Inferential Statistics – 01

1:30

5min

Probability and probability Distributions

Gunther stende

Statistic:

① Descriptive

② Inferential → Hypothesis Testing.

Notes

population mean = SM (sample) is mostly Incorrect.

PM = SM ✗

$\boxed{PM = SM \pm \text{Margin of Error}}$ Can be Correct.

5 years

* Always try to keep No. of sample ~~2000~~ at least 200 or ≥ 200

* According to Central limit theorem each samples must have 30 data points.

* Continuous data col → numerical value → float → decimal

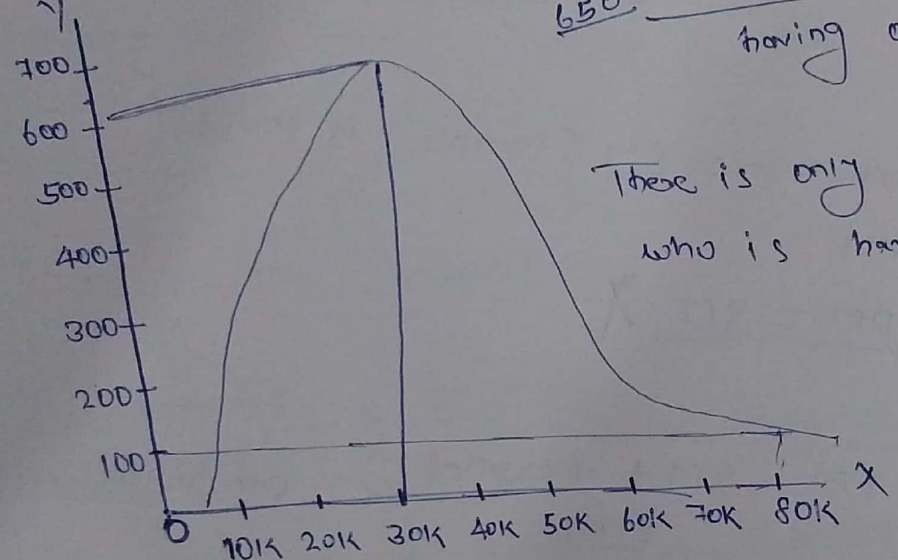
Goodluck !!!

Kde — kernel density estimate.

1:16 * When your have very extreme value at Right. your graph is positively skew.

(C) Positively Skewed

650 No of employee having only 30k Salary



There is only 100 employee who is having 80 Salary.

So, Analysis of graph tell there very less People who is getting ~~less~~ high Salary & maximum people are getting Less Salary.

So, the graph is Negatively Skewed.

* Opposite of this graph is positively Skewed and either Normal / no Skewed

* The bigger the Sample is, the better it is for us.

And

POP

Sample ≥ 30 → data point → Normally distributed.

SM

* if we don't have much resources in this case Sample Size more than or equal to 30. is also greater

$\boxed{SS = \geq 30}$ → proven through. [with Central limit theoram.

Ram Sanyog Kumar
19, 1, 14, 25, 15, 7

→ PM = $\boxed{SM \pm \text{Margin of Error}}$

This is the things so we are studying the Central Limit theoram. 1:30 hour : Min

Ram

Average income of Indian families → (35) (1) 0 (2) 0 (3) 0

Mean

$PM = SM \pm$ Margin of Error

PM = 25K

= 25-1 to 25+1

= 24 to 26 → Range / Interval

And

Pop

↓

Sample ≥ 30 → data point

→ Normally distributed.

↓

SM

* If we don't have much resources in this case

Sample Size more than or equal to 30. is also greate

$\boxed{SS = \geq 30}$ → Proven through.

[with Central limit theorem.

Ram Sanyog Kumar
19, 1, 14, 25, 15, 7

→ PM = $\boxed{SM \pm \text{Margin of Error}}$

↓

This is the things so we are studying the Central Limit theoram. 1:30 hour: Min

Ram

Average income of Indian families → 35

1 ○

2 ○

3 ○

Mean

PM = SM ± Margin of Error

PM = 25k

= 25-1 to 25+1

= 24 to 26 → Range / Interval

PM

fish

river

pop

If PM = SM

Hook

PM / fish $\longrightarrow$ SM $\pm$ Margin

Net

90% $\longrightarrow$ 25K $\pm$ 1K $\longrightarrow$ 24K to 26K

95% $\longrightarrow$ 25K $\pm$ 1.5K $\longrightarrow$ 23.5 to 26.5K

99% $\longrightarrow$ 25K $\pm$ 2.5K

Confidence Interval

90% + 24K to 26K

$\downarrow$

Interval

CI = SM $\pm$ Margin

PM

Formulas to find out Confidence Interval

$$CI = \bar{x} \pm Z^* \times \frac{s}{\sqrt{n}}$$

$\downarrow$

Zt

will tell

Margin of error

1:44:27

$\bar{x}$: Sample Mean

S : Standard deviation of the Sample.

n : Sample Size

Z^* : Z-Score for a Certain Confidence level.

