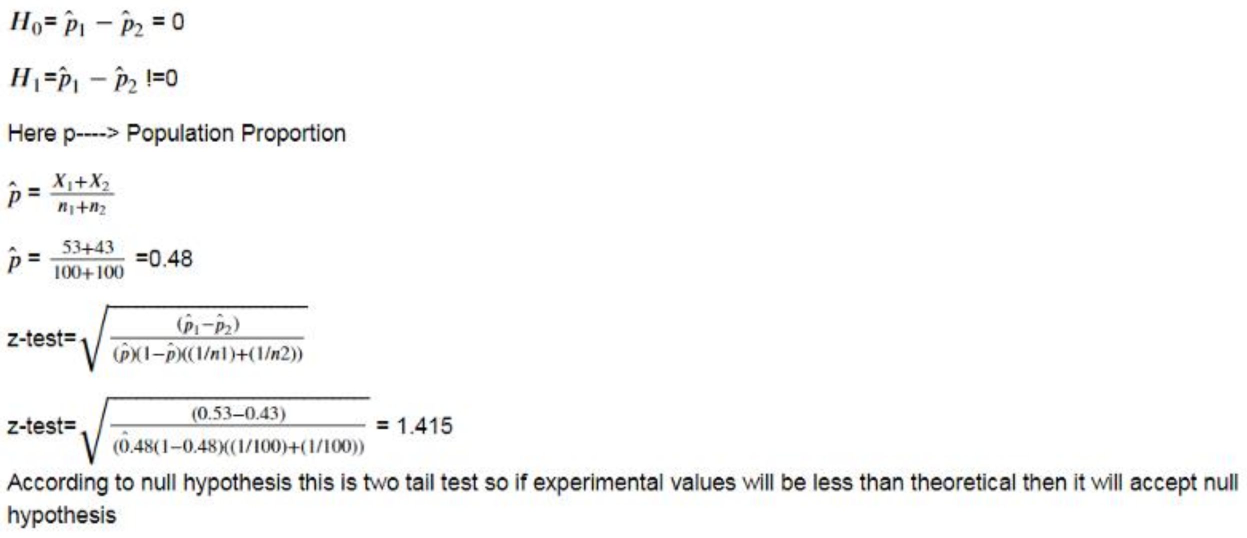
H_0 will be rejected because two tail test does not fall under $Z_{\frac{\alpha}{2}}$

Problem Statement 13:

Pooled estimate of the population variance Does the data provide sufficient evidence to conclude that average percentage increase in the price of sugar differs when it is sold at two different prices?

Population 1: Price of sugar = Rs. 27.50 n1 = 14 x1 = 0.317% s1 = 0.12%

Population 2: Price of sugar = Rs. 20.00 n2 = 9 x2 = 0.21% s2 = 0.11%



H_0= $\hat{p}_1 - \hat{p}_2$ = 0

H_1=$\hat{p}_1 - \hat{p}_2$!=0

Here p----> Population Proportion

$\hat{p}$ = $\frac{X_1+X_2}{n_1+n_2}$

$\hat{p}$ = $\frac{53+43}{100+100}$ =0.48

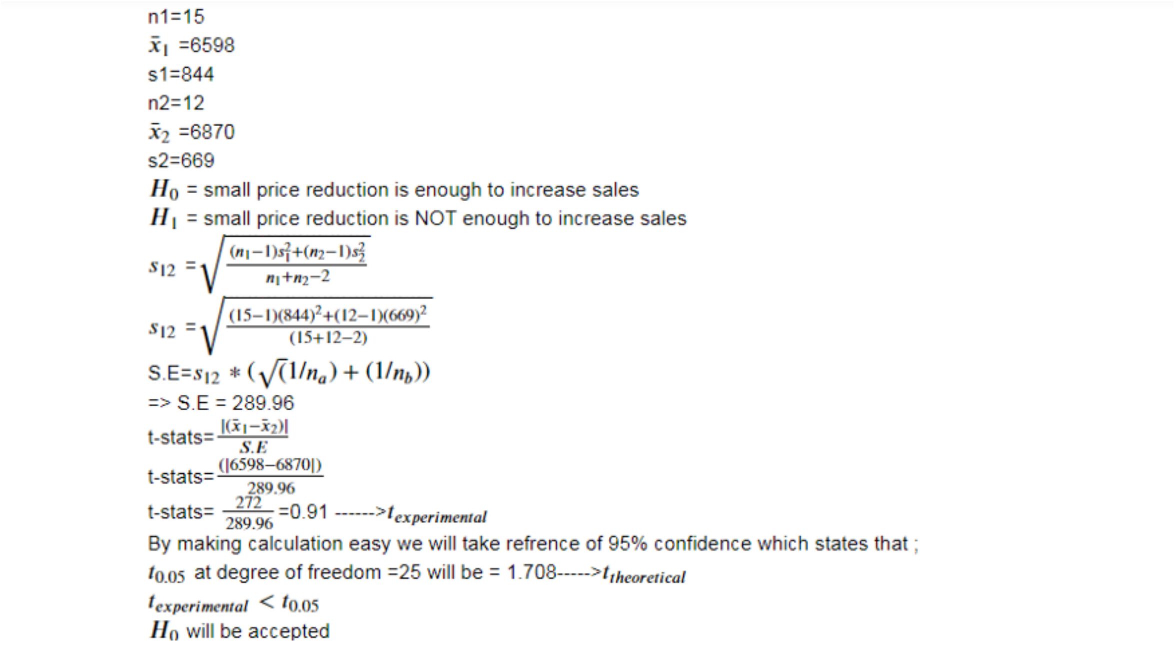
z-test= $\sqrt{\frac{(\hat{p}_1-\hat{p}_2)}{(\hat{p})(1-\hat{p})((1/n1)+(1/n2))}}$

