Hypothesis Testing

Fundamental Test in Statistics.

KASI PRADEEP .J.
NOTES

4 Steps in Data-Driven Decision Making

(i) Formulates a Hypothesis.

(ii) Find the Right Test

(iii) Execute the test

(iv) Make a decision

Hypothesis :- An idea that can be tested.

eg:- Apples in Newyork are Expensive. ~~A~~

↳ Its a statement, or an idea but Not testable

eg:- Trump Vs Hilary Clinton administration :- Who is going to be best?

No data
↓
Can't be tested
↓
Not a Hypothesis.

eg:- Bush ~~Clinton~~ Vs Obama administration :- Who is best.

Data is Present
↓
Can be Tested
↓
Might be Hypothesis.

eg:

Glassdoor Website:- Mean Salaries of Data Scientiest 1,50,000/-
⇒ (Personal Opinion ⇒ Data Scientists doesnot earn that much)

Hypothesis	Notation	
* Null Hypothesis	H_0	→ One to be tested
* Alternative Hypothesis	H_1 or H_A	→ Remaining (Everything)

$H_0 : \mu_0 = 1{,}50{,}000/-$

$H_1 : \mu_0 \neq 1{,}50{,}000/-$

(May < or >)

Accept if : $\bar{x}$ is close to the true Mean.
Reject if : $\bar{x}$ (Mean) is ~~close~~ far from true Mean.

NOTES

The concept of Null Hypothesis is "Innocent Until Proven Guilty"

H_0 is True Until Rejected.

Ex:- $H_0 : \mu_0 = 1,50,000/-$ Two Sided Test

We can also form One sided Tests.

*1. Friend

He says Avg Salaries of Men is more than 1,40,000/-

2. Me

I doubt him. So I have start testing

Status Quo $H_0 : \mu_0 \geq 1,40,000/-$ (Equality Sign)

Change or Innovation $H_1 : \mu_0 < 1,40,000/-$

* Outcome of Tests Refer to Population Parameter Rather than Sample Statistic

* The Researcher is trying to Reject the Null Hypothesis.

* Null Hypothesis is the statement We are trying to Reject.

* Therefore The Null is the Present state of Affairs while the alternative is our Personal Opinion.

Quiz:-

Significance Level (α)

The probability of Rejecting the Null Hypothesis, if it is true.

Typical Values of α is 0.01, 0.05, 0.1

$\uparrow$ Most Commonly used. (0.05)

eg: Cocacola Cool drink Manufacturing?

12 to 12.1 units increasing qty ounces in every bottle.

May Spill, Label cover may damage. Here α is to be more accurate $\cong 0.01$.

How much Coca Cola consumers drink?

Ans:- 12 to 12.1 ounces is Not a problem.

Here the error level can be more ie, Significance level can be > 0.05.

One-Sided Test :-

eg:- Analysis on How students are performing on average?

Dean :- Population Mean grade is 70% (H_0)

Me Researcher (Testing) :- Population Mean grade is not 70% (H_1)

$H_0: \mu_0 = 70\%$

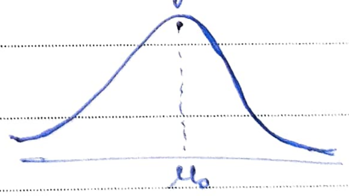
$H_1: \mu_0 \neq 70\%$

Distribution of Grades :- Assuming Population of grades is Normally Distributed.

Sample Mean, Hypothesis Mean

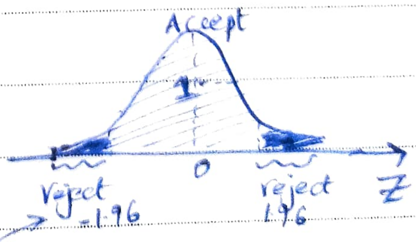
z Test: $Z = \dfrac{\bar{x} - \mu}{(S/\sqrt{n})}$

$(S/\sqrt{n}) \rightarrow$ Std. Error.



