# Lecture 2: Review of Probability Theory

Theodore Hatch Whitfield

January 29, 2018

## 1 Introduction

In today's lecture, we will review the basics of formal probability theory. Technically, this is supposed to be a prerequisite for the class, so we will move quickly. On the other hand, my experience has been that many people do not have a strong foundation in this material, so we will develop everything from first principles. We'll only develop tools that we'll need for later on, so don't confuse this with a comprehensive course in probability theory.

## 2 What is Probability?

At first glance, the section title seems like a weird question. What do you mean, what is probability? This is something that we use everyday, and we feel that we have strong intuitions about this concept. Yet in fact it's a difficult philosophical problem, and there is no definitive answer to the question. Everyone agrees that probabilities are real numbers and should range from 0 to 1, but the disagreement arises when we attempt to say what these numbers represent, and there are at least two different viewpoints as to their interpretation. The first interpretation is called "frequentist" probability, and holds that the probability of an event is essentially just the long-run frequency of that event, over many replications of an experiment. This is intuitively appealing – if we flip a coin a zillion times, and observe half a zillion heads, surely the probability of getting a heads is 50%, right? The second interpretation is called "subjective" probability, and in this view the numerical values of probabilities serve to quantify our degrees of belief about propositions. Thus, if my degree of belief that it will rain tomorrow is the

same as my degree of belief that it will **not** rain tomorrow, then my degree of belief in the proposition "It will rain tomorrow" gets assigned the value 0.5, and this is a probability. Notice that Anita might have a different degree of belief about this proposition, and therefore assign a different probability to it. In the subjectivist viewpoint, that's fine: everybody has their own private degrees of belief, and therefore everybody has their own private set of probabilities.

This is a fascinating debate, and in fact there is no "answer" to it in the sense that we can definitively rule out one interpretation or the other. Some very intelligent people feel very strongly about the frequentist interpretation, and are hostile to the subjectivist approach, but there are other very intelligent people who feel the opposite. However, there is no question that in the development of statistics, especially during the 20th century, the frequentist interpretation has been the dominant one. It's used in almost all formal research, and it's the approach that's taught in practically all introductory courses. As a result, in this course we will be entirely focused on the frequentist definition of probability.

Before we move on, let's discuss the term "frequency", which tends to be used in two different ways. The first is to report the total number of items of a particular type in a collection. Thus, if we have a collection of 20 balls consisting of 5 red balls and 15 white balls, using this definition of "frequency" the frequency of red balls is 5, because that's the total number of red balls. In this course I'm going to call this concept of frequency an "absolute frequency", but beware that this is one of my made-up terms. The other concept associated with the word "frequency" is that it is the *proportion* of items of a particular type in the collection. Thus, in this usage of the word, since we have 5 red balls and the collection is 20 balls, then the "frequency" of red balls is 0.25 or 25%. In this course I'm going to call this concept of frequency a "relative frequency", and again this terminology is unique to this course.

# 3 The Basic Frequentist Setup

In our class, we will focus on a particular framework or setup. In this framework, we have an experiment with a precisely specified protocol or procedure. This protocol states exactly how the experiment is to be conducted, exactly

how much data is to be collected, and exactly how the data should be analyzed. Most importantly, the protocol is determined *before* the experiment is conducted, and we can't change the protocol after we've done the experiment. In short, just like playing a game, the rules are established prior to starting the activity.

Along with an experimental protocol, there is also an experimental *outcome*. This outcome is something that we observe; it's what we typically refer to as the "data". The set of all possible outcomes is called the *sample space*, and we will denote this by the symbol $\mathcal{S}$.

All of this so far has been a little vague and philosophical, so let's look at some concrete examples. The first example is perhaps the simplest experiment imaginable. Our experimental protocol is to take a coin and toss it in the air, let it land and come to rest, and then observe whether the upwards facing side is a Heads or a Tails. Since the outcome is one of the two results Heads or Tails, the sample space is:

$$\mathcal{S} = \{\text{Heads, Tails}\}$$

OK, this is not a very exciting experiment. For our next example, we will still toss a coin, but this time we will do so 4 times. The outcome is the sequence of observed Heads or Tails, and we will denote a particular observed sequence of coin tosses using square brackets and the letters 'H' and 'T'. Thus, if we get a Heads, then a Tails, then another Tails, and finally a Heads, we will descibe this sequence by [H, T, T, H]. Now the sample space $\mathcal{S}$, which is the set of all possible such sequences, is more complicated:

| | | | |
|---|---|---|---|
| [H, H, H, H] | [H, H, H, T] | [H, H, T, H] | [H, H, T, T] |
| [H, T, H, H] | [H, T, H, T] | [H, T, T, H] | [H, T, T, T] |
| [T, H, H, H] | [T, H, H, T] | [T, H, T, H] | [T, H, T, T] |
| [T, T, H, H] | [T, T, H, T] | [T, T, T, H] | [T, T, T, T] |

For this class, we will focus on relatively simple types of experimental protocols. In the real world, these can be extremely elaborate. For instance, for a major clinical trial, the experimental protocol might run to 50 pages

or so, because it has to specify exactly what data is collected at what time points, the criteria for potential subjects to be included or excluded from the trial, what the specific outcomes are, etc.

# 4 The Kolmogorov Probability Axioms

In 1933, the Russian mathematician A.N. Kolmogorov published a set of axioms that define what a valid concept of probability is. One of the key ideas is that a probability is a function that operates on things called "events". We define an *event* to be a subset of the sample space. Thus, for our 4-toss coin flipping experiment, one possible event might be:

$$\mathcal{E}_1 = \{[T, H, H, T],\ [T, T, H, H],\ [H, H, T, H]\}$$

Notice that the curly braces denote a set. Another possible event would be:

$$\mathcal{E}_2 = \{[H, H, T, T],\ [T, H, H, H],\ [H, T, T, H], [T, T, H, T]\}$$

Notice that the empty set, consisting of no elements, is a subset of the sample space, and thus is a perfectly valid event:

$$\mathcal{E}_3 = \{\} = \emptyset$$

Then a *probability function* is a function that takes an event as input and returns a real number between 0 and 1. We denote the probability of an event by $\Pr(\mathcal{E})$. The probability function has to satisfy certain conditions:

- The probability of any event must be a non-negative real number; that is, for any event $\mathcal{E}$, we must have:

$$\Pr(\mathcal{E}) \geq 0$$

- The probability of the entire sample space has to be equal to 1:

$$\Pr(\mathcal{S}) = 1$$

- If two events $A$ and $B$ are disjoint (i.e. have no elements in common) then the probability of the union of $A$ and $B$ is just the sum of the probabilities of $A$ and $B$:

$$\Pr(A \cup B) \ = \ \Pr(A) + \Pr(B)$$

The first two axioms are important because they place limits on the range of the probability function: it can't be less than 0, the probability of any event must be non-negative, and it can't be greater than 1, because nothing can be larger than the entire sample space. But it's really the third axiom that does a lot of work for us in practice, because it enables us to calculate the probability of a complex event by breaking it into smaller pieces.

Let's see a simple example of how Axiom 3 works. We start with an ancient classical Greek urn, and we place a number of colored balls into the urn:

- Six of the balls are red.
- Five of the balls are white.
- Four of the balls are yellow.
- Three of the balls are blue.
- Two of the balls are green.

Let's make a table of the balls in the urn, along with their relative frequencies:

| Color | Absolute Frequency | Relative Frequency |
|---|---|---|
| Red | 6 | 0.30 |
| White | 5 | 0.25 |
| Yellow | 4 | 0.20 |
| Blue | 3 | 0.15 |
| Green | 2 | 0.10 |

What is the probability that if we draw a ball at random from this urn, it will be either White or Green? A ball can't be both White and Green at the same time, so that the events $A$ = "The ball is White" and $B$ = "The ball is Green" are disjoint. Thus the event that ball is either White or Green is the union of the events $A$ and $B$, and since these are disjoint we have:

$$\begin{aligned} \Pr(A \cup B) &= \Pr(A) + \Pr(B) \\ &= 0.25 + 0.10 \\ &= 0.35 \end{aligned}$$

This should be pretty intuitive, and in fact is perhaps even a little underwhelming. But it does show how to use Axiom 3: we can calculate the probability of complex events by breaking them into disjoint pieces, and then adding up the probabilities of these component parts. Remember, in order to use Axiom 3, all the parts have to be disjoint!

