Chapter 3: Descriptive Statistics: Numerical Methods

Outline

In this chapter we study numerical methods for describing the important aspects of a set of measurements. If the measurements are values of a quantitative variable, we often describe
(1) what a typical measurement might be and
(2) how the measurements vary, or differ, from each other.

For example, in the car mileage case we might estimate
(1)a typical EPA gas mileage for the new midsize model
(2) how the EPA mileages vary from car to car. Or, in the marketing research case,

Taken together, the graphical displays of Chapter 2 and the numerical methods of this chapter give us a basic understanding of the important aspects of a set of measurements.

3.1 Describing Central Tendency

Before we dive in, what comes to your mind when you hear the term "central tendency" or "average"?

> The ‘Big three’ of the Central Tendency

Central tendency is typically described using three key measures:

Mean, **Median**, and **mode (Check page 8 for reference)**

1. The mean (Average)

This is the most common measure. Coming from the basic formula for calculating an Average.

- Sample Means $(\bar{x})$:

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

Where $\sum$ means " sum of", x_i represents each value in the sample, and n in the sample size.

- Population Means (μ) :

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Where X_i represents each value in the population, and N is the total population size.

---

2. The Median (M_d)

The median is simple the middle value. To find it , you must first arrange your data in numerical order.

Have you ever had to find the median of a list of numbers?

Here is how we find it!

- Odd number of values: The median is the single value right in the middle.
  - Example : In the set {2, 5, 8, 11, 14}, the median is 8
- Even number of values: The median is the average of the two middlemost values.
  - Example: In the set {3,4,7,9, 12, 15}, the median is (7+9)/2=8

A key feature of the median is that it is resistant to extreme values (outliers). This makes it very useful for skewed data, like household incomes or salaries.

---

3. The mode (M_o)

The mode is the easiest to find. It's simply the value that occurs most frequently in the dataset. A dataset can have one node, more than one mode (bimodal/multimodal), or no mode at all.

- Example: In the set {2, 6, 7,7,7, 9, 10} the mode is 7

3.2 Measures of Variation

Qualifying the Scatter

We use several numbers to measure variation, but the main ones are the **Range, Variance**, and Standard Derivation.

1. The Range

This is the simplest measure of variation. It's the difference between the largest and smallest measurement in a dataset.

- **Range = Largest Value - Smallest Value**

---

2. Variance & Standard deviation

These are the most common and important measures of variation. The **standard deviation** is a measure of how much, on average, each data point deviates from the mean. The **variance** is just the standard deviation squared.

- **Variance (σ^2 for population, s^2 for sample):** The average of the squared deviations from the mean.

(Sample)

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

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

(Population)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- Standard Deviation (σ for population, s for sample): The square root of the variance.

(Sample)

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

(Population)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

A **small** standard deviation means the data points are clustered tightly around the mean. A **large** standard deviation means the data is more spread out.

Putting standard Deviation to Work

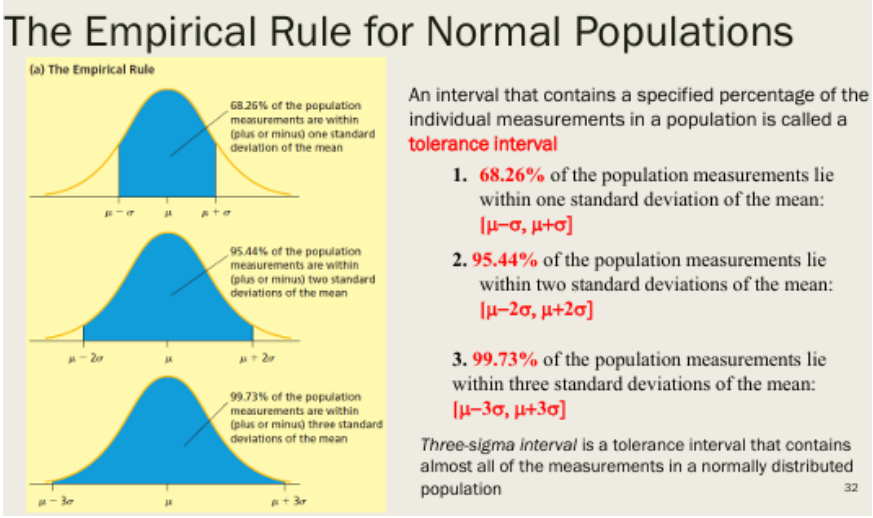
The standard deviation is incredibly useful, especially for data that follows a bell-shaped curve (a normal distribution).

The Empirical Rule Definition

Empirical Rule is a fundamental principle for data that follows a normal distribution (a symmetrical, bell-shaped curve). It formally states that for a **population** with mean μ (mu) and standard deviation **σ** (sigma), specific percentages of the data will fall within certain ranges around the mean.

For data that is approximately bell-shaped, we can use the Empirical Rule to estimate where most of the data lies. Does the "68-95-99.7" rule sound familiar?

- Approximately **68.26%** of data lies within 1 standard deviation of the mean: $[\bar{x} \pm s]$.
- Approximately **95.44%** of data lies within 2 standard deviations of the mean: $[\bar{x} \pm 2s]$.
- Approximately **99.73%** of data lies within 3 standard deviations of the mean: $[\bar{x} \pm 3s]$



Follow :

Gemini - Describing Data: Central Tendency and Variation

Created with Gemini

https://g.co/gemini/share/a3146ac9a2e3