Simulation Approaches

April 29, 2008

Many of the examples that we have encountered have allowed us to use exact calculations or to approximate calculations using the central limit theorem. However, many of the questions present in contemporary statistical questions lead us to make estimates that cannot be accomplished in closed form. Consequently, statisticians have developed a variety of simulation techniques. We present two of them here - **importance sampling** and the **bootstrap**.

These **Monte Carlo methods** use stochastic simulations to approximate solutions to questions too difficult to solve analytically. For example, if $X_1, X_2, \ldots$ are independent random variables uniformally distributed on the interval $[0, 1]$. Then, by the strong law of large numbers

$$\overline{g(X)}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \int_0^1 g(x)\, dx = I(g)$$

with probability 1 as $n \to \infty$. The error in the estimate of the integral is supplied by the central limit theorem

$$I(g) - \overline{g(X)_n} \approx \frac{\sigma}{\sqrt{n}} Z$$

where

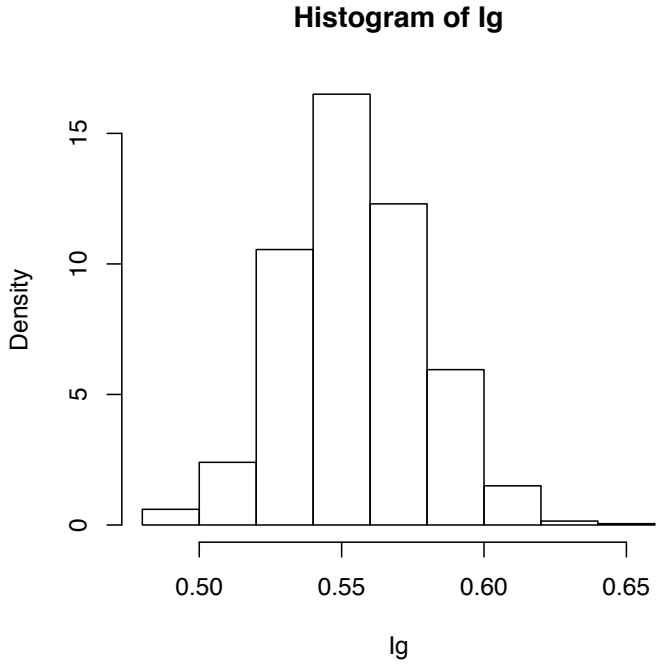
$$\sigma^2 = \int_0^1 (g(x) - I(g))^2\, dx.$$

Example 1. *Let*

$$g(x) = \frac{e^{-x}}{1+x^3} \quad \textit{for } x \in [0, 1]$$

We perform the Monte Carlo integration 1000 times using $n = 100$.

```
> Ig<-rep(0,1000)
> for (i in 1:1000){x<-runif(100);gx<-exp(-x)/(1+x^3);Ig[i]=mean(gx)}
> mean(Ig)
[1] 0.5554112
> sqrt(var(Ig))
[1] 0.02387787
> hist(Ig,probability=TRUE)
```



1 Importance Sampling

Importance sampling methods begin with the observation that we could perform the Monte Carlo integration beginning with $Y_1, Y_2, \ldots$ independent random variables with common density f_Y. This density is called the **importance sampling function** or the **proposal density**. From this, we define the **importance sampling weights**

$$w(y) = \frac{g(y)}{f_Y(y)}.$$

Then

$$\overline{w(Y)}_n = \frac{1}{n}\sum_{i=1}^{n} w(Y_i) \to \int_0^1 w(y) f_Y(y)\ dy = \int_0^1 \frac{g(y)}{f_Y(y)} f_Y(y)\ dy = \int_0^1 g(y)\ dy = I(g).$$

This is an improvement if the variance in the estimator decreases, i.e.,

$$\mathrm{Var}(w(Y)) = \int_0^1 (w(y) - I(g))^2 f_Y(y)\ dy = \sigma_f^2 << \sigma^2.$$

Example 2. *With g as above, we will try to perform the integral*

$$\int_0^1 g(y)\ dy$$

using a proposal density. Recall the probability transform: If Y is a continuous random variable, then $U = F_Y(Y)$ is uniform random variable on $[0, 1]$. Thus, $F_Y^{-1}(U)$ has the same distribution as Y..