There is an important aspect about our experimental protocol that we haven't specified. We will often be interested in performing multiple draws from the urn. Once we've drawn a ball from the urn, what do we do with it? Do we return the ball to the urn, or do we leave it outside the urn? This might seem like a trivial issue, but in fact it's very important. The point is that if we replace the ball in the urn, then the next time we randomly draw a ball we will be sampling from the same probability function as before. However, if we don't replace the ball, then the next time we sample from the urn the probability function will be subtly different. These two modes of sampling have standard names:

- "Sampling with replacement" means that we replace the item we observed back into the population, so that the probability function remains the same across multiple draws.
- "Sampling without replacement" means that we do **not** replace the item we drew, so that the probability function changes with each draw.

In general, in this course we will always be concerned with sampling with replacement. Sampling without replacement is perfectly valid, but it considerably complicates our calculations, because we have to have a separate probability function for each observation.

# 5 Random Variables

Having discussed the concept of the sample space, I have to admit – we generally won't be thinking much about this. Instead, we will be focused on what are called *random variables*.

## 5.1 Definition

The expression "random variable" is unfortunate, because a random variable is not a "variable" at all – it's a function. Specifically, a random variable

is a function that takes an element in the sample space and returns a real number. The real number can be anything – it can be positive or negative or 0, and it can be greater than 1 (or 1,000,000 for that matter). Also, the values don't have to be unique, and the random variable can map different elements of the sample space to the same real number. All that matters is that the random variable takes each element in the sample space and maps it to some real number.

Let's see a concrete example of a random variable, using our ancient classical Greek urn with the colored balls. We will define a random variable $X$ this way:

- If we observe a red ball, then $X$ takes on the value -2.
- If we observe a white ball, then $X$ takes on the value 7.
- If we observe a yellow ball, then $X$ takes on the value 5.
- If we observe a blue ball, then $X$ takes on the value -4.
- If we observe a green ball, then $X$ takes on the value 7.

We can also specify $X$ using conventional function notation:

$$\begin{aligned} X(\text{Red}) &= -2 \\ X(\text{White}) &= 7 \\ X(\text{Yellow}) &= 5 \\ X(\text{Blue}) &= -4 \\ X(\text{Green}) &= 7 \end{aligned}$$

Let's make a table of the sample outcomes, the values of the random variable, and the associated probabilities:

| Color | $X$ | Probability |
|---|---|---|
| Red | -2 | 0.30 |
| White | 7 | 0.25 |
| Yellow | 5 | 0.20 |
| Blue | -4 | 0.15 |
| Green | 7 | 0.10 |

## 5.2 Random Variables and Events

We've defined *events* as subsets of the sample space, and random variables as functions from the sample space to the real numbers, so you might think that these two concepts are different and there is no connection between them. On the contrary, they are very closely related. If you look back at the table, you can see that if $X$ takes on the value -2, then we must have drawn a red ball from the urn, so that the two statements "$X$ = -2" and "The event Red occurred" are really just two different ways of saying the same thing. Likewise, when $X$ is 5, that must mean that we had drawn a yellow ball from the urn, so the two statements "$X = 5$" and "The event Yellow occurred" are the same. Notice however that we have to be careful with the value of 7, because there are two elements of the sample space that can be mapped to 7, White and Green. So the statement "$X = 7$" is the same as the statement "Either White or Green occurred". But White or Green is an event, because it is a subset of the sample space, so everything is still fine. We can use the notation $X^{-1}(x)$ to denote the event associated with the value $x$, and this event is the set of all elements of the sample space that get mapped to $x$. Thus, $X^{-1}(-2) = \{\text{Red}\}$, while $X^{-1}(7) = \{\text{White} \cup \text{Green}\}$.

Now that we know how to associate events with particular values of a random variable, we can use this to idea to assign probabilities to particular values of that random variable. Recall that the probability function takes events as its input. Thus, we will define the probability that the random variable takes on the particular value $x$ to be the probability of the event associated with the value $x$. That is, we first find the set of all elements $\omega$ in the sample space for which $X(\omega) = x$, we then calculate the probability of this set, and finally we can say that this probability is the probability that the random variable takes on the value $x$. So let's redo the table, this time forgetting about the sample space, and just listing the values of the random variable:

| $x$ | $\Pr(X = x)$ |
|---|---|
| -4 | 0.15 |
| -2 | 0.30 |
| 5 | 0.20 |
| 7 | 0.35 |

You might have noticed some new notation in the header for the second column. The expression $\Pr(X = x)$ might look a little funny, because it

seems to be the probability that something is equal to itself, and surely that's always the case, right? But if you look closely, you'll notice that the first $X$ is uppper-case, and the second $x$ is lower-case, and this actually mean something different. The upper-case $X$ is the random variable, which is a function from the sample space to the real numbers, and the lower-case $x$ represents a specific numerical value. So the expression $\Pr(X = x)$ really means "The probability that the random variable $X$ takes on the specific value $x$".

Also, did you notice how I handled the value 7? There were two elements in the sample space that got mapped to 7, namely White and Green, with respective probabilities 0.25 and 0.10. So we collapsed the two sample space categories into the one category $x = 7$, and we added their probabilities together to obtain $\Pr(X = 7)$.

## 5.3 The Support of a Random Variable

Any random variable will have a set of *realized values*, that is, particular real numbers that some element of the sample space is mapped to. For instance, in our example, the set of realized values of $X$ is $\{-4, -2, 5, 7\}$. We also call the set of realized values the *support* of the random variable, and for this course I will use the very cool notation $\Omega_X$ to denote the support of the random variable $X$. We will often want to perform some type of sum over all the realized values of a random variable $X$, and to do this we will use the notation

$$\sum_{x \in \Omega_X}$$

This just means: "sum up over all the values in the support of $X$ i.e. all the realized values of the random variable $X$". For instance, suppose we sum up the probabilities of the realized values of our example random variable $X$:

$$\begin{aligned} \sum_{x \in \Omega_X} \Pr(X = x) &= \Pr(X = -4) + \Pr(X = -2) + \Pr(X = 5) + \Pr(X = 7) \\ &= 0.15 + 0.30 + 0.20 + 0.35 \\ &= 1.00 \end{aligned}$$

This isn't an accident. The collection of events associated with the realized values of a random variable form a partition of the sample space, in the

sense that the sample space is split up into a collection of disjoint sets. (The sets are disjoint because each element of the sample space will be mapped to precisely one real number by the random variable.) By Axiom 3 of the Kolmogorov Probability Axioms, the sum of the probabilities of these events is equal to the probability of their union, which is the entire sample space, and by Axiom 2 this is 1. So, for any random variable, not just the particular example we've been working with, it must be the case that

$$\sum_{x \in \Omega_X} \Pr(X = x) \;=\; 1$$

## 5.4 Discrete and continuous random variables

There are actually two kinds of random variables: discrete and continuous. A *discrete* random variable is one that has either a finite number of realized values, or the set of non-negative integers as its realized values. The random variable $X$ that we've been looking at in our example with the ancient classical Greek urn is a discrete random variable, because it has only a finite number of realized values: -4, -2, 5, and 7. An example of a discrete random variable with an infinite number of realized values is the geometric distribution, which is a random variable which has all the non-zero integers as its realized values: 0, 1, 2, 3, .... However, we can also have random variables that are defined on a range of real numbers, and these are called *continuous random variables*. For instance, we could have a random variable that has the range $[0, 2]$ as its support, and every real number that is between 0 and 2 is a realized value of that random variable. Here the range of support is limited, but we can also work with unlimited ranges such as 0 to $+\infty$ or even $-\infty$ to $+\infty$.