$\bar{x} \cong \mu \Rightarrow Z = 0$

$\alpha = 0.05$



rejection Region $\alpha/2 = 0.025$ (One side)

0.025 (other side) rejection region

$\alpha/2 = 0.975$

96

$\cong 1.96$

19 — 0.975

95% $(-1.96, 1.96)$

1. Calculate a statistic (eg: $\bar{x}$)
2. Scale it (eg: $Z = \dfrac{\bar{x} - \mu}{(S/\sqrt{n})}$)
3. Check if Z is in the rejection Region

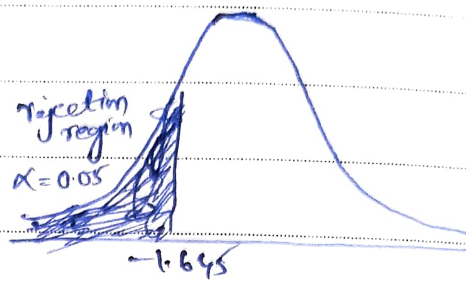
NOTES

Onesided Test :-

① Paul My friend says Min. Salaries of Data Scientist is 1,25,000/-

$H_0 : \mu_0 \geq 1,25,000/-$

$H_1 : \mu_0 < 1,25,000$



$\alpha = 0.05$
$1-\alpha = 0.95$
.04
$\Rightarrow 1.64$
0.95
1.6

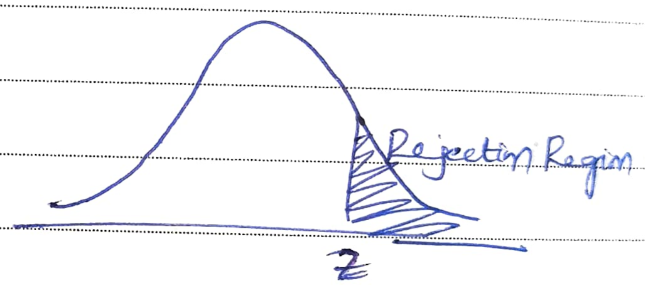
If $z < -1.645$, We Would reject the Null Hypothesis

② University Dean :-
Avg GPA Students get Lower than 70%

$H_0 : \mu_0 \leq 70\%$

$H_1 : \mu_0 > 70\%$

Rejection Region is Right Side



If the test statistic $\underset{(x)}{}$ is bigger than the β Cutoff Z-Score, we would ~~the~~ reject the null, otherwise we wouldn't

* Errors in Hypothesis Testing < Type I Error / Type II Error

~~Type I~~

Type 1 Error :- When we reject a True Null Hypothesis. It is also Called false +ve
probability of this error is α

Type II Error :- Accept a false Null Hypothesis. Denoted by β
Depends on Sample size (n) and Variance (σ)

* probability of Rejecting a false Null Hypothesis $= 1-\beta$

$1-\beta$ is Called **Power of a Test** *

Researchers ↑es Power of a Test by increasing the Sample size eg: U may in Love with One girl. But you are not sure whether that girl is loving or not?

NOTES

Ho :- Status Quo.

The truth.

	Ho is True	Ho is False	
Ho (Status Quo) Accept	✓	Type II Error (False Negative)	P.E = β
She doesn't like you (You shouldn't invite her out) Reject	Type I Error (False Positive) P.E = α	✓	

Ho : She doesn't like You.

The truth

	Ho is true	Ho is False
Accept (Do nthg)		
Reject (Invite her)		

	She doesn't like You	She Likes You	
Accept (Do nthg)	✓	Type II Error (False -ve)	→ Missed Your Chance
Reject (Invite her)	Type I Error (False +ve) (Wrongly Invited Her)	✓	

We don't want to make any of these 2 Errors, but it happens sometimes.

Statistics is useful but not perfect.

