Probability (Introduction)

Ryan Miller

Randomness

In our discussions of study design, *randomness* came up in two contexts:

- **Random sampling**
  - We used randomness to ensure every case in the population had an equal chance of being sampled
  - This prevented *sampling bias*, but we still have to worry about *sampling variability*

Randomness

In our discussions of study design, *randomness* came up in two contexts:

- **Random sampling**
  - We used randomness to ensure every case in the population had an equal chance of being sampled
  - This prevented *sampling bias*, but we still have to worry about *sampling variability*
- **Random assignment**
  - We used randomness to split our sample to into treatment and control groups
  - This protected us against *confounding variables*, but it introduces variability (you can view group assignment as a type of sampling)

We'll spend the remainder of the course learning ways to quantify the variability resulting from randomness, a task that requires us to study *probability*

Terminology

- A **trial** refers to a single instance of a random process
  - For example, sampling 1 individual from a population, or determining if 1 individual is assigned to the treatment or control group

Terminology

- A **trial** refers to a single instance of a random process
  - For example, sampling 1 individual from a population, or determining if 1 individual is assigned to the treatment or control group
- Every trial results in an **outcome**
  - For example, "Ryan Miller" is selected from the population of Xavier faculty, or Subject #1 is assigned to the control group
- The collection of *all possible outcomes* of a trial is called the **sample space**
  - For example, the sample space of selecting a Xavier faculty member would be a list of hundreds of names, while the sample space for assigning Subject #1 is $\{\text{Treatment}, \text{Control}\}$

Events

- Very often we are interested in **events**, which are *combinations of one or more observed outcomes*
  - For example, we might be interested in the event that at least 3 math faculty are sampled
  - Or the event that the 5 oldest subjects are assigned to the control group

Events

- Very often we are interested in **events**, which are *combinations of one or more observed outcomes*
  - For example, we might be interested in the event that at least 3 math faculty are sampled
  - Or the event that the 5 oldest subjects are assigned to the control group
- Recognize that using this definition, a single outcome is itself an event (an event is one, *or more* outcomes)

Probability

- Because these events (and outcomes they are based upon) involve randomness, they are inherently linked to *probability*, but what is probability?
  - That is, everyone agrees the probability of a fair coin landing "heads" is $1/2$, but why?

Probability

- Because these events (and outcomes they are based upon) involve randomness, they are inherently linked to *probability*, but what is probability?
  - That is, everyone agrees the probability of a fair coin landing "heads" is $1/2$, but why?
- **Frequentist** statisticians define probability as the *long-run proportion of an event occurring*
  - Thus, $P(\text{Heads}) = 0.5$ means that if we conducted many *trials* (different coin flips) we'd expect the *event* "Heads" to be observed in half of them

Empirical Probability

- Because probabilities are *long-run proportions*, we sometimes estimate them using proportions finite samples
  - For example, Joey Votto's career batting average is 0.305, so we might estimate he has a 30.5% of getting a hit during any given at-bat, or $P(\text{Hit}) = 0.305$
- This is called an *empirical probability*, it is different from a *theoretical probability* like $P(\text{Heads}) = 0.5$

Unions and Intersections

- As previously mentioned, an event is a combinations of one or more outcomes. There are two ways to combine multiple outcomes, *unions* and *intersections*

Unions and Intersections

- As previously mentioned, an event is a combinations of one or more outcomes. There are two ways to combine multiple outcomes, *unions* and *intersections*
- A **intersection** refers to two (or more) outcomes simultaneously occurring
  - Unions are expressed using the word "and" or the symbol $\cap$
  - Consider rolling a six-sided die, $P(\text{Five and Six}) = P(\text{Five} \cap \text{Six}) = 0$
  - Alternatively, $P(\text{Five and Odd Number}) = P(\text{Five} \cap \text{Odd Number}) = 1/6$

Unions and Intersections

- As previously mentioned, an event is a combinations of one or more outcomes. There are two ways to combine multiple outcomes, *unions* and *intersections*
- A **intersection** refers to two (or more) outcomes simultaneously occurring
  - Unions are expressed using the word "and" or the symbol $\cap$
  - Consider rolling a six-sided die, $P(\text{Five and Six}) = P(\text{Five} \cap \text{Six}) = 0$
  - Alternatively, $P(\text{Five and Odd Number}) = P(\text{Five} \cap \text{Odd Number}) = 1/6$
- An **union** refers to at least one outcome occurring
  - Intersections are expressed using the word "or" or the symbol $\cup$
  - Consider rolling a six-sided die, $P(\text{Five or Six}) = P(\text{Five} \cup \text{Six}) = 2/6 = 1/3$
  - Alternatively, $P(\text{Five or Odd Number}) = P(\text{Five} \cup \text{Odd Number}) = 3/6 = 1/2$

Conclusion

- Probability provides a framework for understanding randomness, something is necessary when our data involve sampling or random assignment (or both)

Conclusion

- Probability provides a framework for understanding randomness, something is necessary when our data involve sampling or random assignment (or both)
- A *trial* described an instance of a random process that resulted in an *outcome*
  - The collection of all possible outcomes was the *sample space*
- An *event* was a combination of one or more outcomes
  - Events can be expressed as *unions* or *intersections* of different outcomes