MATH 108: Elementary Probability and Statistics

Ramapo College of New Jersey

Instructor: Dr. Atul Anurag

Semester: Fall 2025 Date: August 28, 2025

Introduction to the Practice of Statistics

Learning Objectives:

- Understand the definition and importance of statistics
- Distinguish between population and sample
- Differentiate qualitative and quantitative data
- Recognize levels of measurement and their implications
- Distinguish discrete and continuous variables
- Understand the concepts of descriptive and inferential statistics
- Understand variability and lurking variables
- Comprehend statistical certainty vs. mathematical certainty

1. What is Statistics?

Statistics is more than just numbers. It involves understanding where numbers come from, how they are collected, organized, summarized, and analyzed to draw conclusions about reality. Statistics also evaluates how accurately numbers reflect reality.

- **Collection:** Gathering information (data).
- **Organization and Summarization:** Presenting data in manageable forms.
- **Analysis:** Using data to answer questions or draw conclusions.
- **Reporting:** Communicating results with understanding of their reliability.

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make informed decisions.

Branches of Statistics

- **Descriptive Statistics:** Focuses on summarizing data using numbers (e.g., averages) and visuals (e.g., graphs).
- **Inferential Statistics:** Makes generalizations about a population based on a sample and uses probability theory.

Data are facts or propositions used to make decisions or draw conclusions. Data can be numerical (quantitative) or non-numerical (qualitative).

Data are powerful but can be misleading if collected or analyzed improperly. For example, biased polls with non-representative samples do not reflect the true population.

2. Variability and Lurking Variables

Data inherently vary — people, objects, and measurements are rarely identical. This variability is why two polls on the same topic might yield different results.

A *lurking* variable is an unaccounted factor that influences the relationship between studied variables. For example, a study linking breastfeeding to higher IQ might be confounded by the mother's IQ rather than breastfeeding itself.

When Data Show a Relationship — Do We Accept It Immediately?

Sometimes, data show a relationship between two variables. But **does a relationship mean causation**? **Not necessarily.**

Example: Breastfeeding and Child IQ

- A study found that children who were breast-fed had higher IQs than those who were not.
- Does this mean breastfeeding *causes* higher IQ?
- **Not necessarily** — another factor may be at play.

The Lurking Variable

- It turns out that mothers who breastfeed tend to have higher IQs themselves.
- So, the **genetics of the mother** (not breastfeeding) may contribute to the child's higher IQ.

- This hidden factor is called a **lurking variable**.

A good statistical study accounts for possible lurking variables to avoid misleading conclusions.

Variation: A Key Feature of Data

Variation is central to understanding data. Data are not constant — they change and differ.

Variation Within Groups

- Are all students in the class the same height or hair color?
- Clearly, **no** — there is **variation within groups**.

Variation Within Individuals

- Do you eat the same amount every day? Sleep the same number of hours?
- Again, **no** — there is **variation within individuals**.

Goal of Statistics: To **describe and understand** the sources of variation in data.

Statistics vs. Mathematics: Dealing with Uncertainty

In Mathematics

- When a problem is solved correctly, the answer is exact.
- Example: Solve $3x + 5 = 11$
- Both Bob and Jane will find $x = 2$.

In Statistics

- Answers can differ even when procedures are followed correctly.
- Example: Estimate the average commute time in Dallas, Texas.
- Bob and Jane might survey different people, leading to different results.

Conclusion: In statistics, results can vary because **data vary**.

4. Understanding Confidence, Not Certainty

In statistics, we often use confidence levels instead of absolute certainty.

Example

- We may say: "We are 95% confident that the average commute time in Dallas is between 20 and 23 minutes."
- This means the true value likely falls in that range, but not with 100% certainty.

Why Is Statistics Useful If It's Not Certain?

Even without certainty, statistics is incredibly valuable for making informed decisions.

What Statistics Helps You Do:

- Understand and interpret the world
- Recognize and manage variability
- Critically evaluate media and research
- Make informed financial and consumer choices
- Become a better-informed and statistically literate citizen

Example Scenario

You are walking down the street and notice that a person walking in front of you drops $100. Nobody seems to notice the $100 except you. Since you could keep the money without anyone knowing, would you keep the money or return it to the owner?

Goal: *You want to study the morality of students by seeing what percent would return a lost $100 bill.*

- **Population:** All students at the school.
- **Sample:** A subset of 50 students surveyed.
- **Descriptive Statistics:** Summary of the sample (e.g., 78% said they would return the money).
- **Inferential Statistics:** Generalizing from the sample to the population with a confidence interval (e.g., 95% confident that between 74% and 82% of all students would return the money).

Definitions

- **Population:** Entire group being studied.
- **Individual:** A single member of the population.
- **Sample:** A subset of the population.
- **Parameter:** A numerical measure describing a population characteristic (e.g., population mean μ).
- **Statistic:** A numerical measure describing a sample characteristic (e.g., sample mean $\bar{x}$).
- **Descriptive Statistics:** Organizing and summarizing data.
- **Inferential Statistics:** Extending results from sample to population with a measure of reliability.

