

STAT 3011 Discussion 015

Week 3

Talha Hamza
University of Minnesota

Spring 2025

Measures of Central Tendency

- **Mean:** The arithmetic average.
- **Median:** The middle value when data are ordered.
- **Mode:** The most frequently occurring value.

Example Dataset: Let's use the number of coffees sold per hour: 4, 2, 5, 2, 3

Finding the Mean

Step 1: Add (sum) all of the values together. $4+2+5+2+3=16$

Step 2: Divide the sum by the number of values (n).

$$\text{Mean} = \frac{16}{5} = 3.2$$

The Mean

The mean number of coffees sold is **3.2**. On average, the shop sold 3.2 coffees per hour.

The Median

Step 1: Arrange the data in order from smallest to largest. 2 , 2 , 3 , 4 , 5

Step 2: Find the middle value.

- For an odd number of values ($n=5$), the median is the value in position 3.
- **Median = 3**

The Median

The median number of coffees sold is 3. Half of the hours had sales below 3, and half had sales above 3.

The Mode

Step 1: Count how many times each value appears in: **4, 2, 5, 2, 3.**

- 2 appears **2** times
- 3 appears **1** time
- 4 appears **1** time
- 5 appears **1** time

Step 2: The value with the highest frequency is the mode.

The Mode

The modal number of coffees sold is **2**. This was the most frequently occurring sales value in our data.

The Problem with Outliers: Mean vs. Median

New Scenario: You make a mistake and count an entry of 5 as 100 coffees!

Our data is now: 4, 2, **100**, 2, 3

Recalculate:

- **Mean:** $\frac{4+2+100+2+3}{5} = \frac{111}{5} = 22.2$

This is much higher and no longer represents a "typical" hour.

- **Median:** Order the data: 2, 2, 3, 4, **100**

The median number of coffees sold is 3. Half of the hours had sales below 3, and half had sales above 3.

The median is UNCHANGED by the extreme value.

Key Idea

The **median** is **resistant** to outliers, making it a better measure of centre for skewed data.

Why We Need the Mode

The mean and median are great for numerical data, but what about categories?

Example: "What is the most popular coffee size?"

Data: **Small, Medium, Large, Medium, Medium, Large**

- **Mean/Median?** Cannot be calculated! The data is not numerical.
- The **mode** is the **only measure of centre** that can be used for categorical data.

Measures of Spread

- **Range:** The difference between the maximum and minimum values.
- **Interquartile Range (IQR):** The spread of the middle 50% of the data.
- **Standard Deviation:** The average distance of each value from the mean.

Example: Let's use the same dataset of coffees sold per hour: **4, 2, 5, 2, 3**

The Range

Step 1: Identify the smallest and largest values.

Data (ordered): 2, 2, 3, 4, 5

Minimum = 2, Maximum = 5

Step 2: Subtract the minimum from the maximum.

$$\text{Range} = 5 - 2 = 3$$

The Range

The range tells us the spread between the lowest and highest sales.

Range = 3, so the shop's sales varied by 3 coffees per hour.

The Interquartile Range (IQR)

Step 1: Order the data: 2, 2, 3, 4, 5

Step 2: Split into halves around the median (3).

- Lower half: 2, 2
- Upper half: 4, 5

Step 3: Find Q1 (median of lower half) and Q3 (median of upper half).

$$Q1 = 2, \quad Q3 = 4.5$$

Step 4: Compute the IQR.

$$\text{IQR} = Q3 - Q1 = 4.5 - 2 = 2.5$$

The IQR

The IQR measures the spread of the middle 50% of the data.

Here, most sales vary within **2.5 coffees**.

What is Standard Deviation?

