Chapter 3

Discrete Random Variables

- Learning objectives:
  - understand the definition of a discrete random variable
  - be familiar with common types of discrete random variables and their distributions
  - be able to compute probabilities, expected value, and variance for discrete random variables

In this chapter, we will define discrete random variables, study properties of discrete random variables, and learn about several common and useful types of discrete random variables. From our past experience, we know what a variable is. A variable is a quantity that can vary. But what exactly is a discrete random variable? Each word in the term has a specific meaning:

- Variable: a measured quantity that is allowed to take on a variety of values
- Discrete: the possible values form a finite or countable collection
- Random: the value taken on by the variable is "random", i.e., it is not determined by us

Example 3.0.1

A company makes widgets and sells them for $10 each. They can make as many widgets as they want. The number of widgets is therefore not a fixed number but rather a variable chosen by the company. Thus, we would want to use a variable (say x) to represent the number of widgets made.

The revenue generated by these widgets is also a variable (say y) that is dependent on the number of widgets that are made.

In life, there are many quantities that vary (variables), and the value that the variable takes on is not chosen by us or anybody else. Here are some basic examples:

- the length of time it takes a person run to run a mile tomorrow
- the number of spots we will see after a die is rolled
- a person's weight tomorrow morning
- the number of customers that will go to a restaurant for lunch tomorrow
- the number of patients that will be admitted to the emergency room tomorrow
- how long it will take a person to drive to work tomorrow

All of these quantities are variables since each can take on many possible values. The difference between these variables and the number of widgets made in the previous example is that the actual value of the variable is not chosen by us or anybody else. It is random. That is, the value taken on by the variable involves an element of chance or uncertainty. There are many possibilities, but we do not pick the value. Since the value eventually taken on by these variables is random, describing these types of variables involves probability.

Example 3.0.2

A life insurance company insures twenty thousand customers. The number of these customers that will die this year is a variable. We have no control over what value that variable will take on at the end of the year. Thus, this variable is beyond our control and is a random variable.

3.1 What is a Random Variable?

Definition 3.1.1

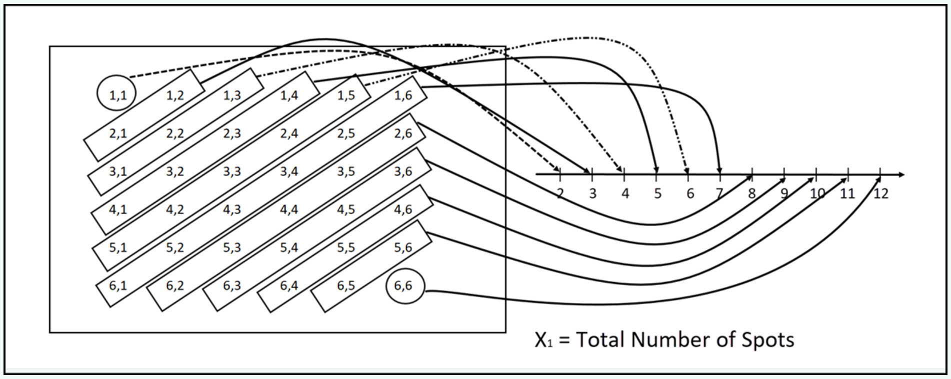
Let $\mathcal{S}$ be a discrete sample space from some experiment. A **random variable** is a function from the sample space to the real numbers. Random variables are always denoted by a capital letter, such as X.

Usually, we would want our random variable to represent some meaningful quantity that we wish to study.

Example 3.1.1

Consider an experiment that involves rolling two dice. The sample space is the collection of all possible pairs in which each member of the pair is one of the integers 1 through 6. Some meaningful random variables in this setting are described below.

Let the random variable X_1 denote the sum of the two dice. A visualization of X_1 is shown below:



We can write what our random variable X_1 does to outcomes in $\mathcal{S}$ using function notation:

$$X_1(1,3) = 4, \quad X_1(2,2) = 4, \quad X_1(3,1) = 4, \quad \text{etc.}$$

We can define other random variables on this sample space that might have an important meaning in certain situations:

- $X_2 =$ the larger of the two values on the dice
- $X_3 =$ the larger number minus the smaller number

As in algebra or calculus, we can use subscripts to distinguish one random variable from another.

Example 3.1.2

Determine $X_1(4,5)$, $X_2(4,5)$, and $X_3(4,5)$.

Example 3.1.3

Suppose that we are going to toss a coin three times. Let X count the number of heads observed in the three tosses. Determine the values of $X(\text{HHH})$, $X(\text{TTH})$, and $X(\text{HTT})$.

In the first three examples, the random variable will choose a number. The randomness is the actual rolling of the dice. Once the dice have stopped moving, the randomness ends. If the dice end up on a 3 and a 4, each of our random variables produces a number. That number is now fixed, and no longer random. The probability that the observed number occurs is something that we usually want to determine in advance.

The same is true for the random variable that counts the number of heads when we toss a coin three times. As the coins rotate through the air, having been flicked with our talented thumb, we have no idea what value the random variable will produce. The value produced is random, since we have no control over its value. It could be noted that we feel confident that $P(X = 0) = \frac{1}{8}$, since we know there are eight equally likely outcomes, and only one of those (TTT) produces $X = 0$. However, we cannot help or hinder the random variable from taking on the value 0, regardless of its probability.

The random variables described above take outcomes from the sample space of a random experiment and turn them into real numbers. We now have a collection of real numbers that are randomly chosen. If we identify each outcome with its corresponding real number, we can view these real numbers as the sample space for an experiment. We would now like to assign probabilities to these numbers in a natural way.

Definition 3.1.2

For every real number x, we define $P(X = x) = \displaystyle\sum_{O_i \in \mathcal{S}; X(O_i)=x} P(O_i)$.

Example 3.1.4

Consider the rolling two dice experiment described above. Recall that X_1 assigns to each outcome the total number of spots rolled. The equally likely probability function seems reasonable to use here, so each of the 36 possible outcomes in the sample space has probability $\frac{1}{36}$ of occurring. Determine the probability that the random variable X_1 takes on the value 4, i.e., determine $P(X_1 = 4)$.

Notice that the only outcomes $O_i \in \mathcal{S}$ that satisfy $X_1(O_i) = 4$ are $(1,3)$, $(2,2)$, and $(3,1)$. Thus, we have

$$P(X_1 = 4) = P((1,3)) + P((2,2)) + P((3,1)) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{3}{36}.$$

Note that if a real number is not assigned to any of the outcomes, then the probability of that number occurring is defined to be 0. For instance, since $X_1(O_i) = 2.75$ is not possible based on the definition of X_1, we have $P(X_1 = 2.75) = 0$.

Definition 3.1.3

The **support** (or sometimes called **space**) of a discrete random variable is defined to be

$$\mathcal{S} = \{x \in \mathbb{R} \mid P(X = x) > 0\},$$

where $\mathbb{R}$ denotes the set of all real numbers.

Note that we often use the same notation $\mathcal{S}$ for the support as we do for the sample space. Since there is an obvious correspondence between these two sets, they are often identified with one another and used interchangeably.

Example 3.1.5

The support of the random variable X_1 from above is $\mathcal{S} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

The support of the random variable X_2 from above is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.

Example 3.1.6

Determine the support of the random variable X_3 from above.

Example 3.1.7

A coin is tossed three times. Let X count the number of heads obtained in the three tosses. Determine the support of the random variable X. Then, determine $P(X = x)$ for each value in the support.

Example 3.1.8

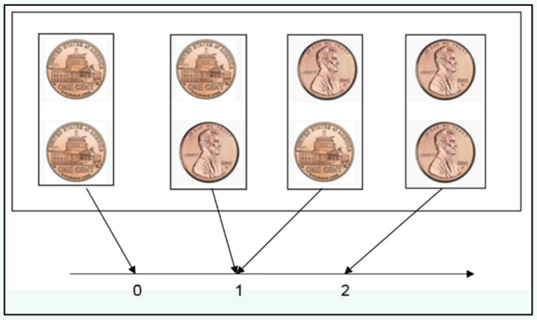
Two coins are tossed, and then a die is rolled. The random variable X will be the number of heads obtained times the value on the die. Determine the support of X.

As noted above, we often identify the support with the sample space. In other words, we view the support as the sample space. Since we have endowed the support with a probability function, the support along with the collection of probabilities is a probability space.

Let's revisit one of the experiments from above. Suppose a coin is tossed twice. How many heads will we get? We can't say for certain. The number of heads obtained is a variable because the value produced can vary among several possibilities. The value the variable takes on is random because we cannot predict in advance what it will be.

Example 3.1.9

Suppose we toss a coin two times. The sample space is shown in the figure below.



Let the random variable X count the number of heads obtained in the two tosses. Thus, X assigns the number of heads tossed to each of the four outcomes as indicated in the figure above. For instance, if our two tosses result in HH, the random variable X assigns the number 2 to that outcome since we ended up with 2 heads.

Note that the sample space is the collection of all possible outcomes of the experiment. In this case, the outcomes are not numbers. A random variable will "turn" these outcomes into numbers according to some rule that we assign. In this case, the rule is to assign each outcome the number of heads it contains.

In this example, X is a variable because the value it takes on can vary (it's one of $0, 1, 2$). The variable X is random because we do not know in advance what value it will take on. Of course, X cannot take on any value randomly. It can only take on values in its support. Here, since the support $\mathcal{S} = \{0, 1, 2\}$ is discrete, X is called a **discrete random variable**.

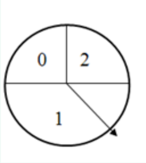
We now make some probability assignments. Of course, these assignments cannot break the three rules for probability functions discussed earlier. We would also like to make assignments that agree with our intuition and loose definition that “probability is a measure of how likely an event is to occur”.

Example 3.1.10

Referring to the previous example, what are reasonable assignments for $P(X = 0)$, $P(X = 1)$, $P(X = 1.75)$, $P(X = 2)$, and $P(X = 3)$?

Example 3.1.11

Suppose a game involves using the spinner below. What is the support of the spinner, and what are reasonable probability assignments?



3.2 The Distribution of a Discrete Random Variable

Consider an experiment in which a coin is tossed three times. Let X count the number of heads. How many heads will we toss? That is, what value will X take on? The answer is “we don’t know, it’s random”. We can, however, produce a collection of ordered pairs of numbers that detail what can happen and what the associated probabilities are. In other words, we can make a list (or develop a formula) that associates the probability of occurrence to each value in the support.

Definition 3.2.1

For a random variable X, the **probability mass function (pmf)** is a function that assigns the probability of occurrence to each value in the support. We often use the notation $f(x)$ to denote this function, so that $f(x) = P(X = x)$.

In some cases, it makes sense to write the pmf using a formula. In other cases, it’s just as easy to display the pmf in tabular format.

Recall the two experiments from earlier: (i) tossing three coins and counting the number of heads and (ii) rolling a single die. Below, we provide the pmfs in both formula and tabular formats.

x	$f(x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

$$f(x) = \binom{3}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{3-x} \text{ for } x = 0, 1, 2, 3$$

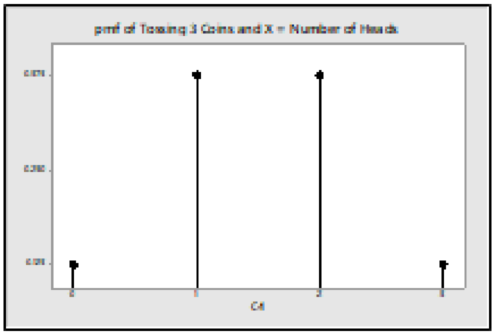
x	$f(x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

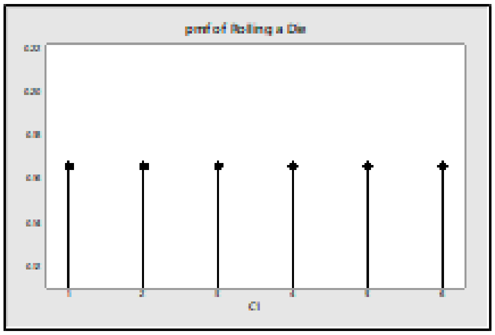
$$f(x) = \frac{1}{6} \text{ for } x = 1, 2, 3, 4, 5, 6$$

When we write the pmf in functional form, we want to be sure to include the support. For instance, if we were to simply write $f(x) = \frac{1}{6}$ in the previous example, we would not be indicating the important fact that $f(x) = \frac{1}{6}$ only if x is in the support of X. For x not in the support of X, the pmf has value $f(x) = 0$. In some cases, this is explicitly indicated as follows:

$$f(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise.} \end{cases}$$

It should be noted that writing out the pmf in tabular form provides the exact same information as if the pmf were written in functional form. We could also graph each pmf as shown below.



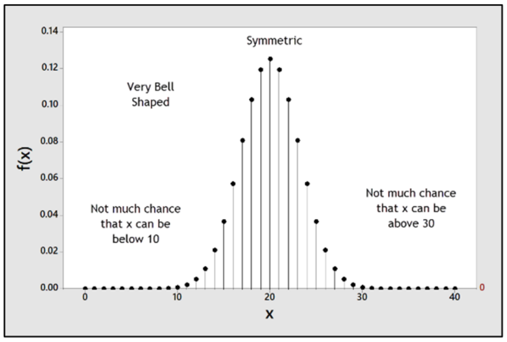


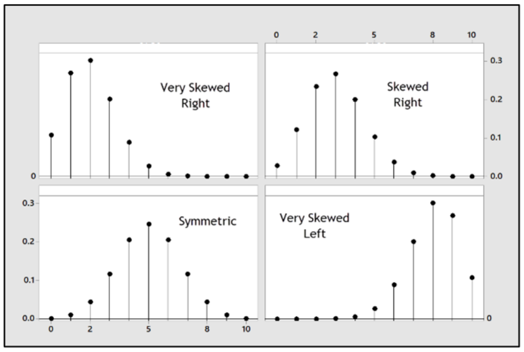
Certainly, if the support was a larger set, a graph like this could be very useful. We purposely use dots with a line projected down. The probabilities here are point masses and are not spread over an interval on the x-axis. In the coin toss example, there is a mass only at $x = 0, 1, 2, 3$. As always, the total mass (probability) must be equal to one (since recall $P(\mathcal{S}) = 1$).

