ST 516: Foundations of Data Analytics

What is Probability?

- What is Probability?
    - Definition of Probability

- Probability in Data Analysis
    - Assessing Probability

What is Probability?

We use or hear the term probability regularly:

- A local, daily weather report often gives the probability of rain
- You might talk with friends about the probability of a certain candidate winning in the next election
- A family member or friend might be offered probabilities of recovery under different therapeutic regimes

But what *is* probability, how is it used in statistics and data analysis, and how do we quantify it?

Definition of Probability

The **relative frequency** of an event is the number of times the event occurs in a set of trials.

- A trial is an action that could lead to the occurrence of an event (for example, the flip of a coin is a trial)
- An event is a possible outcome or a collection of possible outcomes that we observe from a trial (for example, a "head" in one flip of a coin)

The **probability** of an event is the relative frequency of the occurrence of that event in an infinite series of repeated trials.

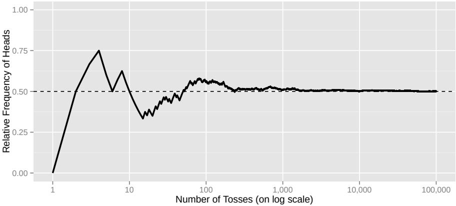
Definition of Probability

In the flip of a coin example, consider the following thought experiment:

1. In 10 flips of a fair coin, would you be surprised to see 7 heads?
2. In 100 flips of a fair coin, would you be surprised to see 70 heads?
3. In 1000 flips of a fair coin, would you be surprised to see 700 heads?

We can use a coin flip simulation in R to observe the behavior of the relative frequency of heads as the number of coin flips (trials) increases.

Definition of Probability



This plot demonstrates **The Law of Large Numbers**: As more trials (in this case, coin flips) occur, the relative proportion of occurrences of a certain event (in this case “heads”) converges to the probability of that event.

Probability in Data Analysis

But probability isn't just about coin tosses!

Statisticians and data analysts use probability to quantify the uncertainties that are inherent in using samples of data to make inferences to, or predictions about populations:

- You will learn how to express a range of plausible values for a quantity of interest, where that range depends on a probability associated with obtaining a sample from a population.

- You will learn how to express the probability of whether a particular event might simply be due to chance, rather than some other mechanism that we could possibly identify.

Assessing Probability

With the relative frequency notion of probability, there are two things you should consider when assessing or evaluating probability:

1. What is the series of repeated trials for which the probability makes sense?

2. Based on a reported probability, how likely is the reported event relative to other events about which you may have a better understanding?

Example

As an example, suppose that you are told that the probability of being killed by lightening in a given year in the United States is 0.0000002.

1. What are the repeated trials?

2. How large or small is 0.0000002? Should you be concerned?

We find it useful to imagine a box with tickets in it for this situation. The population of the US is roughly 400 million, so imagine that the box has 400 million tickets in it, where 80 of the tickets are red and the rest are blue. Now imagine repeatedly drawing a single ticket from the shuffled box of tickets, recording the color of the ticket and replacing it each time. The long-run relative frequency of getting a red ticket is the same as the probability of getting killed lightening.