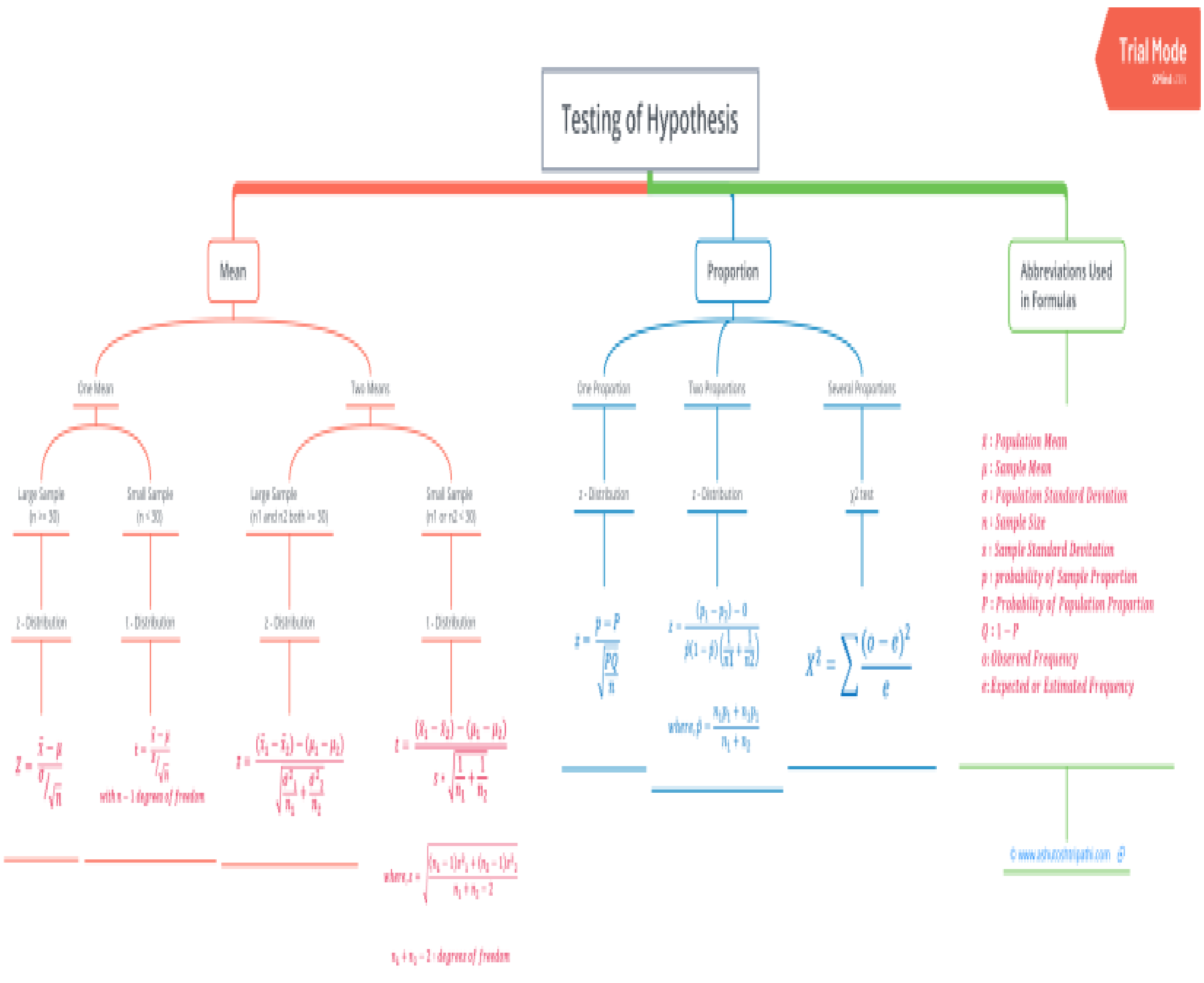
PRACTICE PROBLEMS ON HYPOTHESIS TESTING

Prerequisite to understand Hypothesis testing examples:

