On Probability, Chance, and Likelihood

Neil J. Hatfield
School of Mathematical Sciences,
University of Northern Colorado

Revised May 2019

This reading serves as an opportunity for you to continue to wrestle with the concept of probability that we discussed in class and see how this idea relates to the related concepts of chance, likelihood, and personal certainty. No student comes into an university-level Introductory Statistics course without some past experience with these concepts. Whether you discussed probability in a past class or have just played board games, you have experiences that you can draw upon as you do this reading.

1 The Measurement of Uncertainty

Throughout history, a wide variety of mathematicians, statisticians, and scientists have worked diligently towards the same goal: how do we measure uncertainty? As they wrestled with this question, they found themselves always coming back to four guiding questions:

- What do we mean by "uncertainty"?
- What are we looking at that has "uncertainty"?
- How are we going to measure "uncertainty"?
- What does a measure of "uncertainty" mean once we say that we have one?

Answering these questions has been a long process, and there has not been agreement over the answers to these questions. Even today, there is still some disagreement over how to answer these four questions; this has driven the development of the subfields of "Frequentist", "Bayesian", and "Subjective" approaches to Probability and Statistics. Let us turn our attention to each one of these guiding questions.

1.1 What do we mean by "uncertainty"?

This question as the least amount of disagreement. Simply stated, we think about "uncertainty" as being in a state of not knowing what *exactly* will happen. In each of our lives, we encounter many situations where we aren't sure what will happen. For example, we might ask ourselves questions such as "Will there be a pop quiz today in my Statistics class?", "Will a 6 come up when we roll this fair, standard die?", or "How often will I encounter a person who is at least 6 feet tall?" Each one of these questions involves "uncertainty", but there are some key differences in for each of these situations.

1.2 What are we looking at that has "uncertainty"?

One way in which the three preceding situations differ is in looking at what is the *thing* that has the attribute of "uncertainty". The most natural answer is "Us; we are the ones who are uncertain." However, we do not have to shoulder all of uncertainty. We can choose to move some of the uncertainty burden from us to what we're imagining going on in the situation. If we so choose, we can move most of the burden to what we imagine as being the stochastic process. Recall that we've been thinking/talking a stochastic process as an infinitely repeatable process that allows us to get data that is replicable (but not reproducible), has unfixed outcomes, a fuzzy rule, and the attribute of randomness. Now, keep in mind that the uncertainty is still part of *our* thinking, we're just attributing the uncertainty to our image of a stochastic process for the given context.

	Source of Uncertainty	
Situation	**Us**	**Stochastic Process**
Will there be a pop quiz today in my Statistics class?	Yes	Depends
Will a 6 come up when we roll this fair, standard die?	Yes	Yes; roll the die and record the outcomes
How often will I encounter a person who is at least 6 feet tall?	Yes	Yes; using a lottery to select a person and then measure his/her height

In the last two situations, we can come up with come up with a stochastic process that we can treat as a source of uncertainty. However, for the pop quiz situation, we have a more challenging time. If by "today" we mean one specific day, then there is no way that we can create a stochastic process as we cannot repeat that specific day. However, if we were to view "today" as being "any class day", we can start imagining a stochastic process. When we have a non-repeatable event, we are solely responsible for uncertainty and we enter into the realm of "Subjective" approaches. When we can imagine a stochastic process, we can share the burden of uncertainty and we enter the realms of "Frequentist" and "Bayesian" approaches.

1.3 How are we going to measure "uncertainty"? and What does a measure of "uncertainty" mean?

The last two questions are tightly intertwined and depend upon what we have imagined as the source(s) of uncertainty.

1.3.1 When We are the Only Source of Uncertainty

When we are the only source of uncertainty, we "measure" our degree of belief that something will or will not happen. We refer to this as "personal certainty" or "personal uncertainty". Often our degree of belief is qualitative; for example, you might be "really sure" or "pretty sure" that there will in fact be a pop quiz. Sometimes, you can put a percentage on your statement: "I'm 75% sure that we won't have a pop quiz today". Your degree of belief is entirely dependent upon your past experiences and thus, you and your best friend can come up with entirely different answers that contradict one another for the question of a pop quiz today.

