Some thoughts about Simulation and Statistics

Mai Zhou

University of Kentucky

Abstract

Monte Carlo simulation are done more frequently now due to the fast/cheap computers/softwares. We take a look at some typical examples of statistical use of the simulation and this serve as a starting point for Bootstrap and MCMC.

Key Words: Pivatol, Approximations, Rate of Error.

1. Introduction

A big change in the last 10 years in the research environment is: Cheap PC everywhere and many free softwares for download. Because of the rapid increase/available of computing power, it is easier to carry out many computations and therefore (statistical) simulation or Monte Carlo simulation is getting easier and is done more frequently now.

'Simulation' or 'Monte Carlo simulation' are terms frequently heard in statistics literature. (or MCMC etc.) We want to ask the questions:

(1) what is exactly the (statistical/Monte Carlo) simulation method? (or Monte Carlo method etc.)

(2) what can the simulation method do? (and more importantly what are those things that the usual/classic statistics method cannot do or is difficult to do)? What kind of problem can be solved by simulation?

(3) What are some of the statistical problems that cannot be solved by simulation? Any limitations of simulation?.

We do not discuss simulation in the other fields like engineering, numerical analysis, etc. But only focus on the Statistical use of simulation.

(A1) use the computer/software to mimic a real statistical procedure in action.

(A2) see below for some examples.

2. Examples of Simulation in Action

Example 0: Given an iid sample of size n from a normal distribution, what is the distribution of

$$\frac{\bar{X}-\mu}{\sqrt{s^2/n}}$$

We of course know this statistic follows a t distribution with df n-1. But the rigorous proof would need a lot of probability training. Suppose we do not have the training or do not have the time to work out the distribution theoretically, we may use simulation to find the distribution.

For a specific n (say n=10) we generate 10 observations from the normal distribution with mean m and variance σ^2 (these can be any sensible parameter values, due to the fact that the statistic is a *pivatol*). Compute $t_1 = \frac{\bar{X}-m}{\sqrt{s^2/n}}$, and repeat that N times to get $t_1, \cdots, t_N$.

The distribution of those $t_1, \cdots, t_N$ is what we want. To get the CDF, we can compute the empirical distribution based on $t_1, \cdots, t_N$. To get the density, we can compute a kernel density estimator based on $t_1, \cdots, t_N$. (or a histogram)

The dis-advantage is that for each n we have to repeat the simulation, while the theoretical derivation can be carried out for a general n.

Example 1: In basic statistical courses (like Sta291) we usually carry out the testing hypothesis of binomial population

$$H_0 : p = 0.5 \quad vs. \quad H_a : p \neq 0.5$$

with an iid Bernoulli sample of size $n = 15$ by looking at the statistic

$$\frac{\hat{p}-0.5}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

or

$$\frac{\hat{p}-0.5}{\sqrt{0.5(1-0.5)/n}}$$

We can imaging replace all the 0.5 above by some other value like p_0.

The distribution of the test statistic is approximately normal. But this is only for large samples. We usually do not give a formula solution for smaller samples.

The confidence interval based on this also has the 'only good for large n' pre-condition (n being the sample size the $\hat{p}$ is based on).

Same thing can be said for the (testing and confidence interval of) difference of two p's, ratio of two p's etc.

Simulation can give us the exact distribution of the test statistic above for small sample sizes, under H_0 or under a specific H_a (a simple hypothesis).

Once we knew the exact distribution of the test statistic under H_0, we can get a more accurate P-value then the normal approximation, etc.

```
phats <- rbinom(9000, size=15, prob=0.5)/15
zvals <- (phats-0.5)/sqrt(0.25)
plot.ecdf(ecdf(zvals))
```

Of course, for this simple problem there are other solutions that are not simulation based. But it shows what simulation can do.

Example 2. In logistic regression model, we assume a Bernoulli/binomial random variable with success probability (depend on X)

$$p = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)} \ .$$

The estimation and testing of the parameter β is given by SAS and R and Splus, ok, but the P-value and confidence interval is only valid for large sample sizes. What about smaller sample sizes? Or how accurate is the large sample approximation when the sample is relatively small?

Again, simulation can give us the distribution of the test statistics under null hypothesis $H_0 : \beta = 0$. (in fact any simple hypothesis) for small n, so we get EXACT p-value (not just approx p value)

Here I do not know other solutions that are not simulation based, and can give you exact distribution.

Example 3. When estimating the probability of a random variable larger than a given number $P(X > t_0)$. we can use the sample fraction $\hat{p}$. $n \times \hat{p}$ has a binomial distribution. (but with an unknown parameter!)

So, a simulation could be carried out if we pick a particular distribution to begin with. But if asked why we chose this distribution, the answer is for convieniance.

When the data are right censored. the distribution of the estimator, Kaplan-Meier estimator is only approximately normal for LARGE samples. what about smaller samples?

Example 4. Rank test. Wilcoxon and logrank test? with censoring, no table to look up to. Also, get the distribution of the test statistic under null hypothesis.

Example 5. trimmed mean etc.

Example 6 spatial statistic etc.

Example 7 finding expectations of a r.v.

2. The limitations of Simulation Method

1. Generating random variables from the population can be difficult (even whne the population is completely specified), (this is where MCMC comes in), if not impossible (when the population is not completely specified) (this is where Bootstrap comes in to help).

2. Also, there is the question about the quality of the random variables generated .... are they REALLY iid ? (no, but for all practical purposes, yes, for those high quality PRNG). Note R provides several choices of PRNG.

3. The distribution obtained from the simulation is the result of a limiting process (N goes to infinite). Unless you have infinite time to wait, we have to stop at some finite N, which only gives us an *Approx. Limit Distribution* — defeating the whole purpose of getting the EXACT distribution as we claimed before. But with faster computers (and some patience) we are getting closer to the limit in less time. (as some people use $N = 10,000$, or $N = 5,000$, and in some cases $N = 1,000,000$).

Please note there are: n the original sample size and N the repetitions in the simulation process. They are different.

Usually n is somewhat small and fixed in a given problem and N is up to us to pick and thus usually large (since we have fast computer, we use large N in simulation to get close to the limit).

3. The Theoretical Bases for the Simulation

Why simulation works as claimed? Why in the limit the distribution is the EXACT distribution?

Short answer: The Central Limit Theorem.

Simulation is also used to find the expectation of a random variable, the theoretical base for that is: The Law of Large Numbers.

3. Some Techniques in Simulation

(1) try to converge faster to the limit.

(a) Control functions

(b) Importance sampling

(c) Stratified sampling

(d) Dependent sampling.

(2) checking if a sample follows a given distribution.

(a) Q-Q plot graphically checks this.

4. Error Size Estimation

You have to stop simulation at some point (at some N). How is the finite N affect the accuracy of the simulation?

By central limit theorem, if you use average of iid random variables to estimate the theoretical expectation, the error is of the order $1/\sqrt{N}$.

You may not want to use iid sample.

You may not want to take a simple average.

References

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

Department of Statistics
University of Kentucky
Lexington, KY 40506-0027
mai@ms.uky.edu