Statistical Computing

2. Technique of Statistical Simulation

Sudhakar Kunte

The statistical simulation technique is a very powerful and simple technique for answering complicated probabilistic questions regarding random experiments. The principles on which this technique is based are the Laws of Large Numbers (LLN). These laws say that under fairly general conditions, as the sample size increases, the sample quantities converge, in an appropriate sense, to the corresponding population quantities. (For details regarding LLN, see [1]). Here we illustrate the technique by giving some examples.

Sudhakar Kunte is Professor of Statistics in the Department of Statistics, University of Pune. He obtained his PhD in Statistics from Purdue University, USA in 1973. His current research interests include Bayesian inference and finite population sampling.

Part 1, Understanding Randomness and Random Numbers, appeared in *Resonance*, Vol.4, No.10, p.16, 1999.

Example 1: Aces from a Deck of Cards

Suppose you are playing a game of cards. Thirteen cards are going to be dealt to you randomly from a well-shuffled standard deck of 52 cards. A friend offers the following bet: He will give you Rs. 100 if your hand contains all the four aces, otherwise you pay him Rs. 5. Do you accept this bet? Your answer will obviously depend on your assessment of the probability that a randomly dealt hand contains all the four aces. There are two ways of finding this probability.

1. Work out the probability using the hypergeometric distribution for X: the number of aces in a randomly dealt hand. In particular, $P(X = 4)$, which we want is $\frac{\binom{4}{4}\binom{48}{9}}{\binom{52}{13}}$.

2. Deal the cards yourself a large number of times and obtain the relative frequency of the hands holding all the four aces. Since this probability is going to be a small number, in order to get a reliable estimate of it, you will

have to deal a very large number of times.

This second method of obtaining an estimate of the desired probability is an example of the statistical simulation technique. It was practically impossible to adopt this method before the availability of fast computers with good random number generators. Of course along with the computer, to solve the problem we also need a computer program. Here we briefly describe the steps in an algorithm for such a simulation.

1. Designate the fifty-two playing cards to numbers 1 to 52, the four aces being assigned numbers 1 to 4. Put these numbers 1 to 52 in an array A(1) to A(52).

2. For $i = 1,$ 13, choose a random number R_i between 1 and $52 - i + 1$ and swap the contents of $A(R_i)$ with $A(52 - i + 1)$.

3. The array A(40) to A(52) contains a random deal. Count the number X of times numbers between 1 and 4 appear in this array. This number X denotes the number of aces held in the hand.

Repeat the steps (1) to (3) a large number of times and find the relative frequency of $X = i$, for $i = 0, 1, 2, 3, 4$. This gives us our estimate of the probability distribution of X. Using this probability distribution of X we can answer all questions related to the random variable X. *Table* 1 gives the output of such an exercise for the number of deals $N = 1000, 10000$ and 100000. The last column of the table gives the actual hypergeometric distribution of X. It is interesting to see the closeness of the simulated distribution of X to the actual distribution. Even for the case of $n =$ 100,000, a Pentium 1 computer took about seven minutes of running time. It is obvious from the table that the bet offered by your friend is not a favourable bet to you. The bet would have been approximately fair if he had made the same offer for the event of holding at least three aces.

	Number of Deals			
	1000	10,000	100,000	Actual
P(X = 0)	0.295	0.3060	0.30544	0.3038175
P(X = 1)	0.427	0.4398	0.43596	0.4388475
P(X = 2)	0.235	0.2104	0.21490	0.2134934
P(X = 3)	0.042	0.0407	0.04102	0.0412000
P(X = 4)	0.001	0.0031	0.00268	0.0026410
E(X)	1.027	0.9951	0.99954	1.0
V(X)	0.706271	0.707076	0.7085398	0.7056

Here E denotes expectation and V the variance.

Table 1. X: Number of aces held in a randomly dealt hand.

Example 2: Simulating a Standard Normal Distribution

In many statistical contexts, one needs sampling distributions of certain statistics, based on normal or other distributions. Sometimes, these sampling distributions could be worked out analytically, and other times the analytical derivation appears to be beyond the capacity of the statistician. In such situations, one resorts to simulation for deriving such sampling distributions. A classical example of simulation of this sort is the work of W S Gosset (Student) who derived in essence the theoretical sampling distribution of what is now known as the Student's t-statistic based on samples from a normal distribution. He verified the fit of the t-distribution he derived by repeatedly random sampling from a data set of height and middle-finger measurements of 3000 criminals, the histograms of which were nearly of the form of that of the normal distribution. He did this early in the twentieth century, well before the advent of even mechanical calculators, let alone high-speed computers. Indeed, Student's derivation of the t-distribution was subsequently made more rigorous by other workers.