z-test= $\sqrt{\frac{(0.53-0.43)}{(\hat{0}.48(1-0.48)((1/100)+(1/100))}}$ = 1.415

According to null hypothesis this is two tail test so if experimental values will be less than theoretical then it will accept null hypothesis

Problem Statement 14:

The manufacturers of compact disk players want to test whether a small price reduction is enough to increase sales of their product. Is there evidence that the small price reduction is enough to increase sales of compact disk players? Population 1: Before reduction n1 = 15 x1 = Rs. 6598 s1 = Rs. 844 Population 2: After reduction n2 = 12 x2 = RS. 6870 s2 = Rs. 669



n1=15

$\bar{x}_1$ =6598

s1=844

n2=12

$\bar{x}_2$ =6870

s2=669

H_0 = small price reduction is enough to increase sales

H_1 = small price reduction is NOT enough to increase sales

$s_{12} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

$s_{12} = \sqrt{\frac{(15-1)(844)^2+(12-1)(669)^2}{(15+12-2)}}$

S.E=$s_{12} * (\sqrt{(}1/n_a) + (1/n_b))$

=> S.E = 289.96

t-stats=$\frac{|(\bar{x}_1-\bar{x}_2)|}{S.E}$

t-stats=$\frac{(|6598-6870|)}{289.96}$

t-stats= $\frac{272}{289.96}$ =0.91 ------>$t_{experimental}$

By making calculation easy we will take refrence of 95% confidence which states that ;

