UNIT–1 : Descriptive Statistics (PYQ Solution)

PART–A Answers (≈70 Words Each)

Q1. Define mean, median and mode. Which measure of central tendency is best for skewed data? (Expected)

- **Mean $\bar{x} = \dfrac{\sum x}{n}$ is the arithmetic average; it uses every value but is pulled by extreme values.**
- **Median is the middle value of the ordered data; half the observations lie below it.**
- **Mode is the most frequent value.**

For skewed data the median is best because it is not affected by the long tail. In a positively skewed set (incomes) mean > median > mode.

Q2. Differentiate between population variance and sample variance. (Expected)

Population variance σ^2	**Sample variance s^2**
$\sigma^2 = \dfrac{\sum(x-\mu)^2}{N}$	$s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$
Uses all N members, true mean μ	Uses a sample, estimated mean $\bar{x}$
A fixed parameter	A statistic, varies sample to sample

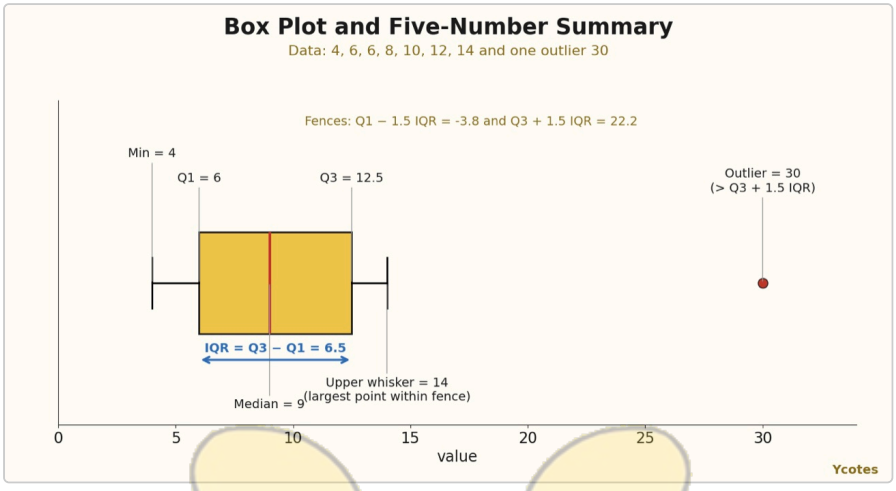
Dividing by $n-1$ (Bessel's correction) makes s^2 an unbiased estimator of σ^2, because the deviations are measured from $\bar{x}$, which is closer to the data than μ.

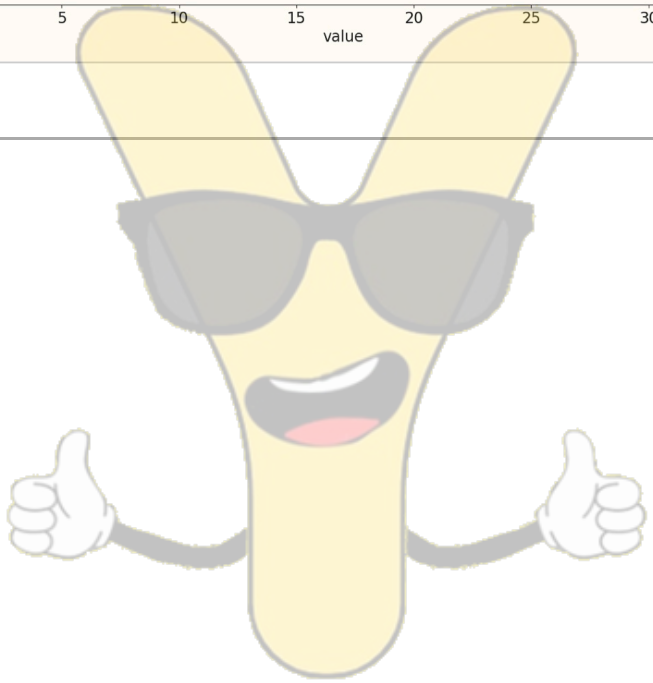
Q3. What is a box plot? Write the five-number summary and the outlier rule. (Expected)

A box plot (box-and-whisker plot) shows the spread and skew of data using five numbers: minimum, Q1, median, Q3, maximum. The box runs from Q1 to

Q3 with a line at the median; whiskers extend to the smallest and largest values inside the fences.

