CHAPTER 3: NUMERICAL MEASURES

Summarizing Location, Variability, and Distribution Shape

I. Measures of Location

These measures tell us where the "center" or specific points of the data reside.

1. Mean ($\bar{x}$ or μ): The average. It is the most common measure but is sensitive to **outliers** (extreme values). **2. Median:** The middle value when data is in ascending order. If n is even, it is the average of the two middle values. It is **robust** against outliers. **3. Mode:** The most frequently occurring value. **4. Percentiles & Quartiles:** Percentiles provide information about how the data is spread over the interval from the smallest to largest value.

- Q_1 (1st Quartile): 25th percentile.
- Q_2 (Median): 50th percentile.
- Q_3 (3rd Quartile): 75th percentile.

II. Measures of Variability

Location isn't enough. Two stocks can have the same average return, but one might be much riskier (more volatile).

1. Range: Largest value minus smallest value. **2. Interquartile Range (IQR):** $Q_3 - Q_1$. It measures the spread of the middle 50% of the data and ignores outliers. **3. Variance (s^2):** The average of the squared deviations from the mean. **4. Standard Deviation (s):** The square root of the variance. It is in the **same units** as the original data, making it the most useful measure of risk/spread.

III. Distribution Shape & Z-Scores

A distribution can be **Symmetric** (bell-shaped), **Skewed Left** (long tail on the left), or **Skewed Right** (long tail on the right).

The Z-Score

The z-score tells us how many standard deviations a data point is from the mean:

$$z = \frac{x_i - \bar{x}}{s}$$

Points with $|z| > 3$ are typically considered **outliers**.

The Empirical Rule

For bell-shaped (normal) distributions:

- **68%** of data falls within $\pm 1s$.
- **95%** of data falls within $\pm 2s$.
- **99.7%** of data falls within $\pm 3s$.

IV. Five-Number Summary & Boxplots

The five-number summary consists of: *Min*, Q_1, *Median*, Q_3, *Max*. A **Boxplot** graphically displays these values. The "box" covers the IQR, and "whiskers" extend to the min/max (unless there are outliers).

V. Measures of Association

To understand how two variables (x and y) move together:

- **Covariance:** Indicates the direction of a linear relationship.
- **Correlation Coefficient (r):** A normalized version of covariance that ranges from -1 to $+1$.
  - $+1$: Perfect positive linear relationship.
  - 0: No linear relationship.
  - -1: Perfect negative linear relationship.

VI. Step-by-Step Example

Problem: Data: 2, 4, 6, 8, 10. Find Mean and Std. Deviation. **Logic:** 1. **Mean:** $(2+4+6+8+10)/5 = 6$. 2. **Deviations:** $(2-6), (4-6), (6-6), (8-6), (10-6) \rightarrow -4, -2, 0, 2, 4$. 3. **Squared Deviations:** $16, 4, 0, 4, 16$. 4. **Sum of Squares:** 40. 5. **Sample Variance (s^2):** $40/(5-1) = 10$. 6. **Std Deviation (s):** $\sqrt{10} \approx 3.16$.

VII. Practice Set

1. If a distribution is skewed heavily to the right, which will be larger: the Mean or the Median? 2. A data point has a z-score of -2.0. What does this mean in plain English? 3. Calculate the IQR: $Q_1 = 45, Q_3 = 70$, Mean $= 50$. 4. A set has a mean of 100 and std dev of 15. According to the Empirical Rule, what % of data is between 70 and 130?

VIII. Answer Key

1. **Mean.** (Extreme high values pull the mean up but don't affect the median as much). 2. The value is **2 standard deviations below** the mean. 3. $70 - 45 = \mathbf{25}$. 4. 100 ± 30 is $\pm 2s$, so **95%**.