CHAPTER 6 Relationships between Categorical Variables

Chapter Outline

6.1 CONTINGENCY TABLES
6.2 BASIC RULES OF PROBABILITY WE NEED TO KNOW
6.3 CONDITIONAL PROBABILITY
6.4 EXAMINING INDEPENDENCE OF CATEGORICAL VARIABLES

6.1 Contingency Tables

Learning Objective

- How to create and use contingency tables.
- How to find the marginal distribution of a variable.

Contingency Tables

Suppose you wanted to evaluate how gender affects the type of movie chosen by movie-goers. How might you organize data on male and female watchers, and action, romance, comedy, and horror movie types, so it would be easy to compare the different combinations?

Contingency tables are used to evaluate the interaction of two different categorical variables. Contingency tables are sometimes called **two-way tables** because they are organized with the outputs of one variable across the top, and another down the side. Consider the table below:

TABLE 6.1:

	Male	**Female**
Chocolate Candy	42	77
Fruit Candy	58	23

This is a contingency table comparing the variable 'Gender' with the variable 'Candy Preference'. You can see that, across the top of the table are the two gender options for this particular study: 'male students' and 'female students'. Down the left side are the two candy preference options: 'chocolate' and 'fruit'. The data in the center of the table indicates the reported candy preferences of the 200 students polled during the study.

Commonly, there will be one additional row and column for totals, like this:

TABLE 6.2:

	Male	**Female**	**TOTAL**
Chocolate Candy	42	77	**119**
Fruit Candy	58	23	**81**
TOTAL	**100**	**100**	**200**

Notice that you can run a quick check on the calculation of totals, since the "total of totals" should be the same from either direction: $119+81=200=100+100$.

A **marginal distribution** is how many overall responses there were for each category of the variable. The marginal distribution of a variable can be determined by looking at the "Total" column (for type of candy) or the "Total" row (for gender). For example, we can see that the marginal totals for type of candy are 119 chocolate and 81 fruit. Similarly, the marginal total for gender tells us there were an equal number of males and females in the study.

Example A

Construct a contingency table to display the following data:

250 mall shoppers were asked if they intended to eat at the in-mall food court or go elsewhere for lunch. Of the 117 male shoppers, 68 intended to stay, compared to only 62 of the 133 female shoppers

Solution

First, let's identify our variables and set up the table with the appropriate row and column headers.

The variables are gender and lunch location choice:

TABLE 6.3:

	Male	***Female***	***TOTAL***
Food Court			
Out of Mall			
TOTAL			

Now we can fill in the values we have directly from the text:

TABLE 6.4:

	Male	***Female***	***TOTAL***
Food Court	68	62	
Out of Mall			
TOTAL	117	133	250

Now we can fill in the missing data with simple addition/subtraction:

TABLE 6.5:

	Male	***Female***	***TOTAL***
Food Court	68	62	130
Out of Mall	49	71	120
TOTAL	117	133	**250**

Example B

Referencing data from Example A, answer the following:

a. What percentage of food-court eaters are female?
b. What is the distribution of male lunch-eaters?
c. What is the *marginal distribution* of the variable "lunch location preference"?
d. What is the marginal distribution of the variable "Gender"?
e. What percentage of females prefer to eat out?

Solution

a. If we read across the row "Food Court", we see that there were a total of 130 shoppers eating "in", and that 62 of them were female. To calculate percentage, we simply divide: $\frac{62}{130} \approx .477$ *or* **47.7%.**
b. The male shoppers were distributed as **68 food court and 49 out of mall.**
c. The *marginal distribution* is the distribution of data "in the margin", or in the TOTAL column. In this case, we are interested in the data on lunch location preference, which is found in the far right column: **130 food**

court and **120 out of mall.**

d. The marginal distribution of gender can be found in the bottom row: **117 males** and **133 females.**
e. Here we are interested in data from the females, so we will be dealing with the 'female' column. From the data in the column, we see that 71 of the 133 females preferred to eat out. This is a percentage of: $\frac{71}{133} \approx .534$ ***or* 53.4%.**

Example C

Using the given data:

a. Construct a contingency table
b. Identify the marginal distributions
c. Identify 3 different percentage-based observations

Out of 213 polled amateur drag racers, 37 drove cars with turbo-chargers, 59 had superchargers, and the rest were normally aspirated. The racers themselves were split between 102 rookies and 111 veterans. The rookies evidently preferred turbos, since 29 of them had turbo-charged vehicles, and avoided superchargers, since there were only 12 of them.

