# MATH 108: Elementary Probability and Statistics

*Ramapo College of New Jersey*

Instructor: Dr. Atul Anurag

Semester: Fall 2025 Date: September 15, 2025

## Review: Empirical Rule

If the distribution of data is roughly bell-shaped (normal distribution), then:

– Approximately 68% of the data will lie within 1 standard deviation of the mean.

$$\mu - \sigma \leq \text{data value} \leq \mu + \sigma$$

– Approximately 95% of the data will lie within 2 standard deviations of the mean.

$$\mu - 2\sigma \leq \text{data value} \leq \mu + 2\sigma$$

– Approximately 99.7% of the data will lie within 3 standard deviations of the mean.

$$\mu - 3\sigma \leq \text{data value} \leq \mu + 3\sigma$$

**Note:** We can apply the Empirical Rule to sample data by using the sample mean $\bar{x}$ in place of the population mean $\mu$, and the sample standard deviation $s$ in place of the population standard deviation $\sigma$.

## Explanation

The Empirical Rule is important because it provides a quick way to understand the spread of data in a normal distribution. It helps us:

– Identify the proportion of data values within certain ranges.

– Detect outliers by seeing which data points fall far outside these ranges.

– Make probabilistic predictions and decisions based on how data clusters around the mean.

In practice, this rule is used widely in statistics, quality control, and various fields to assess the variability and reliability of data sets.

## Problem Statement

The distribution of the length of bolts has a bell shape (i.e., is approximately normal) with a mean of 4 inches and a standard deviation of 0.007 inch.

(a) About 68% of bolts manufactured will be between what lengths?

(b) What percentage of bolts will be between 3.986 and 4.014 inches?

(c) If the company discards any bolts less than 3.986 inches or greater than 4.014 inches, what percentage of bolts manufactured will be discarded?

(d) What percentage of bolts manufactured will be between 4.007 inches and 4.021 inches?

## Given

- Mean: $\mu = 4$ inches
- Standard Deviation: $\sigma = 0.007$ inch
- Distribution: Normal (bell-shaped)

## (a) About 68% of bolts manufactured will be between what lengths?

According to the Empirical Rule, about 68% of values in a normal distribution fall within one standard deviation of the mean.

$$\mu \pm \sigma = 4 \pm 0.007 = (3.993, 4.007)$$

**Answer:** Between **3.993 inches** and **4.007 inches**.

## (b) What percentage of bolts will be between 3.986 and 4.014 inches?

We calculate how many standard deviations these values are from the mean:

$$3.986 = 4 - 2(0.007), \quad 4.014 = 4 + 2(0.007)$$

So these values are 2 standard deviations below and above the mean.
By the Empirical Rule, about 95% of data falls within 2 standard deviations of the mean.
**Answer:** About **95%** of bolts are between **3.986 inches** and **4.014 inches**.

## (c) If the company discards any bolts less than 3.986 inches or greater than 4.014 inches, what percentage will be discarded?

From part (b), 95% of bolts fall between 3.986 and 4.014 inches.
That means:

$$100\% - 95\% = 5\%$$

**Answer:** About **5%** of bolts will be discarded.

## (d) What percentage of bolts will be between 4.007 and 4.021 inches?

We find how far these numbers are from the mean:

$$4.007 = 4 + 1(0.007), \quad 4.021 = 4 + 3(0.007)$$

This range is from 1 to 3 standard deviations *above* the mean.
Using the Empirical Rule:

- About 13.5% of data falls between 1 and 2 standard deviations above the mean.
- About 2.35% of data falls between 2 and 3 standard deviations above the mean.

Adding these gives:

$$13.5\% + 2.35\% = 15.85\%$$

**Answer:** About **15.85%** of bolts are between **4.007 inches** and **4.021 inches**.

## Question

Suppose that a variable is known to follow a bell-shaped distribution with mean 5 and standard deviation 2. What values account for the middle 95% of the data?

1. 5 and 7
2. 3 and 7
3. 1 and 11
4. **1 and 9**

## Solution

Since the distribution is bell-shaped (normal), we apply the empirical rule:

- The middle 95% of the data lies within 2 standard deviations of the mean.

$$\mu = 5, \quad \sigma = 2$$

$$\mu \pm 2\sigma = 5 \pm 2(2) = 5 \pm 4 = (1, 9)$$

## Question

Suppose that a variable is known to follow a bell-shaped distribution with a mean of 5 and a standard deviation of 2. What percent of the variable has a value between 3 and 7?