Keep in mind that when we are the only source of uncertainty, our "measurements of uncertainty" become unique to each one of us. There is not a clear measurement process like what we would see if we were measuring a person's height, the openness of an angle, or the speed of a car. This lack of a reproducible measurement process tends to make people uneasy, especially when we move away from questions about pop quizzes and towards questions of whether a particular cancer treatment saves lives.

1.3.2 When We Share the Uncertainty

When we share the burden of uncertainty between ourselves and our image of the stochastic process, we start laying the foundation for measuring uncertainty in a way that other individuals can reproduce. The key here is that we convey how we have envisioned the stochastic process as clearly as possible to other people so that everyone has a consistent image of the process. To measure uncertainty here, we will need to decide on a particular event of interest (e.g., rolling a six, encountering someone at least 6 feet tall) and then let the stochastic process repeat forever. As we let the stochastic process repeat indefinitely, we keep track of all of the outcomes we get and attempt to count how many times we see our chosen event of interest.

We quickly run into two problems as we attempt to count (i.e., get the absolute frequency): the stochastic process gives outcomes in a haphazard fashion and the process is constantly running. While we can overcome the haphazard nature with some patience, the fact that the absolute frequency of our chosen event will constantly change is harder to overcome. We could move away from absolute frequency and use relative frequency instead. Recall that

$$\text{Rel. Freq.} = \frac{\text{Abs. Freq.}}{\text{Total \# of Trials}}$$

This approach will allow us to account for the many repetitions of the stochastic process. However, with the stochastic process still running, the total number of trials is constantly increasing. This means that the relative frequency for our chosen event is changing with every new iteration/trial of the stochastic process. This was one of the biggest challenges to measuring uncertainty.

However, we can overcome this challenge by extending the notion of relative frequency so that we are no longer bound to the denominator of the total number of

trials. We refer to this extended concept as "long-run relative frequency". We interpret long-run relative frequencies as telling us what percentage of the infinitely many trials of the stochastic process we anticipate having outcomes inline with our event of interest. Keep in mind that the only way that we can use long-run relative frequency is to be imagining that a stochastic process is being repeated an infinite number of times (i.e., constantly running).

We have special names for the long-run relative frequency for different types of events. If we want to talk about the long-run relative frequency of observing particular outcomes (e.g., rolling a six, seeing someone over 6 feet tall), we refer to these long-run relative frequencies as the **probability of a data event**. If we want to talk about the long-run relative frequency of whether some assumptions are true, we speak of the **likelihood of our assumptions**.

For both probability and likelihood, we have standard measurement processes that different people can carry out and get compatible answers—provided everyone has consistent image of the stochastic process. While the mathematical machinery behind these measurement processes is beyond this course, we can make use of technology to get values out of the mathematical machinery. For probability, we'll use distribution functions; for likelihood, we'd use likelihood functions[1].

The following two sections delve into the distinctions between probability, chance, and likelihood.

2 Probability vs. Chance

Something that many individuals have trouble with is the fact that probability and chance are two separate but intimately related ideas. The relationship between probability and chance is like that of the relationship between rectangles and squares. Just as all squares are rectangles, all chance values are probability values, but the reverse is not true. Not all rectangles are squares and not all probability values are chance values. Chance is a much more exclusive concept than probability.

For a probability value to be called a "chance value" there are two conditions that must be met:

1. The underlying stochastic process must have a finite and known number of simple, unique outcomes.

2. The underlying stochastic process must be "fair", that is, each simple outcome must have the same long-run relative frequency (the Principle of Ignorance)

[1]We will not discuss likelihood beyond making a distinction between likelihood and probability in this course.

The Discrete Uniform distribution (in Shorthand: $\mathcal{DU}$) describes the long-run behavior of chance models/processes and therefore yields chance values. For this course, only the Discrete Uniform distribution can be used with the word "chance". Thus, any time you see the word "chance" used, you can assume that the Discrete Uniform distribution is at play.

3 Probability and Chance vs. Likelihood

Just as individuals struggle to separate probability and chance, they also struggle to separate these two ideas from a third related idea: likelihood. If you are struggling with this separation, know that you're not alone. Confounding these three ideas is something that has been happening since the 1700s, and has only exacerbated in recent years by an explosion of introductory textbooks that do not attend to such distinctions. Confounding these three ideas leads to a Circular Meaning for each (i.e., probability means chance, chance means likelihood, likelihood means probability) that students (and instructors) struggle to escape from.

