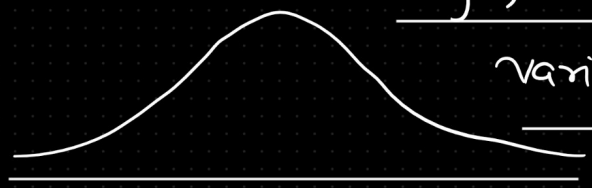
Statistics

- **Descriptive** (Description of the dataset)
- **Inferential** (Infer) — Hypothesis Testing

① Measures of Dispersion ⟷ Data Imputation (ML algorithms → KNN)

(Mean, Median, Mode, Range, Standard deviation, Variance, IQR)

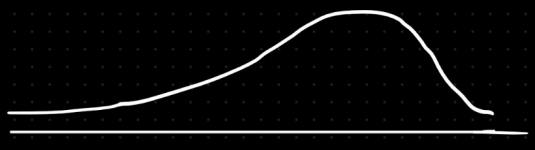


② Data Distribution

- Symmetric / Normal / Gaussian (mean = median = mode)
- Skewed distribution
  - Right — mean > median > mode
  - Left — mean < median < mode

CTC

SNF ⟶ $\mu = 0$, $\sigma = 1$



③ Central Limit Theorem

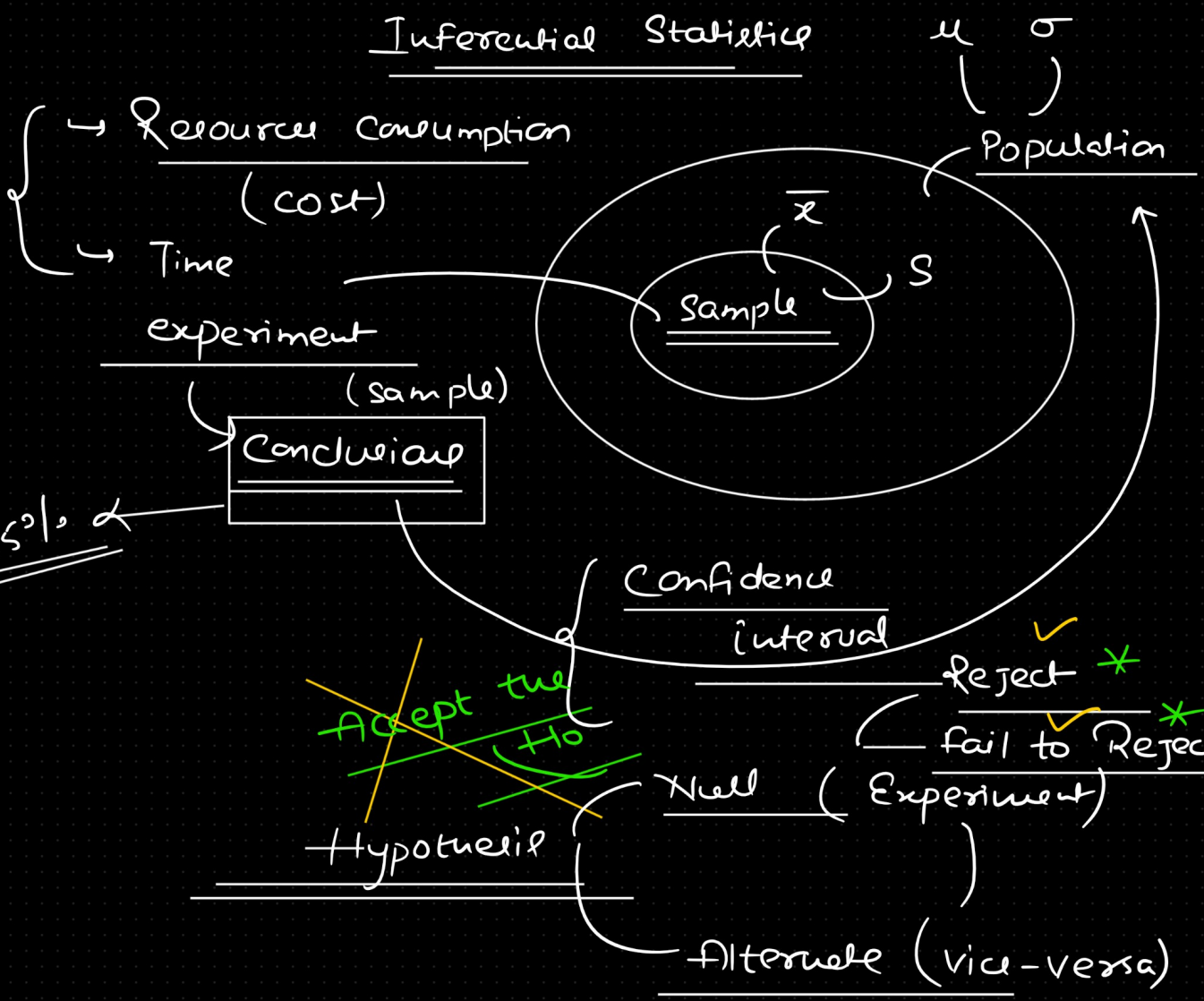
④ Z-score $\frac{x_i - \mu}{\sigma}$

⑤ Sample vs Population

⑥ Random variable
- Discrete → PMF
- Continous → PDF

⑦ Visualization & Plots (Pyplot, catplot)
- ↳ Scatterplot, Barchart,
- Box/Violin Plot (Outliers),
- Correlation heatmap, (feature selection)
