Lecture 2: Review of Probability Theory

MATH E-156: Mathematical Foundations of Statistical Software

Theodore Hatch Whitfield

January 29, 2018

What Is Probability?

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.

What Is Probability?

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?

What Is Probability?

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.

What Is Probability?

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.

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.

What Is 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.

What Is Probability?

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.

The Basic Frequentist Setup

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.

The Basic Frequentist Setup

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 $\mathcal{S}$.

The Basic Frequentist Setup

Let's look at some 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:

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

The Basic Frequentist Setup

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].

The Basic Frequentist Setup

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]

The Basic Frequentist Setup

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]

The Basic Frequentist Setup

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.

The Kolmogorov Probability Axioms

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.

The Kolmogorov Probability Axioms

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:

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

The Kolmogorov Probability Axioms

Now we can define Kolmogorov's axioms for a probability function.

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 $\mathcal{E}$ by $\Pr(\mathcal{E})$.

The Kolmogorov Probability Axioms

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 Kolmogorov Probability Axioms

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.

The Kolmogorov Probability Axioms

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.

The Kolmogorov Probability Axioms

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

The Kolmogorov Probability Axioms

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:

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

The Kolmogorov Probability Axioms

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!

The Kolmogorov Probability Axioms

Tthere 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?

The Kolmogorov Probability Axioms

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.

The Kolmogorov Probability Axioms

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.

Random Variables

Random Variables

In general, in this course we won't be focused so much on the actual sample space.

Instead, we will be working with "random variables".
The expression "random variable" is unfortunate, because a random variable is not a "variable" at all – it's a function.

Random Variables

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.

Random Variables

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.

Random Variables

We can also specify X using conventional function notation:

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

Random Variables

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

Random Variables

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.

Thus, the two statements "$X =$ -2" and "The event Red occurred" are really just two different ways of saying the same thing.

Random Variables

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.

Random Variables

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 ω 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.

Random Variables

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

The notation $\Pr(X = x)$ means "the probability that the random variable X takes on the specific value x".

Random Variables

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 Ω_X to denote the support of the random variable X.

Random Variables

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:

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

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.

Random Variables

However, we can also have random variables that are defined on a range of real numbers (possibly infinite), and these are called *continuous random variables*.

Continuous random variables have to be handled differently than discrete random variables.

We cannot directly assign probabilities to individual values of a continuous random variable.

Random Variables

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?

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.

What is the probability that the random variable X is between $x = 1$ and $x = 2$?

The probability that the random variable X is between $x = 1$ and $x = 2$ is:

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

Random Variables

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.

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

- First, it must be non-negative for all the values in its support.
- 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 $f_X(x)$.

Three Important Functions

Three Important Functions

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

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.

Three Important Functions

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:

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

Three Important Functions

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!

Three Important Functions

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.

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 very much non-standard usage, and it is completely contrary to a very well-established convention.

Three Important Functions

The next important function is called the "survival 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:

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

Three Important Functions

The third important function is called 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.

Three Important Functions

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 $x = 7.2$: $Q_X(0.65) = 7.2$.

Three Important Functions

Thus, we can think of the quantile function as being the inverse of the cumulative probability function:

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

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

Three Important Functions

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".

Three Important Functions

For instance, suppose we 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 you can see in this example how for discrete random variables the inverse relation with the cumulative probability functiont breaks down.

Independence

Independence

Suppose A and B are events.

Then A and B are *independent* if this condition holds:

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

In an intuitive sense, independence means that the events are unrelated.

Independence

We've defined independence in terms of events, but we can extend this definition to random variables.

Two discrete 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.

Independence

The definition of independence for continuous random variables is analogous, except that it is now expressed in terms of densities rather than probabilities:

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

Moments (and the Variance)

Moments (and the Variance)

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

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

Moments (and the Variance)

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 $\mathrm{E}[X] = 2.25$.

Moments (and the Variance)

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:

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

Moments (and the Variance)

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:

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

Moments (and the Variance)

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.

Moments (and the Variance)

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.

Moments (and the Variance)

The issue is that if X is a random variable, then so is $g(X)$.

Let's call this random variable $Y = g(X(\omega))$; then by the definition of the expected value of a random variable, we have:

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

Moments (and the Variance)

Do you see what the problem is?

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?

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

Moments (and the Variance)

It turns out that both expressions give the same answer.

In the notes, I work through a detailed example of why this is.

The proof is a little technical, and not particularly enlightening, and you'll understand more by working through the example.

Moments (and the Variance)

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.

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

Moments (and the Variance)

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.

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

Moments (and the Variance)

We can put these two results together.

Let X be a random variable, and let a and b be constants.

Then:

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

Moments (and the Variance)

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 $\mathrm{E}[V]$:

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

All of the results that we've seen so far are completely general, in the sense that they are true for all random variables, and there are no special conditions that have to be satisfied for these statements to be true (other than the existence of the expectation).

Moments (and the Variance)

There's one more important result about expected values, but it's not completely general.

If X and Y are independent, then

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

If X and Y are not independent, then this relationship does not hold, hence the result is not completely general.

Moments (and the Variance)

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?

Moments (and the Variance)

Let X be a random variable. Then the *kth moment* is defined to be the expected value of X^k i.e. the general kth 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?

Moments (and the Variance)

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 kth moment, but in some special cases this is possible.

Moments (and 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.

Moments (and the Variance)

Here's the definition of the variance:

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

Technically speaking, the variance is not really a moment, because a moment is defined to be something of the form $\text{E}[X^k]$.

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.

Moments (and the Variance)

Often it's much more convenient to use an alternative computation formula:

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

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.

Moments (and the Variance)

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 we have:

$$\text{Var}[Y] \;=\; \text{Var}[X + c] \;=\; \text{Var}[X]$$

Moments (and the Variance)

The second important property is that if we multiply a random variable by a constant, the variance is multiplied by the *square* of the constant.

Suppose $Y = aX$. Then by the linearity of the expectation operator we have:

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

Moments (and the Variance)

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:

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

Moments (and the Variance)

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

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

Moments (and the Variance)

Finally, 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 this also the case for the variance?

In general, no.

But if X and Y are *independent*, then the variance of the sum is indeed the sum of the individual variances:

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

Example In Depth:
The Uniform Distribution

Example In Depth: The Uniform Distribution

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.

Example In Depth: The Uniform Distribution

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:

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

Example In Depth: The Uniform Distribution

This integral is easy to evaluate:

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

Example In Depth: The Uniform Distribution

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:

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

Example In Depth: The Uniform Distribution

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 $1/b$.

Example In Depth: The Uniform Distribution

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.

Example In Depth: The Uniform Distribution

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.

Example In Depth: The Uniform Distribution

The survival function is:

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

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 $F_X(x) = q$.

Example In Depth: The Uniform Distribution

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.

Example In Depth: The Uniform Distribution

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?

Example In Depth: The Uniform Distribution

Now for the fun stuff – calculating the moments!

The density function is sufficiently simple that we can calculate the general kth moment without much difficulty.

Example In Depth: The Uniform Distribution

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

Example In Depth: The Uniform Distribution

To summarize, we have:

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

Now let's calculate the first and second moments, and finally the variance.

Example In Depth: The Uniform Distribution

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.

Example In Depth: The Uniform Distribution

For the second moment, we have:

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

Example In Depth: The Uniform Distribution

Now we can calculate the variance:

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