Math 180A

Introduction to Probability

Spring 2022

Taught by Professor Brett Kolesnik

Table of Contents

1 Discrete Probability Distributions

In this section, we are mostly concerned with discrete sample spaces, or *finite* (or countably infinite) sample spaces.

1.1 Sample Space

Probability is the study of randomness, of uncertainty. Formally, we can think of this process as running a **random experiment**.

Definition 1.1: Sample Space

The **sample space** of an experiment is the set of all possible outcomes of that experiment. Namely, we say that

$$\Omega = \{\omega_1, \ldots, \omega_n\},$$

where each ω_i represents a possible outcome.

(Example.) When flipping a coin, we would have

$$\Omega_{\text{Coin}} = \{\text{Heads}, \text{Tails}\}.$$

The outcomes are assigned **masses** $m(\omega_i) \geq 0$ such that

$$m(\omega_1) + \cdots + m(\omega_n) = 1.$$

Here, we can think of the $m(\omega_i)$ as the **probability** that the outcome ω_i occurs.

(Example.) When rolling a regular die, we would have

$$\Omega_{\text{Die}} = \{1, 2, 3, 4, 5, 6\}.$$

Here, each $\omega \in \Omega_{\text{Die}}$ has a mass of $\frac{1}{6}$; that is,

$$\forall \omega \in \Omega_{\text{Die}}, m(\omega) = \frac{1}{6}.$$

1.2 Random Variables

We can use *random variables* to quantify the outcome.

Definition 1.2: Random Variable

Suppose we have an experiment whose outcome depends on chance. We can represent the outcome of the experiment by a captial Roman letter, such as X, called a **random variable** (RV).

We can think of a random variable X as a function from the sample space Ω to the set of real numbers $\mathbb{R}$; that is,

$$X : \Omega \mapsto \mathbb{R}.$$

If the outcome $\omega \in \Omega$ has a random occurrence, then the value $X(\omega)$ will also be random.

(Example.) Suppose we wanted to flip a fair coin. It's obvious that the sample space Ω is just Heads or Tails. So, we can define a random variable X like

$$X = \begin{cases} 1 & \text{if Heads} \\ 0 & \text{if Tails} \end{cases}.$$

We can also define the random variable

$$X = \begin{cases} 21313 & \text{if Heads} \\ 0 & \text{if Tails} \end{cases}.$$

The point is that your random variable maps your outcomes (from your sample space) to numbers. In other words, you are quantifying your outcomes.

1.3 Events

Often, in probability, we are interested in the probability that a certain event will occur. For example, we might be interested in whether or not it will rain today, or whether or not we will win the lottery, or so on.

Definition 1.3: Event

A subset of a sample space is an **event**.

The probability of an event E is given by

$$\mathbb{P}(E) = \sum_{\omega \in E} m(\omega).$$

(Example.) If we roll a die, then $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event,

$$E = \text{“Roll an even number''},$$

is the subset $E = \{2, 4, 6\} \subset \Omega$.

1.4 Probability Distribution

If we have an outcome ω, we can use the probability distribution function to get the probability that ω occurs.

Definition 1.4: Probability Distribution

Let Ω be a discrete (finite or countable infinite) set. Then, the function

$$\mathbb{P} : \Omega \mapsto [0, 1]$$

is called a **probability distribution** on Ω if the following hold:

1. $\mathbb{P}(\omega) \geq 0$ for all $\omega \in \Omega$, and
2. $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$.

(Example.) If we roll a fair die, then we have

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Then, it follows that

$$\mathbb{P}(i) = \frac{1}{6}$$

for all $i \in \Omega$.

1.5 Probability Mass Function

If X is a random variable on Ω, then we note that $\mathbb{P}(X = x)$ is a probability distribution on the set

$$\Omega_X = \{X(\omega) : \omega \in \Omega\}$$

of all possible values that X can take.

