Probability Notes

Srikar Katta and Edoardo Airoldi

Contents

Probability

The basics of statistics and, by extension, data science are rooted in probability, the study of likelihoods and chances of outcomes and realizations. Using probability, we can predict the likelihood of future events or understand the effect of some treatment or interpolate missing values across space. Probability depends heavily on sets since we are usually interested in understand the chance of a set of outcomes occurring. The set of all possible outcomes is known as the sample space. For example, the sample space of a coin flip is $\{heads, tails\}$ because a coin can either come up as heads or tails. An event is some subset of the same space. For example, $\{heads\}, \{tails\}, \{heads, tails\}, \emptyset$ are the possible events of a coin flipping game. The probability of an event is a representation of the chance of that event occurring, with a probability of 1 meaning that the event is guaranteed while a probability of 0 meaning that the event is guaranteed to not happen.

Definition 0.1: Sample space

A **sample space** is the set (i.e., all elements are unique) of all possible outcomes in a probability space. It is often denoted by Ω.

Definition 0.2: Event

An **event** is some subset of the sample space.

There are some aspects of probability theory that we take as truth and based on these assumptions build the rest of probability. These are known as the axioms of probability. Consider the sets of events A, B, a subset of the sample space Ω, with $\mathbb{P}(A)$ denotes the probability of A. The axioms state

1. Nonnegativity: $\mathbb{P}(A) \geq 0$
2. Normalization: $\mathbb{P}(\Omega) = 1$
3. Finite Additivity: If A and B are disjoint (i.e., $A \cap B$) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

As a consequence of these axioms, $\mathbb{P}(A) \leq 1$, $\mathbb{P}(\emptyset) = 0$, $P(A) + P(A^c) = 1$–where A^c represents the complement of A (i.e., all the elements not in A but in Ω)–$B \subset A$ (i.e., B is a strict subset of A) implies $\mathbb{P}(B) < \mathbb{P}(A)$.

Probability Basics

We often run into situations when we are interested in the probability of the union of two non-disjoint sets, A and B. Then,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

While the $\mathbb{P}(A \cup B)$ may seem unintuitive, recall that sets cannot have duplicated elements. So, since A and B are non-disjoint sets, they share elements. So, $A \cup B$ will have less elements that the sum of the number of elements in A and B individually. By including the $\mathbb{P}(A \cap B)$ term, we account for the redundancy in elements between the two sets. And when A and B are disjoint, $A \cap B = \emptyset$; since $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Example 1 *Suppose we have 100 students in a school: 50 students who have science majors, 60 students who have liberal arts majors, and 30 are both. What are the chances that we randomly select a student who has a science and/or liberal arts major?*

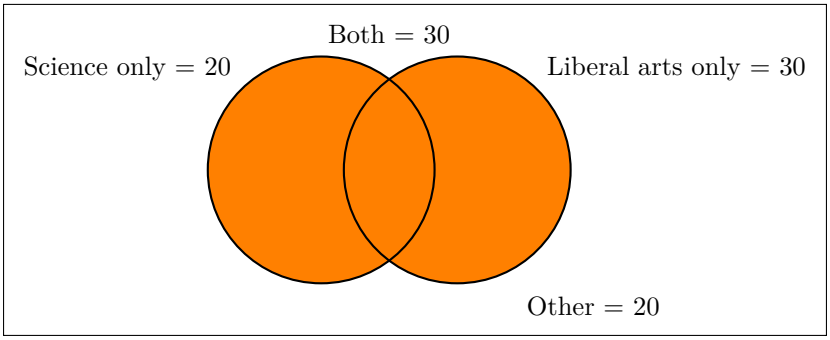


Figure 1: Example **??** Venn Diagram

First, recognize that since we have 100 students, the chances of selecting any one student, irrespective of their major, is $\frac{1}{100}$. *Now, because we have 50 science majors (set A in this situation), the chances of randomly selecting a student who has a science major is* $50 * \frac{1}{100} = 0.5$. *Similarly, because we have 60 liberal arts majors (set B in this situation), the chances of selecting a student who has a liberal arts major is* $60 * \frac{1}{100} = 0.6$. *Lastly, because we have 30 students who are majoring in both science and liberal arts* $(A \cap B)$, *the probability of choosing someone who is majoring in both is* $30 * \frac{1}{100} = 0.3$. *So, the probability of choosing any individual student who has a science or liberal arts major is* $0.5 + 0.6 - 0.3 = 0.8$. *Again, because we are "double counting" students across both majors, we must remove the redundancy, which is why we subtract 0.3.*

We can also validate our answer by understanding the role of the science and liberal arts majors in the whole sample space. In this situation, students can either be science majors, liberal arts majors, both, or neither: there is no other possibility. So if we are interested in the population of science, liberal arts, or both, that means that the only remaining possibility is that a student is majoring in something else. Because we know that $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$ *for some set A, a student majoring in something else is the complement of majoring in science, liberal arts, or both. Since there are 20 students majoring in something else,* $\mathbb{P}(science \cup liberal\ arts) = 1 - \mathbb{P}(other) = 1 - 0.2 = 0.8$, *the answer we arrived at earlier.*

In addition to the inclusion-exclusion principle, we often need to consider conditional probabilities. Considering the setup in Example 1, suppose we are working for the liberal arts school. The school is interested in creating a scholarship for its students and wants to know the probability that a double major will win the award randomly. While it may seem as though there is a 30% chance because 30 of the 100 students in the university are double majors, the liberal arts school only cares about students who have liberal arts majors, which is actually only 60 students. So the chances of awarding the scholarship to a liberal arts student is 50%, not 30%.