Solution

a. Set up the table with the appropriate headers, and fill in the data we know. Note that this time we will need a 3×2 table instead of a 2×2 (it is still a two- *way* table though, as there are only two variables: engine aspiration and driver experience):

TABLE 6.6:

	Turbocharger	Supercharger	Normal Aspiration	TOTAL
Rookie	29	12		102
Veteran				111
TOTAL	37	59	117	213

Now we can update the table with the missing data, calculated using addition or subtraction:

TABLE 6.7:

	Turbocharger	Supercharger	Normal Aspiration	TOTAL
Rookie	29	12	61	102

TABLE 6.7: (continued)

Veteran	8	47	56	111
TOTAL	37	59	117	213

b. The marginal distribution refers to the overall data for each of the two variables:

- Aspiration type is distributed as follows: **37 Turbos, 59 Superchargers, and 117 normally aspirated.**
- Driver experience distribution: **102 Rookies and 111 Veterans.**

c. Three percentage-based observations:

- $\frac{61}{102} = 0.598$ *or* 59.8% of Rookies drive normally aspirated cars.
- $\frac{47}{59} = 0.7966$ *or* 79.66% of the Superchargers were in cars driven by Veterans.
- $\frac{47}{111} = 0.4234$ *or* 42.34% of Veterans use Superchargers.

Vocabulary

A **contingency table** or **two-way** table is used to organize data from multiple categories of two variables so that various assessments may be made.

A **marginal distribution** is the distribution of data "in the margin" of a table. It may also be described as the distribution of the data for a single variable.

Guided Practice

1. Complete the data in the contingency table:

TABLE 6.8:

	A	**B**	**TOTAL**
X	47		
Y		32	100
TOTAL	105		200

2. What is the marginal distribution of the variable consisting of categories A and B?
3. What percentage of B's are Y's?
4. What portion of A's are X's? Express your answer as a decimal.

Solution

TABLE 6.9:

	A	**B**	**TOTAL**
X	47	100 - 47 = 53	200 - 100 = 100
Y	100 - 32 = 68	32	100
TOTAL	105	200 - 105 = 85	200

2. The marginal distribution is distributed as Category A: 105 and Category B: 85.

3. There are 32 B's that are also Y's, out of the total of 100 B's: $\frac{32}{100} = 32\%$
4. 47 of the 100 A's are X's, $\frac{47}{100} = 0.47$

More Practice

Questions 1-9 refer to the following table:

TABLE 6.10:

	Sports Cars	Pickup Trucks	Luxury Cars	TOTAL
Male Drivers	72	67	36	175
Female Drivers	36	71	68	175
TOTAL	108	138	104	350

1. What is the marginal distribution of vehicle types?
2. What is the marginal distribution of driver gender?
3. What decimal portion of male drivers have luxury cars?
4. What percentage of female drivers have pickups?
5. How many drivers were polled?
6. What is the overall most popular vehicle type, by percentage?
7. Which vehicle type has the single largest cell value, and what percentage does it represent of that gender category?
8. What percentage of pickup trucks are driven by females?
9. What percentage of females drive pickup trucks?

Questions 10-18 refer to the following data:

One hundred eighty dogs were studied to determine if breed affected food preference. Of the 70 Huskies, 30 preferred beef flavor and 40 preferred chicken. Of the 50 Poodles, 27 preferred beef, the rest chicken. The rest of the dogs, English Mastiffs, were obviously beef-lovers, as only 19 preferred chicken over beef.

10. Create a contingency table to display the data.
11. What is the marginal distribution of dog breeds?
12. What is the marginal distribution of food types?
13. What percentage of Mastiffs preferred beef?
14. What percentage of beef-lovers were Mastiffs?
15. What flavor/dog combination indicated the strongest preference? What percentage of the breed did it represent?
16. What is the distribution of chicken preference?
17. What is the distribution of beef preference?
18. Which breed shows the least defined preference, as a percentage?

6.2 Basic Rules of Probability We Need to Know

Learning Objectives

- Know basic statistical terminology.
- List simple events and sample spaces.
- Know the basic rules of probability.

Introduction

The concept of probability plays an important role in our daily lives. Assume you have an opportunity to invest some money in a software company. Suppose you know that the company's records indicate that in the past five years, its profits have been consistently decreasing. Would you still invest your money in it? Do you think the chances are good for the company in the future?

Here is another illustration. Suppose that you are playing a game that involves tossing a single die. Assume that you have already tossed it 10 times, and every time the outcome was the same, a 2. What is your prediction of the eleventh toss? Would you be willing to bet \$100 that you will not get a 2 on the next toss? Do you think the die is loaded?

Notice that the decision concerning a successful investment in the software company and the decision of whether or not to bet \$100 on the next outcome of the die are both based on probabilities of certain sample results. Namely, the software company's profits have been declining for the past five years, and the outcome of rolling a 2 ten times in a row seems strange. From these sample results, we might conclude that we are not going to invest our money in the software company or bet on this die. In this lesson, you will learn mathematical ideas and tools that can help you understand such situations.