Definition 1.5: Probability Mass Function

For a discrete random variable X, the function

$$\mathbb{P} : \Omega_X \subset \mathbb{R} \mapsto [0, 1]$$

is called a **probability mass function** (PMF) of said random variable if the following hold:

1. $\mathbb{P}(x) \geq 0$ for all $x \in \Omega_X$, and
2. $\sum_{x \in \Omega_X} \mathbb{P}(x) = 1$.

In other words, given a possible value of the random variable, the probability that the random variable takes that particular value is given by the function above.

We note that, for a random variable X,

$$\mathbb{P}(X = x) = \sum_{\omega : X(\omega) = x} m(\omega).$$

The $X = x$ inside the $\mathbb{P}(X = x)$ is shorthand notation for the **event**

$$\{\omega : X(\omega) = x\} \subset \Omega.$$

Additionally, instead of writing $\mathbb{P}(X = x)$, we may also write $p_X(x)$.

(Example.) If we have a sample space Ω and a random variable $X : \Omega \mapsto \mathbb{R}$, then we can ask questions like "How likely is it that the value of X is equal to 2?" This is the same as the probability of the event

$$\{\omega : X(\omega) = 2\},$$

or, equivalently,

$$\mathbb{P}(X = 2)$$

or

$$p_X(2).$$

(Example.) Suppose again we have a fair die. If X is a random variable that takes the value 1 if the roll is 1 or 2, and the value 0 otherwise, then

$$\mathbb{P}(X = 1) = \frac{1}{3}$$

and

$$\mathbb{P}(X = 0) = \frac{2}{3}$$

is the probability distribution of X on the set $\Omega_X = \{0, 1\}$.

So, effectively, we can think of a probability mass function as a probability distribution function for values that the random variable can take. **We will be using this a lot.**

1.6 Cumulative Distribution Function

While we have the probability mass function, we also have the cumulative distribution function.

Definition 1.6: Cumulative Distribution Function

The **cumulative distribution function** (CDF) of a discrete random variable X is the function given by

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{t : t \leq x} \mathbb{P}(X = t).$$

So, whereas the PMF gives us the probability of a random value taking a specific value, the CDF gives us the probability that a random variable takes on a value *less than or equal to* a specific value.

(Example.) Suppose we're rolling a fair die, and suppose we have the random variable X that takes the value 1 if $\omega \in \{1\}$ and the value 0 if $\omega \in \{2, 3, 4, 5, 6\}$. Then, $X = 1$ if we roll a 1 and $X = 0$ otherwise. Then,

$$\mathbb{P}(X = 1) = m(1) = \frac{1}{6} \text{ and } \mathbb{P}(X = 0) = m(2) + \cdots + m(6) = \frac{5}{6}.$$

(Example.) Suppose we roll a die 3 times in a row. Let X be the number of 1's that we roll. Then,

$$X = X_1 + X_2 + X_3$$

where each X_i has the same distribution, i.e. $\mathbb{P}(X_i = 1) = \frac{1}{6}$ and $\mathbb{P}(X_i = 0) = \frac{5}{6}$ for each $i \in [1, 3]$.

(Example.) Again, suppose we roll a die 3 times in a row. Then,

$$X = X_1 + X_2 + X_3.$$

Suppose we wanted to find the probability that at most *2* of those rolls are 1's. We essentially need to calculate

$$\mathbb{P}(X \leq 2) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2).$$

There are three possibilities:

(a) $\mathbb{P}(X = 0)$: All 3 rolls are not 1's. Here, the probability that we don't get a 1 is $\frac{5}{6}$. So, the probability that all three rolls are not 1's is given by $(5/6)^3$. There is only $\binom{3}{0} = 1$ way we can possibly obtain zero 1's. To visualize this, let x be some non-one integer. Then,

```
x x x.
```

