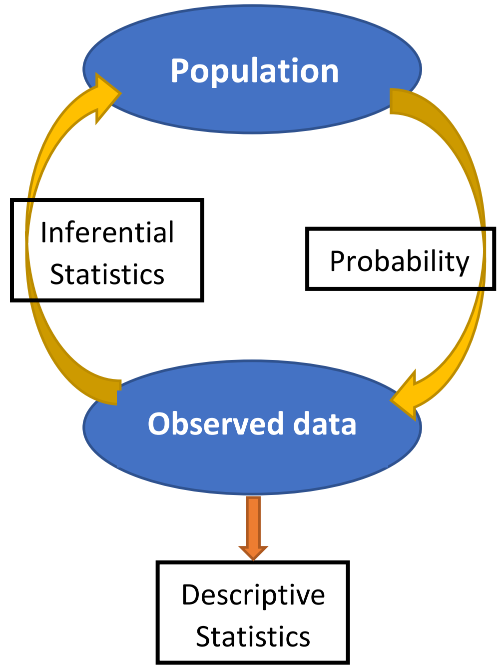
Chapter 4: Elementary Probability

Basic Terminology

- Probability vs. Statistics



- **Sample Space:** The Collection of all possible outcomes
  - Every variable we collect data from has a sample space.

Practice: Identify the sample space for each variable below.

a) We ask students whether or not they are Illinois residents or not. How might we classify the sample space of this

b) We roll a single, 6-sided die and record the result.

c) We measure people's heights in inches

- **Outcomes and Events**
  - **Outcomes** refer loosely to the individual possibilities in my sample space. If I narrow in on one trial, I might record which outcome from the sample space occurred. (e.g., Rolling a 1)
  - **Events** typically include a particular situation (one or more outcomes of interest) that we want to investigate or calculate the probability of (e.g., Rolling a number less than 3).
  - Events are typically notated with a letter: "Let A be the event that I roll a number less than 3."

- **Complement**
  - The complement of an event A includes every possible outcome that is not already included in Event A.
  - Notation: The complement of event A is A^C (may also be symbolized as A' too!)

- Summarizing the basics
  - We can represent probabilities as a value between 0 (0%) and 1 (100%).
  - The sum of the probabilities of all possible outcomes (Sample Space) equals 1 (100%)
  - We typically use the notation P(event) to represent the probability of that event.

Outcome	1	2	3	4	5	6	Total
Probability	1/6	1/6	1/6	1/6	1/6	1/6	1

Practice: For some event A, $P(A) + P(A^C) =$ ____

Simple Events

- **Equally likely outcomes** – *How many out of How many?*
  - ***If*** all possible outcomes are equally likely, then we can use a "how many out of how many" approach: $\frac{\text{\# of outcomes for event to occur}}{\text{Total \# of outcomes possible}}$
  - If all possible outcomes are **not** equally likely, you will need to find the probability for all possible outcomes in the event of interest.

Practice: What is the probability of rolling an odd number on a six-sided dice? *Report as a decimal for all.*

Practice: A can of mixed nuts has 4 types of mixed nuts: Almonds, Peanuts, Pecans, and Cashews. 50% of the nuts are peanuts, 25% are almonds, 15% are cashews, and 10% are pecans. A nut is chosen at random from this can.

What is the sample space for this random draw?

What is the probability that this nut is **not** an almond?

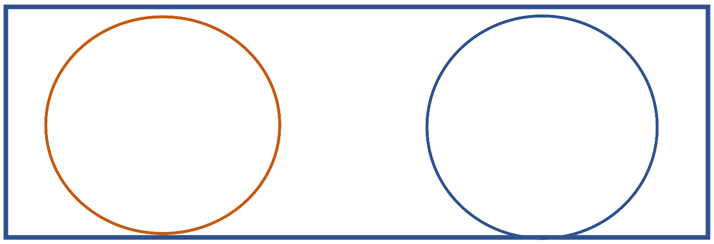
Compound Events

- A Compound Event involves calculating probabilities involving two (or more) events together.
- A first question to ask when assessing probabilities with two events is whether the events are disjoint.
  - **Disjoint** (may also be referred to as "*Mutually Exclusive*")**:** When both events _____________ be true at the same time!