Regardless of the format (functional, tabular, graphical), the total probability of 1 is distributed over the support. We therefore refer to this information that includes the support and associated probabilities as the **distribution of the random variable**.

The stick graphs, like those shown above, are especially helpful when there are more than just a few values in the support. Consider the pmf graphs of the random variables shown in the figure below. These graphs provide much more intuitive information about the distribution of the random variable than we could easily discern from a table of values. While we will use such tables of values to determine answers to probability questions, the graphs can provide a better sense of how the probability is distributed.

In each of the graphs, we can instantly get a feeling for the distribution. If we were working with a random variable with many more possible x-values in the support, the graph of the pmf becomes even more helpful. It is much easier to get a sense of the distribution from a graph of the pmf than it is from a long list of values in a table or from a formula.





Note that the pmf is just one way to convey the distribution of a random variable (sometimes called the "story of probability"). A second method for describing the distribution of a random variable is by using the **cumulative distribution function (CDF)**.

Definition 3.2.2

For a random variable X, we define the **cumulative distribution function (CDF)** as $F(x) = P(X \leq x)$.

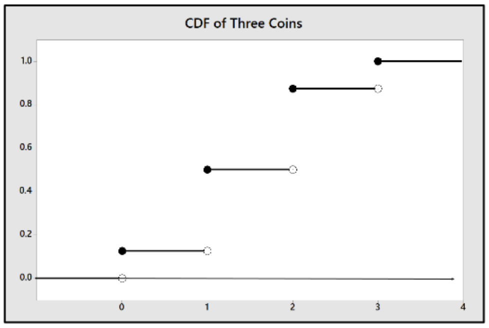
Example 3.2.1

Suppose we needed the probability $P(X \leq 5)$. This probability is equal to the value of the CDF at $x = 5$, i.e., $P(X \leq 5) = F(5)$. If we had a formula for the CDF, or a CDF chart that displays values of $F(x)$ for various x, or even the pmf, we could easily determine this value.

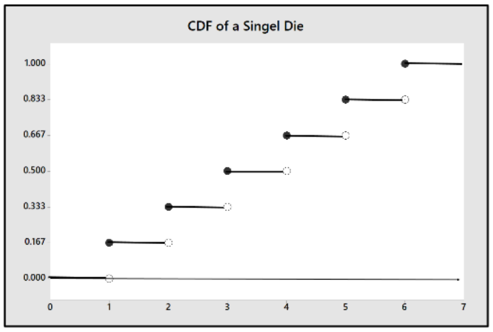
The information contained in the pmf and CDF are identical. Neither tells us more about the story of probability than the other, so both are considered to completely describe the distribution. The distribution in CDF format can be very handy for discrete random variables and is a necessity for continuous random variables that we will discussed in the next chapter.

The CDFs for the two previous examples are provided below, in both tabular and graphical format:

x	$F(x)$
0	$\frac{1}{8}$
1	$\frac{4}{8}$
2	$\frac{7}{8}$
3	1



x	$F(x)$
1	$\frac{1}{6}$
2	$\frac{2}{6}$
3	$\frac{3}{6}$
4	$\frac{4}{6}$
5	$\frac{5}{6}$
6	1



We get from the pmf to the CDF by addition and from the CDF to the pmf by subtraction. When we study continuous distributions, can you guess how we get from one to the other?

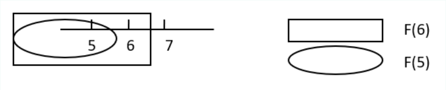
When we are discussing discrete random variables, a graph of the CDF is not as useful as the pmf. We can, however, see some important things from each graph.

Example 3.2.2

Use the CDF to determine the probabilities below. For all problems with CDF charts, you should first write the problem using CDF notation before looking up numbers in the chart.

x	3	4	5	6	7	8	9
$F(x)$	0.20	0.28	0.50	0.55	0.78	0.91	1

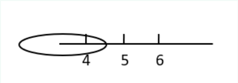
(a) $P(X = 6)$



Note that $F(6) = P(X \leq 6)$ is NOT $P(X = 6)$. It is the probability that has accumulated up to and including 6.

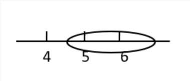
(b) $P(X = 8)$

(c) $P(X < 5)$



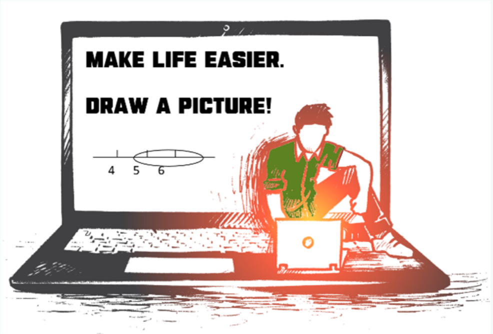
(d) $P(X \leq 7)$

(e) $P(5 \leq X)$



(f) $P(7 \leq X)$

(g) $P(4 \leq X < 8)$



Example 3.2.3

The pmf and CDF of a random variable are given below. In general, the pmf will not be given on exams, so be sure that you are familiar with using the CDF to answer probability questions.

x	$f(x)$	$F(x)$
0	0.0011	0.0011
1	0.0015	0.0026
2	0.0022	0.0048
3	0.0027	0.0075
4	0.0034	0.0109
5	0.0047	0.0156
6	0.0055	0.0211
7	0.0069	0.0280
8	0.0072	0.0352
9	0.0076	0.0428
10	0.0082	0.0510
11	0.0091	0.0601
12	0.0113	0.0714
13	0.0141	0.0855
14	0.0192	0.1047
15	0.0264	0.1311
16	0.0321	0.1632
17	0.0485	0.2117
18	0.0586	0.2703
19	0.0777	0.3480
20	0.0984	0.4464
21	0.1188	0.5652
22	0.0951	0.6603
23	0.0643	0.7246
24	0.0578	0.7824
25	0.0442	0.8266
26	0.0374	0.8640
27	0.0288	0.8928
28	0.0254	0.9182
29	0.0210	0.9392
30	0.0112	0.9504
31	0.0101	0.9605
32	0.0091	0.9696
33	0.0085	0.9781
34	0.0072	0.9853
35	0.0062	0.9915
36	0.0053	0.9968
37	0.0032	1.0000

(a) $P(X = 9)$

(b) $P(X < 19)$

(c) $P(8 \leq X < 25)$

(d) $P(8 < X \leq 25)$

(e) $P(X > 28)$

(f) Find the number k so that $P(X > k) \approx 0.05$.

(g) Find the number k so that $P(X \geq k) \approx 0.05$.

Theorem 3.2.1

Below are some important properties of the CDF for a discrete random variable:

(a) The CDF is not continuous, but it is continuous from the right.

(b) $\lim\limits_{x \to \infty} F(x) = 1$

(c) $\lim\limits_{x \to -\infty} F(x) = 0$

(d) $P(a \leq X \leq b) = F(b) - F(a^-)$, where the notation a^- is used to indicate the x-value immediately preceding a (if one exists)

(e) $P(a \leq X \leq b) = F(b) - F(a) + f(a)$

(f) $P(X = a) = f(a) = F(a) - F(a^-)$

Example 3.2.4

Let X be the result when a single die is rolled. Determine $P(3 \leq X \leq 5)$.

Example 3.2.5

The pmf for some random variable X is given below. Determine the CDF, and use it to compute the probabilities that follow.

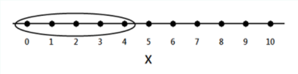
x	$f(x)$	$F(x)$
1	0.15	
2	0.11	
3	0.09	
4	0.35	
5	0.17	
6	0.13	

Note that to get from the pmf to the CDF, we add all probability at or below each value in the support.

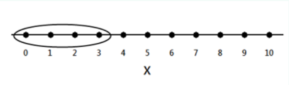
Alternatively, when determining $F(4)$, for instance, we can just add $f(4)$ to $F(3)$.

Notice that the CDF is non-decreasing and ends up with value 1 in the case where the support is finite. If the support is countably infinite, the CDF will never have value 1 but will have a limit of 1.

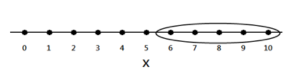
(a) $P(X \leq 4)$



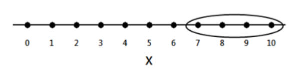
(b) $P(X < 4)$



(c) $P(6 \leq X)$



(d) $P(6 < X)$



Example 3.2.6

Write $P(4 \leq X \leq 8)$ in CDF notation if the support is the set of integers.

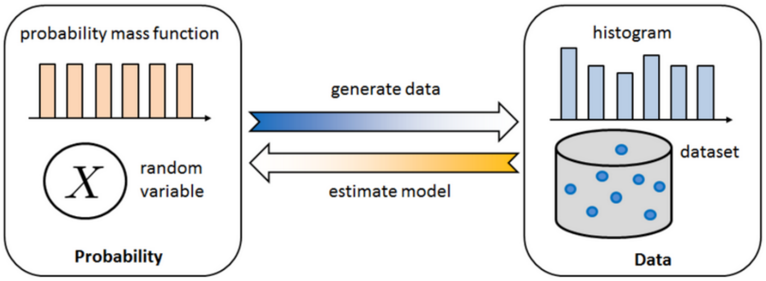
Write $P(4 \leq X < 10)$ in CDF notation if the support is just the set of integers.

Write $P(4 \leq X < 10)$ in CDF notation if the support is NOT just the set of integers.

Write $P(4 \leq X < 10)$ in CDF notation if the support is the set of all integers along with all integers divided by 2.

Write $P(4 \leq X < 10)$ in CDF notation if the support is the set of all integers along with all integers divided by 10.

Note that in this section, we studied the theoretical distribution of a discrete random variable. In practice, we are often working with an actual data set, rather than deriving properties of theoretical distributions. But there is a close connection between the theory and the practice, as illustrated in the figure below (from *Introduction to Probability for Data Science* by Chan):

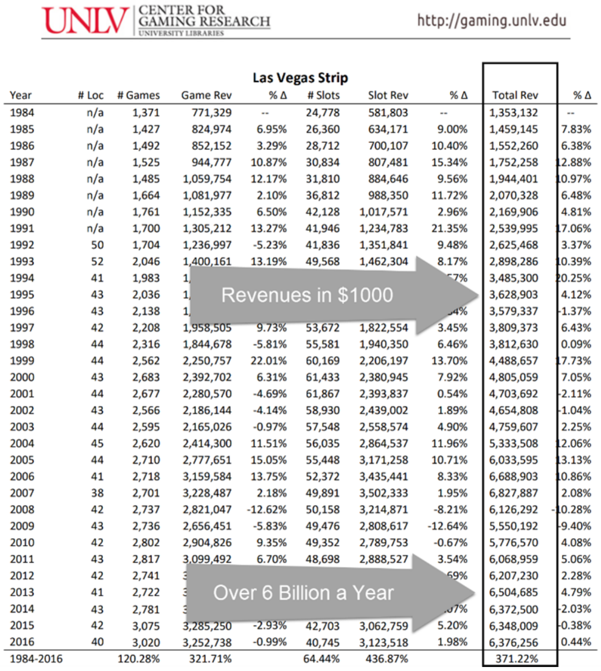


The underlying random variable of interest is assumed to follow some distribution that can be described by a pmf with just a few parameters (for instance, the mean and variance). Using the data set, the goal of statistical inference is to estimate these parameters and thereby estimate the pmf or distribution of the variable of interest. In other words, we try to discern the model information from the available data. The theoretical properties of the estimated distribution can then be used in the context of the variable or process under investigation, for instance, to assist with prediction or recovery in machine learning.

On the other hand, we often find ourselves in a situation where it is reasonable to assume the random variable (or process) of interest follows a certain type of distribution. In this case, we may use the assumed distribution to generate or simulate samples that could have occurred if the process was well-modeled by the assumed distribution.

The samples could be used, for instance, as training data for a deep neural network. The samples could also be used as test data to verify algorithms written to compute statistical properties of the underlying distribution.