(b) $\mathbb{P}(X = 1)$: Exactly 1 of the 3 rolls is a 1. The probability that we get a 1 is $\frac{1}{6}$ and the probability that we don't get a 1 is $\frac{5}{6}$. We note that there are $\binom{3}{1} = 3$ possible ways we can obtain one 1. To visualize this, let x be some non-one integer. Then,

```
1 x x
x 1 x
x x 1.
```

(c) $\mathbb{P}(X = 2)$: Exactly 2 of the 3 rolls is a 1. We note that there are $\binom{3}{2}$ ways we can possibly get two 1's. To visualize this, let x be some non-one integer. Then,

```
1 1 x
x 1 1
1 x 1.
```

Putting this together, we have:

$$\mathbb{P}(X \leq 2) = \binom{3}{0}\left(\frac{5}{6}\right)^3 + \binom{3}{1}\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2 + \binom{3}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) \approx 99.54\%.$$

1.7 Review of Set Theory

Recall that events are subsets. In particular, let $A, B \subset \Omega$ be two events. Then:

- Intersection: $A \cap B = \{\omega \mid \omega \in A \text{ and } \omega \in B\}$.
- Union: $A \cup B = \{\omega \mid \omega \in A \text{ or } \omega \in B\}$.
- Difference: $A \setminus B = \{\omega \mid \omega \in A \text{ and } \omega \notin B\}$.
- Complement: $A^C = \Omega \setminus A$.

Two events A and B are said to be **disjoint** if $A \cap B = \emptyset$. If two events are disjoint, then it is impossible for them to both occur at the same time.

1.8 Properties of Probability Distribution

We now introduce our first theorem of this class.

Theorem 1.1

Suppose that $\mathbb{P}$ is a probability distribution on a discrete set Ω. Then,

1. $\mathbb{P}(E) \geq 0$ for all events $E \subset \Omega$.
2. $\mathbb{P}(\Omega) = 1$.
3. If $E \subset F \subset \Omega$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$.
4. If $A \cap B = \emptyset$ are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
5. $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$ for all events $A \subset \Omega$.

Proof. We'll prove each of the following statements in this theorem.

1. We know that E is a subset of Ω. Take some $\omega \in E$. Then, we know that its mass, $m(\omega) \geq 0$. Thus, it follows that
$$\sum_{\omega \in E} m(\omega) \geq 0.$$
2. Recall that Ω is the set of all possible outcomes. Therefore, it follows that $\mathbb{P}(\Omega) = 1$; if this is false, then this implies that there is at least one outcome that isn't in Ω.
3. Using (1) as a baseline, suppose that $\mathbb{P}(F) = f$. Since E can only have elements from F, it follows that $\mathbb{P}(E) \leq f$; if this statement is false, this implies that E has elements that aren't in F, which cannot be the case.

4. Since $A \cap B = \emptyset$, it follows that
$$\mathbb{P}(A \cap B) = \sum_{\omega \in A \cap B} \mathbb{P}(\omega) = \sum_{\omega \in A} \mathbb{P}(\omega) + \sum_{\omega \in B} \mathbb{P}(\omega) = \mathbb{P}(A) + \mathbb{P}(B).$$
The key here is that we are not double-counting anything.

5. Recall that $A \cup A^C = \Omega$ and $A \cap A^C = \emptyset$. In particular, since $A \cap A^C = \emptyset$, then
$$\mathbb{P}(A \cup A^C) = \mathbb{P}(A) + \mathbb{P}(A^C).$$
But, since $A \cup A^C = \Omega$ and $\mathbb{P}(\Omega) = 1$, we know that
$$1 = \mathbb{P}(A) + \mathbb{P}(A^C).$$
Therefore, it follows that
$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A^C).$$

This concludes the proof. □

Looking at #4 in the previous theorem, we can actually generalize this.

Theorem 1.2

If $A_1, \dots, A_n$ are pairwise disjoint (i.e. $\bigcap_{i \in [1,n] \subset \mathbb{Z}} A_i = \emptyset$), then
$$P\left(\bigcup_{n=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Proof. One way we can go about this is to take advantage of the fact that $A_1, \dots, A_n$ are pairwise disjoint. We will use induction on n.