    1): A randomly selected student is a Freshman

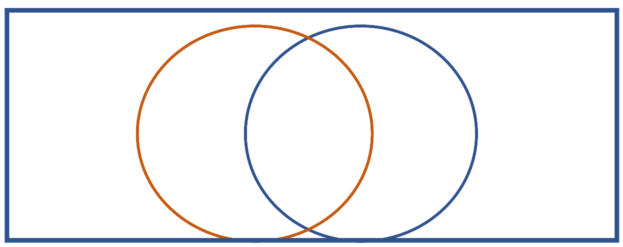
    2): A randomly selected student is applying for graduation

    - With two disjoint events, you can lump the sample space into three spaces: 1) The first event is true, 2) The second event is true, or 3) Neither one is true



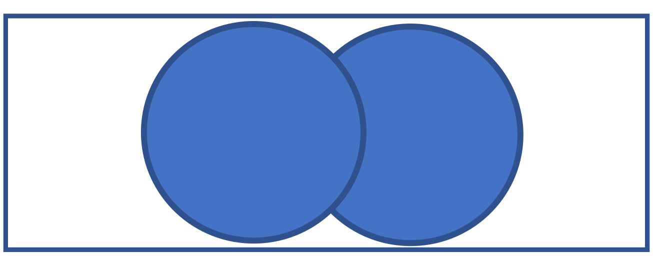
Disjoint Events

  - However, many events are instead **intersecting**.
  - For example: 1) A randomly selected student is a Freshman and 2) A randomly selected student owns a Car.

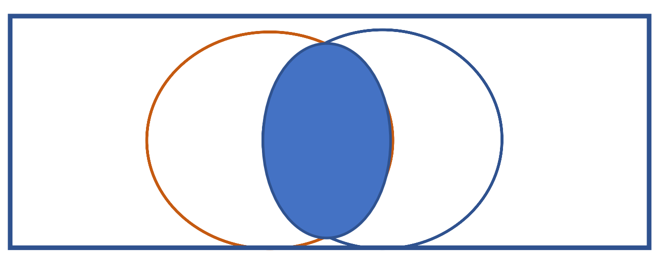


Intersecting Events

  - **Unions & Intersections**
    - The _____________ of two events is the collection of outcomes such that *at least one* of the events takes place. The entire area covered by the red and blue circles is part of the union of events A and B.
    - The _________________ of two events is the collection of outcomes such that *both* events take place. That would simply be the center sector.
    - When events are disjoint, then there is no intersection. But when they are not disjoint, we now have a fourth situation: 4) Both events are true!



Union



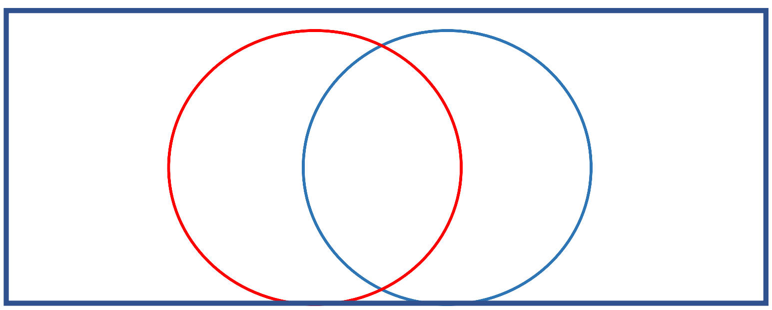
Intersection

Finding the Probability of a Union: P(A ∪ B)

- Quick terminology:
  - ∪ **stands for "union."** A ∪ B would be every possibility where at least one of these two events A and B are true.
  - ∩ **stands for "intersection,"** A ∩ B would be every possibility where BOTH A and B are true.

- **Addition Rule Formulas** (Use for Finding **Unions**)
  - **If two events happen to be disjoint, then their intersection would be nothing…**
  - If two events A and B are disjoint, then the P(A ∪ B) = P(A) + P(B)
  - **If two events are intersecting, be careful not to double count their intersection…**
  - If two events A and B are intersecting, P(A ∪ B) = P(A) + P(B) – P(A ∩ B)

Example: A doctor is trying to help a patient with the flu by prescribing a medication. This doctor knows that 15% of all patients experience fatigue and 10% of all patients experience stomach pain. This also includes 6% of patients who experience both of these side effects together.



First, label each circle as one of the two events of interest, and then fill in the Venn Diagram sectors above with percentages (fatigue only, stomach pain only, intersection, and neither)

The probability of *both* events taking place is _______% (the intersection)

The probability of at least one event taking place is _______% (the union)

The probability no fatigue is _______%

Practice: You're rolling a single die. Let A_1 be the event of getting a 1 or 2 on your first roll, and let A_2 be the event of getting a 4, 5, or 6 on your first roll. What is $P(A_1 \cup A_2)$?