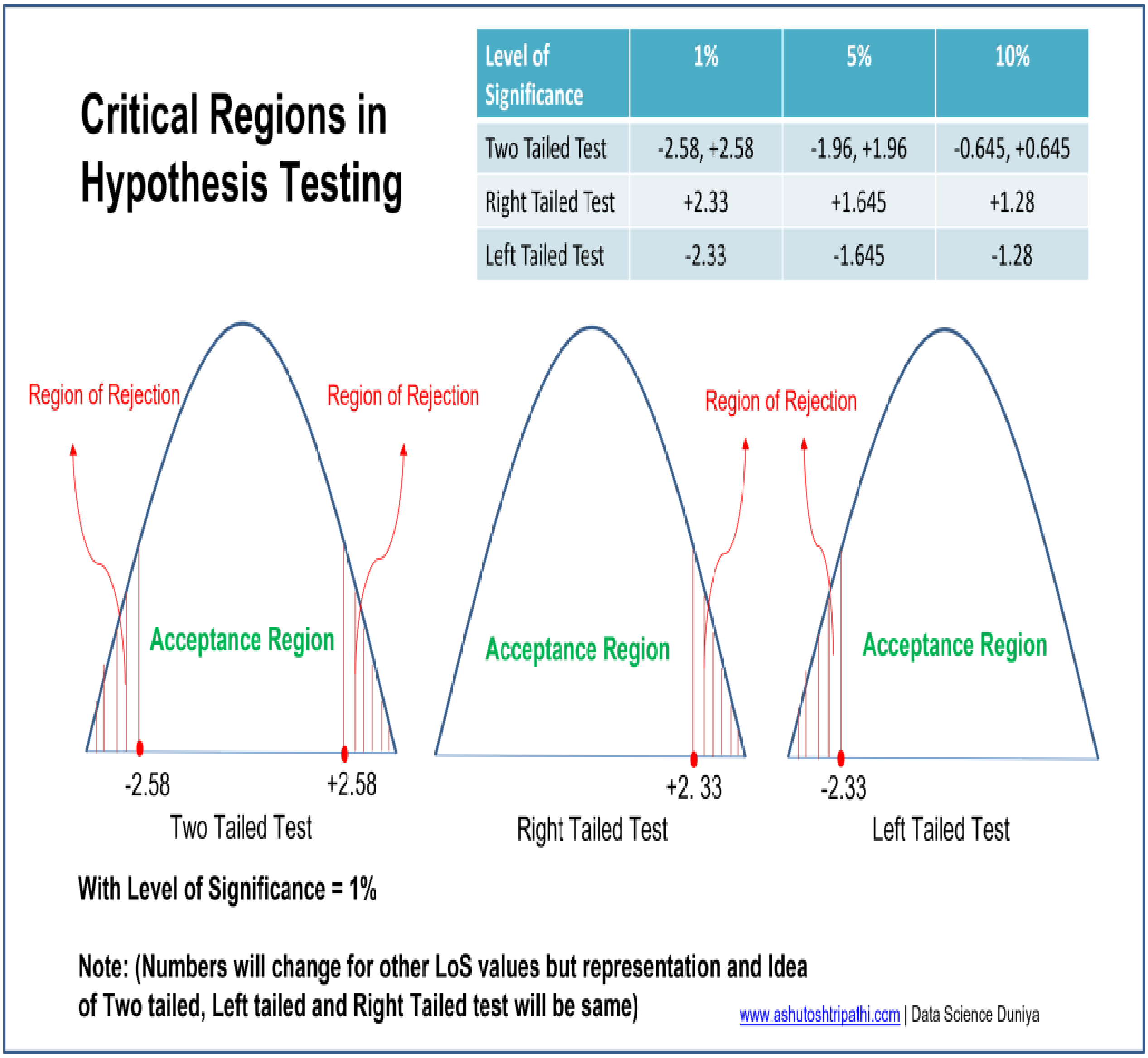
- Understanding of hypothesis testing concepts
- How to use z-table, t-table and chi square table.

Formula list:



Critical Regions

In hypothesis testing, critical region is represented by set of values, where null hypothesis is rejected. So it is also know as region of rejection. It takes different boundary values for different level of significance. Below info graphics shows the region of rejection that is critical region and region of acceptance with respect to the level of significance 1%.



Critical regions in Hypothesis Testing

LoS ->	α = 1%	α = 5%	α = 10%
Two Tailed Test	(-2.58, +2.58)	(-1.96, +1.96)	(-0.645, +0.645)
Right Tailed Test	+2.33	+1.645	+1.28
Left Tailed Test	-2.33	-1.645	-1.28

critical region values for 1% level of significance

Question 1

A Telecom service provider claims that individual customers pay on an average 400 rs. per month with standard deviation of 25 rs. A random sample of 50 customers bills during a given month is taken with a mean of 250 and standard deviation of 15. What to say with respect to the claim made by the service provider?

Solution:

First thing first, Note down what is given in the question:

H_0 (Null Hypothesis) : $\mu = 400$

H_1 (Alternate Hypothesis): $\mu \neq 400$ (Not equal means either $\mu > 400$ or $\mu < 400$ Hence it will be validated with **two tailed test**)

$\sigma = 25$ (Population Standard Deviation)

LoS (α) = 5% (Take 5% if not given in question)

$n = 50$ (Sample size)

xbar $\bar{x} = 250$ (Sample mean)

$s = 15$ (sample Standard deviation)

$n >= 30$ hence will go with z-test

Step 1:

Calculate z using z-test formula as below:

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

$z = (250 - 400) / (25/\sqrt{50})$

$z = -42.42$

Step 2:

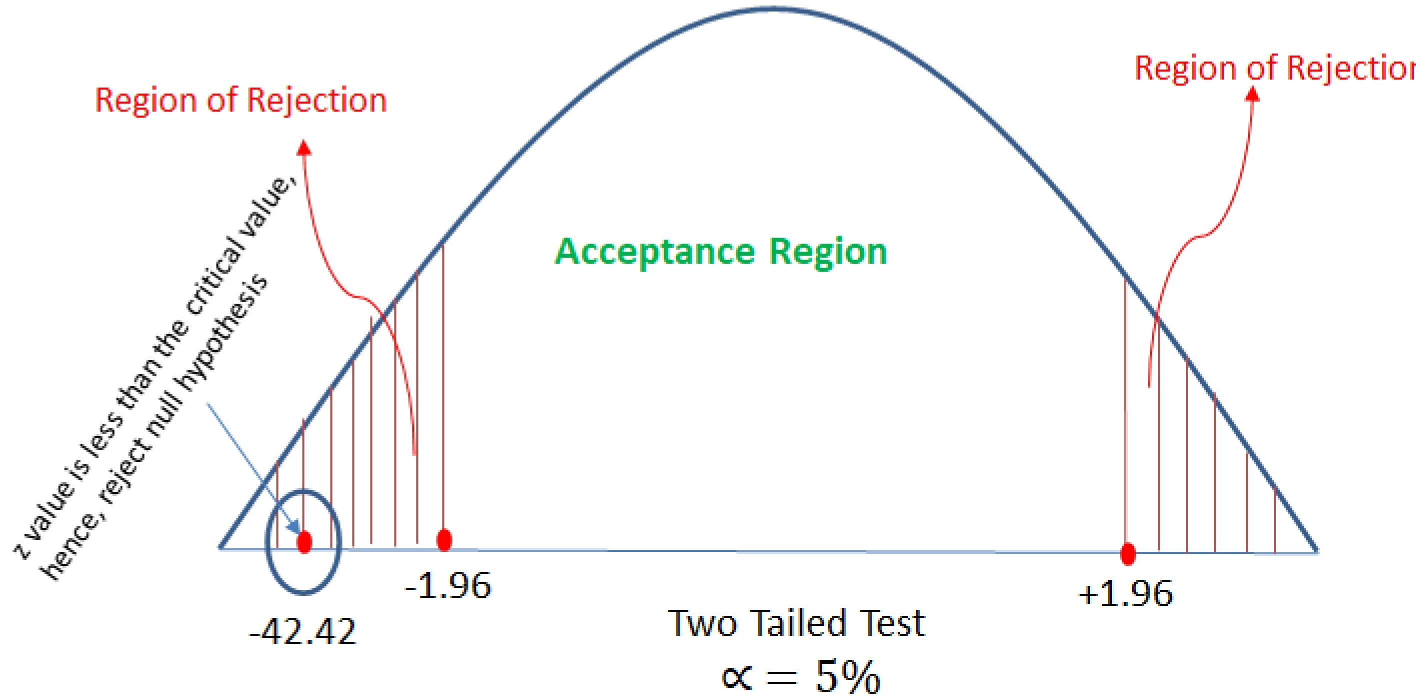
get z critical value from z table for $\alpha = 5\%$