IQR $= Q_3 - Q_1$**. Outlier rule: a value below** $Q_1 - 1.5\,\mathrm{IQR}$ **or above** $Q_3 + 1.5\,\mathrm{IQR}$ **is an outlier and is drawn as a separate dot.**





PART–B Answers (≈250 Words Each)

Q1. Calculate the mean, median and mode of the following grouped frequency distribution. (Expected)

Data (assumed, N = 50):

Class	0-10	10-20	20-30	30-40	40-50
f	5	8	15	16	6
Mid-point m	5	15	25	35	45
fm	25	120	375	560	270
Cum. freq	5	13	28	44	50

Mean: $\bar{x} = \dfrac{\sum fm}{N} = \dfrac{1350}{50} = 27.$

Median: $N/2 = 25$ lies in class 20-30 (cum. freq 28).

$$\text{Median} = L + \frac{N/2 - cf}{f} \times h = 20 + \frac{25 - 13}{15} \times 10 = 28$$

Mode: modal class is 30-40 (f = 16). $\Delta_1 = 16 - 15 = 1,\ \Delta_2 = 16 - 6 = 10.$

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h = 30 + \frac{1}{11} \times 10 = 30.91$$

Check: mean (27) < median (28) < mode (30.9), so the distribution is slightly negatively skewed.

Q2. Define quartiles, deciles and percentiles. Find Q1, Q3 and P90 for grouped data. (Expected)

Definitions: quartiles divide ordered data into 4 equal parts (Q1, Q2 = median, Q3), deciles into 10 parts (D1 to D9), percentiles into 100 parts (P1 to P99).

Formula (grouped):

$$Q_k = L + \frac{kN/4 - cf}{f} \times h, \quad D_k:\ kN/10, \quad P_k:\ kN/100$$

Using the same data (N = 50, cum. freq 5, 13, 28, 44, 50):

- **Q1:** $N/4 = 12.5$ in class 10-20: $10 + \dfrac{12.5 - 5}{8} \times 10 = 19.375$
- **Q3:** $3N/4 = 37.5$ in class 30-40: $30 + \dfrac{37.5 - 28}{16} \times 10 = 35.94$
- **P90:** $0.9 \times 50 = 45$ in class 40-50: $40 + \dfrac{45 - 44}{6} \times 10 = 41.67$

Interquartile range $= 35.94 - 19.375 = 16.56$. Also $D_7 = 34.375$.

Q3. Find the weighted mean of marks 70, 80, 90 with weights 2, 3, 5, and the mean, median, range, variance and standard deviation of 4, 8, 6, 10, 12, 6, 14. (Expected)

Weighted mean: $\bar{x}_w = \dfrac{\sum wx}{\sum w} = \dfrac{2(70) + 3(80) + 5(90)}{10} = \dfrac{830}{10} = 83.$

Data 4, 8, 6, 10, 12, 6, 14 (n = 7): sorted 4, 6, 6, 8, 10, 12, 14.

- Mean $= 60/7 = 8.571$
- Median $= 8$ (4th value); mode $= 6$
- Range $= 14 - 4 = 10$
- $\sum (x - \bar{x})^2 = 77.71$
- **Population variance** $= 77.71/7 = 11.10$, SD $= 3.33$
- **Sample variance** $= 77.71/6 = 12.95$, SD $= 3.60$

Use the sample formula when the data are a sample from a larger population.

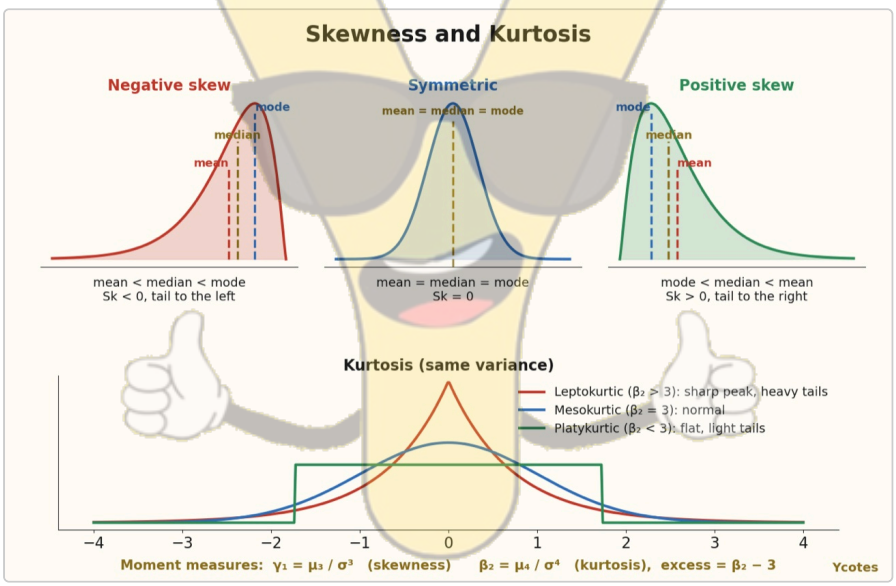
PART–C Answers (≈500 Words Each)

Q1. What are skewness and kurtosis? Compute both for the grouped data and interpret. (Expected)

Skewness measures the asymmetry of a distribution. Positive skew: long right tail (mean > median > mode). Negative skew: long left tail. Zero for symmetric data.