We often denote conditional probabilities with the | sign. Consider two sets A and B. If we want to know the probability of the events in A occurring, knowing that B is true, then we write this as $\mathbb{P}(A|B)$ and it is calculated as $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$. Let us break this down more: because B is guaranteed to be true, we only care about

the outcomes in A that also coincide with outcomes in B, which is the $A \cap B$ term. And because B is true, we are no longer working with the entire population – only the population in B.

Definition 0.3: Conditional Probability

The chance of the set A occurring, conditional upon set B being true, is known as **conditional probability** and is calculated as $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

From conditional probabilities comes the important idea of independence. In probability, we say two events are independent if one event occurring does not change the outcome of another event. So, for example, when flipping a fair coin, the outcome from one flip will not impact the outcome from a subsequent flip: these are independent events. Mathematically, we know the event A is independent from the event B if $\mathbb{P}(A|B) = \mathbb{P}(A)$: the chances of A occurring do not depend on B.

Definition 0.4: Law of Total Probability

Suppose we have two events A and B. Partition B into n disjoint sets: $\{B_1, B_2, \ldots, B_n\}$ The **Law of Total Probability** states that $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$.

From the concept of independence comes an important relationship about set intersections. Suppose the sets A and B are independent of one another. Then, $\mathbb{P}(A|B) = \mathbb{P}(A)$. Recall also that $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$. So, that implies that $\mathbb{P}(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, which in turn can be written as $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. So, if two events are independent, then the probability of their intersection is equivalent to the product of their individual probabilities.

Definition 0.5: Independence

Two events A and B are independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$. From this, we know that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ if A and B are independent.

Because we are discussing calculating the probability of two events co-occurring, notice that $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, so $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$. Since $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$, we can now replace $\mathbb{P}(A \cap B)$ with $\mathbb{P}(A|B)\mathbb{P}(B)$. So,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

This idea, known as Bayes' Rule, is really just converting a realization of an event (B) into the actual occurrence of an event (A).

Definition 0.6: Bayes' Rule

Bayes' Rule states that, given two events A and B,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Example 2 *We are testing students for COVID-19. There are four possible outcomes: the student tests positive and the student actually has COVID-19, the student tests negative and the student actually does not have COVID-19, the student tests positive and the student does not actually have COVID-19, and the student tests negative and the student does actually have COVID-19. The following table illustrates the probability of each of these events occurring.*

We are given the following information: students have a 20% chance of having COVID-19. However, conditional upon having COVID-19, a student has a 98% chance of testing positive and a 2% chance of testing

negative. And conditional upon not having COVID-19, a student has a 1% chance of testing positive and a 99% chance of testing negative.

We know that students will test positive or negative, but using this information, can we backtrack and understand the chances that a student actually has COVID-19 given that they tested positive?

From conditional probability rules, we know that

$$\mathbb{P}(\textit{has COVID-19}|\textit{test positive}) = \frac{\mathbb{P}(\textit{has COVID-19, test positive})}{\mathbb{P}(\textit{test positive})}.$$

However, we have no information on $\mathbb{P}(\textit{has COVID-19, test positive})$*. Instead, we can take advantage of Bayes' Rule to backtrack this information:*

$$\mathbb{P}(\textit{has COVID-19}|\textit{test positive}) = \frac{\mathbb{P}(\textit{test positive / has COVID-19})\mathbb{P}(\textit{has COVID-19})}{\mathbb{P}(\textit{test positive})}.$$

We know that $\mathbb{P}(\textit{test positive / has COVID-19}) = 0.98$ *and* $\mathbb{P}(\textit{has COVID-19}) = 0.2$*. However, what is the probability of a student testing positive? From the law of total probability, we know that* $\mathbb{P}(\textit{test positive}) = \mathbb{P}(\textit{test positive / has COVID-19})\mathbb{P}(\textit{has COVID-19}) + \mathbb{P}(\textit{test positive / does not have COVID-19})\mathbb{P}(\textit{does not have COVID-19})$*. So,* $\mathbb{P}(\textit{test positive}) = 0.98(0.2) + 0.01(0.2) = 0.196 + 0.002 = 0.198$.

Putting all of the factors together, $\mathbb{P}(\textit{has COVID-19}|\textit{test positive}) = \frac{0.98(0.2)}{0.198} = \frac{0.196}{0.198} \approx 0.99$*. So, the chance that a student actually has COVID-19 given that they test positive is 99%.*

Random Variables and Probability Distributions

If we think back to elementary mathematics, we first learned about basic mathematical operations on the integers (e.g., addition and multiplication) and then discussed algebra and solving for missing values. Likewise, now that we understand the basic set theoretic operations for probability, we can expand our toolkit to discuss probabilistic variables, often referred to as random variables.

Definition 0.7: Random variables

A **random variable** is a function that maps the possible values in a sample space to numerical representations. Random variables are typically denoted with capital letters.

Consider the following example for the outcome of a fair coin toss: let X be a random variable for the outcome of a fair coin toss. Then, $X = 1$ if we flip heads up and $X = 0$ if we flip tails up. While random variables are mappings of possible values to numerical representations, they do not detail the exact value of a coin toss, for example. After we toss the coin, the coin takes on some value (i.e., heads or tails) and there is some numeric representation for that outcome (i.e., 0 or 1). The numeric representation of the actual outcome is known as the realization.

