GENERAL ASSEMBLY

PROBABILITY, INDEPENDENCE, & SAMPLING

Matt Brems

Data Science Immersive, GA DC

PROBABILITY, INDEPENDENCE, & SAMPLING

"Mathematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search for truth, must not cast a spell over him."

– California Supreme Court, *People v. Collins* (1968)

LEARNING OBJECTIVES

- Apply five probability rules.
- Define independence and understand its role in probability.
- Explain the relationship between probability and statistics.
- Describe and be able to implement simple random sampling, stratified random sampling, and cluster random sampling.

INTRODUCTION: DEFINITIONS & SETS

DEFINITIONS

- Experiment: A procedure that can be repeated infinitely many times and has a well-defined set of outcomes.

- Event: Any collection of outcomes of an experiment.

- Sample Space: The set of all possible outcomes of an experiment, denoted $\mathcal{S}$.

EXAMPLES

- Experiment: Flip a coin twice.
- Sample Space $\mathcal{S}$:
- Event:

- Experiment: Rolling a single die.
- Sample Space $\mathcal{S}$:
- Event:

DEFINITIONS

- Set: A well-defined collection of distinct objects.
  - $\{Derek\ Jeter, \pi, ☺\}$
    - (Standing on the shoulders of Justin Gash for this one.)

- Element: An object that is a member of a set.
  - Derek Jeter
  - π
  - ☺

SET OPERATIONS

- Union: $A \cup B = the\ set\ of\ elements\ in\ A\ or\ B$
- Intersection: $A \cap B = the\ set\ of\ elements\ in\ A\ and\ B$
- Example:
  - $A =$ even numbers between 1 and 10 $= \{2,4,6,8\}$
  - $B =$ prime numbers between 1 and 10$= \{2,3,5,7\}$
  - $A \cup B =$?
  - $A \cap B =$?

SET OPERATIONS

- Example:
  - $A = \{2,4,6,8\}$ & $B = \{2,3,5,7\}$
  - $A \cup B = \{2,4,6,8\} \cup \{2,3,5,7\} = \{2,3,4,5,6,7,8\}$
  - $A \cap B = \{2,4,6,8\} \cap \{2,3,5,7\} = \{2\}$

BASICS OF PROBABILITY

PROBABILITY BASICS

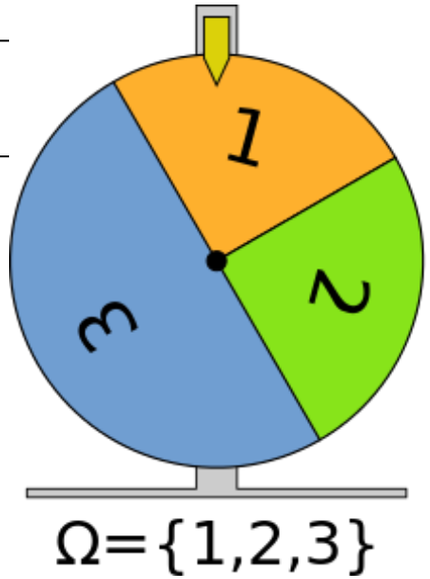
- Given an event A, we say that the probability that A occurs is:

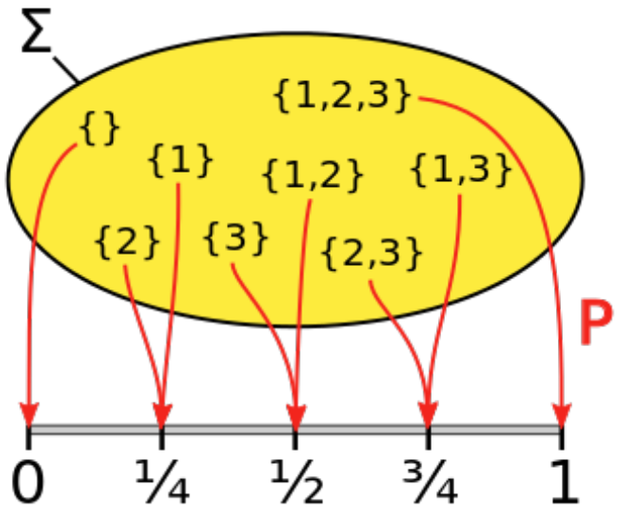
$$P(A) = \frac{\text{number of outcomes in A}}{\text{number of all possible outcomes}}$$

PROBABILITY BASICS

- Probability: $\mathrm{P}(\mathcal{S}, \mathcal{F}) \rightarrow [0,1]$
  - $\mathcal{S}$ is the sample space.
  - $\mathcal{F}$ is the "event space," or set of possible events.
  - P is the probability function, mapping each event to the $[0,1]$ interval.

