

### Activity: Bayes' Theorem

**Concepts:** The Law of Total Probability and Bayes' Theorem

**Prerequisites:** The student should know how to use conditional probabilities, the multiplication rule, and the law of total probability.

#### Scenario 1: AIDS Testing

The ELISA test for AIDS is used in the screening of blood donations. As with most medical diagnostic tests, the ELISA test is not infallible. If a person actually carries the AIDS virus, experts estimate that this test gives a positive result 97.7% of the time. (This number is called the *sensitivity* of the test.) If a person does not carry the AIDS virus, ELISA gives a negative result 92.6% of the time (the *specificity* of the test). Recent estimates are that 0.5% of the American public carries the AIDS virus (the *base rate* with the disease). We will determine the (conditional) probability that a person actually carries the AIDS virus given that he/she tests positive on the ELISA test.

(a) *Prediction!* First, without doing any calculations, take a *guess* for the value of the (conditional) probability that a person who tests positive carries the virus:

#### Solution Through Two-Way Tables:

Before turning to a formal probability rule to address this issue, you can answer the question and develop some intuition for how the probabilities relate to each other by creating a 2x2 table. Imagine a hypothetical population of 1,000,000 people for whom these percentages hold exactly. (The population size is chosen to be so large in order to make the calculations all work out to be integers.) For questions (b)-(e) below, record your answers in the appropriate cells of the table.

(b) Assuming that 0.5% of the population of 1,000,000 people carries AIDS, how many such carriers are there in the population? How many non-carriers are there?

(c) Consider for now just the *carriers*. If 97.7% of them test positive, how many people test positive? How many carriers does that leave who test negative?

(d) Now consider only the *non-carriers*. If 92.6% of them test negative, how many test negative? How many non-carriers does that leave who test positive?

	Positive test	Negative test	Total
Carries AIDS virus	(c)	(c)	(b)
Does not carry AIDS	(d)	(d)	(b)
Total	(e)	(e)	1,000,000

(e) Determine the total number of positive test results and the total number of negative test results.

(f) *Of those who test positive*, what proportion actually carry the disease? How does this compare to your prediction above?

- (g) Explain why this probability turns out to be small compared to the sensitivity and specificity. (Be sure to refer to calculations in the table.)

**Derivation:**

You will use this situation to derive an important probability result known as *Bayes' Theorem*. Let A denote the event that the person carries the disease, P denote a positive test result, and N denote a negative test result. In other words, the events under consideration are:

- $A = \{\text{person carries the AIDS virus}\}$
- $P = \{\text{test result is positive}\}$
- $N = P' = \{\text{test result is negative}\}$

- (h) Express the base rate of .005 as a probability in the notation of these events.
- (i) Express the sensitivity .977 and specificity .926 of the test as conditional probabilities involving these events.
- (j) Consider the conditional probability of having AIDS given a positive test result that you calculated in (f). Express this probability (only its notation) using these symbols.
- (k) Use the definition of conditional probability, to express the probability in (j) as the ratio of two probabilities (again, just report the symbols at this point).
- (l) First examine the numerator. Use the multiplication rule to express the probability of this intersection as the product of a conditional probability and an unconditional probability. [Hints: "Reverse" the conditioning from the probability that you are looking for. Be sure to use only probabilities that are given regarding the ELISA test.]
- (m) Calculate this probability expressed in question (l). Verify that multiplying this by 1,000,000 agrees with the numerator calculated in question (f).
- (n) Now examine the denominator in (k). Use the Law of Total Probability to express this as the sum of two products of two probabilities (just report the notation at this point).
- (o) Calculate this probability in (n). Verify that multiplying by 1,000,000 agrees with the denominator calculated in question (f).

- (p) Combine your answers to (k), (l), and (n) to form an equation expressing  $P(A|P)$  in terms of  $P(A)$ ,  $P(P|A)$ , and  $P(P|A')$ .

This result is known as *Bayes' Theorem*. It provides a mechanism for updating uncertainty about a hypothesis (H) in light of new evidence (E). In its simplest form, Bayes' Theorem can be

$$\text{written as: } P(H | E) = \frac{P(E | H)P(H)}{P(E | H)P(H) + P(E | H')P(H')}.$$

This theorem is exactly what you did with your analysis of the two-way table above. The numerator of the probability in question (f) was  $(.977)(.005)(1,000,000) = 4885$ . The denominator in question (f) was  $(.977)(.005)(1,000,000) + (.074)(.995)(1,000,000) = 78,515$ . When we take the ratio, the arbitrary constant of 1,000,000 in the population drops out.

### **Scenario 2: Presidential Election Votes (cont.)**

Recall your analysis of CNN's exit poll results from the 2000 Presidential election. Now suppose that you want to find the conditional probability that an interviewee is African-American given that he/she voted for Al Gore.

- (q) Express this probability in terms of the events  $A = \{\text{African-American}\}$  and  $G = \{\text{voted for Gore}\}$ .
- (r) Recall and report how you earlier calculated  $P(A \cap G)$ .
- (s) Recall and report how you earlier calculated  $P(G)$ .
- (t) Use the probabilities reported in (r) and (s) to calculate  $P(A|G)$ .

When there are several events being considered, a more general form of Bayes' Theorem applies. Let  $k$  be the number of events and  $i$  designate the hypothesis whose updated probability is to be

calculated. Then Bayes' Theorem asserts that: 
$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{j=1}^k P(B | A_j)P(A_j)}$$

The probabilities  $P(A_i)$  are sometimes called *prior probabilities*, and the conditional probabilities  $P(A_i|B)$  are called updated, or *posterior probabilities*. As you have seen, this theorem follows from the definition of conditional probability, multiplication rule, and law of total probability. It requires that the events  $A_i$  be mutually exclusive and exhaustive.

**Scenario 3: Programming Bugs**

Suppose that you have three programmers designing computer code for a project: Alice has designed 60% of the code, Barb 30% and Chuck 10%. Suppose further that Alice has a bug in 3% of her work, Barb in 7% of her work, and Chuck in 5% of his.

- (u) What percentage of the code written has a bug? In other words, what is the probability that a randomly selected of code has a bug? [*Hint*: Use the law of total probability.]
- (v) Given that you find a bug in a line of code, who is most likely to have written it? Who is least likely? [*Hint*: Use Bayes' Theorem to find each person's conditional probability of having written the line given that it has a bug. Notice that you already calculated the denominator in (u).]

(w) Fill in the following probability table to represent this situation:

	Alice	Barb	Chuck	Total
Bug				
No bug				
Total				1.000

**Scenario 4: Multiple Choice Exams**

Suppose that a student knows (with certainty) the answer to 50% of the questions on a multiple choice exam, while on the other 50% he is clueless and so guesses randomly among the four choices.

- (x) Determine the percentage of questions that he answers correctly, and determine the probability that he actually knew the answer given that he answers correctly. [*Hints*: Define the relevant events carefully. Use the Law of Total Probability and Bayes' Theorem.]
- (y) Now suppose that there are  $k$  choices on each question, where  $k$  is some integer greater than one. Determine the probabilities in (x) as functions of  $k$ .