Parameter versus Statistic

Suppose 48.2% of all students on your campus own a car. This value represents a parameter because it is a numerical summary of a population. Suppose a sample of 100 students is obtained, and from this sample we find that 46% own a car. This value represents a statistic because it is a numerical summary of a sample.

5. The Process of Statistics

1. **Identify the Research Objective**
   Clearly define the specific question(s) to be answered and identify the population of interest. A well-defined objective provides direction for the entire study.
2. **Collect Data**
   Gather relevant data—typically from a sample, not the entire population. Proper sampling methods and data collection are essential; poor data can lead to invalid results.
3. **Describe the Data**
   Use descriptive statistics (such as graphs, tables, and numerical summaries) to organize and summarize the data. This step helps in understanding the data and informs the next phase.
4. **Perform Inference**
   Apply statistical methods to draw conclusions about the population based on the sample data. Include a measure of reliability, such as a confidence interval or margin of error, to communicate the uncertainty in the results.

Example: Statistical Study on Minimum Wage Opinion

CBS News and the New York Times conducted a poll from September 12–15, 2014, asking:
"As you may know, the federal minimum wage is currently $7.25 an hour. Do you favor or oppose raising the minimum wage to $10.10?"
The statistical process used in this study followed the four-step approach:

1. **Research Objective:**
   Determine the proportion of adult Americans who favor raising the federal minimum wage to $10.10 per hour. The population of interest is all adult Americans.
2. **Data Collection:**
   Surveying every adult American is impractical. Instead, a sample of 1,009 adult Americans was surveyed. Of these, 706 respondents stated they favored the wage increase.
3. **Descriptive Statistics:**
   In the sample, 70% (= 706/1009) favored raising the minimum wage. This percentage is a descriptive statistic because it summarizes the sample data.
4. **Inference:**
   The goal is to generalize the sample results to the entire population of adult Americans. To express the uncertainty in this estimate, the researchers reported a margin of error of 3%.

6. Types of Data and Variables

Qualitative (Categorical) Variables

Qualitative, or categorical, variables allow for classification of individuals based on some attribute or characteristic.
Describe attributes or categories. Arithmetic operations are not meaningful.

Quantitative (Numerical) Variables

Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.
Numerical and allow meaningful arithmetic operations.

Examples

- Gender: Qualitative.
- Temperature: Quantitative.

- Number of days studied: Quantitative.
- Zip code: Qualitative (categorical despite numeric form).

7. Discrete vs. Continuous Quantitative Variables

- **Discrete:** Countable values, finite or countably infinite. Examples: Number of heads in 5 coin flips, number of cars arriving.
- **Continuous:** Infinite possible values in an interval, measurable. Example: Distance a car travels on a full tank.

8. Individuals, Variables, and Levels of Measurement

Explain through an example on page 9 of the textbook in Section 1.1.

- **Individuals:** Countries (Australia, Canada, etc.)
- **Variables:**
  - Government type (qualitative, nominal)
  - Life expectancy (quantitative, continuous)
  - Population (quantitative, discrete)
- **Data Example (Poland):** Republic government, 77.40 years life expectancy, 38.6 million population.

Levels of Measurement

- **Nominal:** Categories without order (e.g., government type, hair color).
- **Ordinal:** Categories with a meaningful order (e.g., small, medium, large).
- **Interval:** Ordered categories with meaningful differences; zero does not mean absence (e.g., temperature in Celsius).
- **Ratio:** Interval level with meaningful zero; ratios are meaningful (e.g., weight, height, income, age).

Important: Level of measurement determines the types of analysis that are appropriate.

9. Descriptive Statistics Measures

Mean (Average)

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

Median

The middle value when data is ordered. If even number of data points, average the two middle values.

Mode

The value(s) that occur most frequently. Data may have:

- One mode (unimodal)
- Two modes (bimodal)
- More than two modes (multimodal)
- No mode

10. Example Problem: Quiz Scores

A class of 10 students scored the following on a quiz:

$$8,\ 7,\ 9,\ 10,\ 6,\ 7,\ 8,\ 7,\ 9,\ 10$$

Step 1: Organize the data (ascending order):

$$6,\ 7,\ 7,\ 7,\ 8,\ 8,\ 9,\ 9,\ 10,\ 10$$

- **Mean:** $\bar{x} = \frac{81}{10} = 8.1$
- **Median:** $\frac{8+8}{2} = 8$
- **Mode:** 7 (occurs 3 times)

11. Summary Table of Concepts

Concept	Description
Population	Entire group of interest
Sample	Subset of population
Parameter	Numerical measure from population
Statistic	Numerical measure from sample
Qualitative Data	Non-numeric, categorical
Quantitative Data	Numeric (discrete or continuous)
Levels of Measurement	Nominal, Ordinal, Interval, Ratio

Reading: Sections 1.1 in the textbook

Homework: Problems 1(a, b, c, f, g, h), 2, 11, 13, 15, 17, 19, 21, 23, 25, 35, 37, and 58.

End of Lecture #1