Z^*	Confidence level
1.65	90 % ✓
1.96	95% ✓
2.58	99 % ✓

MVVI

Business Problem

Estimate whether the mean lead content in maggi packets is within the allowed range or not?

Allowed range = 2.5 ppm (part per million)

50 M

100 M

PM

X

Lets take value

$n = 100$ (maggi packet)

$\bar{x} = 2.3$ ppm

$S = 0.3$ ppm

① 2.20 ppm

② 2.43 ppm

③ 2.37 ppm

④ 2.21 pm

⋮

(100) 2.28 ppm

SM → = 2.23 pm

$$CI = \bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

$$= 2.3 \pm 2.58 \times \frac{0.3}{\sqrt{100}}$$

$$= 2.3 \pm 2.58 \times 0.03$$

$$= 2.3 \pm 0.07$$

$$= 2.3 + 0.07 \text{ to } 2.3 - 0.07$$

CI = $\boxed{2.23 \text{ pm to } 2.37 \text{ ppm.}}$

99% Confidence

popm

z^*

90%

95%

99%

pop

maggi

100m

out of this

1

For Study :- Statquest

Financial Analyst

Stock (Apple) → $\boxed{50 - 60\%}$

90% confidence

→ $\boxed{45 - 75\%}$

95%

→ $\boxed{30 - 80\%}$

99%

dib@ 99 5 38881/

Make note of the class because You can revise the video again & Time

Hypothesis Testing 16

So Inferential Statistics is way to PM ⟶ SM ± Margin

CI = SM ± Margin

= 29.5 ± 1

= 28.5 to 30.5

SM

500

SS

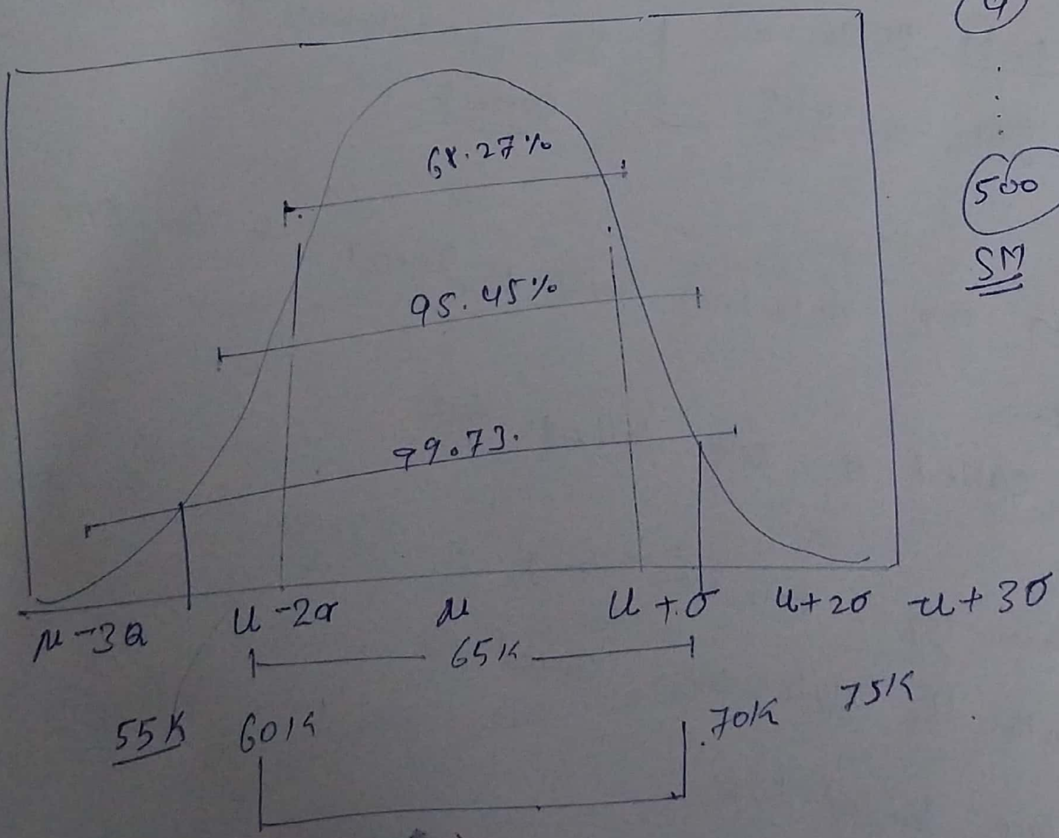
1. 28
2. 27.2
3. 31
4. 30.5

⋮

500

SM = 29.5

The average life = 30 month is - Correct



2024

Salary

20K people

M = 65K

SD = 5K.

→ 68.27% people earning 60K – 70K salary.

65K – 5K

M – 1×SD

↓

60K

M + 1×SD

65 + 5K

70K

* Notes

Through Central limit theorem when my SS ≥ 30 then I will and we plot these 30 values normally distributed.

(50 Types)
SS →

~~POM~~ POM 30
SM = 31.8

So population mean is correct

* When my rejection level is going to be one side.
We called it one tailed test.

* When my rejection level is going to be so both the side we called it two tailed test.

* Rejection area should be of 5% of intere area.

$$Z = \frac{x - \mu}{\sigma}$$

Z = Standard Score

x = Observed value

σ = Standard deviation of sample

Null hyp. Cont: =, <=, >=
Alternate Hyp. Co: ≠, >, <

according to Question

Null hyp <=

Alt >

Note :- Our aim is always to reject the null hypothesis.

Step 1 - BPS

Step 2 NH & AH