Continuous random variables have to be handled differently than discrete random variables. With a discrete random variable, each value in the support is assigned a probability, just like we saw in our example. Thus, with a discrete random variable, we could write an expression such as $\Pr(X = 5)$, and this was fine. But with a continuous random variable that has a range of real numbers for its support, this won't work, because we have too many values. Let's consider the example of a random variable that has a constant value over the range $[0, 2]$, and try to directly assign a probability to each value. In our first attempt, we will assign the probability $1/2$ to each point

in the range and see what happens:

$$\begin{aligned}
\Pr(X = 0.1) &= \frac{1}{2} \\
\Pr(X = 0.2) &= \frac{1}{2} \\
\Pr(X = 0.3) &= \frac{1}{2} \\
\Pr(X = 0.4) &= \frac{1}{2} \\
\Pr(X = 0.5) &= \frac{1}{2}
\end{aligned}$$

Uh oh. We've only looked at a few values of the random variable, and already the sum of the probabilities is much larger than 1. This can't be right.

You might think that the problem was that the value that we assigned was too large. Perhaps we shouldn't have used 1/2 – how about 0.1? But that strategy won't work, because we have this sequence of values:

$$\begin{aligned}
\Pr(X = 0.01) &= 0.1 \\
\Pr(X = 0.02) &= 0.1 \\
\Pr(X = 0.03) &= 0.1 \\
&\vdots \\
\Pr(X = 0.5) &= 0.1
\end{aligned}$$

In this sequence, there are 50 values of the random variable, from 0.01 to 0.5, and even though we used a smaller value for the probability, 0.1, the total now becomes 5. So we are even worse off than before! In fact, if we

were patient enough, we could keep on choosing more values of the random variable, and soon the sum would be 10, or 100, or even 1,000,000.

You should be able to convince yourself that if we try to assign a fixed value for the probability to each value of a continuous random variable, even over a finite range such as from 0 to 1, we can choose enough values to make the sum as large as we want. The moral of this story is: we cannot directly assign probabilities to individual values of a continuous random variable.

If we can't directly assign probabilities to individual values of a continuous random variable, then what sort of thing can we assign to these individual values? The answer comes in an analogy from physics. Think of a wire made out of some material, and assume that the wire has one end at the point $x = 0$ and the other end at the point $x = 1$. There are two related quantities: the *density* of the material, and the *mass* of the wire. The mass is something that has an actual weight, but it only makes sense to speak of the mass of a piece of the wire, say from $x = 0.2$ to $x = 0.5$; we can't really speak of the mass of the wire at a specific point, say $x = 0.4$. The density on the other hand is the mass per length, and it's defined at each individual point of the wire. So it's actually the exact opposite of mass, in the sense that we *can* ask what the density of the wire is at the point $x = 0.4$, while it doesn't really make sense to ask what the density of the wire is from $x = 0.2$ to $x = 0.5$. It's customary to denote the density of the wire at a point by the notation $\rho(x)$. Then we can obtain the mass of the interval from $x = 0.2$ to $x = 0.5$ by integrating the density over the interval:

$$\text{Mass} = \int_{0.2}^{0.5} \rho(x) \cdot dx$$

More generally, we obtain the mass of the piece of the wire from $x = a$ to $x = b$ by integrating the density over this integral:

$$\text{Mass} = \int_{a}^{b} \rho(x) \cdot dx$$

What does the wire have to do with continuous random variables? The answer is remarkably simple: we will think of probability as a mass, and to each individual value of the continuous random variable we will assign a *density*, not a probability; this density will be denoted by $f(x)$. Then the

probability of the event that the continuous random variable $X$ takes a value in the interval $(a, b)$ is just the integral of the density function from $a$ to $b$:

$$\Pr(a \leq X \leq b) = \int_a^b f(x) \cdot dx$$

Let's see an example of this approach. Suppose we have this density function:

$$f(x) = e^{-x}, \quad x > 0$$

Note that the density function is defined for all positive real numbers. Then the probability that the random variable $X$ is between $x = 1$ and $x = 2$ is:

$$\begin{aligned}
\Pr(1 \leq X \leq 2) &= \int_1^2 e^{-x} \cdot dx \\
&= -e^{-x} \Big|_1^2 \\
&= e^{-1} - e^{-2} \\
&= 0.36788 - 0.13534 \\
&= 0.23254
\end{aligned}$$

In general there is a strong analogy between discrete and continuous random variables, with two fundamental principles:

- If we have a formula for a discrete random variable where we sum over the support of a random variable, then the corresponding formula for a continuous random variable uses an integration over the support of the random variable instead.
- If we have a formula for a discrete random variable where we use probabilities as weights for a weighted sum, then the corresponding formula for a continuous random variable uses densities instead.

Here's a simple example of how this works. Previously, we saw the formula

$$\sum_{x \in \Omega_X} \Pr(X = x) \;=\; 1$$

This was for discrete random variables; we know this because it uses a sum over the support, and also because it directly assigns a probability to the value $x$. For the version for a continuous random variable, we replace the sum with an integral, and the probability mass function with a probability density:

$$\int_{\Omega_X} f_X(x) \cdot dx \;=\; 1$$

By the way, did you notice the $X$ subscript on the density function $f_X(x)$? This can often be useful, because we might be working with more than one random variable at a time, and it helps us to keep track of which density function is associated with which random variable. It's entirely conventional, and I promise you I haven't made it up, so you should always do it. In fact, the notation is so well-established that it's typical to use it even when there is only one random variable involved, and I've done that here.

Can any function serve as a probability density function, or does it have to satisfy certain conditions? After all, with the Kolmogorov probability axioms, there are certain requirements that had to be satistified in order for a function to be a valid probability function. As it turns out, this is also true for probability density functions. Since density functions are not actual probability functions, they don't have to satisfy the requirements for a probability function. On the other hand, density functions give rise to probabilities generated by the integral formula

$$\Pr(a \leq X \leq b) = \int_a^b f(x) \cdot dx$$

Because the density function is used to compute the probability, the requirements for a density function have to be such that the resulting values are legitimate probabilities that satisfy the requirements for a probability function.

The first requirement for a probability was that it could not be negative. Now if a density function were negative, even for a short interval, we could

integrate over that interval to obtain a negative probability. Since this would violate the requirements for a probability, we can't allow this to happen. Thus, in order to insure that all probabilities are non-negative, we have to require density functions to be non-negative for all values of $x$.

The second requirement for a probability is that the probability of the entire sample space must be equal to 1. In terms of densities, that means that if we integrate over all possible values of the continuous random variable, we have to get 1:

$$\int_{\Omega_X} f(x) \cdot dx = 1$$

The strange notation at the bottom of the integral is just a way of saying "perform the integration over all the values of the continuous random variable $X$", and in any concrete problem we would just perform the integration with the particular values of the specific random variable we were working with.

Instead of always having to say "all the values of the continuous random variable", which is cumbersome, we can speak of the *support* of the random variable, and this is just the set of values of the random variable. So we could express the second condition on a density as saying that the integral of the density over the support of the random variable must be equal to 1.

Let's take a look at a couple of examples. First, how about the function

$$f(x) = \frac{18}{(x+3)^3}, \quad x > 0$$

Is this a valid probability density? To find out, we need to integrate it over the full range of its support, which in this case is the set of all positive real numbers $x > 0$:

$$\begin{aligned} \int_0^\infty \frac{18}{(x+3)^3} \cdot dx &= -\frac{9}{(x+3)^2} \Bigg|_0^\infty \\ &= -0 + \left(\frac{9}{3^2}\right) \\ &= 1 \end{aligned}$$

So $f(x)$ does indeed integrate to 1, and since it is non-negative for all positive real values of $x > 0$ it is a valid probability density.

