AN EXCEL PACKAGE TO EXPLAIN BAYES' THEOREM

Richard L. Morris
Computer Science and Quantitative Methods
Winthrop University

Abstract

This article presents an Excel package to illustrate Bayes' Theorem. It does so in two ways: First, a graphical approach is presented that represents the various probabilities involved in Bayes' Theorem. Secondly, an intuitive approach is used that to many people is easier to understand than the traditional Bayes' formula.

Introduction

Bayes' Theorem is a very important topic in probability and statistics. Often, conditional probabilities are in one form and they need to be in another. At its simplest, Bayes' Theorem involves switching around these conditional probabilities. Suppose, for example, that a drug test is 98% "accurate" for drug users and 92% "accurate" for non-users. Someone unfamiliar with Bayes' Theorem might say, then, that if someone tests positive on the test, indicating that he or she is a user, the probability that the subject is actually a user is 98%. Likewise, if a person tests negative, indicating that he or she is a non-user, the probability the subject is a non-user is 92%. Neither of these conclusions is correct, however, except in a certain special case, as discussed later. The problem is that the conditional probabilities of 98% and 92% are in one direction and the desired probability is in the other.

Bayes' Theorem is useful in many areas: It is used in oil exploration to revise the probability of a drilling site producing oil given previous experience [1, 2]. In drug testing, it is important in determining the probability of a person's being a drug user given a positive result on a drug test [3]. In medical diagnosis screening tests are given to determine if the probability of a patient's having a disease is high enough to warrant further procedures, such as a biopsy [4]. It is very important in the area of legal evidence. Very often, probabilities introduced into evidence are in one form of a conditional probability and the transposed probability is the appropriate one to use [5, 6, 7].

In addition to having many specific applications, a knowledge of Bayes' Theorem is fundamental to an understanding of statistics and exactly what is meant in interpreting a confidence interval or a test of hypotheses. Furthermore, there is a Bayesian approach to hypotheses testing which offers a completely different approach than the one most commonly used [8, 9].

An Example

Suppose a test is used to detect the presence of drugs. Adopting the following notation:

D = the subject is a drug user
D^c (D-complement) = the subject is a non-user
T = positive result on the test
T^c (T-complement) = negative result on the test

Then the probabilities in the previous section are:

$$P(T|D) = .98$$

and

$$P(T^c|D^c) = .92$$

P(T|D) is referred to as the *sensitivity* of the test; $P(T^c|D^c)$ is called the *specificity*. When a test is conducted and a positive result occurs,

the following conditional probability is the one desired:

$$P(D|T)$$

That is, given a positive result, what is the probability that the subject is actually a user? It would be more convenient, of course, if the original probabilities were in this form. The reason they aren't is that the original likelihoods are dependent only on the accuracy of the test itself. The revised probability, $P(D|T)$, is dependent on both the accuracy of the test *and* the proportion of drug users in the tested population. This is true not only in drug testing. In any medical test, the original likelihoods, $P(T|D)$ and $P(T^c|D^c)$, describe the accuracy of the test, but a physician or patient would like to know the probability of having a condition given a positive result on the test.

The task, then, is to "switch" the conditionals given, and this is the task of Bayes' Theorem. There are many formulations of Bayes' Theorem, but the following, in the opinion of this author, is the most easily understandable:

$$P(D|T) = \frac{P(D \cap T)}{P(T)} \qquad (1)$$

Where $P(D \cap T)$ is the joint probability of the tested subject being a drug user and yielding a positive test result. It is calculated using the following formula:

$$P(D \cap T) = P(D)P(T|D) \qquad (2)$$

If 5% of the population being tested are users, then $P(D) = .05$ and:

$$P(D \cap T) = .05 \times .98 = .049$$

The denominator of Bayes' Theorem is the marginal (or simple) probability of a positive result on the test. It is calculated using the following:

$$P(T) = P(D \cap T) + P(D^c \cap T) \qquad (3)$$

The second term in the above equation is the joint probability of the subject's being a non-user and having a positive result on the test. It is calculated as:

$$P(D^c \cap T) = P(D^c)P(T|D^c) \qquad (4)$$

To solve the above equation, it is easy to find

$$P(D^c) = 1 - P(D) = 1 - .05 = .95$$

and

$$P(T|D^c) = 1 - P(T^c|D^c) = 1 - .92 = .08$$

so that

$$P(D^c \cap T) = .95 \times .08 = .076$$

and

$$P(T) = .049 + .076 = .125$$

and finally:

$$P(D|T) = \frac{.049}{.125} = .392$$

This means that if a person tests positive on the test, the probability that he or she is actually a drug user is 39.2%. Not surprisingly, many people find this hard to accept. After all, the test is 98% accurate for drug users, yet the probability that a person who tested positive is a drug user is only 39.2%. On the face of it, this doesn't seem like a very good test. The explanation lies in the fact that only 5% of the tested population are drug users and 95% are non-users. In this case, the number of non-users who test positive will outnumber the number of users who test positive. So there's nothing wrong with the test. The problem is that the proportion of users is vastly outweighed by the proportion of non-users. The test is, as a matter of fact, very good. It took the original probability of 5% and leveraged (or, in probabilistic terminology, "revised") it into one of 39.2%.

A Graphical Explanation

The preceding is difficult to explain to undergraduates, or for that matter almost anyone else. Therefore the Excel package presented here was undertaken in an effort to explain the problem graphically [10]. Figure 1 shows the input section and resulting graph for the problem described above. The original probabilities are entered into the gray-shaded cells and the other parts are calculated automatically. The comments show each result in probabilistic notation. In the figure here, the comments are shown, although in the spreadsheet itself they appear only when the cursor moves over them. The spreadsheet then takes this information and performs the calculations necessary for the graph. In this graph, the proportion of users in the tested population, 5%, is in the smaller bar to the left. The proportion of non-users, 95%, is in the larger bar to the right. The proportion of true positives, $P(D \cap T) = 4.9\%$, is represented in the larger, darker portion of the bar representing the users. The remainder (.1%) is light-colored but barely visible. Likewise, the smaller portion of the bar for non-users represents the proportion of false positives, [$P(D^c \cap T) = 7.6\%$]. The light portion of this bar is the proportion of true positives, or 87.4%. The test did its job; it correctly identified 98% of the users and 92% of the non-users. However, since there were many more non-users than users, and because there were more false positives than true positives, the final probability that a person who tested positive is actually a user is only 39.2%.

Parenthetically, in order for the final answer, $P(D|T)$, to be the original 98% "accuracy," both the sensitivity and the specificity would have to be 98%, which would be very unlikely, and the proportion of users in the tested population would have to be 50%.

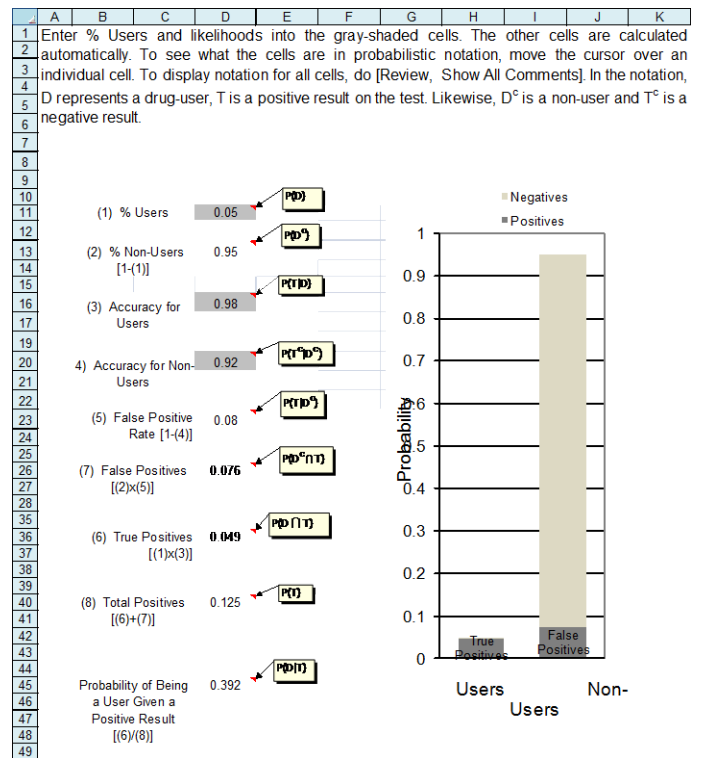In the package, a sheet runs through the calculations above, shown in Figure 2.

Figure 1. Input section and graph.

Figure 2. Calculation section for Bayes' Theorem.

An important side-benefit of this sheet is that students can use it to make up their own problems for additional practice. All they have to do is key in the original information in the

input section and the package goes through the calculations.

Another Explanation Using Frequencies

There is a more intuitive way of applying Bayes' Theorem that many find easier to understand. Suppose that there are 1,000 people in the population to be tested. If that is the case, the expected number of users in the population would be:

.05 x 1,000 = 50

The term "expected" means not there will always be 50 drug users in a population of 1,000, but that if this experiment is repeated many, many times, on average there will be 50 users in a population. Of these 50, the expected number of positive results would be:

50 x .98 = 49

This is the expected number of true positives in the tested population.