3.3 The Expected Value of a Random Variable



UNLV CENTER FOR GAMING RESEARCH UNIVERSITY LIBRARIES http://gaming.unlv.edu

Las Vegas Strip

Year	# Loc	# Games	Game Rev	% Δ	# Slots	Slot Rev	% Δ	Total Rev	% Δ
1984	n/a	1,371	771,329	--	24,778	581,803	--	1,353,132	--
1985	n/a	1,427	824,974	6.95%	26,360	634,171	9.00%	1,459,145	7.83%
1986	n/a	1,492	852,152	3.29%	28,712	700,107	10.40%	1,552,260	6.38%
1987	n/a	1,525	944,777	10.87%	30,834	807,481	15.34%	1,752,258	12.88%
1988	n/a	1,485	1,059,754	12.17%	31,810	884,646	9.56%	1,944,401	10.97%
1989	n/a	1,664	1,081,977	2.10%	36,812	988,350	11.72%	2,070,328	6.48%
1990	n/a	1,761	1,152,335	6.50%	42,128	1,017,571	2.96%	2,169,906	4.81%
1991	n/a	1,700	1,305,212	13.27%	41,946	1,234,783	21.35%	2,539,995	17.06%
1992	50	1,704	1,236,997	-5.23%	41,836	1,351,841	9.48%	2,625,468	3.37%
1993	52	2,046	1,400,161	13.19%	49,568	1,462,304	8.17%	2,898,286	10.39%
1994	41	1,983	1,	[illegible]	[illegible]	[illegible]	57%	3,485,300	20.25%
1995	43	2,036	1,	[illegible]	[illegible]	[illegible]	[illegible]	3,628,903	4.12%
1996	43	2,138	1,	[illegible]	[illegible]	[illegible]	4%	3,579,337	-1.37%
1997	42	2,208	1,958,505	9.73%	53,672	1,822,554	3.45%	3,809,373	6.43%
1998	44	2,316	1,844,678	-5.81%	55,581	1,940,350	6.46%	3,812,630	0.09%
1999	44	2,562	2,250,757	22.01%	60,169	2,206,197	13.70%	4,488,657	17.73%
2000	43	2,683	2,392,702	6.31%	61,433	2,380,945	7.92%	4,805,059	7.05%
2001	44	2,677	2,280,570	-4.69%	61,867	2,393,837	0.54%	4,703,692	-2.11%
2002	43	2,566	2,186,144	-4.14%	58,930	2,439,002	1.89%	4,654,808	-1.04%
2003	44	2,595	2,165,026	-0.97%	57,548	2,558,574	4.90%	4,759,607	2.25%
2004	45	2,620	2,414,300	11.51%	56,035	2,864,537	11.96%	5,333,508	12.06%
2005	44	2,710	2,777,651	15.05%	55,448	3,171,258	10.71%	6,033,595	13.13%
2006	41	2,718	3,159,584	13.75%	52,372	3,435,441	8.33%	6,688,903	10.86%
2007	38	2,701	3,228,487	2.18%	49,891	3,502,333	1.95%	6,827,887	2.08%
2008	42	2,737	2,821,047	-12.62%	50,158	3,214,871	-8.21%	6,126,292	-10.28%
2009	43	2,736	2,656,451	-5.83%	49,476	2,808,617	-12.64%	5,550,192	-9.40%
2010	42	2,802	2,904,826	9.35%	49,352	2,789,753	-0.67%	5,776,570	4.08%
2011	43	2,817	3,099,492	6.70%	48,698	2,888,527	3.54%	6,068,959	5.06%
2012	42	2,741	3	[illegible]	[illegible]	[illegible]	69%	6,207,230	2.28%
2013	41	2,722	3	[illegible]	[illegible]	[illegible]	[illegible]	6,504,685	4.79%
2014	43	2,781	3	[illegible]	[illegible]	[illegible]	.07%	6,372,500	-2.03%
2015	42	3,075	3,285,250	-2.93%	42,703	3,062,759	5.20%	6,348,009	-0.38%
2016	40	3,020	3,252,738	-0.99%	40,745	3,123,518	1.98%	6,376,256	0.44%
1984-2016		120.28%	321.71%		64.44%	436.87%		371.22%	

Las Vegas, Nevada is said to be "the place where dreams come true". This is certainly true for the owners of the casinos. Casinos, just on the strip, have averaged over $6,000,000,000 ($6 billion) in revenue per year from 2005 through 2016.

How do casinos take in so much revenue and consistently come out ahead. Are they cheating? Are they incredibly lucky? Or, do they just expect to win? Why would they expect to win? Perhaps they have chosen or designed the games in such a way that the long-run performance is profitable. In this section, we discuss the expected value of a random variable, which can be used as a measure of this long-run performance.

Each time someone places a bet, they are doing a random experiment. The outcome is the amount won or lost on the bet. We can view the outcome as a random variable that randomly chooses (based on the rules of the game and what happens during the game) how much a player wins or loses.

What would it mean for the casino to have the edge in a game? Does it mean that they have a better chance of winning than the player? In some cases, yes, but not in all. Let's look at some game examples below and decide if we would like to play. Note that these are not actual games in Las Vegas, but rather are designed to introduce the concept of expected value.

Example 3.3.1

A player rolls a single die. If a 1 is rolled, the player wins \$10. For all other outcomes, the player loses \$1.

In this game, the player would lose far more often than she would win. But when she wins, her winnings are much greater than the loss if she were to lose.

If everything goes according to plan, since the probability of rolling a 1 is $\frac{1}{6}$, the player would expect to win \$10 exactly $\frac{1}{6}$ of the time and lose \$1 exactly $\frac{5}{6}$ of the time. Overall, this would make the game profitable for the player.

Of course, things might not go as planned, and the player may not end up rolling a 1 exactly $\frac{1}{6}$ of the time.

The above argument also requires the player to play the game enough times in order for this to be possible. For instance, if the player only played the game once, winning $\frac{1}{6}$ of the time isn't possible. But theoretically, the player would expect to come out ahead *in the long run*. If the player was prepared to play many times, it would be in her best interest to play the game.

Example 3.3.2

Consider a game in which the player has a 50% chance of winning. If the player wins, they collect \$10. If the player loses, they must pay \$11.

Note that in this game, the player will theoretically win as often as they lose.

However, in this case, each win (\$10) does not make up for each loss (\$11).

In the long run, the player should expect to lose money, and so should not want to play this game.

Example 3.3.3

Consider a game in which, if the player wins, they receive \$10. If the player loses, they must pay \$10. Should a person play this game?

If you have already answered "Yes" or "No" to the above question, you have answered too soon. More information is needed. We do not know what the probability of winning is. If the theoretical chance of winning is 50%, the game is called fair, and we may decide to play. If the theoretical chance of winning is 49%, the game is slightly unfair, so we may decide not to play. If the theoretical chance of winning is 40%, the game is very unfair, and playing would be unwise.

Based on these examples, we see that the theoretical expectation involves both the probability of winning along with the amount won or lost.

Definition 3.3.1 (Mean of a Discrete Random Variable)

The **mean** of a discrete random variable X is given by

$$\mu = \sum_{x \in \mathcal{S}} x f(x),$$

provided the sum converges.

If we are in a situation where we have two random variables, say X and Y, we will use subscripts to denote which random variable a mean is connected to: μ_X and μ_Y.

There are times in which we prefer a different name and notation for the mean of a random variable. We can also use the phrase **expected value** to refer to the mean of a random variable. In this case, we often use the notation $\mathrm{E}[X]$, although we could still use μ_X if we prefer.

Remember that a random variable chooses numbers via some random mechanism. We should think of the mean of a random variable as what the data average $\overline{x}$ would be if everything goes according to plan. In other words, if the proportion of times each value in a data set appears is the same as $P(X = x)$, then the average of the data ($\overline{x}$) will be equal to the expected value or mean of the random variable ($\mathrm{E}[X]$ or μ_X). Generally, though, this is not the case. We will learn later that when we have a large amount of data (i.e., a large sample size), the data average $\overline{x}$ will be very close to the mean $\mathrm{E}[X] = \mu_X$ (but probably not exactly equal to it).

Example 3.3.4

A single die is rolled. Let X denote the number of spots viewed. Calculate the mean of X.

Obviously, we can never actually roll a 3.5. So it can happen that the expected value of a random variable is outside the set of possible values. But recall that this number is the theoretical mean, and can be viewed as the long-run average. In other words, if we were to roll the die many, many times, we'd expected the average of the results to be approximately 3.5.

If the proportion in which each outcome occurs agrees with its theoretical probability, then the mean will be equal to the data average as noted above. For instance, consider rolling the die 60 times. Since the probability of each outcome is $\frac{1}{6}$, suppose that the number of times each outcome occurs is 10 (i.e., the proportion of times each outcome occurs agrees with its probability). In this case, the average of the data would be

$$\begin{aligned}\overline{x} &= \frac{1+1+\cdots+1+2+2+\cdots+2+3+3+\cdots+3+4+4+\cdots+4+5+5+\cdots+5+6+6+\cdots+6}{60} \\ &= \frac{10(1)+10(2)+10(3)+10(4)+10(5)+10(6)}{60} \\ &= 3.5\end{aligned}$$

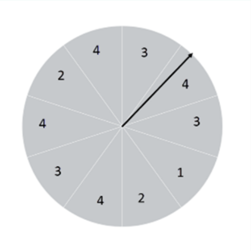
In other words, the data average is equal to the expected value of the random variable. In general, we should not expect this to happen, but it's reasonable to expect that the observed proportions are close to the theoretical probabilities. In this case, the data average will not be equal to the expected value, but it should be close.

Example 3.3.5

A modem transmits a +2 voltage signal into a channel. The channel adds to this signal a noise term X that has support $\mathcal{S}_X = \{0, -1, -2, -3\}$ with probabilities $0.4, 0.3, 0.2, 0.1$, respectively. Let Y represent the output of the channel. Determine (i) the support of Y, (ii) its pmf, (iii) its mean, and (iv) the probability that the input and output are equal.

Example 3.3.6

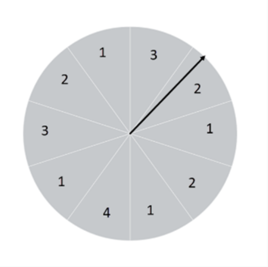
Consider spinning the spinner shown below. What is the expected result?



Suppose the spinner above was used in a game with 300 spaces. In the game, the player moves the number of spaces shown on the spinner. Since the expected number on the spinner is 3, we would expect it to take about 100 spins in order to move through all the spaces.

Example 3.3.7

Consider spinning the spinner shown below. What is the expected result? About how many spins would it take to get through the 300 spaces on the board game?



Note 3.3.1

Not all random variables have a finite mean. Let X be a random variable with pmf $f(x) = \frac{1}{x(x+1)} = \frac{1}{x} - \frac{1}{x+1}$ for $x = 1, 2, 3, \ldots$. It's not hard to check that the sum of the values of the pmf forms a telescoping series with sum 1. In other words, $f(x)$ describes a valid probability function on the set of positive integers.

To determine the mean of this random variable, we calculate

$$\mu = \sum_{x=1}^{\infty} x \cdot \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1},$$

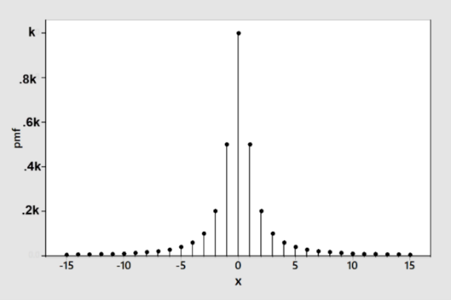
which is essentially the divergent harmonic series (missing the first term). Since this series does not converge, the random variable does not have a finite mean.

While it is possible for a random variable to not have a finite mean, we will not consider any such examples in this course.

Example 3.3.8

Consider a random variable X with support consisting of all integers and pmf given by $f(x) = \frac{k}{x^2+1}$, where k is the value that makes $\displaystyle\sum_{x=-\infty}^{\infty} \frac{k}{x^2+1} = 1$. We know that if $k = 1$, the series converges, so there must be a value k that makes the series converge to 1. The value of k is not important in what follows, so we will not attempt to determine it.

If we make a stick graph of the distribution (or notice that the pmf is an even function), we can see that the distribution is symmetric. Intuitively, what would we expect the mean of X to be?



Example 3.3.9

In the Pick 3 Lottery, a player choose 3 numbers from 0 through 9 (the same number can be selected multiple times). To win, the player must select the correct numbers in the correct order.

There are $10^3 = 1{,}000$ possible permutations, which means each ticket has a probability of winning equal to $\frac{1}{1{,}000} = 0.001$ (assuming each possible winning result is equally likely to occur).

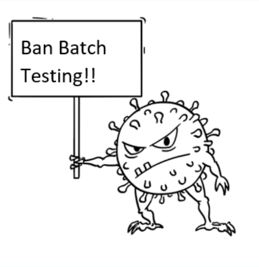
If the player selects the winning permutation, they win \$500. Since the game costs \$1 to play, this means the player profits \$499.

What is the expected value of the Pick 3 Lottery?

Example 3.3.10

Recall our discussion of batch testing for a disease at the end of the previous chapter. In one example, we determined that if the positivity rate is 1% and we use batches of size 10, the probability that a batch comes back negative is 0.9044. Determine the expected number of lab tests needed for a random batch of 10 patients.

Note that without batch testing, we would always need 10 tests for 10 patients. With batch testing, the number of tests needed is either 1 or 11. If the batch comes back negative, we have done only 1 test and are done. If the batch comes back positive, then we would need to test each of the 10 patients individually, and so will have performed a total of 11 tests.



Note 3.3.2

In the previous example, each group of 10 patients would require 10 lab tests if batch testing were not used. But if batch testing is used, each group of 10 patients would only require 1.956 lab tests (on average). So we would be saving, on average, 8.044 lab tests per 10 patients. This is an amazing 80% savings that can provide labs with the ability to not fall behind on testing. Of course, the numbers here will change if the positivity rate and/or batch size changes.

Example 3.3.11

Revisit the previous example using a batch size of 5 and positivity rate 3%.

3.4 The Expected Value of a Function of a Random Variable

Often times, we are in a situation in which we need to look at a function (or transformation) of a variable. For instance, changing feet into meters, degrees Fahrenheit into degrees Centigrade, etc. In this section, our goal is to determine the expected value of a function or transformation of a random variable. If we transform a random variable X by some function $Y = g(X)$, it should be clear that Y itself is a random variable (as the value of Y depends on the value of X, which is random).

