Original Article

Beginning Bayes

Tim Erickson
Epistemological Engineering and Lick-Wilmerding High School, San Francisco, California, USA
e-mail: eepsmedia@gmail.com

Summary Understanding a Bayesian perspective demands comfort with conditional probability and with probabilities that appear to change as we acquire additional information. This paper suggests a simple context in conditional probability that helps develop the understanding students would need for a successful introduction to Bayesian reasoning.

Keywords: conditional probability; Bayes; area diagrams.

INTRODUCTION

Many people were gobsmacked by the results of the recent USA presidential elections, citing polls that seemed to say that Trump would lose. Was this an epic failure for statistics?

Maybe not. To give one example, on the morning of the election, Nate Silver and his fivethirtyeight.com Web site gave Trump a 28% chance of winning (FiveThirtyEight 2016). Should we then be surprised by the result? No. Things with a probability of 1/4 happen all the time.

Still, there were problems. The ongoing saga of the election post-mortem makes interesting reading—but this paper will not discuss issues such as the electoral college, correlation, social media, proper sampling, identifying likely voters, or even the effect of weather. Instead, we will discuss the *form* of that prediction, and what question it seems to answer.

Consider two approaches to predicting Trump's performance. One approach is to construct confidence intervals to estimate the proportion of voters supporting him and report margins of error. This approach tries to answer the question, *what proportion of voters will actually vote for Trump*?

In contrast, Nate Silver and his colleagues try to answer a related but different question: *what's the probability that Trump will win*?

Both approaches are valid. But the second is more foreign to most of us and to our students. How do you come up with a probability for an event like an election result? We know how to calculate the probability of an outcome based on some hypothesis. How do we reverse that and calculate the probability of a hypothesis based on some outcomes? The answer: use Bayesian reasoning.

Bayesian approaches to statistical data analysis are increasingly important across disciplines, in research studies and media reporting, but few introductory courses give more than a passing mention of Bayesian techniques. This is hardly surprising; Bayesian analysis is filled with complex ideas. Priors. Posterior probability. Belief. Likelihood. A specific, different meaning of 'hypothesis'. And of course, the use of Bayes' formula.

Suppose we want to introduce this kind of reasoning. What would students need to understand in order to succeed? One ingredient, certainly, is conditional probability. Considerable effort has gone into making conditional probability more comprehensible (e.g., Gigerenzer and Hoffrage 1995; Martignon and Krauss 2009). Let's use their work to reinforce students' understanding of conditional probability and, at the same time, lay a good foundation for a Bayesian perspective.

We describe a lesson in conditional probability that attempts to do just that. If it is all students ever see of Bayesian analysis, they will still benefit. If they go on to learn more, it could be a touchstone to which they can return to understand elements ranging from vocabulary to probabilities that are updated in light of data and in which a subjective component plays a part (de Finetti 1964). We'll use the simplest imaginable situation and see how far we can get.

ACTIVITY: THE TWO-COINS SITUATION

Suppose I have two coins: a regular, fair coin and one with two heads. I choose one at random.

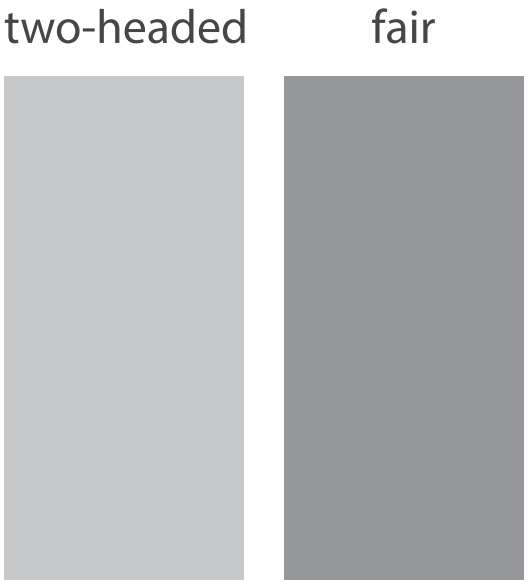

Fig. 1. An area diagram showing the two hypotheses. As we will see, area is proportional to probability

What is the chance that we have chosen the fair coin? The answer is 1/2, or 50%.