PROBABILITY BASICS





- Probability: $\mathrm{P}(\mathcal{S}, \mathcal{F}) \rightarrow [0,1]$
  - $\mathcal{S}$ is the sample space.
  - $\mathcal{F}$ is the "event space," or set of possible events.
  - P is the probability function, mapping each event to the $[0,1]$ interval.

  - In more rigorous treatments of probability:
    - The sample space $\mathcal{S}$ is denoted by Ω.
    - The "event space" is denoted either by $\mathcal{F}$ or Σ, is called a "sigma algebra" or "Borel field," and has a set of very specific properties.

AXIOMS OF PROBABILITY (Kolmogorov Axioms)

- For any event A, $P(A) \geq 0$.
  - Nonnegativity.

- For the sample space $\mathcal{S}$, $P(\mathcal{S}) = 1$.
  - Unit measure.

- For mutually exclusive (or disjoint) E_i, $P\left(\bigcup_{\{i=1\}}^{\infty} E_i\right) = \sum_{\{i=1\}}^{\infty} P(E_i)$
  - Additivity.

- Probability must **ALWAYS** follow these three axioms.

PROBABILITY RULES

- $P(\emptyset) = 0$
  - Note: $\emptyset$ indicates the "empty set," or the event containing zero outcomes from the experiment.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Venn diagrams can help to illustrate this – but remember that Venn diagrams are not proofs!
  - If A and B are disjoint, then $P(A \cap B) = 0 \Rightarrow P(A \cup B) = P(A) + P(B)$.

- $P(A^C) = 1 - P(A)$
  - A^C is known as the "complement of A."