NOTES

Test for the Mean

~~Sample~~ Single Population (1)   Multiple Population

(1) Known Variance (Z)   (2) Unknown (T)

1. Test for the Single Mean. Population Variance Known

eg:- Data scientist salary

Dataset

117,313
104,022
⋮
114,426

Sample Mean = 100,200
Pop. std Known = 15000/-
Std Error = 2739 = $\frac{15000}{\sqrt{30}}$
Sample Size = 30

Glassdoor Website
Mean Salary = 113,000/-
(Based on Self Reported No.)
↓
We have to Check whether this value is correct.

$H_0: \mu_0 = 113000/-$

Two sided Test <a> $H_1: \mu_0 \neq 113000/-$

Testing is done by standardizing the variable at hand and comprising it to the Z.

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{100200 - 113000}{\frac{15000}{\sqrt{30}}}$$

Sample Mean → $\bar{x}$, Mean of Interest → μ_0

Uppercase (Uppercase)
Z → Standardized variable associated with the test Called Z Score.
z → One ~~from~~ from the table and will be refered to as "Critical value"
Lower Case

$Z \sim N(0,1)$

$Z \sim N(\bar{x} - \mu_0, 1)$

~~Compare Means~~ Standardization → Lets compare the means.

$Z = -4.67 = |-4.67| = 4.67$

Now we will compare $|-4.67|$ with $z_{\alpha/2}$

where α is Significance level.

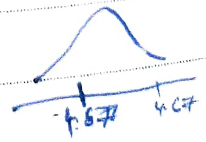
z-tables doesnt include -ve values

As the Standard Normal Distribution is symmetrical around 0, the two Statements are equivalent.

$-4.67 <$ a negative z $\Longleftrightarrow$ $4.67 >$ a positive z

Decision Rule:-

Absolute Value of Z score > Positive Critical Value (z) → Reject



Test at 95%

Using 5% significance

$\alpha = 0.05$

$\boxed{1 - \alpha/2 = 0.975}$

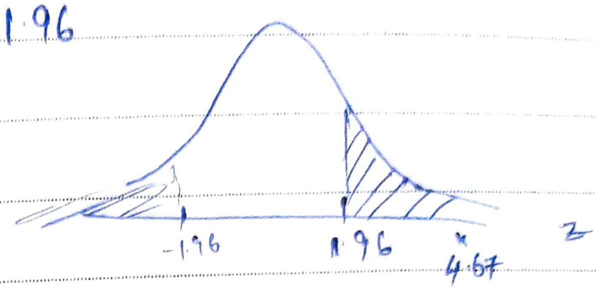
~~$Z_{\alpha/2} = 0.025$~~ $Z_{0.025} = 1.96$

0.06

1.9 — 0.975

$Z_{score} > Z$

4.67   1.96



$\Rightarrow$ We Reject the Null Hypothesis.

Test at 99%

Using 1% significance level.

$\alpha = 0.01$ $Z_{0.005} =$

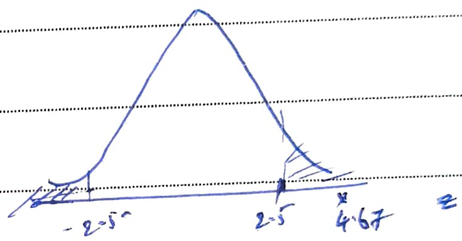
$1 - \alpha/2 = 1 - 0.005$

$= 0.995$

9.8

2.5

$4.67 > 2.58$



$\Rightarrow$ We Reject Null Hypothesis

* Technique for Significance Level Choosing:-

P-value :-

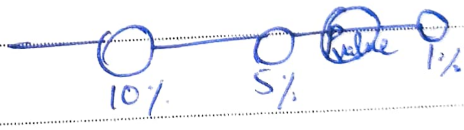
P-value is the smallest level of significance at which we can still reject the Null Hypothesis, given the observed sample statistic.