Put another way: we have two competing hypotheses: we either have a fair coin or a two-headed coin. Initially, these hypotheses are equally likely because we picked the coin at random. Figure 1 shows a simple area diagram illustrating the two hypotheses.

I flip the coin I chose without showing whether it's double-headed or not. It comes up heads. Do we know which hypothesis is true? Of course not. What is the probability that my coin is 'fair'? Most students say that it's [obviously!] still 1/2.

Now suppose I flip the same coin nine more times and get heads on every flip. It's *possible* the coin is fair, but many students will say the coin *must* be double-headed. We can ask: is the probability that this coin is fair still 1/2? Or did the probability change? Here, some students are uneasy: can probability change? After ten heads in a row, the chance that we have the fair coin can't still be 1/2. Or can it?

We explain that there are, in fact, different perspectives on probability. Let's explore one in which probability appears to change based on what happens—based on the data, based on evidence. And let's be quantitative about it.

Consider the first flip. Figure 2 shows the possible results three different ways. On the left, we see a version with separate H and T boxes only for the fair coin on the right. In the centre, we differentiate the two 'heads' of the double-headed coin with subscripts. On the right, an equivalent diagram shows the double-headed result as two identical boxes.

We can now find the probability that we have the fair coin—by enumeration. The left side of Figure 3 shows our one-flip situation. We now have *evidence*: a flip of 'heads'.

To find the probability *given the evidence*, we focus only on the boxes where heads occurred (shaded in the diagram) and ignore the 'tails' box because it didn't happen. Of the three 'heads' boxes, only one (the darker one) is 'fair'. Thus, given 'heads', the probability of 'fair' has decreased from 1/2 to 1/3. The right-hand diagram extends that principle to a second flip. We subdivide each result from the previous step and obtain a probability of 1/5.

By looking only at the H or HH boxes, we are still considering the probability that the coin I chose is the fair coin, but we are using data to update the sample space or the *reference* for the probability. The new probabilities (1/3 or 1/5) are *conditional* on the data we have—on the evidence. We often express this as 'given the data' and use a vertical bar in our notation.

Discussion and extension

We have created our area diagrams sequentially, splitting them into columns according to the

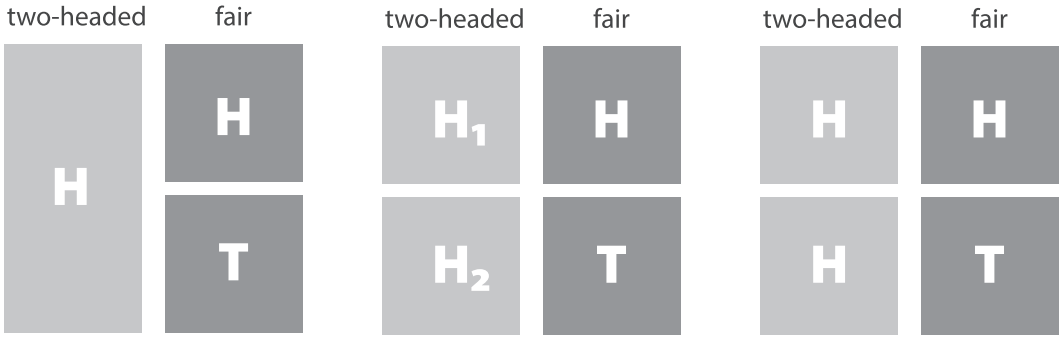

Fig. 2. Equivalent diagrams for the situation after one flip

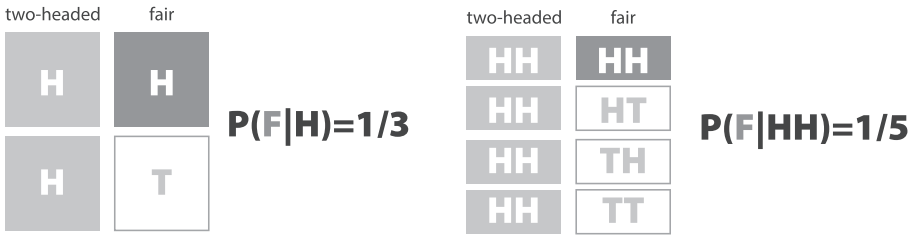

Fig. 3. The probability of our having a fair coin, given "one heads" and "two heads"

hypotheses, that is, the choice of the coin; and then further subdividing the columns so that students see a small set of equally likely outcomes. Finally, we look at the parts of the diagram that match our evidence (one head or two heads) and count the boxes to figure out the probability.

