# Assignment - 01

Part 1 → Descriptive analysis

(1.) Mean → 30, 40, 45, 50, 200

$\frac{30+40+45+50+200}{5} \Rightarrow \frac{365}{5} \rightarrow 73$

(2) Median → 30, 40, 45, 50, 200

→ 45 → median

(3) Mode → 30, 40, 45, 50, 200

Add the salary 40 to the above list

30, 40, 40, 45, 50, 200

→ Mode = 40

(4) interpretation →

Mean → Advantage :- * Take into accounts all data.
* Useful for further calculation

Limitation → * Sensitive to outliers
* May not accurately represent data in skewed distribution

Mode → Advantage → * Simple to understand and Calculate
* Not affected by outliers

Limitation → * Can be multiple mode in dataset
* less informative for continuous data

Median →

Advantage → * Not affected by outliers
* Represent the middle value, providing better central tendency

Limitation → * Does not consider the magnitude of all data points

* less useful in dataset with a small number of observation

Conclusion → I would recommend using median to describe central tendency of the data as it is not affected by outliers and providing better central tendency.

(2) Standard deviation and Variance

(1) Scores from a test ⇒ 5, 10, 10, 20, 30

$$\text{Mean} = \frac{5+10+10+20+30}{5} = \frac{75}{5} = 15$$

Variance $= \frac{|15-5| + |15-10| + |15-10| + |15-20| + |15-30|}{5}$

Variance $= \frac{(10)^2 + (5)^2 + (5)^2 + (5)^2 + (15)^2}{5} \Rightarrow \frac{400}{5} \Rightarrow 80$

S.D $= \sqrt{80} \Rightarrow 8.94$

Interpretation
Standard deviation is closed to the mean that means data is consistance and low spread

③ Box plot Outliers

The ages of a group of people $\Rightarrow$ 10, 12, 13, 15, 16, 18, 25, 35, 80

$Q_2 \Rightarrow$ 10, 12, 13, 15, 16, 18, 25, 35, 80

16 $\leftarrow Q_2$

$Q_1 \Rightarrow$ 10, 12, 13, 15 $\Rightarrow \frac{12+13}{2} \Rightarrow \frac{25}{2} \Rightarrow 12.5$

$Q_3 \Rightarrow$ 10, 25, 35, 80 $\Rightarrow \frac{25+35}{2} \Rightarrow \frac{60}{2} \Rightarrow 30$

IQR $= Q_3 - Q_1$
$= 30 - 12.5 \Rightarrow 17.5$

(ii) IQR Rule

Lower bound $= Q_1 - 1.5 \times IQR$

$= 12.5 - 1.5 (17.5)$

$= 12.5 - 26.25 = -13.75$

Upper bound $= Q_3 + 1.5 \times IQR$

$= 30 + (1.5 \times 17.5)$

$= 30 + 26.25 = 56.25$

Now any values $< -13.75$ and $> 56.25$ are outliers

Here 80 is an outliers

Part-2 Inferential Statistics

Simple Linear Regression

Hours Studied ⇒ 1, 2, 3, 4, 5
Exam Scores ⇒ 50, 55, 65, 70, 75

(1) Equation of Simple linear regression

$$y = B_0 + B_1 x + \epsilon$$

where $B_0$ → $B_0$ is the intercept
$B_1$ → $B_1$ is the slope
$\epsilon$ → Epsilone is the 'error term'

$$B_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

⇒

| hours Studied | Exam Scores | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|
| 1 | 50 | (3-1)(63-50) | 4 |
| 2 | 55 | (3-2)(63-55) | 1 |
| 3 | 65 | (3-3)(63-65) | 0 |
| 4 | 70 | (3-4)(63-70) | 1 |
| 5 | 75 | (3-5)(63-75) | 4 |
| | | 26+8+0+7+24 | 10 |

$$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} \Rightarrow 3$$

$$\bar{y} = \frac{50+55+65+70+75}{5} = 63$$

$$B_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{65}{10} = 6.5$$

$$B_0 = \bar{y} - B_1 \cdot \bar{x}$$

$$= 63 - 6.5 \times 3$$

$$= 63 - 19.5 = 43.5$$

Interpretation of Slope (6.5)

The Slope of 6.5 means that for every additional hour studied, the Exam Score increases by 6.5 points on Average

This shows a positive linear regression between hours studied and exam performance.

