23-08-2025

Agenda: statistics

z-stat :

used → hypothesis test

→ $n \geq 30$ (CLT)

→ standardised how far the sample mean is from the population mean in unit standard deviate.

$z = 1.50$ (score) $\quad \alpha = 0.05$

$z_{critical} = 1.645$ $\quad z_{score} < z_{critical} \rightarrow$ Fail to reject H_0

→ True/False

p-value:

used → hypothesis test

pvalue ≈ small value → evidence against H_0

pvalue ≈ large value → not enough evidence.

$p = 0.067$ → 6.7% chance result are this extreme if null true

$\alpha = 0.05$

→ $p > \alpha$ (fail to reject H_0)
$< \alpha$ (reject H_0)

$p = 0.01$ → 1% prob that mean of your sample is due random.

$p \downarrow$ more $\uparrow$

$0.01 < 0.02 < 0.03 < 0.04 < 0.05$

reject H_0

Type I & Type II Error :-

TP, TN, FP, FN → Type II Error (Missed Detection)

↓

Type I error (False Alarm)

H_0 : the defendant in innocent.

H_1 : The defendant is guilty.

Image:

Real	Predicted	
Apple	Apple	→ TP
None	None	→ TN
Apple	None	→ FN
None	Apple	→ FP

H_0: patient is Healthy

H_1: patient has disease

	Healthy (H_0 true)	Sick (H_1 true)
Reject H_0	Type I error	Correct
Fail to reject H_0	Correct	Type II error

Z-Test

- → population σ (known)
- → sample size ($n \geq 30$)
- → Data $\sim$ normal (CTL applies)

T-test

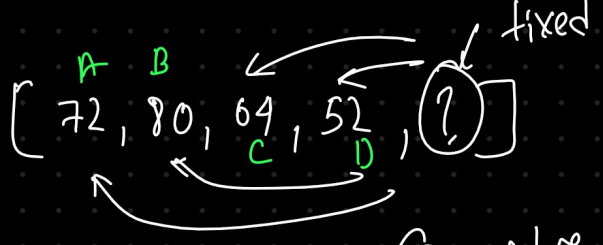
- → population σ is unknown instead we use sample σ
- → works even with small sample size ($n < 30$)
- → Data should be $\sim$ normal

Degree of freedom :—

(df)

5 students :

avg marks of 5 students : 72



$$72 = \frac{72 + 80 + 64 + 52 + ?}{5}$$

solve for this

$$df = n - 1$$

T-test :

Pizza problem :

H_0 : avg delivery time : 30 minutes : $\mu = 30$

H_1 : takes longer than they claim : $\mu > 30$

$n = 25$

$\bar{x} = 32$ minutes

$s = 4$ minutes (sample s)

T-test formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{32 - 30}{4/\sqrt{25}} = 2.5$$

$df = n - 1 = 25 - 1 = 24$

$\alpha = 0.05$, $df = 24$ ⟶ lookup on t-table → $t_{(0.05, 24)}$ → 1.711

(one tailed)

$t_{calc} = 2.5$, $t_{critical} = 1.711$

$t_{calc} > t_{critical}$ → reject H_0

Chi-squared distribution :-

→ categorical data (counts, frequencies)
→ variance → σ^2

(1) Goodness of fit / Independence Test

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O → observerved freq

E → expected frequency

(2) Test of variance (spread)

$$\chi^2 \rightarrow \frac{(n-1)s^2}{\sigma^2}$$

$s^2 \rightarrow$ sample variance

$\sigma^2 \rightarrow$ population variance

$n \rightarrow$ no. of sample

Z/T → is my mean different from a claim

χ^2 → is my spread (variance) different

H0

population $\sigma^2 = 16 \text{ minutes}^2$ mean ± 2·STD

popul $\sigma = 4$ minutes

0.05 → 95% → from the mean

95% deliveries should fall with in ± 8 minutes of the average delivery time

H_1

Customers are saying deliveries are more inconsistent.