Definition 0.8: Realization

The value that a random variable actually takes on is called the **realization**.

So far, we have discussed random variables as simply mappings from the sample space to a numerical representation; however, we know that in a probability space, we have more than just the sample space. We also have the probability function. For example, in our discussion of coin tosses, the probability function maps heads to 50% and tails to 50%. We extend this probability function to now work for random variables; namely, the probability distribution is the mapping of a random variable's outcomes to a probability. Reconsidering the random variable X for a coin toss (i.e.,$X = 1$ for heads up and $X = 0$ for tails up), the probability that $X = 1$ is the same as saying the probability of flipping a coin heads up. The probability that $X = 0$ is the same as saying the probability of flipping tails up.

Definition 0.9: Probability Distribution

A **probability distribution** is a function that maps the outcomes of a random variable to the probability of observing that outcome.

Discrete Random Variables There are two types of random variables: discrete and continuous. Discrete random variables typically deal with sample spaces whose possible outcomes we can write in a list, and their probability distribution functions are called probability mass functions. For example, let X be a random variable representing the face of a single die roll. There are six possible values of X: $\{1, 2, 3, 4, 5, 6\}$. Because there is a finite number of outcomes, the random variable of X is a discrete distribution. Additionally, suppose Y is the number of times we flip a coin heads up in an infinite number of tries. Then, the possible values of Y are $\{0, 1, 2, 3, \dots \infty\}$. Even though there is an infinite number of possible values for Y, we can write all of the outcomes in a list. So, Y is a discrete random variable. Now, assume that Z is a random variable that selects a real number in the space $[0, 1]$. No matter how hard we try, we can never list out all the possible outcomes of Z. By way of contradiction, let us assume that we can list out the possible outcomes of Z: $\{a_1, a_2, \dots, a_\infty\}$, where a_i is some possible outcome for Z. Because Z's domain is over the space $[0, 1]$, there will definitely be a number between any two consecutive elements, i.e. $\frac{a_i + a_{i+1}}{2}$ is also in Z. So no matter how hard we try, we will have an uncountably infinite number of elements in Z's co-domain, so Z is not a discrete random variables.

Definition 0.10: Probability Mass Function

A **probabilty mass function** (pmf) is the probability distribution function for a discrete random variable. Let X be a discrete random variable. The pmf of X, which we call $f(x)$, is given by $f_X(x) = \mathbb{P}(X = x)$; in other words, it describes the probability that the disrecete random variable is equal to a specific value.

Let us consider a few examples of discrete random variables.

Example 3 *Suppose we are rolling a six-sided, fair die. Let N be a random variable representing the number of times we roll the die until we roll our first 6. We want to find the support and the probability mass function of N. Let us first recall what the support and pmf of a random variable are. The pmf is the probability of realizing an event c, i.e. $\mathbb{P}(N = n)$. And the support is all values of n such that $\mathbb{P}(N = c) > 0$. Because we can roll $1, 2, 3, \dots, \infty$ number of times before rolling a 6, the values of N that have probabilities over 0 are $\{1, 2, \dots, \infty\}$. These values are known as the support of a random variable.*

Definition 0.11: Support

*Let X be a random variable with probability distribution function $f_X(x)$. The **support** of X is all the possible values of X such that $f_X(x) > 0$.*

Now, finding the pmf is a little trickier but doable. To help us keep track of which roll we are discussing, let us introduce new random variables $X_1, X_2, \dots, X_n$, each of which represent the outcome of a single roll; for any random variable in this set, represented by X_i,

$$X_i = \begin{cases} 1 & \text{if we roll a 6} \\ 0 & \text{if we do not roll a 6.} \end{cases}$$

Then, the pmf of X_i would be given by

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1 \\ \frac{5}{6} & \text{if } x = 0. \end{cases}$$

First, realize that $\mathbb{P}(N = n)$ is the same as saying that we do not roll a 6 for $n-1$ times and roll a 6 the last time. In other words, $\mathbb{P}(N = n) = \mathbb{P}(X_1 = 0, X_2 = 0, \ldots, X_{n-1} = 0, X_n = 1)$. While the joint probability may seem intimidating, we can easily simplify it by recalling that because the outcome of one roll does not impact the probability of another roll, each of the X_i's are independent of one another. This means that for two events A and B, $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$. So, because they are independent,

$$\begin{aligned}
\mathbb{P}(N = n) &= \mathbb{P}(X_1 = 0, X_2 = 0, \ldots, X_{n-1} = 0, X_n = 1) \\
&= \mathbb{P}(X_1 = 0)\mathbb{P}(X_2 = 0)...\mathbb{P}(X_{n-1} = 0)\mathbb{P}(X_n = 1) \\
&= \mathbb{P}(X_n = 1) \prod_{i=1}^{n-1} \mathbb{P}(X_i = 0) \\
&= \frac{1}{6} \prod_{i=1}^{n-1} \frac{5}{6} \\
&= \frac{1}{6} \left(\frac{5}{6}\right)^{n-1}.
\end{aligned}$$

