Descriptive Statistics UNIT 2

Dr. Praveen Barapatre

Arithmetic Mean

- Arithmetic Mean (Average): It's a way to find the central tendency of a dataset by adding all the values and dividing by the number of values.
- Sample vs. Population Mean:
  - Sample mean: Represents a smaller group of data points (sample) taken from a larger population. It's often estimated using the sample mean denoted by $\bar{x}$ (x bar) or $\hat{\mu}$ (mu hat).
  - Population mean: Represents the entire population of data points. It's denoted by the symbol μ (mu).
- Weighted Mean: When each data point has an importance or weight associated with it, we use a weighted mean. Weights are denoted by $\{f_i\}$.

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

The frequency weighted version (in which groups of x-values that are the same are each assigned a frequency value, f) is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}, \text{ or } \bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N} \text{ where } \sum_{i=1}^{n} f_i = N, \text{ or } \bar{x} = \sum_{i=1}^{n} x_i f_i \text{ if } \sum_{i=1}^{n} f_i = 1$$

- If X is a random variable with probability density function $f(x)$ then the arithmetic mean is the expected value of X, also written as E(X), or the Expected value of X:

$$E(X) = \sum_{x} xf(x) \text{ , or for the continuous case } E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

- **Power Means:** It mentions that many variants of mean values are special cases of power means. Power means are a broader category of averages where each data point is raised to a certain power (r) before being averaged.
- **Expected Value and Moments:** The expected value (E(X)) characterizes a probability distribution using its arithmetic mean. However, it's limited in describing the entire distribution. Other measures like moments (expected values of X raised to different powers) are crucial to understand the spread and shape of the data.

- R code examples: The text showcases how to calculate the mean in R using the mean(x) function for a single value and colMeans(swiss) for column-wise means of a matrix (using the SWISS dataset as an example).

Mean (harmonic)

- Definition: The harmonic mean is calculated by taking the reciprocal of the average of reciprocals of the data points. In simpler terms, it emphasizes smaller values in the dataset.
- Formula: $H = n / (1/x1 + 1/x2 + ... + 1/xn)$ where H is the harmonic mean, n is the number of data points, and xi are the individual data values.

Comparison with Arithmetic Mean:

- Arithmetic mean gives equal weight to all data points.
- Harmonic mean gives more weight to smaller values.
- Harmonic mean is generally less than or equal to the geometric mean, which is further less than or equal to the arithmetic mean.

- **Applications:**
  - Useful when dealing with rates or ratios (e.g., speed, price-to-earnings ratio).
  - Preferred when small denominators in ratios can significantly skew the arithmetic mean.
- **Example:** The passage uses a travel speed scenario to illustrate the difference between arithmetic mean (average speed) and harmonic mean (considering time spent at each speed).

- Reason for using the harmonic mean - avoiding situations where very small denominators in ratios disproportionately influence the arithmetic mean.
- Financial applications like calculating average price-to-earnings ratios for companies or portfolios are mentioned as areas where harmonic mean or weighted harmonic mean might be preferred.

Mean (geometric)

- The geometric mean (GM) of a set of data values {x1, x2, ..., xn} is the nth root of the product of all those values. In other words:

- GM = (x1 * x2 * ... * xn)^(1/n)

Applications:

The geometric mean is particularly useful when dealing with:

- **Rates of change:** When values are multiplied over time (e.g., investment growth rates, population growth rates).
- **Proportions or ratios:** It reflects the central tendency when multiplying ratios or percentages makes more sense than adding them (e.g., average return on investment over multiple periods).

Comparison with Arithmetic Mean:

- The geometric mean is generally **lower** than the arithmetic mean, especially when the data has a large spread or outliers.
- The arithmetic mean is more sensitive to very large or very small values, while the geometric mean gives more weight to values closer to the middle.

Example:

Imagine two investments: Investment A grows 20% in year 1 and 10% in year 2. Investment B grows 15% each year.

- **Arithmetic Mean:** Average annual growth for A: (20% + 10%) / 2 = 15%
- **Geometric Mean:** Growth rate of A considering both years together: ((1.2) * (1.1))^(1/2) ≈ 1.148 or 14.8% (compounded growth)

The geometric mean (14.8%) better reflects the actual overall growth of investment A because it considers the compounding effect.