Likewise, there would be 950 non-users in the population, calculated either as:

1,000 - 50 users = 950

or:

1,000 x (1 - .05) = 950

Of these 950 non-users, there are expected to be 874 negative results (950 multiplied by the 92% accuracy for non-users) and 76 (950 – 874) positive results. The figure of 76 represents the expected number of false positives.

Then the total number of positive results is:

Total positives = True positives + False positives or:

Total positives = 49 + 76 = 125

Finally, the proportion of true positives to total positives is:

49/125 = .392

This is the probability of a subject who tests positive being a user, or P(D|T), the number previously calculated using Bayes' Theorem. It will be noticed that each of the expected frequencies in the above solution mirrors one of the probabilities used in the formulaic approach to Bayes' Theorem in Figure 1. For example, the 49 true positives corresponds to P(D∩T) = .049.

There is a sheet showing the calculations for the intuitive approach in the spreadsheet. This is shown in Figure 3.

Figure 3. Calculation section for the intuitive approach to Bayes' Theorem.

In this sheet, the probabilities specified in the original entries (Figure 1) are automatically carried over so that the user can see how they are used in the intuitive approach. The population size in cell E18 is determined so that all the frequencies are integer-valued. First, the LEN function is used to find the length beyond the decimal point of each of $P(D \cap T)$ and $P(D^c \cap T)$. Then the larger of these (in this case, both are the same) is taken to determine the sample size, n, using the following relationship:

$$n = 10^{\text{Max length}[P(D \cap T), \ P(D^c \cap T)]}$$

Since $P(D \cap T) = .049$ and $P(D^c \cap T) = .076$, and $\text{Max length}[P(D \cap T), \ P(D^c \cap T)] = 3$,

$$n = 10^3 = 1{,}000$$

There is one last sheet in the package, showing a graph of the expected frequencies. Since this graph is essentially a repeat of Figure 2, it is not shown here.

Conclusion

Since Bayes' Theorem is one of the more difficult topics in statistics, it is important to present it in an easy-to-understand manner. The graphical approach illustrates why it often occurs that a revised probability is different from what common sense would suggest. The intuitive approach using frequencies will often make more sense to a student than Bayes' Theorem itself. Both of these methods combined have helped students understand the concepts behind Bayes' Theorem more easily.

References

1. L. Benkherouf and J. A. Bather, "Oil Exploration: Sequential Decisions in the Face of Uncertainty," *Journal of Applied Probability*, vol. 25, no. 3, Sept. 1988.

2. K. Hendricks and R. H. Porter, "The Timing and Incidence of Exploratory Drilling on Offshore Wildcat Tracts," *American Economic Review*, vol. 86, no. 3, June 1996.

3. D. T. Barnum and J. M. Gleason, "The Credibility of Drug Tests: A Multi-Stage Bayesian Analysis," *Industrial and Labor Relations Review*, vol. 47, no. 4, July 1994.

4. J. L. Gastwirth, "The Statistical Precision of Medical Screening Procedures: Application to Polygraph and AIDS Antibodies Test," *Statistical Science*, vol. 2, no. 3, Aug. 1987.

5. D. L. Faigman and A. J. Baglioni, Jr., "Bayes' Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence," *Law and Human Behavior, v*ol. 12, no. 1, Mar. 1988.

6. W. C. Thompson and E. L. Schumann, "Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy," *Law and Human Behavior*, vol. 11, no. 3, Sept. 1987.

7. S. E. Fienberg and M. J. Schervish, "The Relevance of Bayesian Inference for the Presentation of Statistical Evidence and for Legal Decisionmaking," *Boston University Law Review*, vol. 66, 1986.

8. I. J. Good, "Interface Between Statistics and Philosophy of Science," *Statistical Science*, vol. 3, no. 4, Nov. 1988.

9. B. Efron, "Why Isn't Everyone a Bayesian?" *The American Statistician*, vol. 40, no. 1, Feb. 1986.

10. http://faculty.winthrop.edu/morrisr/VisualBasic.htm

Biographical Information

Richard Morris is a Professor of Quantitative Methods at Winthrop University. He has a B.S. in mechanical engineering from West Virginia University, an M.B.A. from the College of William and Mary and a Ph.D. in Quantitative Methods from Virginia Tech. His current research interests are in statistics, quality management, and applications of Visual Basic and Excel.