Now let's consider the function

$$g(x) = e^{-3x}, \quad x > 0$$

Is this a valid probability density? To find out, we integrate it over its support, which once again is the set of all positive real numbers $x > 0$:

$$\begin{aligned} \int_0^\infty e^{-3x} \cdot dx &= -\frac{1}{3} \cdot e^{-3x} \Big|_0^\infty \\ &= -0 - \left(-\frac{1}{3}\right) \\ &= \frac{1}{3} \end{aligned}$$

Since $g(x)$ does not integrate to 1, it is not a valid probability density.

The third requirement for a valid probability function is that if $A$ and $B$ are two disjoint events, the probability of their union is just the sum of the probabilities of the individual events:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

This is automatically satisfied by the properties of the integral. For instance, suppose $A$ is the event that the random variable $X$ is between 1 and 2, and $B$ is the event that $X$ is between 3 and 4. Since these intervals don't overlap, the events $A$ and $B$ are disjoint, and thus

$$\begin{aligned} \Pr(A \cup B) &= \int_1^2 f(x) \cdot dx + \int_3^4 f(x) \cdot dx \\ &= \Pr(A) + \Pr(B) \end{aligned}$$

So, because of the way that integrals work, the third condition for a valid probability function is automatically satisfied, and we don't need to place any additional requirements on the density to achieve this.

To summarize, a valid probability density must satisfy two requirements:

1. First, it must be non-negative for all the values in its support.

2. Second, the integral of the density over the support must be equal to 1.

As long as these conditions are satisfied, the density function will always give rise to a valid probability function.

Finally, it is conventional to write density functions with a subscript to remind us which random variable they are associated with. Thus, we typically denote the density function for the random variable $X$ by $f_X(x)$.

### 5.5 An Important Point

Before we move on, let me make a very important point. We've seen how to associate random variables with probabilities, so that even though probabilities are, properly speaking, assigned to events in the sample space, in practice we can treat them as being assigned to realized values in the support of the random variable. This raises the interesting question as to whether or not we really need the sample space or not – couldn't we just work directly with the random variable and its associated probabilities, and forget about the sample space entirely? The answer is that for theory to be fully general, we do need to think about the sample space, but in practice we can indeed dispense with the sample space and work with the random variable directly. In fact, that's basically what we will do for the rest of the course (with perhaps a few exceptions), and for the rest of this lecture we will usually work directly with random variables.

## 6 Three Important Functions

There are three important functions associated with any random variable, discrete or continuous.

### 6.1 The Cumulative Probability Function

The first important function is commonly called the *cumulative distribution function*, or often more simply the *distribution* function, and denoted $F_X(x)$. The word "cumulative" essentially means "so far"; when people talk about their "cumulative GPA", they mean their GPA over all the courses they've taken so far in their college career. Thus, the cumulative distribution function

$F_X(x)$ is defined as the probability "so far", that is, the probability that the random variable is less than or equal to the particular value $x$:

$$F_X(x) \;=\; \Pr(X \leq x)$$

For a discrete random variable, the cumulative distribution function is computed by adding up all the probabilies less than or equal to $x$:

$$F_X(x) = \sum_{k \in \Omega_X}^{x} \Pr(X = k)$$

For a continuous random variable, we use our two principles: the sum is converted to an integral, and the probabilities are converted to densities:

$$F_X(x) = \int_{-\infty}^{x} f_X(s) \cdot ds$$

Remember that $F_X(x)$ is defined to be the probability that the random variable is less than **or equal to** the particular value $x$. For discrete random variables, this is very important! For instance, if a random variable has non-zero probability for the number $x = 2$, then the probability $\Pr(X < 2)$ is different from the probability $\Pr(X \leq 2)$, because the latter probability includes the probability that $X = 2$. So for discrete random variables, you must always remember, and although it might seem like a technical point, it really does make a difference in computations. For a continuous random variable, this doesn't matter, because the difference between $\Pr(X < 2)$ and $\Pr(X \leq 2)$ is simple $\Pr(X = 2)$, and as we've seen this will always be 0. Thus, for a continuous random variable, $\Pr(X < 2) = \Pr(X \leq 2)$, and we don't have to worry about this issue.

I mentioned at the beginning that the function $F_X(x)$ is conventionally called the "cumulative distribution function", or even the "distribution function" for short. I am not a fan of this, because it doesn't really describe what the function is. Also, I want to use the word "distribution" to describe the whole package of a random variable and its associated probability mass or density function, and calling something the "cumulative distribution function" just makes things confusing. It's even worse with the term "distribution function", which is totally unclear. What is that? So, for this course, I'm going to call this the "cumulative probability function", which I think is much easier to understand. However, I must warn you that this is non-standard, and it is contrary to a very well-established convention.

## 6.2 The Survival Probability Function

The *survival probability function*, or the shorter version *survival function*, is denoted by the symbol $S_X(x)$, and is defined as $S_X(x) = \Pr(X > x)$. It's not hard to prove that

$$F_X(x) + S_X(x) = 1$$

We use this identity to obtain the common computation formula for the survival function:

$$S_X(x) = 1 - F_X(x)$$

That is, to calculate the survival function at a point $x$, calculate the cumulative probability function $F_X(x)$ and then subtract from 1. Once again, pay attention to the inequality sign here. Since $F_X(x)$ is defined as the probability of the event $X \leq x$, then $S_X(x)$ must be defined as the probability of the event $X > x$ in order for the two to add up to one. In other words, the survival probability function does **not** include the event $X = x$. For continuous random variables this doesn't make a differnce, but for discrete random variables it's an important distinction.

## 6.3 The Quantile Function

The *quantile function*, denoted $Q_X(p)$, is a little subtle. Recall that for the cumulative probability function, we take a value of $x$ as the input, and the function returns the total cumulative probability $\Pr(X \leq x)$ i.e. all the probability "up to" the point $x$. The quantile function $Q_X 9x)$ does the opposite: we are given a cumulative probability $p$ as the input, and the function returns the value of $x$ which will give the specified cumulative probability. For instance, suppose for some random variable the value 7.2 has a cumulative probability of 0.65, so that $F_X(7.2) = 0.65$. Then the quantile function for the probability 0.65 is the point $x = 7.2$: $Q_X(0.65) = 7.2$. Thus, we can think of the quantile function as being the inverse of the cumulative probability function:

$$Q_X(F_X(x)) = x$$

$$F_X(Q_X(p)) = p$$

Well ... kinda sorta. In general, for continuous random variables, these two equations are true, and there is no difficulty. The problem is with discrete

random variables, because the probability mass function comes in "lumps". So, we might have a situation where $X$ has the integers as support and $F_X(2) = 0.47$, and $F_X(3) = 0.61$. Now what's $Q_X(0.50)$? There is no value of $x$ for which the cumulative probability function is exactly 0.50, so in a strict sense the quantile function should really be considered as undefined for this probability. In practice, this is too restrictive, and some sort of interpolation scheme is typically employed, but here the inverse relation with the cumulative probability function might break down.

# 7 Independence

Suppose we have two events $A$ and $B$, and a probability function $\Pr(X)$. Then the probabilities of the events $\Pr(A)$ and $\Pr(B)$ are well-defined. Also, since events are subsets of the sample space, they are sets, and we can perform set operations on them. Consider the event $A \cap B$: this is the *intersection* of $A$ and $B$, which means that it is the set consisting of all the elements that are members of both $A$ and $B$. It's a subset of the sample space, so it has a probability $\Pr(A \cap B)$. Can we say anything about the probability of $A \cap B$ in terms of $A$ and $B$? In general, the answer is no. Let's consider a standard die, and define two events:

- $A$ is the event that the value on the die is less than or equal to 3.
- $B$ is the event that the value on the die is a prime number.