1. 68%
2. 34%
3. 5%
4. 95%

## Solution

We are given:

$$\mu = 5, \quad \sigma = 2$$

Now calculate how far the endpoints are from the mean:

$$3 = \mu - \sigma = 5 - 2, \quad 7 = \mu + \sigma = 5 + 2$$

So, the interval $(3, 7)$ represents one standard deviation on either side of the mean.

According to the Empirical Rule (68-95-99.7 rule), approximately:

$$68\% \text{ of values lie within } \mu \pm \sigma$$

## Approximate the Mean of a Variable from Grouped Data

Before learning how to compute the mean from grouped data, it is important to understand how to determine the **class midpoint** of each class.

### Class Midpoint

The class midpoint is the value that represents all data points in a class interval. For grouped data where classes are defined by their lower limits, the class midpoint is calculated by:

$$\text{Class midpoint} = \frac{\text{Lower class limit of current class} + \text{Lower class limit of next class}}{2}$$

This value represents the center of the class.

### Formula for Approximate Mean from Grouped Data

The mean can be approximated by:

$$\mu = \frac{\sum(x_i f_i)}{\sum f_i}$$

where

- $x_i$ = midpoint of the $i$th class,
- $f_i$ = frequency of the $i$th class,
- $n$ = number of classes.

The term $x_i f_i$ approximates the sum of all data values in the $i$th class. The formula sums these approximations and divides by the total frequency.

### Steps to Approximate the Mean

1. Determine the class midpoint of each class by adding consecutive lower class limits and dividing by 2.
2. Compute the sum of the frequencies, $\sum f_i$.

3. Multiply each class midpoint by its frequency to obtain $x_i f_i$ for each class.
4. Sum all the $x_i f_i$ values.
5. Calculate the mean using the formula:

$$\text{Population mean,} \quad \mu = \frac{\sum(x_i f_i)}{\sum f_i} \quad \text{and} \quad \text{Sample mean,} \quad \bar{x} = \frac{\sum(x_i f_i)}{\sum f_i}$$

### Example: Five-Year Rate of Return of Mutual Funds

Consider the frequency distribution of the five-year rate of return of a sample of 40 large-blended mutual funds shown in Table **??**.

Table 1: Five-Year Rate of Return and Frequencies

| **Class (Five-Year Rate of Return)** | **Frequency** |
|---|---|
| 8 – 8.99 | 2 |
| 9 – 9.99 | 2 |
| 10 – 10.99 | 4 |
| 11 – 11.99 | 1 |
| 12 – 12.99 | 6 |
| 13 – 13.99 | 13 |
| 14 – 14.99 | 7 |
| 15 – 15.99 | 3 |
| 16 – 16.99 | 1 |
| 17 – 17.99 | 0 |
| 18 – 18.99 | 0 |
| 19 – 19.99 | 1 |

**Step 1: Calculate Class Midpoints**

Using consecutive lower class limits, the midpoints are:

| Class | Midpoint |
|---|---|
| $8 - 8.99$ | $\frac{8+9}{2} = 8.5$ |
| $9 - 9.99$ | $\frac{9+10}{2} = 9.5$ |
| $10 - 10.99$ | $\frac{10+11}{2} = 10.5$ |
| $11 - 11.99$ | $\frac{11+12}{2} = 11.5$ |
| $12 - 12.99$ | $\frac{12+13}{2} = 12.5$ |
| $13 - 13.99$ | $\frac{13+14}{2} = 13.5$ |
| $14 - 14.99$ | $\frac{14+15}{2} = 14.5$ |
| $15 - 15.99$ | $\frac{15+16}{2} = 15.5$ |
| $16 - 16.99$ | $\frac{16+17}{2} = 16.5$ |
| $17 - 17.99$ | $\frac{17+18}{2} = 17.5$ |
| $18 - 18.99$ | $\frac{18+19}{2} = 18.5$ |
| $19 - 19.99$ | 19.5 (approximate) |