So, the probability mass function of N is

$$f_N(n) = \frac{1}{6} \left(\frac{5}{6}\right)^{n-1}.$$

This example covers a few important topics. Besides offering an example of discrete random variables, we also introduced the idea of the support of a distribution. Implicitly, we also discussed independently and identically distribution, or IID for short, variables. When we have a collection of random variables that all follow the same probability distribution function and are independent of one another, we call them independently and identically distributed or IID for short. The idea of IID variables is crucial in applied statistics and data science. <!– First, the probability mass function of the form $p(1-p)^{n-1}$ for some p between $[0, 1]$ and a random variable with the support being the whole numbers (i.e., $0, 1, 2, \ldots, \infty$) is known as the geometric distribution. These are typically used to model "time-to-event" problems in discrete settings, like the number of times we need to roll a die to see a 6. –>

<!– Implicitly through the problem, we also learned about the Bernoulli distribution: when the probability mass function is of the form $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some p in $[0, 1]$ and a random variable with support $\{0, 1\}$, the pmf is called a Bernoulli distribution. These are typically used to model "indicator" problems that deal with binary responses. –>

Definition 0.12: IID

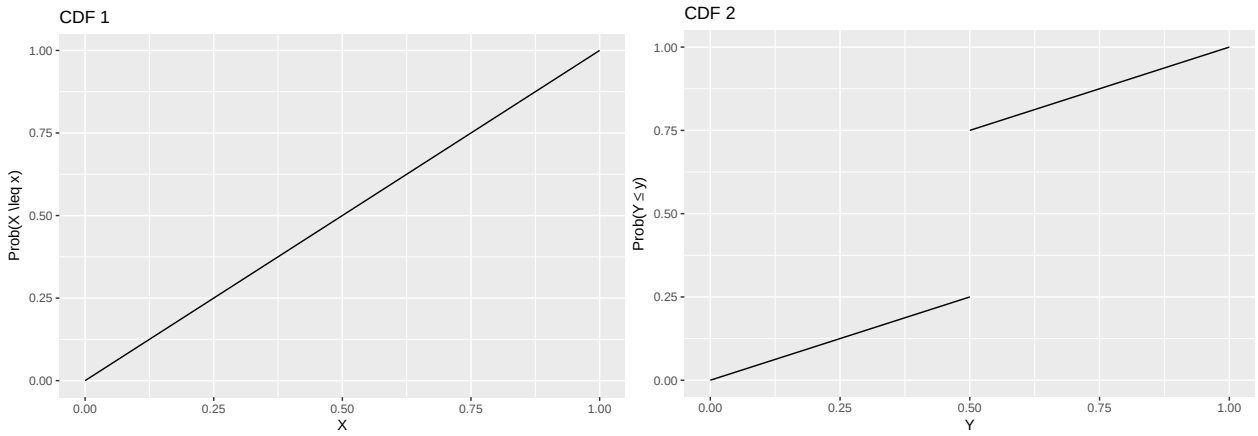
*A collection of random variables are called independently and identically distributed, or **IID** for short, if all the random variables follow the same probability distribution and are independent of one another.*

Continuous Random Variables To understand continuous random variables, we first need to introduce the cumulative distribution function, a function that maps the possible outcomes of a random variable to values between 0 and 1. The cumulative distribution function asks "what is the probability that the realization of some random variable is less than or equal to a specified value?" While it is not important to our discussion now, because the CDF only looks at "less than or equal to" values, the CDF is an increasing function. In other words, let x_1, x_2 be elements of the sample space of the random variable X. If $x_1 < x_2$, then $F_X(x_1) \leq F_X(x_2)$ where $F_X(x)$ represents the CDF of X.

Definition 0.13: Cumulative Distribution Function

Let X be a random variable. Then, the **cumulative distribution function** (CDF) $F_X(x)$ is a function that maps X to $[0,1]$ defined by $F_X(x) = \mathbb{P}(X \leq x)$. The general notation of the CDF is $F_{[\text{random variable}]}(\text{realization})$.

Now, the formal definition of a continuous probability distribution is a random variable whose cumulative distribution function is differentiable. Consider the following cumulative distribution functions:



First, it is easy to recognize that the cumulative distribution function of X is differentiable because it is smooth and there are no breaks or kinks in the function. So, X's probability distribution is continuous. On the other hand, notice that CDF2 has a sudden jump at $Y = 0.5$, so its CDF is not differentiable and Y is therefore not a continuous random variable. However, not continuous does not imply discrete. Because we cannot "list out" the possible values of Y, Y's probability distribution is not discrete either.

While probability mass functions calculate $\mathbb{P}(X = x)$ for a discrete random variable X, probability density functions are not as straightforward because in an uncountably infinite set, the possibility of choosing exactly one value is 0. Let Y be a random variable chosen between $[0,1]$, and suppose we were interested in the probability that Y is 0.3; then, $\mathbb{P}(Y = 0.3) = \frac{1}{|[0,1]|}$, where $|A|$ represents the number of units in the set A. Because $[0,1]$ is continuous, it has an uncountably infinite number of values. So, $\mathbb{P}(Y = 0.3) = 0$, which would mean that it is impossible for us to observe 0.3. Instead, we will ask what is the probability that Y lies in the range $[0.3, 0.3 + \delta]$, where δ is some infinitesimal number. For the same reason that $[0,1]$ has an uncountably infinite number of values, ($[0.3, 0.3 + \delta]$ has an uncountably infinite number of possible values; $[0.3, 0.3 + \delta]$ simply has a smaller uncountably infinite number of values. So we should be able to quantify the probability that $Y = 0.3$.

