Random variables: theory and practice

Contents

1 Defining and generating randomness

You may recall that random variables live on sample space, but has anyone told you the address of this sample space? You have generated random variables on R, how does it do that? It turns out that these two issues are very closely related.

2 Theory

Recall that a **probability space** can be expressed as a triple, $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the **sample space** and $\mathcal{F}$ is the set subsets of Ω whose elements are called **events**, and $\mathbb{P}$ is a countably additive **probability measure** which assigns each event a number in $[0, 1]$. A (real-valued) **random variable** is map $X : \Omega \to \mathbb{R}$,

such that $X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{F}$ for all open intervals $I \subset \mathbb{R}$. It's **law** or **distribution** is given the **cumulative distribution function** given by

$$F(x) = \mathbb{P}(X \le x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \le x\}).$$

Many properties about X just depend on its law, so in elementary modules on probability it is easy to forget that random variables live on sample space. For example, if X is an integer-valued random variable, we can compute the probability mass function for X as

$$f(n) = F(n) - F(n-1),$$

from which we can compute the mean of X by

$$\mathbb{E}X = \sum_{n \in \mathbb{Z}} n f(n).$$

3   Can we forget about sample space?

One reason why you haven't heard much about sample spaces is because it is not easy to prove the existence of a sample space that can support a random variable that is uniformly distribution on the unit interval. However, if you are ready to assume that existence of even one such random variable, on some sample space, this is enough to do most of probability theory. Any collection of random variables can be thought of as a deterministic function of a single random variable that is uniformly distributed on the unit interval.

4   The inverse transform method

We demonstrate how to generate any (continuous) random variable from a random variable U that is uniformly distributed on the unit interval.
Let F is the cdf of a continuous real-valued random variable. For simplicity, assume that F is increasing, so that F^{-1} exists. Notice that the random variable $X = F^{-1}(U)$ has cdf F:

$$
\begin{aligned}
\mathbb{P}(F^{-1}(U) \le x) &= \mathbb{P}(U \le F(x)) \\
&= F(x).
\end{aligned}
$$

If we need to generate say a Bernoulli p random variable, then we can set $\phi(U) = 1$ if $U \in [0, p]$ and $\phi(U) = 0$ if $U \in [p, 1]$. Other discrete distributions can be dealt with similarly.

5   Independent random variables

Here we show that we can generate i.i.d. sequences of random variables, from a single uniform.
Recall that functions of independent random variables are again independent. Thus by the inverse transform method, it suffices to have independent uniform random variables.

If U is uniformly distributed on $[0, 1]$, express U as

$$U = X_1 X_2 X_3 X_4 \cdots$$

using the usual base-10 expansion of U, so that X_i are random variables that take integer-values $0, 1, 2, \ldots, 9$. It is not hard to show that X_i are independent and uniformly distributed. By rearranging and recombining

these digits X_i, from one U uniform random variables, you can produce as many independent ones as you like. For example, by taking odd and even terms:

$$V_1 = X_1 X_3, \ldots$$

and

$$V_2 = X_2 X_4, \ldots$$

are uniformly distributed on $[0, 1]$ and independent.

6   What does R do?

We already saw that the key is uniform random variables, which can be thought as a sequence of independent digits. R simulates a uniform random variable. However, it is not truly random as no dice are rolled! R does something in spirit close to the following, which is completely deterministic!

The Lehmer random number generator, generates a deterministic sequence that behaves like a random one. Take m to be prime. Let a be a primitive root modulo m. Set $0 < X_0 < m$:

$$X_{k+1} = a \cdot X_k \bmod m$$

The X_k behave like independent random variables that are uniformly distributed in $0, 1, \ldots, m-1$.

7   Box-Mueller for normal random variables

It turns out that the the inverse transform method is often too slow. The Box-Muller method for generating normal random variables. Ann. Math. Stat. 1958 is a more efficient way generating normal random variables.

Let U and V be independent random variables that uniformly distributed on the unit interval. Set

$$X = \sqrt{-2 \log U} \cos(2\pi V)$$

and

$$Y = \sqrt{-2 \log U} \sin(2\pi V).$$

Then X and Y are independent standard normals.

Proof. We solve for (U, V) to obtain that

$$U = \exp[-(X^2 + Y^2)/2]$$

and

$$V = \frac{1}{2\pi} \arctan(Y/X).$$

Thus the map

$$(u, v) \mapsto \left(\sqrt{-2 \log u} \cos(2\pi v), \sqrt{-2 \log u} \sin(2\pi v) \right)$$

is a bijection from $(0, 1) \times (0, 1)$ to $\mathbb{R}^2 \setminus \{(0, 0)\}$.

