Intro to Probability

Probability Theory: Foundation for Data Science

with **Anne Dougherty**

Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning Goals for Module 1

In this module, we'll learn about the difference between a population and a sample and why probability is the foundation for statistics and data science. At the end of this Module, students should be able to:

- **Explain why probability theory is relevant to statistics and data science.**
- **Describe what it means to predict the outcome of an experiment and organize the outcomes into sample spaces.**
- Calculate probabilities of events using the Axioms of Probability.
- Understand permutations and combinations and be able to calculate probabilities when each simple event is equally likely.

What is Statistics?

Statistics is the science of using data effectively to gain new knowledge. We need data to learn something new. We need to collect and analyze the data ethically.

Population: Those individuals or objects from which we want to acquire information or draw a conclusion. Most of the time, the population is so large, we can only collect data on a subset of it. We will call this our **sample**.

In **probability** we assume we know the characteristics of the entire population. Then, we can pose and answer questions about the nature of a sample. In **statistics**, if we have a sample with particular characteristics, we want to be able to say, with some degree of confidence, whether the whole population has this characteristic, or not.

Sample Spaces and Events

Probability studies randomness and uncertainty by giving these concepts a mathematical foundation.
For example, we want to understand how to find the probability

- of getting at least 2 heads in 5 coin flips,
- that a customer will buy milk if they are also buying bread,
- that the price of a stock will be in a certain range on a certain date in the future.

Probability gives us the framework to quantify uncertainty.

Terminology

- An **experiment** is any action or process that generates observations.
- The **sample space** of an experiment, denoted S, is the set of all possible outcomes of an experiment.
- An **event** is any possible outcome, or combination of outcomes, of an experiment.
- The **cardinality** of a sample space or an event, is the number of outcomes it contains. $|S|$ represents the cardinality of the sample space.

Examples

For each of the following, describe the sample space, S, and give its cardinality.

- Experiment 1: Flip a coin twice
- Experiment 2: Flip a coin until you get a tail.
- Experiment 3: Select a car coming off an assembly line and inspect it for 3 different defects (engine problem, seat belt problem, bad paint job).
- Experiment 4: Measure the arrival time between two customers.

Set Notation

For events A and B,

- $A \cup B$, the **union** of A and B, means an outcome in A or an outcome of B occurs.
- $A \cap B$, the **intersection** of A and B, is all the outcomes that are in both A and B
- A^c, the **complement** of A, means the set of all events in $\mathcal{S}$ that are not in A
- A and B are mutually exclusive, or disjoint, if they have no events in common. We write $A \cap B = \emptyset$.

Examples continued

$S = \{000, 100, 010, 001, 110, 101, 011, 111\}$ Consider the following events:

- A is the event that there is an engine problem (defect 1). In set notation: $A = \{100, 110, 101, 111\}$
- B is the event that there is exactly one defect. In set notation: $B = \{100, 010, 001\}$
- C is the event that there are exactly two defects, so $C = \{110, 101, 011\}$
- $A \cap B =$
- $A^c =$
- $A^c \cup B =$
- $B \cap C =$

Venn Diagrams

Venn diagrams can be used to help us visualize unions, intersections, and complements.