We know that $F_Y(0.3) = \mathbb{P}(Y \leq 0.3)$ and $F_Y(0.3 + \delta) = \mathbb{Y} \leq \nvDash \nVdash + \delta$. So, we should be able to find $\mathbb{P}(0.3 \leq Y \leq 0.3 + \delta)$ with $F_Y(0.3 + \delta) - F_Y(0.3)$, which is actually a calculation for the change in F_Y from 0.3 to $0.3 + \delta$. Now, we can get even more precise values by finding $\lim \delta \to 0 F_Y(0.3 + \delta) - F_Y(0.3)$, which is just the instantaneous rate of change, which is just the derivative. So, we define the probability density function of Y as $f_Y(y) = \frac{d}{dy} F_Y(y)$.

Definition 0.14: Probability density function

Let Y be a continuous random variable with CDF $F_Y(y)$. Then, the **probability density function** of Y is given by $f_Y(y) = \frac{d}{dy} F_Y(y)$.

We know that probability distributions are functions with domains and co-domains. However, we often care about something beyond the domain in probability: the support of a probability distribution. As an illustration, suppose we are trying to relate income to a random variable. We can define a variable Z that maps the sample space $(-\infty, \infty)$ to (∞, ∞); because it maps the entire sample space to the real numbers, Z is a random variable. However, income (in this very simplistic setting) cannot be negative. So, $\mathbb{P}(Z < 0) = 0$. Even though, $(-\infty, 0)$ is a part of the domain of Z's probability distribution, we care more about $[0, \infty)$

because those values are possible.

Named Distributions

Every probability distribution has its own functional form, but some are so well known and so widely used that they have their own names. In this portion, we will discuss a few of the main distributions and demonstrate how to sample values from the distribution in `R`.

Named Discrete Random Variables and Distributions

Bernoulli Random Variable

If a random variable X's support is $\{0, 1\}$ (i.e., $\mathbb{P}(X = x) > 0$ if and only if $x = 0$ or $x = 1$), then X is said to follow the Bernoulli distribution.

Definition 0.15: Bernoulli distribution

Let X be a random variable that follows the **Bernoulli distribution**. Then, the probability mass function of X is

$$f_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} \tag{1}$$

for some p in the range $(0, 1)$. We denote X as a Bernoulli random variable with parameter p by writing X $Bernoulli(p)$. We read the symbol as "is distributed as."

We often use Bernoulli random variables to indicate whether an outcome occurred or not. The outcome of interest occurring is represented by the random variable taking on a value of 1 and a value of 0 if the outcome did not occur. We also use Bernoulli random variables in experiments where the only outcomes are either “successes” or “failures” but not both. In this type of experiment, known as a Bernoulli trial, we can consider the Bernoulli random variable to be an indicator of success, such that the random variable is 1 if the success occurs or the random variable is 0 if the success does not occur. Because the random variable is 1 with probability p, the parameter p is sometimes referred to as the “success probability.”

We actually saw Bernoulli random variables in Example 3. We used them to denote whether the outcome of a single die roll was a six or not. In this example too, the Bernoulli random variable was used as an indicator variable of sorts.

Binomial Random Variable

Let N be some positive integer and p be a real number in the range $(0, 1)$. Let $X_1, \ldots, X_N$ be N random variables all simulated from $Bernoulli(p)$. Let $X = \sum_{i=1}^{N} X_i$; in other words, X represents the number of successes in N independent Bernoulli trials with the same success probability p. X follows the Binomial distribution, and we therefore call it a binomial random variable.

We know that probability mass functions have two major components: the support (i.e., values of X such that $\mathbb{P}(X = x) > 0$) and the functional form (i.e., $\mathbb{P}(X = x)$). So let us intuitively devise both of these components.

Starting with the support, we know that we have N independent trials and that X represents the number of successes. The lower bound for X is having no trials succeed, which would mean $X = 0$. And the upper bound for X is having all the trials succeed, which would mean $X = N$. Additionally, it is impossible for us to have a fractional number of successes (e.g., 2.5 successes), so the support of X is all integer values between 0 and N, inclusive.

The pmf is a little more difficult. But suppose $X = 0$; this would imply that $X_1 = 0, X_2 = 0, \ldots, X_N = 0$. And because $X_1, \ldots, X_N$ are IID random variables, they are independent, which means that the joint probability of all these events occurring is the product of the probabilities of the individual events occurring:

$$\begin{aligned}\mathbb{P}(X_1 = 0, X_2 = 0, \ldots, X_N = 0) &= \mathbb{P}(X_1 = 0)\mathbb{P}(X_2 = 0)\ldots\mathbb{P}(X_N = 0)\\ &= (1-p)(1-p)\ldots(1-p)\\ &= (1-p)^N,\end{aligned}$$

because the probability of any single trial failing is $(1-p)$, and we multiply $(1-p)$ N times. So, $\mathbb{P}(X = 0) = (1-p)^N$.

Now, suppose $X = 1$. Then, that means exactly 1 of the X_i's is 1. This could be when $X_1 = 1$, all else 0; or $X_2 = 1$, all else 0; or $X_3 = 1$, all else 0; etc. Notice, we have a sequence of *or* statements, which means we can apply the inclusion-exclusion principle (i.e., given events A and B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$). Notice, we can never realize two sequences together. For example, the events "$X_1 = 1$, all else 0" and "$X_2 = 1$, all else 0" can never occur together because they are contradictory (X_2 cannot be 1 if all else, which includes X_2, is 0 when $X_1 = 1$). So in this case, we can just sum the probability of each sequence of outcomes together. Now suppose we are looking at one individual sequence: $X_i = 1$, all else 0. Then, because each of the X_i are independent, we have $(1-p)$ multiplied $N-1$ times for the $N-1$ failures multiplied by p for the 1 success: $p(1-p)^{N-1}$. Now, realize that we have N possible sequences because each of the N random variables can be 1. So, we sum $p(1-p)^{N-1}$ together N times. This means, $\mathbb{P}(X = 1) = Np(1-p)^{N-1}$.