Events, Sample Spaces, and Probability

An **event** is something that occurs, or happens. For example, flipping a coin is an event, and so is walking in the park and passing by a bench. Anything that could possibly happen is an event.

Every event has one or more possible outcomes. While tossing a coin is an event, getting tails is the outcome of that event. Likewise, while walking in the park is an event, finding your friend sitting on the bench is an outcome of that event.

Suppose a coin is tossed once. There are two possible outcomes, either heads, H, or tails, T. Notice that if the experiment is conducted only once, you will observe only one of the two possible outcomes. An **experiment** is the process of taking a measurement or making an observation. These individual outcomes for an experiment are each called **simple events**.

Example A

A die has six possible outcomes: 1, 2, 3, 4, 5, or 6. When we toss it once, only one of the six outcomes of this experiment will occur. The one that does occur is called a simple event.

Example B

Suppose that two pennies are tossed simultaneously. We could have both pennies land heads up (which we write as HH), or the first penny could land heads up and the second one tails up (which we write as HT), etc. We will see that there are four possible outcomes for each toss, which are HH, HT, TH, and TT.

What we have accomplished so far is a listing of all the possible events of an experiment. This collection is called the **sample space** of the experiment. The sample space is the set of all possible outcomes of an experiment, or the collection of all the possible simple events of an experiment. We will denote a sample space by S.

Example C

We want to determine the sample space of throwing a die and the sample space of tossing a coin.

Solution

As we know, there are 6 possible outcomes for throwing a die. We may get 1, 2, 3, 4, 5, or 6, so we write the sample space as the set of all possible outcomes:

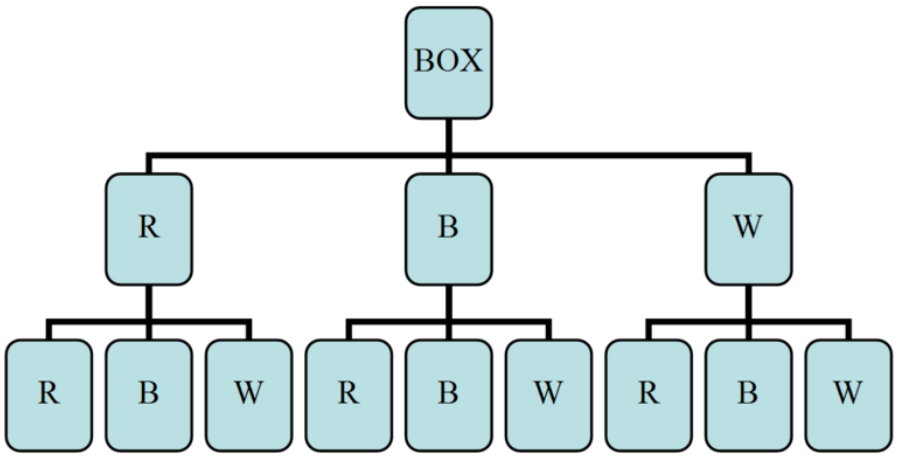
$$S = \{1,2,3,4,5,6\}$$

Similarly, the sample space of tossing a coin is either heads, H, or tails, T, so we write $S = \{H,T\}$.

Example D

Suppose a box contains three balls, one red, one blue, and one white. One ball is selected, its color is observed, and then the ball is placed back in the box. The balls are scrambled, and again, a ball is selected and its color is observed. What is the sample space of the experiment?

It is probably best if we draw a **tree diagram** to illustrate all the possible selections.



As you can see from the tree diagram, it is possible that you will get the red ball, R, on the first drawing and then another red one on the second, RR. You can also get a red one on the first and a blue on the second, and so on. From the tree diagram above, we can see that the sample space is as follows:

$$S = \{RR, RB, RW, BR, BB, BW, WR, WB, WW\}$$

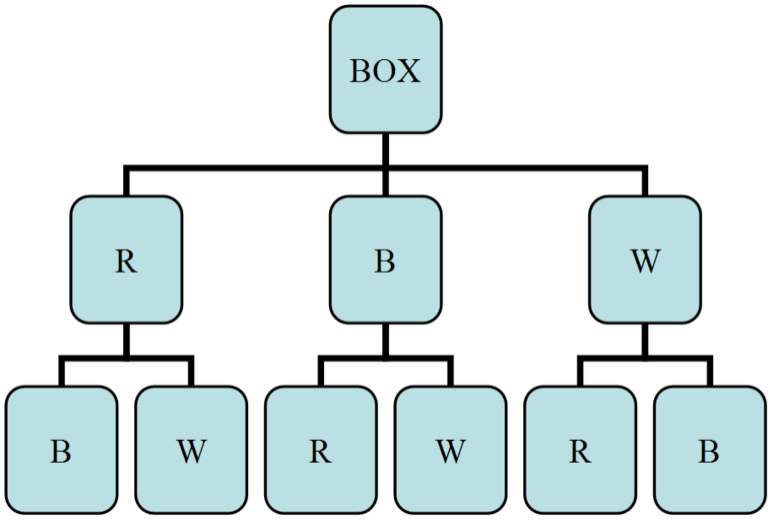
Each pair in the set above gives the first and second drawings, respectively. That is, RW is different from WR.