z critical values = $(-1.96, +1.96)$

to accept the claim (significantly), calculated z should be in between

$-1.96 < z < +1.96$

but calculated z $(-42.42) < -1.96$ which mean reject the null hypothesis



z-test example 1

Question 2

From the data available, it is observed that 400 out of 850 customers purchased the groceries online. Can we say that most of the customers are moving towards online shopping even for groceries?

Solution:

Note down what is given:

400 out of 850 which indicates that this is a proportion problem.

Proportion p (small p) = 400/850 = 0.47

H_0 (Null Hypothesis): P (capital P) > 0.5 (claim is that most of the customers are moving towards online shopping even for groceries which mean at least 50% should do online shopping)

H_1 (Alternate Hypothesis): P < = 0.5 **left tailed**

n = 850

LoS (α) = 5% (assume 5% as it not given in question)

n > = 30 hence will go with z-test

Step 1:

calculate z value using the z-test formula

z = (p - P)/√(P*Q/n)

z = (0.47 - 0.50)/√(0.5*0.5/850)

z = 1.74

Step 2:

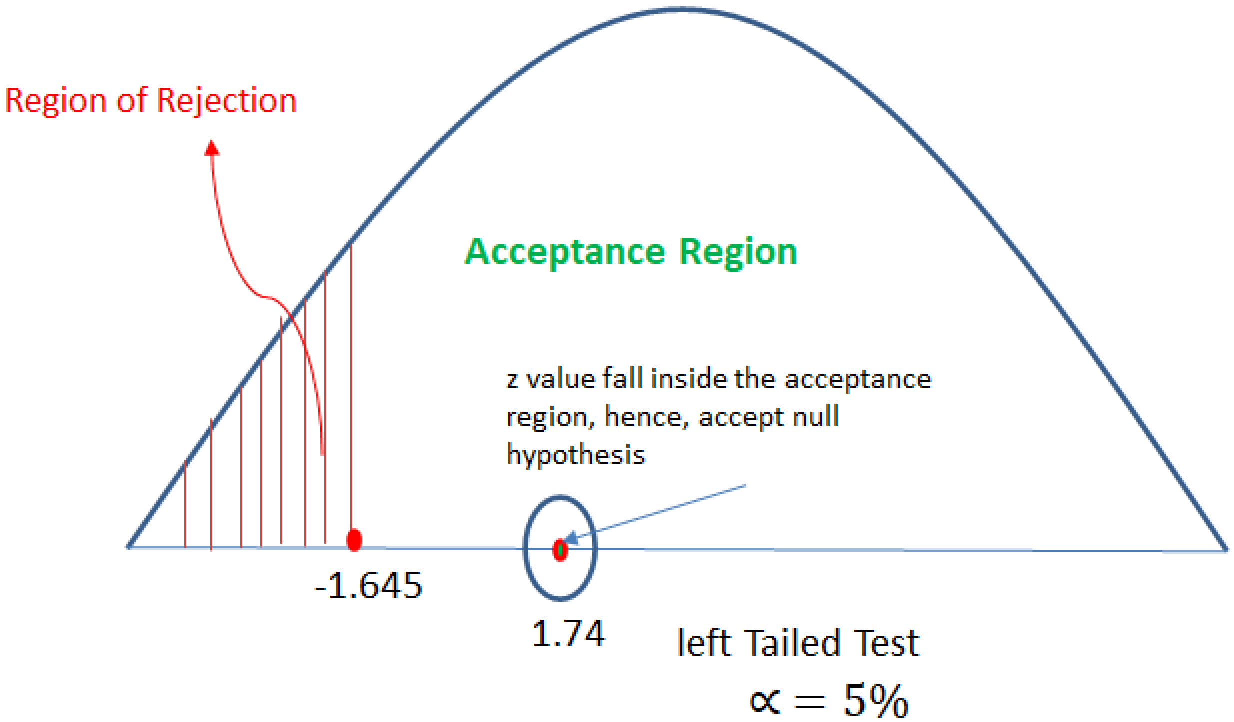
get z value from z table for α = 5%

From z-table, for α = 5%, z = -1.645 (one value as it is one (left) tailed problem)

z(calculated) 1.74 > -1.645 (z from z-table with α = 5%)

<u>Conclusion</u>:

Hence accept the null hypothesis that mean, with given data we can validate significantly that most of the customers are moving towards online shopping even for groceries.



z-test example 2

Question 3

It is found that 250 errors in the randomly selected 1000 lines of code from Team A and 300 errors in 800 lines of code from Team B. Can we assume that team B's performance is superior to that of A.