To help you make the distinction between probability and likelihood, we'll use another analogy. Consider a divided highway or interstate; you have two sets of traffic lanes that traverse the same terrain (typically) but the two sets are different. One set of lanes are for traffic all headed in one direction and the other set are headed in the opposite direction. The same is true for probability (and chance) and likelihood; they are both covering the same phenomenon but are looking in opposite directions[2]. Probability makes use of our assumptions along with the infinitely running stochastic process to provide us with a measure of the long-run relative frequency of a particular data event. Likelihood starts from the data that we have collected (in other words, lots of data events) and provides a measurement of the long-run relative frequency of our assumptions being true.

4 Interpreting Values

To help you make the distinction between probability, chance, and likelihood, you should routinely practice interpreting values of each. To help with this, here are three example interpretation sentences:

- **Probability:** A probability value of 0.72 for a given data event means that if we repeat the stochastic process indefinitely, we would see the given data event occur 72% of the time.
- **Chance:** A chance value of 0.2 for a given data event means that if we repeat the stochastic process with the assumption of fairness indefinitely, we will see the given data event occur 20% of the time.
- **Likelihood:** A likelihood value of .65 for a set of assumptions means that if we repeat the stochastic process indefinitely, the data yielded will be consistent with our assumptions (i.e., our assumptions will be true) 65% of the time.

[2] A second analogy is that they are the two sides of the same coin.

All three interpretations hinge upon the idea that we're repeating the stochastic process indefinitely. The major difference between the interpretation for probability and the interpretation for chance is the phrase "with the assumption of fairness". (This reflects that chance is a more restrictive type of probability.) The reversal of perspective between probability and likelihood appears with our focus on the assumptions in the interpretation of likelihood values.

To adapt these statements for any situation is fairly easy:

1. Pick which of the situations (probability, chance, or likelihood) you're currently dealing with,
2. Replace the given value (0.72, 0.2, 0.65) with your actual value (convert to a percent for the second usage),
3. Replace "given data event" (or "a set of assumptions") with the actual data event you're focused on (for likelihood, you can either list out the assumptions– which are typically parameter values, or you could just write "our assumptions" provided you've listed them previously).

You don't have to get any fancier than this when writing. To practice, feel free to make up your own contexts, events, and values for you to use; don't worry about whether or not the values are the correct ones.
Here are a few practice situations:

- The probability of selecting a US adult man who is under 6 ft tall is 2/3.

- The chances of flipping a fair coin twice and getting both heads is 0.25.

- The likelihood of our process being fair is 0.01.

The following sections of the reading are going to focus on conceptualizing data events, working with probability notation, and discussing the formal probability rules.

5 Data Events

Recall that we find probability values through distribution functions and each probability value is a measure of the long-run relative frequency some data event occurring when we imagine the stochastic process running forever. If we know that the probability of observing a US adult man who is 6 ft tall or taller is 0.33, then we know that 33% of the time we carry out a stochastic process to select a US adult man and measure his height, we'll select a man whose height is at least 6 feet. In this example, we're examining the attribute **height** of US adult men, *with a particular/special interest* in heights that are 6 feet or more. The set of all heights that of particular/special interest (i.e., 6+ feet) is what we refer to as a **data event** (or more simply, an event).

There are several ways that we can define a data event. I will discuss two of them: the Classical Approach and the Data Event Approach.

5.1 Classical Approach

The Classical Approach to defining a data event is to use set notation. This approach is what most students encounter in school mathematics (elementary, middle, and high school) and even in many university-level Statistics courses. The general idea of this approach is for a person to first imagine the set of all possible outcomes of the stochastic process and then make a subset of just the outcomes of interest. The set of all possible outcomes is referred to as the **Sample Space** (denoted as $\mathcal{S}$) and the subset is the event. By convention, we denote events with capital letters, starting with A. Each event will be a subset of the sample space (mathematically, $A \subseteq \mathcal{S}$).