- Base Case: Suppose $n = 1$. Trivially, A_1 is pairwise disjoint since it's the only set and so $\mathbb{P}(A_1) = \mathbb{P}(A_1)$. Likewise, $n = 2$ is satisfied by the previous theorem.

- Inductive Step: Suppose that this holds for n. We now want to show that this holds for $n + 1$. To do so, we note that
$$A_1 \cap \dots \cap A_n = \emptyset$$
and
$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \mathbb{P}(A_1) + \dots + \mathbb{P}(A_n).$$
So, we can define $A = A_1 \cup \dots \cup A_n$. We now introduce the set A_{n+1}; suppose that $A_{n+1} \cap A = \emptyset$. Then, it follows that
$$\mathbb{P}(A \cup A_{n+1}) = \mathbb{P}(A) + \mathbb{P}(A_{n+1}) = \mathbb{P}(A_1) + \dots + \mathbb{P}(A_n) + \mathbb{P}(A_{n+1})$$

This concludes the proof. □

Remarks:

- The following consequence of the previous theorem is an extremely useful tool for calculating the probability of events.

- Often, it is difficult to find $\mathbb{P}(E)$ directly, and it is easier to "split the job up" into doable subtasks.

1.9 Law of Total Probability

This brings us to the *Law of Total Probability*.

Theorem 1.3: Law of Total Probability (LoTP)

Let $E \subset \Omega$ be an event, and let $A_1, \ldots, A_n$ be a partition of Ω (that is, a pairwise disjoint collection of sets that "cover" the sample space $\bigcup_{i=1}^{n} A_i = \Omega$). Then, we have that

$$P(E) = \sum_{i=1}^{n} P(E \cap A_i)$$

Proof. Note that E is the pairwise disjoint union of the sets $E \cap A_1, \ldots, E \cap A_n$. Thus, we can just apply the previous theorem. □

Remark: While it might be difficult to find $P(E)$ directly, if you pick the A_i's wisely, it can become easy to find each of the $P(E \cap A_i)$'s.

Corollary 1.1

For any two events A and B,

$$P(A) = P(A \cap B) + P(A \cap B^C)$$

Remark: This holds since B, B^C is a parition of Ω.

Theorem 1.4: I

A and B are subsets of Ω, then

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

Proof. Recall that

$$P(A \cup B) = \sum_{\omega \in A \cup B} m(\omega).$$

Now, if ω is in exactly one of the two sets, then it is only counted once (hence the first and second term on the right-hand side). However, if ω is in both A and B, then we would be double-counting since it's counted for in $P(A)$ and $P(B)$. So, we need to subtract it (hence, the last term on the right-hand side). □

Remark: If $A \cap B = \emptyset$, then $P(A \cap B) = 0$.

1.10 Example Probability Distributions

We now talk about two types of distributions: uniform and geometric distributions.

1.10.1 Uniform Distribution

Definition 1.7: Uniform Distribution

The *uniform distribution* on a finite sample space Ω containing n elements is the function m defined by

$$m(\omega) = \frac{1}{n}$$

for every outcome $\omega \in \Omega$.

For example, when flipping a fair coin, there is only two possiblities: heads or tails. So,

$$m(\text{Heads}) = \frac{1}{2}.$$

A nice property of the uniform distribution is that, for all events $E \subset \Omega$, we simply have that

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|}.$$

1.10.2 Infinite Sample Sizes & Geometric Distribution

The same definitions that we discussed earlier apply the same way in the infinite case. However, notice that the rule

$$\sum_{i=1}^{\infty} \mathbb{P}(\omega_i) = 1$$

means that this infinite series *converges*, and it converges to 1 (it is not just a usual sum). Now, when Ω is countably infinite, we further assume, in the definition of probability distribution, that

$$\mathbb{P}\left(\bigcup_{i\in I} E_i\right) = \sum_{i\in I} \mathbb{P}(E_i)$$

for all (possibly countably infinite) collections of pairwise disjoint sets $\{E_i \mid i \in I\}$.