Before giving an example of simulation of a sampling distribution, we discuss the generation of random samples from the standard normal distribution. We describe a method due to Box and Muller [2]. Here we start with

a pair of independent uniform (0, 1) random variables U_1, U_2. Let

$$W = R^2 = -2\log U_1; \quad V = 2\pi U_2.$$

Then let

$$X = R\cos V; \quad Y = R\sin V.$$

Now the density of U_1 is $f(u_1) = 1, \quad 0 \leq u_1 \leq 1$. The transformation to W produces $u_1 = e^{-w/2}$, with $du_1 = \frac{1}{2}e^{-w/2}dw$ giving rise to the density of W to be $\frac{1}{2}e^{-w/2}$, $0 \leq w < \infty$, an exponential density with mean 2. Now V is uniform in $(0, 2\pi)$, thus producing the joint density of independent W, V to be

$$f(w,v) = \frac{1}{2\pi}\frac{1}{2}e^{-\frac{w}{2}} \quad 0 \leq w < \infty,\ 0 \leq v \leq 2\pi.$$

Now the transformation

$$x = \sqrt{w}\cos v;\ y = \sqrt{w}\sin v$$

has its inverse

$$w = x^2 + y^2;\ \tan v = \frac{y}{x}$$

giving rise to the Jacobian matrix

$$2\begin{bmatrix} x & y \\ -\frac{y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix}$$

with Jacobian 2, leading to the joint distribution of X, Y to be

$$f(x,y) = 2\frac{1}{2\pi}\frac{1}{2}e^{-\frac{1}{2}(x^2+y^2)}$$

$$= \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)},$$

the joint density of independent standard normal variates. Thus from a pair of random numbers from the uniform (0, 1) distribution, we can construct a pair of independent random observations from the standard normal distribution using the transformation from U_1, U_2 to

X, Y as indicated above. To get a pair of random observations from a normal distribution with mean μ and variance σ^2, we take $\mu + \sigma X$ and $\mu + \sigma Y$

Although this method is nice in theory, there can be some problems depending upon the method used to generate the uniform random numbers. See [3], for details.

Example 3: Q-Q plots for testing normality

Most of classical statistical theory and methods are based on assumptions of independence, homoscedasticity (uniformity of variance) and normality. For instance, in a simple linear regression problem, where x is an independent variable and y a dependent variable (to be predicted using x), the model used is

$$E(Y|x) = \alpha + \beta x,$$

the observations being (y_i, x_i), $i = 1, 2, \quad , n$, the assumptions made being that $(Y_i|x_i)$ are independently normally distributed with expectation as above and with the same variance σ^2. A complete statistical exercise involves checking these assumptions, besides estimating α, β and σ^2. A standard method of checking the normality assumption is to carry out what is called the Q–Q plot (Quantile–Quantile plot) or the normal probability plot of the residuals and then make a judgement as to whether the plot is linear. We can look at it this way:

Suppose we have a sample of n observations and we want to test whether the sample comes from a normal distribution with some unknown mean and variance. One could consider the χ^2 test of goodness of fit and the Kolmogorov–Smirnov test. A simpler alternative is to use the Quantile-Quantile (Q-Q) plot. One proceeds as follows:

1. Form the order statistics $z_{(1)}, z_{(2)}, ..., z_{(n)}$ from the sample.

2. For $i = 1, 2, 3, ..., n$ obtain the normal quantiles, m_i, defined by

$$\Phi(m_i) = \frac{i + .5}{n + 1},$$

where Φ is the distribution function of the standard normal distribution.

3. Plot the points $(m_i, z_{(i)})$ for $i = 1, 2, ..., n$.

4. If the plot is close to being a straight line the assumption of normality is reasonable, otherwise we need to worry.

Figure 1 below gives such a Q-Q plot for twenty six observations simulated from the normal distribution with mean 8 and standard deviation 4.

Is the plot straight enough? Visual inspection is OK, but we would perhaps like to have a better criterion to judge.

Filliben's test is a more formal test for the composite hypothesis of such normality and uses the normal probability plot correlation coefficient as a test statistic. It was

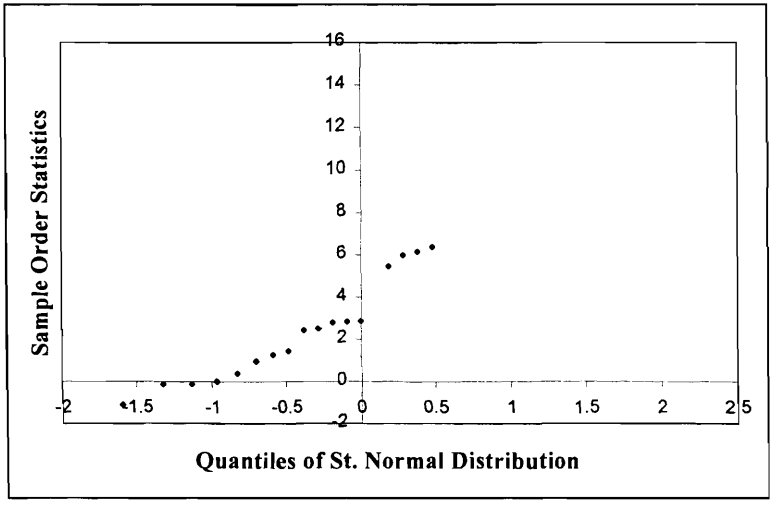

Figure 1. Q–Q plot of 26 observations.

was suggested by J J Filliben [4]. It is increasingly getting popular in various contexts, especially in testing model assumptions in linear models, using the normal probability plot, since it is found to have nice properties vis a vis its competitors.