Solution:

Note down what is given in the question:

There are two samples : Team A and Team B

for each Team some proportion is given in terms of line of error out of total line of code.

Hence this problem can be solved using two proportion z-test.

For one or two proportion type problem we use z-test. (in case of multi-proportion we use χ2 that is chi square test)

Team A (Sample A):

proportion $\mathbf{p_A}$ (small p) = 250/1000 = 0.25

$\mathbf{n_A}$ = 1000

Team B (Sample B):

proportion $\mathbf{p_B}$ (small p) = 300/800 = 0.375

$\mathbf{n_B}$ = 800

Take α = 5% (assume α = 5% if not given in question)

Claim: Team B's performance is superior than Team A which means:

H_0 (Null Hypothesis): overall mean error of Team B $\mu_B < \mu_A$ overall mean error of Team A (with respect to population)

H_1 (Alternate Hypothesis) : $\mu_B >= \mu_A$ (one right tailed test)

Step 1:

calculate z value from two proportion z-test formula as below:

$z = (p_A - p_B)/[p\hat{}(1-p\hat{})(1/n_A + 1/n_B)]$

where **p^** (p hat) = $(n_A*p_A + n_B*p_B)/(n_A + n_B)$

p^ (p hat) = (1000*0.25 + 800*0.375) / (1000 + 800) = 0.305

z = (0.25 - 0.375) / [0.305*(1-0.305)*(1/1000 + 1/800)]

z = -0.125/[0.212*0.00135]

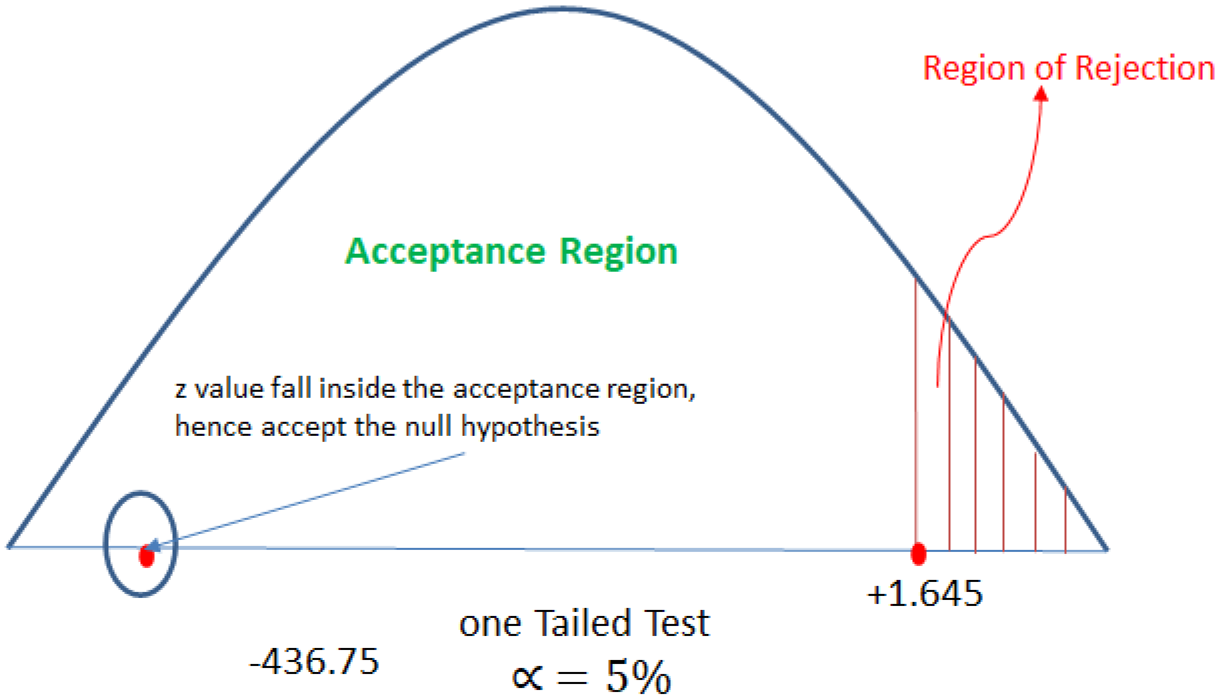
z = -436.75

Step 2:

get z using z-table for α = 5% which is z = +1.645

Now calculated z -436.75 < +1.645

Hence will conclude that null hypothesis is true which mean from given data it is proven significantly that team B's performance is better that team A's performance.



z-test example 3

Question 4

Following is the record of number of accidents took place during the various days of the week.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunc
120	140	200	90	140	120	180

Accidents took place in various days of the given week | Data Science Duniya

Can we conclude that accident s are independent of the day of week?

Solution:

H_0 (Null Hypothesis): Accidents are independent of the day

H_1 (Alternate Hypothesis): Not independent

Here each day will represent a sample and observed accident data is proportion.

Hence this problem can be categorized as multi proportion problem and will be solved using $\chi 2$ chi square test.

Below table shows the Observed values of accident in first column

As we want to validate that accidents are independent of the day of week.

for that average accidents on each day should be different. Hence we need to calculate average accidents on each day and this will be called as expected value in $\chi 2$ test.

Take **α = 5%** (assume α = 5% if not given in question)

Step 1: calculate expected values and $\chi 2$ values using $\chi 2$ formula as shown below in the table

Observed (o)	**Expected (e = average of Observed values)**	**$\chi 2 = \Sigma[(o-e)^2]/e$**
120	110	0.909
140	110	8.181
200	110	73.636
90	110	3.636
140	110	8.181

Observed (o)	**Expected (e = average of Observed values)**	**$\chi 2 = \Sigma[(o-e)^2]/e$**
120	110	0.909
180	110	44.545
Total = 990		**Total $\chi 2$ = 139.99**

χ2 calculation example | χ2 test in hypothesis testing

Step 2: use χ2 table for **α = 5%** and get χ2 value from the table.

