

4.1: Random Variables

Learning Objectives

- Define and construct random variables
- Establish that the sum of probabilities across all values of a random variable is 1
- Define and distinguish discrete random variables from continuous random variables
- Introduce the discrete uniform distribution

□ [Section 4.1 Excel File](#) (contains all of the data sets for this section)

Review and Preview

In inferential statistics, we seek to understand a population by studying a sample randomly selected from it. In particular, we are trying to estimate the value of a population parameter based on our study of a random sample, including many sample statistics. Hopefully, at this point, we recognize that random sampling is a random experiment; when random sampling is conducted, the outcome is a particular sample, and sample statistics are values computed from the outcome. The distinction between the outcome of random sampling (the sample) and the values that describe the outcome (statistics) is essential to remember. We are interested in the likelihood that our sample is representative of the population. To determine this likelihood, we consider all possible values that sample statistics may take on and the probabilities of such values occurring. As such, we can understand sample statistics as random variables.

It is crucial to grasp that, when we view a sample statistic as a random variable, we are not just seeing a number. We are exploring a method that can generate a value from any random sample, and diving into the realm of all the possible values that could be produced. This shift in perspective is not just a technicality, but a fundamental concept that necessitates understanding.

Random Variables

A **random variable** is a quantitative variable that assigns a number to each outcome in the sample space of a given random experiment. We generally denote random variables using capital letters, like X , and the particular values that they take on (the values that are assigned to the outcomes of a random experiment) with the same letter but lowercase and with indices: $x_1, x_2, x_3, x_4, \dots, x_n$ (if there are n possible values). As we indicated above, we are most interested in connecting the values of a random variable with the probability that they occur. The values of the random variable, together with their probabilities, form the **probability distribution** of the random variable. In general, our interest lies in the probability distributions of random variables, and as you will see, we have studied some of them already.

Consider a familiar example of a random experiment: rolling two fair dice. There are many different ways in which we could construct a random variable. One intuitive way is to consider the pairs of values and then assign each of the 36 outcomes in the sample space a number. For our first example, let X be the "sum of the two values that land face up." This process allows us to assign a number to each outcome in our sample space, illustrated below.











































						
						
						
						
						
						

Figure 4.1.1: Sum of the two values that land face up when rolling two fair dice

As we can see in the figure above, the random variable X has 11 possible values: $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Our next task is to determine the probabilities for each value that X can be. For our example, we have already computed some of the probabilities; see Text Exercises 3.1.3.3 and 3.3.2.1. By selecting a value of our random variable, say $X = 3$, we define a specific

event in the sample space of our random experiment (namely, rolling a 1 and 2) and rely on the content of the previous chapter. So $P(X = 3) = P(1 \text{ and } 2) = P(1 \text{ FIRST and } 2 \text{ SECOND or } 2 \text{ FIRST and } 1 \text{ SECOND}) = \frac{1}{36} + \frac{1}{36} - \frac{0}{36} = \frac{2}{36} = \frac{1}{18}$. We encourage the reader to confirm each of the probabilities in the table below.

Table 4.1.1: Probability distribution of the random variable X

$X = x_j$	$P(X = x_j)$
2	$\frac{1}{36} \approx 2.7778\%$
3	$\frac{2}{36} = \frac{1}{18} \approx 5.5556\%$
4	$\frac{3}{36} = \frac{1}{12} \approx 8.3333\%$
5	$\frac{4}{36} = \frac{1}{9} \approx 11.1111\%$
6	$\frac{5}{36} \approx 13.8889\%$
7	$\frac{6}{36} = \frac{1}{6} \approx 16.6667\%$
8	$\frac{5}{36} \approx 13.8889\%$
9	$\frac{4}{36} = \frac{1}{9} \approx 11.1111\%$
10	$\frac{3}{36} = \frac{1}{12} \approx 8.3333\%$
11	$\frac{2}{36} = \frac{1}{18} \approx 5.5556\%$
12	$\frac{1}{36} \approx 2.7778\%$

? Text Exercise 4.1.1

When rolling a pair of fair dice, each die lands with a value face up. Construct the probability distribution for the random variable Y , defined to be "the maximum value of the two dice rolled."

Answer

Visualizing the sample space may help identify possible values and determine probabilities for some random variables. We will want to use deeper reasoning when our sample spaces are much larger. We use both visualization and reasoning in this example. To construct a probability distribution, we must first determine the possible values of our random variable Y and then determine their probabilities.



















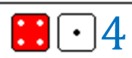
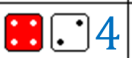




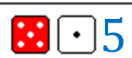
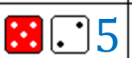





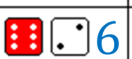




 1	 2	 3	 4	 5	 6
 2	 2	 3	 4	 5	 6
 3	 3	 3	 4	 5	 6
 4	 4	 4	 4	 5	 6
 5	 5	 5	 5	 5	 6
 6	 6	 6	 6	 6	 6

Figure 4.1.2 Maximum value of two fair dice rolled

From the figure above, we can tell that our random variable Y has 6 possible values: $\{1, 2, 3, 4, 5, 6\}$. With 6 occurring most frequently. Since we are rolling fair dice, each outcome is equally likely, so we produce our probability distribution by counting the number of occurrences of each value. We make the probability in the table below.