- Distribution plot,
- Histogram

Inferential Statistics



$N = 10{,}000$ $n = 100$

$N \geq n$

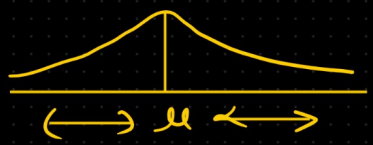
Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

$N = 10{,}000$ $n = 100$

Population Standard Deviation	Sample Standard Deviation
$\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X_i - \bar{x})^2}{n-1}}$
σ = population standard deviation N = the size of the population X_i = each value from the population μ = the population mean	s = sample standard deviation n = the size of the sample X_i = each value from the sample $\bar{x}$ = the sample mean

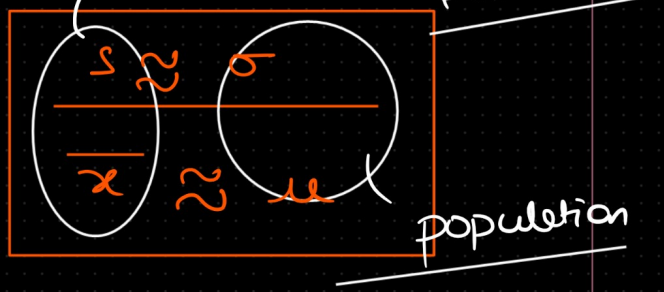
SD → measure of dispersion

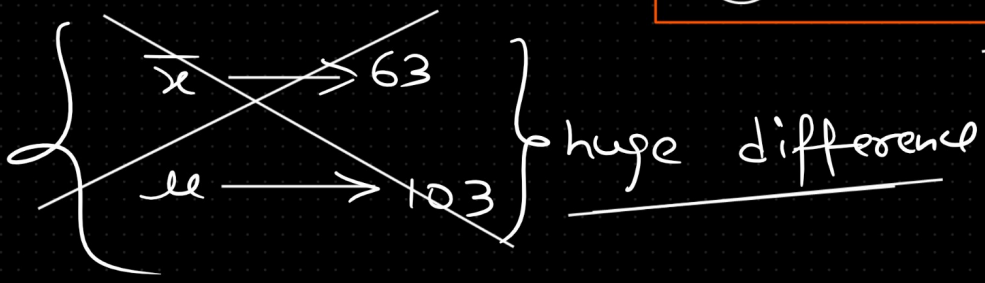
$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$



Why sample SD, division by $n-1$??

