Exercises 7 · Probability warm-up

Due Monday, March 28, 2016

Problem 1: conditional probability

You have just been appointed to a Very Important Role in one of the state's largest health-insurance companies. Your first assignment is to study the cost-effectiveness of instituting a universal test for a disease called SOS. Your firm is thinking about making the test free and mandatory for all 10 million of its clients in Texas. The tests themselves are a significant expense. Yet if caught early, before the onset of its worst symptoms, SOS can be treated much more cost-effectively. This could potentially save your firm a large amount of money down the road. You are charged with making a policy recommendation to people holding Even More Important Roles in the company.

(A) We know that SOS afflicts roughly 1 Texan out of every 1000, and let's assume that your 10 million clients are a representative sample of all Texans. No medical test is perfect, but this one is reasonably accurate: it gives a positive result for 95% of people who have SOS, and a negative result for 99% of people who do not have SOS. What is the probability that a patient has SOS, given that he or she tests positive for the disease?

(B) Is the proposed policy of free, universal SOS testing a smart financial move for your company? Use the following facts about the test/treatment costs for SOS to help you decide.

- Each test for SOS costs your company \$10.
- Each false positive—that is, where the test gives a positive result for someone who does not actually have the disease—costs your company \$50 in follow-up costs associated with discovering that the initial test was wrong.
- On average, each correctly identified case of SOS saves your company $\$10,000$ in net future medical expenses.[1]

Hint: you may find it useful to recapitulate the steps in Part A for the people who test negative.

(C) In Part B, how much larger or smaller than $\$10,000$ would the savings have to be from a correct identification of SOS in order for the proposed policy to be a "break-even" proposition?

[1] For you accounting majors, assume this is adjusted to net present value.

Problem 2: sequential betting

Suppose you have w_0 dollars to your name, and that you are offered the following bet.

- With probability $p = 0.52$, you will win the bet.
- With probability $1 - p = 0.48$, you will lose the bet.

It sounds like a good bet: you are more likely to win than to lose. Moreover, you get to choose what fraction c of your total wealth w_0 to wager on the outcome of the bet.

(A) Suppose that your current wealth is $w_0 = \$1000$, and you decide to risk $c = 0.10$ (i.e. 10%) of your wealth on this bet. What is your expected final wealth, w_1, after the outcome of the bet becomes known?

(B) Imagine now that you get to repeat the bet as many times as you want. After every single round of betting, you decide to risk the same fraction $c = 0.1$ of your current wealth on the next bet. In other words, if you have w_t dollars after round t of betting, you place $0.1 w_t$ dollars on the next round's wager. If you win, you will wealth will be $w_{t+1} = (1 + c) \cdot w_t$. If you lose, it will be $w_{t+1} = (1 - c) \cdot w_t$.

Suppose you start with \$1000. Simulate 10,000 rounds of this bet.[2] (I recommend that you build on our R scripts from the walk-through on Monte Carlo for sequential events. But you can also make a spreadsheet with 10,000 rows.) Make a plots of your simulated trajectory of wealth w_t over every round from $t = 1$ to $t = 10000$ (the betting round t should be on the x-axis). What happens after 10,000 rounds of betting? Are you rich or broke? (Repeat the simulation a few times if you need to, in order to convince yourself of what happens here.) In light of the answer from Part A, do you find this surprising?

[2] Why 10,000? Because that's roughly the number of trading days over a 40-year period of investment.

(C) Now repeat the simulation—except this time, only risk 0.5% of your current wealth (that is, $c = 0.005$, or 1 part in 200) at every round of betting. As before, plot your simulated value of current wealth w_t at every step from 1 to 10,000. This time, what happens after 10,000 rounds? Are you rich or broke?

(D) Let's focus on the fraction c. If you have total wealth w_t at round t, and bet a fraction c of this wealth on the outcome for round t, what are the expected value and standard deviation of w_{t+1} of your total wealth at step $t + 1$, i.e. after the outcome of the bet has been realized?

(E) Experiment with your Monte Carlo simulation to find a value of c that you like best in order to maximize the long-term growth of your portfolio.

(F) Comment on the wisdom of the following statement: "If your goal is to ensure the long-term growth of your capital, you should make bets that carry the highest possible expected return."

Problem 3: probability in statistical inference

In the question, you will use probability theory to derive one of the most basic (and important) results in all of statistical inference: how does the variability of the sample mean relate to the size of the sample? (Make sure to read Harold Wainer's article on "The Most Dangerous Equation", also linked through the class website.)

Suppose that $X_1, X_2, \ldots, X_N$ are independent samples drawn from some common probability distribution P. The only things you are told about P are its mean μ, and its variance σ^2. That is: $E(X_i) = \mu$ and $\mathrm{var}(X_i) = E\{(X-\mu)^2\} = \sigma^2$ for all samples X_i. We know nothing else about P.

Let $\bar{X}_N$ denote the sample mean of the observed data points $X_1, \ldots, X_N$. This sample mean is obviously an estimator of μ, the underlying true mean of the probability distribution P. As an estimator of μ, $\bar{X}_N$ has a sampling distribution (which you could use bootstrapping to approximate).

Derive an explicit formula for the mean and standard error for $\bar{X}_N$, using results from Chapter 7 of the course packet.[3] Now go back and pick any variable in any data set we've looked at this semester that has at least 30 observations. Compute the sample mean of that variable. Then compare the standard error of the sample mean that you get by applying your formula[4] to the standard error of the sample mean that you get from bootstrapping the sample. Comment.

[3] Pay particular attention to the results on linear combinations of random variables.

[4] You may use the sample variance of that variable to approximate σ^2, the variance of the underlying probability distribution.