$H_a: \sigma^2 > 16$

$n = 36$

$s^2 = 28$

$$\chi^2 = \frac{(36-1)20}{16} = 43.75$$

$df = 36 - 1 = 35$

$\alpha = 0.05$

$$\chi^2_{0.05,35} = \frac{43.773 + 55.758}{2}$$

$$= 49.765$$

$$= 49.8$$

Chi score with chi score critical :—

$43.75 < 49.8$ ← Fail to reject H_0 / accept H_0

↓

not enough evidence.

Goodness of fit / Independence Test

→ Examine whether two categorical variable are associated

	Apple	Banana	Orange	Total
Male	30	20	10	60
Female	20	20	10	50
Total	50	40	20	110

H_0 : Fruit preference and gender are independent.

H_1 : there is an association

step1: Calc. Expected count.

step2: compute the chi-square statistic.

step3: degree of freedom

step4: p-value

step 1: Expected-count

→ (Row count × Column total) / Grand total

* M-A → (60×50) / 110 → 27.27
* M-B → (60×40) / 110 → 21.82
* M-O → (60×20) / 110 → 10.91
* F-A → (50×50) / 110 → 22.73
* F-B → (50×40) / 110 → 18.18
* F-O → (50×20) / 110 → 9.09

step 2: Compute the chi-square statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

* M-A → $(30 - 27.27)^2 / 27.27 \rightarrow 0.273$
* M-B → $(20 - 21.82)^2 / 21.82 \rightarrow 0.152$
* M-O → $(10 - 10.91)^2 / 10.91 \rightarrow 0.076$
* F-A → $(20 - 22.73)^2 / 22.73 \rightarrow 0.329$
* F-B → $(20 - 18.18)^2 / 18.18 \rightarrow 0.182$
* F-O → $(10 - 9.09)^2 / 9.09 \rightarrow 0.091$

$x^2 = 1.102$

Step3: DOF

(df) → (number of row - 1) * (number of col - 1)

→ (2-1) * (3-1) → 1 x2 = 2

Step: x^2, df, p →

p → 0.576

$p > \alpha$ → fail to reject H_0

↓

insufficient evidence.

Confidence interval (CI)

→ range of values derived from a sample statistic., likely to obtain from parameter (μ).

→ It is always associated with confidence level. which quantifies the level of conf that the parameter lies within the interval.

→ point estimates are imperfect, we don't get range from this.

formula:

$$CI = \bar{x} \pm (\text{critical value}) \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

$\bar{x}$ → sample mean

σ → Population std deviation

n → sample size

conf level → (95%) → critical Z → 1.96

$\bar{x} = 31$ minutes

$\sigma = 4$ minutes

sample size - 36 deliveries

$z_{critical} = 1.96$

MOE: $z_{crit} \times SE \rightarrow 1.96 \times \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{4}{\sqrt{36}} \approx 1.307$

Construct the CI:

$CI = \bar{x} \pm MOE$

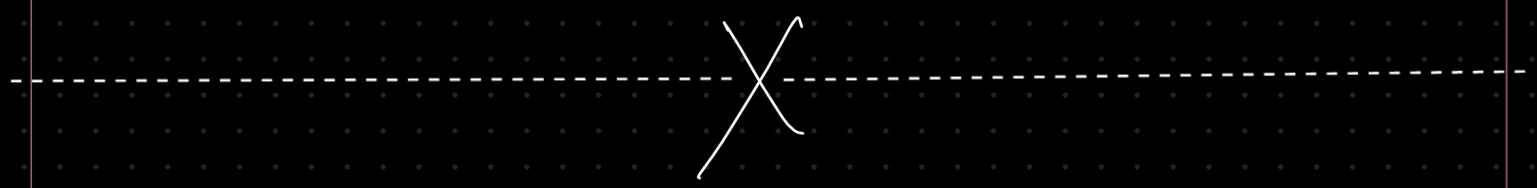
$= 31 \pm 1.307$

Lower bound = 29.693

Upper bound = 32.307

Final: The 95% CI for the true average delivery time

(29.693, 32.307) minutes.



python → Tools → pyfeat → stats module → EDA → ML

ML:

Data gathering

process

EDA

Feature Engineer

Train / Test split

Algorithm

Testing

Expert

Use