Random Numbers and Simulation

- **Generating random numbers:** Generating truly random numbers is not possible
- Programs have been developed to generate pseudo-random numbers:
  - Values are generated from a complicated deterministic algorithm, which can pass any *statistical test* for randomness
  - They *appear* to be independent and identically distributed.
- Random number generators for common distributions are built into R.
- For less common distributions, more complicated methods have been developed (e.g., Acceptance Sampling, Metropolis-Hastings Algorithm)
  - STAT 740 covers these.

(Monte Carlo) Simulation

Some Common Uses of Simulation

1. Modelling Stochastic Behavior
2. Calculating Definite Integrals
3. Approximating the Sampling Distribution of a Statistic (Ex: Max of a random sample)

Modelling Stochastic Behavior

- Buffon's needle
- Random walk
  - Observe $X_1, X_2, \ldots,$ where $p = P(X_i = 1) = 1 - P(X_i = -1) = 1 - p$ and study $S_1, S_2, \ldots$, where $S_i = \sum_{j=1}^{i} X_j$.
  - This is also called *Gambler's ruin*; each X_i represents a \$1 bet with a return of \$2 for a win and \$0 for a loss. The properties of a fair game ($p = .5$) are alot more interesting than the properties of unfair games ($p \neq .5$).
  - Some properties of this process are easy to anticipate (E(S)).
  - Some properties are difficult to anticipate, and can be aided by simulation (The number of returns; average winning or losing streak).

Calculating Definite Integrals

In statistics, we often have to calculate difficult definite integrals (posterior distributions, expected values)

$$I = \int_a^b h(x)\,dx$$

(here, $\mathbf{x}$ could be multidimensional)

Example 1: Find:

$$\int_0^1 \frac{4}{1+x^2}\,dx$$

Example 2: Find:

$$\int_0^1 \int_0^1 (4 - x_1^2 - 2x_2^2)\,dx_2\,dx_1$$

Hit-or-Miss Method

Example 1:

$$h(x) = \frac{4}{1+x^2}$$

$$\left(\int_0^1 \frac{4}{1+x^2}\,dx = 4(\arctan(1) - \arctan(0)) = 4\,\pi/4 = \pi\right)$$

- Determine c such that $c \geq h(x)$ across entire region of interest. (Here, $c = 4$)
- Generate n random uniform (X_i, Y_i) pairs, X_i's from $U[a,b]$ (here, $U[0,1]$) and Y_i's from $U[0,c]$ (here, $U[0,4]$)
- Count the number of times (call this m) that Y_i is less than $h(X_i)$
- Then $I \approx c(b-a)\frac{m}{n}$

[This is (height)(width)(proportion in shaded region)]

Classical Monte Carlo Integration

$$I = \int_a^b h(x)\, dx$$

- Take n random uniform values $U_1, \ldots, U_n$ (could be vectors) over $[a, b]$

  Then
  $$I \approx \frac{b-a}{n} \sum_{i=1}^{n} h(U_i)$$

- This method seems more straightforward than Hit-or-Miss Monte Carlo, but it is actually more efficient.

Expected Value of a Function of a Random Variable

Suppose X is a random variable with density f.

Find $E[h(X)]$ for some function h, e.g.,

$$E[X^2]$$

$$E[\sqrt{X}]$$

$$E[\sin(X)]$$

- Note $E[h(X)] = \int_{\mathcal{X}} h(x)f(x)\,dx$ over the support of f.
- Take n random values $X_1, \ldots, X_n$ from the distribution of X (i.e., with density f)
- Then
$$E[h(X)] \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

Examples

Example 3: If X is a random variable with a $N(10, 1)$ distribution, find $E(X^2)$.

Example 4: If Y is a beta random variable with parameters $a = 5$ and $b = 1$, find $E(-\ln Y)$.

- There are more advanced methods of integration using simulation (Importance Sampling)
- `integrate()` does numerical integration for functions of a *single* variable (*not* using simulation techniques)
- `adapt()` in the "`adapt`" package does multivariate numerical integration

Approximating the Sampling Distribution of a Statistic

To perform inference based on sample statistics, we typically need to know the sampling distribution of the statistics.

Example: $X_1, \ldots, X_n \sim iid\ N(\mu, \sigma^2)$.

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a $t(n-1)$ distribution.

If σ^2 known,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a $N(0, 1)$ distribution.

Then we can use these sampling distributions for inference (CIs, hypothesis tests).

What if the data's distribution is not normal?

1. Large sample: Central Limit Theorem
2. Small sample: Nonparametric procedures based on permutation distribution

- If the population distribution is known, we can approximate the sampling distribution with simulation.
- Repeatedly (m times) generate random samples of size n from the population distribution.
- Calculate a statistic (say, S) each time.
- The empirical distribution of S-values approximates its true distribution.

Example 1: $X_1, \ldots, X_4 \sim Expon(1)$

- What is the sampling distribution of $\bar{X}$?
- What is the sampling distribution of the sample midrange?

$$\frac{X_{(n)} + X_{(1)}}{2}$$