Suppose that we have a discrete random variable X with pmf given below. We now let $Y = 2X + 5$. What is the expected value of Y? By definition, $\mathrm{E}[Y] = \sum_{\text{all } y} y f_Y(y)$. So to determine the expected value of Y, we will start by finding the pmf $f_Y(y)$ of Y, and then use the aforementioned formula.

x	$f_X(x)$
-2	$\frac{1}{8}$
-1	$\frac{1}{8}$
0	$\frac{2}{8}$
1	$\frac{1}{8}$
2	$\frac{1}{8}$
3	$\frac{1}{8}$
4	$\frac{1}{8}$

y	$f_Y(y)$
1	$\frac{1}{8}$
3	$\frac{1}{8}$
5	$\frac{2}{8}$
7	$\frac{1}{8}$
9	$\frac{1}{8}$
11	$\frac{1}{8}$
13	$\frac{1}{8}$

The expected value of Y can then be calculated as follows:

$$\begin{aligned}\mathrm{E}[Y] &= \sum_{\text{all } y} y f_Y(y)\\ &= 1\cdot\frac{1}{8} + 3\cdot\frac{1}{8} + 5\cdot\frac{2}{8} + 7\cdot\frac{1}{8} + 9\cdot\frac{1}{8} + 11\cdot\frac{1}{8} + 13\cdot\frac{1}{8}\\ &= \frac{54}{8}\\ &= 6.75\end{aligned}$$

Notice that we could also compute the expected value of Y as follows:

$$\begin{aligned}\mathrm{E}[Y] &= 6.75\\ &= \frac{54}{8}\\ &= 1\cdot\frac{1}{8} + 3\cdot\frac{1}{8} + 5\cdot\frac{2}{8} + 7\cdot\frac{1}{8} + 9\cdot\frac{1}{8} + 11\cdot\frac{1}{8} + 13\cdot\frac{1}{8}\\ &= (2(-2)+5)\cdot\frac{1}{8} + (2(-1)+5)\cdot\frac{1}{8} + (2(0)+5)\cdot\frac{2}{8} + (2(1)+5)\cdot\frac{1}{8} + (2(2)+5)\cdot\frac{1}{8} + (2(3)+5)\cdot\frac{1}{8} + (2(4)+5)\cdot\frac{1}{8}\\ &= \sum_{\text{all } x}(2x+5)f_X(x)\\ &= \sum_{\text{all } x} g(x) f_X(x)\end{aligned}$$

Theorem 3.4.1

Given a discrete random variable X and the linear transformation $Y = g(X) = aX + b$, the expected value of Y is

$$\mathrm{E}[Y] = \mathrm{E}[g(X)] = \sum_{\text{all } x} g(x) f_X(x).$$

Proof: Since $Y = g(X) = aX + b$ is a one-to-one function, the proof is rather straightforward:

$$\mathrm{E}[Y] = \sum_{\text{all } y} y f_Y(y) = \sum_{\text{all } x} (ax+b)P(Y = ax+b) = \sum_{\text{all } x} (ax+b)P(X = x) = \sum_{\text{all } x} g(x) f_X(x).$$

Note that the argument above would work with any function $Y = g(X)$ that is one-to-one. □

Theorem 3.4.2

Given a discrete random variable X and the linear transformation $Y = g(X) = aX + b$,

$$\mathrm{E}[Y] = \mathrm{E}[aX+b] = a\,\mathrm{E}[X] + b.$$

Proof:

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}[aX+b] \\
&= \sum_{\text{all } x} (ax+b) f_X(x) \\
&= \sum_{\text{all } x} ax f_X(x) + \sum_{\text{all } x} b f_X(x) \\
&= a \sum_{\text{all } x} x f_X(x) + b \sum_{\text{all } x} f_X(x) \\
&= a\,\mathrm{E}[X] + b(1) \\
&= a\,\mathrm{E}[X] + b
\end{aligned}$$

□

Because of the previous theorem, the expected value of a random variable is a called a **linear operator**.

Example 3.4.1

If a random variable X has mean $\mu_X = \mathrm{E}[X] = 4$, determine the mean of $Y = 2X + 3$.

If we think about the last example, it seems like a no-brainer that the mean of Y should be 11. For instance, suppose the average score on a quiz is 4 (out of 15 points, let's say). The professor realizes the quiz may have been too hard due to the low average score, so decides to apply a curve by doubling everyone's score. It should be clear that the new average score is 8. After some further consideration, the professor believes an average score of 8 out of 15 is still too low, so decides to add 3 points to everyone's score. It's obvious that the new average score is now 11. The effect on the average score by doubling and adding 3 points is intuitively obvious, and agrees with our calculation in the previous example.

In the case where Y is a linear transformation of X, i.e., $Y = g(X) = aX + b$, the previous example suggests that the mean of Y is obtained by applying the same linear transformation that defines Y to the mean of X, i.e., $\mathrm{E}[Y] = \mathrm{E}[g(X)] = a\,\mathrm{E}[X] + b = g(\mathrm{E}[X])$. A natural question to ask is whether this is always true, i.e., if $Y = g(X)$, is it always true that $\mathrm{E}[Y] = g(\mathrm{E}[X])$? It turns out the answer to this question is no, in general. However, we have seen that it is true in the special case where $Y = g(X)$ is a linear transformation of X. Below is an example to demonstrate that this property does not hold in general.

Example 3.4.2

Consider a discrete random variable X with pmf given below. The expected value of X is easily shown to be $\mathrm{E}[X] = \sum_{\text{all } x} x f_X(x) = \frac{7}{8}$.

x	$f_X(x)$	$x f_X(x)$
-2	$\frac{1}{8}$	$\frac{-2}{8}$
-1	$\frac{1}{8}$	$\frac{-1}{8}$
0	$\frac{2}{8}$	0
1	$\frac{1}{8}$	$\frac{1}{8}$
2	$\frac{1}{8}$	$\frac{2}{8}$
3	$\frac{1}{8}$	$\frac{3}{8}$
4	$\frac{1}{8}$	$\frac{4}{8}$

Now consider $Y = g(X) = X^2$. The pmf of Y is computed in the first table below, and then summarized in the second table (so that each value of Y appears only once in the table). Using this, we can determine the expected value of Y as $\mathrm{E}[Y] = \sum_{\text{all } y} y f_Y(y) = \frac{35}{8}$.

y	$f_Y(y)$	$y f_Y(y)$
4	$\frac{1}{8}$	$\frac{4}{8}$
1	$\frac{1}{8}$	$\frac{1}{8}$
0	$\frac{2}{8}$	0
1	$\frac{1}{8}$	$\frac{1}{8}$
4	$\frac{1}{8}$	$\frac{4}{8}$
9	$\frac{1}{8}$	$\frac{9}{8}$
16	$\frac{1}{8}$	$\frac{16}{8}$

y	$f_Y(y)$	$y f_Y(y)$
0	$\frac{2}{8}$	0
1	$\frac{2}{8}$	$\frac{2}{8}$
4	$\frac{2}{8}$	$\frac{8}{8}$
9	$\frac{1}{8}$	$\frac{9}{8}$
16	$\frac{1}{8}$	$\frac{16}{8}$

Note that in this case, it's clear that $\mathrm{E}[Y] \neq g(\mathrm{E}[X]) = (\mathrm{E}[X])^2$.

So we see that, in general, if $Y = g(X)$, it is not true that $\mathrm{E}[Y] = g(\mathrm{E}[X])$. We do, however, notice something useful when considering the pmf tables in the previous example. Note that if we compare the pmf table of X with the first version of the pmf table of Y, the probabilities are identical and each y-value is the square of an x-value, i.e., $y = x^2$. This leads us to the following theorem that will help us determine the expected value of a transformation of a random variable X.

Theorem 3.4.3

Given a discrete random variable X and a function (or transformation) $Y = g(X)$, the expected value of Y can be calculated as follows:

$$\mathrm{E}\left[Y\right] = \mathrm{E}\left[g(X)\right] = \sum_{\text{all } x} g(x) f_X(x).$$

The importance of this theorem is much more than it seems. It allows us to determine the expected value of a transformation of a random variable without determining the distribution of the new random variable. We were able to easily determine $f_Y(y)$ in the previous example, but in many situations, determining $f_Y(y)$ is not as straightforward (we will especially see this later when we discuss continuous random variables). In cases where determining $f_Y(y)$ is difficult, the above theorem allows us to compute $\mathrm{E}\left[Y\right]$ without having to determine $f_Y(y)$.

3.5 Some Important Functions of Random Variables (Moments and Variance)

Definition 3.5.1

The **variance** of a discrete random variable X is defined by the formula

$$\sigma_X^2 = \sum_{\text{all } x} (x - \mu)^2 f_X(x),$$

provided the sum converges.

Note that we could also express the variance as the expected value of the random variable $Y = g(X) = (X - \mu)^2$, i.e., $\sigma_X^2 = \mathrm{E}\left[Y\right] = \mathrm{E}\left[(X - \mu)^2\right]$.

Definition 3.5.2

The **standard deviation** of a random variable X, denoted by σ_X, is the square root of the variance: $\sigma_X = \sqrt{\sigma_X^2}$.

When working with a data set, we use the standard deviation to measure of the variation or "spread" of the data (as seen in Chapter 1). In other words, if the values in the data set are close in value (and thus close to the mean), the standard deviation will be small. On the other hand, if the values in the data set are very spread out or far apart, then some values will be far from the mean and the standard deviation will be large.

The standard deviation of a random variable can be used in mostly the same way. It provides a measure of how spread out the support is with respect to the associated probabilities. In other words, if we observed a data set where the proportion of times each value occurs exactly matches its probability, the standard deviation of the random variable will equal the standard deviation of the data set.

Example 3.5.1

Let X be the result when a single die is rolled. Determine the standard deviation of X.

Theorem 3.5.1 Computational Formula for Variance

When computing the variance of a random variable, we often make use of the following computational formula:

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

Proof:

$$\begin{aligned}
\sigma_X^2 &= \mathrm{E}\left[(X-\mu)^2\right] \\
&= \sum_{\text{all } x} (x-\mu)^2 f_X(x) \\
&= \sum_{\text{all } x} (x^2 - 2x\mu + \mu^2) f_X(x) \\
&= \sum_{\text{all } x} x^2 f_X(x) - 2\mu \sum_{\text{all } x} x f_X(x) + \mu^2 \sum_{\text{all } x} f_X(x) \\
&= \mathrm{E}\left[X^2\right] - 2\mu\,\mathrm{E}\left[X\right] + \mu^2(1) \\
&= \mathrm{E}\left[X^2\right] - 2\left(\mathrm{E}\left[X\right]\right)^2 + \left(\mathrm{E}\left[X\right]\right)^2 \\
&= \mathrm{E}\left[X^2\right] - \left(\mathrm{E}\left[X\right]\right)^2
\end{aligned}$$

□

Example 3.5.2

Let X be the result when a single die is rolled. Recalculate the standard deviation of X using the above computational formula.

In general, this theorem makes the calculation of the variance of a random variable much easier than using the definition. In many problems, we will see that this computational formula will be a necessity.

Note that not all random variables have a finite variance. We saw earlier that it's possible for a random variable to not have a finite mean. But even if a random variable has a finite mean, it's possible that the variance is not finite. While we will not consider such random variables in this course, we include one example here for completeness.

Example 3.5.3

Consider a random variable X with pmf $f(x) = \frac{k}{x^3}$ for $x = 1, 2, 3, \ldots$, where k is a constant that makes $f(x)$ a valid pmf, i.e., k is a constant such that $\displaystyle\sum_{x=1}^{\infty} \frac{k}{x^3} = 1$.

This random variable has a finite mean since $\displaystyle \mathrm{E}[X] = \sum_{x=1}^{\infty} x \cdot \frac{k}{x^3} = \sum_{x=1}^{\infty} \frac{k}{x^2} = \frac{k\pi^2}{6}$.

To determine the variance of X, we need to determine $\mathrm{E}\left[X^2\right]$. But since $\displaystyle \mathrm{E}\left[X^2\right] = \sum_{x=1}^{\infty} x^2 \cdot \frac{k}{x^3} = \sum_{x=1}^{\infty} \frac{k}{x}$, we see that $\mathrm{E}\left[X^2\right]$ diverges so that X does not have a finite variance.

Theorem 3.5.2

Given a discrete random variable X and the linear transformation $Y = g(X) = aX + b$, the variance of Y is

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

Equivalently, we could write $\sigma_Y^2 = a^2 \sigma_X^2$.

Proof:

$$\begin{aligned}
\mathrm{Var}[Y] &= \mathrm{Var}[aX + b] \\
&= \mathrm{E}\left[\left((aX + b) - \mathrm{E}[aX + b]\right)^2\right] \\
&= \mathrm{E}\left[\left((aX + b) - (a\,\mathrm{E}[X] + b)\right)^2\right] \\
&= \mathrm{E}\left[\left(aX - a\,\mathrm{E}[X]\right)^2\right] \\
&= \mathrm{E}\left[a^2\left(X - \mathrm{E}[X]\right)^2\right] \\
&= a^2\,\mathrm{E}\left[\left(X - \mathrm{E}[X]\right)^2\right] \\
&= a^2\,\mathrm{Var}[X]
\end{aligned}$$

□

Example 3.5.4

Given a random variable X with variance $\sigma_X^2 = 8$, determine the variance of $Y = 3X - 5$.

Example 3.5.5

Given a random variable X with variance $\sigma_X^2 = 8$, determine the variance of $Y = 5 - 3X$.

The expected values that appear in the computational formula for variance occur frequently and are given special names. $\mathrm{E}\left[X\right]$ is called the **first moment of** X and $\mathrm{E}\left[X^2\right]$ is called the **second moment of** X. In general, we have the following definition.

Definition 3.5.3

The $k^{\mathbf{th}}$ **moment of** X is defined to be the expected value of the k^{th} power of X, i.e., $\mathrm{E}\left[X^k\right]$.

These moments are important features of a random variable. For instance, knowing the first and second moments of a random variable allows us to quickly obtain the mean and variance. These moments are sometimes referred to as **moments about the origin**.

In certain situations, we may be interested in another type of moment known as a **moment about the mean**. In general, the k^{th} moment about the mean is the quantity $\mathrm{E}\left[(X - \mu)^2\right]$. Notice that the second moment about the mean is actually the variance of X.

Example 3.5.6

A voltage X is uniformly distributed on the support $\mathcal{S}_X = \{-3, -2, \ldots, 3, 4\}$.

(a) Determine the mean and variance of X.

(b) Determine the mean and variance of $Y = -2X^2 + 3$.

3.6 The Moment Generating Function (mgf)

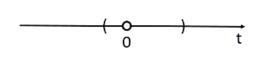
We now look at a very special function (or transformation) of a random variable X: $Y = g(X) = e^{tX}$, where t denotes any real number. This transformation turns out to be quite useful in many areas of probability theory. Notice that as the value of t changes, so does the transformation. We are especially interested in this transformation for values of t near 0.