If we use a density proportional to $\exp(-y)$*, we obtain*

$$f_y(y) = \frac{e}{e-1}\exp(-y),\ y \in [0,1].$$

The distribution function

$$F_Y(y) = \frac{e}{e-1}(1-\exp(-y)),\ y \in [0,1].$$

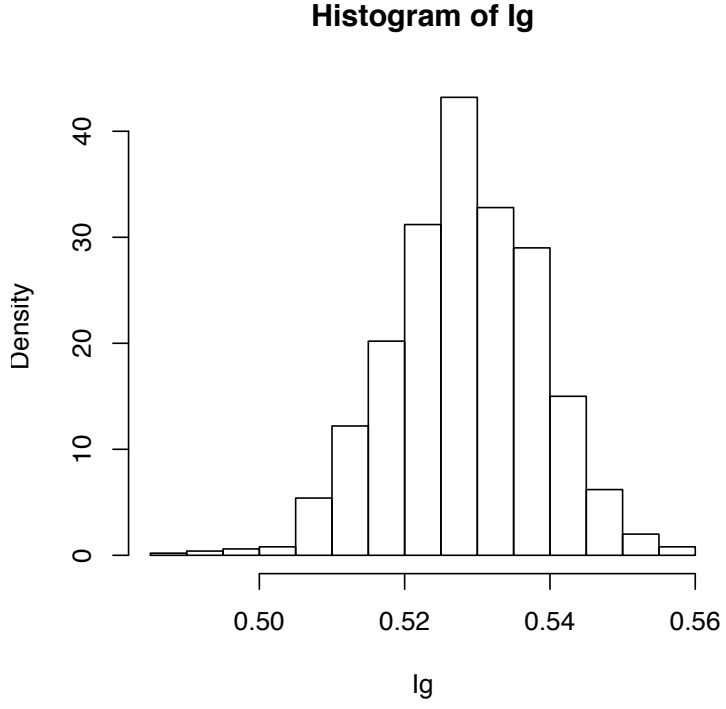
Its inverse

$$F_Y^{-1}(u) = 1 - \ln(e-(e-1)u)$$

The weight function

$$w(y) = \frac{g(y)}{f_Y(y)} = \frac{e-1}{e(1+y^3)}.$$

```
> Ig<-rep(0,1000)
> for (i in 1:1000){u<-runif(100);y<-1-log(exp(1)-(exp(1)-1)*u);
wy<-((exp(1)-1)/(exp(1)*(1+u^3)));Ig[i]=mean(wy)}
> mean(Ig)
[1] 0.5283857
> sqrt(var(Ig))
[1] 0.01015961
> hist(Ig,probability=TRUE)
```



2 The bootstrap

The strategy of the **bootstrap** is to perform a calculation using the **empirical cumulative distribution function** $\hat{F}_n$ as an estimate of the calculation one would like to perform using the distribution function F.

Let $X = (X_1, \cdots, X_n)$ be a simple random sample of S valued random variables

If the empirical distribution function

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) = \frac{1}{n}\#\{X_i \leq x\}$$

is used, then the method is the **nonparametric bootstrap**.

If $\hat{\theta}_n$ is an estimate of θ and $\hat{F}_n(x) = F_{X_1|\Theta}(x|\hat{\theta}_n)$ is used, then the method is the **parametric bootstrap**.

Let $\mathcal{F}$ be a set of cumulative distribution functions and let

$$R : S^n \times \mathcal{F} \to \mathbb{R}$$

be some function of interest, e.g., the difference between the sample median of X and the median of F. Then the bootstrap replaces

$$R(X, F) \qquad \text{by} \qquad R(X^*, \hat{F}_n).$$

Here,

- X is a simple random sample of size n from the distribution function F, and
- X^* is a simple random sample of size n from the empirical cumulative distribution function $\hat{F}_n$.

The bootstrap was originally designed as a tool for estimating bias and standard error of a statistic.