Table 4.1.2 Probability distribution of the random variable Y

$Y = y_j$	$P(Y = y_j)$
1	$\frac{1}{36} \approx 2.7778\%$
2	$\frac{3}{36} = \frac{1}{12} \approx 8.3333\%$
3	$\frac{5}{36} \approx 13.8889\%$
4	$\frac{7}{36} \approx 19.4444\%$
5	$\frac{9}{36} = \frac{1}{4} = 25\%$
6	$\frac{11}{36} \approx 30.5556\%$

Let us reason our way to the probability distribution without listing all the outcomes. Knowing that the values on our dice range from 1 to 6, we can restrict our considerations for the possible values of Y to these; that is, Y has 6 possible values: $\{1, 2, 3, 4, 5, 6\}$.

To determine $P(Y = 6)$, we note that 6 is the largest value on our dice. So if a 6 is rolled, it is the maximum. So $P(Y = 6) = P(6 \text{ IS ROLLED FIRST OR SECOND}) = \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$.

To determine $P(Y = 5)$, we note that the only number larger than 5 is 6. So $P(Y = 5) = P(5 \text{ IS ROLLED and } 6 \text{ IS NOT ROLLED}) = \frac{11}{36} \cdot \frac{9}{11} = \frac{9}{36} = \frac{1}{4}$. Notice that $P(6 \text{ IS NOT ROLLED} | 5 \text{ IS ROLLED}) = \frac{9}{11}$ because there are 11 ways to roll a 5 and 9 of them do not contain a 6.

A similar process could continue through all of the possible values, but we might notice an easier way to count; for any particular value y_j of Y , we have the outcome of rolling double y_j s, and the remaining outcomes come in pairs. The number of pairs equals the number of values less than y_j . We develop a formula for our probabilities: $P(Y = y_j) = \frac{2 \cdot (y_j - 1) + 1}{36}$. Check that our reasoning produces the same probability distribution as above.

Note: Probabilities Across all Possible Values

Recall that the sum of the probabilities of all outcomes from a sample space is 1. This is true because when conducting a random experiment, something must happen; a single outcome from the sample space must occur, and no two outcomes in the sample space are the same. If we have n outcomes, $1 = P(\text{OUTCOME}_1 \text{ or } \text{OUTCOME}_2 \text{ or } \dots \text{ or } \text{OUTCOME}_n) = P(\text{OUTCOME}_1) + P(\text{OUTCOME}_2) + \dots + P(\text{OUTCOME}_n)$.

A similar line of reasoning follows for random variables. Since a random variable assigns a number to every outcome, and an outcome must occur when a random experiment is conducted, we are sure that some value will occur and no outcome will return two values. If we have n values for a random variable X , $1 = P(X = x_1 \text{ or } X = x_2 \text{ or } \dots \text{ or } X = x_n) = P(X = x_1) + P(X = x_2) + \dots + P(X = x_n)$. The sum of probabilities across all possible values of a random variable must always equal 1. This means the sum of all the values in the $P(X = x)$ column of a probability distribution must add up to 1.

Let us confirm the sum of the probability column of a probability distribution is 1 for the random variables that we have discussed thus far, X and Y .

$$X: \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{36}{36} = 1$$

$$Y: \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{9}{36} + \frac{11}{36} = \frac{36}{36} = 1$$

? Text Exercise 4.1.2

Consider the random variable D , loosely based on this [article](#) from 2009, which returns the number of days adults exercise in a week and its incomplete probability distribution below.

Table 4.1.3: Incomplete probability distribution for the random variable D

$D = d_j$	$P(D = d_j)$
0	0.28
1	0.11
2	
3	0.14
4	0.10
5	0.15
6	0.08
7	0.04

1. Complete the probability distribution by determining $P(D = 2)$.

Answer

The sum of all the probabilities in a probability distribution must equal 1. We can compute $P(D = 2)$ by figuring out what value makes the sum 1. $0.28 + 0.11 + 0.14 + 0.10 + 0.15 + 0.08 + 0.04 = .90$. So $P(D = 2) = 1 - 0.90 = 0.10$. Ten percent of adults exercise for just 2 days a week.

2. What is the probability that a randomly selected adult exercises at least 5 days a week?

Answer

We are trying to determine the probability that a randomly selected adult exercises 5, 6, or 7 days a week. which we can denote as $P(D \geq 5) = P(D = 5 \text{ or } D = 6 \text{ or } D = 7)$. Since each event is assigned a single value, they are mutually exclusive; thus, we can simply add the probabilities of each event. $P(D \geq 5) = P(D = 5 \text{ or } D = 6 \text{ or } D = 7) = P(D = 5) + P(D = 6) + P(D = 7) = 0.15 + .08 + .04 = 0.27$. We understand this to mean that 27% of adults work out at least 5 days a week.