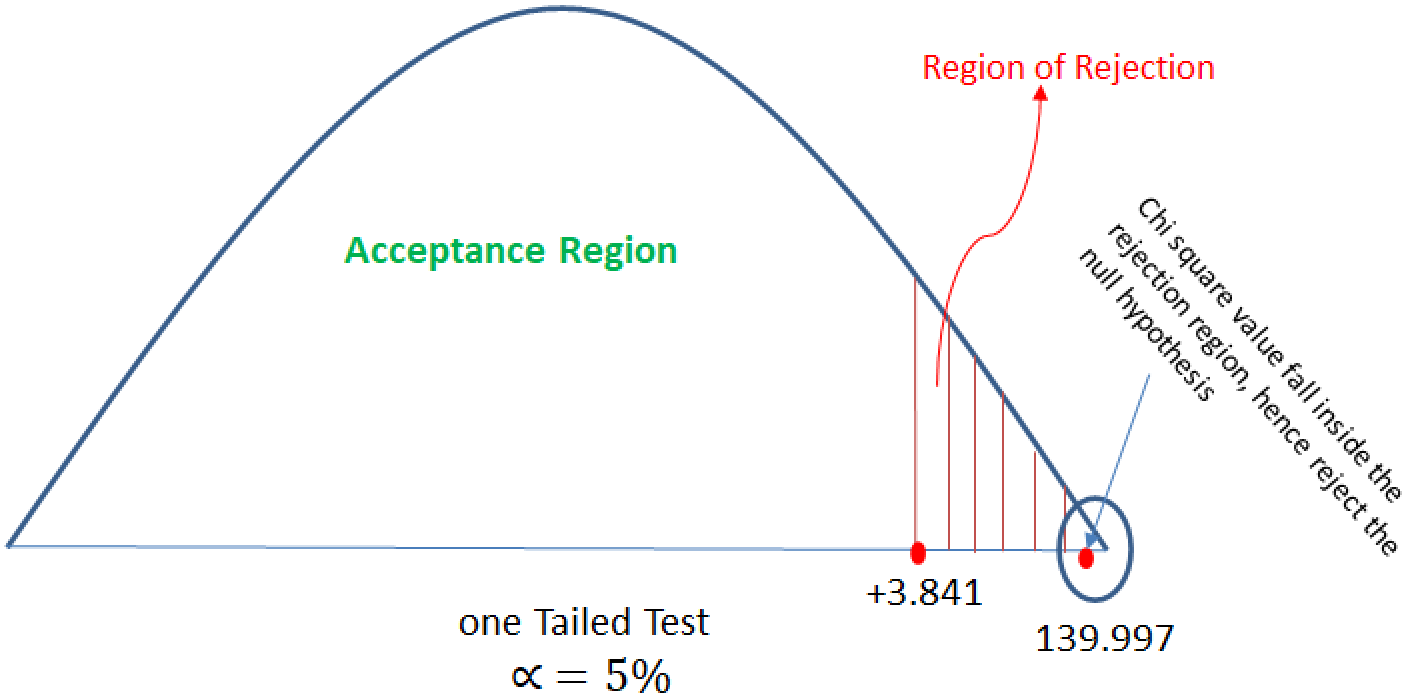
from table we got χ2 (critical value at **α = 5%**) = 3.841

Step 3: compare both χ2 values.

The chi-square value of 18.99 is much larger than the critical value of 3.84, so

the null hypothesis can be rejected.

It means, reject the null hypothesis and accept the alternate hypothesis. Which means with given data we can conclude significantly that accidents are not independent of the day of week. [might not look realistic but with given data is concluding this]



chi square test example 1

Question 5

Analyze the below data and tell whether you can conclude that smoking causes cancer or not?

Category	Diagnosed as Cancer	Without Cancer	Tot
Smokers	400	300	700
Non-Smokers	300	500	800
Total	700	800	1500

chi square test check the independence of the two categorical variable. Here in this question we need to test whether smoking and cancer are independent or dependent to each other. Hence will perform chi square test.

Solution:

Step 1:

H0 (Null Hypothesis): Cancer is dependent on smoking

H1 (Alternate Hypothesis): cancer is not dependent on smoking

Step 2:

Calculate the expected value for each cell of the table (when null hypothesis is true)

The expected values specify what the values of each cell of the table would

be if there is no association between the two variables.

The formula for computing the expected values requires the sample size, the

row totals, and the column totals.

expected value (e) = (row total * column total)/table total

Now lets create another table with observed and expected values both:

Category	Diagnosed as Cancer	Without Cancer	T
Smokers	o = 400, e = 700*700/1500 = 326	o = 300, e = 700*800/1500 = 373	70

Category	Diagnosed as Cancer	Without Cancer	T
Non-Smokers	o = 300, e = 800*700/1500 = 373	o = 500, e = 800*800/1500 = 426	80
Total	700	800	15

Step 3:

calculate the chi square value:

$$\chi 2 = \boldsymbol{\Sigma}[(o-e)^2]/e$$

$\chi 2 = (400-326)^2/326 + (300-373)^2/373 + (300-373)^2/373 + (500-426)^2/426$

$\chi 2 = 16.79 + 14.28 + 14.28 + 12.85$

$\boldsymbol{\chi 2 = 58.2}$

Step 4:

Decide if $\chi 2$ is statistically significant.

The final step of the chi-square test of significance is to determine if the value

of the chi-square test statistic is large enough to reject the null hypothesis.