- Karl Pearson: $Sk = \dfrac{\text{Mean} - \text{Mode}}{SD}$
- Moment form: $\gamma_1 = \dfrac{\mu_3}{\sigma^3}$

Kurtosis measures peakedness and tail weight: $\beta_2 = \dfrac{\mu_4}{\sigma^4}$. $\beta_2 = 3$ mesokurtic (normal), > 3 leptokurtic (sharp, heavy tails), < 3 platykurtic (flat). Excess kurtosis $= \beta_2 - 3$.



Computation for the data (m = 5, 15, 25, 35, 45; f = 5, 8, 15, 16, 6; $\bar{x} = 27$):

m	f	$m - \bar{x}$	$f(m-\bar{x})^2$	$f(m-\bar{x})^3$	$f(m-\bar{x})^4$
5	5	-22	2420	-53240	1171280
15	8	-12	1152	-13824	165888
25	15	-2	60	-120	240
35	16	8	1024	8192	65536
45	6	18	1944	34992	629856
	50		6600	-24000	2032800

$\mu_2 = 6600/50 = 132$, $\sigma = 11.49$; $\mu_3 = -24000/50 = -480$; $\mu_4 = 2032800/50 = 40656$.

- Pearson $Sk = (27 - 30.91)/11.49 = -0.34$
- $\gamma_1 = -480/11.49^3 = -0.317$ (moderately negative skew)
- $\beta_2 = 40656/132^2 = 2.33$ (platykurtic, flatter than normal; excess $= -0.67$)

Interpretation: the left tail is longer and the curve is flatter than a normal curve.

Q2. Explain the measures of dispersion. Draw a box plot for 4, 6, 6, 8, 10, 12, 14, 30 and identify the outlier. (Expected)

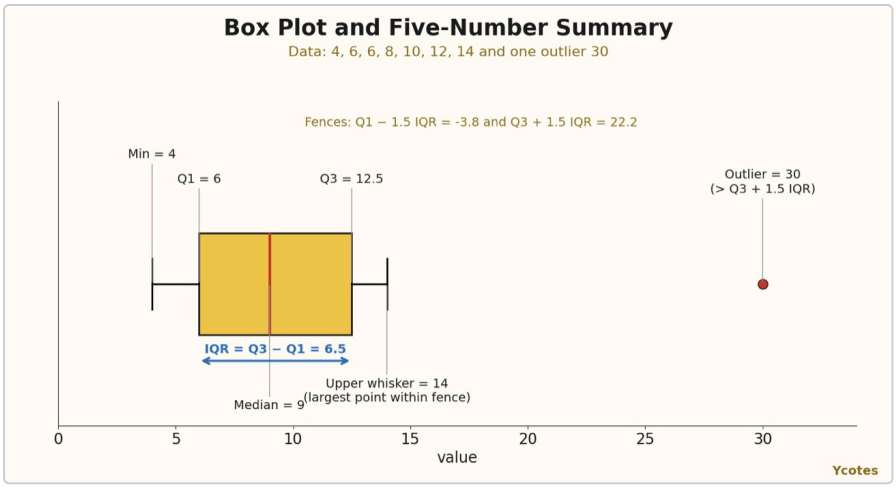
Measures of dispersion show how far data spread around the centre:

1. **Range** = max - min (quick, uses two values only)
2. **Interquartile range** $Q_3 - Q_1$ (robust to outliers)
3. **Mean deviation** $\sum |x - \bar{x}|/n$
4. **Variance** σ^2 and **standard deviation** σ (most used, same unit as data)
5. **Coefficient of variation** $CV = \dfrac{\sigma}{\bar{x}} \times 100\%$ to compare variability of two series

Box plot data (n = 8): sorted 4, 6, 6, 8, 10, 12, 14, 30.

- Median $= (8 + 10)/2 = 9$
- Using position $1 + (n-1)p$: Q_1 at 2.75 gives 6, Q_3 at 6.25 gives $12 + 0.25 \times 2 = 12.5$
- IQR $= 12.5 - 6 = 6.5$
- Lower fence $= 6 - 9.75 = -3.75$, upper fence $= 12.5 + 9.75 = 22.25$

Since **30 > 22.25**, the value 30 is an **outlier**; the upper whisker stops at 14.



Effect of the outlier: mean $= 11.25$ is pulled above the median 9, so for such data prefer the median and IQR.