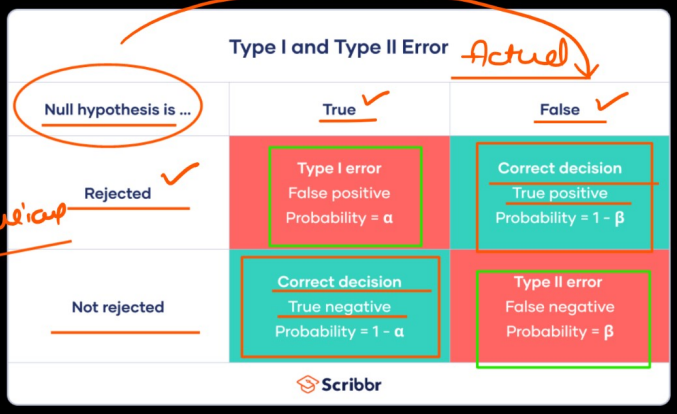
→ Task 1





(ML)

→ Confusion Matrix



$H_0 \Rightarrow$ Heart attack has no link with drinking coffee

False (Reality)

(Domain Experts)

$H_0 \rightarrow$ Actually (True)

Experiment → fails to reject H_0 (TN)

(Good or bad)

Reject H_0 ← Experiment

Domain experts

Reality

- $H_0 \rightarrow$ True
- Exp → Reject H_0
  - Bad decision
  - Type 1 error (α) ↓↓

(Reality)

- $H_0 \rightarrow$ false
- Exp → fails to Reject H_0
  - Bad decision
  - Type 2 error (β) ↓↓

Hypothesis
- H_0 → Building is not safe
- H_1 → Building is safe

α is more Critical to handle

Type 1 error → ↓ (α)
- H_0 → True (Reality)
- Exp → Reject the H_0

Type | error is committed if we reject "Building is not safe" when it is not safe.

Type 2 Error → ↓ (β)
- H_0 → false (Reality)
- Exp → fail to reject H_0

Type || error is committed if we fail to reject "Building is not safe" when it is safe.

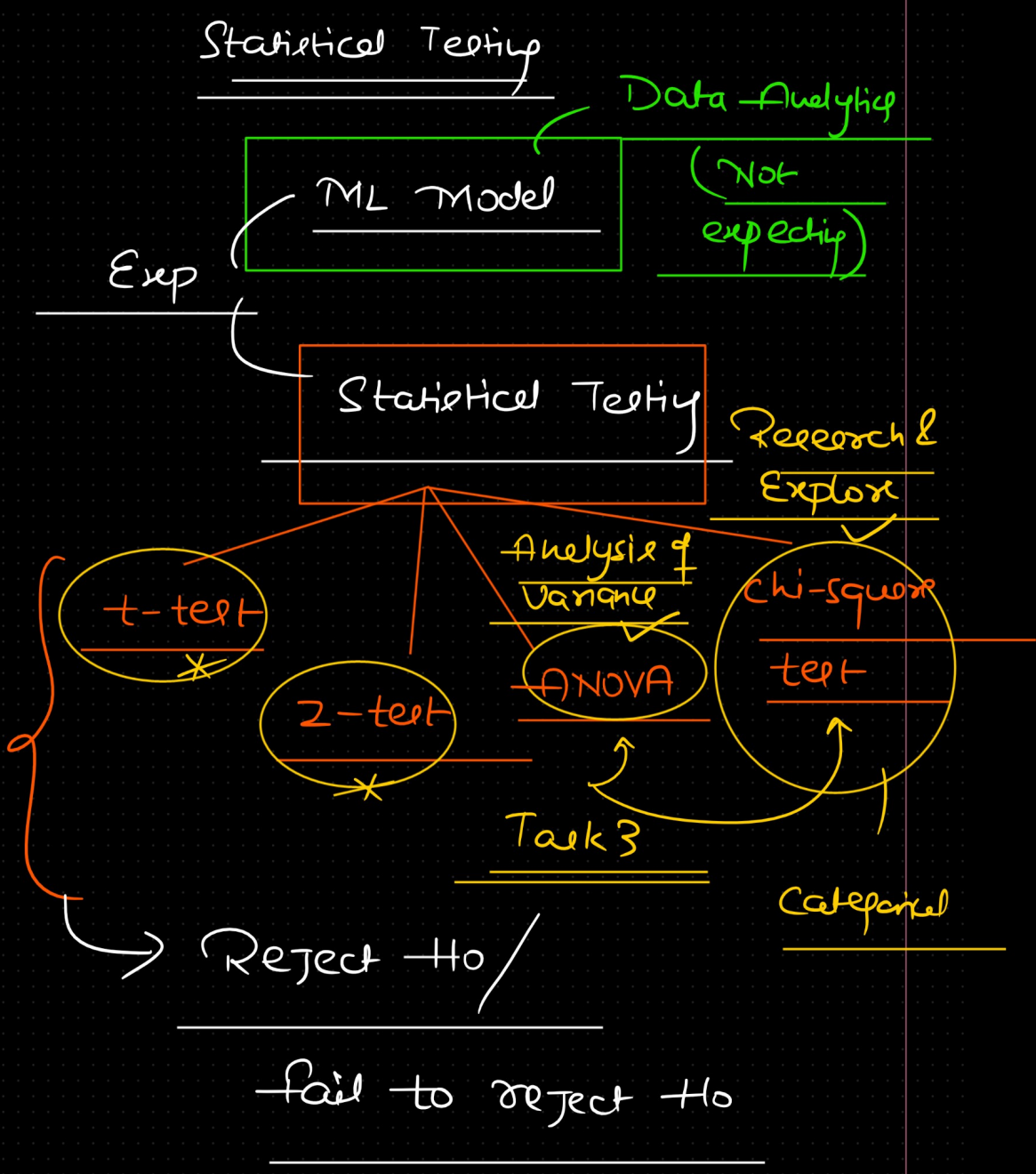
To detect whether the patient is diabetic or not.