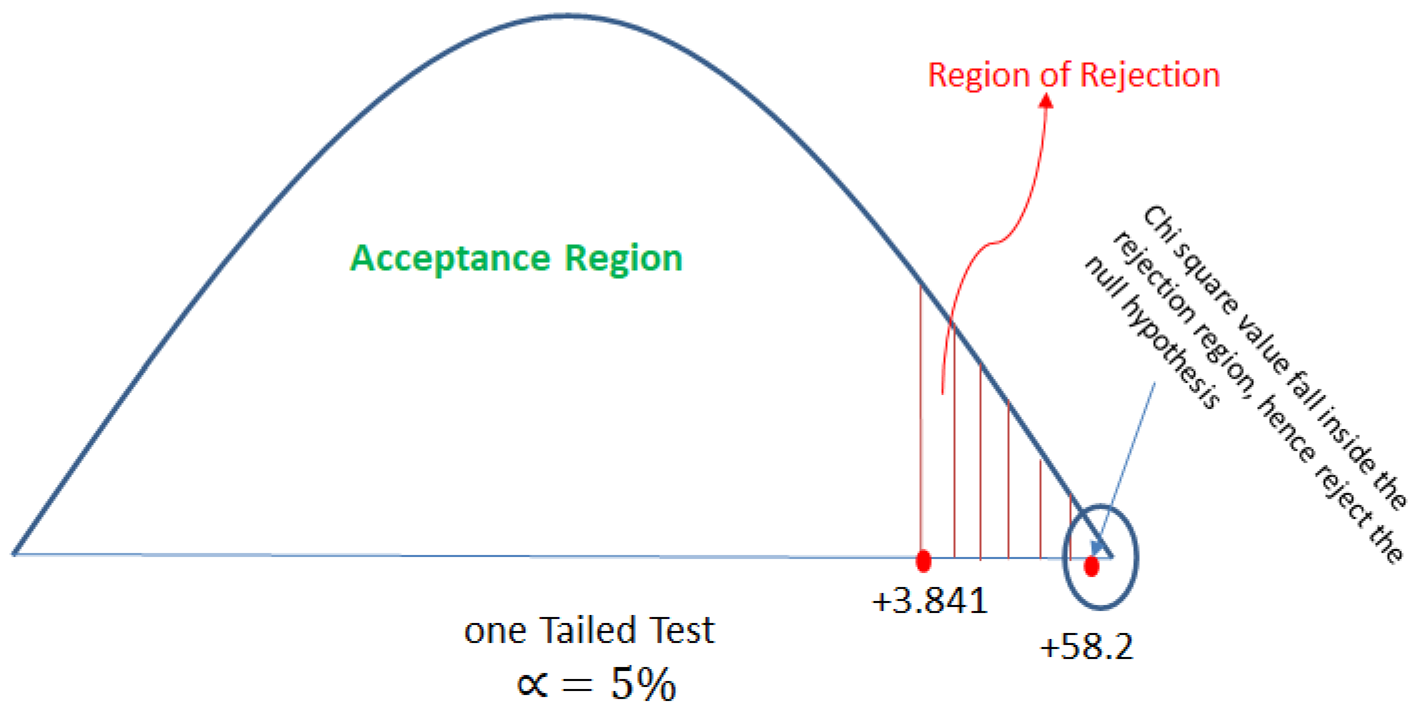
Now will check $\chi 2$ table for the critical value with $\boldsymbol{\alpha = 5\%}$

So from table we got $\chi 2$ (critical value at $\boldsymbol{\alpha = 5\%}$) = 3.841

The chi-square value of 58.2 is much larger than the critical value of 3.84, so

the null hypothesis can be rejected.

Which means with given data, it can be significantly concluded that cancer is not dependent on smoking.



chi square test example 2

Question 6

It is claimed that the mean of the population is 67 at 5% level of significance. Mean obtained from a random sample of size 100 is 64 with SD 3. Validate the claim.

Solution:

First thing first, Note down what is given in the question:

H_0 (Null Hypothesis) : $\mu = 67$

H_1 (Alternate Hypothesis): $\mu \neq 67$ (Not equal to means either $\mu > 67$ or $\mu < 67$ Hence it will be validated with **two tailed test**)

LoS (α) = 5%

n = 100 (Sample size)

xbar $\bar{x}$ = 64 (Sample mean)

s = 3 (sample Standard deviation)

n > = 30 hence will go with z-test

Step 1: Calculate z using z-test formula as below:

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

$z = (64 - 67) / (3/\sqrt{100})$ (in question population standard deviation is not given, in that case take sample standard deviation)

$z = -10$

step 2:

calculate z critical value for $\alpha = 5\%$ from z-table.

so from z-table Z critical value = -1.96, +1.96 (will get two values due two tailed test)