Let's consider some examples:

- Coin Flip: Given a stochastic process of flipping a coin with two different sides and recording the side that lands up, we denote the sample space as $\mathcal{S} = \{H, T\}$. If we're interested in Tails, we could define the event B as $B = \{T\}$.
- Blood Type: Given a stochastic process of selecting a person in the world and classifying/recording his/her blood type, we denote the sample space as $\mathcal{S} = \{A+, A-, B+, B-, AB+, AB-, O+, O-\}$. If we're interested in whether a person is the universal red cell donor or the universal plasma donor, we'll define the event C as $C = \{O-, AB+, AB-\}$.
- 6-Sided Die[3]: Given a stochastic process of rolling a standard, 6-sided die and recording the number of pips showing on the top face when the die stops, we

[3]If you want to get technical, we should define the sample space for the 6-Sided Die example as $\mathcal{S} = \{y \in \mathbb{Z} | 1 \leq y \leq 6\}$ and the event as $A = \{y \in \mathcal{S} | 0 \equiv y \mod 2\}$. However, since there are only six elements, we just list all six. Suppose that we instead rolled a 34-sided die, a 100-sided die, or even a 120-sided die. Rather than writing out all of the appropriate integers using this more formal notation speeds up our writing.

denote the sample space as $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$. If we're interested in getting an even number of pips, then we'll define the event A as $A = \{2, 4, 6\}$.

- Height of Men: Given a stochastic process of selecting a US adult man and then measuring/recording his height (in inches), we denote the sample space as $\mathcal{S} = \{x \in \mathbb{R} | 21.51 \leq x \leq 107.1\}$. If we want to consider heights that are at least 6 feet (72 inches), then we define the event D as $D = \{x \in \mathcal{S} | x \geq 72\}$.

For the first three examples, the sample space is fairly easy to imagine and write down. Writing out the event is just as easy. This is partly due to the fact that each of these examples has a **finite number of outcomes**; the stochastic process yields a fixed number of unique values. In the case of Coin Flip and Blood Type, we can use characters rather than numbers. However, in the last example (Height of Men), the set notation becomes more involved. This is due to the fact that we're dealing with an attribute that has **infinitely many values**; the stochastic process can yield infinitely many unique outcomes. You'll notice that there was an x that suddenly appeared when I defined the sample space and the event. When dealing with numeric data, we have to first specify which space of numbers we're working in, which requires us to use a variable. Thus, the $x \in \mathbb{R}$ tells us that we're dealing with Real-valued numbers (continuous for our purposes). However, the Reals include all positive numbers, zero, and all negative numbers; zero and negative numbers do not make much sense for heights. Thus, in our definition of the sample space, we add a condition. The condition is an expression that appears to the right of a pipe (|; which you can read as "given that..."). In the case of heights, I used the condition of the heights of the smallest man and the tallest man ever recorded. When we define our event, we must do something similar, but instead of saying that we're looking at all Reals, we'll specify that we're looking at the defined sample space ($x \in \mathcal{S}$).

However, defining events like the Height of Men example is rarely (if ever) covered in upper division (second, third, etc.) university courses let alone first-courses. For the most part, authors and instructors introduce the Classical approach and never mention how to apply such an approach to non-finite, non-discrete cases. They either sweep the issue under the rug or expect students to figure out what to do themselves.

The Classical approach is a valid approach, but can be confusing, daunting, and messy to use. To properly use this approach, an individual needs to not only know set notation, but be fairly fluent in the use of set notation. Further, this approach requires the spontaneous use of stochastic variables without formally defining the stochastic variable. This can be a tripping point. The Classical approach also invites people to internalize the Principle of Ignorance and treat all probability as chance, even when there is clear evidence that chance does not apply. You might have thought about the Coin Flip and 6-sided die examples as being about a fair coin and fair die. However, there is no change to either example whether we use a fair coin/die or an unfair coin/die.

5.2 The Data Event Approach

The Data Event approach to defining a data event is to use a properly defined stochastic variable and at least one relation (or logical operation) to describe what outcomes we're interested in. In this approach we leave the idea of the Sample Space behind and instead use the domain of the stochastic variable (either explicitly or implicitly stated). Further, the Data Event approach uses the same set of intuitive tools for any situation or context. To demonstrate this approach, we'll look at the same examples as before:

- Coin Flip: Let Y represent the face of a two-sided coin that lands up when we flip the coin. The domain of Y is $\{Heads, Tails\}$ (alternatively, the domain of Y could be $\{1, 2\}$ where 1 is a Heads and 2 is Tails). If we're interested in Tails, we could define the event $Y = Tails$.

- Blood Type: Let X represent the blood type of an individual. If we're interested in whether a person is the universal red cell donor or the universal plasma donor, we'll define the event as
  $X = \text{O–} \cup X = \text{AB+} \cup X = \text{AB–}$. (The $\cup$ symbol means OR.)