- H_0 → Patient is not diabetic (0)
- H_1 → Patient is diabetic (1)

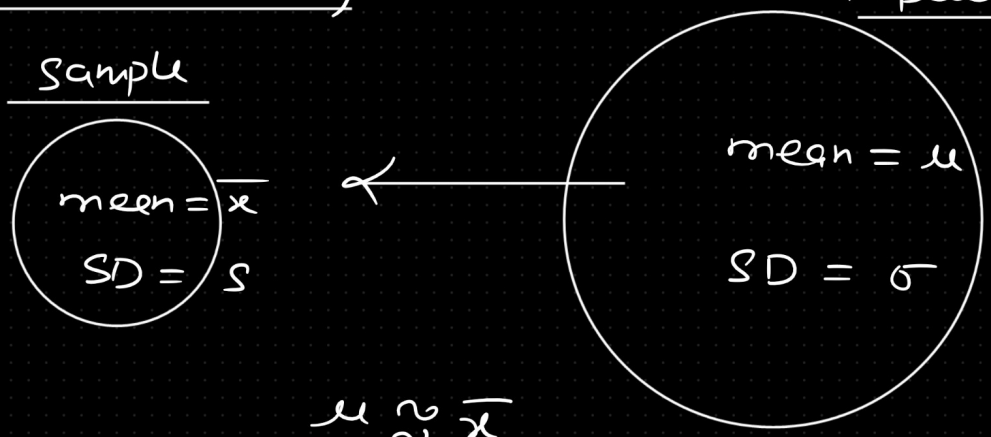
Ho / Exp	Reality: 1	Reality: 0
Reject	Correct	Type 2 Error (Type 1)
fail to Reject	Error (Type 2) (Type 1) → Talk 2	Correct