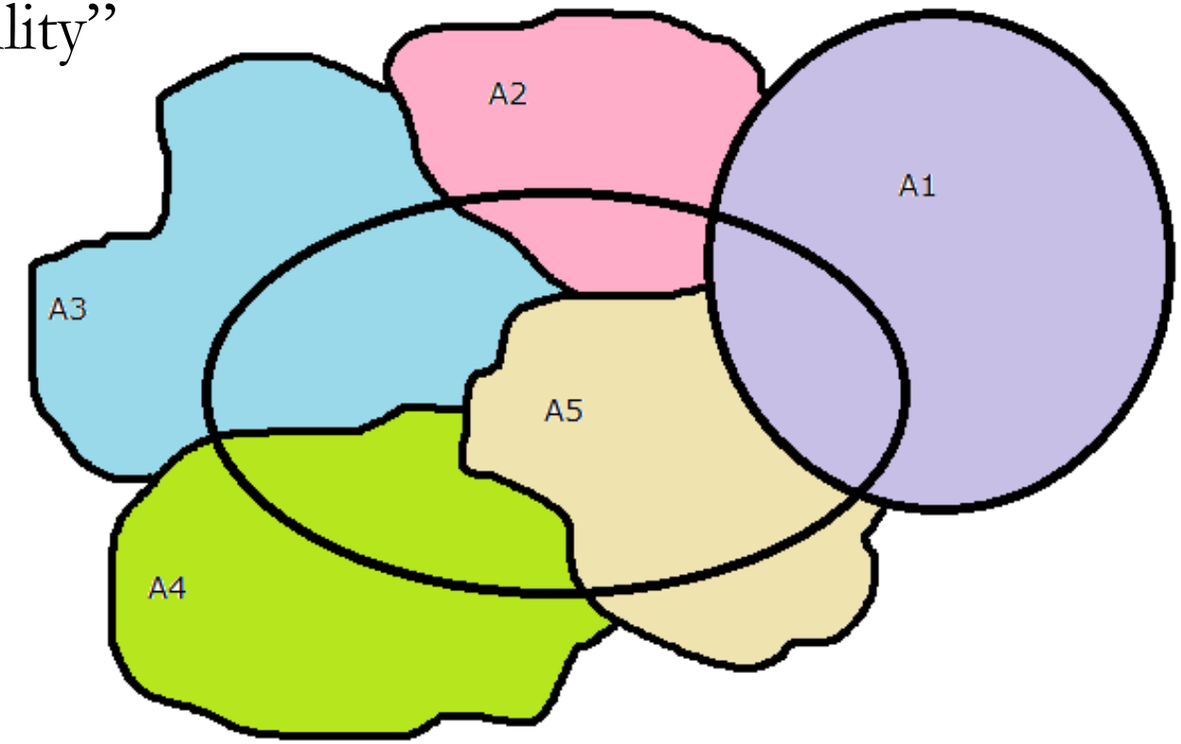
PROBABILITY RULES

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - Note: $A|B$ means "A given B" or "A conditional on the fact that B happens."
  - Example:
    - $A =$ roll a 2 $\Rightarrow P(A) = \frac{1}{6}$
    - $B =$ roll an even number $\Rightarrow P(B) = \frac{1}{2}$
    - $P(A \cap B) = P(\text{roll 2 and roll even number}) = \frac{1}{6}$
    - $P(A|B) =$ given that I roll an even, what is the probability of rolling a 2? $= \frac{1/6}{1/2} = \frac{1}{3}$

- $P(A \cap B) = P(A|B)P(B)$
  - We took the first rule on this slide, multiplied both sides of $P(B)$, and voila!
  - $P(A \cap B \cap C) = P(A|B, C)P(B|C)P(C)$

PROBABILITY RULES

- $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$
  - "Law of Total Probability"



PROBABILITY RULES – SUMMARY

- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A^C) = 1 - P(A)$
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A|B)P(B)$
  - $P(A \cap B \cap C) = P(A|B, C)P(B|C)P(C)$
- $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$

PROBABILITY RULES – PRACTICE

- A = {a U.S. birth results in twin females}
- B = {a U.S. birth results in identical twins}
- C = {a U.S. birth results in twins}

- In words, what does $P(A \cap C)$ mean?

- In words, what does $P(A \cap B \cap C)$ mean?

PROBABILITY RULES – PRACTICE

- A = {a U.S. birth results in twin females}
- B = {a U.S. birth results in identical twins}
- C = {a U.S. birth results in twins}

- A twin birth occurs approximately 1 in every 90 births.
- Roughly $\frac{1}{3}$ of all human twins are identical and $\frac{2}{3}$ are fraternal.
- Identical twins are necessarily the same sex and are male with probability 50%.
- Among fraternal twins, $\frac{1}{4}$ are both female, $\frac{1}{4}$ are both male.
- Find the values of $P(A \cap C)$ and $P(A \cap B \cap C)$.
  - Note any assumptions that you make along the way.

PROBABILITY RULES – PRACTICE

- Suppose the probability that an infant dies from sudden infant death syndrome (SIDS) is approximately 0.001%.

- What is the probability that a family has two children who die from SIDS?

PROBABILITY RULES – PRACTICE

- $P(A \cap B) = P(A|B)P(B)$

- What does this mean?

PROBABILITY RULES – PRACTICE

- $P(A \cap B) = P(A|B)P(B)$

- What does this mean?
  - This means that the probability of events A and B occurring is calculated by finding the probability of B, then finding the probability of A given that B already occurred, then multiplying those two.

PROBABILITY RULES – PRACTICE

- $P(A \cap B) = P(A|B)P(B)$

- What does this mean?
  - This means that the probability of events A and B occurring is calculated by finding the probability of B, then finding the probability of A given that B already occurred, then multiplying those two.
  - This gets us to the notion of dependence versus independence.

INDEPENDENCE

INDEPENDENCE

- Two events A and B are said to be independent if $P(A|B) = P(A)$.