3. Determine and explain the meaning of $P(2 < D \leq 4)$.

Answer

We are trying to determine the probability that a randomly selected adult exercises more than 2 times a week but no more than 4 times a week or equivalently that a randomly selected adult exercises 3 or 4 days a week. $P(2 < D \leq 4) = P(D = 3) + P(D = 4) = 0.14 + 0.10 = 0.24$. So 24% of adults exercise 3 or 4 days a week.

Now consider the random experiment of rolling two fair dice from a slightly different perspective and arrive at another type of random variable. We could define a random variable Z to be "the time (in seconds) it takes both dice to come to a complete stop after one die leaves our hands." We understand our random variable by examining the possible values that our random variable takes on. Determining the precise values is difficult. What is the shortest time? Does it always take at least 1 second? What is the longest time? Can it ever exceed 5 seconds? We cannot give definitive answers. However, after a moment or two of thought, we recognize that the possible values will take on any numerical value in an interval of positive real numbers. Hopefully, this last description reminds us of a type of variable. In [Chapter 1](#), we defined two types of quantitative variables: discrete and continuous. Here, we make similar designations: **discrete random variable** and **continuous random variable**, based on the possible outcomes. Examples X , Y , and D from above are discrete random variables while Z is a continuous random variable. Our

understanding of probability needs further development to handle continuous random variables. This will take place in the latter portion of this chapter; for now, we restrict ourselves to the study of discrete random variables.

? Text Exercise 4.1.3

Classify each random variable as either discrete or continuous. Explain.

1. P = the prescription count of a randomly chosen patient.

Answer

We understand a patient's prescription count to be the number of medicines prescribed. While a patient may be prescribed a half or double dosage, this is not half of a prescription. There are gaps between each possible value that P takes on, making P a discrete random variable.

2. V = the appraisal value of a randomly chosen coin collection.

Answer

An appraisal value must be given in some currency, perhaps U.S. dollars. Currencies have a smallest denomination. Therefore, there must be gaps between the possible values in the appraisal value, making V a discrete random variable.

3. T = the total distance traveled on the campaign trail of a randomly chosen politician.

Answer

A politician's total distance on the campaign trail (in any standard unit) may be any nonnegative number within a reasonable magnitude. T is a continuous random variable.

Discrete Uniform Distribution

As the name indicates, the probability distribution of a random variable explains how probabilities are distributed. We say a random variable has a **discrete uniform distribution** if the random variable is discrete and each outcome has equal probability. If we consider rolling a single fair die and define a random variable S to be the number that lands face up, the random variable S has a discrete uniform distribution. There are only six possible values making S discrete, and since the die is fair, each value is equally probable. $P(S = s) = \frac{1}{6}$ for any s in $\{1, 2, 3, 4, 5, 6\}$.

? Text Exercise 4.1.4

1. Consider the discrete random variable R , with 10 values, that has a discrete uniform distribution, and determine the probability of each value of R .

Answer

Since R has a discrete uniform distribution and takes on 10 values we have that $P(R = r_1) = P(R = r_2) = \dots = P(R = r_{10})$ and $\sum_{j=1}^{10} P(R = r_j) = 1$. Combining these yields that $1 = \sum_{j=1}^{10} P(R = r_j) = \sum_{j=1}^{10} P(R = r_1) = 10 \cdot P(R = r_1)$ meaning $P(R = r_1) = \frac{1}{10}$ and thus $P(R = r_j) = \frac{1}{10}$ for any j in $\{1, 2, 3, \dots, 10\}$.

2. Consider the discrete random variable R which takes on k values and has a discrete uniform distribution, determine the probability of each value that R takes on.

Answer

Since R has a discrete uniform distribution and takes on k values we have that $P(R = r_1) = P(R = r_2) = \dots = P(R = r_k)$ and $\sum_{j=1}^k P(R = r_j) = 1$. Combining these yields that $1 = \sum_{j=1}^k P(R = r_j) = \sum_{j=1}^k P(R = r_1) = k \cdot P(R = r_1)$ meaning $P(R = r_1) = \frac{1}{k}$ and thus $P(R = r_j) = \frac{1}{k}$ for any j in $\{1, 2, 3, \dots, k\}$.

3. Consider the random variables X , Y , and D which have been recurring this section. For each of them, determine, with justification, if they are a uniform random variable or not.

Answer

First consider X . The probability that X is 7 is not the same as the probability that X is 12. Therefore, X is not a uniformly distributed random variable. We should be able to convince ourselves, using similar reasoning, that Y and D are also both not uniformly distributed.

4.1: [Random Variables](#) is shared under a [Public Domain](#) license and was authored, remixed, and/or curated by LibreTexts.

- [3.3: Measures of Central Tendency](#) by [David Lane](#) is licensed [Public Domain](#). Original source: <https://onlinestatbook.com>.