Eg:-   Standard Error = 2739

$n = 30$

Data Scientist ex: Pop. std = 15000

$N \sim (\mu, \sigma^2)$



We rejected the Null Hypothesis at 0.01 & 0.05 Significance Level

But We wanted to know How much lower We can go?

p-value = 1 - (number of Table) = 1 - 0.9990 = 0.001

(Least Value)

$Z_{score} = -4.67$ $\rightarrow$ Can't be table, As value is large.

Rule:- You should reject the null Hypothesis, if p-value < α ~~$1 - \alpha/2 = 0.999$~~

$\alpha/2 = \alpha$

	p-value α	
Test at 90% :	$0.001 < 0.1$	Rejects the Null Hypothesis.
Test at 95% :	$0.001 < 0.05$	
Test at 99% :	$0.001 < 0.01$	

NOTES

Eg: ~~Rule:~~

Z score = 2.12

0.017 < 0.001 ✓ reject

0.017 < 0.05 ✓ reject.

0.017 < 0.01 ✗ can't reject

Test at 90% } ⇒ Reject
Test at 95% }
Test at 99% } → Cannot Reject.

2.1 — 0.02 — 2.12 $\boxed{0.9830}$

One sided Test p value = 1 − No from Table = 1 − 0.983 = 0.017

Two Sided Test p value = 2 × (1 − No. from Table) = 2 × 0.017 = 0.034.

Where & How p-values are used?

* Most statistical softwares calculations p-values for each test.
* Researchers decide whether the variable statistically significant or not
* p-values usually found in 3 digits after the dot x.xxx.
* The closer to 0.000 the p value, the more significant the result we obtain.

* P-value is a Universal Concept that works with Every Distribution (Z, T, Uniform, Binomial)
* If the p-value is lower than the level of significance
  ⇒ We reject the Null Hypothesis

Use online p-value Calculators to Support $\boxed{000}$

*

2. Test for Single Mean, Population Variance Known

eg:- You are a Market Analyst of One Company & You are asked to Compare email Open Rate above of our competitor company.

Your company has email open rate of 40%.

Email Open Rate : Measure of How many People on Email list actually open the emails they have received.

Your Competitor Company employee posts a something on facebook regarding appreciating email Management software of that company. In Background an image clearly sees ~~10~~ last 10 email campaigns and their corresponding email open Rates.

Sol:-

$H_0 : \mu_{OR} \leq 40\%$

$H_1 : \mu_{OR} > 40\%$

Significance level 0.05%

Open Rate

26%

:

40%

Sample Mean = 37.7%

Sample Std Dev = 13.74%

$n = 10$

Std Error = 4.34%

Null Hypothesis Value = 40%

We don't Pop. Variance & Small No. of Sample ⇒ T Statistics.

$$T = \frac{\bar{x} - \mu}{(S/\sqrt{n})} = |-0.53| = \frac{37.7 - 40}{4.34} \quad 0.53$$

degrees of freedom = 10 - 1 = 9

0.05 one sided significance.

0.05

9 1.833

$t = 1.83$

$t < T$

Tscore < t

0.53 < 1.83

$|-0.53| < 1.83$ ⇒ Can't reject Null Hypothesis

NOTES

α (Significance Level) < p-value

T score < t_value

** More { p value > α significance level ~~Calculated~~ T score
(or) t value (Value from only & Sample) α > T score.

We can't reject.
Hypothesis is Correct

p-Value = 0.304 (Online p-value Calculator)

> 0.05

~~n = 30~~ d of = 9
T score = 0.53
α = 0.05

Can't reject.