These area diagrams—also called mosaic plots (Friendly 2001)—are common in teaching conditional probability. We carefully defined the situation so that we could (1) easily split the columns into identically sized boxes and (2) emphasize how to update the probability when we acquire additional data.

We can use this same technique with more complex situations. For example, suppose that we agreed, after one flip, that the probability of having the fair coin was 1/3. (We must distinguish that probability from the probability of heads; this latter probability—which has been either 0.5 or 1.0 so far—is a *parameter* that characterizes the hypotheses. We'll call it π.)

To analyse the consequences of obtaining one *additional* head instead of analysing two heads at once—as we did in the right side of Figure 3—we can use diagrams like the ones in Figure 4. Again we obtain a probability of 1/5. That is, with each additional head, the probability that we have the fair coin decreases. This makes sense.

The area diagrams in Figure 4 help confirm that the probability is 1/5, and also, surreptitiously, let us introduce some vocabulary. Before the first flip, back in Figure 3, we naturally split the diagram in half. We call this 1/2 a *prior* probability: the probability we assign to a hypothesis *before* we acquire additional information. Here in Figure 4, after the first flip, we believe the probability of 'fair' is now 1/3, so we have a new prior, 1/3, and that's what we use in this diagram. (By the way: prior to what? Prior to the second flip.)

The diagram shows that if we obtain heads again, the new, updated probability that the coin is fair will be 1/5. Before we flip, 1/5 is called a *posterior* probability. It will become the prior if we decide to flip a third time.

Note that the Law of Large Numbers still holds; if we performed this experiment many times and recorded all the times we obtain two heads, we would in fact have the fair coin in about one-fifth of those trials.

What about the Bayes formula?

We have intentionally avoided the formula so far, but students who go further with Bayes will eventually need it. There are any number of ways to develop Bayes' rule. This two-coins situation is a new opportunity: students can see the formula

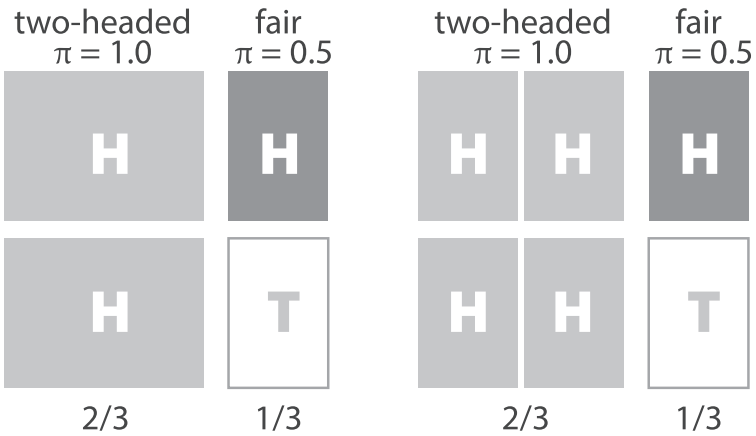

Fig. 4. Two alternative, equivalent diagrams for the second flip, starting after a single "heads" with unequal "prior" probabilities

as an abstraction of the way they have been counting up boxes.

Using the right side of Figure 4, we found the probability 1/5 by dividing the number of boxes of 'H' in the 'fair' column by the total number of 'H's. That is,

$$P(fair|heads) = \frac{1}{1+4} = \frac{P(heads\,\&\,fair)}{P(heads\,\&\;\;fair) + P(heads\,\&\,\sim fair)}.$$

The denominator is simply *P(heads)*, which is 5/6. Looking at the diagram, we can express the numerator as the prior probability of obtaining a fair coin (i.e. the width of the 'fair' column, which is 1/3) times the 'likelihood'—the probability that you obtain another head given that the coin is fair (which is 1/2). This gives us

$$P(fair|heads) = \frac{P(fair)P(heads|fair)}{P(heads)} = \frac{(1/3)(1/2)}{(5/6)} = \frac{1}{5}$$

This is Bayes' rule in context. It is worth noting how much students accomplish from this formulation. They see how different approaches give the same result; and they see how the process of counting rectangles and making ratios can turn into a formula.