Example E

Consider the same experiment as in the last example. This time we will draw one ball and record its color, but we will not place it back into the box. We will then select another ball from the box and record its color. What is the sample space in this case?

Solution

The tree diagram below illustrates this case:



You can clearly see that when we draw, say, a red ball, the blue and white balls will remain. So on the second selection, we will either get a blue or a while ball. The sample space in this case is as shown:

$$S = \{RB, RW, BR, BW, WR, WB\}$$

Now let us return to the concept of probability and relate it to the concepts of sample space and simple events. If you toss a fair coin, the chance of getting tails, T, is the same as the chance of getting heads, H. Thus, we say that the probability of observing heads is 0.5, and the probability of observing tails is also 0.5. The probability, P, of an outcome, A, always falls somewhere between two extremes: 0, which means the outcome is an impossible event, and 1, which means the outcome is guaranteed to happen. Most outcomes have probabilities somewhere in-between.

Property 1: $0 \leq P(A) \leq 1$, for any event, A.

The probability of an event, A, ranges from 0 (impossible) to 1 (certain).

In addition, the probabilities of all possible simple outcomes of an event must add up to 1. This 1 represents certainty that one of the outcomes must happen. For example, tossing a coin will produce either heads or tails. Each of these two outcomes has a probability of 0.5. This means that the total probability of the coin landing either heads or tails is $0.5 + 0.5 = 1$. That is, we know that if we toss a coin, we are certain to get heads or tails.

Property 2: $\sum P(A) = 1$ when summed over all possible simple outcomes.

The sum of the probabilities of all possible outcomes must add up to 1.

Notice that tossing a coin or throwing a die results in outcomes that are all equally probable. That is, each outcome has the same probability as all the other outcomes in the same sample space. Getting heads or tails when tossing a coin produces an equal probability for each outcome, 0.5. Throwing a die has 6 possible outcomes, each also having the same probability, $\frac{1}{6}$. We refer to this kind of probability as classical probability. **Classical probability** is defined to be the ratio of the number of cases favorable to an event to the number of all outcomes possible, where each of the outcomes is equally likely.

Probability is usually denoted by P, and the respective elements of the sample space (the outcomes) are denoted by A, B, C, etc. The mathematical notation that indicates the probability that an outcome, A, happens is $P(A)$. We use the following formula to calculate the probability of an outcome occurring:

$$P(A) = \frac{\text{The number of outcomes for } A \text{ to occur}}{\text{The size of the sample space}}$$

Example F

When tossing two coins, what is the probability of getting a head on both coins, HH? Is the probability classical?

Since there are 4 elements (outcomes) in the sample space set, $\{HH, HT, TH, TT\}$, its size is 4. Furthermore, there is only 1 HH outcome that can occur. Therefore, using the formula above, we can calculate the probability as shown:

$$P(A) = \frac{\text{The number of outcomes for } HH \text{ to occur}}{\text{The size of the sample space}} = \frac{1}{4} = 25\%$$

Notice that each of the 4 possible outcomes is equally likely. The probability of each is 0.25. Also notice that the total probability of all possible outcomes in the sample space is 1.

Example G

What is the probability of throwing a die and getting $A = 2, 3,$ or 4?

There are 6 possible outcomes when you toss a die. Thus, the total number of outcomes in the sample space is 6. The event we are interested in is getting a 2, 3, or 4, and there are three ways for this event to occur.

$$P(A) = \frac{\text{The number of outcomes for 2, 3, or 4 to occur}}{\text{The size of the sample space}} = \frac{3}{6} = \frac{1}{2} = 50\%$$

Therefore, there is a probability of 0.5 that we will get 2, 3, or 4.

Example H

Consider tossing two coins. Assume the coins are not balanced. The design of the coins is such that they produce the probabilities shown in the table below:

TABLE 6.11:

Outcome	Probability
HH	$\frac{4}{9}$
HT	$\frac{2}{9}$
TH	$\frac{2}{9}$
TT	$\frac{1}{9}$

What is the probability of observing exactly one head, and what is the probability of observing at least one head?

