

Developing Bayes' Theorem

In this activity, you will work in teams of 3–4 students to learn about a very important concept in probability and statistics called Bayes' Theorem. Bayes' Theorem has applications in many social and scientific settings. You will explore applications from medicine and public health in this activity.

Content Learning Objectives

After completing this activity, students should be able to:

- Describe conditional probability
- Interpret probabilities, specificity and sensitivity rates, and prevalence
- Construct Bayes Theorem

Process Skill Goals

During the activity, students should make progress toward:





















































- Evaluating, interpreting, manipulating, or transforming information (Information Processing)



Copyright © 2021 Angela Ebeling, Katie Fitzgerald, and Olga Glebova. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Model 1 Review probability using a cards example

Example set of 52 playing cards; 13 of each suit: clubs, diamonds, hearts, and spades

	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													

Questions (5 min)

Start time:

1. Based on the screenshot above:
 - a) How many cards are there in a standard deck of cards?
 - b) How many red cards are there?
 - c) Is the number of black cards the same as the number of red?
 - d) How many cards are in the suit hearts?
 - e) Are there the same number of cards in the suit clubs?
 - f) When you add up the number of hearts, clubs, diamonds, and spades, what number do you get?
 - g) Is this the same as the total number of red + black cards?

2. If someone shuffled the deck of cards, and you picked one card...

- a) ...what is the probability it will be a red card? *Hint: think how many red cards are there out of the total possible cards.*
- b) ...what is the probability of a card being hearts?

3. Before knowing any information about your card, you just calculated that the probability of the card being red was $\frac{1}{2}$ or 50% and the probability of the picked card being hearts was $\frac{1}{4}$ or 25%. Now let's say you picked a card from the standard deck of 52 cards. Your friend looks at your card and tells you that it is red (your friend looked at the card but you didn't so you do not know what it is, neither suit nor denomination).

- a) Given the information that you chose a red card from the deck, find the probability the red card is a King? Show your work. *Hint: how many Kings are there among the red cards?*
- b) Given the information that you chose a red card from the deck, find the probability that the red card is of suit hearts. Show your work. *Hint: recall how many hearts are there out of the total number of red cards.*

Conditional Probability:

We use the notation

$$P(\text{hearts}|\text{red})$$

to mean

“probability of a card being hearts, given that it is a red card”

This is called *conditional probability*. Conditional probability allows us to use new information (e.g. the card is red) to update our hypothesis about the probability of an event happening (e.g. the card is hearts). Before we knew anything about the card, there was a 25% probability the card was hearts. But once we knew the card was red, the probability it was hearts changed to 50%. We incorporated new information (we “conditioned” on it) to update our hypothesis about the probability of the card being hearts.

4. Using the notation we introduced above as a guide, try writing notation for the probability you found in question 3a. That is, what notation would we use to describe the “probability that a card is a King, given that it is a red card”?

5. How would you write the notation “P(black | Queen)” in English?

Model 2 Using conditional probabilities

The following example was inspired by [AskGoodQuestions blog](#), by Allan Rossman [3].

The ELISA test for HIV was developed in the mid-1980s during the peak of the AIDS epidemic in the United States. Blood samples were tested to detect whether or not an HIV-infection was present. As with any medical diagnostic test, the results will sometimes be wrong. That is, sometimes a person will actually have an HIV infection and the ELISA test will return a negative result, and other times a person will NOT have an HIV infection but the ELISA test will return a positive result. In the early stages of ELISA's development, the test was found to return correct results more than 90% of the time both when a person was actually HIV-infected and when they were NOT HIV infected.

6. Thought question (do not spend much time on this.. do not use the table or any data, just think about what was discussed in the paragraph above): Make a prediction for the percentage of blood samples with positive test results that are actually infected with HIV ($\Pr(\text{HIV} | +)$). In other words, if you got a positive test result, what's the probability you are actually infected with HIV?