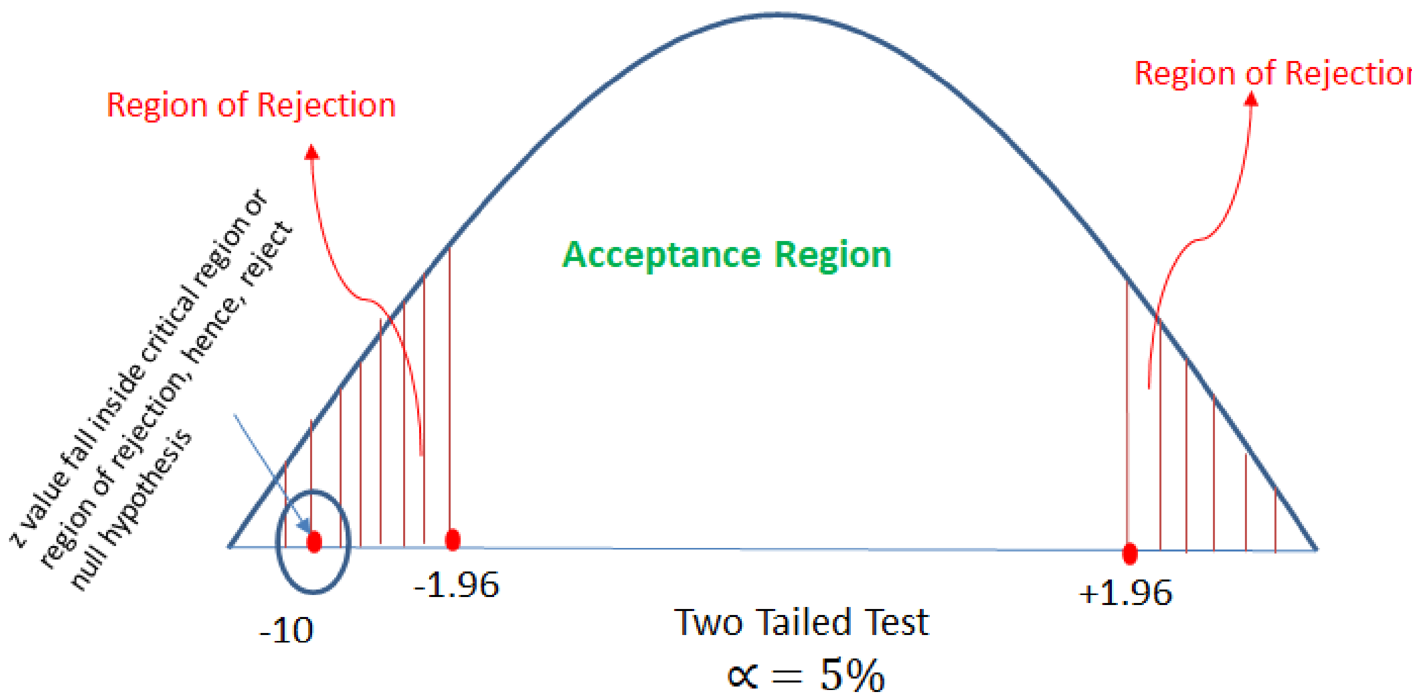
step 3:

check if calculated z value is in between z critical value then accept the null hypothesis if z calculated is outside z critical then reject the null hypothesis.

Here, z calculated value = -10 which is much lesser than the left side z critical value -1.96, hence will reject the null hypothesis.

Conclusion:

with given data it is significantly proven that population mean is not equal to 67.



z-test example 4

Question 7

There is an assumption that there is no significant difference between boys and girls with respect to intelligence. Tests are conducted on two groups and the following are the observations

	Mean	**Standard Deviation**	**Size**
Girls	75	8	60
Boys	73	10	100

Validate the claim with 5% LoS (Level of Significance)

Solution:

First thing first, Note down what is given in the question:

H_0 (Null Hypothesis) : No difference between boys and girls in terms of intelligence. (**μ_2** = **μ_2**)

H_1 (Alternate Hypothesis): Boys and girls are different in terms of intelligence (**μ_2** ≠ **μ_2**) => two tailed test

x1bar = 75 (boys sample mean)

x2bar = 73 (girls sample mean)

LoS (α) = 5%

In question, we have two sample mean. Boys sample mean and girls sample mean. Hence this can be solved with two mean problem.

Next both samples size n1 = 60 and n2 = 100 are greater than 30 hence will use z-test.

Step 1:

calculate z value from the two mean z test formula as below:

$z = [(x_1bar - x_2bar) - (\mu_2 - \mu_2)]/\sqrt{}(s_1^2/n_1 + s_2^2/n_2)$

$\mu_2 - \mu_2 = 0$ assuming null hypothesis is true

$z = (75-73)/\sqrt{}(8^2/60 + 10^2/100)$

z = 1.39

step 2:

calculate z critical value for $\alpha = 5\%$ from z-table.

so from z-table Z critical value = -1.96, +1.96 (will get two values due two tailed test)

step 3:

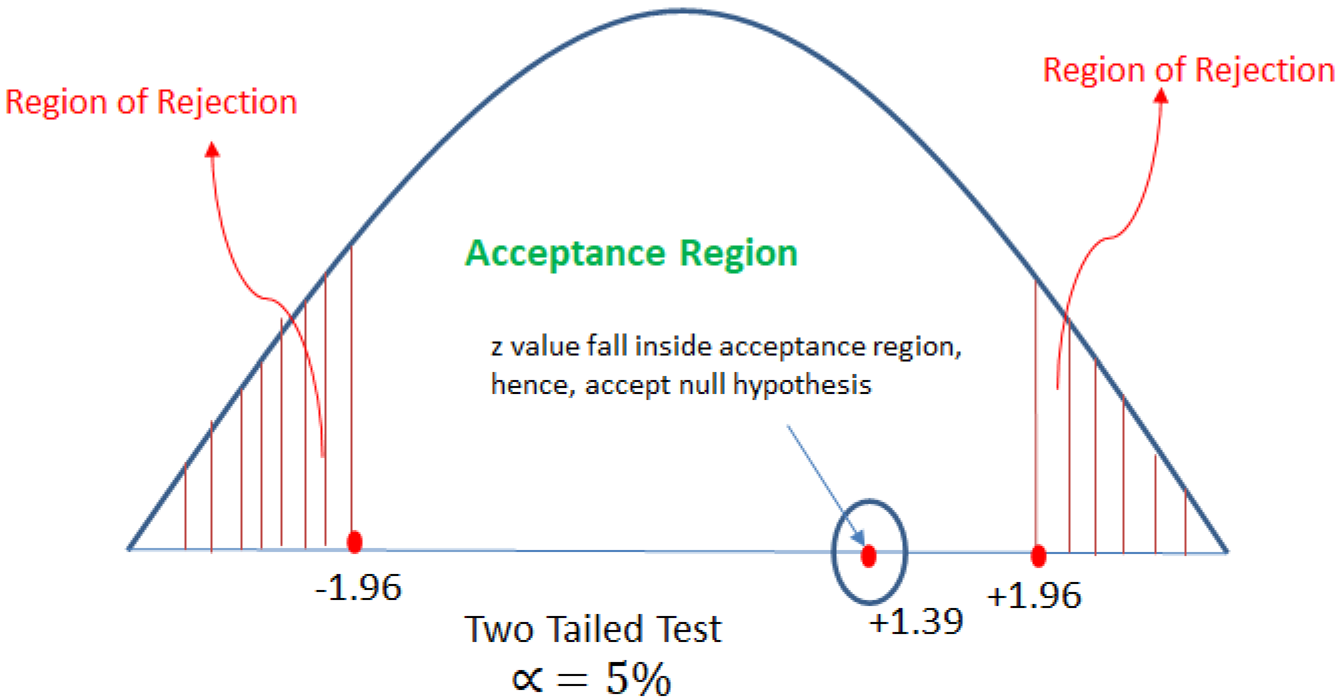
check if calculated z value is in between z critical value then accept the null hypothesis if z calculated is outside z critical then reject the null hypothesis.