$t_{0.05}$ at degree of freedom =25 will be = 1.708----->$t_{theoretical}$

$t_{experimental} < t_{0.05}$

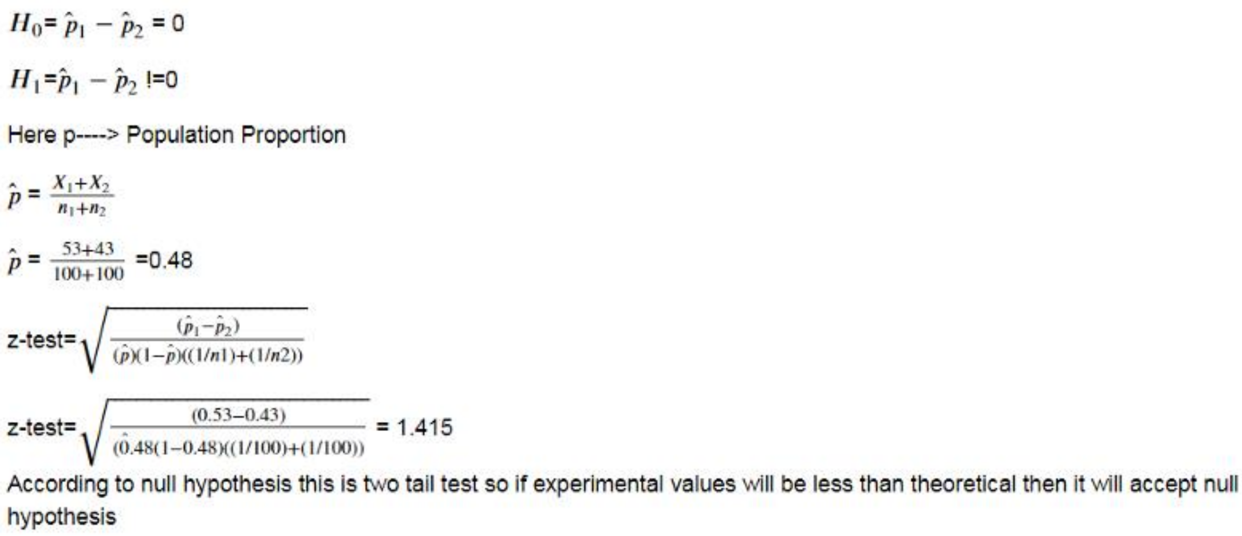
H_0 will be accepted

Problem Statement 15:

Comparisons of two population proportions when the hypothesized difference is zero Carry out a two-tailed test of the equality of banks' share of the car loan market in 1980 and 1995.

Population 1: 1980 n1 = 1000 x1 = 53 p 1 = 0.53

Population 2: 1985 n2 = 100 x2 = 43 p 2= 0.53



H_0= $\hat{p}_1 - \hat{p}_2$ = 0

H_1=$\hat{p}_1 - \hat{p}_2$!=0

Here p----> Population Proportion

$\hat{p}$ = $\frac{X_1+X_2}{n_1+n_2}$

$\hat{p}$ = $\frac{53+43}{100+100}$ =0.48

z-test=$\sqrt{\frac{(\hat{p}_1-\hat{p}_2)}{(\hat{p})(1-\hat{p})((1/n1)+(1/n2))}}$

z-test=$\sqrt{\frac{(0.53-0.43)}{(\hat{0}.48(1-0.48)((1/100)+(1/100))}}$ = 1.415

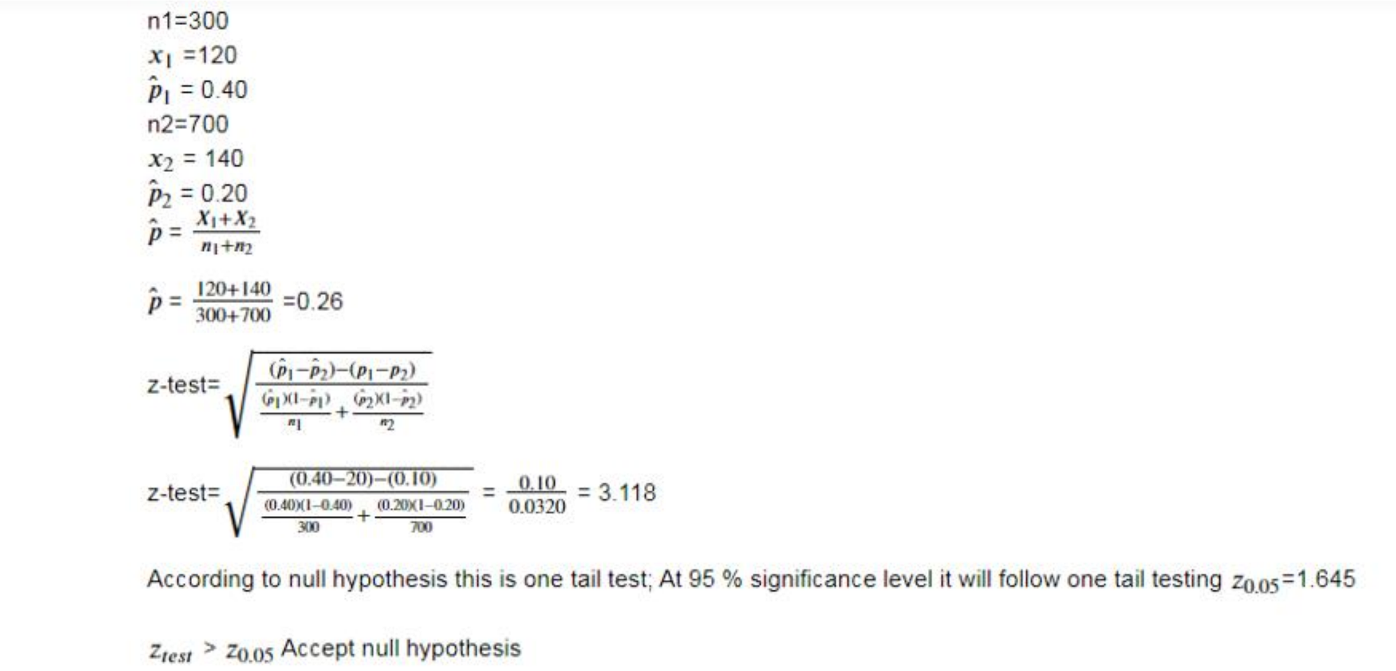
According to null hypothesis this is two tail test so if experimental values will be less than theoretical then it will accept null hypothesis

Problem Statement 16:

Carry out a one-tailed test to determine whether the population proportion of traveler's check buyers who buy at least $2500 in checks when sweepstakes prizes are offered as at least 10% higher than the proportion of such buyers when no sweepstakes are on.