- 6-Sided Die: Let Z represent the number of pips showing on the top face of a standard, 6-sided die when rolled. If we're interested in getting an even number of pips, we'll define the event
  $Z = 2 \cup Z = 4 \cup Z = 6$.

- Height of Men: Let W represent the height (inches) of a US adult man. If we want to consider heights that are at least 6 feet (72 inches), then we define the event as $W \geq 72$.

For each one of these examples we did the exact same thing:

1. We defined a stochastic variable,
2. We either implicitly or explicitly gave the domain, based on the context,
3. We stated what we were interested in as a sentence,
4. We converted that statement by writing a mathematical expression that used our stochastic variable and a relation.

We used this approach regardless of data type (character vs. numeric), we kept focused on the outcomes of interest (i.e., the event), and our notation did not suddenly become more complex/different looking for a new situation. In the Classical approach, we have to split our focus between the sample space, the event, and the name of the event (the letter assigned to the subset). In the Event approach, we let the definition of the stochastic variable handle the sample space (i.e., the variable's domain) and rather than introducing a new symbol to keep track of, we jump straight to writing out the event.

There are six basic types of events. For each of these we'll treat X as a properly defined stochastic variable and x as a specific realization (i.e., observed value) of that variable.

- Applicable to all stochastic variables, regardless of data and modeling types:

    1. Equality: the event that the stochastic variable takes on a specific value; denoted as $X = x$.
    2. Not Equal: the event that the stochastic variable is any value BUT the specified value; denoted as $X \neq x$.

- Applicable to all stochastic variables that are at least ordinal modeling type:

    3. Strictly Less Than: the event that the stochastic variable takes on a value that is less than/smaller than/of lower order than a specific value; denoted as $X < x$.
    4. Strictly Greater Than: the event that the stochastic variable takes on a value that is greater than/larger than/of higher order than a specific value; denoted as $X > x$.
    5. Less Than or Equal to: the event that the stochastic variable takes on a value that is less than/smaller than/of lower order than a specific value OR is equal to that value; denoted as $X \leq x$.
    6. Greater Than or Equal to: the event that the stochastic variable takes on a value that is greater than/larger than/of higher order than a specific value OR is equal to that value; denoted as $X \geq x$.

Strictly speaking, only the Equality Event is a "simple" event; all of the other events are actually examples of "compound" events, but students are so familiar with them that we can treat them as being simple events. A **compound event** is the combination of several simple events and/or operations on simple events. In the examples that follow, the $\equiv$ can be read as "equivalent to" and indicates that the expression on the left side has the same effect as the expression on the right.

- Negation/Logical NOT: the event of looking at every outcome but those specified by stated event; can be denoted as
    - $\neg (X = x) \equiv X \neq x$
    - $\neg (X \neq x) \equiv X = x$
    - $\neg (X < x) \equiv X \not< x \equiv X \geq x$
    - $\neg (X > x) \equiv X \not> x \equiv X \leq x$
    - $\neg (X \leq x) \equiv X \not\leq x \equiv X > x$
    - $\neg (X \geq x) \equiv X \not\geq x \equiv X < x$
- Complement: another name for Negation; can be denoted as $(X = x)^C$ or as $\overline{X = x}$.
- Union/Logical OR: the event of looking at all outcomes that match any of the stated events; denoted as $X = x_1 \cup X \geq x_2$
- Intersection/Logical AND: the event of looking at all outcomes that satisfy ALL stated events; denoted as $X \geq x_1 \cap X \leq x_2$

For the remainder of this document (and in the course), I'm going to use the Data Event approach to writing data events[4].

6 Probability Notation

There is a fundamental truth about probability that is sadly glossed over by every textbook currently on the market and by nearly every instructor of introductory statistics:

All probabilities are conditional probabilities.

This is to say that any probability value that we reason about/with, calculate, etc. depends entirely on the assumptions that we make about the stochastic process and our approach to arrive at that value. Probability values come from cumulative probability which in turn comes from distribution functions. In order to invoke a distribution function we must make assumptions about the long-run behavior of the stochastic variable (i.e., the distribution). We must make assumptions regarding aspects of the stochastic variable's domain, the data and modeling types, which named distribution might apply, and the values of any necessary parameters.