Formulas:

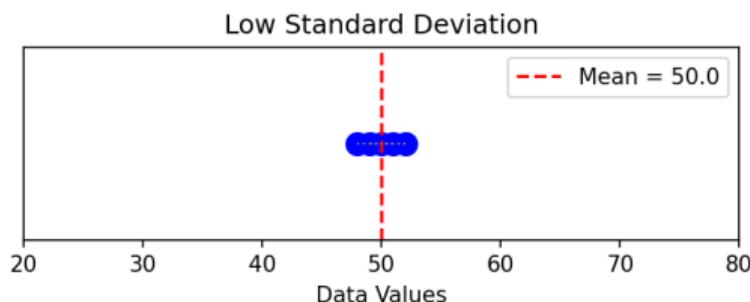
Population SD

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

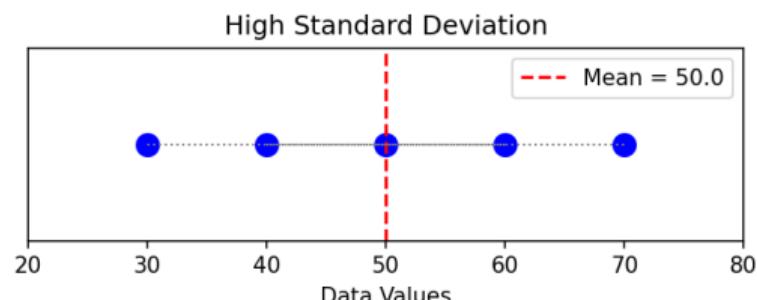
Sample SD

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

Visual Intuition:



Low SD: Values close to the mean



High SD: Values spread far from the mean

The Standard Deviation

Step 1: Recall the data: 4, 2, 5, 2, 3. **Mean** = 3.2

Step 2: Find each deviation from the mean.

$$(4 - 3.2) = 0.8, (2 - 3.2) = -1.2, (5 - 3.2) = 1.8, (2 - 3.2) = -1.2, (3 - 3.2) = -0.2$$

Step 3: Square the deviations and average them.

$$0.8^2 + (-1.2)^2 + 1.8^2 + (-1.2)^2 + (-0.2)^2 = 6.8$$

$$\text{Variance} = \frac{6.8}{5} = 1.36$$

Step 4: Take the square root.

$$\text{SD} = \sqrt{1.36} \approx 1.17$$

The Standard Deviation

The standard deviation is about **1.17**. On average, coffee sales vary about 1.17 from the mean of 3.2.

Comparing Measures of Spread

- **Range:** Simple but sensitive to outliers.
- **IQR:** Resistant to outliers; focuses on the middle 50%.
- **Standard Deviation:** Uses all data; shows typical distance from the mean.

Probability Formulas

General Addition Property of Probability:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Independence Rule:

If events A and B are independent, then:

$$P(A \cap B) = P(A) \cdot P(B)$$

Conditional Probability Formula:

Read as Probability of event A occurring GIVEN that B has already occurred

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{where } P(B) > 0$$

Understanding $P(A \cup B)$ and $P(A \cap B)$

$P(A \cup B)$: The probability of **either A or B** occurring.

- This includes the possibility of only A occurring, only B occurring, or both A and B occurring.
- Formula:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$P(A \cap B)$: The probability of **both A and B** occurring.

- Formula:

$$P(A \cap B) = P(A) \cdot P(B) \quad (\text{if A and B are independent})$$

Lab 2: R Cheat Sheet

Basics:

- `names(cereal)` Get column names
- `head(cereal)` View first 6 rows
- `nrow(cereal)` Get number of observations
- `str(cereal)` Check structure of dataset/variables

Summarizing Data:

- `table(cereal$Type)` Frequency table (Adult vs Child cereals)
- `summary(cereal$Sugar)` Five-number summary of Sugar
- `mean(cereal$Sugar)` Mean of Sugar variable
- `sd(cereal$Sugar)` Standard deviation of Sugar

Lab 2: R Cheat Sheet Continued

Visualizations:

- `hist(cereal$Sugar)` Histogram of Sugar
- `hist(cereal$Sugar, breaks=10)` Histogram with 10 bins
- `boxplot(cereal$Sugar)` Boxplot of Sugar
- `boxplot(Numerical variable ~ Categorical variable)` Side-by-side boxplots

Useful:

- Add labels: `main=""`, `xlab=""`, `ylab=""`
- Compare mean vs median to check skewness.
- Use boxplots to spot outliers.

Questions?