There are three faces on the die that are less than 3, out of a total of 6 faces, and since by hypothesis the die is fair, the probability of event $A$ is $3/6 = 0.5$. Likewise, there are three prime numbers less than or equal to 6 (2, 3, and 5), so the probability of $B$ is also $3/6 = 0.5$. Thus:

$$\Pr(A) \cdot \Pr(B) = 0.50 \times 0.50 = 0.25$$

Now what is the event $A \cap B$? It's the event that the value on the die is both less than or equal to 3 and also a prime number. There are two possibilities here: the values 2 and 3. Thus, the probability of $A/capB$ is $2/6 = 1/3$. So we have:

$$\Pr(A \cap B) = \frac{1}{3} \neq 0.25 = \Pr(A) \cdot \Pr(B)$$

We've defined independence of events in a strictly mathematical form: it's just the condition that for two events $A$ and $B$ we have:

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

However, it would be nice to have some intuitive mental models for this concept. One way to think about independence is that event $A$ has no influence over event $B$, and vice versa. For instance, if we flip a coin once, and then a second time, the first coin flip does not influence the outcome of the second coin flip. Therefore, it makes sense to treat these two coin flips as independent events. The second way to think about independence is in terms of information: by knowing about the first flip, do we gain any insight into the second flip? Again, assuming that you know that the coin is fair, then the answer is no, you can't get any information about the second flip from the first flip. Therefore, because the first flip doesn't give you any information about the second flip, the two events are independent.

So far, we've defined independence in terms of events. But we can extend this definition to random variables; let's start by working with discrete random variables. Remember that every value in the support of a random variable implicitly defines an event, and thus can be assigned a probability. That is, if the value 1 is in the support of $X$, then $X = 1$ describes an event, which is the set of elements $\omega$ in the sample space for which $X(\omega) = 1$. Then the probability of $X = 1$ is just the probability of this event, and we can write $\Pr(X = 1)$. Likewise, let $Y$ be another random, and suppose the support of $Y$ contains the value 3; then $\Pr(Y = 3)$ denotes the probability of the event consisting of those elements $\omega$ in the sample space for which $Y = 3$. Finally, we also have the event $(X = 1) \cap (Y = 3)$, which is just the set of those elements *omega* in the sample space for which both $X(\omega) = 1$ and also $Y(\omega) = 3$. Then independence holds if:

$$\Pr(X = 1 \ \cap \ Y = 3) = \Pr(X = 1) \cdot \Pr(Y = 3)$$

That's independence for the events $X = 1$ and $Y = 3$. How about if the value 2 is also in the support of $X$? Then independence would hold for the events $X = 2$ and $Y = 3$ if:

$$\Pr(X = 2 \ \cap \ Y = 3) = \Pr(X = 2) \cdot \Pr(Y = 3)$$

Likewise, suppose the value 4 is in the support of $Y$? Then the events $X = 1$ and $Y = 4$ are independent if:

$$\Pr(X = 1 \ \cap \ Y = 4) = \Pr(X = 1) \cdot \Pr(Y = 4)$$

Now we can define the concept of two random variables being independent: two random variables $X$ and $Y$ are independent if for every realized value $x$ in the support of $X$ and every realized value $y$ in the support of $Y$ we have:

$$\Pr(X = x \ \cap \ Y = y) \ = \ \Pr(X = x) \cdot \Pr(Y = y)$$

This definition is a little verbose; all it's really saying is to take all the possible values of $X$ and all the possible values of $Y$, form all possible pairs of these values, and then make sure that the independence condition holds for all of these.

We've conducted this whole discussion in terms of discrete random variables and their probabilities; what about continuous random variables? As you might imagine, the definition of independence is analogous, except that it is now expressed in terms of densities rather than probabilities:

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$$

# 8 Moments (and the Variance)

## 8.1 The Expected Value $\mathbf{E}[X]$

The *expectation* or *expected value* of a discrete random variable $X$, denoted $\mathrm{E}[X]$, is defined as:

$$\mathrm{E}[X] = \sum_{x \in \Omega_X} x \cdot \Pr(X = x)$$

We can think of the expected value of a random variable as a weighted sum of all the realized values in the support of $X$, where the weights are the corresponding probabilities. Thus, values that occur with relatively high frequency (i.e. have a relatively high probability) are weighted more heavily that values that occur with a low frequency (i.e. have a relatively low probability).

Let's calculate the expected value of our example random variable from the ancient classical Greek urn. For discrete random variables with a small

number of realized values, it can be helpful to use a tabular form for the calculation:

| $x$ | $\Pr(X = x)$ | $x \cdot \Pr(X = x)$ |
|---|---|---|
| -4 | 0.15 | -0.60 |
| -2 | 0.30 | -0.60 |
| 5 | 0.20 | 1.00 |
| 7 | 0.35 | 2.45 |
| | Total | 2.25 |

So the expected value of this random variable is $\mathrm{E}[X] = 2.25$.

If $X$ is a continuous random variable, we still denote the expectation of $X$ by $\mathrm{E}[X]$, but now we modify the computational procedure in two ways:

- Instead of summing over all the values of the support, we integrate over the values of the support.
- Instead of weighting with probabilities, we weight with densities.

So now the computational formula is:

$$\mathrm{E}[X] = \int_{\Omega_X} x \cdot f_X(x) \cdot dx$$

There's a catch here: the integral might not exist. That might seem impossible, especially since the expected value is what we call the "average". How can a random variable not have an "average value"? The standard example of this is the *Cauchy* distribution, which has this density function:

$$f_X(x) = \frac{1}{\pi \cdot (1 + x^2)}, \quad -\infty < x < +\infty$$

This seems incredible, because the function is symmetric about the $y$-axis, and so surely the negative values should cancel out the positive values and we should end up with an expected value of 0, right? And in fact if we construct the integral for the expected value, we have:

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\pi \cdot (1 + x^2)} \cdot dx \\ &= \int_{-\infty}^{+\infty} \frac{x}{\pi \cdot (1 + x^2)} \cdot dx \end{aligned}$$

Now this seems to confirm our intuitions. The integrand here is an *odd* function, which means that $f(-x) = -f(x)$, and we are integrating over the symmetric range $-\infty$ to $+\infty$, and a standard result in integration theory holds that the integral of an odd function over a symmetric range is necessarily 0. So what's the problem? The problem is that the limits of integration are not numbers like 0 or 27.6, but are $-\infty$ and $+\infty$, and these are shorthand for a special limiting process. This type of integral is called a *doubly improper Riemann integral*, and it can be shown that, when the theory of Riemann doubly improper interals is used, this integral is undefined. So, the expected value doesn't exist. If this all seems too technical for you, then don't worry about it, and just remember that the integral for the expected value of a random variable is not guaranteed to exist, and there are such things as random variables that do not have an expected value. On the other hand, if you're interested in the details, see the appendix.

## 8.2 The Law of the Unconscious Statistician (LOTUS)

The *Law of the Unconsious Statistician* is a result that is usually taken for granted, hence the use of the word "unconscious". It's very important, and we use it all the time, but it really does need to be justified. The name "Law of the Unconscious Statistician" is unwieldy, but it does have the delightful acronym LOTUS, so that's nice.

Suppose we have a discrete random variable $X$, and some function $g(X)$. What is the expected value of this function of $X$? What everybody does is this:

$$\mathrm{E}[g(X)] = \sum_{x \in \Omega_X} g(x) \cdot \Pr(X = x)$$

It turns out that this is OK, but we do need to justify this. To see what the problem is, remember that a random variable is really a function from the sample space to the real numbers, so we should properly write $X(\omega)$ to make this clear. Now suppose we apply a function $g$ to $X$; this is also a random variable, because $X$ maps elements in the sample space to the real numbers, and then $g$ maps this to another real number, so the composite function $g(X(\omega))$ maps the element in the sample space to a real number, and that is just the definition of a random variable. Let's call this random variable $Y = g(X(\omega))$; then by the definition of the expected value of a

random variable, we have:

$$\mathrm{E}[Y] = \sum_{y \in \Omega_Y} y \cdot \Pr(Y = y)$$

Do you see what the problem is here? $Y$ is the same thing as $g(X)$, but we have two seemingly different expressions for the expected value of this random variable. Is the right formula this:

$$\mathrm{E}[Y] = \sum_{x \in \Omega_X} g(x) \cdot \Pr(X = x)$$

Or is it this?

$$\mathrm{E}[Y] = \sum_{y \in \Omega_Y} y \cdot \Pr(Y = y)$$

Who's in charge here?