Here, z calculated value is in between the z critical values. $-1.96 < 1.39 < 1.96$

Hence will **accept the null hypothesis.**

Conclusion:

with given data it is significantly proven that there is no significant difference between the intelligence of boys and girls.



z-test example 5

Question 8

An automobile tyre manufacturer claims that the average life of a particular grade of tyre is more than 20,000 km. A random sample of 16 tyres is having mean 22,000 km with a standard deviation of 5000 km.

Validate the claim of the manufacturer at 5% LoS.

Solution:

First thing first, Note down what is given in the question:

H_0 (Null Hypothesis) : $\mu > 20000$

H_1 (Alternate Hypothesis): $\mu <= 22000$ (less than mean one tailed test)

LoS (α) = 5% (Take 5% if not given in question)

n = 16 (Sample size)

$\bar{x} = 22000$ (Sample mean)

s = 5000 (sample Standard deviation)

n < 30 hence will go with t-test

step 1:

calculate t value from the t-test formula:

$t = (\bar{x} - \mu)/ (s/\sqrt{n})$

$t = (22000 - 20000) / 5000/\sqrt{16}$

t = 1.60

step 2:

get t critical value from t-table for α = 5% and degree of freedom = 16-1 = 15.

t critical value = 1.753

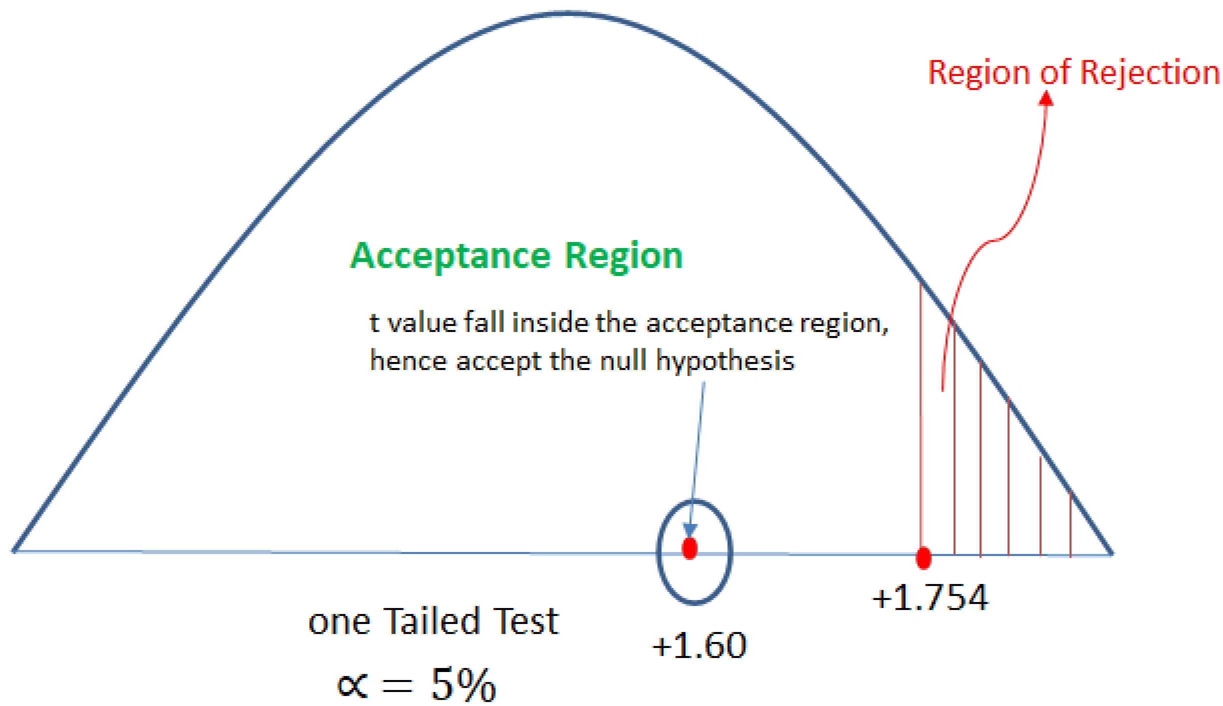
step 3:

check if t calculate < t critical then accept the null hypothesis else reject the null hypothesis.

Here, t calculated 1.60 < t critical 1.753, hence will accept the null hypothesis.

Conclusion:

from the data given, it is significantly proven that average life of the tyres is more than 20000.



t-test example 1