Math notation, statistics measure central tendency

μ = Population mean (divides Total number, n) → Average of entire group

$\bar{x}$ = Sample mean (divides Total number n-1) → Average of subset of the population

σ = standard deviation (SD)

σ^2 - variance

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

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

Sample SD

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Population S.D

$$\sqrt{\frac{1}{N} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Sample variance

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

Population Variance

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

⇒ Class interval = upper limit - lower limit

Frequency → Count of occurance (how of then a value appers)

Range → diff b/w largest and smallest value
= max - min

~~Measure of~~

Measure of central tendency

1 Mean(μ) = $\frac{\text{Sum of values}}{\text{no: of values}}$

2 Median → arrange descending / ascending order, for odd no:of data point - middle value (median) else even - average of 2 - middle values

3) mode - most frequently occuring values in the same population

Measure of dispersion

1. Range = max - min
2. Variance (σ^2) = $\frac{\Sigma(x-\mu)^2}{n}$
3. S.D (σ) = $\sqrt{(\sigma)^2}$
4. Interquartile Range (IQR)

$IQR = Q_3 - Q_1$

5. Quartiles (Q) - divides the dataset into 4 equal points

Q_1 - Median of the lower 25%

Q_2 - median (50%)

Q_3 - Median of the upper 25%

Outlier - extrem values (far from the actual values)

6. Coefficient of variance (v)

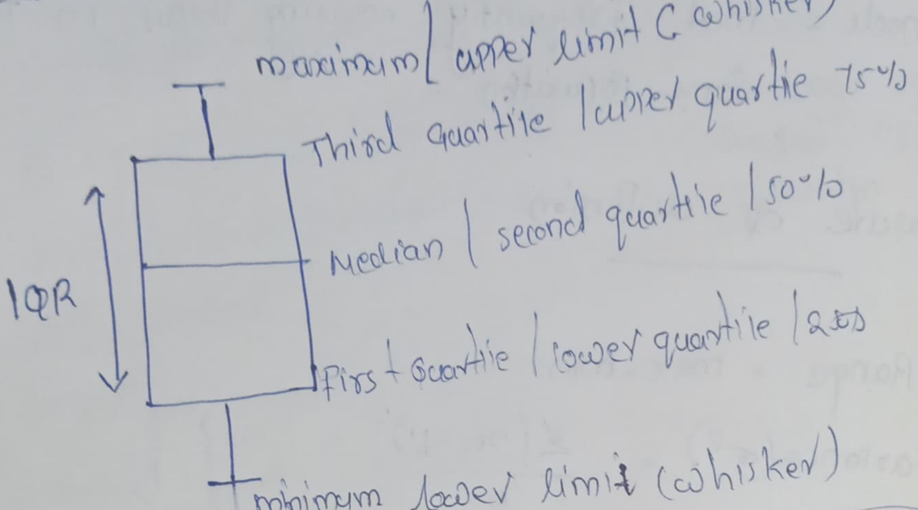
$CV = \frac{\sigma}{\mu} \times 100$ σ = S.D, μ - mean

7. Mean Absolute Deviation

$MAD = \sum_{i=1}^{n} \frac{|x_i - \mu|}{n}$

Box Plot

data visualization that summarizes a dataset distribution



upper limit = $Q_3 + 1.5 \times IQR$

lower limit = $Q_1 - 1.5 \times IQR$

condition for outlier
outlier < lower limit
or
outlier > upper limit

Covarience

$$Cov(x,y) = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n}$$

Correlation

$$\rho(x,y) = \cancel{\sum(x_i-\bar{x})} \; \frac{cov(x,y)}{\sigma_x \sigma_y}$$

range → -1 to 1

0 → light

-1 → highly -ve corelated

+1 → highly +ve correlated.

Z-Score

$$\text{Z-Score} = \frac{x_i - \mu}{\sigma}$$

x — data value
μ → mean
σ = SD

Hypothesis Testing.

Null hypothesis (H_0)

There is no significant difference or effect

Alternative hypothesis (H_1)

There is a significant effect i.e the given statement can be false.

level of significance (α)

It is a threshold used to determine statistical significance (value common - 0.05, 0.01, or 0.10)

↓ commonly. → median

Z- test

used to compare means when sample size is large $n > 30$ and population variance - known.

Steps

1. State null hypothesis
2. State Alternative hypothesis.
3. Choose your significance level (α)
4. Calculate your z-test statistics

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

$\bar{x}$ = simple mean
μ = Population mean
σ = Population SD
n = Sample size.

5. Find P-value using z-table and z-test statics computed ~~if P~~

6. If P-value > α

Then we fail to reject null hypothesis (Accept Alternative hypothesis)

7. There are one tailed and two tailed tests.

Two tailed, Alternative hypothesis look like this

$$H_0: \mu = \mu_0 \quad v/s \quad H_1: \mu \neq \mu_0$$

one tailed, look like (A.H)

Right tailed, $H_0: \mu \leq \mu_0 \quad v/s \quad H_1: \mu > \mu_0$

Left tailed, $H_0: \mu \geq \mu_0 \quad v/s \quad H_1: \mu < \mu_0$

Eg. A particular companies choclate bars are supposed to have an average weight of 50 ~~grams~~ grams according to the manufacturer we want to test if a sample of choclate bars deviates significantly from this weight data:

50.8, 49.5, 50.2, 51, 49.7 50.3 49.8 50.5, 49.6 50.1 population standard deviation 1.5 grams?