INDEPENDENCE

- Two events A and B are said to be independent if $P(A|B) = P(A)$.
  - Intuitively, this means that the probability that A occurs is not affected by knowing whether or not B occurs.

INDEPENDENCE

- Independence is huge in statistics and machine learning – often we assume that our observations are independent.

INDEPENDENCE

- Independence is huge in statistics and machine learning – often we assume that our observations are independent.

- Making this assumption when it the assumption is obviously violated can have disastrous effects on our results.
  - In this case, we won't get an error message. It is up to us to keep an eye out for whether or not our assumptions are justified.
  - *People v. Collins*, Sally Clark, Lucia de Berk
  - Time series, spatial data, etc.

INDEPENDENCE

- Independence is huge in statistics and machine learning – often we assume that our observations are independent.

- Making this assumption when it the assumption is obviously violated can have disastrous effects on our results.
  - In this case, we won't get an error message. It is up to us to keep an eye out for whether or not our assumptions are justified.
  - *People v. Collins*, Sally Clark, Lucia de Berk
  - Time series, spatial data, etc.

- Deciding whether or not data are independent is, unfortunately, a judgment call that is contingent upon your specific use-case.

INDEPENDENCE

- Areas where independence is important:
  - Considering joint probabilities. (i.e. $P(A \cap B)$)
  - Most modeling techniques.
  - Sampling.
  - Training/testing sets. (More on this next week!)

WHEN BY HAND IS TOUGH…

- Oftentimes, we won't evaluate probabilities by hand.
  - It's still very important to understand the ideas behind probability – as we move forward, it's critical to:
    - a) know probability's relationship with statistics and machine learning.
    - b) identify potentially bad assumptions.

WHEN BY HAND IS TOUGH…

- Oftentimes, we won't evaluate probabilities by hand.
  - It's still very important to understand the ideas behind probability – as we move forward, it's critical to:
    - a) know probability's relationship with statistics and machine learning.
    - b) identify potentially bad assumptions.

- We can often use simulations to give us a good approximation of the true probability of some event.

PROBABILITY & STATISTICS

PROBABILITY AND STATISTICS

- Sometimes it is less convenient to work with one particular event and more convenient to describe all possible outcomes.

PROBABILITY AND STATISTICS

- Sometimes it is less convenient to work with one particular event and more convenient to describe all possible outcomes.

- For example, rather than finding the probability that someone has an IQ of exactly 100, we might be interested in looking at all possible IQ scores and how frequently we observe each IQ value.

PROBABILITY AND STATISTICS

- Sometimes it is less convenient to work with one particular event and more convenient to describe all possible outcomes.

- For example, rather than finding the probability that someone has an IQ of exactly 100, we might be interested in looking at all possible IQ scores and how frequently we observe each IQ value.

- Recall: a distribution is the set of all possible values of a variable and how frequently the variable takes on each value.

PROBABILITY AND STATISTICS

- Every variable has a distribution.
  - The probabilities associated with each distribution must add up to 1.
  - The probability of any variable's value can never be negative.

PROBABILITY AND STATISTICS

- Every variable has a distribution.
  - The probabilities associated with each distribution must add up to 1.
  - The probability of any variable's value can never be negative.

- Let X = IQ score.
  - We might say that X follows a Normal distribution with mean 100 and standard deviation 15.

PROBABILITY AND STATISTICS

- Every variable has a distribution.
  - The probabilities associated with each distribution must add up to 1.
  - The probability of any variable's value can never be negative.

- Let X = IQ score.
  - We might say that X follows a Normal distribution with mean 100 and standard deviation 15.

- Now let Y = time it takes all American workers to get to work.
  - What do we do here?

PROBABILITY AND STATISTICS

- This gets to the relationship between probability and statistics.

PROBABILITY AND STATISTICS

- This gets to the relationship between probability and statistics.
- In probability, we know the values of these parameters (measures of a population) and can thus completely define the probability distribution.
- In statistics, we don't know the values of these parameters, so we have to estimate them.

PROBABILITY AND STATISTICS

- This gets to the relationship between probability and statistics.
- In probability, we know the values of these parameters (measures of a population) and can thus completely define the probability distribution.
- In statistics, we don't know the values of these parameters, so we have to estimate them.
- We gather a sample to learn about the population.
- We calculate statistics to learn about parameters.