Example 3. *Assume that the sample is real values having distribution function F. Let*

$$R(X, F) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 - \left(\int x \; dF(x)\right)^2,$$

then

$$R(X^*, \hat{F}_n) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^*\right)^2 - (\bar{x}_n),$$

where $\bar{x}_n$ is the observed sample average. Use

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

as an estimate of the variance. Now

$$E[R(X, F)] = \frac{1}{n}\sigma^2, \qquad E[R(X^*, \hat{F}_n)|X = \mathbf{x}] = \frac{1}{n}s_n^2.$$

Example 4. *We will look to see if there is a correlation between a major league baseball player's salary and his performance based on his batting average. In this case the, F is the joint distribution function of the salary X and the batting average Y of major league baseball players.*

A joint empirical distribution function

$$\hat{F}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) I_{(-\infty,y]}(Y_i) = \frac{1}{n}\#\{X_i \le x, Y_i \le y\}$$

is obtained from a random sample of 50 salaries and lifetime batting averages:

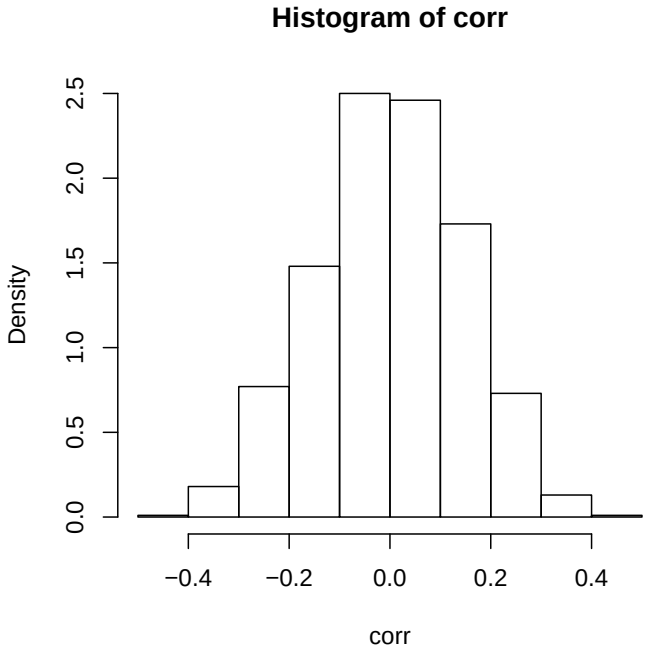
```
> salary
 [1] 9.500 8.000 7.330 7.250 7.166 7.086 6.375 6.250 6.200 6.000 5.825
[12] 5.625 5.000 4.900 4.500 4.000 3.625 3.450 3.150 3.000 2.500 2.400
[23] 2.250 2.125 2.100 1.800 1.500 1.088 1.000 0.950 0.800 0.750 0.720
[34] 0.675 0.630 0.600 0.500 0.325 0.320 0.305 0.285 0.232 0.227 0.221
[45] 0.220 0.220 0.217 0.202 0.202 0.200
> average
 [1] 0.269 0.282 0.327 0.259 0.240 0.270 0.253 0.238 0.300 0.247 0.213
[12] 0.238 0.245 0.276 0.268 0.221 0.301 0.242 0.273 0.250 0.208 0.306
[23] 0.235 0.277 0.227 0.307 0.276 0.216 0.289 0.237 0.202 0.344 0.185
[34] 0.234 0.324 0.200 0.214 0.262 0.207 0.233 0.259 0.250 0.278 0.237
[45] 0.235 0.243 0.297 0.333 0.301 0.224
> cor(salary,average)
[1] 0.1067092
```

The question "Is this value, 0.1067092, for the correlation statistically significantly above 0?" leads to the hypothesis

$$H_0 : \rho \le 0 \quad \text{versus} \quad H_1 : \rho > 0.$$

The bootstrap strategy is to choose 50 salaries with replacement and independently 50 batting averages with replacement and to compute the correlation between these bootstrapped salaries and bootstrapped batting averages. Here is the R *command that give 1000 bootstrap correlations. We take this as the distribution of correlations under the null hypothesis.*

```
> for (i in 1:1000){bavg<-sample(average,50,replace=TRUE);
bsal<-sample(salary,50,replace=TRUE);corr[i]=cor(bavg,bsal)}
> hist(corr,probability=TRUE)
```



The P-value for this test is the probability that correlation, under the null hypothesis, is above the obseerved correlation, 0.1067092. Of the 1000 bootstrap sample, 526 bootstrap correlations were below 0.1067092 and 474 were above. Thus, the bootstrapped P-value for the test is 0.474 and we cannot reject the null hypothesis of no correlation between the salary and batting average of major league baseball players.