

Hypothesis : A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.

28/10/2019

## 10. Hypothesis Testing

— Manoj M

### \* Inferential Statistics

\* Null Hypothesis  
Alternate Hypothesis

Alpha Value

Critical value

p value

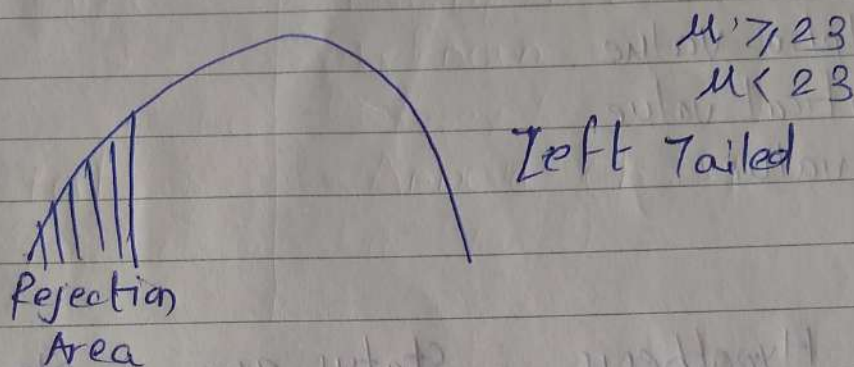
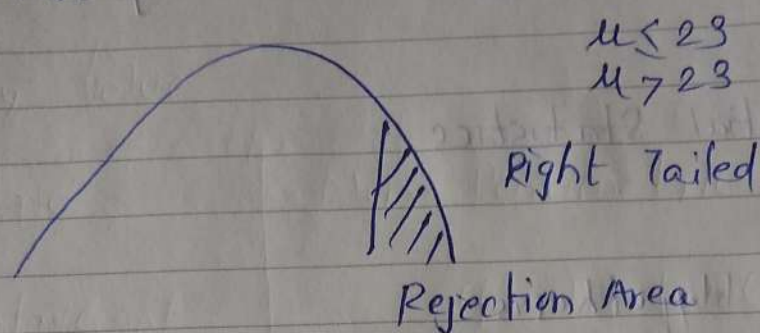
\* Null Hypothesis : Status quo or No difference.

e.g. I am assuming a machine gives 5 Gold coins a day. And someone tells me that it does not give 5 Gold coins.

Hypothesis	Two Tailed Test	One Tailed Test	One Tailed Test
Null	mean = 5 coins	mean $\geq$ 5 (Right) (left)	mean $\leq$ 5 (left) (right)
Alternate	mean is not equal to 5 ( $>$ or $<$ )	mean $<$ 5	mean $>$ 5

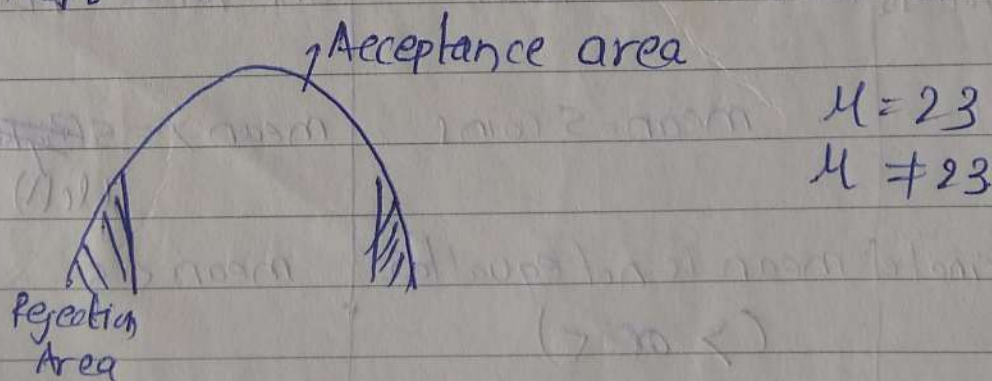


## \* One Tailed Test



→ In one Tailed Test, Rejection area falls on only one side.

## \* Two Tailed Test



→ In Two Tailed Test, Rejection area falls on both the sides.



\* Alpha Value (Margin of error)

→ level of significance.

→ Rejection Area

→ A cut off or Threshold level / limit for rejecting Null Hypothesis.

→ 5%

→ If Margin of error is 5% then confidence level is 95%.