Population 1: With sweepstakes n1 = 300 x1 = 120 p = 0.40

Population 2: No sweepstakes n2 = 700 x2 = 140 p 2= 0.20



n1=300

x_1 =120

$\hat{p}_1$ = 0.40

n2=700

x_2 = 140

$\hat{p}_2$ = 0.20

$$\hat{p} = \frac{X_1+X_2}{n_1+n_2}$$

$$\hat{p} = \frac{120+140}{300+700} = 0.26$$

$$\text{z-test}= \sqrt{\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\frac{(\hat{p}_1)(1-\hat{p}_1)}{n_1}+\frac{(\hat{p}_2)(1-\hat{p}_2)}{n_2}}}$$

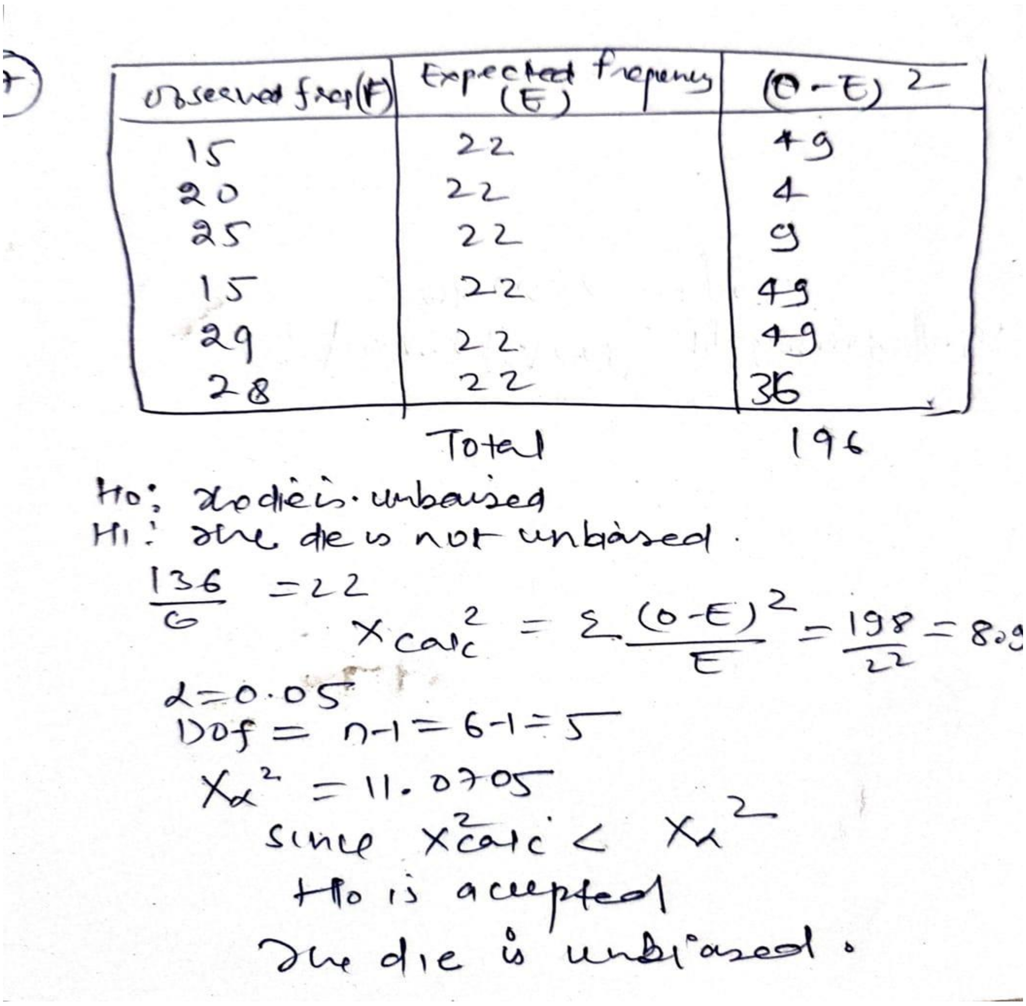
$$\text{z-test}= \sqrt{\frac{(0.40-20)-(0.10)}{\frac{(0.40)(1-0.40)}{300}+\frac{(0.20)(1-0.20)}{700}}} = \frac{0.10}{0.0320} = 3.118$$

