Generating Samples for Simulation

If we know the distribution of a random variable X, how may samples of its values be generated as inputs for a simulation?

"We know its distribution" has at least two possible meanings. (1) There is an analytical expression for X; (2) We can generate histograms for X even if we do not have an analytical expression for it.

Example of (1): X is a normal distribution with known mean and standard deviation.

Example of (2): X is the number of vehicles entering Gate 10 on Mondays.

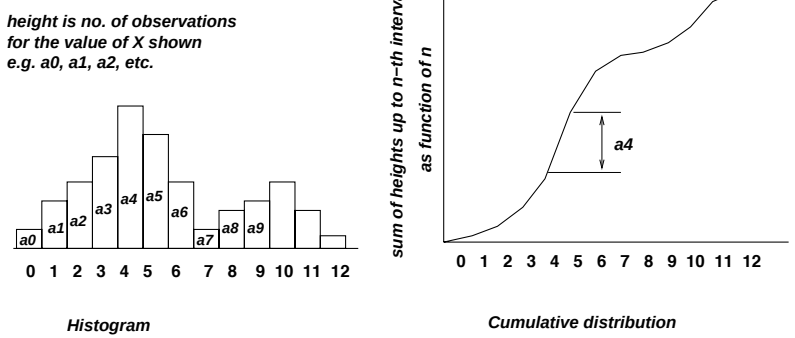
Histograms of Frequencies

Suppose we are able to carry out *experiments* or *observations* on the underlying event space on which X is defined.

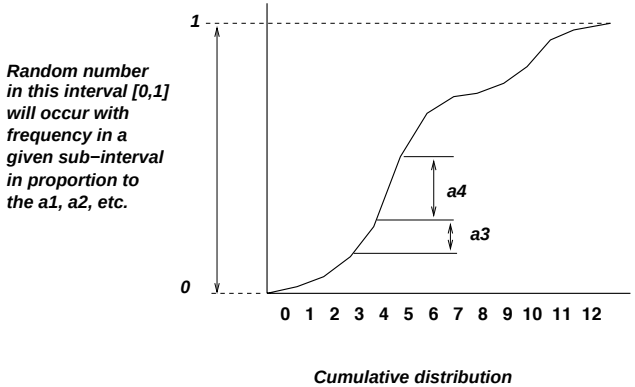
Assume the values of X are integers between 0 and 12. We record the number of occurences of such values as a *histogram*. These are the *frequencies* of the values.

We want a method to produce values of X that respect these frequencies.

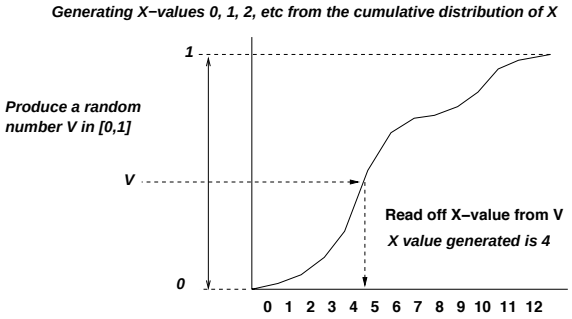
Cumulative Distribution



Random Numbers on Y-axis of Cumul. Dist.



Generating Samples from Cumulative Disrtibution



Successive samples of X−values, e.g, 4,5,0,4,7,2,4,5,12, 4,3, 5,4, ... from repeated generation of random numbers in this way. The relative frequency of events so generated is consistent with the original histogram

Analytically known distributions

This is no problem. Rather than empirically obtained histograms, we have a probability function F for X-values, i.e. the graph of F is interpreted as a *probability density* in the sense that the area under F between v and $v + \delta v$ is the probability of X taking a value in that interval.

In fact, had we *normalized* the histogram by its area, we obtain a similar interpretation.

Then the cumulative distribution G is obtained from F by $G(s) = \int_a^s F(v)dv$, assuming that the least value of the RV X is a.

The same game can then be played to generate X-values respecting the probability density F of X.

Justification

This is the *standard practice* in simulation, including "Monte Carlo" methods for estimation.

The Frequentist is happy with this technique because it can be shown that "in the long run", the sequence of generated values for X will exhibit proportions of its values will better and better approximate the probability distribution specified by either a histogram or by a density function.

Your CSE lecturers are probably agnostic about this in theory, but pragmatic in practice. The question of "randomness" is at the core of this agnosticism.

How many samples?

We glossed over an important point about simulation. You saw tables for simulation experiments on Hashing and AVL trees. How many samples did they use to produce the results? What confidence do they have that these are enough?

The answers to such questions are treated in courses on *Experimental Design* and *Statistical Inference*.

Empirical computer scientists should at least have some awarenesss of these areas.