Probability Notation is a set of conventions that we've adopted to represent a probability value and communicate to other people what data event we're speaking about as well as all of the assumptions that we might be making. Each instance of Probability Notation represents a probability value, but does not provide a way to calculate this value; calculation of the value requires us to do more work[5].

The basic format of Probability Notation is

$$P\left[\text{data event}|\text{listing of assumptions}\right]$$

where we place any basic or compound data event to the left of the pipe ("|") and to the right, we list out all of our assumptions. Let's return to our Height of Men example. We will assume that the attribute is numeric and continuous, that the stochastic variable's long-run behavior is that of a normal (Gaussian) distribution with an Expected Value of 63.7 inches and a Variance of 35.5 inches-squared. We'll remain interested in the event of heights of at least 6 feet ($W \geq 72$). To represent the probability of this data event we would write

$$P\left[W \geq 72 \middle| \begin{array}{c}\text{the attribute is numeric and continuous,}\\ \text{normal (Gaussian) distribution,}\\ \text{Expected Value of 63.7 inches,}\\ \text{Variance of 35.5 inches-squared}\end{array}\right]$$

Notice that the listing of assumptions takes up a lot of space. This is where Distribution Shorthand becomes very helpful. Remember that Distribution Shorthand is

[4]All of the basic events and compound events still apply under the Classical approach; the notation is similar but the event notation (e.g., $X = x$) is replaced with the name of the defined subset (e.g., A). Thus, $\neg(X = x)$ becomes $\neg A$.

[5]This is similar to how a formula represents a value but requires us to do all of the calculations while function notation takes care of all of the calculations for us.

a way of writing all of our assumptions in a very concise way. If we use Distribution Shorthand, we can re-write the Probability Notation for our height example as

$$P\left[W \geq 72 | W \sim \mathcal{N}(63.7, 35.5)\right]$$

which is much easier (and shorter) to write.

7 Formal Rules

When dealing with probability there are certain rules that we've established over history. Most of the rules are intuitive and you have worked with these rules in the past, you just may not have realized that you were. Many of the rules developed through the process of quantification; i.e., trying to decide 1) what we're measuring, 2) how we go about getting a measure, and 3) what a measure means once we get one. While these are often referred to as "Probability Rules" they equally apply to Chance and Likelihood.

7.1 Bounds on Values

No matter what event you're dealing with, the probability value will always be in the closed interval $[0, 1]$ (i.e., between zero and one, inclusive). Recall that probability is the long-run relative frequency of seeing the event. That is to say that probability measures the accumulation of outcomes of interest when we repeat a stochastic process indefinitely. If we haven't accumulated any of outcomes of interest, the probability value is zero (0% of of the infinitely many outcomes). If all of the accumulated outcomes are what we're interested in, then the probability is 1 (100% of the infinitely many outcomes). Thus, the smallest possible value of probability is zero and the largest is 1.

7.2 Total Probability

If our event is the entire domain of our stochastic variable, that is, we're interested in every possible outcome, then the probability should be 1. (In the Classical approach this would be the same as setting the event to be the sample space.) Since we're looking at the long-run relative frequency of *every possible outcome* combined, then we should accumulate everything, which is 1 (100% accumulation of all outcomes as we imagine the stochastic process running forever).

7.3 Probability Value of Zero

A probability value that is zero can be tricky. For continuous stochastic variables, the probability of being exactly any one value is zero, but we can observe outcomes that are that exact value. For example, the probability of selecting an adult living in the USA whose height is exactly 71" is zero. However, I happen to be exactly 71" tall. Thus, we need to avoid interpreting probability values of zero as "we never see this value". Keep in mind that "improbable" and "impossible" are not synonyms.

7.4 Probability of Inadmissible/Impossible Values

A common practice related to the Probability Value of Zero is to say that values that cannot happen have a probability value of zero. However, this confuses the issue at hand. Suppose that we're dealing with our 6-Sided Die example and we're interested in the event that Z = blue. Hopefully when you read that event you thought "What?!" On a standard, 6-sided die there is no "blue" number of pips. This is an example of an **inadmissible event** or **impossible event**; an event that does not make sense in the given context and is beyond the domain of the stochastic variable. This is an event that cannot happen in the context of rolling a standard, 6-sided die. Rather than saying that the probability value here is zero, we'll say that the probability "Does Not Exist (DNE)". If we were to say the probability value is zero, then we're essentially saying that "blue" is part of the domain of our stochastic variable. By saying that the probability value does not exist we raise attention to the fact that we have an inadmissible/impossible data event.