→ In Two Tailed Test, if  $\alpha$  I assume margin of error is 5%. then ~~2.5%~~ Rejection will be 2.5% on both the sides. <sub>area</sub>

→ Cut-off or Threshold <sup>limit</sup> will vary depending on complexity ~~an~~, study & outcome.

→ 5% is universally accepted



## Two data Types in statistics

Quantitative : Continuous  
Discrete

Qualitative : Categorical  
(e.g. Gender)

\* p value

→ probability value

→ Parametric

Non-Parametric

Quantitative

Qualitative (Categorical)

about  
Much of the data  
is known

Not much about data  
is known.

T, F and z Test

ANOVA, chi sq.

## \* T-Test

(Parametric Test)

- Framing a "Null Hypothesis" is important
- If Hypothesis is like assuming a "Null Hypothesis" and then trying to find the evidence to prove "Alternate Hypothesis" is correct.
- t-Test is used to determine means between 2 groups are same or not.
- It talks about degree of randomness
- Hypothesized Mean Difference  
Difference of mean between sample and population
- In Inferential Statistics, every machine will consider dataset as sample. (or program)



→ Mean : Average

→ Variance : How spread the data is spread from the mean

→ Observation : Count

→ Pearson Correlation : How strongly two variable/datasets are co-related  
`Correl (Array 1, Array 2)`

→ df : degree of freedom ( $n-1$ )

$(n-1)$  values are closely distributed and are used in statistical calculation

\* If p value  $\leq$  alpha value : Reject Null Hypothesis  
(statistically significant)

\* If p value  $>$  alpha value : Accept Null Hypothesis  
(significantly insignificant)  
statistically

\* If test stat is  $<$  critical value : Accept Null

\* If test stat is  $>$  Critical value : Reject Null



→ p value : T-TESTC Array1, Array2, tails, type)

tails = 1 or 2

type = paired

Two sample equal variance

Two sample unequal variance

→ t-critical ~~to~~ one tail (Cut off value)

Calculated from t-Distribution Table

but have know alpha value & degree of freedom

→ t-statistics

The t-statistics is the ratio of the departure of the estimated value of a parameter from its hypothesized value to its standard error.

It is used in hypothesis testing via student's t-Test.

F-Test

The purpose of the F-Test is to check if the variance of the 2 separate groups are equal or not.

Variable 1 Range: Group with higher variance.

Null = Equal Variance

Alternate = Unequal Variance

} always same for  
F-Test



## Z-Test Versus T-Test

- Z-Test is a statistical hypothesis test that follows a normal distribution while T-Test follows a Student's T-distribution.
- A T-Test is appropriate when you are handling small samples ( $n < 30$ ) while a Z-Test is appropriate when you are handling moderate to large samples ( $n > 30$ ).
- T-Test is more adaptable than Z-Test since Z-Test will often require certain conditions to be reliable. Additionally, T-Test has many methods that will suit any need.
- T-Tests are more commonly used than Z-Tests.
- Z-Tests are preferred than T-Tests when standard deviations are known.

→ How to choose Null Hypothesis?

⇒ e.g. A Criminal is convicted for a crime.

If the Judge thinks, he is not guilty then no need for a trial. (Means no need to collect the evidence).

Null: Guilty

Alternate: Not Guilty

So, I will the evidence of criminal is Guilty or not.



31 / 10 / 2019

## 11. Hypothesis Testing Case Study

— (Manoj M)

### \* Formula for skewness

= Skew	(For Sample)	(n-1) observations
= Skew · P	(For Population)	n observations

### \* Histogram

- Continuous data
- formation of bins (clustering the data)
- Very much presentable as compared to line chart, bar chart or scatter plot

\* We can use "frequency" formula

= frequency (data-array, bins-array)

data-array  $\Rightarrow$  data

bins-array = bins

Select All  $\rightarrow$  Write formula  $\rightarrow$  click  
ctrl + shift + Enter (Array fill)

$$F \text{ Stat value} = \frac{\text{Highest Variance}}{\text{Lowest Variance}}$$

\* Formula for standard deviation

Standard deviation (For sample)  $\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$   
 Standard deviation (For population)  $\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$

\* Histogram

→ Comparison of data

→ Formation of bins (dividing the data)

→ Very much preferable as compared to line chart, bar chart or scatter plot

\* The code for "pandas" formula

- pandas.DataFrame(array, index=array)

array = data

index = index