$H_0: \mu = 50$

$H_1 = \mu \neq 50$

$\alpha = 0.05$

$\bar{x} = 50.15$

$SD = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{10}} = 0.47434$

$$z = \frac{\bar{x} - \mu_0}{SD} = \frac{50.15 - 50}{0.47434}$$

$z = 0.8162$

Pvalue - 2 tailed

$P = 0.7518$

Here $\alpha = 0.05$, $P = 0.7518$, $z = 0.8162$

$0.05 < 0.7518$

So fail to reject H_0

$\therefore$ No significant deviation from 50 g.

T-test

- We need T-test because population standard deviation σ is not always available.
- We will computing T-test statistic (based on Problem statement and type of t-test)

Type of T Test	Null hypothesis	Alternative hypothesis	DF
One sample T-test	Sample mean = reference mean	Sample mean $\neq$ Reference mean	$df = n-1$
Independent sample T-test	mean value in both group are same.	mean value in both group not same	$df = n_1 + n_2 - 2$
Paired sample t-test	mean value of the difference b/w the pairs is zero	mean value of difference b/w the pairs is not zero	$df = n-1$

one sample T-test steps

compares the sample mean to a known population mean

steps

1. state the null hypothesis
2. state the alternative hypothesis
3. choose your significance level(α)
4. calculate your T-test statics

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

$\bar{x}$ = samplemean
μ = population mean
s = sample SD

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

5. critical t-value find from T-table using α
6. if t statics < critical t-value, we fail to reject null hypothesis

Q. same eq as z test

$H_0 : \mu = 50$

$H_1 : \mu \neq 50$

$\bar{x} = 50.15$, $s = 0.5104$ $n = 10$

$$SE = \frac{s}{\sqrt{n}} = 0.1614$$

T-test statist $t = \dfrac{\bar{x} - \mu_0}{SE}$

$= \dfrac{50.15 - 50}{0.1614}$

$= 0.929$

$df = 10-1 = 9$

Value (2-tailed) T-value = 2.262

t-test static Pvalue

0.929 < 0.262

∴ we fail to reject null hypothesis.

Independent Sample T-test

compare means of 2 independent groups

steps

~~compare means of 2 independent groups~~

1. state null hypothesis
2. state alternative hypothesis
3. Choose your α
4. Calculate your T-test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ For one tailed.

$$t = \frac{\overline{x_1} - \overline{x_2}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$ For 2 tailed

(Pooled SD) $$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

5. Find the critical t-value from T-table using 2

6. If t-statics < critical t value, we fail to reject Null hypothesis.

Eg: Brand A = 49.5, 50.1, 49.8, 50.3, 50

Brand B = 50.4, 49.9, 50.6, 50.2, 50.1

$H_0 : \mu_A \neq \mu_B$

$H_1 : \mu_A \neq \mu_B$

$\overline{x}_A = 49.94$ $n_A = 5$

$\overline{x}_B = 50.24$ $n_B = 5$

$\overline{x}_A - \overline{x}_B = -0.30$

$S_A = 0.316$

$S_B = 0.265$

$$S_p = \sqrt{\frac{(5-1)(0.316)^2 + (5-1)(0.265)^2}{5+5-2}}$$

$= 0.292$

$$t = \frac{49.94 - 50.24}{0.292\sqrt{1/5 + 1/5}}$$

$$= \frac{0.3}{0.1847} = -1.624$$

$df = n_1 + n_2 - 2 = 8$

$\alpha = 0.05$

critical t-value = 2.306

t statistic   critical t value

-1624   < 2.306

Here we fail to reject the null hypthesis

z-test

z test Statiss $\xrightarrow{\text{z-table}}$ p-value Vs α

T-test

T-test statiss v/s Alpha (α) df tailedness

↓

critical t-value.

Matrix

- $r \times c = m \times n$ (order of a matrix)

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

- $a_{ij} = a_{ji}$ → Symmetric matrix
- Identity matrix = $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Eg: $A = \begin{bmatrix} 6 & 3 \\ 9 & 4 \\ 0 & 7 \end{bmatrix}_{3\times 2}$ $C = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 11 & 31 \end{bmatrix}_{3\times 2}$

⇒ $A+C = \begin{bmatrix} 7 & 5 \\ 12 & 8 \\ 11 & 38 \end{bmatrix}_{3\times 2}$

~~Singular~~

- Single value x matrix → scalar matrix

- $A_{m\times 2} \times B_{2\times n}$ → Result $_{m\times 2}$

[1st matrix column = second matrix row]

- Transpose (A^T)

If $A = A^T$ symmetric Matrix (only for square matrix)

Identity matrix is also symmetric

- Determinant

$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ $|M| = (ad - bc)$

- Inverse Matrix (M^{-1})

$M \cdot M^{-1} = I$

- M^{-1} does not exist when $|M| = 0$

~~Eigen:~~

Eigen Values / Vectors

- eigen values are unique scalar values associated with a square matrix of linear transformation.
- They indicate how much an eigen vector is stretched of compressed during the transformation.
- Represented by $AV = \lambda V$

  A = Matrix
  V = eigen vector
  λ = eigen value.

- 2x2 = 2 eignvalues
  3x3 → 3 eigenvalues.

Probability - chance of happen an event

- event - get an outcome from Sample space.
- probability = $\frac{\text{event (count)}}{\text{(S) count}}$

event → independent
→ dependent.

~~Joint Probability~~

~~P(C) = $\frac{15}{32}$ P(T) = $\frac{17}{32}$~~

~~P(m) = 2/32 P(n) = $\frac{30}{32}$.~~