Univariate Statistics

1. Data Types and Representation

- **Variables:**
  - **Qualitative:** Nominal, Ordinal
  - **Quantitative:** Discrete, Continuous
- **Measurement Scales:** Nominal, Ordinal, Interval, Ratio
- **Graphs:** Bar diagram (simple, composite, stacked), Line diagram, Pie diagram, Dot plots, Stem-and-leaf plots.

2. Frequency Distributions (Grouped Data)

- **Number of classes (k):**
  - 2^k **rule:** $2^k \geq n$, where n is the number of observations.
  - **Sturges' Rule:** $k = 1 + 3.322 \times \log_{10}(n)$
- **Graphs for Grouped Data:**
  - **Histogram:** X-axis: Class intervals/boundaries, Y-axis: Frequency.
  - **Frequency Polygon:** X-axis: Class marks (midpoints), Y-axis: Frequency.
  - **Ogive (Cumulative Frequency Polygon):** X-axis: Upper class boundaries, Y-axis: Cumulative frequency.
- **Unequal Class Intervals:**
  - **Relative Frequency** $= \frac{\text{Frequency}}{\text{Total Frequency}}$
  - **Frequency Density** $= \frac{\text{Frequency}}{\text{Class Width}}$
  - For Histograms with unequal classes, the Y-axis must be Frequency Density.

3. Measures of Central Tendency

Ungrouped Data

- **Arithmetic Mean (AM):**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  Property: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- **Median:** The middle value of an ordered dataset.
- **Mode:** The most frequent value.
- **Weighted Mean:**

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Geometric Mean (GM):**

$$GM = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \text{antilog}\left\{\frac{1}{n}\sum_{i=1}^{n}\log x_i\right\}$$

- **Harmonic Mean (HM):**

$$HM = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Grouped Data

Let x_i be the class mark (midpoint) of the i^{th} class, f_i be its frequency, and $n = \sum_{i=1}^{k} f_i$ be the total frequency.

- **Arithmetic Mean (AM):**

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$$

- **Geometric Mean (GM):**

$$GM = \text{antilog}\left\{\frac{1}{n}\sum_{i=1}^{k} f_i \log x_i\right\}$$

- **Harmonic Mean (HM):**

$$HM = \frac{n}{\sum_{i=1}^{k}\frac{f_i}{x_i}}$$

- **Relationship:** $AM \geq GM \geq HM$. Also, $GM^2 = AM \times HM$.
- **Median:**

$$\text{Median} = L_{\text{med}} + \left(\frac{\frac{n}{2} - F}{f_{\text{med}}}\right) \times c$$

  where:

  - L_{med} = Lower boundary of the median class
  - n = Total frequency
  - F = Cumulative frequency *before* the median class
  - f_{med} = Frequency of the median class
  - c = Class width of the median class

- **Mode:**

$$\text{Mode} = L_{\text{mod}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times c$$

  where:

  - L_{mod} = Lower boundary of the modal class
  - $\Delta_1 = f_{\text{mod}} - f_1$ (frequency of modal class - frequency of preceding class)
  - $\Delta_2 = f_{\text{mod}} - f_2$ (frequency of modal class - frequency of succeeding class)
  - c = Class width of the modal class

4. Measures of Dispersion

Ungrouped Data

- **Range:** $R = x_{\max} - x_{\min}$
- **Interquartile Range (IQR):** $IQR = Q_3 - Q_1$
- k^{th} **Quantile Position:** $\frac{k(n+1)}{q}$ -th observation (e.g., for Quartile Q_k, $q = 4$).
- **Mean Deviation (MD):**

$$MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

- **Sample Variance (s^2):**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

- **Sample Standard Deviation (s):** $s = \sqrt{s^2}$

Grouped Data

- **Mean Deviation (MD):**

$$MD = \frac{1}{n}\sum_{i=1}^{k} f_i|x_i - \bar{x}|$$

- **Sample Variance (s^2):**

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

- **Sample Standard Deviation (s):** $s = \sqrt{s^2}$
- **Coefficient of Variation (CV):**

$$CV = \left(\frac{s}{|\bar{x}|}\right) \times 100\%$$

- **Median Absolute Deviation (MAD):**

$$MAD = \text{Median}\left(|x_i - \tilde{x}|\right) \quad \text{(where } \tilde{x} \text{ is the data median)}$$

Range Rules and Standardized Score

- **Normal Rule (Empirical Rule):**
  - $[\bar{x} \pm s]$ contains approx. 68% of data.
  - $[\bar{x} \pm 2s]$ contains approx. 95% of data.
  - $[\bar{x} \pm 3s]$ contains approx. 99.7% of data.
- **Chebyshev's Rule:** For any $k > 1$, the proportion of data within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$.