Definition 3.6.1

The **moment generating function** of a discrete random variable X is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] = \sum_{\text{all } x} e^{tx} f(x),$$

provided the sum converges for all values of t in some open interval around $t = 0$, except possibly at $t = 0$.



Theorem 3.6.1

The first derivative of the moment generating function with respect to t, evaluated at $t = 0$, is the first moment of X. In other words,

$$M'_X(t)\Big|_{t=0} = \mathrm{E}\left[X\right].$$

Proof: Differentiate with respect to t to get $M'_X(t) = \sum_{\text{all } x} x e^{tx} f(x)$. Substituting $t = 0$, we see that

$$M'_X(t)\Big|_{t=0} = \sum_{\text{all } x} x f(x) = \mathrm{E}\left[X\right].$$

□

Theorem 3.6.2

The second derivative of the moment generating function with respect to t, evaluated at $t = 0$, is the second moment of X. In other words,

$$M''_X(t)\Big|_{t=0} = \mathrm{E}\left[X^2\right].$$

Proof: Differentiate with respect to t twice to get $M''_X(t) = \sum_{\text{all } x} x^2 e^{tx} f(x)$. Substituting $t = 0$, we see that

$$M''_X(t)\Big|_{t=0} = \sum_{\text{all } x} x^2 f(x) = \mathrm{E}\left[X^2\right].$$

□

It should be clear that this result can be generalized, as shown in the theorem below.

Theorem 3.6.3

The k^{th} derivative of the moment generating function with respect to t, evaluated at $t = 0$, is the k^{th} moment of X. In other words,

$$M^{(k)}_X(t)\Big|_{t=0} = \mathrm{E}\left[X^k\right].$$

This theorem justifies the name “moment generating function”. Note that the function $M_X(t)$ can be used to “generate the moments” of X by differentiating and substituting $t = 0$. What better name for this function than the moment generating function? Note that not all random variables have a moment generating function, since, for instance, it’s possible for the sum in the definition above to diverge.

Example 3.6.1

Suppose that the mgf of X is $M_X(t) = (1 - p + pe^t)^n$. Determine the mean and variance of of X.

Example 3.6.2

Suppose that the mgf of X is $M_X(t) = e^{4(e^t-1)}$. Determine the mean and variance of of X.

3.7 The Mean and Variance of Mixture Distributions

Suppose that we have two random variables X_1 and X_2 with probability mass functions $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. Suppose these random variables have means and variances $\mathrm{E}[X_1]$, $\mathrm{E}[X_2]$, $\mathrm{Var}[X_1]$, and $\mathrm{Var}[X_2]$. Given a number $0 < b < 1$, we can define a new random variable X with pmf $f_X(x) = bf_{X_1}(x) + (1-b)f_{X_2}(x)$. In other words, X is a **mixture** of X_1 and X_2 with support $\mathcal{S}_X = \mathcal{S}_{X_1} \cup \mathcal{S}_{X_2}$.

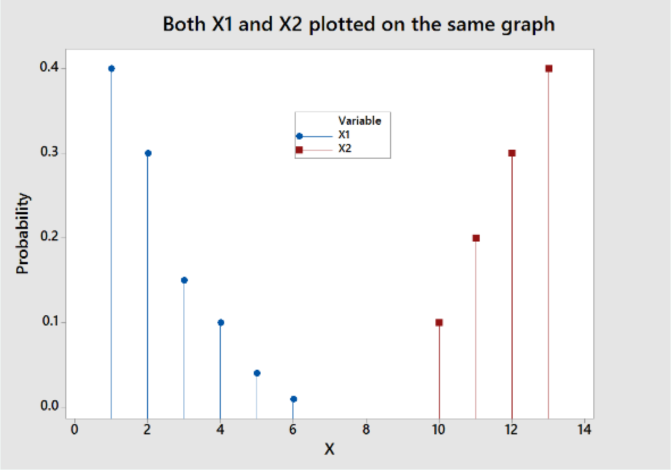
If we think of X_1 and X_2 as sub-populations, then X combines these two sub-populations into one larger population with proportion (or probability) b from X_1 and proportion (or probability) $1-b$ from X_2.

Example 3.7.1

Let X_1 and X_2 be discrete random variables with pmfs given below.

x_1	$f_{X_1}(x_1)$
1	0.40
2	0.30
3	0.15
4	0.10
5	0.04
6	0.01

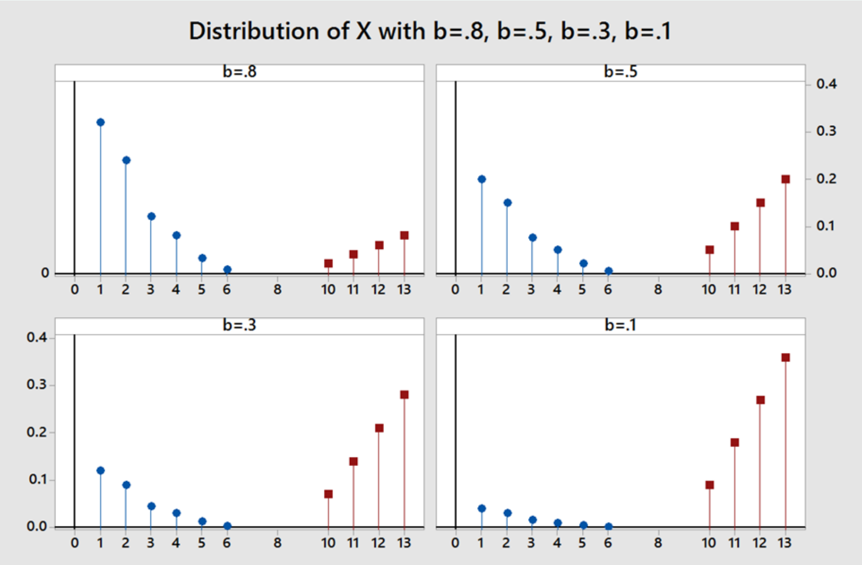
x_2	$f_{X_2}(x_2)$
10	0.10
11	0.20
12	0.30
13	0.40



We can choose a few values of b to illustrate how the distribution of X changes as b changes. It should be clear that as the value of b changes, the mean and variance of X also change.

Large values of b give more weight to X_1 and makes $\mathrm{E}[X]$ closer to $\mathrm{E}[X_1]$. Small values of b give more weight to X_2 and makes $\mathrm{E}[X]$ closer to $\mathrm{E}[X_2]$.

If b were approximately 1, then $\mathrm{Var}[X] \approx \mathrm{Var}[X_1]$. As the value of b decreases, $\mathrm{Var}[X]$ should increase since we are sliding probability away from the current mean. Of course, the variance would eventually stop increasing since the same logic holds for $b \approx 0$.



Theorem 3.7.1

For X defined as a mixture of X_1 and X_2 with parameter b as above,

$$\mathrm{E}[X] = b\,\mathrm{E}[X_1] + (1-b)\,\mathrm{E}[X_2].$$

Proof:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{x\in\mathcal{S}_X} xf(x) \\
&= \sum_{x\in\mathcal{S}_X} x\left[bf_{X_1}(x) + (1-b)f_{X_2}(x)\right] \\
&= \sum_{x\in\mathcal{S}_{X_1}} xbf_{X_1}(x) + \sum_{x\in\mathcal{S}_{X_2}} x(1-b)f_{X_2}(x) \\
&= b\sum_{x\in\mathcal{S}_{X_1}} xf_{X_1}(x) + (1-b)\sum_{x\in\mathcal{S}_{X_2}} xf_{X_2}(x) \\
&= b\,\mathrm{E}[X_1] + (1-b)\,\mathrm{E}[X_2]
\end{aligned}$$

□

Theorem 3.7.2

For X defined as a mixture of X_1 and X_2 with parameter b as above,

$$\mathrm{E}\left[X^2\right] = b\,\mathrm{E}\left[X_1^2\right] + (1-b)\,\mathrm{E}\left[X_2^2\right].$$

Proof:

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \sum_{x\in\mathcal{S}_X} x^2f(x) \\
&= \sum_{x\in\mathcal{S}_X} x^2\left[bf_{X_1}(x) + (1-b)f_{X_2}(x)\right] \\
&= \sum_{x\in\mathcal{S}_{X_1}} x^2bf_{X_1}(x) + \sum_{x\in\mathcal{S}_{X_2}} x^2(1-b)f_{X_2}(x) \\
&= b\sum_{x\in\mathcal{S}_{X_1}} x^2f_{X_1}(x) + (1-b)\sum_{x\in\mathcal{S}_{X_2}} x^2f_{X_2}(x) \\
&= b\,\mathrm{E}\left[X_1^2\right] + (1-b)\,\mathrm{E}\left[X_2^2\right]
\end{aligned}$$

□

In other words, $\mathrm{E}[X]$ is a weighted average of $\mathrm{E}[X_1]$ and $\mathrm{E}[X_2]$, and $\mathrm{E}\left[X^2\right]$ is a weighted average of $\mathrm{E}\left[X_1^2\right]$ and $\mathrm{E}\left[X_2^2\right]$.

Note that it is NOT true that $\mathrm{Var}[X] = b\,\mathrm{Var}[X_1] + (1-b)\,\mathrm{Var}[X_2]$. Instead, we must use the computational formula $\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2$ and the two theorems above to determine the variance of X.

To emphasize that the variance of X is not simply a weighted average of the variances of X_1 and X_2, we work out the following example.

Example 3.7.2

Consider a random variable X obtained as a mixture of X_1 and X_2 using $b = 0.8$.

Determine the variance of X, and show that it is NOT the weighted average of the variances of X_1 and X_2.

x_1	$f_{X_1}(x_1)$	$x_1 f_{X_1}(x_1)$	$x_1^2 f_{X_1}(x_1)$
1	0.40		
2	0.30		
3	0.15		
4	0.10		
5	0.04		
6	0.01		

x_2	$f_{X_2}(x_2)$	$x_2 f_{X_2}(x_2)$	$x_2^2 f_{X_2}(x_2)$
10	0.10		
11	0.20		
12	0.30		
13	0.40		

We now extend the idea of a mixture to more than two populations.

Theorem 3.7.3

Suppose we are forming a mixture of k populations. The probabilities of choosing populations $1, 2, \ldots, k$ are $b_1, b_2, \ldots, b_k$, respectively, where $\sum_{i=1}^{k} b_i = 1$. Let $\mathrm{E}[X_1]$, $\mathrm{E}[X_2]$, …, $\mathrm{E}[X_k]$ denote the means of the k populations, as usual. If X is the mixture of the populations using the weights provided above, then

$$\mathrm{E}[X] = b_1 \, \mathrm{E}[X_1] + b_2 \, \mathrm{E}[X_2] + \cdots + b_k \, \mathrm{E}[X_k] \, .$$

Theorem 3.7.4

Suppose we are forming a mixture of k populations. The probabilities of choosing populations $1, 2, \ldots, k$ are $b_1, b_2, \ldots, b_k$, respectively, where $\sum_{i=1}^{k} b_i = 1$. Let $\mathrm{E}\left[X_1^2\right]$, $\mathrm{E}\left[X_2^2\right]$, …, $\mathrm{E}\left[X_k^2\right]$ denote the second moments of the k populations, as usual. If X is the mixture of the populations using the weights provided above, then

$$\mathrm{E}\left[X^2\right] = b_1\,\mathrm{E}\left[X_1^2\right] + b_2\,\mathrm{E}\left[X_2^2\right] + \cdots + b_k\,\mathrm{E}\left[X_k^2\right].$$

Example 3.7.3

Suppose that we have a population that is a mixture of 3 sub-populations. The mixture proportions (or probabilities) are $b_1 = 0.5$, $b_2 = 0.3$, and $b_3 = 0.2$. Suppose $\mathrm{E}[X_1] = 20$, $\mathrm{E}[X_2] = 15$, and $\mathrm{E}[X_3] = 10$. Determine $\mathrm{E}[X]$.

Example 3.7.4

Suppose that we have a population that is a mixture of 3 sub-populations. The mixture proportions (or probabilities) are $b_1 = 0.5$, $b_2 = 0.3$, and $b_3 = 0.2$. Suppose $\mathrm{E}\left[X_1^2\right] = 410$, $\mathrm{E}\left[X_2^2\right] = 232.5$, and $\mathrm{E}\left[X_3^2\right] = 105$. Determine $\mathrm{Var}[X]$.

Example 3.7.5

Consider an experiment in which we roll a single die.

If we roll a 1, 2, or 3, we toss 20 coins and let X_1 denote the number of heads obtained in the 20 tosses.

If we roll a 4 or 5, we toss 15 coins and let X_2 denote the number of heads obtained in the 15 tosses.

If we roll a 6, we toss 10 coins and let X_3 denote the number of heads obtained in the 10 tosses.

Let X denoted the number of heads obtained.

Determine $\mathrm{E}[X]$.

Note that prior to the start of the experiment, we do not know how many coins will be tossed. To compute $\mathrm{E}[X]$ directly, we would need to determine the pmf of X. While not overly difficult, this would be time-consuming, as we'd need to calculate $P(X = x)$ for $x \in \mathcal{S}_X = \{0, 1, 2, \ldots, 20\}$.

Instead, let's use the fact that X can be viewed as a mixture of X_1, X_2, and X_3 with weights $b_1 = P(1,2,3) = \frac{3}{6}$, $b_2 = P(4,5) = \frac{2}{6}$, and $b_3 = P(6) = \frac{1}{6}$. It should be clear that $\mathrm{E}[X_1] = 10$, $\mathrm{E}[X_2] = 7.5$, and $\mathrm{E}[X_3] = 5$. Using the theorem from above, the expected number of heads is

$$\mathrm{E}[X] = \frac{3}{6}(10) + \frac{2}{6}(7.5) + \frac{1}{6}(5) = \frac{50}{6} \approx 8.333$$

To find the variance of X, we would need to determine the value of $\mathrm{E}\left[X^2\right]$, which would require the values of $\mathrm{E}\left[X_1^2\right]$, $\mathrm{E}\left[X_2^2\right]$, and $\mathrm{E}\left[X_3^2\right]$. We are not yet equipped with the ability to easily determine these values, but we will be soon.