I think the best to see the answer is to work through an example; the formal proof is a little technical, and it's hard to obtain a good grasp of what's going on. Let's go back to our ancient classical Greek urn with the colored balls, and define a new random variable:

| Color | $X$ | Probability |
|---|---|---|
| Red | -5 | 0.30 |
| White | -2 | 0.25 |
| Yellow | 0 | 0.20 |
| Blue | 2 | 0.15 |
| Green | 5 | 0.10 |

Now define the random $Y = X^2$, so that $g(x) = x^2$. Now we have:

| Color | $X$ | $Y$ | Probability |
|---|---|---|---|
| Red | -5 | 25 | 0.30 |
| White | -2 | 4 | 0.25 |
| Yellow | 0 | 0 | 0.20 |
| Blue | 2 | 4 | 0.15 |
| Green | 5 | 25 | 0.10 |

Now let's calculate the expected value of $Y$ using the formula:

$$\mathrm{E}[Y] = \sum_{x \in \Omega_X} g(x) \cdot \Pr(X = x)$$

That is, we take each realized value of $X$, apply the function $g(x)$ to it (i.e. square it), weight it by the probability of that value of $X$, and finally add everything up. Here's the calculation:

| Color | $X$ | $Y$ | Probability | $Y\times$ Probability |
|---|---|---|---|---|
| Red | -5 | 25 | 0.30 | 7.5 |
| White | -2 | 4 | 0.25 | 1.0 |
| Yellow | 0 | 0 | 0.20 | 0.0 |
| Blue | 2 | 4 | 0.15 | 0.6 |
| Green | 5 | 25 | 0.10 | 2.5 |
| | | | **Total** | 11.6 |

So using this method, we find that the expected value of $Y$ is $\mathrm{E}[Y] = 11.6$.

Now let's use the other approach:

$$\mathrm{E}[Y] = \sum_{y \in \Omega_Y} y \cdot \Pr(Y = y)$$

In this formula, we work directly with the random variable $Y$, so let's construct the table for $Y$:

| Color | $Y$ | Probability | $Y\times$ Probability |
|---|---|---|---|
| Red, Green | 25 | 0.40 | 10.0 |
| White, Blue | 4 | 0.40 | 1.6 |
| Yellow | 0 | 0.20 | 0.0 |
| | | **Total** | 11.6 |

Look at that! We got the same answer as before.

Why did this work out? Let's take a look at the colors Red and Green. When we did the first calculation, we took the value of $X$ for Red, which was -5, squared it, and multiplied by the probability of Red, which was 0.30. Then we took the value of $X$ for Green, which was +5, squared that, multiplied by the probability of Green, which was 0.10, and then added them together. So we ended up with a computation like this:

$$((-5)^2 \times 0.3) \; + \; (5^2 \times 0.1)$$

However, when we did the calculation with $Y$ directly, we combined the two probabilities for Red and Green, because $g(X)$ gave us the same value of $Y$,

and we ended up with $\Pr(Y = 25) = 0.3 + 0.1$. Thus, we ended up with a calculation of the form:

$$25 \times (0.3 + 0.1)$$

You should be able to see that these two expression are the same.

This calculation indicates the reason why the Law of the Unconscious Statistician holds. Suppose there are two realized values $x_1$ and $x_2$ of $X$ such that $g(x_1) = g(x_2)$ i.e. g(x) maps the two distinct $x$ values to the same $y$ value. Then we could weight $g(x_1)$ with the probability $\Pr(X = x_1)$ and $g(x_2)$ with the probability $\Pr(X = x_2)$. Or we could work directly with $Y$, in which case there is just one realized value, $y = g(x_1) = g(x_2)$, and the probability $\Pr(Y = y)$ is just the sum of $\Pr(X = x_1)$ and $\Pr(X = x_2)$. In both cases we end up with the same value. Of course if $g(x)$ maps some element $x$ to a unique $y$ value, then $\Pr(X = x)$ is the same as $\Pr(Y = g(x))$, so the calculation of the weighted sum is exactly the same.

We've done all our calculations in this section using a discrete random variable. However, the analogous result also holds for continuous random variables. That is, if we have a random variable $X$ and a function $g(X)$, then the expected value of the function can be calculated as:

$$\mathrm{E}[g(X)] = \int_{\Omega_X} g(x) \cdot f_X(x) \cdot dx$$

As I mentioned, the formal proof of this result isn't that instructive. But in case you're interested, I've included it in the appendix.

## 8.3 Three Important Properties of the Expected Value

The expectation of a random variable $X$ has three important properties.

The first of the important properties is that if we add a constant to $X$, then the expected value of this new random variable is just the expectation

of $X$ plus the constant. To see this, let $Y = X + c$. Then

$$\begin{aligned}
\mathrm{E}[Y] &= \sum_{x \in \Omega_X} Y(x) \cdot \Pr(X = x) \\
&= \sum_{x \in \Omega_X} (x + c) \cdot \Pr(X = x) \\
&= \sum_{x \in \Omega_X} x \cdot \Pr(X = x) \; + \; \sum_{x \in \Omega_X} c \cdot \Pr(X = x) \\
&= \sum_{x \in \Omega_X} x \cdot \Pr(X = x) \; + \; c \cdot \sum_{x \in \Omega_X} \Pr(X = x) \\
&= \mathrm{E}[X] + c \cdot 1 \\
&= \mathrm{E}[X] + c
\end{aligned}$$

So when we created a new random variable $Y = X + c$ by adding a constant $c$ to $X$, the expected value of $Y$ is just the expected value of $X$ plus the constant:

$$\mathrm{E}[Y] = \mathrm{E}[X] + c$$

This is a geometrically intuitive result: when we add a constant to a random variable, we are just shifting everything over, and it's natural that the "average value" should shift over by the same amount.

By the way, in that long derivation, did you notice that we used LOTUS in the first line?

The second result on expected values is similar to the first. If $X$ is a random variable and we now create a new random variable $W$ by multiplying $X$ by a constant, then the expected value of $W$ is just the expected value of $X$ multiplied by the constant. This is even easier to prove than the first

result. Let $W = c \cdot X$. Then:

$$\begin{aligned} \mathrm{E}[W] &= \sum_{x \in \Omega_X} W(x) \cdot \Pr(X = x) \\ &= \sum_{x \in \Omega_X} (c \cdot x) \cdot \Pr(X = x) \\ &= c \cdot \sum_{x \in \Omega_X} x \cdot \Pr(X = x) \\ &= c \cdot \mathrm{E}[X] \end{aligned}$$

Thus when we created a new random variable $W = c \cdot X$ by multiplying $X$ by a constant $c$, the expected value of $W$ is just the expected value of $X$ times the constant $c$:

$$\mathrm{E}[W] = c \cdot \mathrm{E}[X]$$

Did you notice that we used LOTUS in the first line of this derivation as well?

