Statistics 601, Spring 2018

Assignment 3

In class we discussed the problem of estimating a probability that one variable is less than another independent variable based on simulated values from the distributions of those variables. Our context was estimating the probabilty that a logistic regression for one group had a slope that was greater than that for another group. We arrived at two potential ways to estimate this probability.

The problem can be reduced to one in which we would like to produce a Monte Carlo approximation to

$$Pr(X \leq Y),$$

where X is a continuous random variable with density $p(x);\ -\infty < x < \infty$ and Y is an independent random variable that is continuous with density $p(y);\ -\infty < y < \infty$. Note that $p(\cdot)$ is being used as notation for a generic density function (X and Y do not necessarily have the same density function).

We assume we have samples of values $\{x_m :\ m = 1, \ldots, M\}$ and $\{y_m :\ m = 1, \ldots, M\}$ that have been simulated from distributions with densities $p(x)$ and $p(y)$, respectively. One MC approximation would be

$$P_{1,M} = \frac{1}{M} \sum_{m=1}^{M} I(x_m \leq y_m), \tag{1}$$

where $I(A)$ is the indicator function that assumes a value of 1 if A is true and 0 otherwise. This approximation comes from consideration of the integral

$Pr(X \leq Y) =$

$$E[I(X \leq Y)] = \int_{-\infty}^{\infty} I(x \leq y)\, p(x,y)\, dx\, dy \; = \int_{-\infty}^{\infty} I(x \leq y)\, p(x)\, p(y)\, dx\, dy.$$

We can, however, also write the desired probability as,

$Pr(X \leq Y) =$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{y} p(x,y)\ dx\, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{y} p(x)\, p(y)\ dx\, dy = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{y} p(x)\, dx\right]\, p(y)\, dy.$$

Consideration of this second form led us to the approximation

$$P_{2,M} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{M}\sum_{m=1}^{M} I(x_m \leq y_j). \tag{2}$$

For this assignment, conduct a Monte Carlo investigation of the precision of these two possible approximations to the desired probability. This is not meant to be a horribly long assignment, but keep in mind the following:

1. What you want to approximate in your Monte Carlo exercise are the expected values and variances of (1) and (2). Thus, you will have two Monte Carlo sample sizes to worry about, the samle size denoted as M in expressions (1) and (2) and the Monte Carlo sample size you will use in your assessment of the accuracy and precision of these approximation, lets call it N. To make this exercise a little easier, just use $N = 2000$. You will vary M so lets make these values $M \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000\}$.

2. The easiest situation to set up is one in which you know the answer (think about the probability one independent normal is less than another if both have mean zero).

3. Keep in mind the following question. We generally assess the precision of a Monte Carlo approximation as the sample variance of Monte Carlo terms divided by Monte Carlo sample size M. Here, given that the terms are indicators, we would most likely use the from $P_M(1 - P_M)/M$. Could this be done here, and bypass the need for an "outer" Monte Carlo procedure? If you get confused, convince yourself that this is fine for a single variable. Consider simulating values $x_m;\ m = 1, \ldots, M$ from some distributon, and then computing a Monte Carlo approximation to $Pr(X \leq a)$ for some fixed value a. This approximation will be

$$P_M = \frac{1}{M}\sum_{m=1}^{M} I(x_m \leq a).$$

You might want to try to determine whether or not the value of $P_M(1-P_M)/M$ for any one sample is close to the variance of P_M obtained from repeated samples, each of size M. We might hope that the variance in repeated samples will be an approximation to the expected value of $P_M(1-P_M)/M$). Does this appear to be true for a single variable? Does this appear to be true for our objective of approximating $Pr(X \leq Y)$ for two random variables X and Y?

4. Finally, recall that relative precision is generally expressed as a ratio of variances. With this in mind, is the relative precision of (1) to (2) affected by the Monte Carlo sample size M?