3.8 The Binomial Distribution

The most important thing about a random variable is its distribution (i.e., its story of probability). In this section and those that follow, we study discrete random variables that have special types of distributions.

Definition 3.8.1

A **Bernoulli experiment** (or Bernoulli trial) is an experiment that has only two outcomes. The outcomes are often called "success" and "failure". Usually, our focus is on one of the two outcomes, and this is the outcome we label a success. We denote the probability of a success by $P(\text{success}) = p$ and the probability of failure by $P(\text{failure}) = q = 1 - p$.

Definition 3.8.2

A random variable X that assigns the value 1 to a success and 0 to a failure on the sample space of a Bernoulli experiment is said to be a **Bernoulli random variable**. That is, $X(\text{success}) = 1$ and $X(\text{failure}) = 0$.

Theorem 3.8.1

Let X be a Bernoulli random variable. The distribution of X is given by the pmf $f(x) = p^x q^{1-x}$ with support $x = 0, 1$.

Theorem 3.8.2

The mean and variance of a Bernoulli random variable X are $\mu_X = p$ and $\sigma_X^2 = pq$.

Proof: Using the usual formulas, the mean is $\mu_X = \sum_{\text{all } x} x f(x) = 0(q) + 1(p) = p$. The second moment is $\mathrm{E}\left[X^2\right] = \sum_{\text{all } x} x^2 f(x) = 0^2(q) + 1^2(p) = p$, so that the variance is $\sigma_X^2 = \mathrm{E}\left[X^2\right] - (\mathrm{E}\,[X])^2 = p - p^2 = p(1-p) = pq$. □

Definition 3.8.3

A **binomial experiment** is a collection of n independent Bernoulli trials with a common probability of success p. By independent trials, we mean the probability of success on any given trial does not depend on previous trials.

Example 3.8.1

Consider selecting three cards from a deck of 52 cards. The outcome of interest is drawing an ace, so we will consider getting an ace to be a success. Prior to selecting any cards, $P(\text{success}) = P(\text{ace}) = \frac{4}{52}$. If we have yet to select a card, the probability that the second card selected from our well-shuffled deck is an ace is also $\frac{4}{52}$. But once we select a card, the probability the second card is an ace will instantly change to $\frac{3}{51}$ if the first card was an ace, or it will change to $\frac{4}{51}$ if the first card was not an ace. Therefore, the three selections (or trials) are not independent, and this is NOT a binomial experiment.

Note, however, that if we were selecting cards with replacement, then it would be true that $P(\text{ace}) = \frac{4}{52}$ for each selection. So in this case, the trials would be independent and we'd be describing a binomial experiment.

Definition 3.8.4

Let X be a random variable that counts the number of successes in a binomial experiment with n trials. We refer to X as a **binomial random variable**. Notice that the support for X is $\mathcal{S}_X = \{0, 1, 2, \ldots, n\}$.

Theorem 3.8.3

The distribution of a binomial random variable is given by the pmf

$$f(x) = P(X = x) = b(x; n, p) = b(x) = \binom{n}{x} p^x q^{n-x}$$

for $x \in \mathcal{S}_X = \{0, 1, 2, \ldots, n\}$.

Proof: We need to show that the probability of x successes is $P(X = x) = b(x) = \binom{n}{x} p^x q^{n-x}$. First note that since there are n trials and x success, there must be $n - x$ failures. Consider the particular pattern in which the x successes occur in the first x trials, and the $n - x$ failures occur in the last $n - x$ trials, i.e., SS$\cdots$SFF$\cdots$F. Because the trials are independent, it's easy to see that the probability of this pattern is simply $p^x q^{n-x}$ (i.e., just multiply the probabilities of the individual outcomes since the trials are independent). It's not hard to see that any rearrangement in which there are x successes and $n - x$ failures will also have probability $p^x q^{n-x}$. The number of such rearrangements is $\binom{n}{x}$, so that $b(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}$ for $x \in \mathcal{S}_X = \{0, 1, 2, \ldots, n\}$. □

The values n and p are called the **parameters** of the distribution. These parameters distinguish one binomial distribution from another. The collection of binomial distributions is said to be a 2-parameter family of distributions. Note that q is not considered a parameter. Why?

Theorem 3.8.4

The mean of a binomial random variable X is $\mu_X = np$.

Proof:

$$\begin{aligned}
\mu_X &= \sum_{x=0}^{n} x b(x) \\
&= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x} \\
&= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!(n-x)!} p^x q^{n-x} \\
&= \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x} (np) \\
&= np \sum_{y=0}^{n-1} \frac{(n-1)!}{y!((n-1)-y)!} p^y q^{(n-1)-y} \\
&= np,
\end{aligned}$$

where we have made the substitution $y = x - 1$ and used the facts (i) $n - x = (n-1) - y$ and (ii) the sum in the second-to-last line can be viewed as the sum of the probabilities of a binomial random variable Y with parameters $n - 1$ and p, and so must be 1. □

Theorem 3.8.5

The variance of a binomial random variable X is $\sigma_X^2 = npq = np(1-p)$.

Proof: The same manipulation as used in the proof above can be used here to determine $\mathrm{E}\left[X^2\right]$. Then use $\mathrm{Var}\left[X\right] = \mathrm{E}\left[X^2\right] - \left(\mathrm{E}\left[X\right]\right)^2$ to determine the variance of X. □

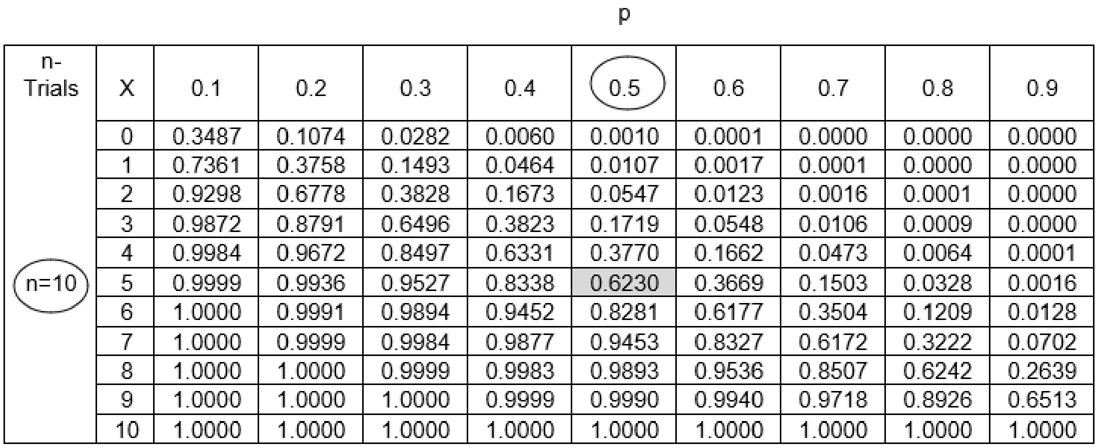
Example 3.8.2

A single die is to be rolled 10 times. Let X count the number of fives rolled. Determine $P(X = 3)$, as well as the mean and variance of X.

Example 3.8.3

A coin is weighted so that $P(\text{H}) = 0.6$. The coin will be tossed 25 times. Let X count the number of heads obtained in the 25 tosses. Determine $P(X = 15)$, as well as the mean and variance of X.

Binomial CDF charts can be very useful for determining certain probabilities. Our CDF notation for the binomial distribution will be $B(x; n, p) = B(x) = P(X \leq x)$. The CDF for certain combinations of n and p are provided in tabular format. To use these charts, first identify the section corresponding to the appropriate value of n. Then find the column that is labeled with the appropriate value of p. Finally, look up the value corresponding to the desired x-value.



		p								
n-Trials	X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
n=10	0	0.3487	0.1074	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000
	1	0.7361	0.3758	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	0.0000
	2	0.9298	0.6778	0.3828	0.1673	0.0547	0.0123	0.0016	0.0001	0.0000
	3	0.9872	0.8791	0.6496	0.3823	0.1719	0.0548	0.0106	0.0009	0.0000
	4	0.9984	0.9672	0.8497	0.6331	0.3770	0.1662	0.0473	0.0064	0.0001
	5	0.9999	0.9936	0.9527	0.8338	0.6230	0.3669	0.1503	0.0328	0.0016
	6	1.0000	0.9991	0.9894	0.9452	0.8281	0.6177	0.3504	0.1209	0.0128
	7	1.0000	0.9999	0.9984	0.9877	0.9453	0.8327	0.6172	0.3222	0.0702
	8	1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.6242	0.2639
	9	1.0000	1.0000	1.0000	0.9999	0.9990	0.9940	0.9718	0.8926	0.6513
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Example 3.8.4

A fair coin is tossed 10 times. Determine the probability that we obtain 5 or fewer heads.

Example 3.8.5

A fair coin is tossed 10 times. Let X count the number of heads obtained in the 10 tosses. Determine $P(3 \leq X \leq 7)$.

Example 3.8.6

An audio player uses a low-quality hard drive. The probability that the hard drive fails after being used for one month is $\frac{1}{12}$. If it fails, the manufacturer offers a free-of-charge repair for the customer. The manufacturer offers this repair service for a period of one year, during which the customer can ask for repair up to 12 times (once per month).

(a) Let X be the number of repairs in one year for a randomly selected hard drive. Determine the pmf of X.

(b) The initial build of the hard drive costs \$50 and each repair costs \$20. What is the average or expected cost per hard drive?

3.9 The Geometric and Negative Binomial Distributions

In the last section, we considered experiments in which repeated independent Bernoulli trials were performed. The number of trials, denoted by n, was fixed in advance, and the number of successes X in the binomial experiment was the random variable (i.e., the number of successes varies randomly).

In this section, we consider repeated independent Bernoulli trials in which the number of successes, denoted by r, is fixed in advance (so r is a parameter), and the number of trials that it takes to get the desired number of successes is a random variable X. This type of random variable is called a **negative binomial random variable**, and will be defined more formally below.

In other words, a binomial random variable counts the number of successes in a fixed number of trials, while a negative binomial random variable counts the number of trials it takes to observe a fixed number of successes.

Definition 3.9.1

Consider an experiment in which we perform repeated independent Bernoulli trials until we obtain the r^{th} success, where the probability of success is p for each trial. Such an experiment is called a **negative binomial experiment**.

The **negative binomial random variable** X counts the number of trials it takes to get the r^{th} success. In this case, we say that X follows the **negative binomial distribution**.

We temporarily postpone the discussion of the general properties of the negative binomial distribution while we consider the most elementary case, the case where $r = 1$.

Definition 3.9.2

Consider an experiment in which we perform repeated independent Bernoulli trials until we obtain the 1^{st} success, where the probability of success is p for each trial. Such an experiment is called a **geometric experiment**.

The **geometric random variable** X counts the number of trials it takes to get the 1^{st} success. In this case, we say that X follows the **geometric distribution**.

Note 3.9.1

Note that a geometric random variable is really a special type of negative binomial random variable for which $r = 1$. In other words, the collection of all geometric random variables is a subset of the collection of all negative binomial random variables.

Theorem 3.9.1

Let X be a discrete random variable that follows the geometric distribution. The distribution of X is given by the pmf

$$f(x) = P(X = x) = pq^{x-1},$$

with support $\mathcal{S}_X = \{1, 2, 3, \ldots\}$.

Proof: If the first success occurs on trial x, then the first $x - 1$ trials must have been failures. So $f(x) = P(X = x)$ is just the probability of the sequence FF$\cdots$FS, where the first $x - 1$ trials are failures. Since the trials are independent, it should be clear that the probability of this sequence is $f(x) = P(X = x) = pq^{x-1}$ for $x = 1, 2, 3, \ldots$. □

It may not be obvious that the above function defines a valid pmf, but note that the sum of the probabilities forms a geometric series with sum 1:

$$\begin{aligned}
\sum_{x=1}^{\infty} P(X = x) &= p + pq + pq^2 + pq^3 + \cdots \\
&= p(1 + q + q^2 + q^3 + \cdots) \\
&= \frac{p}{1-q} \\
&= \frac{p}{p} \\
&= 1,
\end{aligned}$$

where we have used the fact that $|q| = q < 1$ since q is the probability of failure.

Theorem 3.9.2

The mean of a geometric random variable X is $\mu_X = \dfrac{1}{p}$.

Proof: The usual formula for the mean of a random variable gives $\mu_X = \displaystyle\sum_{x=1}^{\infty} xpq^{x-1}$. Trying to determine this sum directly is not as straightforward as we'd like. Instead, we will use a somewhat common technique in probability theory. Note that $\displaystyle\sum_{x=1}^{\infty} q^x = \frac{q}{1-q}$, as the former is a convergent geometric series. Differentiate both sides of this equation with respect to q to obtain $\displaystyle\sum_{x=1}^{\infty} xq^{x-1} = \frac{1}{(1-q)^2} = \frac{1}{p^2}$. Multiply both sides of this equation by p to obtain $\mu_X = \displaystyle\sum_{x=1}^{\infty} xpq^{x-1} = \frac{p}{p^2} = \frac{1}{p}$. □

While we omit the details, a similar technique could be used to obtain the second moment $\mathrm{E}\left[X^2\right]$, which could then be used to determine the variance of X.

Theorem 3.9.3

The variance of a geometric random variable X is $\sigma_X^2 = \dfrac{1-p}{p^2}$.

Note 3.9.2

Care must be taken if you are consulting other textbooks or resources regarding the geometric (or negative binomial) distribution. Here, we have defined a geometric random variable X as the number of trials it takes to obtain the 1^{st} success. Some texts define a geometric random variable as the number of failures until the 1^{st} success (rather than the number of trials). Note that such a random variable would have a value 1 less than the value of our geometric random variable. While the pmf would essentially be defined in the same way, we'd have to carefully consider the effect on the support, mean, and variance of the random variable. It should be obvious that the support would be $\{0, 1, 2, \ldots, \}$ rather than $\{1, 2, 3, \ldots, \}$ and the mean would be $\frac{1}{p} - 1$ rather than $\frac{1}{p}$. What would be the effect on the variance?