Notice that the simple events HT and TH each contain only one head. Thus, we can easily calculate the probability of observing exactly one head by simply adding the probabilities of the two simple events:

$$\begin{aligned} P &= P(HT) + P(TH) \\ &= \frac{2}{9} + \frac{2}{9} \\ &= \frac{4}{9} \end{aligned}$$

Similarly, the probability of observing at least one head is:

$$\begin{aligned} P &= P(HH) + P(HT) + P(TH) \\ &= \frac{4}{9} + \frac{2}{9} + \frac{2}{9} = \frac{8}{9} \end{aligned}$$

Lesson Summary

An event is something that occurs, or happens, with one or more possible outcomes.

An experiment is the process of taking a measurement or making an observation.

A simple event is the simplest outcome of an experiment.

The sample space is the set of all possible outcomes of an experiment, typically denoted by S.

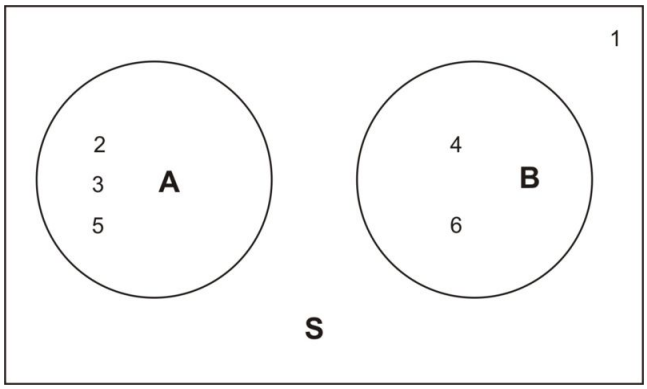
Review Questions

1. Consider an experiment composed of throwing a die followed by throwing a coin.
   a. List the simple events and assign a probability for each simple event.
   b. What are the probabilities of observing the following events?

(a)
   i. A 2 on the die and H on the coin
   ii. An even number on the die and T on the coin
   iii. An even number on the die
   iv. T on the coin

2. The Venn diagram below shows an experiment with six simple events. Events A and B are also shown. The probabilities of the simple events are:

$$P(1) = P(2) = P(4) = \frac{2}{9}$$
$$P(3) = P(5) = P(6) = \frac{1}{9}$$



1. a. Find $P(A)$
   b. Find $P(B)$
2. A box contains two blue marbles and three red ones. Two marbles are drawn randomly without replacement. Refer to the blue marbles as $B1$ and $B2$ and the red ones as $R1$, $R2$, and $R3$.
   a. List the outcomes in the sample space.
   b. Determine the probability of observing each of the following events:

(a) i. Drawing 2 blue marbles
ii. Drawing 1 red marble and 1 blue marble
iii. Drawing 2 red marbles

6.3 Conditional Probability

Learning Objective

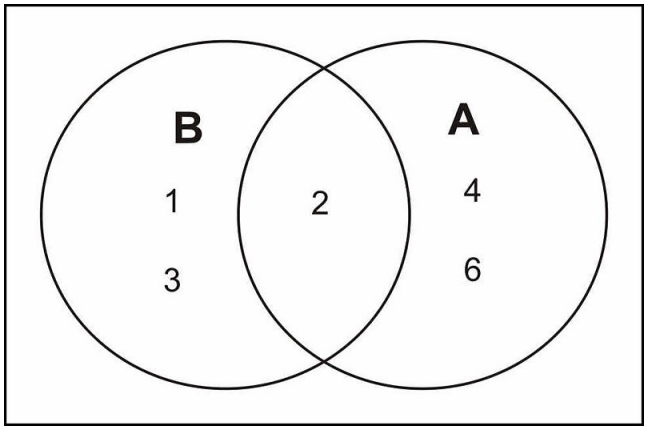
- Calculate the conditional probability that event A occurs, given that event B has occurred.
- Understand the difference between independent and dependent events.

Introduction

In this lesson, you will learn about the concept of conditional probability and be presented with some examples of how conditional probability is used in the real world. Once we understand the basics of conditional probability in this chapter, we can begin to think about how these concepts can be used to determine whether two categorical variables are related or, in the language of probability, whether they are independent.

Notation

We know that the probability of observing an even number on a throw of a die is 0.5. Let the event of observing an even number be event A. Now suppose that we throw the die, and we know that the result is a number that is 3 or less. Call this event B. Would the probability of observing an even number on that particular throw still be 0.5? The answer is no, because with the introduction of event B, we have reduced our sample space from 6 simple events to 3 simple events. In other words, since we have a number that is 3 or less, we now know that we have a 1, 2 or 3. This becomes, in effect, our sample space. Now the probability of observing a 2 is $\frac{1}{3}$. With the introduction of a particular condition (event B), we have changed the probability of a particular outcome. The Venn diagram below shows the reduced sample space for this experiment, given that event B has occurred:



The only even number in the sample space for B is the number 2. We conclude that the probability that A occurs, given that B has occurred, is 1:3, or $\frac{1}{3}$. We write this with the notation $P(A|B)$, which reads "the probability of A, given B." So for the die toss experiment, we would write $P(A|B) = \frac{1}{3}$.