Now that we know the basics of infinite sample size, we can now consider the **geomtric distribution**[1]

$$P(X = k) = p(1-p)^{k-1}$$

for $k = 1, 2, \ldots$ and $P(X = x) = 0$ for all other x.

(Example: Geometric Distribution.) To see this, suppose a coin flips "Tails" with probability p. Then, the random variable, X is the number of flips until we flip "Tails" for the first time, has this distribution. For example, if we flip "Heads" twice and get "Tails" on the third attempt, then $X = 3$. Indeed, for this to happen on flip k, we need all of the previous $k-1$ flips to be "Heads," and then the next flip to be "Tails."

Thus, the probability that we only get "Tails" on the third attempt (i.e., "Heads" on the first two attempts) is given by

$$\mathbb{P}(X = 3) = (1-p)(1-p)p = (1-p)^2 p = (1-p)^{3-1}p.$$

To check that this is a bonafide probability distribution, we note that

$$\sum_{k=1}^{\infty} P(X = k)$$

[1] This will be discussed in detail later on.

is equal to

$$\sum_{k=1}^{\infty} p(1-p)^{k-1} = p\sum_{k=0}^{\infty}(1-p)^k = \frac{p}{1-(1-p)} = 1.$$

Note that the second-to-last step is from the geometric series

$$\sum_{i=0}^{\infty} \alpha^i = \frac{1}{1-\alpha}$$

for $|\alpha| < 1$.

2 Continuous Probability Distributions

Now, we will discuss the case where there is a continuum of possible values that a RV can take. For example, rather than discrete choices like 1, 2, 3, 4, 5, we will instead he dealing with things like the time until the first customer appears at a store, or the lifetime of a lightbulb.

2.1 Probability Density Function

Recall that, in the discrete case, a random variable's probability mass function (PMF)

$$p_X(x) = \mathbb{P}(X = x)$$

has the property that

$$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x).$$

We now define the analog to the PMF. In particular, we want to find a probability density function f such that

$$\mathbb{P}(X \in A) = \int_A f(x)dx,$$

where $A \subseteq \mathbb{R}$ is some arbitrary region. Note that $f(x)$ is not a probability.

Definition 2.1: Probability Density Function

Let X be a continuous, $\mathbb{R}$-valued random variable. A **probability density function** (PDF) for X is a $\mathbb{R}$-valued, non-negative function f that satisfies

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx$$

for all $a, b \in \mathbb{R}$.

We note that

$$\mathbb{P}(X = x) = \int_x^x f(x)dx = 0$$

for any x. This means that

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b).$$

For example, a uniform random variable on an interval $I \subset \mathbb{R}$ has PDF

$$\boxed{f = \frac{1}{\text{length}(I)}}.$$

2.2 Cumulative Distribution Function

While we have the probability density function, we also have the cumulative distribution function.

Definition 2.2: Cumulative Distribution Function

The **cumulative distribution function** (CDF) of a continuous random variable X is the function given by

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(t)dt.$$

2.3 Relationship Between PDF and CDF

We can relate the PDF and the CDF by the following theorem:

Theorem 2.1

Let X have CDF F_X and PDF f_X. Then, $F'(x) = f(x)$.

Proof. Note that $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$. Hence, by the Fundamental Theorem of Calculus, it follows that

$$F'_X(x) = f_X(x),$$

as desired. □

Important Note 2.1

The PDF is the derivative of the CDF.

(Example.) Suppose we have a dart board with unit radius. Suppose we throw a dart at the target. The sample space is the unit disk

$$D = \{(x, y) \mid x^2 + y^2 \leq 1\}.$$

The unit circle has area $\pi(1)^2 = \pi$. Supposing a dart lands at uniformly random position on the target, we would have the PDF

$$f(x, y) = \frac{1}{\pi}$$

for a random throw (X, Y). Note that, because this is two-dimensional, we consider the *area* as opposed to the length of the interval.