Practice: Let's say that 80% of the population loves chocolate ice cream. Let's say that 60% love strawberry, and let's say that 10% love neither chocolate nor strawberry. First, create a Venn Diagram to help organize the information we have.

What is the probability that a randomly chosen person will love at least one of chocolate or strawberry?

What is the probability that a randomly chosen person loves both strawberry and chocolate?

Finding the Probability of an Intersection: $P(A \cap B)$

- **Independent vs. Dependent Events**
  - Two events are **independent** if the occurrence of one event has <u>no effect</u> on the probability of the other event. *<u>The probabilities will remain constant</u>*
    - **Example:** the event that I get a "tails" when I flip a coin doesn't change the likelihood of me getting heads or tails on the next coin flip.
  - Two events are **dependent** if the probability of one event changes/updates based on knowing the outcome of the other
    - **Example:** If drawing a marble from an urn "without replacement," then knowing that yellow was picked on draw 1 should decrease the probability of yellow on draw 2.
    - **Example:** 6.6% of people in the U.S. have Type O blood. Blood types are inherited, so if you know that your father has Type O, then that information increases the probability that you might have Type O compared to when you don't know anything.

Practice: For each example, identify whether you expect it to be independent or dependent, as well as disjoint or intersecting.

A = Parking in a metered space without paying. B = Getting a parking ticket.

A = You draw one card from a deck, and the card is red. B = You draw one card from a separate deck, and the card is red.

A = You take the bus to work. B = You walk the entire way to work

- **The Gambler's Fallacy**
  - Believing that the likelihood of a particular outcome increases if it hasn't happened for a while…"this slot machine is 'due' for a win" or "We've had several heads in a row, so it's more likely to be tails on the next toss" etc.
  - *In other words, it's **not** recognizing when trials are **independent**!*

Practice: A fair coin has one side heads and one side tails. I have flipped the coin twice and gotten tails twice in a row. I flip it a third time. Is the result of the next coin flip independent or dependent of the previous flips?

What is the probability that I get tails on the third flip?

More than 50% About 50% Less than 50%

Practice: You are sampling cards ***without replacement*** from a standard deck of cards. This deck has 13 hearts, 13 clubs, 13 spades, and 13 diamonds. You have drawn 2 hearts in a row. Is the suit of the next card you draw independent or dependent on the previous draws?

What is the probability that the 3rd card you draw will also be a heart?

More than 25% About 25% Less than 25%

FIRST Multiplication Rule Formula (Use for finding **Intersections** for **Independent** events)
- If two events A and B are independent, then the P(A ∩ B) = P(A) x P(B)
- Consider coin flips: Probability of Heads first, then Tails second is 0.5 x 0.5 = 0.25

Practice: A hospital is trying to gather more blood donations for type O negative, the "universal donor." Let's say we know from a past event that 20% of people we ask on the street are willing to give blood. We also know that approximately 6.7% of the population has O- blood. Let's *assume* willingness to give blood and type of blood are **independent** events.

What is the probability that a randomly selected individual will fulfill both of these events?

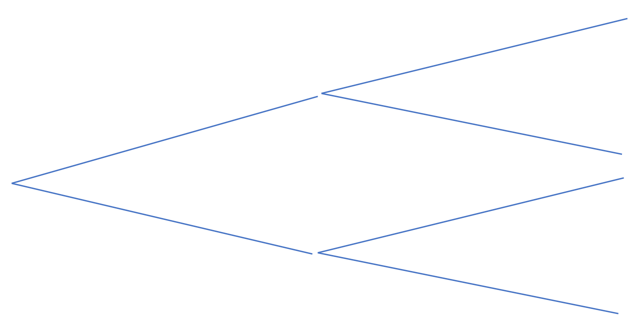
- **Conditional Probability**
  - Conditional probability represents the probability of one event *given* known information about another event.

Conditional Probability: The conditional probability of "B given A" is notated P(B|A).

We can represent this in a formula. P(B|A) = $\frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} = \frac{\text{(Probability of both A and B happening)}}{\text{(Probability of A happening)}}$

  - **Tree Diagrams** are a great way to visualize compound events when wanting to identify conditional probabilities.
    - Choose one event to be represented on the first stem layer (along with its complement as the other stem).
    - Choose the second event and its complement to be represented in the second layer of stems

Practice: Let's create a tree diagram to represent the probability of fatigue conditional on whether the patient experiences stomach pain.



Question: What would we see in a tree diagram if the two events were *independent*?

