Drawing a Sample from a Given Distribution

Dhiren Ghosh[1] and Andrew Vogt[2]
[1]Synectics for Management Decisions, Inc., 1901 North Moore Street, Arlington, VA 22209
[2]Georgetown University, Washington, DC 20057-1233

Abstract

Four methods of sampling from a given distribution are considered: natural, inversive, rejective (especially, Lahiri's method), and geometric. In natural sampling, equally-likely sampling is done on a finite population that obeys the distribution approximately. Inversive sampling refers to choosing a random sample from a uniform distribution and applying the inverse of the cumulative distribution function (cdf). This can also be done with joint distributions in higher dimensions. Rejective sampling refers to choosing a random sample from a product of independent uniform, distributions and rejecting units that violate a constraint. Lahiri's method, in particular, does not require computation of the cdf and its inverse. Geometric sampling exploits features of the distribution to transform the sampling into rejective and/or uniform sampling. One application of this is equal area/volume sampling from a surface/body.

Key Words: Rejective sampling, inversive sampling, Lahiri's method, cdf, simulation

1. Introduction

We review four different methods for drawing a sample that obeys a given distribution. Such methods are fundamental to statistics and thus some of what we report will be quite familiar. Consider the problem of selecting a point $(x_1, x_2, \ldots,x_n)$ in R^n according to the probability distribution $f(x_1, x_2, \ldots,x_n)$ where f is either discrete or continuous. How do we go about doing this? We consider four different methods: the natural method, the inversive method, the rejective method (Lahiri), and the geometric method. The last is actually a loose collection of methods.

2. The Methods

2.1 The Natural Method

Briefly, the idea is to find a natural population that obeys the distribution (approximately) and sample from it. Choose units in the population in such a way that each unit has an equal probability of selection. Then determine the values of $x_1, x_2, \ldots,x_n$ on each unit. Enumerating the population or preparing a sample frame can be difficult, and it may be necessary to resort to cluster sampling with weights, etc. However, an advantage of this method is that we need not possess a formula for the distribution. For example, there may be no simple formula for the distribution of heights of male freshmen at American colleges but we can certainly design a survey to sample from this distribution. If, on the other hand, we have a formula and know that some natural population obeys it exactly or approximately, we may proceed in the same way. If the formula has unknown parameters in it, then there are additional complications that can be resolved by such methods as maximum likelihood.

Examples of the natural method include counting the number of emissions from a radioactive substance, or measuring the time between emissions from such a substance. Another example is diastolic blood pressure at mid-day of healthy resting females aged 30 to 40 of a particular ethnicity. These examples call to mind an issue that is rarely addressed in statistics, namely, the postulate of time invariance. Usually the time at which data are drawn from a population is chosen for convenience rather than on a random basis. Thus radioactivity is thought to be without memory, and when we measure blood pressure, we do not stipulate the time of year or times of year when the measurements are made, and thus we assume that seasonal rhythms play no appreciable role. By stipulating "mid-day," we acknowledge the possibility of circadian rhythms.

2.2 The Inversive Method

The inversive method requires that we have a formula for the distribution, including the values of all parameters. In addition, we must be able to generate independent uniform random variables on [0,1] and perform certain computations, either exactly or approximately. The simplest case is that of a single scalar variable X with cumulative

distribution function $F(x) = P(X \leq x)$. Pick a number t uniformly in [0,1], and take $x = F^{-1}(t)$, where F^{-1} is the inverse of the cdf F.

Consider, for example, an exponential distribution $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, with $\lambda > 0$. Its cdf is $F(x) = 1 - e^{-\lambda x}$ for $x > 0$. The inverse cdf is $x = F^{-1}(t) = (1/\lambda) \ln(1/(1 - t))$ for $0 \leq t \leq 1$. If t is selected uniformly from [0,1], it is easily seen that x obeys the exponential distribution.

Two fine points should be mentioned.