Now that we have gone through a couple specific examples, we can intuit the general functional form of the pmf. Suppose $X = m$ for some integer m between 0 and N, inclusive. Let us first find the probability of a single sequence occurring: the first m terms are successes and the last \$N - m \$ terms are failures. First recall that since the X_i's are independent of one another, the joint probability is the product of the probabilities of the individual events. Now, since the probability of a single success is p and the probability of a single failure is $1-p$, that means we take the product of p, m times, and $(1-p)$, $N-m$ times. So, $\mathbb{P}$(the first m terms are successes and the last $N-m$ terms are failures) $= p^m(1-p)^{N-m}$. Now, we have to figure out how many ways can we have m successes in N possible trials.

First, recognize that each of the X_i are independent of one another, so $\mathbb{P}(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} \mathbb{P}(X_i)$. Since there are m successes and $N-m$ failures, the probability of this occurring is $p^m(1-p)^{N-m}$. Now, we have to consider how many possible ways we can have m successes and $N-m$ failures though.

We know that there are N possible places to keep 1 success. Let us fix where that is. Now, we have $N-1$ random variables that could be the 2nd success. And then we have $N-2$ random variables that could be the 3rd success. And now notice a pattern emerging: we have $N-k+1$ possible places to keep our k^{th} success. So, when we consider all possible permutations of m successes and $N-m$ failures, we have $N(N-1)(N-2)\ldots(N-m+1)$ arrangements. But now, we have to consider that the order in which we place the successes actually does not matter. Whether X_2 was labeled as a success first or second is of no relevance to our problem since we simply care about the number of successes. We have m possible orderings for the first success, $m-1$ possible orderings for the second success, $m-2$ possible orderings for the third success, and $m-k+1$ orderings for the k^{th} success. So, the total number of combinations is actually $\frac{N(N-1)(N-2)\ldots(N-m)}{m(m-1)(m-2)\ldots1}$.

To make this more compact, let us introduce some notation on the idea of the factorial.

Definition 0.16: Factorial

Let n be an integer. The **factorial** of n, written as $n!$, is the product of all positive integers less than or equal to n:

$$\begin{aligned}n! &= n(n-1)(n-2)\ldots1\\ &= \prod_{i=1}^{n} i.\end{aligned}$$

Now, we can rewrite the total number of combinations as

$$\begin{aligned}\frac{N(N-1)(N-2)\dots(N-m)}{m(m-1)(m-2)\dots 1} &= \frac{N(N-1)(N-2)\dots(N-m)(N-m-1)\dots(3)(2)(1)}{m(m-1)(m-2)\dots(1)(N-m)(N-m-1)\dots(3)(2)(1)} \\ &= \frac{N!}{m!(N-m)!}.\end{aligned}$$

So, we can write the pmf of X as

$$\mathbb{P}(X=m) = \frac{N!}{m!(N-m)!}p^m(1-p)^{(N-m)}.$$

Definition 0.17: Binomial random variable

A **binomial random variable** represents the sum of N independent and identically distributed Bernoulli trials for some integer N. We would denote X being a binomial random variable with success probability p for N independent trials as $X \sim Binomial(N,p)$. And the pmf of X is given by

$$\mathbb{P}(X=x) = \frac{N!}{x!(N-x)!}p^x(1-p)^{N-x}.$$

Let us consider a few examples to illustrate the practicality of binomial random variables.

Example 4 *Suppose we are flipping a fair coin three times. What are the chances that we flip two heads?*

First, let X be a random variable denoting the number of times we flip the coin heads-up. Because the coin flips are independent (i.e., the outcome of one coin flip does not affect the outcome of another), we have 3 independent Bernoulli trials, each with success probability $\frac{1}{2}$, where success is given by flipping the coin heads-up. Because X is the sum of the outcomes, $X \sim Binomial(3, \frac{1}{2})$. So, the probability of flipping two heads is given by

$$\begin{aligned}\mathbb{P}(X=2) &= \frac{3!}{2!(3-2)!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{(3-2)} \\ &= \frac{3!}{2!1!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 \\ &= \frac{3\cdot 2\cdot 1}{2\cdot 1\cdot 1}\left(\frac{1}{2}\right)^3 \\ &= 3\left(\frac{1}{8}\right) \\ &= \frac{3}{8}.\end{aligned}$$

Now, we can confirm this answer by simplifying the problem. Let H mean that we flipped the coin heads-up and T mean that we flipped the coin tails-up. Let a triplet of letters represent the outcome of three coin flips; for example, we would represent flipping heads then tails then heads as HTH. Now, our sample space for this problem is $\{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$. Notice, there are only three possibilities in our sample space of eight elements in which we have exactly two heads-up: $\{HHT, HTH, THH\}$. So, the probability of us seeing two heads is $\frac{3}{8}$ from this approach as well.

