Section 3
Probability and Random Variables

3.1 – Core Probability Concepts

Sample spaces

In class this week, we began examining probability through simulating random processes in the TinkerPlots sampler. This section aims to formalize probability content by examining the theory of random processes. We call such a process that leads to one of several possible outcomes an ____________________. All such possible outcomes form a set called the ______________________, usually denoted S.

Example: Write out the sample spaces for the following experiments:

Flip a coin. S =

Roll a six-sided die. S =

Flip two coins in sequence. S =

Flip three coins in sequence. S =

When examining the outcomes of an experiment, we often focus on specific outcomes of interest. We call such a subset of a sample space an ____________. We can then turn to probability to find how likely something interesting to us is!

Example: Find the probability of getting exactly two heads when flipping three coins in sequence.

Now, find the probability of getting three heads.

Properties of probabilities

There are three basic axioms of probability that define how probability works more generally. Let E be an event of a sample space S.

1. $0 \le P(E) \le 1$
2. $P(S) = 1$
3. For any disjoint events E_1, E_2

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$$

Note: Two events are *disjoint* or *mutually exclusive* if $P(A \text{ and } B) = 0$, that is, they cannot happen together.

The *complement* of an event A, denoted A^c or A', is defined as the opposite of the given event A, that is, all elements of the sample space not included in A.

Complement Rule: $P(A^c) =$

Conditional probability

Example: Consider the following table of counts below:

		Income Level			
		High (H)	Mid (D)	Low (L)	
Marital Status	Not Married (N)	30	40	30	100
	Married (M)	60	50	40	150
		90	90	70	250

Find the following probabilities:

$P(N) =$

$P(H) =$

$P(N \text{ and } H) =$

We found the probability of selecting an unmarried person, a person of high income, and an unmarried person of high income. How would we go about finding the probability of selecting someone of high income, given that we know the person is unmarried?

$P(H \mid N) =$

The *conditional probability* of A given B is

$$P(A \mid B) =$$

We can think of these probabilities as working with known information. The event provided after the "|" is something that we know is true, and are now working to find the probability knowing this.

Notice that the definition of conditional probability gives us information about the intersection too, if we simply rearrange the terms from the conditional probability above.

$$\mathrm{P}(A \text{ and } B) =$$

But sometimes, knowing some given event happens doesn't actually affect the probability of another event. Perhaps there's a certain probability that you attend this class on a particular day. There are probably many things that will change this probability – a rainy day may make you less likely to attend, or maybe other personal events in your life may occur that would lower the probability of your attendance. But there are some events that may not affect this too much: it's likely that the event of you receiving a spam call or the event that someone else in the USA winning the Powerball lottery will not have an impact on how likely you attend class. If that is the case that knowing one event happens doesn't affect the probability of another event, we can say two events are *independent.* Mathematically, we can write the condition for independence as:

$P(A \mid B) =$ (for independent events A, B)

We can also write out an alternative rule for independence, based on the multiplication rule for intersections above.

$P(A \text{ and } B) =$ (for independent events A, B)

Important: While they sound similar, independent events are not the same as mutually exclusive or disjoint events! In fact, mutually exclusive events are very dependent. If A and B are mutually exclusive and A occurs, then you know for a fact that B did not happen!

As your instructor, I would sincerely hope that the event you attend this class and the event that it is rainy are independent events, but experience has unfortunately told me that this isn't true. 🙃

3.2 – Discrete Random Variables

Definition of random variable

We now define new tools to convey a numerical description of items in the sample space and their associated probabilities.

Let S be a sample space. A ____________________ is a function $X : S \longrightarrow \mathbb{R}$, equating outcomes from a sample space to a numerical value.

> *Example:* Consider the experiment of flipping 3 coins, and let X be a random variable for the number of heads.

Random variables give us a way to relate the outcomes of an experiment to a numeric value, but it doesn't tell us anything about probability. We need to assign another function to determine the probability of certain numeric values for a random variable.

The type of function we assign to a variable differs depending upon what kind of variable we are measuring in the first place. There are two types of random variables: ________________ and ________________. In this section, we will only talk about discrete random variables.

Probability mass functions

If the random variable we are working with has a ________________ range, that is, the possible numeric values the random variable can be is ________________, we say that it is a *discrete random variable*.

To talk about probabilities for a discrete random variable, we define a *probability mass function (pmf)*. The probability mass function for a discrete random variable is defined $p_X : \mathbb{R} \longrightarrow [0, 1]$, where $p_X(x) = P(X = x)$, and must satisfy the property:

$$\sum_{all\ x} p_X(x) = 1$$

All of this mathematical jargon is really just masking the main axioms of probability we discussed previously in terms of a new probability function. Specifically, these just make sure that for our probability function that all probabilities should always be between 0 and 1, and that the probabilities of all outcomes must add up to 1.