First, there is the matter of generating a uniform random variable on [0.1]. One way to do this is to flip a fair coin n times independently, generating a sequence of 0s and 1s that form the binary expansion of a number between 0 and 1. For a decimal expansion select n random numbers from 0 to 9. Of course, in this manner we do not generate all numbers between 0 and 1 but merely integer multiples of $1/(2^n)$ or $1/10^n$. However, the numbers so generated approximate a uniform distribution.

A second issue is that the inversive method requires adjustment if the distribution has a discrete component. In this case take x to be the infimum of the set $\{x: F(x) \geq t\}$.

Finally we note that this method can also be applied to joint distributions although it is somewhat cumbersome. We illustrate for the continuous bivariate case. Let f(x,y) be the joint density, and let $f_1(x)$ be the marginal density for X with cdf F_1. Choose uniform random variables u and v independently in [0,1]. Let $x = F_1^{-1}(u)$. Given x let f(y/x) be the conditional density of y, and let $F_2(y/x)$ be the corresponding conditional cdf. Then let $y = F_2^{-1}(v)$. The pair (x,y) obeys the distribution f(x,y).

2.3 The Rejective Method

Our third method is the rejective method. Like the inversive method, this requires that the formula for the distribution be known. Suppose X is a scalar random variable taking values in the interval [a, b] according to the continuous probability density function f(x). Let M be an upper bound for f on [a, b], M assumed finite. Choose x uniformly in [a, b] (for example, $x = a + t(b - a)$ where t is uniform in [0,1]). Then choose u uniformly in [0, M]. If $u \leq f(x)$, we select x. Otherwise we reject x and start over. Obviously this method will reject fewer points the closer M is to the least upper bound of f on [a, b].

This method is called Lahiri's method. See [1, p. 251]. It has the advantage that it does not require the computation of the cdf F or its inverse F^{-1}. On the other hand, since some elements are rejected, there is a cost per unit.

In case f is unbounded and/or its support is infinite, we may truncate and replace f by an approximant f_1 that is bounded and supported on a finite interval with arbitrarily small effect on the probabilities. We omit tails of arbitrarily small probability and bound the density so that the probabilities of certain values or subintervals are reduced by an arbitrarily small amount.

If f has a discrete component, we first choose a number v uniformly in [0, 1]. If $v \leq$ probability of the continuous component of f, we work only with that component as above. If v > this probability, we pick randomly one of the discrete values x_i (after truncating to get a finite number of such) and keep it if $u \leq f(x_i)$ where u is chosen uniformly from [0,M] where $M = \max\{f(x_j)\}$.

In higher dimensions the extension is straight-forward. For example, in the continuous bivariate case, let f(x,y) be supported on the bounded set Ω, a subset of R^2. Include Ω in a rectangle [a, b] x[c, d], and let M be an upper bound for f. Choose x uniformly in [a, b], y uniformly in [c, d], and u uniformly in [0, M]. Keep (x,y) if (x,y) is in Ω and $u \leq f(x,y)$. Otherwise reject (x,y) and start over.

2.4 The Geometric Method

This method is a miscellaneous collection of ideas based on geometry and partly motivated by Lahiri's method.

.

2.4.1 Uniform distribution on an irregular region

Suppose we want to generate a uniform distribution on an irregular bounded region Ω in R^n. Include Ω in a generalized coordinate rectangle $[a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$. Choose independent uniform random variables $x_1, x_2, \ldots, x_n$ in the respective intervals. Keep $(x_1, x_2, \ldots, x_n)$ if this point lies in Ω, and reject the point otherwise and start over.

