IEOR 4701: Stochastic Models in Financial Engineering

Summer 2007, Professor Whitt

Class Lecture Notes: Tuesday, July 10.

Monte Hall and Statistics

Now that you are experts on conditional probability, you are ready to consider the following problem:

1. Suppose you are on a **game show**, and you are given the choice of three doors. Behind one door is a brand new (desirable) car; behind the other two, **goats**. You pick a door, say door number 1, and the host, who knows what is behind all the doors, opens another door, say door number 3, which has a goat. He says to you, "Do you want to change your pick to door number 2?"

Is it to your advantage to switch your choice of doors? Why?

This is the famous **Monte Hall Game Show Problem**, from the show, **Let's Make a Deal!** Like many of our problems, this too easily gives a headache upon first thought.

This problem is included for a reason. The reason is that **non-mathematical issues are crucial here**. The assumptions you make are critical. In particular, what are the host's motives? Will the host always offer the contestant a choice? Or will the host, trying to trick the contestant, only offer the contestant a choice when the contestant has picked the car (i.e., the door with the car behind it)? Or, on the other hand, will the host, trying to help the contestant, only offer the contestant a choice when the contestant has picked a goat (i.e., a door with a goat behind it)?

It is natural that we assume that the host will always open a door, and focus on the clean mathematics problem, but it is important to pay attention to assumptions in problems. You do not want to make assumptions too quickly. Or when they are not justified. In this problem, I believe the assumptions are the most important part.

Given that we assume that the host will *always* open one of the doors, and that he knows what is behind each door, so that he always shows you a door with a goat behind it, then it can be shown that it is indeed to your advantage to switch. If you do not switch, then you win in one of three cases. If you do switch, then you win in two of three cases.

But how to convince you? Here is a try: You can consider the strategy that says *decide in advance to always switch*. It is easy to see that, if we take that strategy as given, we win in 2 cases out of 3, whereas if we decide not to switch, we win in 1 out of 3. Our chances do not improve to 1 out of 2, because the host always picks a door with a goat. We do not simply reduce the choices from 3 to 2.

For more on this fascinating problem, Google the Monte Hall Problem. You will find simulators to check this out. Also you might Google Marylin vos Savant; she wrote about it. My feeling is that there is not enough discussion about the importance of basic assumptions concerning the host's motives.

2. A Statistics Example

Suppose that we are trying to **identify whether a particular author wrote a manuscript with unknown author**. The known author uses the word **"moreover,"** $2,000$ times out of $400,000$ words, which is 1 out of 200. The word "moreover" is used 2 times in the new text of $4,000$ words. If we selected 4000 words at random from this known author, then the expected number of times "moreover" would appear is $4000 \times (1/200) = 20$, so 2 is a lot less than we would expect. The question is: Is the observed value 2 so much less than the expected value 20 that it could not reasonably have come from the author in question?

To answer the question, we **make a probability model** (somewhat less realistic than the model for the coin tossing). We assume that successive words are randomly selected from all possible words with certain probabilities that depend on the author. We ask what is the probability that there would be 2 or fewer occurrences of the word "moreover" in a text of 4000 words by the known author under the assumption that the successive words are independent and identically distributed (IID) with probability

$$p = \frac{2000}{400,000} = \frac{1}{200} = 0.005$$

for the word "moreover"?

With the probability model we have chosen, the number of occurrence of the word "moreover" in the text of $4,000$ words would have a binomial distribution. Let N denote the number of occurrences of the word "moreover" in a sample of 4000 words. The exact probability of 2 or fewer occurrences would be

$$P(N \leq 2) = b(0; n, p) + b(1; n, p) + b(2; n, p) \ ,$$

where $b(k; n, p)$ denotes the binomial probability of outcome k with parameters n and p, i.e.,

$$b(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \ .$$

Here $n = 4000$ and $p = 0.005$. If you need to, look this up in the book for further explanation and background. This is something good to review.

This calculation is messy. It is significant that we can use approximations that greatly simplify the calculation. In this case, where n is big and p is small, it is appropriate to use a Poisson approximation of a binomial distribution. (See pages 32-33 of the text for the justification.) The Poisson parameter λ (which is the mean and the variance) is obtained by matching the mean, i.e.,

$$\lambda = np \ .$$

Thus, with that approximation

$$b(k; n, p) \approx Poisson(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \ .$$

With the Poisson approximation, we see that the probability is very small. Indeed,

$$P(N \leq 2) \approx P(Pois(\lambda) \leq 2) = e^{-20}(1 + 20 + 200) \ ,$$

which is very very small. ($e^{-20} = 2.06 \times 10^{-9}$) Be sure you understand this!

Since the mean of the Poisson distribution is not too small (greater than or equal to 5, say), we can also use a normal approximation. The variance of a Poisson distribution is equal to its mean. Thus the random variable N has mean 20 and approximately variance 20 too. (The variance of a binomial distribution with parameters n and p is exactly $np(1-p)$, which is approximately np when p is small.) From the normal table on p. 81 of Ross, we see that with probability about 0.95 a normal random variable is within two standard deviations of its mean. With probability about 0.998, it is within 3.1 standard deviations of its mean. We see that 2 is more than 3 standard deviations from the mean of 20. Specifically, a standard deviation is $\sqrt{20} = 4.47$ and 2 is 4.02 standard deviations from the mean. So the observed outcome is unlikely to have occurred by chance (according to our rough probability model). Based on this probability analysis, we conclude that it is unlikely that the unkonwn 4000-word document was written by the author of the $400,000$ words.

That last example is a "short course" on Statistics. We show the "statistical way of thinking."