> *Example*: For the previous experiment of flipping 3 coins, write out the pmf of X.

Describing and summarizing discrete random variables

We now explore several summary measures of a random variable. The first one we will examine is the *expected value*, which is synonymous with the mean value of the random variable. We define the expected value as

$$\mu = E(X) =$$

Example: For the previously discussed experiment of flipping three coins, find the expected value of X.

We can also look at measures of variability on random variables. Similar to the sample variance measure we defined previously on a set of data, we can define the *variance* of a random variable as

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

Computationally, this is not too fun to do. But we can make it slightly easier on ourselves by doing some rearranging of the definition above:

Thus, we get the alternative formula for the variance of X below.

$$\mathrm{Var}(X) =$$

When we defined the sample variance, we defined the sample standard deviation as the square root of the sample variance. That relationship holds up in the random variable case as well!

$$\sigma = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Example: For the previously discussed experiment of flipping three coins, find the variance and standard deviation of X.

Example: For the pmf given in the table below, find $E(X)$ and $\text{Var}(X)$.

x	0	1	2	3	4
$p(x)$	0.4	0.2	0.15	0.15	0.1

When we're working with expected values and variances, there are nice properties we can leverage with linear functions of a random variable. Let a, b be real numbers, and X a random variable.

$E(aX + b) =$

$\text{Var}(aX + b) =$

These properties also hold for linear combinations of random variables:

$$E(X \pm Y) = E(X) \pm E(Y)$$

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$$

Notice the ± in the equation for variance above doesn't stay on the right side of the equation. Why is that the case – specifically, why would taking the difference of two random variables result in a larger variance?

The binomial distribution

When modeling real world experiments, we can often map these to common probability models or distributions. There are many distributions we will not cover in this class (e.g. geometric, Poisson), but we will focus on the binomial distribution.

The binomial distribution is commonly used to model random sampling from a population where outcomes are categorical in nature. This model is something that we've been looking at informally too with some of the examples we've already seen!

The binomial distribution: $X \sim \text{Bin}(n, p)$

$$p_X(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x},\ x = 0, 1, \ldots, n$$

Used to count the number of successes in n independent trials that have a probability p each of succeeding.

Characteristics of a binomial experiment:

- There are always a fixed number of trials, n.
- Each trial has two outcomes: success and failure.
- The probability of success is fixed from trial to trial.
- Each trial is independent – that is, results from previous trials don't influence the outcome of other trials.

Example: A bike shop is ordering new bikes from a supplier, and they always order in shipments of 3 bikes. Before selling bikes, this shop always does an inspection of the bikes to ensure all components are aligned and working properly. Based on their history with the manufacturer, 90% of all bikes shipped to them are ready to be sold without any further work done.

If X is a random variable for the number of bikes ready to be sold immediately, does X represent a binomial experiment?

While we can use the equation for the pmf above to find binomial probabilities, this can be cumbersome, especially if you need to find the probability for an inequality probability. Thus, we will primarily focus on computing binomial probabilities in R. The various functions you can use with the binomial distribution are given below.

```
dbinom(x, n, p) #gives the probability P(X = x)
pbinom(x, n, p) #gives the probability P(X ≤ x)
qbinom(q, n, p) #finds the value x value where P(X ≤ x) = q
rbinom(k, n, p) #randomly generates k data points that come
                  from a Bin(n, p) distribution
```

Example: The bike shop has two customers that have placed pre-orders for this bike. What is the probability that this shipment will have at least two bikes ready to sell with no additional work to be done?

In our last activity in TinkerPlots, we examined the probability experiments of flipping 10 coins and drawing 10 cards from a deck of playing cards. Let's revisit these scenarios now in light of the binomial distribution.

Example: Are counting the number of heads and counting the number of hearts binomial experiments?

Example: For the coins scenario, find the probability that you get exactly 5 heads, exactly 8 heads, and exactly 2 heads. How do these probabilities compare to your TinkerPlots simulation from last activity?

To find the mean and variance of a binomial random variable, we can start with the case of when $n = 1$. In this case, there would only be two possible outcomes for X, 0 and 1. Thus, we can easily find the following:

$E(X) =$

$\text{Var}(X) =$

Now, we can see that to get to a binomial distribution with any n, we could consider adding up many binomial random variables with $n = 1$. So long as these outcomes are independent, their sum would equal the total number of successes (or 1's) that occurred in each of the individual trials. Thus, using our rules for expected values of sums, we can find that for any binomial distribution:

$E(X) =$

$\text{Var}(X) =$

Example: Find the expected value and variance for the number of bikes ready to be sold from the previous example.

3.3 – Continuous Random Variables

Probability density functions

Last section, we talked about discrete random variables and common models like the binomial distribution. This chapter, we will talk about continuous random variables.