- The geometric mean is a valuable tool for analyzing data where multiplication or ratios are more relevant than simple addition. It provides a different perspective on the "average" value compared to the arithmetic mean, especially for growth rates or proportional changes.

Mean (power)

- **Power mean**, also known as a **generalized mean** or **Hölder mean** (named after Otto Hölder). It's a broad category of averages that encompasses the arithmetic mean, geometric mean, and harmonic mean as special cases.

Here's the gist of power means:

- Formula: A power mean of order p for a set of data {x1, x2, ..., xn} is calculated as:
- **(x1^p + x2^p + ... + xn^p)^(1/p)**

Where:

- * p is the power that determines the type of mean (explained below).
- * n is the number of data points.

Varying the power (p) gives different means:

- **p = 1:** This reduces to the **harmonic mean** (focuses on smaller values).
- **p = 0:** As p approaches 0, the power mean approaches the **geometric mean** (emphasizes product of values).
- **p = 2:** This is the familiar **arithmetic mean** (gives equal weight to all values).
- **p > 2:** The mean gives more weight to larger values as p increases.
- **p < 1:** The mean gives more weight to smaller values as p gets closer to negative one (but not reaching -1).

Power means offer flexibility in calculating an "average" based on the data and the desired emphasis. They can be particularly useful when:

- The data has a large spread or outliers.
- You want to focus on either the larger or smaller values in the dataset.
- Neither the arithmetic nor geometric mean is ideal for your specific analysis.

Example:

Imagine income levels in a city: {10000, 50000, 100000, 1000000}.

- **Arithmetic Mean:** This would be skewed by the very high income, not reflecting the typical income level.
- **Geometric Mean:** This might underestimate the wealth gap in the city.
- **Power Mean (p = 0.5):** This would give more weight to lower and middle incomes, providing a better picture of the income distribution.

Mean (circular)

- The circular mean, also known as the angular mean, is a specific type of mean designed for analyzing data that represents angles, cyclical quantities, or rotations. Unlike the arithmetic mean used for everyday averages, the circular mean considers the circular nature of the data.

- **Focuses on Angular Relationships:** It considers the difference between angles and their periodic nature (e.g., 0 degrees and 360 degrees are equivalent).
- **Computation:** There are different methods to calculate the circular mean, but they generally involve converting the data points to unit vectors on a circle and then averaging those vectors.

Applications: Circular means are used in various fields where data is cyclical or angular, such as:

- Wind direction analysis
- Daily/seasonal variations in biological data
- Time series data with periodic patterns (e.g., sleep/wake cycles)
- Animal navigation patterns

Mode

Mode Definition:

- The mode is the most frequent value (or group of values) in a dataset.
- It can be applied to numerical or categorical data (e.g., hair color).

Types of Modes:

- **Unimodal:** The data has one dominant value or range of values.
- **Multimodal:** The data has several commonly occurring values or ranges.

Properties of the Mode:

- **Robust:** Less sensitive to outliers compared to mean or median.
- **Can be Non-Unique:** Multimodal data or uniform distributions can have multiple modes.
- **May Not Exist:** Certain distributions might not have a mode at all.

Challenges with Continuous Data:

- In continuous distributions, each value occurs only once, making the mode technically undefined.

Solutions:

- Discretization: Divide the data into intervals and find the most frequent interval. (Choice of interval width can affect results.)
- Slope-Based Approach: Identify points where the slope of the curve is 0 and transitions from positive to negative (local maxima).

Median

Definition:

- The median is the "middle" value in a dataset arranged in ascending or descending order. It represents the value that separates the higher half of the data from the lower half.

Calculation:

- **Odd Number of Data Points:** The median is the middle value when the data is ordered.
- **Even Number of Data Points:** The median is the average of the two middle values when the data is ordered.

Example:

- Data set: {2, 3, 5, 7, 9} (ordered)
  - Median: 5 (the middle value)
- Data set: {1, 4, 6, 8} (ordered)
  - Median: (4 + 6) / 2 = 5 (average of the two middle values)

Applications:

- The median is particularly useful for datasets with skewed distributions or outliers.
- It provides a clear understanding of the "typical" value in the data when the mean might be misleading.
- Used in various fields like income inequality, housing prices, etc.