Probability and Inference

1 Introduction

In this section, we will be dealing with an operator (probability) applied to collections of events (sets). In terms of the logical steps dealing with sets, we are implicitly going back to Socrates and Aristotle. But we will be using the notation of George Boole (1815 - 1864). The probabilistic arguments really goes back into the realm of folklore, for human beings have dealt with probabilities, crudely or well, for millenia.

What exactly does the television weather forecaster mean when stating,"There is a 20% probability that it will rain today"? Hopefully, she is making the following sort of analysis, "Based on all the information I have at this time, around 20% of the time that conditions are as they are today, it will rain." But, it could be argued, surely the conditions are in place which will cause rain or not. The earth is pretty much a closed system. Consequently, the weatherwoman should be able to say of the next 24 hours, "It will rain," or "It will not rain."

Suppose we wished to measure the amount of water which will be required to fill a one liter bottle. The answer is "one liter." Here, there is no reason to invoke probabilities. But, it could be argued, the weather is determined by all the factors which go to make the weather behave as it does. So, if we have the right model, then there is no reason to have probabilities in the weatherwoman's statement as to whether it will rain or not within the next day.

But that is precisely the problem. We do not have a very good model for predicting the weather. We have good models for many physical processes, e.g., for example, force really is equal

to mass times acceleration to a very close approximation (unless the object is moving close to the speed of light). But we do not have good models for many processes. For example, we do not know how to predict very well whether a stock (Blahblahbiotech) will be, say, 10% higher one year from today.

But even with the stock market, there are useful models which will enable us to say things like, "There is a 50% chance that, a year from today, BBBT will have increased at least 10% in value." These models are by no means of Newtonian reliability, but they are getting better. Weather models are getting better too. Today, unlike 20 years ago, we have a much better chance of predicting how many hurricanes will hit the shoreline of the United States.

2 What Is Probability?

But what does it mean, this "20%" probability for rain within the next 24 hours? As we have noted above, it could well mean that conditions are such that, with the crude model available to the weather forecaster, she can say that it will rain in 20% of days with such conditions. The weather forecaster is making a statement about the reliability of her model in the present circumstances. Of course, whether it will rain or not is a physical certainty based on the conditions. But she has no model which will capture this certainty.

Probability, then, can be taken to be a measure of the quality of our models, of our knowledge and ignorance in a given situation. We can, in toy examples, make statements rather precisely. For example, if I toss a coin, it can be said to have a 50% probability of coming up heads. Again, however, if we really knew the precise conditions concerning the tossing of the coin, we could bring the result to a near certainty.

Over many years, a frequency interpretation of probability has become rather common, as in the meaning of the statement of the weather forecaster: "If the conditions of today were duplicated many times in 20% of the cases, it would rain."

There are problems with such an interpretation. Suppose we ask the question whether the People's Republic of China will

launch a nuclear attack against the United States within the next five years? It is hard to consider such a unique event from a frequency standpoint. We could, of course, postulate a fiction of 10,000 parallel worlds just like this one. A 1% chance of the nuclear attack would imply a nuclear attack in roughly 100 of the 10,000 parallel worlds. Such a fictional construction may be useful, but it is fictional nonetheless.

Another interpretation of probability can be made from the standpoint of betting. Let $\mathbf{A}$ be the event that a PRC nuclear attack is launched against the USA during the next five years. Let $\mathbf{A^c}$, that is $\mathbf{A}$ *complement*, be the event that $\mathbf{A}$ does not happen. Suppose a Swiss gambler is willing to put up 100 Swiss francs in order to receive 10,000 Swiss francs in the event the PRC launches the attack (and nothing if there is no attack). Then, it could be argued that $P(\mathbf{A})$ might be computed via

$$100 = P(\mathbf{A}) \times 10,000 + P(\mathbf{A^c}) \times 0. \tag{1}$$

A "fair game" would say that the *expected value* of the bet (right hand side of (1)) should be equal to the amount one pays to play it. By such a rule, $P(\mathbf{A})$ is equal to .01, for

$$100 = .01 \times 10,000 + .99 \times 0. \tag{2}$$

The fact is that we can develop many definitions of what we mean by *probability*. However, there are certain common agreements concerning probability. For example, consider an election where there are n candidates. Suppose the election gives the victory to the candidate with the most votes (plurality rule). Then $P(A_i)$ is the probability that the ith candidate wins the election. One of the candidates must win the election. The probability that two candidates can both win is zero. Therefore.

