# R

*Scientists should always state the opinions upon which their facts are based.*

—Author Unknown

## Random Numbers

Random numbers are useful for a variety of purposes, such as generating data encryption keys, simulating and modeling complex phenomena, and selecting random samples from larger data sets. They have also been used aesthetically (e.g., in literature and music) and are, of course, ever popular for games and gambling. When discussing single numbers, a random number is one that is drawn from a set of possible values, each of which is equally probable (i.e., a uniform distribution). When discussing a sequence of random numbers, each number drawn must be statistically independent of the others.

There are two main approaches to generating random numbers using a computer. Pseudorandom number generators (PRNGs) are algorithmic approaches that use mathematical formulae (e.g., the linear congruential method) or simply precalculated tables to produce sequences of numbers that appear random. PRNGs are efficient and deterministic, meaning that they can produce many numbers in a short time and a given sequence of numbers can be reproduced at a later date if the starting point in the sequence is known. PRNGs are typically periodic, meaning that the sequence will eventually repeat itself. These characteristics make PRNGs suitable for applications where many numbers are required and where it is useful that the same sequence can be replayed easily, such as in simulation and modeling applications. PRNGs are not suitable for applications where it is important that the numbers be really unpredictable, such as data encryption and gambling.

In comparison, true random number generators (TRNGs) extract randomness from physical phenomena, such as quantum events or chaotic systems. For example, random numbers can be generated by measuring the variations in the time between occurrences of radioactive decay (quantum events) or the variations in amplitude of atmospheric noise (caused by the planet's chaotic weather system). TRNGs are generally much less efficient than PRNGs, taking a considerably longer time to produce numbers. They are also nondeterministic, meaning that a given sequence of numbers cannot be reproduced, although the same sequence may, of course, occur several times by chance. TRNGs have no period. These characteristics make TRNGs suitable for roughly the set of applications for which PRNGs are unsuitable, such as data encryption, games, and gambling. Conversely, the poor efficiency and nondeterministic nature of TRNGs make them less suitable for simulation and modeling applications.

Many statistical tests exist that can be used to assess the randomness of numbers generated with either approach. Examples include the Chi-Square Test, the Run Test, the Collision Test, and the Serial Test. However, testing for randomness is not straightforward, because each possible sequence is equally likely to appear, and good random number generators therefore will also produce sequences that look nonrandom and fail the statistical tests. Consequently, it is impossible to prove definitively whether a given sequence of numbers (and the generator that produced it) is random. Rather, random numbers from a given generator are subjected to an increasing number of tests, and as the numbers pass more tests, the confidence in their randomness (and the generator that produced them) increases.

*—Mads Haahr*

***See also*** Chance; Monte Carlo Methods; Random Sampling

### Further Reading

Knuth, D. E. (1997). *The art of computer programming, Volume 2: Seminumerical algorithms* (3rd ed.). Reading, MA: Addison-Wesley.

The National Institute for Standards and Technology's guidelines for true and pseudorandom number generation and testing: http://csrc.nist.gov/rng/

Theory and practice of pseudorandom number generation: http://random.mat.sbg.ac.at/

True random number service based on atmospheric noise: http://www.random.org/

True random number service based on radioactive decay: http://www.fourmilab.ch/hotbits/

## RANDOM SAMPLING

Three concepts that are relevant to understanding "random sampling" are population, sample, and sampling. A *population* is a set of elements (e.g., scores of a group of prisoners on a personality scale, weights of newborns of mothers younger than 18, longevity of smokers in New York City, and changes in the systolic blood pressure of hypertensive patients exposed to relaxation training) defined by the researcher's interests. Populations can be defined narrowly or broadly and can consist of a few or many elements. For example, a researcher concerned with effects of relaxation training on systolic blood pressure might be interested only in pre- to posttreatment systolic blood pressure changes (i.e., pre–post) in the small group of patients included in her study, or she might be interested in the efficacy of relaxation training in the collection of all persons who could be classified as "elderly obese hypertensive patients" by some operational definitions of these terms.

A *sample* is a subset of a population, and *sampling* refers to the process of drawing samples from a population. Sampling is generally motivated by the unavailability of the entire population of elements (or "data") to the researcher and by her interest in drawing inferences about one or more characteristics of this population (e.g., the mean, $\mu$, or variance, $\sigma^2$, of the population). Thus, in the hypertension example, the researcher might be interested in estimating from her sample data the mean ($\mu$) or variance ($\sigma^2$) of pre- to posttreatment changes in systolic blood pressure of the population consisting of persons who fit the operational definition of "elderly obese hypertensive patients." She might also be interested in using the sample data to test (the tenability of) some hypothesis about $\mu$, such as the hypothesis that the average change in systolic blood pressure ($\mu$) in the population of hypertensives exposed to relaxation training is zero. This hypothesis will be called the "null hypothesis" and will be abbreviated "$H_0$" below. Because characteristics of a population, such as $\mu$ and $\sigma^2$, are defined as "parameters" of the population, the principal objectives of random sampling can often be described in terms of (a) estimation of population parameters and (b) testing of hypotheses about population parameters. Parameters that are often of interest are means ($\mu$s) and linear combinations of means that provide information about mean differences, trends, interaction effects, and so on.

*Simple random sampling* refers to a method of drawing a sample of some fixed size $n$ from the population of interest, which ensures that all possible samples of this size ($n$) are equiprobable (i.e., equally likely to be drawn). Sampling can be carried out with