7.5 Disjoint (Mutually Exclusive) Events

Suppose that we have two data events that we're interested in; using our Height example $W \leq 62$ and $W \geq 75$. When there are no outcomes that are common to both events, we say that the events are **disjoint** or **mutually exclusive** from each other. One way in which we can decide whether we're dealing with disjoint events is to try to imagine an object/living being who simultaneously fits both data events. Using our Men's Height example, can we imagine a single man who is both 62" tall or shorter AND is over 75" tall? No, each man only has one height; he is either 62" or shorter, 75" or taller, or not part of either event.

Let us now think of our Blood Type example and the events of being Rh positive as well as being Type B. Can we imagine a person who is simultaneously both Rh positive and Type B? Yes, a person can have the blood type B+ and thus fulfills both data events. This would mean that being Rh positive and having Type B are not mutually exclusive events.

7.6 Addition Rule

If we're interested in the Union of two events, we can add the probability values for the individual events, in the following ways:

- Disjoint (Mutually Exclusive) Events

$$P\left[W \leq 62 \cup W \geq 75 | W \sim \mathcal{N}(63.7, 35.5)\right]$$

$$= P\left[W \leq 62 | W \sim \mathcal{N}(63.7, 35.5)\right] + P\left[W \geq 75 | W \sim \mathcal{N}(63.7, 35.5)\right]$$

- Non-disjoint Events (General Rule-works in all cases)

$$P\left[Z \text{ is prime} \cup Z \text{ is even} | Z \sim \mathcal{DU}(6)\right]$$

$$= P\left[Z \text{ is prime} | Z \sim \mathcal{DU}(6)\right] + P\left[Z \text{ is even} | Z \sim \mathcal{DU}(6)\right] - P\left[Z \text{ is prime AND even} | Z \sim \mathcal{DU}(6)\right]$$

Notice that for the non-disjoint events we have to subtract of the probability of the Intersection event. This accounts for the double counting of the underlying objects/living beings that happens with the addition of the two individual events. In the 6-sided die case, the value 2 is both a prime and an even; thus all occurrences of 2 are included in the probability value for primes as well as in the probability value for events. We've counted all occurrences of 2 twice. By subtracting out the probability of the Intersection, we correct our over-counting.

7.7 Complement Rule

All data events have a complement (logical negation). Any data event and the complement of that data event must be disjoint. Together, they must account for all possible outcomes. Thus, the union of a data event and that event's complement results in a probability value of 1. This allows us to use the Total Probability Rule, the idea of Disjoint Events, and the Addition Rule to establish that

$$P\left[(W \leq 62)^C \,|W \sim \mathcal{N}(63.7, 35.5)\right] = 1 - P\left[W \leq 62 | W \sim \mathcal{N}(63.7, 35.5)\right]$$

In general,

$$P\left[\text{complement of a data event}|\text{assumptions}\right] = 1 - P\left[\text{the data event}|\text{assumptions}\right]$$

7.8 Independent Events

Two data events are independent if the occurrence (or lack thereof) one event has no impact on the occurrence (or lack thereof) the second event. Examples include successive rolls of a standard, 6-sided die and flipping a two-sided coin. Events that are not independent are called "dependent events".

7.9 Conditional Events

When we're interested in a data event and we know something about another data event, we say that we're looking at a **conditional event**[6]; that is, we're only interested in the first data event when we already know the second event happened. For instance, we might be interested in the rolling an even number of pips on a standard, 6-sided die when we already know that we've gotten Tails on the a flip of the two-sided coin. Here the first event is the die roll resulting in an even, while the second event is that we've gotten Tails on the coin flip. We express conditional events as

$$P\left[\text{first data event}|\text{second data event}\middle|\text{assumptions}\right]$$

In the example given, we'd write

$$P\left[Z \text{ is even}|Y = Tails\middle|Z \sim \mathcal{DU}(6), Y \sim \mathcal{DU}(2)\right]$$

[6]Most textbooks call this "conditional probability" but in actuality they are talking about conditional events; plus, all probabilities are conditional.

Notice that we use two pipes (||); the meaning for both symbols is the same (i.e., "given that..."). Most of the time when people talk about "conditional probability" they mean that they are interested in the probability of a conditional event.