Significance level = 0.01

0.304 > 0.01 We can't reject Null Hypothesis

0.05 > 0.01

If we cannot reject the test at 0.05 Significance level, we cannot reject it at smaller levels either.

t-distribution for 2 sided Test

df	One sided α	Two sided $\alpha/2$
	0.05	0.025 ✓
	0.01	0.005 ✓ (We Consider in T-table)
p-value	0.304	0.608 2 × (0.304)

NOTES

Test for the 2 Means, Dependent Samples :-

eg:

Patient	Drug Taken Before	Drug Taken After	Diff (B-A)
1	2	1.7	0.3
2	:		:
:	:		:
10	1.5	2.4	-0.9

Sample Mean = -0.33
Std. Deviation = 0.45
Std. Error = 0.14

Population Mean before ≥ Pop. Mean after

$H_0: \mu_B \geq \mu_A \Rightarrow H_0: \mu_B - \mu_A \geq 0 \Rightarrow H_0: D_0 \geq 0$

$H_1: \mu_B < \mu_A \Rightarrow H_1: \mu_B - \mu_A < 0 \Rightarrow H_1: D_0 < 0$

1. Small sample
2. Variance unknown
3. We assume Distribution is Normal

$\Rightarrow$ t-statistic.

$$T_{score} = \frac{\bar{d} - \mu_0}{\text{Std. Error}} = \frac{-0.33 - 0}{0.14} = -2.29 = |-2.29| = 2.29$$

~~5% Signifi~~

p-value = 0.023 (Online p-calculator)

5% Signifi p value < ~~T score~~ 0.05 $\Rightarrow$ Can reject the Null Hypothesis.

The drug qty in body without taking ≤ After taken.

1% Signi.

p-value = 0.023 < 0.01 (X) ~~(False)~~ $\Rightarrow$ We can't reject the Null Hypo

$\Rightarrow$ Accept the Null Hypothesis.

The lowest value we can reject is 0.023 only.

~~Note:~~

NOTES

Test for 2 Means, Independent samples with Known Variance.

Engg. M.B.A students in University.

~~Size~~ Dean announced Engg is toughest Dept. Engg. Students will get lower grades. Dean believes On, Avg. Management Students Outperforms Engg students by 4%.

Ans: $H_0: \mu_E - \mu_M = -4\%$

$H_1: \mu_E - \mu_M \neq -4\%$ ⇒ Two sided test.

	Engg.	Management	diff
Size	100	70	-
Mean	58%	65%	-7%
Pop Std	10%	6%	$1.23\% = \sqrt{\frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}}$

Hyp. diff -4%

Big Samples, Known Variances ⇒ Z statistics.

Small samples, Unknown Variances ⇒ T-Statistic

Big Samples, Unknown Variances ⇒ Upto Researcher. But generally Z-preferable

$$Z_{score} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}}} = \frac{-7\% - (-4\%)}{1.23\%} = -2.44$$

p-value (online) $0.0146 \cong 0.015$

~~95~~ $0.015 < 0.05$ (We reject Null Hypothesis)

⇒ The diff. is not 4%

*** But <4% or >4% ???

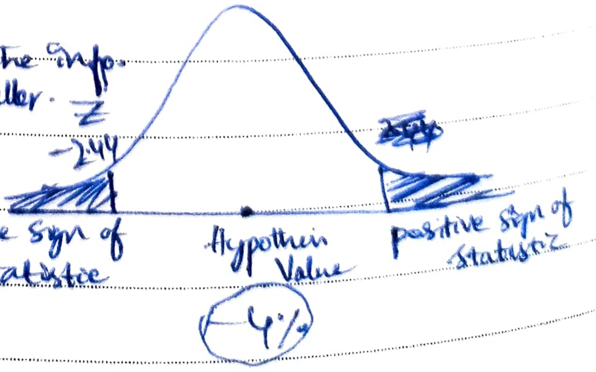
* Sol:- A sign of Test Statistic (Z-score) will give the info.

A minus sign of Test Statistic means its Smaller.

~~True~~ True value is lower than Hypothesis Value.

lower than -4%

i.e -5%, -6%



2^{nd} Question :- Is the price of apples in NY is 20% higher than that in L.A ?

