Monte Carlo Simulation Studies

* One very important way in which Monte Carlo is used in statistical research is through simulation studies.

* This is a method through which we can gain understanding of the behaviour of an inference technique.

* In frequentist inference, the statements we can make regarding sampling distributions, coverage of intervals or power of tests rely on the concept of repeated sampling from the underlying distribution.

* In some cases we can get exact results but more often we rely on asymptotic theory as the sample size goes to infinity.

* Simulation studies can examine the finite sample properties of an inference procedure when these are not available analytically.

An Example

* Suppose we have data from a right skewed distribution such as the exponential with mean μ.

* It is easy to show that $\overline{X}$ is an unbiased estimator of μ and get the sampling distribution of this statistic.

* Because of the skewness, however, we may prefer to instead try to estimate μ based on the sample median.

* Consider the estimator $\widehat{\mu} = \text{median}(X_1, \ldots, X_n)/\log(2)$.

* The distribution of this estimator can be written down but is hard to work with.

Simulation Study for the Example

* Choose values of n and μ.
* Generate N samples each of size n from an exponential distribution with mean μ.
* Calculate $\widehat{\mu}$ for each of these samples.
* $\widehat{\mu}_1, \ldots, \widehat{\mu}_N$ form a random sample from the sampling distribution of $\widehat{\mu}$.
* We can estimate the bias in $\widehat{\mu}$ by comparing the mean of $\widehat{\mu}_1, \ldots, \widehat{\mu}_N$ with the true value of μ.
* As with any Monte Carlo estimator we should also give a measure of error in our bias estimate.

Simulation Study for the Example

* In general we will want to examine the behaviour of the estimator for different sample sizes and compare it to the sample mean.

* We may also want to examine if the pattern changes for different values of the mean.

* We could therefore run the simulation study on a grid of values of n and μ.

* We can then examine the results of the simulation study to see if we can find any general patterns.

* Often simulation studies can suggest some results which we can then prove theoretically.

Examining Robustness

* Most statistical procedures come with assumptions.

* We often wish to examine the robustness of the methodology to violations of these assumptions since they may not hold in practice.

* Simulation studies allow us to to do this by generating samples known to violate the assumptions and examining how the behaviour changes.

* Another form of robustness is when there are outliers in the data. One way to simulate data with outliers is using mixture distributions.

* As with all simulation studies, we will generally want to look at a range different departures from assumptions or contamination with outliers.

Assessing Convergence to Asymptotic Results

* For a great many statistical procedures we can only find asymptotic results.

* In practice, however, we only ever have finite samples.

* Many simulation studies are designed to examine the loss incurred by the procedure when the sample size is finite.

* As always we will generally consider a sequence of sample sizes to see how large a sample is needed to make the finite sample behaviour close enough to the asymptotic behaviour.

Likelihood Ratio Inference Example

* Suppose that we wish to test

$$H_0 : \theta = \theta_0 \quad \vee \quad H_1 : \theta \neq \theta_0$$

where θ_0 is some specified value of a population parameter θ.

* The likelihood ratio test statistic is defined to be

$$\Lambda(\boldsymbol{X}) \;=\; 2\left[\max_{\theta} f(\boldsymbol{X} \mid \theta) - f(\boldsymbol{X} \mid \theta_0)\right]$$

where $f(\boldsymbol{x} \mid \theta)$ is the joint density of $X_1, \ldots, X_n$ when the true parameter value is θ.

* Asymptotic theory says that $\lambda(\boldsymbol{X})$ converges in distribution to a chi-squared distribution with one degree of freedom.

* This result can be used to test the hypothesis.

Likelihood Ratio Inference Example

* In this example we would want to look at the actual size of a test with asymptotic size α for finite values of n.
* This will involve constructing many samples of each sample size from the joint density $f(\boldsymbol{x} \mid \theta_0)$ and for each one calculating $\Lambda(\boldsymbol{x})$ and deciding if we reject H_0.
* We can estimate the size of the test using the proportion of times we (incorrectly) reject H_0.
* Plots of the density histogram of the observed likelihood ratio statistic with the asymptotic chi-squared distribution superimposed can also be very illuminating.
* We may also be interested in the power function which will involve generating many samples from distributions with parameter $\theta \neq \theta_0$ and finding the proportion of times we reject H_0.

Simulation Study Considerations

* What are the questions we wish to answer?
* To answer those questions which aspects of the model should be varied and which can be held constant?
* What range of values should we use for the varying aspects?
* What calculations need to be done for each simulation?
* What results based on all simulations for each model are required?
* How will the overall results be summarized and presented?

The Question(s) of Interest

* It is very important to clearly define the questions of interest prior to implementing a simulation study.

* In general a single simulation study should not try to answer more than a small number of defined questions.

* A clear definition of the questions will tell us what characteristics of the model need to be varied over separate runs in the study.

* We then need to decide on the values of these factors that we will use. The range of values should be as realistic as possible but we do not want too many values as that increases the length of the study and so may result in fewer replications at each chosen value.

Nuisance Factors

* Usually we are not interested in all possible factors in a given system.

* For other factors we have two choices:
  - Keep them fixed at a chosen value;
  - Randomize them over all possible values by simulating them.

* Which we choose will depend on how generalizable we require our results to be. Randomization increases generalizability at the expense of increased variation.

* Sometimes careful thought can completely remove the dependence on some of these extraneous factors in the study.

Computing Considerations

* The ease with which we can simulate the required random variables;

* The computational time of a single run which will help define the number of different settings and also the number of replicate runs within each setting.

* What we need to store from each run. Keeping too much can make it hard to organize and summarize the results. Keeping too little may mean you later find you are missing some vital information and need to run the simulation again!

* Clearly storage capacity is also factor in how much information you save from each run.

Computing Considerations

* Computers will often stop in the middle of a run! Make sure you have many checkpoints at which you save your output during the run.

* A good rule of thumb is that you should save your current R workspace around every 1-2 hours.

* Can any aspects of the simulation be run in parallel on separate processors?

* If this is possible, make sure that you set up your R code so different R processes are writing to different workspaces or one process will overwrite the others and you will lose data!

* Only when all of these aspects have been decided on and properly documented should you proceed to programming the study and setting it running.

Analysis and Presentation of Conclusions

* A simulation study results in lots of data.

* An essential part of any study is a careful analysis of the data.

* Standard statistical techniques for data description and summarization are useful here.

* Proper presentation of the results is one of the most important aspects.

* A lot of thought must be given to what summary statistics, tables and figures best summarize the results of the study and answer the original questions of interest.