We can find the value of a conditional event in the following way:

$$P\left[\text{first data event}|\text{second data event}\middle|\text{assumptions}\right] =$$

$$\frac{P\left[\text{first data event} \cap \text{second data event}\middle|\text{assumptions}\right]}{P\left[\text{second data event}\middle|\text{assumptions}\right]}$$

For our example,

$$P\left[Z \text{ is even}|Y = Tails\middle|Z \sim \mathcal{DU}(6), Y \sim \mathcal{DU}(2)\right] =$$

$$\frac{P\left[Z \text{ is even} \cap Y = Tails\middle|Z \sim \mathcal{DU}(6), Y \sim \mathcal{DU}(2)\right]}{P\left[Y = Tails\middle|Y \sim \mathcal{DU}(2)\right]}$$

7.10 Multiplication Rule

If we're interested in the Intersection of two events, we can multiply the probability values for the individual events in the following ways:

- Independent Events

$$P\left[Z = 2 \cap Y = Tails|Z \sim \mathcal{DU}(6), Y \sim \mathcal{DU}(2)\right]$$

$$= P\left[Z = 2|Z \sim \mathcal{DU}(6)\right] \cdot P\left[Y = Tails|Y \sim \mathcal{DU}(2)\right]$$

- Dependent Events (General Rule-works in all cases)

$$P\left[Z = 2 \cap Y = Tails|Z \sim \mathcal{DU}(6), Y \sim \mathcal{DU}(2)\right]$$

$$= P\left[Z = 2|Z \sim \mathcal{DU}(6)\right] \cdot P\left[Y = Tails|Z = 2\middle|Y \sim \mathcal{DU}(2), Z \sim \mathcal{DU}(6)\right]$$

$$= P\left[Y = Tails|Y \sim \mathcal{DU}(2)\right] \cdot P\left[Z = 2|Y = Tails\middle|Z \sim \mathcal{DU}(6), Y \sim \mathcal{DU}(2)\right]$$

8 Bayes' Theorem

Related to Conditional Events is the idea of Bayes' Theorem. The central tenant of this theorem is that we can use new information/additional evidence to update our probabilities values. This can be applied to both conditional events as well as non-conditional events.

8.1 Bayes and Conditional Events

For conditional events (the more standard formation of Bayes' Theorem), we can make the following statement:

$$P\left[\text{1st event}|\text{2nd event}\middle|\text{assumptions}\right] =$$

$$\frac{P\left[\text{1st event}\middle|\text{assump.}\right] \cdot P\left[\text{2nd event}|\text{1st event}\middle|\text{assump.}\right]}{P\left[\text{2nd event}\middle|\text{assumptions}\right]}$$

The essence of this application of Bayes' Theorem is that we can alter the order of our conditional event by using additional information.

8.2 Bayes' and Non-conditional Events

A more powerful use of Bayes' Theorem deals more with non-conditional events by leveraging updates to our assumptions. We would express Bayes' Theorem in this case as

$$\mathcal{L}\left(\text{assumptions}\middle|\text{data event}\right) =$$

$$\frac{P\left[\text{data event}\middle|\text{assumptions}\right] \cdot \mathcal{L}\left(\text{assumptions}\middle|\text{prior to getting data}\right)}{P\left[\text{data event}\middle|\text{all possible assumptions}\right]}$$

Notice that in the above formation of Bayes' Theorem that we've introduced likelihood with $\mathcal{L}$ as the name of the likelihood function. Bayes' Theorem allows us to connect the long-run relative frequency of our assumptions being true given our data (i.e., likelihood) with the long-run relative frequency of a data event given our assumptions (i.e., probability). You'll also notice that we have two instances where the conditions differ from what we typically work with. For $P\left[\text{data event}\middle|\text{all possible assumptions}\right]$ we are looking that probability that we get our data across all possible assumptions that we might make; for $\mathcal{L}\left(\text{assumptions}\middle|\text{prior to getting data}\right)$, we refer to this as the "prior" likelihood of our assumptions being true before we look at any of our data. In other words, we're trying to express what the long-run relative frequency is that we get this data event given *any* set of assumptions and what's the long-run relative frequency of our assumptions being true given *any* data collection? This formation of Bayes' Theorem is what underpins and drive Bayesian Statistical Inference. We refer to $\mathcal{L}\left(\text{assumptions}\middle|\text{data event}\right)$ as the "posterior".