Students can also generalize, using the terms we have used above. The possibilities for the unknown coin are hypotheses (*H*), and the observed results the flips are evidence (*E*). Using these terms, Bayes' formula is

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}.$$

Diagnosis example

Let's do another example, in a diagnostic setting. Whether you're diagnosing a patient or a car or a piece of software, you're presented with symptoms (the evidence) but need to figure out the root causes (the hypotheses). Here is an example:

Suppose that in our school, one-tenth of the people have colds. Among the people who have colds, four out of five sneeze. Among those who don't have colds, only one out of five sneezes. (Having a cold or not are the hypotheses. Sneezing is evidence.)

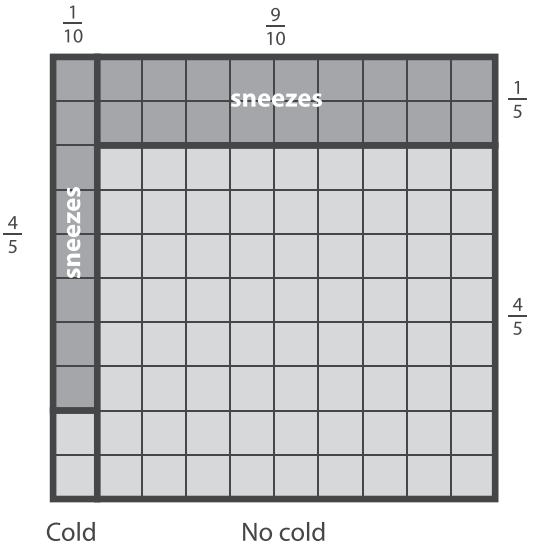

Fig. 5. Area diagram for colds and sneezing. One box is 1%

Someone sneezes. What's the chance that they have a cold? Figure 5 is an area diagram for this situation.

Using the diagram, we pay attention to the darker, 'sneezes' regions and ignore the rest. What proportion of the sneezers have colds? Count the boxes (or compute areas) to find the answer: 8/26 or 4/13. We could also use the formula:

$$P(cold|sneeze) = \frac{P(cold)P(sneeze|cold)}{P(sneeze)} = \frac{(1/10)(4/5)}{(26/100)} = \frac{4}{13}$$

Let's make several observations:

- This illustrates how the Bayesian approach can be useful, showing how much a probability can change when the condition or evidence changes.
- As we would expect, the probability of having a cold increases (from 0.10 to about 0.31) given the evidence of a sneeze.
- Note that despite that increase, even though sneezes are characteristic of colds—i.e. most people with colds do sneeze—less than half of the people who sneeze have colds. We can use this to emphasize the vital asymmetry of conditional probability, such as in this situation or in the prosecutor's fallacy, where *P*(*innocence*l *evidence*) ≠ *P*(*evidence*l*innocence*).
- Note that the above analysis starts with assumptions: one-tenth of the people have colds, four out of five cold sufferers sneeze and one-fifth of the rest also sneeze. These could be based on