We can put these two results together. Let $X$ be a random variable, and let $a$ and $b$ be constants. Then:

$$\begin{aligned} \mathrm{E}[aX + b] &= \mathrm{E}[aX] + b \\ &= a \cdot \mathrm{E}[X] + b \end{aligned}$$

The third important result is that if have two random variable $U$ and $V$, and we add them together to obtain a new random variable $S = U + V$, then the expected value of the sum $S$ is equal to the sum of the expected values $\mathrm{E}[U]$ and $\mathrm{E}[V]$:

$$\mathrm{E}[S] = \mathrm{E}[U] + \mathrm{E}[V]$$

The expected value of a function of two random variables is defined similarly to our previous definition for one random variable:

$$\mathrm{E}[f(U, V)] = \sum_{u \in \Omega_U} \sum_{v \in \Omega_V} f(u, v) \cdot \Pr(U = u \textbf{ and } V = v)$$

## 8.4 Independence and Expectation

If $X$ and $Y$ are independent, then

$$\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$$

We haven't worked with two variables together so far, but the ideas are similar to what we've seen for a single variable. Recall that we calculated the expected value of a random variable by computing a weighted sum over all the realized values in the support of the random variable, where the weights were the probabilities:

$$\mathrm{E}[X] = \sum_{x \in \Omega_X} x \cdot \Pr(X = x)$$

For two discrete random variables $X$ and $Y$, we do the same thing: we form a weighted sum of the product of $X$ and $Y$, summing over all the realized values in the supports of both random variables, and we weight this by the joint probability of the two random variables:

$$\mathrm{E}[X \cdot Y] = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} xy \cdot \Pr(X = x \textbf{ and } Y = y)$$

Since by hypothesis $X$ and $Y$ are independent, we can re-write the joint probability $\Pr(X = x \textbf{ annd } Y = y)$ as the product of the two marginal probabilities $\Pr(X = x)$ and $\Pr(Y = y)$:

$$\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} xy \cdot \Pr(X = x \textbf{ and } Y = y) \;=\; \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} xy \cdot \Pr(X = x) \cdot \Pr(Y = y)$$

Now we can factor the double summation into a product of two sums over the individual random variables:

$$\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} xy \cdot \Pr(X = x \textbf{ and } Y = y) \;=\; \left( \sum_{x \in \Omega_X} x \cdot \Pr(X = x) \right) \cdot \left( \sum_{y \in \Omega_Y} y \cdot \Pr(Y = y) \right)$$

But now the expressions inside the parentheses are just the formulas for the expected values of $X$ and $Y$:

$$\left( \sum_{x \in \Omega_X} x \cdot \Pr(X = x) \right) \cdot \left( \sum_{y \in \Omega_Y} y \cdot \Pr(Y = y) \right) \;=\; \mathrm{E}[X] \cdot \mathrm{E}[Y]$$

Putting this all together, we have:

$$\begin{aligned}
\mathrm{E}[X \cdot Y] &= \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} xy \cdot \Pr(X = x \textbf{ and } Y = y) \\
&= \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} xy \cdot \Pr(X = x) \cdot \Pr(Y = y) \\
&= \left( \sum_{x \in \Omega_X} x \cdot \Pr(X = x) \right) \cdot \left( \sum_{y \in \Omega_Y} y \cdot \Pr(Y = y) \right) \\
&= \mathrm{E}[X] \cdot \mathrm{E}[Y]
\end{aligned}$$

We just proved our result for discrete random variables; how about for continuous random variables? Of course, it's basically the same, as long as we replace sums by integrals and probabilities by densities:

$$\begin{aligned}
\mathrm{E}[X \cdot Y] &= \int_{\Omega_X} \int_{\Omega_Y} xy \cdot f_{XY}(x, y) \cdot dx \; dy \\
&= \int_{\Omega_X} \int_{\Omega_Y} xy \cdot f_X(x) \; f_Y(y) \cdot dx \; dy \\
&= \left( \int_{\Omega_X} x \cdot f_X(x) \cdot dx \right) \cdot \left( \int_{\Omega_Y} y \cdot f_Y(y) \cdot dy \right) \\
&= \mathrm{E}[X] \cdot \mathrm{E}[Y]
\end{aligned}$$

## 8.5 The General $k$th Moment $\mathbf{E}[X^k]$

So far we've been working with the expected value or expectation of a random variable, yet the title of this section was "Moments (and the Variance)". We'll get to the variance in a little bit, but what's up with this "moments" business? What's that all about?

First, let's have a formal definition. Let $X$ be a random variable. Then the *$k$th moment* is defined to be the expected value of $X^k$ i.e. the general $k$th moment is just $\mathrm{E}[X^k]$. The computational formula for discrete random variables is:

$$\mathrm{E}[X^k] = \sum_{x \in \Omega_X} x^k \cdot \Pr(X = x)$$

For continuous random variables the formula is:

$$\mathrm{E}[X^k] = \int_{\Omega_X} x^k \cdot f_X(x) \cdot dx$$

Hey!! Did you notice that we're using the Law of the Unconscious Statistician here?

Notice that the first moment is just the expected value:

$$\mathrm{E}[X^1] = \mathrm{E}[X]$$

In practice we are typically interested in the first and second moments of a random variable, and that's what we will focus on in this course.

Usually, we won't be able to find a nice general formula for the general $k$th moment, but in some special cases this is possible.

## 8.6 The Variance

Any discussion of the moments of a random variable will inevitably bring us to the concept of the variance of a random variable. Modern statistical theory is fundamentally concerned with the variance, and many of the more sophisticated aspects of the theory are based on this concept. Indeed, one of the most powerful of all statistical techniques is called ANOVA, which is short for the ANalysis Of VAriance.

Here's the definition of the variance of a random variable $X$:

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$$

Technically speaking, the variance is not really a moment, because a moment is defined to be something of the form $\mathrm{E}[X^k]$. However, the variance is very closely related to moments, and so often it's convenient to be a little sloppy and think of it as a moment – perhaps it's a "kinda sorta" moment.

Often it's much more convenient to use an alternative computation formula. We can derive this computational formula by multiplying out the

square, and then applying the linearity of the expectation operator. Note that $\mathrm{E}[X]$ is just a number, and so it's a constant. Here we go:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2 - 2\mathrm{E}[X] \cdot X + (\mathrm{E}[X])^2\right] \\
&= \mathrm{E}[X^2] - \mathrm{E}[2 \cdot \mathrm{E}[X] \cdot X] + \mathrm{E}[(\mathrm{E}[X])^2] \\
&= \mathrm{E}[X^2] - 2 \cdot \mathrm{E}[X] \cdot \mathrm{E}[X] + (\mathrm{E}[X])^2 \\
&= \mathrm{E}[X^2] - 2 \cdot (\mathrm{E}[X])^2 + (\mathrm{E}[X])^2 \\
&= \mathrm{E}[X^2] - (\mathrm{E}[X])^2
\end{aligned}$$

So, just to summarize, we have:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

So you can see why it's so tempting to think of the variance as some sort of moment – it's just a simple function of the first and second moments. Also, you will find that it's often much easier to compute the first and second moments and then use this formula than to use the original definition of the variance, and in fact this will usually be the approach that we employ in this course to calculate variances.

This result is very important, and not just because it's computationally useful – it's also an excellent demonstration of how to work with the linearity of the expectation operator. I encourage you to learn this derivation, so that you can do it without looking at any notes. Don't memorize the sequence of symbols – if you really understand the ideas that underly the various steps, you won't need to memorize anything.

There are two important results concerning the variance. The first is that if we add a constant to a random variable, that doesn't affect the variance. Suppose $Y = X + c$. Then by the linearity of the expectation operator we have:

$$\mathrm{E}[Y] \;=\; \mathrm{E}[X + c] \;=\; \mathrm{E}[X] + c$$

Now let's calculate the variance of $Y$; in this case it's actually easiest to go back to the original definition:

$$\begin{aligned}
\mathrm{Var}[Y] &= \mathrm{E}[(Y - \mathrm{E}[Y])^2] \\
&= \mathrm{E}[((X + c) - \mathrm{E}[X + c])^2] \\
&= \mathrm{E}[(X + c - (\mathrm{E}[X] + c))^2] \\
&= \mathrm{E}[(X + c - \mathrm{E}[X] - c)^2] \\
&= \mathrm{E}[(X - \mathrm{E}[X])^2] \\
&= \mathrm{Var}[X]
\end{aligned}$$

So even though we added a constant onto $X$, resulting in the new random variable $Y = X + c$, we still had $\mathrm{Var}[Y] = \mathrm{Var}[X]$, and thus the variance remained unchanged. Intuitively, this should make sense: the variance of a random variable $X$ measures how much $X$ is spread out, and if we add a constant to $X$ then we just shift everything, but the spread remains the same.