**Step 2 and 3: Calculate $f_i$ and $x_i f_i$**

**Step 4: Calculate the Mean**

Using the formula:

| Class | Frequency $f_i$ | Midpoint $x_i$ | $f_i \times x_i$ |
|---|---|---|---|
| 8 – 8.99 | 2 | 8.5 | 17.0 |
| 9 – 9.99 | 2 | 9.5 | 19.0 |
| 10 – 10.99 | 4 | 10.5 | 42.0 |
| 11 – 11.99 | 1 | 11.5 | 11.5 |
| 12 – 12.99 | 6 | 12.5 | 75.0 |
| 13 – 13.99 | 13 | 13.5 | 175.5 |
| 14 – 14.99 | 7 | 14.5 | 101.5 |
| 15 – 15.99 | 3 | 15.5 | 46.5 |
| 16 – 16.99 | 1 | 16.5 | 16.5 |
| 17 – 17.99 | 0 | 17.5 | 0.0 |
| 18 – 18.99 | 0 | 18.5 | 0.0 |
| 19 – 19.99 | 1 | 19.5 | 19.5 |
| **Totals** | **40** | | **524.5** |

$$\mu = \frac{\sum(x_i f_i)}{\sum f_i} = \frac{524.5}{40} = 13.11$$

## Compute the Weighted Mean

When data values have different importance, or **weight**, associated with them, we compute the **weighted mean**. For example, your grade-point average (GPA) is a weighted mean, where the weights are the number of credit hours for each course, and the values are the grade points earned.

**Definition:** The weighted mean, denoted by $\bar{x}_w$, of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. The formula is:

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \tag{2}$$

where:

- $w_i$ = weight of the $i$th observation,
- $x_i$ = value of the $i$th observation,
- $n$ = number of observations.

This formula calculates the average value considering the relative importance (weight) of each observation.

## Example: Computing the Weighted Mean

**Problem:** Marissa just completed her first semester in college. She earned the following grades:

- A in her 4-hour statistics course,
- B in her 3-hour sociology course,
- A in her 3-hour psychology course,
- C in her 5-hour computer programming course,

- A in her 1-hour drama course.

Determine Marissa's grade-point average (GPA).

**Approach:** Assign point values to each grade:

$$\text{A} = 4 \text{ points}, \quad \text{B} = 3 \text{ points}, \quad \text{C} = 2 \text{ points}$$

The number of credit hours for each course is the weight $w_i$. Multiply the weight of each course by the points earned, sum these products, and divide by the sum of the weights (total credit hours).

**Solution:**

$$\text{GPA} = \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{4 \times 4 + 3 \times 3 + 3 \times 4 + 5 \times 2 + 1 \times 4}{4 + 3 + 3 + 5 + 1} = \frac{51}{16} = 3.19$$

Marissa's grade-point average for her first semester is $\boxed{3.19}$.

# Measures of Position and Outliers

## Objectives

By the end of this section, you will be able to:

- Determine and interpret **z-scores**.
- Interpret **percentiles**.
- Determine and interpret **quartiles**.
- Determine and interpret the **interquartile range (IQR)**.
- Check a set of data for **outliers**.

## Introduction

Measures of position help us understand the relative standing of a data value within a data set. They provide context by comparing a value to the rest of the data, rather than just focusing on the center or spread.

We will study the following important measures of position:

- **Z-scores:** Indicate how many standard deviations a data value is from the mean.
- **Percentiles:** Indicate the relative position of a data value in a dataset by showing the percentage of values below it.
- **Quartiles:** Divide the dataset into four equal parts.
- **Interquartile Range (IQR):** Measures the spread of the middle 50% of the data.
- **Outliers:** Data points that lie far outside the typical range of the dataset.

Each of these measures helps us analyze data more deeply and identify unusual observations that may require further attention.

## Z-Scores

The **z-score** represents the distance that a data value is from the mean in terms of the number of standard deviations. It is calculated by subtracting the mean from the data value and then dividing this result by the standard deviation.

There are two types of z-scores: one for a population and one for a sample.

$$\text{Population z-score:} \quad z = \frac{x - \mu}{\sigma}$$

$$\text{Sample z-score:} \quad z = \frac{x - \bar{x}}{s}$$

where:

- $x$ = data value,
- $\mu$ = population mean,
- $\bar{x}$ = sample mean,
- $\sigma$ = population standard deviation,
- $s$ = sample standard deviation.

The z-score is **unitless** and has a mean of 0 and a standard deviation of 1.

Interpreting z-scores:

- A positive z-score indicates the data value is above the mean.
- A negative z-score indicates the data value is below the mean.
- The magnitude of the z-score indicates how many standard deviations the value is from the mean.

## Percentiles

The **kth percentile**, denoted $P_k$, of a set of data is a value such that $k$ percent of the observations are less than or equal to that value.

In other words, the $k$th percentile divides the data so that approximately $k\%$ of the data lies at or below $P_k$, and $(100 - k)\%$ lies above $P_k$.