Conditional Probability of Two Events

If A and B are two events, then the probability of event A occurring, given that event B has occurred, is called **conditional probability**. We write it with the notation $P(A|B)$, which reads "the probability of A, given B."

To calculate the conditional probability that event A occurs, given that event B has occurred, take the ratio of the probability that both A and B occur to the probability that B occurs. That is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

For our example above, the die toss experiment, we proceed as is shown below:

A : observe an even number

B : observe a number less than or equal to 3

To find the conditional probability, we use the formula as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(2)}{P(1)+P(2)+P(3)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Example

A medical research center is conducting experiments to examine the relationship between cigarette smoking and cancer in a particular city in the USA. Let A represent an individual who smokes, and let C represent an individual who develops cancer. This means that AC represents an individual who smokes and develops cancer, AC' represents an individual who smokes but does not develop cancer, and so on. We have four different possibilities, or simple events, and they are shown in the table below, along with their associated probabilities.

TABLE 6.12:

Simple Events	Probabilities
AC	0.10
AC'	0.30
$A'C$	0.05
$A'C'$	0.55

These simple events can be studied, along with their associated probabilities, to examine the relationship between smoking and cancer.

We have:

A : individual smokes

C : individual develops cancer

A' : individual does not smoke

C' : individual does not develop cancer

A very powerful way of examining the relationship between cigarette smoking and cancer is to compare the conditional probability that an individual gets cancer, given that he/she smokes, with the conditional probability that an individual gets cancer, given that he/she does not smoke. In other words, we want to compare $P(C|A)$ with $P(C|A')$.

Recall that $P(C|A) = \frac{P(C \cap A)}{P(A)}$.

Before we can use this relationship, we need to calculate the value of the denominator. $P(A)$ is the probability of an individual being a smoker in the city under consideration. To calculate it, remember that the probability of an event is the sum of the probabilities of all its simple events. A person can smoke and have cancer, or a person can smoke and not have cancer. That is:

$$P(A) = P(AC) + P(AC') = 0.10 + 0.30 = 0.4$$

This tells us that according to this study, the probability of finding a smoker selected at random from the sample space (the city) is 40%. We can continue on with our calculations as follows:

$$P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{P(AC)}{P(A)} = \frac{0.10}{0.40} = 0.25 = 25\%$$

Similarly, we can calculate the conditional probability of a nonsmoker developing cancer:

$$P(C|A') = \frac{P(A' \cap C)}{P(A')} = \frac{P(A'C)}{P(A')} = \frac{0.05}{0.60} = 0.08 = 8\%$$

In this calculation, $P(A') = P(A'C) + P(A'C') = 0.05 + 0.55 = 0.60$. $P(A')$ can also be found by using the Complement Rule as shown: $P(A') = 1 - P(A) = 1 - 0.40 = 0.60$.

From these calculations, we can clearly see that a relationship exists between smoking and cancer. The probability that a smoker develops cancer is 25%, and the probability that a nonsmoker develops cancer is only 8%. Keep in mind, though, that it would not be accurate to say that smoking causes cancer. However, our findings do suggest a strong link between smoking and cancer.

Independence

Suppose you are flipping a coin and at the same time rolling a dice. Obviously, the probability of rolling a 3 has nothing to do with whether the coin lands heads or tails. Such events are known as **independent**.

Event B is said to be independent of event A if P(BlA) = P(B). Alternatively, P(AlB) = P(A)

If the above is not true, then the events are said to be **dependent**. There are other less obvious examples that we frequently encounter. Suppose your math teacher was recently at an event featuring door prizes. The prizes varied in value from water bottles to a kayak. Suppose there are 200 names in the drawing. If you are like you math teacher and never win anything, would you prefer them to start the drawing with the kayak or the water bottles? How does the probability of getting you name drawn change as they draw more names?

Lesson Summary

If $A$and $B$are two events, then the probability of event Aoccurring, given that event Bhas occurred, is called conditional probability. We write it with the notation $P(A|B)$, which reads "the probability of A, given B."

Conditional probability can be found with the equation $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Vocabulary

Conditional Probability: Probability that is predicted based on specific criteria or conditions.

Independent Events: The outcome of one event has no effect on the outcome of a second event.

Dependent Events: If the outcome of one event has an effect on the outcome of another event they are dependent events.