1. $0 \leq P(\mathbf{A_i}) \leq 1$.
2. $P(\mathbf{A_1}+\mathbf{A_2}+\ldots+\mathbf{A_n}) = P(\mathbf{A_1})+P(\mathbf{A_2})+\ldots+P(\mathbf{A_n}) = 1$.
3. For any collection of the candidates, say, i, j, ..., $P(\mathbf{A_i} + \mathbf{A_j} + \ldots) = P(\mathbf{A_i}) + P(\mathbf{A_j}) + \ldots$.

3 The 2000 Election

In the year 2000, of the four leading candidates for the presidency of the United States had a residential address in a southern state (Buchanan, Bush and Gore). Two of the candidates (Gore and Nader) were liberals. (Let us suppose, that other third party candidates are dismissed as persons whose election in 2000 were beyond the realm of possibility.) Let us suppose that we would like to find the probability of the winner being a southern liberal.

Table 1.Candidate Probabilities

Candidate	Southerner	Liberal	Prob. of Election
Buchanan	Yes	No	.0001
Bush	Yes	No	.5300
Gore	Yes	Yes	.4689
Nader	No	Yes	.0010

Table 1 is both a probability table and a "truth table," i.e., it shows the probability of each candidate winning, and it also shows by "yes" and "no" answers (T and F) whether a candidate has the property of being a southerner and/or a liberal. In order to compute the probability of an *event*, we need to decompose the event into *primitive events*, i.e., events which cannot be decomposed further. In this case, we note that the event of a southern liberal is satisfied only by Gore. Thus, we have

$$P(\text{Southerner} \cap \text{Liberal}) = P(\text{Gore}) = .4689 \tag{3}$$

On the other hand, suppose that we seek the probability that a southern nonliberal wins. The set of southern nonliberals winnings includes two primitive events: Buchanan wins and Bush wins. Then

$$\begin{aligned} P(\text{Southerner} \cap \text{NonLiberal}) &= P(\text{Buchanan + Bush}) \\ &= P(\text{Buchanan}) + P(\text{Bush}) \\ &= .0001 + .5300 \\ &= .5301 \end{aligned} \tag{4}$$

During the campaign, one poll showed

Table 2.Percentage Support in Poll	
Candidate	Percentage in Poll
Buchanan	1%
Bush	48%
Gore	45%
Nader	6%

Based on this poll, can we compute the probability that a particular candidate will win? Some might (wrongly) look at the poll and suppose that Nader's chance of winning is 6%: after all, he seems to have 6% of the electorate behind him. In fact, given the profile in Table 2, the chance of Nader winning is essentially zero. A candidate wins the US presidential election by garnering a majority of the electoral vote. A candidate wins the electoral votes (obtained by summing the number of congresspersons plus two) Generally speaking, a profile like that in Table 2 will guarantee that only Bush and Gore have a chance of winning. Nader and Buchanan will probably gain a plurality in not one single state. If Nader stays in the race, then Bush probably wins, since his votes are largely Democratic.If Nader drops out, then Gore probably wins. And of course, we are looking at a poll taken some time before the election. It would be a stretch to make a guess about the probability that Gore will win the election. And, again, we note that we are going to have a hard time making a rigid frequency interpretation about this probability. There is only one US presidential election in 2000. On the other hand, there were presidential elections which have had similarities to the situation in 2000. And there were elections for governors and congressmen which are relevant, and the opinions of experts who study elections. To make a rigorous logical statement about the probability Gore will win is extremely difficult. But practicality will require that people attempt to make such statements. Part of living in a modern society requires a practical and instinctive grasp of probability. An acquisition of such a practical understanding is part of the motivation for this course.

4 Conditional Probability

Now, let us examine, briefly the *Venn diagram* in Figure 1. John Venn (1834-1923) gave us the Venn diagram as a means of graphical visualization of logical statements. In the above example, there are four primitive events all of which are described by whether **A** is true or false and whether **B** is true or false. (The nonshaded area is not**B** or **B** complement: $\mathbf{B^c}$). More generally, if we have a number of events, say 5 of them, then the number of primitive events will be $2 \times 2 \times 2 \times 2 \times 2 = 32$—everything is characterized by whether an event happens or does not happen. So, then, in the full generality, for n events, there are 2^n primitive events.