According to null hypothesis this is one tail test; At 95 % significance level it will follow one tail testing $z_{0.05}$=1.645

$z_{test} > z_{0.05}$ Accept null hypothesis

Problem Statement 17:

A die is thrown 132 times with the following results: Number turned up: 1, 2, 3, 4, 5, 6 Frequency: 16, 20, 25, 14, 29, 28 Is the die unbiased? Consider the degrees of freedom as $p - 1$



7)

Observed freq(F)	Expected frequency (E)	$(O-E)^2$
15	22	49
20	22	4
25	22	9
15	22	49
29	22	49
28	22	36
	Total	196

Ho; The die is unbaised
H1: The die is not unbiased.

$\frac{136}{6} = 22$

$X_{calc}^2 = \Sigma \frac{(O-E)^2}{E} = \frac{198}{22} = 8.9$

$\alpha = 0.05$

$Dof = n-1 = 6-1 = 5$

$X_\alpha^2 = 11.0705$

since $X_{calc}^2 < X_\alpha^2$

Ho is accepted

The die is unbiased.

Problem Statement 18:

In a certain town, there are about one million eligible voters. A simple random sample of 10000 eligible voters was chosen to study the relationship between sex and participation in the last election. The results are summarized in the following 2X2 (read two by two) contingency table:

	Men	Women
Voted	2792	3591
Didn't vote	1486	2131

We want to check whether being a man or a woman (columns) is independent of having voted in the last election (rows). In other words is "sex and voting independent"?

Solution:

In order to answer the question we need to build a test of hypothesis as usual. We have

> Null := `Sex is independent of Voting`:
> Alternative := `Sex and Voting are dependent`:

After specifying the Null hypothesis we need to compute the expected table under the assumption that rows and columns are in fact independent. To compute the expected table we use the product rule for chances:

chance of (row_i,col_j) = (chance row_i) * (chance col_j)

From here we deduce that the expected number of counts in (row_i,col_j) is given by:

N*(chance row_i)*(chance col_j) = (Sum row_i)*(Sum col_j) / N

The observed table with totals included is:

OBSERVED TABLE

	Men	Women	Total
Voted	2792	3591	6383
Didn't vote	1486	2131	3617
Total	4278	5722	10000

The associated expected table under the assumption that sex and voting are independent is given by

EXPECTED TABLE

	Men	Women	Total
Voted	2731	3652	6383
Didn't vote	1547	2070	3617
Total	4278	5722	10000

X2 := (2792-2731)^2/2731 + + + (2131- We now have the observed table and the expected table under the null hypothesis of independence. After that we need to compute the X2 statistic. The X2 statistic measures how far away is the observed table from the expected one. The X2 statistic has as many terms as there are cells in the observed table (4 in our case):

```
> c11 := (2792-2731)^2/2731.:
> c12 := (3591-3652)^2/3652.:
> c21 := (1486-1547)^2/1547.:
> c22 := (2131-2070)^2/2070.:
```

The X2-statistic is the sum of each of the contributions from each cell:

```
> X2 := c11+c12+c21+c22;

                X2 := 6.584283457
```

The last part is to compute the P-value. This is done by looking under the Chi-square table with (rows-1)*(cols-1) degrees of freedom. In the case of a 2x2 table (our case) the number of degrees of freedom is (2-1)(2-1)=1*1=1. The table gives the tail areas at:

Degrees of freedom	99%	...	10%	5%	1%
1	0.00016		2.71	3.84	6.64
2	0.020		4.60	5.99	9.21

Since the observed X2 = 6.58 and thus,

$$3.84 < X2 < 6.64$$

we conclude that:

$$1\% < \text{P-value} < 5\%$$

and we reject the NULL. The data supports the hypothesis that sex and voting are dependent in this town.

Problem Statement 19:

1. A sample of 100 voters are asked which of four candidates they would vote for in an election. The number supporting each candidate is given below:

Higgins	Reardon	White	Charlton
41	19	24	16

Do the data suggest that all candidates are equally popular? [Chi-Square = 14.96, with 3 d.f.: $p<0.05$].