Review Questions

1. If $P(A) = 0.3, P(B) = 0.7$, and $P(A \cap B) = 0.15$, Find $P(A|B)$ and $P(B|A)$.
2. Two fair coins are tossed. i. List the possible outcomes in the sample space. ii. Two events are defined as follows:

   $$\text{A: } \{\text{At least one head appears}\}$$
   $$\text{B: } \{\text{Only one head appears}\}$$

   Find $P(A), P(B), P(A \cap B), P(A|B)$, and $P(B|A)$

3. A box of six marbles contains two white, two red, and two blue. Two marbles are randomly selected without replacement and their colors are recorded. i. List the possible outcomes in the sample space. ii. Let the following events be defined:

   $$\text{A: } \{\text{Both marbles have the same color}\}$$
   $$\text{B: } \{\text{Both marbles are red}\}$$
   $$\text{C: } \{\text{At least one marble is red or white}\}$$

   Find $P(B|A), P(B|A'), P(B|C), P(A|C)$, and $P(C|A')$

Review Answers

1. $0.21, 0.5$
2. $3/4, 1/2, 1/2, 1, 2/3$
3. $1/3, 0, 1/14, 1/7, 1$

6.4 Examining Independence of Categorical Variables

Learning Objectives

- Apply what you have learned about conditional probabilities to determine if two categorical variables influence each other or not.

Bivariate Relationships for Categorical Data

Suppose you conducted a survey where you asked each person two questions: "Do you have Cable TV?" and "Did you go on vacation in the past year?." You now have data on two **categorical** variables for each person. As we have seen, whenever you have two pieces of data from each person, you can organize the data into a **two-way frequency table,** or **contingency table**.

Here is data collected from a group of individuals who answered those two questions in a contingency table:

TABLE 6.13:

	Took a Vacation	**No Vacation**	***Total***
Have Cable TV	97	38	*135*
Don't Have Cable TV	14	17	*31*
Total	*111*	*55*	*166*

The numbers in the frequency table show the numbers of people that fit each pair of preferences. For example, 97 people have cable TV and took a vacation last year. 38 people have cable TV but did not take a vacation last year. The totals of the rows and columns have been added to the frequency table for convenience. From the far right column you can see that 135 people have cable TV and 31 people don't have cable TV for a total of 166 people surveyed. From the bottom row you can see that 111 people took a vacation and 55 people did not take a vacation for a total of 166 people surveyed.

You can use the two-way frequency table to calculate probabilities about the people surveyed. For example, you could find:

a. The probability that a random person selected from this group took a vacation last year.
b. The probability that a random person from this group who has cable TV took a vacation last year.
c. Whether or not "choosing a person with cable TV" and "choosing a person who took a vacation" are independent events for this population of 166 people.

Example A

Suppose you choose a person at random from the group surveyed below. Let A be the event that the person chosen took a vacation last year. Find $P(A)$.

TABLE 6.14:

	Took a Vacation	**No Vacation**	*Total*
Have Cable TV	97	38	*135*

TABLE 6.14: (continued)

Don't Have Cable TV	14	17	*31*
Total	*111*	*55*	*166*

Solution

There were 166 people surveyed, so there are 166 outcomes in the sample space. 111 people took a vacation last year.

$$P(A) = \frac{111}{166} \approx 0.67 \text{ or } 67\%$$

Example B

Suppose you choose a person at random from the group surveyed below. Let A be the event that the person chosen took a vacation last year. Let B be the event that the person chosen has cable TV. Find $P(A|B)$.

TABLE 6.15:

	Took a Vacation	**No Vacation**	*Total*
Have Cable TV	97	38	*135*
Don't Have Cable TV	14	17	*31*
Total	*111*	*55*	*166*

Solution

You are looking for the probability that the person took a vacation *given* that they have cable TV. Since you know that the person has cable TV, the sample space has been restricted to the 135 people with cable TV. 97 of those people took a vacation.

$$P(A|B) = \frac{97}{135} \approx .72 \text{ or } 72\%$$

Suppose you wanted to use the conditional probability formula for this calculation.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{97}{166}}{\frac{135}{166}} = \frac{97}{135} \approx 0.72 \text{ or } 72\%$$

With the conditional probability formula, each probability is calculated with the sample space of 166. The two 166's cancel each other out, and the result is the same. Sometimes it makes sense to use the conditional probability formula, and sometimes it is easier to think logically about what is being asked.

Example C

Suppose you choose a person at random from the group surveyed below. Let A be the event that the person chosen took a vacation last year. Let B be the event that the person chosen has cable TV. Are events A and B independent?