The table below gives the results of the ELISA test for a hypothetical population of 1 million people. The numbers are based on real data about the effectiveness of the test found in this 1987 article: <https://projecteuclid.org/journals/statistical-science/volume-2/issue-3/The-Statistical-Precision-of-Medical-Screening-Procedures-Application-to/10.1214/ss/1177013215.full>.

	Positive Test Result	Negative Test Result	Total
Actually HIV infected	4885	115	5000
Actually NOT HIV infected	73630	921370	995,000
Total	78515	921485	1,000,000

Questions (20 min)

Start time:

Guided questions to develop percentages/probabilities:

7. Use the Table above to answer the following questions:

- How many people are actually HIV-infected?
- Among those who are actually HIV-infected, how many tested positive?
- Among those who are actually HIV-infected, how many tested negative?

- d) If you add your answers to b and c, you should obtain the total number of people who actually have the disease in the population. Does this match your answer to a? *Hint: it should.*
- e) How many people are in this population in total?
- f) Now that we calculated 5,000 people are actually HIV-infected and there are in total 1,000,000 people in the population, what proportion of the population are HIV-infected?

This is called the *prevalence*, or sometimes the *base rate* or the *prior* information.

Prevalence:

the number of people in the sample with the characteristic of interest, divided by the total number of people in the sample. This is also called the *base rate* or the *prior*.

We use the notation

$$P(\text{HIV})$$

to mean

“the probability of having HIV.”

8. Now that we know from Question 7b that 4885 people who are actually HIV-infected tested positive, we can investigate that row.

- a) Focusing on the people with a positive HIV result, what proportion of HIV-infected people tested positive?

This is called the *sensitivity* of the test: the probability of getting a (correct) positive result when one is HIV-infected.

Sensitivity:

the proportion of people with a condition who are correctly identified by a screening test as indeed having that condition.

We use the notation

$$P(+ | \text{HIV})$$

to mean

“the probability of testing positive given that you have HIV.”

- b) Would you want the sensitivity of the test to be high or low? Justify your answer.

A 100% sensitivity would mean the test correctly detects an HIV infection every time it is present. The lower the sensitivity, the more often the test returns false negatives, which we want to avoid.

9. Now let's focus on yet another aspect.

- a) How many people are NOT HIV-infected? What proportion of the population does this represent?
- b) Among those who are NOT actually HIV-infected, how many tested negative? What proportion of non-HIV-infected people does this represent? That is, what proportion of non-HIV-infected people tested negative?
- c) What notation would you use for this probability?

This is called the *specificity* of the test: the probability of getting a (correct) negative result when you are NOT HIV-infected.

Specificity:

Specificity is the proportion of people without a condition who are correctly identified by a screening test as indeed not having the condition. We use the notation

$$P(- | \text{Not HIV})$$

to mean

“the probability of testing negative given that you do NOT have HIV. ”

d) Would you want the specificity of the test to be high or low? Justify your answer.

A 100% specificity would mean that every person who does not have HIV would correctly receive a negative result. The lower the specificity, the more often the test returns false positives.

e) How many people received a positive test result?

f) Among those people (people receiving a positive test result), are there more people that are actually HIV infected or actually NOT HIV infected?

10. Return to the question you tried to make a prediction for at the beginning of this model. *“Thought question: Make a prediction for the percentage of blood samples with positive test results that are actually infected with HIV. In other words, if you got a positive test result, what’s the probability you are actually infected with HIV?”* Let’s investigate that using the information in the table.

a) Among those who tested positive, what percentage of blood samples are actually infected with HIV?

b) What is the notation for this probability?

- c) Try to guess why $P(\text{HIV} \mid +)$ is so low, despite the fact that the test is correct 97.7% of the time when someone is actually HIV infected, and correct 92.6% of the time when someone is NOT actually HIV infected. *Hint: consider your answers to 9f, 9a, and 7f*
- d) Now summarize the information you have discovered so far. Write out the notation and the calculated value for each of the following: base rate (prior), sensitivity, specificity, and the updated probability (posterior) after knowing there is a positive test result.