To find the probability of landing in a certain region, we would have to integrate over that region; more specifically, we have to integrate that uniform density $\frac{1}{\pi}$ by that region. For instance, if the "bullseye" region of the target is the center circle B of radius $\frac{1}{5}$, then the probability of getting a "bullseye" would be

$$\frac{\text{area}(B)}{\text{area}(D)} = \frac{\pi(1/5)^2}{\pi} = \frac{1}{25}.$$

So, we should expect approximately 1 in every 25 throws to be a "bullseye."

Now, let D be the distance from the center to the point (X, Y) where a uniformly thrown dart lands. We note that

$$D = \sqrt{X^2 + Y^2} \in [0, 1].$$

What is the distribution of this random variable? We should not expect D to have a uniform distribution. For instance, notice that there are more points at distance $\geq 1/2$ from the center than there are points at distance $\leq 1/2$ from the center. So, we expect

$$\mathbb{P}(D \in [0, 1/2]) < \mathbb{P}(D \in [1/2, 1])$$

although both of these sub-intervals of $[0, 1]$ have the same length. Recall that $f(d) = F'(d)$. Then,

$$F(d) = \mathbb{P}(D \leq d) = \mathbb{P}(X^2 + Y^2 \leq d^2) = \frac{\pi d^2}{\pi} = d^2.$$

Notice here that πd^2 is the area of the *inner* circle. Therefore, $f(d) = 2d$.

(Example Problem.) Let U be a uniform random variable on $[0,1]$, and consider the random variable

$$X = U^2.$$

Find the PDF of X.

We know that X has PDF

$$F_X(x) = \mathbb{P}(X \leq x).$$

We also know that U is a uniform RV on $[0,1]$, so its PDF is given by

$$f_U(u) = \begin{cases} \frac{1}{1-0} = 1 & \text{if } u \in [0,1] \\ 0 & \text{otherwise} \end{cases}.$$

So, we have that

$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}(U^2 \leq x) \\ &= \mathbb{P}(U \leq \sqrt{x}) \\ &= \int_{-\infty}^{\sqrt{x}} f_U(t)dt \\ &= \int_0^{\sqrt{x}} 1dt \\ &= \sqrt{x}. \end{aligned}$$

This tells us that the CDF is

$$F_X(x) = \sqrt{x}.$$

Then, to find the PDF, we can just take the derivative of the CDF, like so:

$$\frac{d}{dx}F_X(x) = \frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}.$$

2.4 Exponential Random Variable

This random variable is useful when studying events occurring at random times. For example, lifetime of a lightbulb, time until the next customer, time until the next earthquake, etc.

Recall that $F(x) = \mathbb{P}(X \leq x)$. Hence,

$$1 - F(x) = \mathbb{P}(X > x).$$

This is sometimes denoted by $S(x) = 1 - F(x)$ and is referred to as the **survival function** of the random variable X. Note that if X is a random time, e.g. the lifetime of a lightbulb, then $S(x)$ is the probability of "surviving" until time x.

A very special type of continuous random variable is the exponential random variable with rate $\lambda > 0$. This is the random variable with survival function

$$S(x) = e^{-\lambda x}.$$

That is, the probability of survival until time X decays exponentially with rate λ. Note that $F(x) = 1 - e^{-\lambda x}$, and so $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $f(x) = 0$ otherwise is its PDF.

2.4.1 Memoryless Property

The exponential RV is very important because it is the only continuous RV with a special property, known as the **memoryless property**. Then, given that an exponential RV has survived until time x, the probability that it survives for y amount of time longer (i.e. until time $x + y$) is the same as the probability of just surviving until time y.

For example, if an exponential lightbulb has survived until time x, then given this, the probability of surviving until time $x + y$ is the same as the probability of a brand new lightbulb.

3 Combinatorics