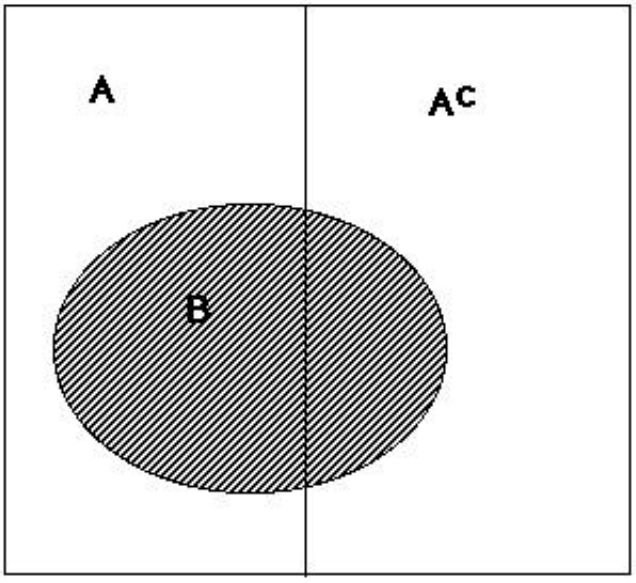

Figure 1.Venn Diagram.

Another matter we will need to investigate is that of causality. We might ask the question: Will there be a big turnout this afternoon at an outdoor political rally. Let us denote this event as **A**. The turnout of the rally is probably dependent on the weather this afternoon. Suppose, then, we consider another event: it starts to rain by one hour before the rally. Call this

event **B**. Now, we can say that

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{B})P(\mathbf{A}|\mathbf{B}). \tag{5}$$

Reading (5) partly in English and partly in mathematical symbols, we are saying:

> The Probability **A** happens and **B** happens **equals** the Probability **B** happens multiplied by the Probability **A** happens given that **B** happens.

Recalling what **A** and **B** represent, we are saying (completely in English this time):

> The probability there is a big turnout at the rally and that it rains is equal to the probability it rains multiplied by the probability there is a big turnout if it rains.

Stated in this way, it is clear that we are implying that rain can have an effect on the rally. If there were no effect of rain on the rally, then (5) would become simply

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{B})P(\mathbf{A}). \tag{6}$$

In such a case, we would say that **A** and **B** are *stochastically independent* and that the *conditional probability* $P(\mathbf{A}|\mathbf{B})$ is simply equal to the *marginal probability* $P(\mathbf{A})$.

Returning to the situation where **A** and **B** are not independent of each other let us note that the following is logically true:

$$P(\mathbf{B} \cap \mathbf{A}) = P(\mathbf{A})P(\mathbf{B}|\mathbf{A}). \tag{7}$$

In English,

> The probability there is a big turnout at the rally and that it rains is equal to the probability there is a big turnout at the rally multiplied by the probability that it rains if there is a big turnout.

This might sound as though we were talking about the turnout having an effect on the climate. Really, we are not. We are talking here about *concurrence* rather than *causation*. If there

is a natural *causal* event and a natural *effect* event, then the kind of statements in (5) and (7) speak of concurrence, which may be causation but need not be. The weather may well effect the turnout at the rally (5) (concurrency and causation), but the turnout at the rally does not affect the weather (concurrency only) (7). All these matters can be written rather simply but require some contemplation times before it becomes clear.

Returning to our original problem, we might write the probability of a big turnout at the rally as:

$$P(\mathbf{A}) = P(\mathbf{B})P(\mathbf{A}|\mathbf{B}) + P(\mathbf{B^c})P(\mathbf{A}|\mathbf{B^c}). \tag{8}$$

Reading (8) in English, we are saying

> The Probability **A** happens **equals** The Probability **B** happens times the Probability **A** happens given that **B** happens **plus** The Probability **B** does not happen times the Probability **A** happens given that **B** does not happen.

An experienced political advisor gives us the information that probability of a big turnout in the case of rain is 40%, but the probability of a big turnout in the case of no rain is 90%. Making a good guess about $P(\mathbf{A})$ is important, for the public relations team can then know whether to prepare to get the media to cover the event or not. The weather consultant informs the probability of rain in the afternoon is 20 %. We wish to compute our best estimate as to whether the turnout will be big or not.

$$P(\mathbf{A}) = .20 \times .40 + .80 \times .90 = .08 + .72 = .80. \tag{9}$$

It would appear that the public relations team might well try and get the media to come to the rally.