$$P(\bar{x} - ks \leq X \leq \bar{x} + ks) \geq 1 - \frac{1}{k^2}$$

- **Z-score (Standardized Variable):**
$$z = \frac{x - \bar{x}}{s}$$

5. Positional Measures and Data Summary

Positional Measures

- **Quantiles (Ungrouped):** Position of k^{th} q-quantile is $\frac{k(n+1)}{q}$.
- **Quartile Deviation (QD):**
$$QD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}$$
- **k-th q-Quantile (Grouped):**
$$Q_{k/q} = L + \left(\frac{\frac{k \cdot n}{q} - F}{f} \right) \times c$$
  where:
  - L = Lower boundary of the quantile class
  - n = Total frequency
  - F = Cumulative frequency *before* the quantile class
  - f = Frequency of the quantile class
  - c = Class width

Data Summary

- **Five-Number Summary:** (Min, Q_1, Median, Q_3, Max)
- **Fences for Outlier Detection:**
  - **Inner Fences:**
    * Lower Inner Fence (LIF) = $Q_1 - 1.5 \times IQR$
    * Upper Inner Fence (UIF) = $Q_3 + 1.5 \times IQR$
  - **Outer Fences:**
    * Lower Outer Fence (LOF) = $Q_1 - 3 \times IQR$
    * Upper Outer Fence (UOF) = $Q_3 + 3 \times IQR$
- A Box-Whisker plot is used to represent the five-number summary.

6. Moments, Skewness, and Kurtosis

Moments

(Using n in the denominator for descriptive moments)

- r^{th} **Raw Moment (about origin):**
$$\mu'_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r$$

- r^{th} **Raw Moment (about 'a'):**

$$\mu'_r(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r$$

- r^{th} **Central Moment (about mean $\bar{x}$):**

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

  Note: $\mu_1 = 0$, $\mu_2 = \sigma^2$ (population variance, or s^2 if using $n-1$).

- **Relationship (Raw vs. Central):**
  - $\mu_2 = \mu'_2 - (\mu'_1)^2$
  - $\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$
  - $\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4$
- **Effect of Transformation ($y_i = a + bx_i$):**
  - $\bar{y} = a + b\bar{x}$
  - $\mu_r(y) = b^r\mu_r(x)$ (Central moments are independent of origin 'a')
- **Moments for Grouped Data:**
  - Raw: $\mu'_r = \frac{1}{n}\sum_{i=1}^{k} f_i x_i^r$
  - Central: $\mu_r = \frac{1}{n}\sum_{i=1}^{k} f_i(x_i - \bar{x})^r$

Sheppard's Correction (for Grouped Data Moments)

- $\mu_2(\text{corrected}) \approx \mu_2 - \frac{c^2}{12}$
- $\mu_4(\text{corrected}) \approx \mu_4 - \frac{c^2}{2}\mu_2 + \frac{7c^4}{240}$
- μ_1 and μ_3 need no correction.

Measures of Shape: Skewness

(Absence of symmetry)

- **Pearson's First Coefficient (SK_1):**

$$SK_1 = \frac{\bar{x} - \text{Mode}}{s}$$

- **Pearson's Second Coefficient (SK_2):**

$$SK_2 = \frac{3(\bar{x} - \text{Median})}{s}$$

- **Kelley's Coefficient (Decile-based):**

$$SK = \frac{D_9 + D_1 - 2D_5}{D_9 - D_1} \quad \text{(where } D_5 = \text{Median)}$$

- **Bowley's Coefficient (Quartile-based):**

$$SK = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \quad \text{(where } Q_2 = \text{Median)}$$

- **Moment-based Coefficient (γ_1):**

$$\gamma_1 = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{\mu_3}{s^3}$$

Measures of Shape: Kurtosis

(Peakedness or tailness of the distribution)

- **Leptokurtic** (high peak, fat tails), **Mesokurtic** (normal), **Platykurtic** (flat peak, thin tails).
- **Quantile-based Coefficient (K):**

$$K = \frac{Q.D.}{P_{90} - P_{10}} \quad \text{(where } Q.D. = \frac{Q_3 - Q_1}{2}\text{)}$$

- **Kelly's Coefficient (β):**

$$\beta = \frac{P_{75} - P_{25}}{P_{90} - P_{10}} = \frac{Q_3 - Q_1}{P_{90} - P_{10}}$$

- **Moors' Coefficient (Octile-based):**

$$K = \frac{(O_7 - O_5) + (O_3 - O_1)}{O_7 - O_1}$$

- **Moment-based Coefficient (β_2):**

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{\mu_4}{s^4}$$

- **Excess Kurtosis (γ_2):**

$$\gamma_2 = \beta_2 - 3$$

  ($\gamma_2 > 0$ is Leptokurtic, $\gamma_2 < 0$ is Platykurtic, $\gamma_2 = 0$ is Mesokurtic)