Model 3 Discovering Bayes' Theorem

Let's consider a different infectious disease that is more prevalent than HIV: influenza A (the common flu). We'll assume the *prevalence* is 8%. The flu is usually diagnosed via Rapid Influenza Diagnostic Tests (RIDTs). We'll assume these tests have a 97.7% sensitivity and 92.6% specificity just like the ELISA HIV test; these rates are on par with what the FDA requires for RIDTs (<https://www.cdc.gov/flu/professionals/diagnosis/rapidlab.htm>). Often, this is the only information we have in real life - we don't observe all the counts or results for everyone in the population, but we do have estimates of the prevalence, sensitivity, and specificity.

	Positive Test Result	Negative Test Result	Total
Actually Flu infected			
Actually NOT Flu infected			
Total			1,000,000

Questions (40 min)

Start time:

11. Since Flu has a higher prevalence than HIV, make a prediction for whether the $P(\text{disease} \mid +)$ will be higher or lower for the Flu compared to HIV? Note, remember in Model 2, the last question (10), you calculated the $P(\text{HIV} \mid +)$. This question is asking whether you think $P(\text{Flu} \mid +)$ similar or different than $P(\text{HIV} \mid +)$ based on these diseases' different prevalence rates.

12. Now be more specific, make a prediction for the percentage of positive RIDT samples that are actually infected with Influenza A. In other words, if you got a positive test result, what's the probability you are actually infected with influenza A?

13. For each of the following, calculate the answer and fill in the proper square in the table:
 - a) If 8% of the population (of 1,000,000 people) is infected with the flu, how many people are actually flu-infected?
 - b) How many people are not infected?
 - c) If the test gives a (correct) positive result for 97.7% of samples that are infected with the flu, how many flu-infected people will get a (correct) positive result?
 - d) How many flu-infected will get a (false) negative result?
 - e) If the test gives a (correct) negative result for 92.6% of samples that are NOT infected with the flu, how many non-flu-infected people will get a (correct) negative result?

f) How many people who are NOT flu-infected will get a (false) positive result?

g) How many people tested positive altogether?

h) How many people tested negative altogether?

14. Return to the question you tried to make a prediction for at the beginning of this model: What percentage of samples with positive test results are actually infected with Influenza A? Answer this using what you now know about the table.

15. What you just found was $P(Flu|+)$. How does this compare to $P(HIV|+)$ that you found in Model 2? Give an explanation for why you think these numbers are so different.

16. Let's work to establish a formula for how we got to $P(Flu|+) = \frac{78160}{146240} = 0.5344$.

a) **Numerator:** What 3 numbers did we multiply to get 78,160, the number of true positive results in the population? (Hint: we first multiplied two numbers to get 80,000, then we multiplied 80,000 by one more number to get 78,160.) (Note: there are other ways to obtain the numerator, but this is designed to help you discover an important relationship.)

b) **Denominator:** To get the denominator (146,240), we can add 2 numbers from the table. Which two numbers are those?

- c) Check to make sure your answers to parts a and b match the information below. Then, using the same pattern for representing the numerator as 3 numbers being multiplied together, fill in the blanks to rewrite the denominator in the same format.

$$P(Flu|+) = \frac{78160}{146240} = \frac{(0.977)(0.08)(1,000,000)}{78,160 + 68,080}$$

$$= \frac{(0.977)(0.08)(1,000,000)}{(0.977)(0.08)(1,000,000) + (\quad)(\quad)(\quad)}$$

- d) Notice that each term has the population size 1,000,000 in it. These cancel out. Fill in the blanks to rewrite the formula without the population size:

$$P(Flu|+) = \frac{(\quad)(\quad)}{(\quad)(\quad) + (\quad)(\quad)}$$

- e) The following numbers should have appeared in your answer above:

- A 0.977 (appears twice)
- B 0.08 (appears twice)
- C 0.074
- D 0.92

Match the above numbers with the appropriate notation:

P(+ | NOT Flu)

P(+ | Flu)