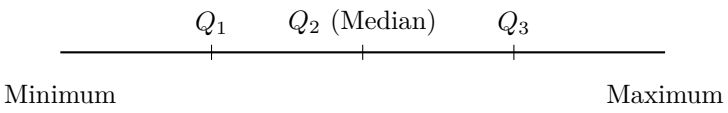
This illustration shows the percentiles arranged along the data range, with $P_1$ near the smallest values and $P_{99}$ near the largest.

## Quartiles

**Quartiles** divide a data set into four equal parts, each containing approximately 25% of the data.

- The **first quartile** ($Q_1$) is the 25th percentile ($P_{25}$). It separates the lowest 25% of data from the rest.
- The **second quartile** ($Q_2$) is the 50th percentile ($P_{50}$), which is also the **median** of the data.
- The **third quartile** ($Q_3$) is the 75th percentile ($P_{75}$), separating the lowest 75% of data from the highest 25%.

These quartiles provide important information about the spread and center of the data.




**Interpretation:**

- About 25% of data lies below $Q_1$.
- About 50% of data lies below $Q_2$ (median).
- About 75% of data lies below $Q_3$.

## Finding and Interpreting Quartiles

The Highway Loss Data Institute routinely collects data on collision coverage claims. Collision coverage insures against physical damage to an insured individual's vehicle. The data in Table 16 represent a random sample of 18 collision coverage claims.

**Problem:** Find and interpret the first, second, and third quartiles for collision coverage claims.

**Table 16** Collision Coverage Claims (in dollars):

| | | | | | |
|---|---|---|---|---|---|
| 6751 | 9908 | 3461 | 2336 | 21147 | 2332 |
| 189 | 1185 | 370 | 1414 | 4668 | 1953 |
| 10034 | 735 | 802 | 618 | 180 | 1657 |

**Step 1: Sort the data in ascending order**

| | | | | | |
|---|---|---|---|---|---|
| 180 | 189 | 370 | 618 | 735 | 802 |
| 1185 | 1414 | 1657 | 1953 | 2332 | 2336 |
| 3461 | 4668 | 6751 | 9908 | 10034 | 21147 |

**Step 2: Find the median $Q_2$**

There are $n = 18$ data points (even), so the median is the average of the 9th and 10th values:

$$Q_2 = \frac{1657 + 1953}{2} = \frac{3610}{2} = 1805$$

**Step 3: Split the data into lower and upper halves**

- Lower half (first 9 values):
  $180, 189, 370, 618, 735, 802, 1185, 1414, 1657$
- Upper half (last 9 values):
  $1953, 2332, 2336, 3461, 4668, 6751, 9908, 10034, 21147$

**Step 4: Find the first quartile $Q_1$**

The median of the lower half (9 values) is the 5th value:

$$Q_1 = 735$$

**Step 5: Find the third quartile $Q_3$**

The median of the upper half (9 values) is the 5th value:

$$Q_3 = 4668$$

**Final Results:**

$$Q_1 = 735, \quad Q_2 = 1805, \quad Q_3 = 4668$$

**Interpretation:**

- About 25% of collision coverage claims are less than or equal to \$735.
- About 50% (the median) of claims are less than or equal to \$1805.
- About 75% of claims are less than or equal to \$4668.

## Interquartile Range (IQR)

The **interquartile range (IQR)** measures the spread of the middle 50% of the data. It is calculated as the difference between the third quartile ($Q_3$) and the first quartile ($Q_1$):

$$\text{IQR} = Q_3 - Q_1$$

The IQR is useful for understanding the variability of the central portion of the data and is less affected by extreme values or outliers than the overall range.

## Outliers

**Outliers** are data values that lie far outside the typical range of the data. They may indicate variability in measurement, experimental errors, or novel insights.
A common method to detect outliers uses the IQR:

- Calculate the **lower boundary**:
$$Q_1 - 1.5 \times \text{IQR}$$
- Calculate the **upper boundary**:
$$Q_3 + 1.5 \times \text{IQR}$$

Any data value below the lower boundary or above the upper boundary is considered an outlier.

**Summary:**

$$\text{Outlier if } x < Q_1 - 1.5 \times \text{IQR} \quad \text{or} \quad x > Q_3 + 1.5 \times \text{IQR}$$

Detecting outliers is important because they can affect measures of central tendency and variability, and may indicate errors or interesting phenomena in the data.

---

*End of Lecture #5*