→ try to the reject the null hypothesis through some test.

Step: 3 sample data collection

Step 4 SM → Z - value of SM

$$\left(Z = \frac{x - \mu}{\sigma (SD)} \right)$$

(5) SM VS PM (with margin)

(6) if SM's Z- value fall into rejection region. (population mean claim is rejected) or Null hypothesis will be rejected

---

Rejection Region Method (old)

* Always Go with p-value

~~1:20~~

Steps (Critical value)

⇒ SM ⇒ SM Z value ⇒ Critical values, critical value VS SM Z value

⇒ then deciding we are rejecting the null hypothesis or not

* Notes Significance leave gives us 5% rejection area.

* Notes:- when our p-value is less than the significant value means %5 or 0.05 then only we can reject the null hypothesis.

or Let's

Z- test

SD

↓

P- value (0.03)

3% → 97% → ANH

Here it mean 97% proof is against the null hypothesis and 3% in favour of null hythesis.

And

We can reject the null hypothesis when we have the proof against the null hypothesis, and it should be major [%] percentage.

→ Journey of hypothesis testing:-

1.) BPS

(2.) NH & AH

(3) Sample data → ~~Hyp test.~~

(4) Sample data → Hypothesis Test.

(5) proof against NH in the form of p-value or critical value.

(6) we should use the p-value to make the final decision.

(7) if the p-value is ≤ SL (5%)
(α)

alpha
α
β → Beta

Then the proof in favour of NH is too weak & the proof against NH is is too strong. So we can reject the null hypothesis

P-value $\nless$ (0.05)

↓

<u>we fail to reject NH</u>

0.43 $\nless$ 0.05 so we don't have enough proof to reject the null hypothesis.

In the favour of NH

Z- test (Assumptions)

1. Sample size ≥ 30
2. Data should be normally distributed
3. population Standard Deviation must be given to us.
4. Sample should be randomly Collected.

Notes for 2

for checking this the data is normally distributed or not ploting will not help.

So to verify this we will use another test which

→ Shapiro wilk test.

T- test

① Sample size ≥ 30 or Sample size ≤ 30. → so data is normally distributed (Asumption)

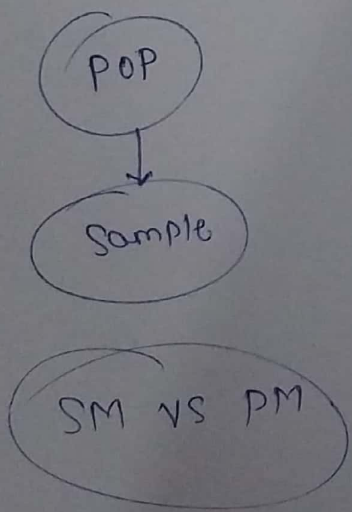
② Data should be normally distributed.
→ Shapiro will test for normality.