Now that we have considered the special case where $r = 1$, we return to our study of the general case.

Theorem 3.9.4

Let X be a negative binomial random variable with parameters r and p, as described above. The distribution of X is given by the pmf

$$f(x) = P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

for $x = r, r+1, r+2, \ldots$.

Proof: First notice that the support is $x = r, r+1, r+2, \ldots$ because in order to obtain r successes, the number of trials has to be at least r.

To determine $P(X = x)$, we note that the only way the r^{th} success can occur on the x^{th} trial is if there were exactly $r-1$ successes in the first $x-1$ trials. If there were more than $r-1$ successes in the first $x-1$ trials, then the experiment would have ended prior to trial x (since the experiment ends once the r^{th} success is obtained). If there were less than $r-1$ successes in the first $x-1$ trials, there is no way the r^{th} success could occur on the x^{th} trial since there would be fewer than $r-1$ successes with only one trial to go.

Below is an example of an outcome that would satisfy $X = x$. Note that any outcome in which success r occurs on trial x requires $r - 1$ successes in the first $x - 1$ trials.

$$\overbrace{\underline{S}\ \underline{S}\ \underline{F}\ \underline{F}\ \underline{S}\cdots\underline{S}\ \underline{F}}^{\text{First x-1 slots: Exactly r-1 successes}}\ \ \underline{S}$$

The number of successes in the first $x - 1$ trials can be viewed as a binomial random variable Y with $n = x - 1$ and success probability p. The probability of $r - 1$ successes in the first $x - 1$ trials is then

$$P(Y = r - 1) = \binom{x-1}{r-1} p^{r-1}(1-p)^{x-r}.$$

Since the final trial has to be a success, we simply multiply the above by the success probability p to obtain the desired pmf for the negative binomial random variable X:

$$\begin{aligned}
f(x) &= P(X = x) \\
&= P(r - 1 \text{ successes in the first } x - 1 \text{ trials} \cap \text{success on trial } x) \\
&= P(Y = r - 1) \cdot P(\text{success on trial } x) \\
&= \binom{x-1}{r-1} p^{r-1}(1-p)^{x-r} \cdot p \\
f(x) &= \binom{x-1}{r-1} p^{r}(1-p)^{x-r}
\end{aligned}$$

for $x = r, r + 1, r + 2, \ldots$. □

Example 3.9.1

A die is to be rolled until the 3rd two is obtained. Determine the probability that the 3rd two occurs on the 8th roll.

While we omit the proofs of the following two theorems, these results can easily be obtained using methods similar to those used in the proofs of some of our previous results.

Theorem 3.9.5

The mean of a negative binomial random variable X is $\mu_X = \dfrac{r}{p}$.

Theorem 3.9.6

The variance of a negative binomial random variable X is $\sigma_X^2 = \dfrac{r(1-p)}{p^2}$.

3.10 The Hypergeometric Distribution

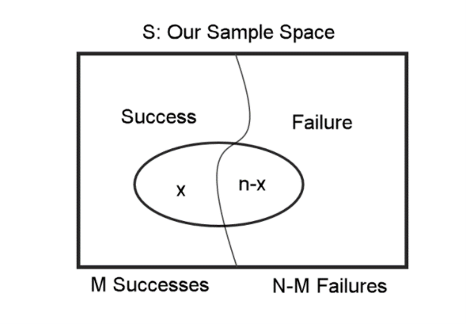
Definition 3.10.1

Suppose we have a finite sample space with N outcomes that consist of M successes (or Type 1 outcomes) and $N - M$ failures (or Type 2 outcomes).

We now select n outcomes without replacement.

Let X count the number of successes (or Type 1 outcomes).

The random variable X is called a **hypergeometric random variable** and is said to follow a **hypergeometric distribution**.



Theorem 3.10.1

The pmf of a hypergeometric random variable is given by

$$h(x; N, M, n) = h(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

The support depends on the values of the parameters N, M, and n.

We have actually already utilized the above pmf in some of our counting problems in Chapter 2. An example of such a problem is given below.

Theorem 3.10.2

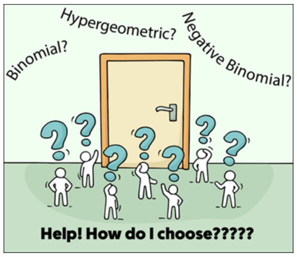
The mean of a hypergeometric random variable X is $\mu_X = \dfrac{nM}{N}$.

Theorem 3.10.3

The variance of a hypergeometric random variable X is $\sigma_X^2 = n\left(\dfrac{M}{N}\right)\left(\dfrac{N-M}{N}\right)\left(\dfrac{N-n}{N-1}\right)$.

Example 3.10.1

Suppose we are to draw 7 cards from a standard deck. Determine the probability that we obtain 5 hearts. Also determine the mean and variance of the number of hearts selected.



Note 3.10.1

There are obvious similarities and subtle differences amongst our three new distributions (binomial, negative binomial, and hypergeometric). It is common to get confused and frustrated when deciding which type of distribution to use in a given situation. The strategy below can be used to help decide which distribution is most appropriate to use when solving a problem.

Start by carefully deciding what the variable is.

- If the variable is the **number of trials**, then use the **negative binomial distribution**.
- If the variable is the **number of successes**, then do one of the following:
  - Use the **binomial distribution** if the trials are independent (i.e., trials with replacement).
  - Use the **hypergeometric distribution** if the trials are not independent (i.e., trials without replacement).

3.11 The Poisson Distribution

Definition 3.11.1

A **counting process** is a process that counts events (arrivals) over a fixed length time period. Examples of counting processes include the following everyday scenarios:

- counting the number of customers entering a building in a 20-minute period
- counting the number of phone calls coming into a call center in a 10-minute period
- counting the number of cars passing through an intersection in a 5-minute period

Definition 3.11.2

A counting process that counts the number of events in a time interval of length t that satisfies the three conditions below is said to be a **Poisson counting process**:

- the rate of events is constant over time (the rate is denoted by λ)
- the number of events that occur in disjoint time intervals are independent
- the probability that two or more events occur in a time interval of length t goes to 0 as the length of the interval goes to 0

Consider the following counting process that does not satisfy all three of the above conditions and so is NOT a Poisson counting process.

Example 3.11.1

An event is defined to occur when an e-mail is received from a student. Over the next 10 hours, each of 500 students in a course must send the instructor an e-mail. The theoretical arrival rate is 50 e-mails per hour. Suppose that 230 e-mails arrive in the first hour. Since there are 270 emails that need to arrive in the next 9 hours, the theoretical arrival rate decreases to 30 e-mails per hour. Thus, the probability of an arrival in the next hour is dependent on what happened in the past. This is certainly a counting process, but does not satisfy the conditions necessary to be considered a Poisson counting process.

Theorem 3.11.1

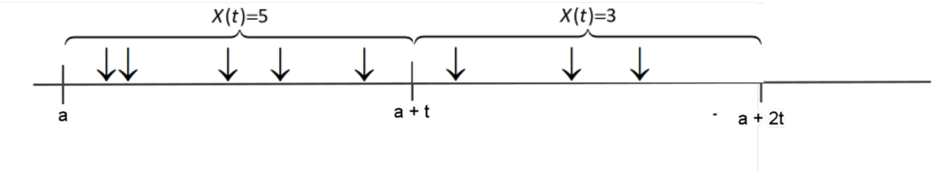
Consider a Poisson counting process, i.e., a counting process that satisfies the three conditions listed above.

Let X be a random variable that counts the number of arrivals in the Poisson process during a time interval of length t. Let λ denote the rate of arrivals, i.e., the average number of arrivals in an interval of length t.

X is said to be a **Poisson random variable** and follows the **Poisson distribution** described by the pmf

$$P(X = x) = \pi(x; \lambda) = \pi(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for $x = 0, 1, 2, 3, \ldots$.



Note that λ is a theoretical arrival rate. Not every time period of length t will have λ arrivals. Rather, we can view λ as the average number of arrivals if we were to observe the process over many time periods of length t.

Example 3.11.2

Suppose the random variable X follows a Poisson distribution with rate parameter $\lambda = 1.4$. Determine $P(X = 2)$.

We should check that the given pmf defines a valid probability distribution. That $\pi(x) \geq 0$ should be clear. The fact that the sum of the probabilities is 1 is less obvious and is shown in the proof of the next theorem.

Theorem 3.11.2

For a Poisson distribution with rate parameter λ and pmf $\pi(x)$, we have

$$\sum_{x=0}^{\infty} \pi(x) = 1.$$

Proof:

$$\begin{aligned} \sum_{x=0}^{\infty} \pi(x) &= \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} e^{\lambda} \\ &= 1, \end{aligned}$$

where we have made use of the Taylor series for the function e^x from calculus, namely that $e^x = \displaystyle\sum_{n=0}^{\infty} \frac{x^n}{n!}$. □

Theorem 3.11.3

The mean and variance of a Poisson random variable X are $\mu_X = \lambda$ and $\sigma_X^2 = \lambda$, respectively.

Proof: We determine the mean of X as follows:

$$\begin{aligned} \mathrm{E}\left[X\right] &= \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\ &= \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} \\ &= \lambda \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \lambda(1) \\ &= \lambda, \end{aligned}$$

where we have made the substitution $y = x - 1$ in the third-to-last line and used the result of the previous theorem, i.e., the sum of the values of the pmf is equal to 1.

Next, we determine $\mathrm{E}\left[X^2\right]$ as follows:

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \sum_{x=0}^{\infty} x^2 \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\
&= \lambda \sum_{x=1}^{\infty} x \cdot \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} \\
&= \lambda \sum_{y=0}^{\infty} (y+1) \cdot \frac{\lambda^y e^{-\lambda}}{y!} \\
&= \lambda \sum_{y=0}^{\infty} y \cdot \frac{\lambda^y e^{-\lambda}}{y!} + \lambda \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} \\
&= \lambda(\lambda) + \lambda(1) \\
&= \lambda^2 + \lambda,
\end{aligned}$$

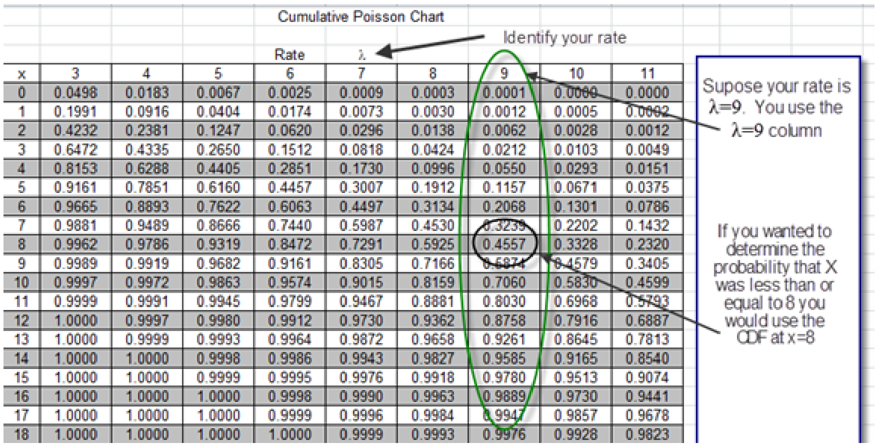
where we have made the substitution $y = x-1$ in the fourth-to-last line and used that the two series in third-to-last line sum to the mean of a Poisson distribution with rate parameter λ and 1, respectively.

Finally, the variance of X can be calculated as $\mathrm{Var}\left[X\right] = \mathrm{E}\left[X^2\right] - \left(\mathrm{E}\left[X\right]\right)^2 = \lambda^2 + \lambda - (\lambda)^2 = \lambda$. □

Example 3.11.3

To determine $P(X \leq 2)$ for a Poisson random variable X, we would need to add up $\pi(0) + \pi(1) + \pi(2)$. Instead, we can make use of CDF charts, if available. Note that since the support of a Poisson random variable is infinite and there are an infinite number of possible values of the parameter λ, there are infinitely many CDF charts. Obviously a textbook cannot include all CDF charts, so only a selection of those possible are included.

Using a CDF chart is very straightforward. First, identify the column or section that contains the relevant value of the rate parameter λ. Note that if the value of λ needed is not found in the chart, then you must use the distribution itself (or some type of software) to solve the problem. Above, we used the notation $\pi(x)$ for the pmf of a Poisson distribution. We will change to a capital letter and use $\Pi(x)$ when working with the CDF of a Poisson distribution.



Cumulative Poisson Chart

Identify your rate

Rate λ

x	3	4	5	6	7	8	9	10	11
0	0.0498	0.0183	0.0067	0.0025	0.0009	0.0003	0.0001	0.0000	0.0000
1	0.1991	0.0916	0.0404	0.0174	0.0073	0.0030	0.0012	0.0005	0.0002
2	0.4232	0.2381	0.1247	0.0620	0.0296	0.0138	0.0062	0.0028	0.0012
3	0.6472	0.4335	0.2650	0.1512	0.0818	0.0424	0.0212	0.0103	0.0049
4	0.8153	0.6288	0.4405	0.2851	0.1730	0.0996	0.0550	0.0293	0.0151
5	0.9161	0.7851	0.6160	0.4457	0.3007	0.1912	0.1157	0.0671	0.0375
6	0.9665	0.8893	0.7622	0.6063	0.4497	0.3134	0.2068	0.1301	0.0786
7	0.9881	0.9489	0.8666	0.7440	0.5987	0.4530	0.3239	0.2202	0.1432
8	0.9962	0.9786	0.9319	0.8472	0.7291	0.5925	0.4557	0.3328	0.2320
9	0.9989	0.9919	0.9682	0.9161	0.8305	0.7166	0.5874	0.4579	0.3405
10	0.9997	0.9972	0.9863	0.9574	0.9015	0.8159	0.7060	0.5830	0.4599
11	0.9999	0.9991	0.9945	0.9799	0.9467	0.8881	0.8030	0.6968	0.5793
12	1.0000	0.9997	0.9980	0.9912	0.9730	0.9362	0.8758	0.7916	0.6887
13	1.0000	0.9999	0.9993	0.9964	0.9872	0.9658	0.9261	0.8645	0.7813
14	1.0000	1.0000	0.9998	0.9986	0.9943	0.9827	0.9585	0.9165	0.8540
15	1.0000	1.0000	0.9999	0.9995	0.9976	0.9918	0.9780	0.9513	0.9074
16	1.0000	1.0000	1.0000	0.9998	0.9990	0.9963	0.9889	0.9730	0.9441
17	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9947	0.9857	0.9678
18	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9976	0.9928	0.9823

Supose your rate is λ=9. You use the λ=9 column

If you wanted to determine the probability that X was less than or equal to 8 you would use the CDF at x=8

Example 3.11.4

Suppose the rate parameter is $\lambda = 9$ for a Poisson random variable X. Determine $P(X \leq 8)$ and $P(X = 6)$.