NOTES

Test for 2 Means, Independent samples, with Unknown Variance & Assumed to be equal :-

eg:	Newyork	L.A. - Apples Shop
1.	3.8/-	3.1/-
2.	3.4/-	
		3.2/-
10.	3.62/-	

$H_0: \mu_{NY} = \mu_{LA} \Rightarrow H_0: \mu_{NY} - \mu_{LA} = 0$

$H_1: \mu_{NY} \neq \mu_{LA} \Rightarrow H_1: \mu_{NY} - \mu_{LA} \neq 0$

	NY	LA
Mean	3.94	3.25
Std. dev.	0.18	0.27
Sample size	10	8

Pooled Variance

$$S_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2} = \frac{(10-1)0.0324 + (8-1)(0.0727)}{10+8-2}$$

$= 0.05$

~~Std. Error = $\sqrt{0.05}$ = 0.22~~

Pooled Std. Dev = ~~$\sqrt{0.05}$~~ $\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} = \sqrt{\frac{0.05}{10} + \frac{0.05}{8}} = 0.11$

Small sample
Variance Unknown $\Rightarrow$ T statistic.

Degree of freedom $= 10 + 8 - 2 = 16.$

$$T = \frac{\bar{d} - \mu_0}{\text{Std. Error}} = \frac{(3.94 - 3.25) - 0}{0.11} = 6.53$$

~~P value~~

Rule of Thumb :- Reject the Null Hypothesis when the T score is bigger than 2.

Generally for Z and T score, a value Higher than 4 is extremely significant.

P-Value 0.000
(online Calculator).

Prob. of making Type I is 0.

We reject the Null Hypothesis at all common and uncommon levels of significance

We could easily say that the prices are different by seeing. No Testing required.

Mean cost of Apples in Newyork > L.A

NOTES

Practical ~~exa~~ of Hypothesis Testing

Gender Pay Gap

Company :-

$N = 5000+$ (Total employees)

$n = 174$ (We have sample data of 174)

Name	Gender	Salary	Age	designation
	M			
	M			
	⋮			
	F	More highest	63	CEO
	F			

Sample

Male | Female

Independent samples

$H_0: \mu_m - \mu_f = 0$

$H_1: \mu_m - \mu_f \neq 0$

t test to be used ~~for~~ here. Variance unknown
We assume it is equal

Overall	n	Mean	Sample Variance	Pooled Variance	T Score	p value
female	98	65,736	935,705,380	1,025,188,119	1.34	0.18
male	76	72,300	1,144,799,128			

$$T = \frac{(\bar{x}-\bar{y}) - (\mu_m - \mu_f)}{\sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}}} = \frac{(72300 - 65736) - 0}{\sqrt{\quad}}$$

p value < ~~T Score~~ α

~~→ Reject Null Hypothesis~~

Test 99% → p value 0.182 > 0.01

95% → 0.05

90% → 0.1

$\left.\begin{matrix} 0.01 \\ 0.05 \\ 0.1 \end{matrix}\right\}$ Can't reject Null Hypothesis.

Avg Male Salary ≈ Female Salaries

Now we are dividing further.

>35 yrs age → Male, Female

<35 yrs age → Male, Female

Below 35	n	Mean	Sample Var	Pooled Vari	T-Score	p-value
Female	46				0.43	0.668
Male	37					

Concl: There is no wage gap on Gender basis

Over 35	n	Mean	Sample Var	Pooled Var	T-score	p-value
Female	52				2	0.048
Male	40					

~~Below 35:~~ 95% $0.048 < 0.05$ (reject Null Hyp.)

Older employees ⇒ There is wage gap.

Its a 2 sided Test.

We dont know who will get more money

My Equation:

$$\boxed{\mu_m - \mu_f = T = 2 \ (+ve)}$$

$\mu_m > T_s + \mu_f$

If T_s +ve. $\mu_m > \mu_f$ (Here) ✓

If T_s -ve $\mu_m < \mu_f$

Avg. Salaries of Male will get more.