## (5). Hypothesis Testing T-Test

Given ⇒ Avg. completion rate = 75%

Company claim that a new interactive feature has increased the average Completion rate = 80%

you collect a sample = 50 users

Avg. Completion rate = 78%

Std. deviation = 5%

(a) → • $H_0$ := Std. 75%
• $H_1$ ≠ Std. 75%
→ hypothesis

(2.) Test statistic :

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

$t$ = test statistics
$\bar{x}$ = Average Sample mean
$\mu$ = hypothesized mean
$s$ = Std. deviation
$n$ → Sample size

$$t = \frac{78\% - 75\%}{5\%/\sqrt{50}}$$

$$= \frac{+3\%}{5\%/5\sqrt{2}} \rightarrow \frac{+3\%}{\frac{5\%}{5\sqrt{2}}}$$

$$= \frac{+3\% \times 5\sqrt{2}}{5\%}$$

$$= \frac{+3 \times 5 \times 1.414}{5} = +4.242$$

Sample size = 50
degrees of freedom (df) = n-1 = 49
Significance level ($\alpha$) = 0.05

Test is one tailed

Critical value ≈ 1.677

4.247 > 1.677 → Null Hypothesis rejected

(6) → Chi Square test

| | Like | Dislike | Row total |
|---|---|---|---|
| Male | 40 | 60 | 100 |
| female | 50 | 50 | 100 |
| Total | 90 | 110 | 200 |

$H_0$ = There is no relation with Gender and preferences

$H_1$ = There is relations with Gender and preferences

Formula For Expected Frequencies

$$= \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand total}}$$

Male-Like = $\frac{(100 \times 90)}{200} = 45$

male DisLike = $\frac{100 \times 110}{200} \Rightarrow 55$

female-Like = $\frac{100 \times 90}{200} \Rightarrow 45$

female-DisLike = $\frac{100 \times 110}{200} \Rightarrow 55$

| | Like | Dislike female | Total |
|---|---|---|---|
| male | 45 | 55 | 100 |
| female | 45 | 55 | 100 |
| | 90 | 110 | 200 |

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

O = observed frequency

E = Expected frequency

| | observed freq. | Expected freq. | $(O-E)^2/E$ |
|---|---|---|---|
| male (Like) | 40 | 45 | 0.556 |
| male (Dislike) | 60 | 55 | 0.455 |
| female (Like) | 50 | 45 | 0.556 |
| female (Dislike) | 50 | 55 | 0.455 |
| | | | $\chi^2 = 2.02$ |

df → degree of freedom

$df = (rows-1)(column-1)$
$= (2-1)(2-1)$
$1 \times 1 \rightarrow 1$

$\alpha = 0.05$ for $df = 1$ is 3.841

Since $x^2 = 2.022 < 3.841$ hence
we fail to reject the null hypothesis

## ANOVA

⇒ Three different teaching methods result in different exam scores:

| | | | |
|---|---|---|---|
| Group A ⇒ | 70, | 75, | 80 |
| Group B ⇒ | 60, | 65, | 70 |
| Group C ⇒ | 85, | 90, | 95 |

Anova ⇒ Anova is a statistical method used to compare the means of three or more groups.

### One-way - Anova

Compare the means of three or more independent group based on one factor.

hypotheses :- $H_0$ → All group means are equal
$H_1$ → At least one group mean is different

Step-1
Calculate Group mean

Group A mean → $\frac{70+75+80}{3} = \frac{225}{3} \rightarrow 75$

Group B mean = $\frac{60+65+70}{3} = \frac{195}{3} \rightarrow 65$

Group C mean = $\frac{85+90+95}{3} = \frac{270}{3} \rightarrow 90$

Step-2
Calculate Grand mean

$GM = \frac{70+75+80+60+65+70+85+90+95}{9}$

$= \frac{690}{9} \rightarrow 76.67$

Step-3 Sum of squares between (SSB)