Example 3.11.5

Determine $P(7 \leq X)$ for a Poisson random variable with $\lambda = 4$.

An important property of a Poisson distribution is that the rate parameter can be "scaled" to fit a desired time interval. For instance, if the rate of arrivals in a Poisson process is 120 per hour, we can scale this (if necessary) to 60 arrivals per 30 minutes, 30 arrivals per 15 minutes, 2 arrivals per minute, etc.

Theorem 3.11.4 The Scaling Property

If the number of events in a time interval of length t follows a Poisson distribution with mean λ, then the number of events in a time period of length kt follows a Poisson distribution with mean $k\lambda$.

Example 3.11.6

Suppose that X counts the number of arrivals in a 20-minute period in a Poisson process with $\lambda = 8$. Thus, the rate for X is 8 per 20-minute time period. Let Y count the number of arrivals in a 25-minute time period. Determine $\mathrm{E}[Y]$ and $P(Y = 6)$.

Example 3.11.7

In the context of the previous example, determine $P(9 \leq Y)$.

Example 3.11.8

The number of patients requiring a certain type of treatment in a given day is well-modeled by a Poisson distribution with a mean rate of $\lambda = 6$.

The facility currently has 4 machines that are required for the treatment.

Each machine can be used on 2 patients per day.

Determine the probability that the facility has enough machines to treat all patients on a given day.

How many machines would the facility need if they wanted the probability of having enough machines to treat all patients on a given day to exceed 0.99?

Example 3.11.9

A person tosses a coin every 10 seconds. Each time a heads is tossed, the person rings a bell. Let X count the number of rings per minute. Clearly X has a binomial distribution with parameters $n = 6$ and $p = 0.5$.

Does X have a Poisson distribution, too? All three conditions seem to be met as (i) we have a constant arrival rate of three rings per minute; (ii) we have independent increments (i.e., the probability the bell is rung in one period is unaffected by the events of other periods); and (iii) there is no way to get two or more rings in a very short period of time.

At first glance, it may seem that we could model the random variable X using either a binomial distribution with $n = 6$ and $p = 0.5$ or a Poisson distribution with $\lambda = 3$. Below, we compare the pmfs of these two distributions:

x	$b(x; n = 6, p = 0.5)$	$\pi(x; \lambda = 3)$
0	0.0156	0.0498
1	0.0938	0.1494
2	0.2344	0.2240
3	0.3125	0.2240
4	0.2344	0.1680
5	0.0938	0.1008
6	0.0156	0.0504
7	0	0.0216
8	0	0.0081
9	0	0.0027
10	0	0.0008

The two distributions $b(x; n = 6, p = 0.5)$ and $\pi(x; \lambda = 3)$ are clearly very different. So it appears that X cannot be modeled by both distributions. So which is correct? Consider the support of X. If X is modeled by a binomial distribution with $n = 6$, the support is $x = 0, 1, 2, 3, 4, 5, 6$, which agrees with the context. But if X is modeled by a Poisson distribution, the support is $x = 0, 1, 2, \ldots$, which disagrees with the context. For instance, the bell cannot be rung more than 6 times in one minute. So the distribution of X cannot be modeled by a Poisson distribution. Which of the three Poisson conditions were violated, then?

Suppose instead that the person is tossing a weighted coin with $P(\mathrm{H}) = 0.25$ every 5 seconds, i.e., the probability of a head has been cut in half and the number of tosses per minute has doubled. Again, let X count the number of rings per minute. Like earlier, it still seems reasonable to model X with a binomial distribution, but this time the parameters are $n = 12$ and $p = 0.25$. It still is not clear why modeling X with a Poisson distribution with $\lambda = 3$ is not appropriate, so we again consider this distribution as well. Below, we compare the pmfs of these two distributions (and also include the binomial distribution considered earlier for comparison):

x	$b(x; n = 6, p = 0.5)$	$b(x; n = 12, p = 0.25)$	$\pi(x; \lambda = 3)$
0	0.0156	0.0317	0.0498
1	0.0938	0.1267	0.1494
2	0.2344	0.2323	0.2240
3	0.3125	0.2581	0.2240
4	0.2344	0.1936	0.1680
5	0.0938	0.1032	0.1008
6	0.0156	0.0401	0.0504
7	0	0.0115	0.0216
8	0	0.0024	0.0081
9	0	0.0004	0.0027
10	0	0	0.0008

Note that the binomial distribution $b(x; n = 12, p = 0.25)$ seems to be "closer" to the Poisson distribution $\pi(x; \lambda = 3)$, but there are still noticeable differences in the probabilities, as well as the support.

Consider another modification where the probability of a head p is reduced, and the number of flips per minute n is increased, but in such a way that $np = 3$ remains fixed. For instance, above we used $(n = 6, p = 0.5)$ and $(n = 12, p = 0.25)$, but now we also consider $(n = 30, p = 0.1)$, $(n = 60, p = 0.05)$, $(n = 600, p = 0.005)$, and $(n = 6{,}000, p = 0.0005)$. Note that these new values of n correspond to tossing the coin every 2 seconds, every 1 second, every one-tenth of a second, and every one-hundredth of a second, respectively. The pmfs of these distributions (rounded to 4 decimal places) are provided in the table below (along with the pmfs of the binomial distributions considered earlier):

x	$b(x; 6, 0.5)$	$b(x; 12, 0.25)$	$b(x; 30, 0.1)$	$b(x; 60, 0.05)$	$b(x; 600, 0.005)$	$b(x; 6{,}000, 0.0005)$	$\pi(x; \lambda = 3)$
0	0.0156	0.0317	0.0424	0.0461	0.0494	0.0497	0.0498
1	0.0938	0.1267	0.1413	0.1455	0.1490	0.1493	0.1494
2	0.2344	0.2323	0.2277	0.2259	0.2242	0.2241	0.2240
3	0.3125	0.2581	0.2361	0.2298	0.2246	0.2241	0.2240
4	0.2344	0.1936	0.1771	0.1724	0.1685	0.1681	0.1680
5	0.0938	0.1032	0.1023	0.1016	0.1009	0.1008	0.1008
6	0.0156	0.0401	0.0474	0.0490	0.0503	0.0504	0.0504
7	0	0.0115	0.0180	0.0199	0.0214	0.0216	0.0216
8	0	0.0024	0.0058	0.0069	0.0080	0.0081	0.0081
9	0	0.0004	0.0016	0.0021	0.0026	0.0027	0.0027
10	0	0.0000	0.0004	0.0006	0.0008	0.0008	0.0008
11	0	0.0000	0.0001	0.0001	0.0002	0.0002	0.0002
12	0	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001
13	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
14	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
15	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
16	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
17	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
18	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
19	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
20	0	0	0.0000	0.0000	0.0000	0.0000	0.0000

Based on the results above, it seems like the Poisson distribution is a good estimate of the binomial distribution if n is large and p is small. The reason we initially saw such a discrepancy is based on the way the events (ringing the bell) we able to occur. Initially, when $n = 6$, the events could only occur at time multiples of 10 seconds, i.e., at 10 seconds, 20 seconds, ..., 60 seconds, since the coin was tossed every 10 seconds. In other words the event could only occur at specific discrete time points. A Poisson distribution is not a good model for a situation like this, because in a Poisson process, events are able to occur in a continuous way, i.e., the bell should be able to be rung at any time in the interval $[0, 60]$. So as n increased, the frequency of the coin tossing increased, which allowed the bell to be rung at many more possible times. In other words, the set of "discrete" possibilities we initially had started to approach the "continuous" set of possibilities needed in a Poisson process. This is why the binomial probabilities began to approach the Poisson probabilities as n increased. And of course, comparing these probabilities wouldn't exactly make sense if the expected number of events differed, which is why p was decreased so that the expected number of events remained fixed at 3 in all of the distributions considered.

Theorem 3.11.5

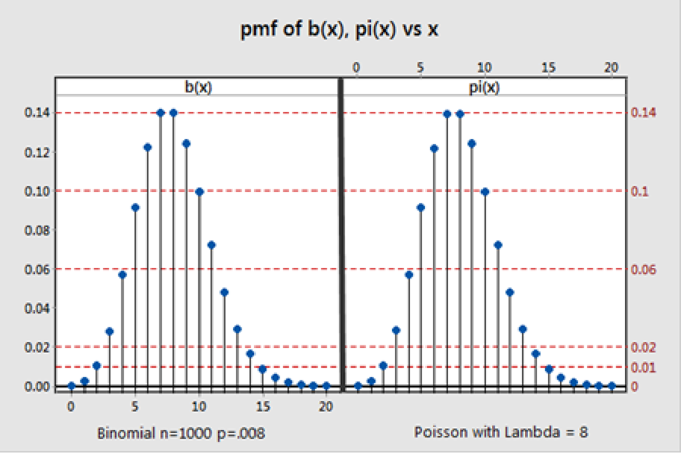
Let the binomial parameters n and p be such that the product np remains constant at some value λ. Then the binomial distribution becomes more and more similar to the Poisson distribution with parameter λ as n gets large:

$$\lim_{n\to\infty} b(x; n, p) = \pi(x; \lambda = np).$$

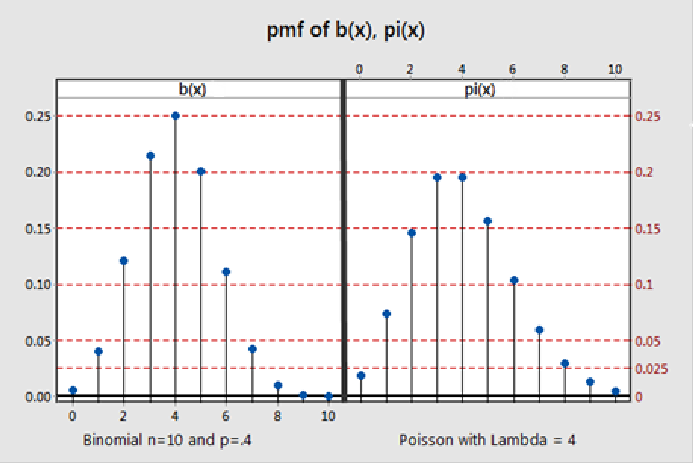
The reason this theorem is important is that when n is large, performing a binomial calculation can be very difficult. Even a computer can struggle with certain calculations, and many calculators are unable to handle very large n. The Poisson calculation, on the other hand, is quite straightforward, even when n is large.

For instance, to compute $b(8; n = 6{,}000, p = 0.0005) = 0.0081$ on a calculator, you'd need to be able to do $\binom{6{,}000}{8}$. But notice that calculating $\pi(8; \lambda = 3) = \frac{e^{-3} \cdot 3^8}{8!} = 0.0081$ is no problem. In certain situations, for large x, a computer may also have trouble computing a binomial probability, but can quickly calculate any Poisson probability.

Below are side-by-side plots of a binomial distribution with $n = 1{,}000$ and $p = 0.008$ and a Poisson distribution with $\lambda = np = 8$. In this case, n is large and p is small. Visually, we see that the probabilities match almost perfectly. Based on this, it would seem reasonable to use the Poisson distribution to approximate the binomial distribution.



Below are side-by-side plots of a binomial distribution with $n = 10$ and $p = 0.4$ and a Poisson distribution with $\lambda = np = 4$. In this case, n is not large and p is not small. Visually, we see that the probabilities are very different. Based on this, it would seem like a poor decision to use the Poisson distribution to approximate the binomial distribution. But in this case, since n is not large, it's not difficult to calculate directly using the binomial distribution, so the Poisson approximation is unnecessary.



Example 3.11.10

Suppose we have a binomial distribution problem with $n = 60{,}000$ and $p = 0.00031$. What value of λ should we use if we want to approximate the binomial distribution with a Poisson distribution?

Example 3.11.11

Nationwide, a certain disease occurs in only a small proportion of the population ($p = 0.00024$). A certain town has 50,000 residents. Use an appropriate Poisson approximation to the binomial distribution to answer the questions below:

(a) What is the expected number of residents with the disease?

(b) Determine the probability that the town has 4 or fewer residents with the disease.

(c) Determine the number of disease “cut-offs” that provide us with an approximate 95% to 5% split. Note that we can interpret these numbers as the cut-offs that define an “abnormally high” number of cases and an “abnormally low” number of cases.

(d) Suppose the town has 13 residents with the disease. Is this unusual?

(e) Suppose the town has 21 residents with the disease. Is this unusual?

Example 3.11.12

A binary communication channel has a probability of bit error of $p = 10^{-6}$. Suppose that transmission occurs in blocks of 10,000 bits. Let X be the number of errors introduced by the channel in a transmission block.

(a) Determine the pmf of X.

(b) Determine $P(X = 0)$ and $P(X \leq 3)$.

(c) For what value of p will the probability of at least 1 error in a block be 99%?