Type I and Type II Error		
Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

Scribbr

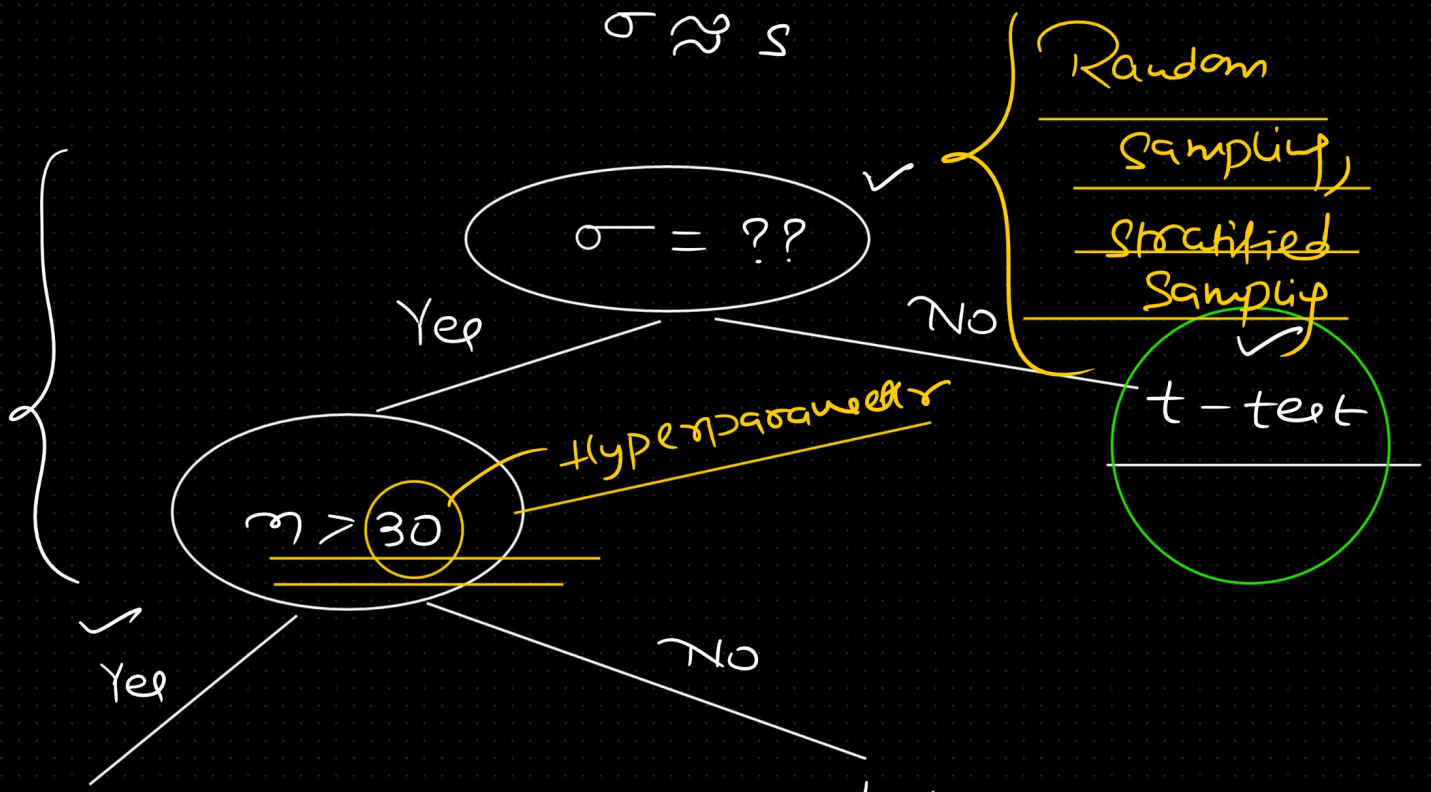

Statistical Testing
Data Analytics
ML Model
(Not expecting)
Exp
Statistical Testing
Research & Explore
t-test
Analysis of Variance
ANOVA
Chi-square test
2-test
Task 3
Categorical
Reject Ho /
fail to reject Ho