$SSB = n \sum (\bar{x}_i - GM)^2$

$= 3[(75-76.67)^2 + (65-76.67)^2 + (90-76.67)^2]$

$3[2.79 + 136.19 + 177.69]$

$= (316.67)\ 3$

$= 950.01$

Step-3

Sum of squares within (SSW)

$$SSW = \sum (x_{ig} - \bar{x}_i)^2$$

Group A → mean = 75

$$(70-75)^2 + (75-75)^2 + (80-75)^2$$
$$= 25 + 0 + 25 = 50$$

Group B

$$(60-65)^2 + (65-65)^2 + (70-65)^2$$
$$25 + 0 + 25 = 50$$

Group C

$$(85-90)^2 + (90-90)^2 + (95-90)^2$$
$$= 25 + 0 + 25 = 50$$

Group A + Group B + Group C = 150

Step-4 Calculate → F-ratio

$df_{between} = K-1$

$df_{within} = N-K$

K → Total no. of groups

n → Total no. of observation

$df_{between} = K-1 = 3-1 = 2$

$df_{within} = N-K = 9-3 = 6$

Mean Square between (MSB)

$$= \frac{SSB}{df \text{ between}} = \frac{950.01}{2} \rightarrow 475.005$$

Mean Square within (MSW)

$$= \frac{SSW}{df \text{ within}} = \frac{150}{6} \rightarrow 25$$

F ratio $\rightarrow$

$$\frac{MSB}{MSW} = \frac{475.005}{25} \times \frac{1}{1000}$$

$$\rightarrow 19.0002$$

$$= 19.002$$

Significance level of F at $\alpha = 0.05$

$df_1 = 2, \quad df_2 = 6$

Significance $F \approx 5.14$

$F = 19.0002 > 5.14$

we reject null hypothesis

Part - 3 Parametric vs. Non parametric Tests

(i) → Parametric and non-parametric tests are two categories of statistical tests. parametric test rely on assumptions about the distribution of population data, typically requiring a normal distribution while non-parametric test does not require any assumption

Difference between parametric and non-parametric tests

Parametric tests

These tests assume that the data comes from a population with a specific distribution, often a normal distribution. They also assumed data is measured.

* They also assumed data is an interval or ratio.
* The variance of the grap being compared are roughly equal.

Non - parametric tests

These tests do not make assumption about the distribution of the population data

* They are suitable for data measured on ordinal or nominal scales.

* They can handle outliers and non-normal distribution.

(ii) Examples

parametric test

* T-test ⇨ Comparing the average height of student in two different classes.

* Anova ⇨ Comparing the Average test scores of student across multiple teaching methods.

Non-parametric test

Mann-Whitney U test :-
Comparing the satisfaction levels (ranked on a scale) of patients receiving two different treatment

Kruskal-Wallis test :-
Comparing the preferences ranking of different brands of coffee.

Page No.
Date

## Non-normal data

If the data is not normally distributed, non-parametric tests are generally preferred. This is because parametric tests are sensitive to violations of the normality assumption and may produce inaccurate results.

Part-4 T-test

(9) Z-test vs. T-test

① What is the difference between a z-test and a T-test

Z-tests are statistical calculations that can be used to compare population means to samples.

T-test are calculation used to test a hypothesis but they are most useful when we need to determine if there is a statistically significant difference between two independent sample groups

| Bases | z-test | T-test |
|---|---|---|
| population std. deviation | Known | Uknown |
| Sample size | Generally used for larger samples. | Generally used for smaller samples. |
| Distribution | Assume a normal distribution of the population | Assume a normal distribution of the population, but also accounts for the uncertainty in estimating the population std. deviation |
| Test-statistics | Z-statistics | T-statistics |

Case Study

(a) P- value interpretation

A p-value is the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true, It quantifies the evidence against the null evidence against the null hypothesis.

A small p-value suggests that the observed data is unlikely under the null hypothesis indicating stronger evidence against the null

A large p-value suggest the data is Consistent with the null - hypothesis, providing less evidence against it.

(b) interpreting a p-value of 0.03 at $\alpha = 0.05$

If you obtain a p-value of 0.03 and are using a significance level ($\alpha$) of 0.05.

since $0.03 < 0.05$ you reject the null hypothesis

That means there is statistically significant evidence at the 5% level to suggest that the observed effect is not due to random chance.