The second important result concerning the variance is that if we multiply a random variable by a constant $a$, then the variance is multiplied by the *square* of that constant. Suppose $Y = aX$. Then

$$\mathrm{Var}[Y] = \mathrm{Var}[aX] = a^2 \cdot \mathrm{Var}[X]$$

Again, by the linearity of the expectation operator, we have

$$\mathrm{E}[Y] = \mathrm{E}[aX] = a \cdot \mathrm{E}[X]$$

Then:

$$\begin{aligned}
\mathrm{Var}[Y] &= \mathrm{E}[(Y - \mathrm{E}[Y])^2] \\
&= \mathrm{E}[(a \cdot X - \mathrm{E}[a \cdot X])^2] \\
&= \mathrm{E}[(a \cdot X - a \cdot \mathrm{E}[X])^2] \\
&= \mathrm{E}[a^2 \cdot (X - \mathrm{E}[X])^2] \\
&= a^2 \cdot \mathrm{E}[(X - \mathrm{E}[X])^2] \\
&= a^2 \cdot \mathrm{Var}[X]
\end{aligned}$$

As with the variance, we can combine these two results into one formula. Let $X$ be a random variable, and let $a$ and $b$ be constants. Then:

$$\begin{aligned}
\mathrm{Var}[aX + b] &= \mathrm{Var}[aX] \\
&= a^2 \cdot \mathrm{E}[X]
\end{aligned}$$

Let's compare our results on the expectation and the variance of a random variable $X$:

$$\mathrm{E}[aX + b] = a \cdot \mathrm{E}[X] + b$$

$$\mathrm{Var}[aX + b] = a^2 \cdot \mathrm{Var}[X]$$

Previously we showed that the expectation of the sum is equal to the sum of the expectations:

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

Is there a similar result for the variance? There is, if we can assume that $X$ and $Y$ are independent. First, let's do some algebra with $\mathrm{E}[(X + Y)^2]$:

$$\begin{aligned}
\mathrm{E}[(X + Y)^2] &= \mathrm{E}[X^2 + 2XY + Y^2] \\
&= \mathrm{E}[X^2] + 2\mathrm{E}[XY] \cdot \mathrm{E}[Y^2] \\
&= \mathrm{E}[X^2] + 2\mathrm{E}[X] \cdot \mathrm{E}[Y] + \mathrm{E}[Y^2]
\end{aligned}$$

Do you see where we needed to use independence? It was when we went from $\mathrm{E}[XY]$ to $\mathrm{E}[X] \cdot \mathrm{E}[Y]$. Next, let's do some algebra on $(\mathrm{E}[X+Y])^2$:

$$\begin{aligned}(\mathrm{E}[X+Y])^2 &= (\mathrm{E}[X]+\mathrm{E}[Y])^2 \\ &= (\mathrm{E}[X])^2 + 2\mathrm{E}[X]\cdot\mathrm{E}[Y] + (\mathrm{E}[Y])^2\end{aligned}$$

Now we can calculate the variance:

$$\begin{aligned}\mathrm{Var}[X+Y] &= \mathrm{E}[(X+Y)^2] - (\mathrm{E}[X+Y])^2 \\ &= \big(\mathrm{E}[X^2] + 2\mathrm{E}[X]\cdot\mathrm{E}[Y] + \mathrm{E}[Y^2]\big) - \big((\mathrm{E}[X])^2 + 2\mathrm{E}[X]\cdot\mathrm{E}[Y] + (\mathrm{E}[Y])^2\big) \\ &= \big(\mathrm{E}[X^2] - (\mathrm{E}[X])^2\big) \;+\; \big(\mathrm{E}[Y^2] - (\mathrm{E}[Y])^2\big) \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y]\end{aligned}$$

# 9 Example In Depth: A Continuous Random Variable

Now let's look at what is perhaps the simplest continuous random variable, the *uniform* distribution. This random variable is defined on a finite interval of the form $[0, b]$, and the density function has the same value for all values in this interval (that's why it's called "uniform"). Let's call this constant value $c$. Our first step is to determine the value of this constant, and we can do this by remembering that a probability density must integrate to 1 over the full support. Thus, we have:

$$\int_0^b c \cdot dx \;=\; 1$$

This integral is easy to evaluate:

$$\begin{aligned}\int_0^b c \cdot dx &= cx\big|_{x=0}^{x=b} \\ &= c\cdot b - c\cdot 0 \\ &= cb\end{aligned}$$

Thus our equation becomes

$$cb = 1$$

Solving, we obtain

$$c = \frac{1}{b}$$

So the density function for a uniform random variable on $[0, b]$ is:

$$f_X(x) = \frac{1}{b}$$

If you think about it, we actually didn't need calculus to prove this, and we could have used a simple argument from elementary geometry: if a rectangle has a total are of 1, and the width is $b$, then the height of the rectangle must be $1/b$.

Once we have the density function $f_X(x)$ for our random variable, we can figure out the cumulative probability function and the survival function. First, let's do the cumulative probability function:

$$\begin{aligned} F_X(x) &= \int_0^x \frac{1}{b} \cdot ds \\ &= \frac{x}{b} \Big|_{x=0}^{x=b} \\ &= \frac{x}{b} \end{aligned}$$

Again, by elementary geometry, this makes sense: if we have a rectangle with width $x$ and height $1/b$, then the total area of the rectangle is $x/b$. The survival function is:

$$\begin{aligned} S_X(x) &= 1 - F_X(x) \\ &= 1 - \frac{x}{b} \\ &= b - xb \end{aligned}$$

We can also solve for the quantile function. Recall that the quantile function $Q_X(q)$ for a given value of $q$ is the particular value $x$ that solves the equation $F_X(x) = q$. In our case, this equation becomes:

$$F_X(x) = \frac{x}{b} = q$$

Thus, the quantile function is

$$Q_X(q) = bq$$

Once again, we can obtain this result from elementary geometry. If we have a rectangle with height $1/b$ and total area $q$, this means that the width must be $bq$.

Notice the relationship between the cumulative probability function and the quantile function:

- For the cumulative probability function, we have a rectangle with height $1/b$ and width $x$, and we want to calculate the total area.
- For the quantile function, we again have a rectangle with height $1/b$, but now we fix the total area to be $q$ and ask what width will achieve this area?

Now for the fun stuff – calculating the moments! The density function is sufficiently simple that we can calculate the general $k$th moment without much difficulty:

$$\begin{aligned}
\mathrm{E}[X^k] &= \int_{\Omega_X} x^k \cdot f_X(x) \cdot dx \\
&= \int_0^b x^k \cdot \frac{1}{b} \cdot dx \\
&= \left. \frac{x^{k+1}}{(k+1) \cdot b} \right|_0^b \\
&= \frac{b^k}{k+1}
\end{aligned}$$

To summarize, we have:

$$\mathrm{E}[X^k] = \frac{b^k}{k+1}$$

We can obtain the first moment from this formula by setting $k = 1$:

$$\begin{aligned}
\mathrm{E}[X^1] &= \frac{b^1}{1+1} \\
&= \frac{b}{2}
\end{aligned}$$

Again, this is easy to understand from elementary geometry. If we have a rectangle of width $b$ and height $1/b$, then it is symmetric about the line $x = b/2$, so this must be the mean.

For the second moment, we have:

$$\begin{aligned} \mathrm{E}[X^2] &= \frac{b^2}{2+1} \\ &= \frac{b^2}{3} \end{aligned}$$

Now we can calculate the variance:

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\ &= \frac{b^2}{3} - \left(\frac{b}{2}\right)^2 \\ &= \frac{b^2}{3} - \frac{b^2}{4} \\ &= \frac{b^2}{12} \end{aligned}$$

# 10 Appendix: Additional Material

Under construction!