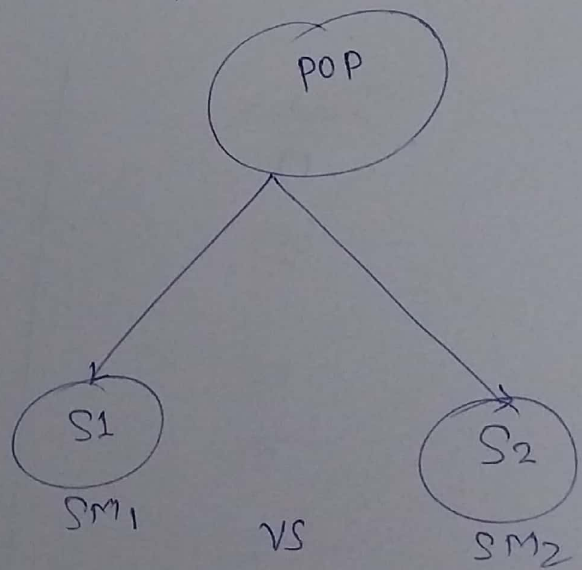
③ Sample Standard deviation.

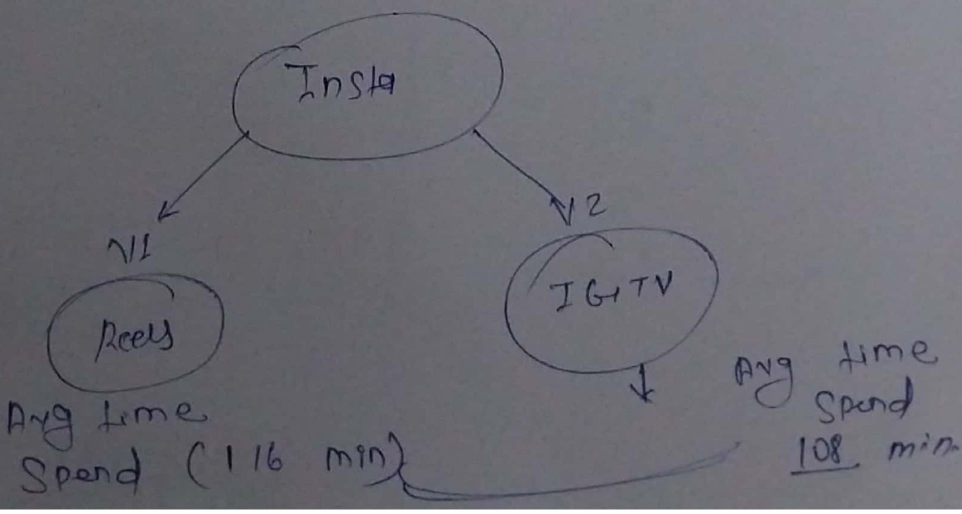
Note: T-test is most use test rather than any other.

One sample Test or Z- test

Two sample t-test or Z-test







ANOVA $\longrightarrow$ Numerical Data

for 3 or more than 3 samples.

$\rightarrow$ Analysis of Variances

Chi – Squared Test of Ind.

$\downarrow$

use on Categorical Data

Suppose

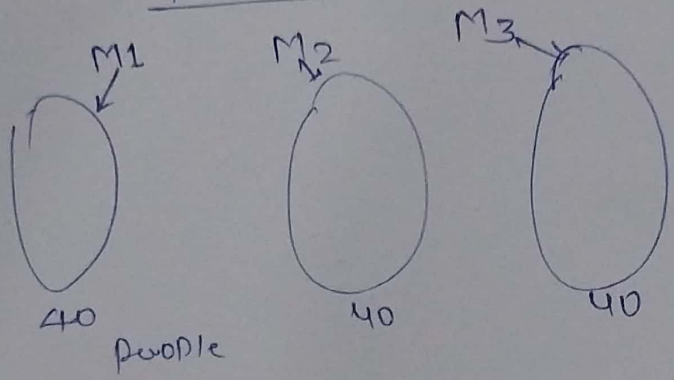
Starbucks

Gender	Preference.
M	T
F	C
	C
F	C
M	T
M	C
M	T
F	C
F	T
M	

Example of ANOVA

Take help of Shikhar Sir

Plasma



Book for DS in Shikhar Sir, Github.

Statistical thinking for programmers.