SAMPLING

SAMPLING

- Speaking broadly, in statistics we will take measurements on a sample to learn about the population. We must, however, be *very* careful about how we define this population, **as any inferences we make based on a sample are extended only to the sampled population**.

- **Sampled Population:** The population about which we *can* make inferences; the population from which we sample.

- **Target Population:** The population about which we want to make inferences; the population of interest.

SAMPLING

- My goal is to estimate the proportion of people in New Hampshire who plan to vote for each candidate in the Republican and Democratic Presidential primaries. I have a voter file containing the names, contact information, and demographic data of everyone registered to vote within the state of New Hampshire - roughly 900,000 people. I choose to call about 10,000 people. Of those called, about 6% responded the question "For whom do you plan to vote in the upcoming Presidential primary?"

- What is the **sampled population**?

- What is the **target population**?

SAMPLING

- **Simple Random Sample (SRS):** A sample is a simple random sample when every possible subset of n units in the population has the same chance of being the sample.

- **Stratified Random Sample:** A sample is a stratified random sample when the population is broken into subgroups (called strata), a simple random sample is pulled within each subgroup, and those simple random samples are combined into one larger "stratified" sample.

- **Cluster Random Sample:** A sample is a cluster sample when observation units are grouped into larger sampling units (called clusters), a sample of larger sampling units are selected, and then observation units within the selected larger sampling units are selected.

SAMPLING

- **Simple Random Sample (SRS):** A sample is a simple random sample when every possible subset of n units in the population has the same chance of being the sample.

- **Example:** I want to estimate the proportion of individuals in Baltimore who support a particular piece of legislation about gun control. I use a table of random digits to select 500 phone numbers from the Baltimore phone book.

SAMPLING

- **Stratified Random Sample:** A sample is a stratified random sample when the population is broken into subgroups (called strata), a simple random sample is pulled within each subgroup, and those simple random samples are combined into one larger "stratified" sample.

- **Example:** I want to estimate the average height of Ohio State undergraduates. I find that 52% of undergraduates are female and 48% are male. I get a list of all students, their sexes, and their email addresses. I gather a simple random sample of 480 men and a simple random sample of 520 women, then contact these 1,000 people via email.

SAMPLING

- **Cluster Random Sample:** A sample is a cluster sample when observation units are grouped into larger sampling units (called clusters), a sample of larger sampling units are selected, and then observation units within the selected larger sampling units are selected.

- **Example:** I want to estimate the effect of an SAT prep program in D.C. Optimally, I'd offer an SAT prep program at each of the 50 high schools in D.C. and compare SAT scores of those who did and didn't take the program within each school. This is expensive, though, so I take a random sample of 10 schools, offer the SAT prep program there, then compare the SAT scores at these 10 schools relative to the other 40.

SAMPLING

- Some points about sampling:
  - Simple random samples are the simplest with which we can work. Their formulas are straightforward. (The standard calculations for confidence intervals, etc. assume a simple random sample.)
  - Stratified random samples allow for more precise inference than other sampling techniques. (Confidence intervals are smaller.)
  - Cluster random samples are less precise than simple random samples, but allow us to be more cost-effective in cases where we logistically must do so.

REPLICATION

- Whether it be for an academic journal, needing to replicate results for your supervisor, or in-house testing among peers, it is often desirable to be able to replicate your work. When drawing a random sample, however, the computer relies on “randomness” and this is difficult to replicate.

- However, computers cannot do anything that is *truly* random. They can, however, generate pseudorandom numbers. Because of this, you can actually set what is called a "seed" so that you can return the computer to a particular state before generating a random number. Thus, if you set your seed to be 6, then generate a random sample, set your seed to be 6 again, and generate another random sample, you'll notice that the exact same random sample was generated!

RECAP

- Probability is a building block to learning about statistics.
  - Samples help us to learn about populations.
  - Statistics help us to learn about parameters.

- How we sample matters!

- Independence is a huge consideration that will depend on your use-case.

- Probability is complicated, but being familiar with the basics will go a long way in understanding how pieces fit together.