5 Bayes' Theorem

Now, it is reasonable to suppose that rain can have an effect on the size of an outdoor rally. It is less clear that the size of the outdoor rally can have an effect on the weather. However, let us suppose that some years have passed since the rally. Reading in a pile of newspaper clippings, a political scientist reads that the

turn-out at the rally was large. Nothing is found in the article about what the weather was. Can we make an educated guess as to whether it rained or not? We can try and do this using the equation

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{B})P(\mathbf{A}|\mathbf{B})}{P(\mathbf{B})P(\mathbf{A}|\mathbf{B}) + P(\mathbf{B^c})P(\mathbf{A}|\mathbf{B^c})} \tag{10}$$

This equation is not particularly hard to write down: no fancy mathematical machinery is necessary. But it is one of the most important equations in science. Interestingly, it was not discovered by Newton or Descartes or Pascal or Gauss, great mathematicians all. Rather it was discovered by an 18th century Presbyterian pastor, Thomas Bayes (1702-1761). The result bears his name **Bayes' Theorem**. Let's prove it.

First of all, returning to our Venn diagram in Figure 1, we note that it either rains or it does not, that is to say,

$$\mathbf{B} + \mathbf{B^c} = \mathbf{\Omega}. \tag{11}$$

$\mathbf{\Omega}$ is the *universal set*, the set of all possibilities. Clearly, then

$$P(\mathbf{B}) + P(\mathbf{B^c}) = P(\mathbf{\Omega}) = 1. \tag{12}$$

Rewriting (7), we have

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{B} \cap \mathbf{A})}{P(\mathbf{A})}. \tag{13}$$

Next, we substitute the equivalent of $\mathbf{B} \cap \mathbf{A}$ from (5) into (13) to give:

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{B})P(\mathbf{A}|\mathbf{B})}{P(\mathbf{A})}. \tag{14}$$

Returning to the Venn diagram, we note that

$$\begin{aligned} \mathbf{A} = \mathbf{A} \cap \mathbf{\Omega} &= \mathbf{A} \cap (\mathbf{B} + \mathbf{B^c}) \\ &= \mathbf{A} \cap \mathbf{B} + \mathbf{A} \cap \mathbf{B^c}. \end{aligned} \tag{15}$$

Combining (5) with (15), we have

$$P(\mathbf{A}) = P(\mathbf{B})P(\mathbf{A}|\mathbf{B}) + P(\mathbf{B^c})P(\mathbf{A}|\mathbf{B^c}). \tag{16}$$

Substituting (16) into the denominator of (14), we have **Bayes' Theorem**

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{B})P(\mathbf{A}|\mathbf{B})}{P(\mathbf{B})P(\mathbf{A}|\mathbf{B}) + P(\mathbf{B^c})P(\mathbf{A}|\mathbf{B^c})} \tag{17}$$

Let us see how we might use it to answer the question about the probability it rained on the day of the big rally. Substituting for what we know, and leaving symbols in for what we do not, we have

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{B})(.40)}{P(\mathbf{B})(.40) + P(\mathbf{B^c})(.90)} \ . \tag{18}$$

And here we see a paradox. In order to compute the *posterior* probability that it was raining on the afternoon of the big rally, given that there was a big rally that afternoon, we need to have a *prior* guess as to $P(\mathbf{B})$, that is to say a guess prior to our reading in the old clipping that there had been a big afternoon rally that day. The philosophical implications are substantial. Bayes wants us to have a prior guess as to what the probability of rain on the afternoon in question was *in the absence of the data at hand, namely that there was a big rally.* This result was so troubling to Bayes that he never published his results. They were published for him and in his (Bayes') name by a friend after his death (a nice friend, many results get stolen from living people, not to mention dead ones).Why was Bayes troubled by his theorem? Bayes lived during the "Enlightenment." After Descartes it was assumed that everything could be reasoned out "starting from zero." But Bayes' Theorem does not start from zero. It starts with guesses for the probability which might simply be the prejudices of the person using his formula. Bayes' Theorem, then, was politically incorrect by the standards of his time. Then, as now, political correctness is highly damaging to human progress.

To use Bayes' Theorem, we need to be able to estimate $P(\mathbf{B})$. Of course, if we know $P(\mathbf{B})$, then we know $P(\mathbf{B^c}) = 1$ - $P(\mathbf{B})$. What to do? Well, suppose the rally was in October. We look in an almanac and find that for the area where the rally was held, it rained in 15% of the days. (We are actually looking for rain in the afternoon, but almanacs are usually not that detailed.)

