Assignment:

Part-1: Descriptive statistics:

1.

a. Mean $= \frac{30+40+45+50+200}{5} = \frac{365}{5} = 73$

b. Median = 30, 40, $\boxed{45}$, 50, 200

middle value = 45

c. Mode (add another 40)

New list = 30, $\boxed{40, 40}$, 45, 50, 200

Mode = 40

d. Interpretation

The median is the best measure because the mean is heavily affected by the outlier 200. median represent the central salary value in skewed data

2. Variance & S.D (Scores : 5, 10, 20, 30)

mean $= \frac{5+10+20+30}{5} = 15$

Variance $= \frac{100+25+25+25+225}{5} = 80$

S.d $= \sqrt{80} = 8.94$

b. Interpretation: A standard deviation of 8.94 means the score are spread out fairly widely around 15.

3. Box plot & Outliers (Ages: 10, 12, 13, 15, 16, 18, 25, 35, 80)

a. Quartiles: Sorted: 10, 12, 13, 15, 16, 18, 25, 35, 80

$Q_1 = 13$

$Q_3 = 25$

$IQR = Q_3 - Q_1 = 25 - 13 = 12$

b. Outlier Rule: lower bound $= Q_1 - 1.5 \times IQR$

$= -5$

Upper bound $= Q_3 + 1.5 \times IQR = 43$

Anything above 43 is an Outlier → 80 is an outlier

Part-2: Inferential statistics

4. Regression analysis

a. Hours studied: 1, 2, 3, 4, 5, $\bar{x} = 3$

Exam score: 50, 55, 65, 70, 75 $\bar{y} = 63$

$\boxed{Y = 43.5 + 6.5x}$, $\beta_1 = 6.5$

$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$

$\beta_0 = 43.5$

5. Hypothesis testing (T-test)

$\mu_0 = 80\%$, sample mean $(\bar{x}) = 78\%$

$n = 50$, $SD = 5\%$, $\alpha = 0.05$

a. hypothesis: $H_0: \mu = 80\%$ (no improvement)

$H_1: \mu \neq 80\%$ (There is a change)

b. Test statistic:

$$t = \frac{78 - 80}{5\sqrt{50}} = \frac{-2}{0.707} \approx -2.83$$

c. Decision

t-critical (two-tailed, $df = 49$, $\alpha = 0.05$) ≈ ±2.01

Since $-2.83 < -2.01$ → Reject H_0

6. Chi-square test

	Like	Dislike	total
Male	40	60	100
Female	50	50	100
total	90	110	200

a. Null hypothesis (H_0): Gender & preference are Independent

b. Expected frequencies:

* Male - like = $(100 \times 90)/200 = 45$
* Male - dislike = $(100 \times 110)/200 = 55$
* Female like = 45
* Female dislike = 55

c. $\chi^2 = \Sigma (O-E)^2/E$,

$$\chi^2 = \frac{25}{45} + \frac{25}{55} + \frac{25}{45} + \frac{25}{55} \approx 2.02$$

d. Conclusion: χ^2 critical ($df = 1$, $\alpha = 0.05$) ≈ 3.84

≈ 2.02 < 3.84 → Fail to reject H_0 → no significant link between gender & preference

7. ANOVA (Group A, B, C)

A	B	C
70, 75, 80	60, 65, 70	85, 90, 95

Grand mean = $(225 + 195 + 270)/9 = 76.67$

a. Hypothesis

⇒ H_0: All means are equal

⇒ H_1: At least one mean is different

b. SSb (between group)

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

$= 954.84$

SSW (within group):

A = 25+0+25 = 50

B = 50

C = 50

SSW = 150

c. F - Ratio ⇒ dfb = 2, dfw = 6

⇒ MSb = SSb / 2 = 477.42

⇒ MSW = 150/6 = 25

F = 477.42 / 25 = 19.10

d. Conclusion: F - critical (df = 2, 6) ≈ 5.14

→ 19.10 > 5.14 → Reject H_0 → teach methods do affect exam score.

Part 3: Parametric vs Non Parametric

a. Difference ⇒ parametric: assumes normal distribu

(eg: t-test, ANOVA)

⇒ Non-parametric: no distribution assump

(eg: Mann-whitney, chi-square)

b. Example:

→ parametric : t- test, ANOVA

→ Non- parametric : Wilcoxon test, Chi-Square test

c. If data is not normal → Use non parametric

because They are safer and don't require strict distribution assumption

Part - 4: Z-test ~~and~~ Vs T- test

a. Difference : Z-test: population SD known, large sample

T-test: population SD, unknown, small sample

b. Use Z-test when:

⇒ population standard deviation is known

⇒ Sample size is large

* P-Value Interpretation:

a. p-value: The probability of getting results at least as extreme as observed, assuming H_0 is true

→ 0.03 < 0.05 → Reject H_0

→ There is significant evidence against The null hypothesis.