The idea behind Q-Q plot is that under the hypothesis of normality, the plot of $z_{(i)}$ versus m_i will be linear. Filliben's method tests this linearity using the sample product moment correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(z_{(i)} - \bar{z})(m_i - \bar{m})}{\sqrt{[\sum_{i=1}^{n}(z_i) - \bar{z})^2][\sum_{i=1}^{n}(m_i - \bar{m})^2]}}$$

between X and M, where $\bar{z}$ and $\bar{m}$ are the means of z_i and m_i, respectively. Let us denote the random variable of which r is a sample by R. Small values of r indicate departure from normality.

One can see that the derivation of the theoretical sampling distribution of R will be exceptionally difficult. So Filliben used simulation to work out the cut-off points for R for various sample sizes n. In his original work, Filliben used sample sizes $n = 3(1)50(5)100$, used published random normal deviates from [5], and simulated percentage points for p =0.005, 0.01, 0.025, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.975, 0.99, 0.995. If you are using a computer program for this simulation exercise, including generation of the normal samples, then you might use a method like the Box–Muller method. Some results of Filliben's simulation are given below:

Table 2. Cut-off points of Filliben's statistic.

Sample size (n)	Level						
	0.000	0.01	0.05	0.50	0.90	0.95	0.995
3	0.866	0.869	0.879	0.966	0.999	1.000	1.000
11	0.556	0.883	0.922	0.972	0.988	0.990	0.995
26	0.412	0.939	0.959	0.984	0.992	0.993	0.996
100	0.252	0.981	0.987	0.994	0.997	0.998	0.998

For the sample of *Figure* 1 the calculated value of r is .983, which is large enough not to cause us any worry.

Example 4: Acceptance-Rejection Sampling

If for a survey we want to draw a unit at random from a list of 878 units, what we might do is to sample one unit X from 000 to 999 (which is done easily) and reject the sample if it is > 877 and select the unit $X+1$ otherwise. The rejection method is a method in the same vein, a general statement of which is as follows:
Suppose we want to simulate X from a continuous density f. Let g be another density which is easier to simulate from. Let c be a constant such that

$$\frac{f(y)}{g(y)} \leq c \quad \text{for all} \quad y.$$

Then the steps of the rejection method are as follows:

1. Simulate Y from density g. Simulate independently a number U from uniform (0, 1).

2. If $U \leq \frac{f(Y)}{cg(Y)}$, then let $X = Y$; else go to step 1.

It can be fairly easily shown that the generated X has density f.

Let us apply this method to simulate X from a standard normal distribution. First note that $z = |X|$ has density

$$f(z) = \frac{2}{\sqrt{2\pi}} e^{-z^2/2} \quad 0 < z < \infty.$$

Let us first simulate z. Let us choose g to be the exponential density

$$g(z) = e^{-z},$$

which is easy to simulate from. (Work this out for yourself.) Notice that

$$\frac{f(z)}{g(z)} = \sqrt{2e/\pi} e^{-(z-1)^2/2} \leq \sqrt{2e/\pi}.$$

Hence the rejection method for this case is as follows:

[1] Generate Y from the exponential with mean 1 and U independently from uniform (0,1).

[2] If $U \leq e^{-(Y-1)^2/2}$, that is, $-\log U \geq (Y-1)^2/2$, then let $z = Y$; else go to step 1.

Once we have simulated z, we choose X to be $+z$ or $-z$ with probability $\frac{1}{2}$ each.

Conclusion

With the advent of very high-speed and efficient computing facilities, the theory and practice of statistics has changed dramatically. Nonparametric methods and methods based on randomisation and resampling are gaining popularity. More complex and more realistic models are being used. Bayesian inference carries on without necessarily attempting to use natural-conjugate and such analytically-convenient priors. All these depend upon very heavy computation as replacement for analytical work and depend on a great deal of simulation. Two of these classes of techniques are the *bootstrap* and the *Markov Chain Monte Carlo*, both of which depend heavily on simulation, requiring newer and more complex methods. In subsequent articles, such topics as bootstrap, Monte Carlo integration, Markov Chain Monte Carlo and other related methods will be discussed.

Suggested Reading

[1] R L Karandikar, *Resonance*, Vol. 1, No. 2, 1996.

[2] G E P Box and M E Muller, A note on the generation of random normal deviates, *Annals of Mathematical Statistics*, 29, 610–611, 1958.

[3] B D Ripley, *Stochastic Simulation*, John Wiley & Sons, New York, 1997.

[4] J J Filliben, The probability plot correlation coefficient tests for normality, *Technometrics*, 17, 111–117, 1975.

[5] *A Million Random Digits with 100,000 Normal Deviates*, Glencoe, Ill.: The Free Press Publishers, Rand Corporation, 1955.

Address for Correspondence
Sudhakar Kunte
Department of Statistics
University of Pune
Pune 411 007, India.
Email:
skunte@stats.unipune.ernet.in