SECOND Multiplication Rule Formula (Use for finding **Intersections** from **Dependent** events)

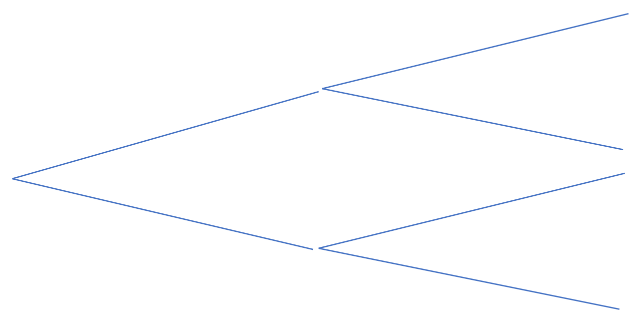
- If two events A and B are dependent, Pr(A ∩ B) = <u>Pr(A) x Pr(B|A)</u>

Practice: Let's say that the CDC estimates based on their disease modeling that approximately 4% of people in your community have COVID-19. You take an at-home COVID test with the following accuracy:

- 95% of people with COVID-19 will correctly test positive.
- 90% of people who don't have COVID-19 will correctly test negative.
- Let's assume there are no "inconslusive" results. Everyone gets a positive or negative result.

Organize your information into a tree diagram.

Are the events of having COVID-19 and testing positive independent or dependent events?



What's the probability that you don't have have COVID-19, but get a positive test result (a false positive)?

What's the probability that you get an incorrect test result? *(hint: there are two ways to get an incorrect test result. Each of these scenarios is disjoint!)*

We will come back to answer this: What is the sensitivity of this test?

We will come back to answer this: What is the negative predictive value of this test?

Theoretical vs. Empirical Probabilities

- So far, we have focused on calculating **theoretical probabilities**
  - Known (or assumed) probabilities associated with a population or random process
- When working with real data, we often estimate these measures using **empirical probabilities**
  - Estimated probabilities based on how many of each outcome we observed.

Practice: DNA testing has allowed some cases to be re-examined (many years after a court finding) to determine if there was evidence of a crime by the defendant. Consider this *fictional* data that shows results of 104 re-examined cases, along with the original finding by the court.

- 23 of the 25 truly innocent people were acquitted
- 17 of the 79 truly guilty people were acquitted

Fill in the contingency table, and then answer the following questions:

	Acquitted	**Convicted**	**Totals**
Truly Innocent			
Truly Guilty			
Totals			

What is the probability that a randomly selected case from this sample resulted in acquittal?

What is the probability that an acquitted person was truly guilty?

What is the probability that someone was convicted given that they were innocent?

Based on the data, would you say that acquittal status and innocence/guilt is independent or dependent?

Measuring the Effectiveness of a Test

- In Biostatistics, conditional probability is commonly used to assess the effectiveness of a test.
- Let's talk about four common measures to understand how effective a test is:
  - **Positive Predictive Value (PPV):** Percentage of positive test results that are correct
    - *Probability of a true result **given** the test result is "positive"*
  - **Negative Predictive Value (NPV):** Percentage of negative test results that are correct
    - *Probability of a true result **given** the test result is "negative"*
  - **Sensitivity:** Percentage of people who have the "condition" who correctly test positive
    - *Probability of a true result **given** the person has "condition of interest"*
  - **Specificity:** Percentage of people who don't have the "condition" who correctly test negative
    - *Probability of a true result **given** the person does not have the "condition of interest"*

To understand these measures, consider this question: 555 individuals completed drug tests, some of whom had or had not used drugs in the last 24 hours. The following table reveals how accurate those tests were.

	Positive Test Result (Test shows drug use)	**Negative Test Result** (Test shows no drug use)	**Totals**
Subject uses drugs	45 (True Positive)	5 (False Negative)	50
Subject does not use drugs	25 (False Positive)	480 (True Negative)	505
Totals	70	485	555

Practice: Let's estimate each of these measures for the drug-testing example based on the data collected.

What is the sensitivity of this test?

What is the specificity of this test?

What is the positive predictive value of this test?

What is the negative predictive value of this test?

- **Achieving Confidence with Inaccurate Tests**
  - In reality, every test will have weaknesses. For example, the PPV of a test will often be low in the context of low incidence conditions because there will be so many non-infected people taking the test—*there will be a lot of false positives!*
  - A common way to combat that is to order multiple tests.

If the patient is not infected, then their percent negatives should match the test's _________________

If the patient is infected, then their percent positives should match the test's _________________