If the random variable we are working with has a _______________ range, that is, the possible numeric values the random variable can be is _______________, we say that it is a *continuous random variable*.

For discrete random variables, we assigned to each possible outcome a probability using the probability mass function. However, we now have an uncountable set of possible values for the range of our random variable, like all of $\mathbb{R}$ or an interval like [0, 1]. If we were to try to assign probabilities to an uncountably infinite number of values, we wouldn't be able to make the probabilities add up to 1 – they would greatly exceed that!

Thus, we need a new tool to allow us to take probabilities over ranges of values. We now use functions known as *probability density functions* to do this. These functions allow you to find probabilities for ranges of values for the distribution by calculating the area underneath the function within that range. To preserve the idea that all of the probability must "add up" to 1, a probability density function must have all area under the function equal to 1.

Let's start with a basic example of a probability density function!

> *Example*: A student takes MTD's Green (5) bus into campus. They haven't memorized the schedule, but they know that during the middle of the day that the bus leaves every 15 minutes. Assuming that the bus arrives on time, what is the probability that they will wait no more than 8 minutes for the bus to arrive?

So, areas under functions are easy to find when they're familiar geometric shapes like rectangles. But limiting our probability modeling to geometric area formulas is not very versatile. Unfortunately, we need calculus to compute the area under functions generally, and this is not a prerequisite for this class. (I'm guessing most people are thinking this is more fortunate than unfortunate, and I can respect that. Enjoy your calculus in Stat 400, stat majors!)

However, R is quite good at finding areas under functions, especially functions for well-known distributions. And the most well-known continuous distribution is…

The normal distribution

The normal distribution is probably the most loved, known, used and misused distribution of them all. You've probably heard of bell-curves before in relation to modeling test scores, human heights, and many other natural phenomena. You might have also heard it in terms "curving" grades in a class – in practice, this has little to do with making grades look like a normal distribution, but instead, ends up in the instructor being nice and giving out more points than originally earned.

Just for fun, let's take a look at the function that defines the normal distribution:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty$$

In the normal distribution above, the parameters μ and σ are the mean and standard deviation of that normal distribution. In fact, there is no way to use calculus to determine probabilities with this function in a closed form, so we have to use other tools to compute probabilities.

Before the widespread use of computers, the primary method was to convert everything to a *standard normal distribution*. Usually denoted with the random variable Z, the standard normal distribution is a normal distribution with mean 0 and standard deviation 1. When drawn, this looks something like the picture below:

The usefulness of a standard normal distribution is realized by the following fact: If a random variable X is distributed $N(\mu, \sigma)$, then the random variable Z, defined as

$$Z =$$

has a standard normal distribution. This quantity is typically referred to as a ___________________, which tells you how many standard deviations one data value is from the mean.

Thus, for any normal distribution, we could use this to convert something from any normal distribution to a standard normal distribution, and then use a probability table to find the solution. Such methods are irrelevant with the use of computers! Like with the binomial distribution, R gives ways to calculate the probabilities of a normal distribution.

```
dnorm(x, μ, σ) #gives the height of the density function
pnorm(x, μ, σ) #gives the probability P(X ≤ x)
qnorm(q, μ, σ) #finds the value k where P(X ≤ k) = q
rnorm(k, μ, σ) #randomly generates x data points that
                come from a N(μ, σ) distribution
```

If you leave the fields for the mean and standard deviation blank, R will assume a standard normal distribution.

Let's try some examples of normal probabilities that use these R functions.

Example: Assume that for a specific population, heights are normally distributed with μ = 68 inches and σ = 2 inches. What percentage of the population is taller than 72 inches?

Example: Assume that speeds on Interstate 57 are normally distributed with μ = 68 mph and σ = 4 mph. Find the 85^{th} percentile of these speeds.

Example: For the Interstate 57 speed example, find the 2 speeds that contain the middle 90% of all speeds on Interstate 57.

3.4 – Additional Practice

Example: An archer can hit the bullseye with an arrow 40% of the time. If the archer takes 6 shots in a given round, what is the probability they hit at most 1 bullseye?

What is the probability they hit 4 or more bullseyes?

What is the expected number of bullseyes the archer will hit? What is the variance?

Example: Measuring blood pressure provides many challenges due to the variation in measurements depending on the time within a cardiac cycle it is taken. Adults are often diagnosed for treatment of high blood pressure when they report a blood pressure of 140 mm Hg. For a patient whose average blood pressure is 130 mm Hg and standard deviation is 13 mm Hg, what is the probability that a doctor will diagnose this patient with high blood pressure? Assume that this patient's distribution of blood pressure measurements is normally distributed.

What blood pressure reading would represent the largest measurement among the lowest 25% of measurements? What blood pressure reading would represent the smallest measurement among the highest 25% of measurements?