Then, we have from (18)

$$P(\mathbf{B}|\mathbf{A}) = \frac{.15 \times .40}{.15 \times .40 + .85 \times .90} = .0727. \tag{19}$$

(19) is the way to proceed when we have a reasonable estimate of the *prior probability* $P(\mathbf{B})$. And, in very many cases, we do have an estimate of $P(\mathbf{B})$.

What troubled Bayes was the very idea that in order to use a piece of data, such as the knowledge that on the day in question there was a big rally, if we are to estimate the probability that it rained on that day, then we need to have knowledge of the probability of rain *prior to the use of the data*. By Enlightenment standards, one should be able to "start from zero." A prior assumption such as $P(\mathbf{B})$ was sort of like bias or prejudice.

Bayes came up with a politically correct way out of the dilemma, but the fix always troubled him. The fix is referred to as **Bayes' Axiom**:

> In the absence of prior information concerning $P(\mathbf{B})$, assume that $P(\mathbf{B})$=$P(\mathbf{B^c})$.

When we take this step in (18), then notice that $P(\mathbf{B})$ and $P(\mathbf{B^c})$ cancel from numerator and denominator, and we are left simply with:

$$\begin{aligned} P(\mathbf{B}|\mathbf{A}) &= \frac{P(\mathbf{B})P(\mathbf{A}|\mathbf{B})}{P(\mathbf{B})P(\mathbf{A}|\mathbf{B}) + P(\mathbf{B^c})P(\mathbf{A}|\mathbf{B^c})} \\ &= \frac{P(\mathbf{A}|\mathbf{B})}{P(\mathbf{A}|\mathbf{B}) + P(\mathbf{A}|\mathbf{B^c})} \\ &= \frac{.40}{.40 + .90} \\ &= .3077. \end{aligned} \tag{20}$$

The difference in the answers in (19) and (20) is substantial. To us, today, it would appear that (19) is the way to go, and that the assumption of prior ignorance is not a good one to make unless we must. But, as a matter of fact, many of the statistical computations in use today are based on (20) rather on than (19). And, as it turns out, as we gain more and more data, our guesses as to the prior probabilities become less and less important.

Let us now consider a more practical use of Bayes' Theorem. Suppose a test is being given for a disease at a medical center. Historically, 5% of the patients tested for the disease at the center actually have the disease. In 1% of the cases when the patient has the disease, the test (incorrectly) gives the answer that the patient does not have the disease. Such an error is called a *false negative*. In 6% of the cases when the patient does not have the disease, the test (incorrectly) gives the answer that the patient does have the disease. Such an error is called a *false positive*. Let us suppose that the patient tests positive for the disease. What is the posterior probability that the patient has the disease? We will use the notation $\mathbf{D}^+$ to indicate that the patient has the disease, $\mathbf{D}^-$ that the patient does not have the disease. $\mathbf{T}^+$ indicates the test is positive, $\mathbf{T}^-$ indicates the test is negative. Then we have

$$\begin{aligned} P(\mathbf{D}^+|\mathbf{T}^+) &= \frac{P(\mathbf{T}^+|\mathbf{D}^+)P(\mathbf{D}^+)}{P(\mathbf{T}^+|\mathbf{D}^+)P(\mathbf{D}^+) + P(\mathbf{T}^+|\mathbf{D}^-)P(\mathbf{D}^-)} \\ &= \frac{.99 \times .05}{.99 \times .05 + .06 \times .95} \\ &= .4648. \end{aligned} \tag{21}$$

Suppose that a physician finds that a patient has tested positive. It is still very likely—80%—that the patient does not have the disease. So, it is likely the physician will tell the patient that it is quite likely the disease is not present, but to be on the safe side, the test should be repeated. Suppose this is done and the test is again positive. The new prior probability of the disease being present is now $P(\mathbf{D}^+) = .4648$. So, we have

$$\begin{aligned} P(\mathbf{D}^+|\mathbf{T}^+) &= \frac{P(\mathbf{T}^+|\mathbf{D}^+)P(\mathbf{D}^+)}{P(\mathbf{T}^+|\mathbf{D}^+)P(\mathbf{D}^+) + P(\mathbf{T}^+|\mathbf{D}^-)P(\mathbf{D}^-)} \\ &= \frac{.99 \times .4648}{.99 \times .4648 + .06 \times .5352} \\ &= .9348. \end{aligned} \tag{22}$$

At this point, the physician advises the patient to enter the hospital for more detailed testing and treatment.