Solution:

A Chi-Squared Goodness-of-Fit test is appropriate here. The null hypothesis is that there is no preference for any of the candidates: if this is so, we would expect roughly equal numbers of voters to support each candidate. Our expected frequencies are therefore 100/4 = 25 per candidate.

O	**41**	**19**	**24**	**16**
E	**25**	**25**	**25**	**25**
(O-E)	**16**	**-6**	**-1**	**-9**
(O-E)^2	**256**	**36**	**1**	**81**
(O-E)^2 **---------** **E**	**10.24**	**1.44**	**0.04**	**3.24**

Adding together the last row gives us our value of c^2 :

(O - E)^2
å ----------------- = 10.24+ 1.44 + 0.04 + 3.24 = **14.96**, with 4 - 1 = 3 degrees of freedom.
E

The critical value of Chi-Square for a 0.05 significance level and 3 d.f. is 7.82. Our obtained Chi-Square value is bigger than this, and so we conclude that our obtained value is unlikely to have occurred merely by chance. In fact, our obtained value is bigger than the critical Chi-Square value for the 0.01 significance level (13.28). In other words, it is possible that our obtained Chi-Square value is due merely to chance, but highly unlikely: a Chi-Square value as large as ours will occur by chance only about once in a hundred trials. It seems more reasonable to conclude that our results are not de to chance, and that the data do indeed suggest that voters do not prefer the four candidates equally.

Problem Statement 20:

Children of three ages are asked to indicate their preference for three photographs of adults. Do the data suggest that there is a significant relationship between age and photograph preference? What is wrong with this study? [Chi-Square = 29.6, with 4 d.f.: $p<0.05$].

		Photograph:		
		A	B	C
Age of child:	5-6 years:	18	22	20
	7-8 years:	2	28	40
	9-10 years:	20	10	40

Solution:

	photograph:			
age of child:	**A:**	**B:**	**C:**	**row totals:**
5-6 years	**18**	**22**	**20**	**60**
7-8 years	**2**	**28**	**40**	**70**
9-10 years	**20**	**10**	**40**	**70**
column totals:	**40**	**60**	**100**	**200**

(a) Work out the row, column and grand totals (as shown in the shaded parts of the table, above).
(b) Work out the expected frequencies, using the formula:

$$E = \frac{\text{(row total * column total)}}{\text{grand total}}$$

For each cell of the above table, this gives us:

O:	**18**	**22**	**20**	**2**	**28**	**40**	**20**	**10**	**40**
E:	**12**	**18**	**30**	**14**	**21**	**35**	**14**	**21**	**35**

Next, work out (O - E):

(O-E):	**6**	**4**	**-10**	**-12**	**7**	**5**	**6**	**11**	**5**

Square each of these, to get (O - E)^2 :

(O - E)^2:	**36**	**16**	**100**	**144**	**49**	**25**	**36**	**121**	**25**

Divide each of the above numbers by E, to get (O - E)^2 / E:

$\frac{(O - E)^2}{E}$	**3**	**0.89**	**3.33**	**10.29**	**2.33**	**0.71**	**2.57**	**5.76**	**0.71**

Chi-squared is the sum of these:

c^2 = **29.60**.

d.f. = (rows - 1) * (columns - 1) = 2 * 2 = 4.

The critical value of Chi-Square in the table for a 0.001 significance level and 4 d.f. is 18.46. Our obtained Chi-Square value is bigger than this: therefore we have a Chi-Square value which is so large that it would occur by chance only about once in a thousand times. It seems more reasonable to accept the alternative hypothesis, that there is a significant relationship between age of child and photograph preference.

Problem Statement 21:

. A study of conformity using the Asch paradigm involved two conditions: one where one confederate supported the true judgement, and another where no confederate gave the correct response

	Support:	No Support:
Conform:	18	40
Not Conform:	32	10

Is there a significant difference between the "support" and "no support" conditions in the frequency with which individuals are likely to conform? [Chi-Square = 19.87, with 1 d.f.: $p<0.05$. *OR: Chi-Square = 18.1. See the comment at the end of this handout*].

Solution:

Here we have a 2x2 contingency table. Chi-Square is the appropriate test to use, but since we have 1 d.f., we will modify the formula to include "Yates' correction for continuity".

	support	no support	row totals:
conform:	18	40	58
not conform:	32	10	42
column totals:	50	50	100

(a) Calculate the row, column and grand totals.

(b) Calculate the expected frequency for each cell of the table, by multiplying together the appropriate row and column totals and then dividing by the grand total.

(c) Subtract each expected frequency from its associated observed frequency; but then apply Yates' correction, by subtracting 0.5 from the absolute value of each O-E value. (The vertical bars in the formula mean "ignore any minus signs").