Statistical Testing



$\mu \approx \bar{x}$

$\sigma \approx s$



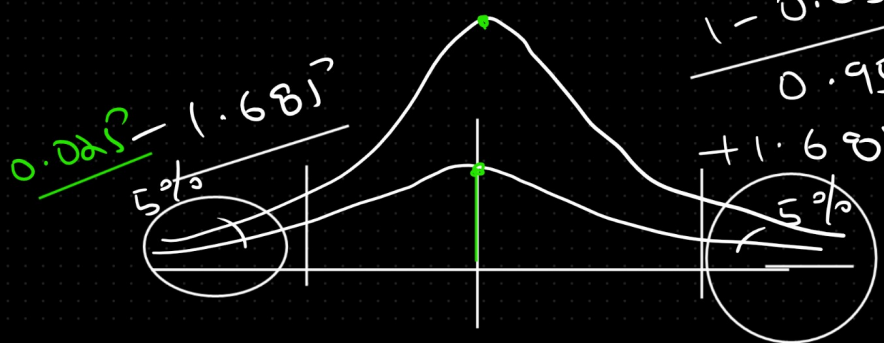
z-test

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

t-test

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

$\alpha = 0.05$



Numerical Example: The average height of adults in a certain country is known to be 168 cm. A scientist believes that the average height of adult residents in a certain city in that country is different. The scientist measures the heights of 40 randomly selected adult residents of the city and finds an average of 171 cm with a standard deviation of 7 cm. Test the scientist's hypothesis at the 0.10 significance level.

Null → $H_0 \rightarrow \mu = 168\ cm$

$H_1 = \mu \neq 168\ cm$ → Alternate

$\mu = 168\ cm$

$n = 40$

$\bar{x} = 171\ cm$

$S = 7\ cm$

t-test

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$t = \frac{171 - 168}{7/\sqrt{40}}$$

$$t = 2.71$$

$\alpha = 0.10$ (Risk)

CI $\Rightarrow 1 - \alpha = 1 - 0.10 = 0.90$

Critical value → ± 1.685 ← Internet

→ t-table

$df = n - 1 = 39$

$2.71 > 1.68$

t-test > Critical value

→ Reject H_0

↓

t-test < Critical value

↳ fail to Reject Ho

The height is different from 168 cm

Z-test, t-test ⟹ value ✓

Critical value ✓

(Internet-table)

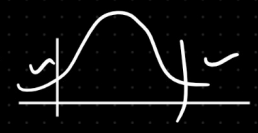
exp

Value < critical value

↳ fail to reject Ho

exp val > critical value

↳ reject Ho

two - tailed test → 

one - tailed test

↳ $\alpha = 0.10$

$1 - \alpha = CI$

A sample of 50 students using the new teaching method had an average score of 78 with a standard deviation of 10. The average score for students using the traditional method is known to be 75. We want to test the hypothesis at a 5% significance level ($\alpha = 0.05$)

α
=

$\mu = 75$ — H_0

$\mu \neq 75$ — H_1

Risk factor

↳ Domain experts

$n = 50$

$\bar{x} = 78$

$s = 10$

$\mu = 75$

$\alpha = 0.05$

$CI = 1 - \alpha$

$= 0.95$ ✓

$df = 49$ ✓

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

$= \frac{78 - 75}{10 / \sqrt{50}}$

≈ 2.12

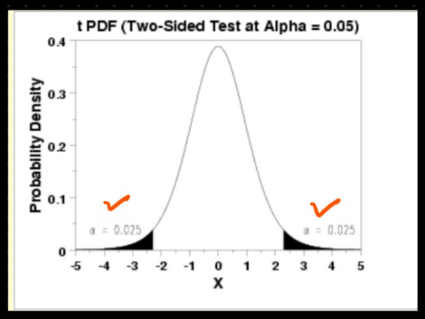
Refer the table

evaluated value ↑ — critical value

$2.12 \not> 2.010$

↳ Reject H_0

one-sided | two-sided (by default)



$\alpha = 0.05$

$1 - 0.025$

$= 0.975$

49, 0.975 → 2.010

Diabetic

5000 → 0

5000 → 1

Random Sampling

100%

1000

500 → 0

500 → 1