P(Flu)

P(NOT Flu)

f) Using the notation from part e and your results from part d, re-write $P(\text{Flu} \mid +)$ using notation only.

$$P(\text{Flu} \mid +) = \frac{(\quad)(\quad)}{(\quad)(\quad) + (\quad)(\quad)}$$

You have constructed a formula called *Bayes' Theorem!*

- $P(+ \mid \text{Flu})$ is called the *likelihood*, which refers to the likelihood (aka probability) of testing positive if someone has the flu (Note: in this context, this is sensitivity.)
- $P(\text{Flu})$ is called the *prior* (Note: in this context, this is disease prevalence.)
- The denominator, $P(+ \mid \text{Flu})P(\text{Flu}) + P(+ \mid \text{NOT flu})P(\text{NOT flu})$ is called the *marginal probability*.
- The denominator can also be written as $P(+)$, which refers to "probability of testing positive", and this notation is widely used.
- $P(\text{Flu} \mid +)$ is called the *posterior*

Note that $P(\text{Flu})$, the **prior**, tells us the probability a person has the Flu *before* they've taken a test (i.e. *prior* to any additional evidence), whereas the **posterior** tells us the probability a person has the Flu *after* they've received a positive test. We can think of this as being an *updated* probability, after we've observed some data (the result of the test).

The general notation for Bayes' Theorem is:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

17. Write out Bayes' Theorem using notation for the Flu. (A = test positive, B = Flu)

18. Returning to the HIV example from Model 2,

- a) Write out a formula for $P(\text{HIV} \mid +)$, using the notation above as a guide. (A = test positive, B = HIV)

- b) What is the *likelihood* for the HIV example (sensitivity)? Provide the correct notation and the correct numerical value.
- c) What is the *prior* for the HIV example (prevalence)? Provide the correct notation and the correct numerical value.
- d) What is the *posterior* for the HIV example? Provide the correct notation and the correct numerical value.
- e) Now directly compare the posterior probability for Model 2 ($P(\text{HIV} | +)$) with Model 3 ($P(\text{Flu} | +)$). Which is higher? Does this match your answer to the first question in Model 3?

19. Explain in words how the *prior* of a disease (e.g. the prevalence = $P(\text{disease})$) influences how worried you should be after receiving a positive test result.

In real life we do not usually have all of the information from the population, thus we don't usually have the full two-way table from which to calculate $P(\text{disease} | +)$ directly. So Bayes' Theorem is a useful way to use information we might have access to (base rate: $P(\text{disease})$, sensitivity ($P(+ | \text{disease})$), and proportion of positive tests ($P(+)$) to still be able to calculate the probability of having the disease if we test positive: $P(\text{disease} | +)$. We've also learned that Bayes' Theorem allows us to update our beliefs once we have observed data - that is, we can use information that we have before (the prior - prevalence of the disease in the population ($P(\text{disease})$)) - and update it with some data (the result of the test) to give you an updated probability on the likelihood of having the disease.

If your curiosity is peaked, consider reading more about Bayesian reasoning through these freely accessible online textbooks.

- Bayes Rules! An Introduction to Applied Bayesian Modeling (by Alicia Johnson, Miles Ott, and Mine Dogucu). [Bayes Rules](#)[2]
- Probability and Bayesian Modeling (by Jim Albert and Jingchen Hu). [Probability and Bayesian Modeling](#)[1]

References

- [1] Jim Albert and Jingchen Hu. *Probability and Bayesian Modeling*. 2020. URL: <https://bayesball.github.io/BOOK/probability-a-measurement-of-uncertainty.html>.
- [2] Alicia Johnson, Miles Ott, and Mine Dogucu. *Bayes Rules! An Introduction to Applied Bayesian Modeling*. 2021. URL: <https://www.bayesrulesbook.com/>.
- [3] Allan Rossman. *Ask Good Questions*. URL: <https://askgoodquestions.blog/2019/09/09/10-my-favorite-theorem/>. (accessed: 05.29.2024).