O:	**18**	**40**	**32**	**10**
E:	**29**	**29**	**21**	**21**

Next, work out (O - E):

(|O-E|- 0.5):	**10.5**	**10.5**	**10.5**	**10.5**

Square each of these, to get $(O - E)^2$:

$(\|O-E\|- 0.5)^2$:	**110.25**	**110.25**	**110.25**	**110.25**

Divide each of the above numbers by E, to get $(O - E)^2$ / E:

$(O - E)^2$ **-----------** **E**	**3.80**	**3.80**	**5.25**	**5.25**

Chi-squared is the sum of these:

c^2 = **18.10.**

d.f. = (rows - 1) * (columns - 1) = 1 * 1 = 1.

Our obtained value of Chi-Squared is bigger than the critical value of Chi-Squared for a 0.001 significance level. In other words, there is less than a one in a thousand chance of obtaining a Chi-Square value as big as our obtained one, merely by chance. Therefore we can conclude that there is a significant difference between the "support" and "no support" conditions, in terms of the frequency with which individuals conformed.

Problem Statement 22:

We want to test whether short people differ with respect to their leadership qualities (Genghis Khan, Adolf Hitler and Napoleon were all stature-deprived, and how many midget MP's are there?) The following table shows the frequencies with which 43 short people and 52 tall people were categorised as "leaders", "followers" or as "unclassifiable". Is there a relationship between height and leadership qualities? [Chi-Square = 10.71, with 2 d.f.: p<0.01]

Solution

Expected frequencies are in brackets:

	height:		
	short	**tall**	**row totals:**
leader:	**12 (19.92)**	**32 (24.08)**	**44**
follower:	**22 (16.29)**	**14 (19.71)**	**36**
unclassifiable:	**9 (6.79)**	**6 (8.21)**	**15**
column totals:	**43**	**52**	**95**

Chi-Square = 3.146 + 2.602 + 1.998 + 1.652 + 0.720 + 0.595 = **10.712**, with 2 d.f.

10.712 is bigger than the tabulated value of Chi-Square at the 0.01 significance level. We would conclude that there seems to be a relationship between height and leadership qualities. Note that we can only say that there is a relationship between our two variables, not that once causes the other. There could be all kinds of explanations for such a relationship.

Problem Statement 23:

Each respondent in the Current Population Survey of March 1993 was classified as employed, unemployed, or outside the labor force. The results for men in California age 35-44 can be cross-tabulated by marital status, as follows:

	Married	Widowed, divorced, or separated	Never married
Employed	679	103	114
Unemployed	63	10	20
Not in labor force	42	18	25

Men of different marital status seem to have different distributions of labor force status. Or is this just chance variation? (you may assume the table results from a simple random sample.

We have:

```
> Obs_table := matrix(3,3,[679,103,114,63,10,20,42,18,25]);

                                [679    103    114]
                                [                 ]
                  Obs_table :=  [ 63     10     20]
                                [                 ]
                                [ 42     18     25]
> R1 := 679+103+114:R2:=63+10+20:R3:=42+18+25:
> C1:=679+63+42:C2:=103+10+18:C3:=114+20+25:N:=evalf(R1+R2+R3):
> Exp_table := matrix(3,3,(i,j)-> round(R.i*C.j/N));
                                [654    109    133]
                                [                 ]
                  Exp_table :=  [ 68     11     14]
                                [                 ]
                                [ 62     10     13]
```

```
                        2                    2
          (679 - 654)             (25 - 13)
   X2 :=  ------------ + ...    + ----------
              654                     13
```

> X2 := 30.96:

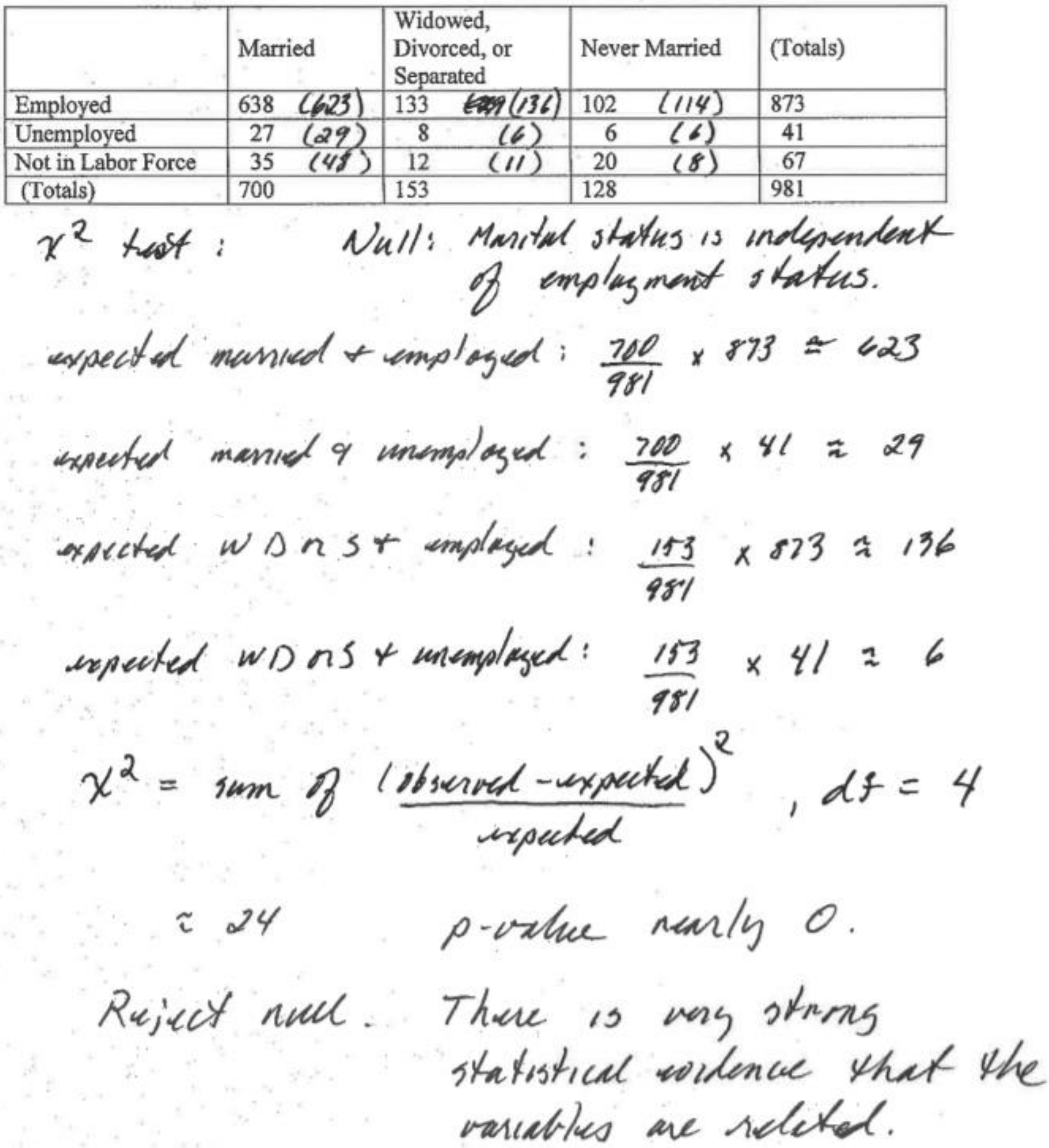
Looking at the table of the Chi-sqare distribution with (3-1)(3-1)=2*2=4 degrees of freedom we get:

```
Degrees of
 freedom      99%  ...      10%     5%      1%
_____________________________________________________
1             0.00016               2.71    3.84    6.64
2             0.020         4.60    5.99    9.21
3             0.12          6.25    7.82    11.34
4             0.30          7.78    9.49    13.28
5             0.55          9.24    11.07   15.09
```

since 30.96 > > 13.28 we conclude from the table that:

P < < 1%

so we reject with all confidence. Conclussion: Marital Status seems to be related to Job Status in this town.



	Married	Widowed, Divorced, or Separated	Never Married	(Totals)
Employed	638 (623)	133 ~~639~~ (136)	102 (114)	873
Unemployed	27 (29)	8 (6)	6 (6)	41
Not in Labor Force	35 (48)	12 (11)	20 (8)	67
(Totals)	700	153	128	981

χ^2 test : Null: Marital status is independent of employment status.

expected married + employed : $\frac{700}{981} \times 873 \approx 623$

expected married & unemployed : $\frac{700}{981} \times 41 \approx 29$

expected WDorS + employed : $\frac{153}{981} \times 873 \approx 136$

expected WDorS + unemployed : $\frac{153}{981} \times 41 \approx 6$

$\chi^2 = \text{sum of } \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$, $df = 4$

≈ 24 p-value nearly 0.

Reject null. There is very strong statistical evidence that the variables are related.