TABLE 6.16:

	Took a Vacation	**No Vacation**	*Total*
Have Cable TV	97	38	*135*
Don't Have Cable TV	14	17	*31*
Total	*111*	*55*	*166*

Solution

Let's remind ourselves what it means for two events to be independent. As we learned in the last chapter, two events are independent if the following statement is true:

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

From Example A, you know that $P(A) = 67\%$. From Example B, you know that $P(A|B) = 72\%$ because these probabilities are not equal, the two events are NOT independent (they are dependent). People with cable TV are more likely to have taken a vacation as opposed to people without cable TV, so knowing that a person has cable TV increases the probability that they took a vacation.

Vocabulary

The **probability** of an event is the chance of the event occurring.

Two events are **independent** if one event occurring does not change the probability of the second event occurring. $P(A \cap B) = P(A)P(B)$ **if and only if** A **and** B **are independent events. Also,** $P(A|B) = P(A)$ **and** $P(B|A) = P(B)$ **if and only if** A **and** B **are independent events.**

Two events are **dependent** if one event occurring causes the probability of the second event to go up or down.

The **conditional probability** of event A given event B is the probability of event A occurring given event B occurred. The notation is $P(A|B)$, which is read as "the probability of A given B".

A **two-way table,** or **contingency table,** organizes data when two categories are associated with each person/object being classified.

Guided Practice

A group of 110 students was surveyed about what grade they were in and whether they preferred dogs or cats. 20 9th graders preferred dogs, 5 9th graders preferred cats, 16 10th graders preferred dogs, 4 10th graders preferred cats, 22 11th graders preferred dogs, 6 11th graders preferred cats, 30 12th graders preferred dogs, and 7 12th graders preferred cats.

1. Construct a two-way frequency table to organize this data.
2. Suppose a person is chosen at random from this group. Let C be the event that the student prefers cats. Let T be the event that the student is in 10th grade. Find $P(C)$ and $P(C|T)$.
3. Are events C and T independent?

Solutions

TABLE 6.17:

	Dogs	Cats	*Total*
9th Grade	20	5	*25*

TABLE 6.17: (continued)

10th Grade	16	4	*20*
11th Grade	22	6	*28*
12th Grade	30	7	*37*
Total	*88*	*22*	*110*

2. There are 110 students total. 22 of them prefer cats. $P(C) = \frac{22}{110} = 20\%$. $P(C|T)$ means the probability that the student prefers cats given that they are in 10th grade. Restrict the sample space to the 20 10th grade students. 4 of them prefer cats. $P(C|T) = \frac{4}{20} = 20\%$.
3. The events are independent because $P(C) = P(C|T)$. Being in 10th grade does not affect the probability of the student preferring cats.

More Practice

For 1-5, use the following information:

A group of 64 people were surveyed about the type of movies they prefer. 12 females preferred romantic comedies, 10 females preferred action movies, and 3 females preferred horror movies. 8 males preferred romantic comedies, 25 males preferred action movies, and 6 males preferred horror movies.

1. Construct a two-way frequency table to organize this data.

   Suppose a person is chosen at random from this group.

2. Let F be the event that the person is female. Find $P(F)$.
3. Let R be the event that the person prefers romantic comedies.

   Find $P(R)$.

2. Find $P(F|R)$ and $P(R|F)$. Explain how these two calculations are different.
3. Are events F and R independent? Justify your answer.

For 6-10, use the following information:

The middle school students in your town were surveyed and classified according to grade level and response to the question "how do you usually get to school"? The data is summarized in the two-way table below.

TABLE 6.18:

	Walk	**Bus**	**Car**	*Total*
6th Grade	30	120	65	*215*
7th Grade	25	170	25	*220*
8th Grade	40	130	41	*211*
Total	*95*	*420*	*131*	*646*

6. If a student is chosen at random from this group, what is the probability that he or she is a 6th grade student who takes the bus?
7. If a 6th grade student is chosen at random from this group, what is the probability that he or she takes the bus?
8. If a student who takes the bus is chosen at random from this group, what is the probability that he or she is in

6th grade?

9. The previous three questions each have to do with 6th grade and taking the bus. Why are the answers to these questions different?
10. Are the events "being in 6th grade" and "taking the bus" independent? Justify your answer.

For 11-15, use the following information:

A hospital runs a test to determine whether or not patients have a particular disease. The test is not always accurate. The two-way table below summarizes the numbers of patients in the past year that received each result.

TABLE 6.19:

	Positive Result on Test	**Negative Result on Test**	*Total*
Has Disease	100	4	*104*
Does Not Have Disease	12	560	*572*
Total	*112*	*564*	*676*

11. If a patient is chosen at random from this group, what is the probability that he or she has the disease?
12. A patient from this group received a positive test result. What is the probability that he or she has the disease?
13. A patient from this group has the disease. What is the probability that he or she received a positive result on the test?
14. A "false positive" is when a patient receives a positive result on the test, but does not actually have the disease. What is the probability of a false positive for this sample space?
15. How many of the 676 patients received accurate test results?