Example 5 *Suppose we own an e-commerce business and each product at the store can be rated with a thumbs-up or a thumbs-down. If everyone has the same chance of giving a product a thumbs-up at $\frac{7}{10}$, what are the chances that at least 60 out of the 100 people who bought the product will give the product a thumbs-up? To simplify the problem, assume that one person's vote does not impact another person's review.*

Let X represent the number of people who rated the product with a thumbs-up. Because we assume that votes do not affect each other and that each vote can either be a thumbs-up or not, the votes are independent Bernoulli trials. And since the probability of anyone giving a product a thumbs-up is $\frac{7}{10}$, $X \sim Binomial\left(100, \frac{7}{10}\right)$.

Now, we need to know the probability that at least 60 people give the product a thumbs-up: $\mathbb{P}(X \geq 60)$. Notice, if $X = x_1$ and $X = x_2$, then $x_1 = x_2$ because it is impossible for us to have both x_1 and x_2 thumbs-up at the same time. So,

$$\begin{aligned}\mathbb{P}(X \geq 60) &= \mathbb{P}(X = 60 \cap X = 61 \cap \ldots \cap X = 100) \\ &= \mathbb{P}(X = 60) + \mathbb{P}(X = 61) + \ldots + \mathbb{P}(X = 100).\end{aligned}$$

Now, we could either calculate the probability of each of these, or we could use `R` and simplify our work. There is a function called `pbinom` that returns the output of the cumulative distribution function (i.e., $\mathbb{P}(X \leq x)$, if given x, the success probability, and the number of trials. While it may be tempting to simply to take the answer from `pbinom`, it would be incorrect because `pbinom(60, ...)` would return $\mathbb{P}(X \leq 60)$, when we are actually interested in $\mathbb{P}(X \geq 60)$. To over come this issue, we can realize that because the probability of the entire sample space is 1,

$$\begin{aligned}\mathbb{P}(X \geq 60) &= 1 - \mathbb{P}(X < 60) \\ &= 1 - \mathbb{P}(X \leq 59).\end{aligned}$$

So, we can find the solution by calculating

So, $\mathbb{P}(X \geq 60) \approx 0.988$, which means we are almost guaranteed for at least 60 of 100 people to rate the problem with a thumbs up if our success probability is 0.7.

Now, we can also confirm our `R` answer. We can create a function that calculates the pmf $\mathbb{P}(X = x)$ and then take the sum over a sequence:

We have confirmed our answer.

```
binomial_pmf <- function(N, p, x) {
  ## N represents the number of trials
  ## p represents the success probability
  ## x is the value at which we are evaluating the pmf: P(X = x)

  ## calculate (N!)/[x!(N - x)!] * p^x * (1-p)^(N - x)
  return_value <- (factorial(N)/(factorial(x) * factorial(N - x))) *
    (p^x) * ((1 - p)^(N - x))

  return(return_value)
}

sum_pmf_60_to_100 <- 0
for(x in 60:100) {
  sum_pmf_60_to_100 <- sum_pmf_60_to_100 + binomial_pmf(N = 100, p = 0.7, x = x)
}
print(sum_pmf_60_to_100)
```

```
## [1] 0.9875016
```

Multinomial Random Variables

We can also generalize the idea motivating binomial random variables to cases in which we have more than just two outcomes. For example, suppose we are interested in not just the number of 6s and non-6s rolled in a six-sided die in N trials; what would be the pmf of a random variable that accounted for the the number of times each side was rolled? The multinomial distribution generalizes the binomial distribution to answer exactly those questions.

Recall that we first introduced the idea of the Bernoulli distribution to motivate the source of the binomial distribution. The multinomial distribution has its own version of a Bernoulli trial, known as the categorical distribution.

Definition 0.18: Categorical distribution

Let X be a random variable following the **categorical distribution** with k outcomes for some positive integer k. Let $p_i = \mathbb{P}(X = i)$. The pmf of X is then given by

$$\mathbb{P}(X = x) = p_1^{\mathbb{1}(x=1)} p_2^{\mathbb{1}(x=2)} \dots p_k^{\mathbb{1}(x=k)},$$

where $\mathbb{1}(\cdot)$ is the indicator function, which returns 1 if its argument is true and 0 otherwise. Let j be a possible outcome. Then, $\mathbb{1}(x = 1) = \mathbb{1}(x = 2) = \dots = \mathbb{1}(x = j - 1) = \mathbb{1}(x = j + 1) = \dots = \mathbb{1}(x = k - 1) = \mathbb{1}(x = k) = 0$ and $\mathbb{1}(x = j) = 1$. Since anything to the power of 0 is 1, $\mathbb{P}(X = j) = p_j$.

Let X be a multinomial random variable with k categories and N independent trials. We now let X_i represent the number of outcomes in which category i was realized. In the other distributions we have seen, $\mathbb{P}(X = x)$ refers to the probability that a random variable is a scalar (e.g., $\mathbb{P}(X = 2)$). However, in the multinomial distribution, $\mathbb{P}(X = x)$ is equivalent to $\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$. And similar to the categorical distribution, p_i represents the success probability for a single category i. We leave the derivation of the pmf to the reader, but it follows a similar logic to the the binomial distribution:

$$\begin{aligned}\mathbb{P}(X = x) &= \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) \\ &= \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}.\end{aligned}$$

Definition 0.19: Multinomial random variable

Let X be a multinomial random variable with N independent trials and k categories such that the probability of realizing each category i is p_i. Then,