6 Multistate Version of Bayes' Theorem

Typically, there will be more than two possible states. Let us suppose there are n states, $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_n}$. Let us suppose that $\mathbf{H}$ is some piece of information (data). Consider the Venn diagram in Figure 2.

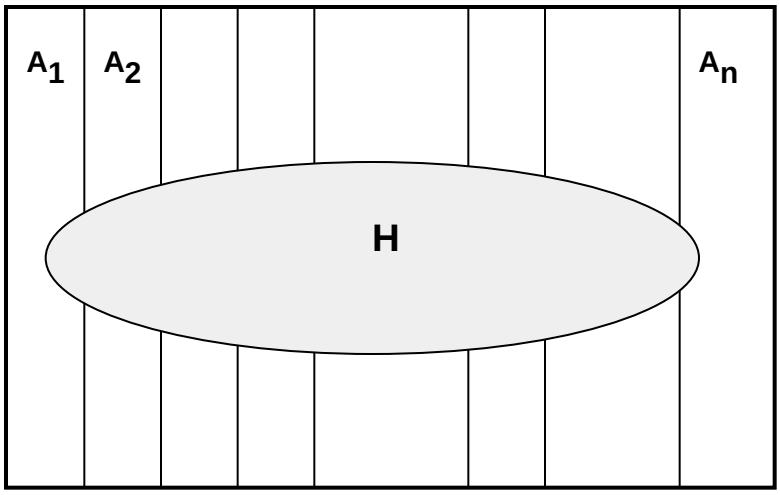

Figure 2.Inferential Venn Diagram.

Then it is clear (really) that (17) becomes

$$P(\mathbf{A_1}|\mathbf{H}) = \frac{P(\mathbf{A_1})P(\mathbf{H}|\mathbf{A_1})}{\sum_{j=1}^{n} P(\mathbf{A_j})P(\mathbf{H}|\mathbf{A_j})} \tag{23}$$

As an example, consider the case where a mutual fund manager is deciding where to move 5% of the fund's assets (currently in bonds). She wishes to decide whether it is better to invest in chips $\mathbf{A_1}$, Dow listed large cap companies $\mathbf{A_2}$, or utilities $\mathbf{A_3}$. She is inclined to add the chip sector to the fund, but is not certain. Her prior feelings are that she is twice as likely to be better invested in chips than in Dow stocks, and three times as likely to be better invested in chips than in utilities. This gives her prior probabilities

$$P(\mathbf{A_1}) + \frac{1}{2}P(\mathbf{A_1}) + \frac{1}{3}P(\mathbf{A_1}) = 1. \tag{24}$$

Solving, we have, for the prior probabilities,

$$\begin{aligned} P(\mathbf{A_1}) &= \frac{6}{11} \\ P(\mathbf{A_2}) &= \frac{3}{11} \\ P(\mathbf{A_3}) &= \frac{2}{11} \ . \end{aligned}$$

She receives information **H** that the prime interest rate is likely to rise by one half percent during the next quarter.

She feels that

$$\begin{aligned} P(\mathbf{H}|\mathbf{A_1}) &= .1 \\ P(\mathbf{H}|\mathbf{A_2}) &= .4 \\ P(\mathbf{H}|\mathbf{A_3}) &= .5 \end{aligned}$$

How, then, should she revise her estimates about the desirability of the various investments, given the new information on interest rate hikes? From (23), we have

$$\begin{aligned} P(\mathbf{A_1}|\mathbf{H}) &= \frac{6/11 \times .10}{6/11 \times .10 + 3/11 \times .40 + 2/11 \times .50} = .21(25) \\ P(\mathbf{A_2}|\mathbf{H}) &= \frac{3/11 \times .40}{6/11 \times .10 + 3/11 \times .40 + 2/11 \times .50} = .42(86) \\ P(\mathbf{A_3}|\mathbf{H}) &= \frac{2/11 \times .50}{6/11 \times .10 + 3/11 \times .40 + 2/11 \times .50} = .35(71) \end{aligned} \tag{28}$$

Based on the new information, she reluctantly abandons her prior notion about investing in chips. The increase in interest rates will have a chilling effect on their R&D. She decides to go with a selection of large cap Dow stocks.