2.4.2 Uniform distribution on a triangular region or the interior of a tetrahedron

Suppose we seek a uniform distribution on a triangular region in R^2. The triangle can be replaced by two congruent triangles forming a parallelogram. The parallelogram can be sliced at one end and the slice can be added to the other end, so that the reassembled parallelogram is a rectangle. Since a rectangle is a product of intervals, a point can be selected uniformly in the rectangle without difficulty. We then map back to the parallelogram by a pair of affine isometries and thus to one or the other of our triangles. However, corresponding points in each triangle can be identified, and this gives us a map back to the original triangular region that is uniform. Note that in this method no rejection occurs.

Similar constructions can be performed in higher dimensions. For example, given a tetrahedron, six copies of it with the same base area and same altitude (but not necessarily congruent) can be assembled into a parallelepiped, which in turn can be sliced and reassembled as a rectangular solid. A point can be chosen uniformly in the rectangular solid and mapped back to a unique point inside the original tetrahedron. The mappings used are affine transformations that preserve volume.

These constructions can perhaps be simplified. A triangle, for example, can be mapped to a single rectangle (rather than two triangles) by a set of isometries, and a tetrahedron, we conjecture, may also be reassembled into a single rectangular solid (rather than six tetrahedra forming the solid).

2.4.3 Uniform distribution on a polygonal or polyhedral region

A polygonal region can be decomposed into a finite number of triangular regions. If we know the area of each triangle, we can select a triangle with probability proportional to its area and proceed as above. Likewise a polyhedron can be decomposed into tetrahedra (begin by decomposing the polygonal faces into triangular regions), and if the volume of each tetrahedron is known, then we select a tetrahedron with probability proportional to its volume and proceed as above.

2.4.4 Uniform distribution on the boundary of a polygon or polyhedron that circumscribes a circle or ball

Suppose we have a polygon that circumscribes a circle. An example is any regular polygon. However, there are many others, merely select a finite set of tangent lines to a circle and consider the region so defined and its bounding polygon. The circle has a center. If we pick a point uniformly inside the polygon and project it from the center to the polygon's boundary, we will obtain a point on the polygonal curve that is uniform with respect to length along the curve.

In three dimensions consider a polyhedron circumscribed about a ball. The five regular Platonic solids have this property as do some of the thirteen Archimedean (or so-called semiregular) polyhedra. Any solid defined by a finite set of tangent planes to a ball is in the class of interest. Pick a point uniformly inside the polyhedron, project this point from the center of the ball to the polyhedral surface, and you obtain a point uniformly distributed with respect to area on that surface.

2.4.5 Uniform distribution on the boundary of a convex polygon or polyhedron

The method of 2.4.4 can be applied with any interior point of a convex polygon or convex polyhedron replacing the center used previously. The only requirement is that we know the perpendicular distance of this point form each of the polygonal edges or polyhedral faces. We then use this distance to weight the probabilities of our points to ensure that they are uniform with respect to length/area on the bounding curve/surface.

2.4.6 Nonuniform distributions derived from geometry

Consider the density function $f(x) = 3x^2$ for $0 \leq x \leq 1$. This can easily be handled by the inversive or rejective methods.

.

However, it can also be treated geometrically in the following way. Choose a point (x, y, z) uniformly inside the unit ball (perhaps by the method of 2.4.1). Then convert to spherical coordinates (r, φ, θ) and throw away φ and θ. It will be found that the variable $r = \{x^2 + y^2 + z^2\}^{1/2}$ obeys the distribution $f(r) = 3r^2$.

Similar devices can be used with such densities as $f(r) = (n+1)r^n$, and trigonometric densities such as $(1/2) \cos \varphi$.

3. Remarks

Our motivation for studying this subject was to find ways to sample uniformly from surfaces. An n-dimensional surface can be regarded as a manifold patched together by charts from regions in n-dimensional Euclidean space. We can equip these regions with density functions that correspond to n-dimensional volume or to some non-uniform distribution. If we know the n-volume of each region, we can often use one or more of the methods discussed here to sample the manifold.

References

[1] Cochran, William G. Sampling techniques. Third Edition. John Wiley & Sons, inc. 1977.