1. $\sum_{i=1}^{k} p_i = 1$: the sum of the probabilities of class realizations must be 1
2. $\sum_{i=1}^{k} X_i = N$: the total number of outcomes across all classes must be equal to the number of trials
3. The pmf is given by

$$\begin{aligned}\mathbb{P}(X = x) &= \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) \\ &= \frac{N!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}.\end{aligned}$$

Notationally, X is distributed multinomially is written as

$$X \sim Multinomial(p_1, p_2, \dots, p_k, N)$$

Example 6 *Let* $D \sim Multinomial\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 100\right)$. *How many categories does this problem have? How many independent trials do we run? And what is* $\mathbb{P}(D_1 = 20, D_2 = 20, D_3 = 25, D_4 = 10, D_5 = 5, D_6 = 20)$*?*

In order to first calculate the number of categories, we have to recall what each of the parameters of the multinomial distribution represent. We first have $p_1, p_2, \dots, p_k$, which represent the probability of each class being realized in one trial. Since we have 6 class realization probabilities, we have 6 classes.

In the multinomial distribution's construction, after we specify the class realization probabilities, we then have the number of independent trials. In this case, we have 100 independent trials.

Lastly, to evaluate $\mathbb{P}(D_1 = 20, D_2 = 20, D_3 = 25, D_4 = 10, D_5 = 5, D_6 = 20)$, we have to find the pmf of D:

$$\begin{aligned}\mathbb{P}(D = d) &= \frac{N!}{d_1!d_2!\dots d_6!}p_1^{d_1}p_2^{d_2}\dots p_k^{d_k} \\ &= \frac{100!}{20!20!25!10!5!20!}\left(\frac{1}{6}\right)^{20}\left(\frac{1}{6}\right)^{20}\left(\frac{1}{6}\right)^{25}\left(\frac{1}{6}\right)^{10}\left(\frac{1}{6}\right)^{5}\left(\frac{1}{6}\right)^{20} \\ &= \frac{100!}{20!20!25!10!5!20!}\left(\frac{1}{6}\right)^{100}.\end{aligned}$$

We can now calculate this value in R:

```
print(factorial(100)/(factorial(20)*factorial(20)*factorial(25)*factorial(10)*factorial(5)*factorial(20
```

```
## [1] 1.468638e-09
```

So there is a near 0 chance of these outcomes occurring.

Example 7 *Suppose we are rolling a fair, six-sided die. What is the probability that in 100 rolls, we roll twenty 1s, twenty 2s, twenty-five 3s, ten 4s, five 5s, and twenty 6s?*

Because we are dealing with the number of outcomes of multiple classes across independent trials, we should model this problem using the multinomial distribution. First, recall that the parameters of the multinomial distribution are the probabilities of each class being realized in one trial and the number of independent trials. Since we are working with a fair, six-sided die, we have six classes, each of which has a class realization probability of $\frac{1}{6}$. And since we roll the die 100 times, we have 100 independent trials. So, let X be the random variable we use in this problem. Then,

$$X \sim Multinomial\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 100\right).$$

So, we want to know

$$\mathbb{P}(X_1 = 20, X_2 = 20, X_3 = 25, X_4 = 10, X_5 = 5, X_6 = 20).$$

Notice, this problem has the same exact form as Example 6. So, we already know the answer: the probability of these outcomes occurring is near 0. However, we will confirm this answer using R.

We previously discussed the `pbinom` function in example 5. There is another, very similar function for the multinomial distribution: `dmultinom`. `dmultinom` has arguments `x` and `prob`: `x` is a list that contains the number of times each class was observed in all the independent trials while `prob` contains the class realizations for the corresponding class in `x`. So, in R, we can easily find the probability of our outcome occurring:

```
print(dmultinom(x = c(20, 20, 25, 10, 5, 20), prob = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)))
```

```
## [1] 1.468638e-09
```

The result from R is the same as what we had calculated previously in Example 6, confirming our computation.

Example 8 ?? *Similar to Example 5, we will consider another example studying the probability of a product having certain ratings. Rather than having thumbs-up or thumbs-down ratings, we have five stars. We want to know the probability that out of 2021 ratings, we will have 103 one-star ratings, 200 two-star ratings, 505 three-star ratings, 507 four-star ratings, and 706 five-star ratings. The probability of anyone rating the product with x stars is given below:*

Let us assume that one person's rating does not affect another's. Since the ratings are independent of one another and we have multiple categories, it makes sense to model this problem as a multinomial distribution. Let $R \sim Multinomial(0.05, 0.1, 0.25, 0.25, 0.35, 2021)$ represent the ratings results. Using, R, we can calculate $\mathbb{P}(R_1 = 103, R_2 = 200, R_3 = 505, R_4 = 507, R_5 = 706)$:

Table 1: Multinomial Ratings Probability

Rating	Probability
1	0.050
2	0.100
3	0.250
4	0.250
5	0.350

```
print(dmultinom(x = c(103, 200, 505, 507, 706), prob = c(0.05, 0.1, 0.25, 0.25, 0.35)))
```

```
## [1] 5.694475e-07
```

So, the probability that out of 2021 ratings, we will have 103 one-star ratings, 200 two-star ratings, 505 three-star ratings, 507 four-star ratings, and 706 five-star ratings is about $5.6 \cdot 10^{-7}$.

Geometric Random Variables

Poisson Random Variables

Negative Binomial Random Variables

Normal Random Variables

log-Normal Random Variables

Exponential Random Variables

Gamma Random Variables

Expectation and Variance Operators