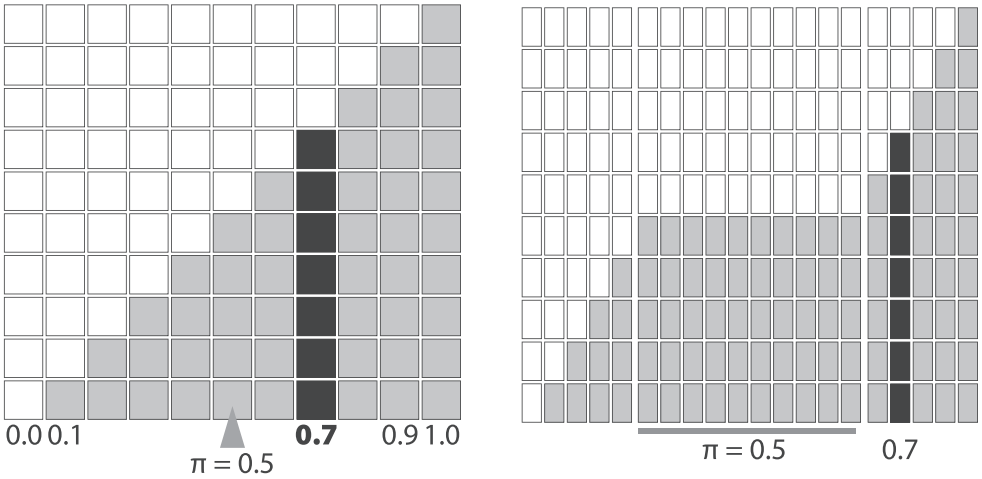

Fig. 6. More elaborate area diagrams, highlighting $\pi = 0.7$

current observations, earlier data or informed guesses. The one-tenth is a prior probability: the probability that someone has a cold *in the absence of any additional data* (e.g. sneezing). We start with this estimate and then obtain the probability that a person who sneezes has a cold based on this prior—and the evidence.

- For instructional purposes in this scenario, we picked 'nice' ratios so that students could make the diagrams by hand. Because the denominators are 2, 5 or 10, students can use a 10-by-10 grid and easily obtain percentages.
- In this example, we saw that hypotheses do not have to be equally likely. Suppose we flipped what we assumed was an ordinary coin. We would assign a very high probability to 'fair'. But what probability? 0.999? 0.99999? Our subjective choice, our *belief*, would affect the subsequent calculations.

Going further

We can extend the same strategies—using area diagrams where we can count boxes—to more complex situations. For example, Figure 6 shows diagrams you might make if you had eleven hypotheses for a mystery coin, where π, the probability of heads, is $\pi = \{0, 0.1, 0.2, \ldots, 0.9, 1.0\}$. Each column becomes divided up according to the probability of heads or tails in that hypothesis. In our diagrams, 'heads' is shaded grey or black.

In the left-hand diagram, each column represents a hypothesis; here, each hypothesis is equally probable, so each column has the same width: 1/11. If it's actually a coin, though, we might use past experience or our personal preferences to give more weight to the possibility that $\pi = 0.5$ and draw the right-hand diagram. In Bayes-speak, we create a non-uniform prior.

We flip our mystery coin once and obtain heads. Using the left side of Figure 6, we can calculate that the new (posterior) probability for $\pi = 0.7$ has increased from 1/11 (about 0.09) to 7/55 (about 0.13). Other hypotheses similarly change their probabilities; note that with one head, the new probability that $\pi = 0$ is … zero.

In the diagram on the right in Figure 6, since we have 20 columns, the priors are all 0.05 except for $\pi = 0.5$, which occupies 10 columns, giving a probability of 0.5. After one heads, the probability for $\pi = 0.7$ increases from 0.05 to 0.07 (7 out of 100 shaded rectangles). Of course, diagrams like these foreshadow the idea that the posterior is a function representing belief, the use of continuous probability distributions and the transition to calculus.

To go even further, note that in this basic introduction, we have only considered flipping heads. What would these diagrams look like if you flipped the coin ten times and got seven heads and three tails? Now you have to think more deeply about probabilities and combinations.

CONCLUSION

We've proposed activities and approaches to give students some relevant background for Bayesian reasoning.

Our use of area diagrams extends the idea of natural frequencies, and the lesson helps us emphasize the update in probability (or belief) rather than focusing exclusively on the updated or 'posterior' probability. Other features of this situation let the student grapple with ideas of

different representations and slippery concepts such as subjective assignment of probabilities.

ACKNOWLEDGEMENT

The author is grateful for excellent guidance and numerous suggestions from an anonymous reviewer and from the editors.

References

de Finetti, B. (1964). Foresight: its logical laws, its subjective sources. In: Kyburg, H.E. and Smokler, H.E. (ed.) *Studies in Subjective Probability*, New York: Wiley.

FiveThirtyEight. (2016). 2016 election forecast: Who will win the presidency? http://projects.fivethirtyeight.com/2016-election-forecast/

Friendly, M. (2001). A brief history of the mosaic display. http://www.datavis.ca/papers/moshist.pdf.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, **102**(4), 684–704.

Martignon, L. and Krauss, S. (2009). Hands on activities with fourth-graders: A tool box of heuristics for decision making and reckoning with risk. *International Electronic Journal for Mathematics Education*, **4**, 117–148.