It has Jacobian is given by

$$J(x, y) = -\frac{1}{2\pi} e^{-(x^2 + y^2)/2}$$

Hence the (joint) pdf for (X, Y) is given by

$$(x, y) \mapsto \left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \right),$$

as desired. $\square$

8 Acceptance/Rejection methods: Sampling on a disc

A powerful method of generating random variables, goes back to von Neumann.

We simulate a single uniformly distributed point on the disc, by simulating a uniform point on a square, and *accept* it if lands inside an inscribed disc, or *reject* otherwise, and repeat until we have acceptance.

Consider the following function

```
point <- function(){
  z=2
  while(length(z)==1){
  x = 2*runif(1) -1
  y = 2*runif(1) -1
if (x^2 + y^2 <1) { z<-c(x,y)}
  }
  z
}
```

applied 500 times

```
re=replicate(500, point())
plot(re[1,], re[2,], xlim=c(-1.1,1.1), ylim=c(-1.1,1.1), asp=1)
```
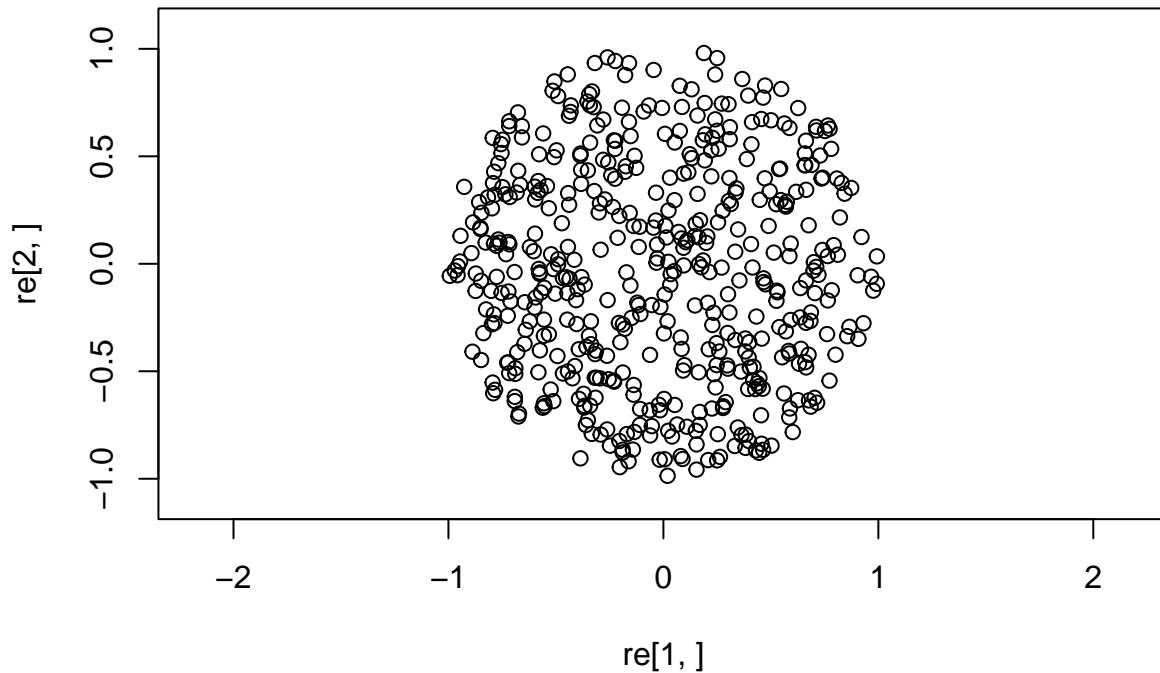
8.1 Acceptence/Rejection: Sampling from a pdf

The acceptance/rejection method can used to sample from a pdf.

- Suppose we want to generate a random variable X with pdf f.

- Suppose f is a complicated pdf with an inverse that is difficult to compute.

- Suppose we can easily generate Y which has pdf g.

- Suppose there exists M so that $f(x) \leq Mg(x)$.

- Then you are in good shape!

8.2 Procedure

- Let U be uniformly distirbuted on $[0, 1]$ and independent of Y
- Sample from $(U, Y) = (u, y)$
- Check if

$$u \leq \frac{f(y)}{M \cdot g(y)}$$

- If the inequality holds, accept the value of y, and report $X = y$; reject otherwise, and repeat.

8.3 Proof of method

Observe that by the change of variables formula and Fubini, we have

$$
\begin{aligned}
\mathbb{P}(Y \le x, MUg(Y) \le f(Y)) &= \int_{-\infty}^{x} \int_{0}^{1} \mathbf{1}[Mug(y) \le f(y)]g(y)dudy \\
&= \int_{-\infty}^{x} \mathbb{P}[U \le f(y)/Mg(y)] \cdot g(y)dy \\
&= \int_{-\infty}^{x} f(y)/Mdy \\
&= F(x)/M.
\end{aligned}
$$

Thus

$$
\mathbb{P}(Y \le x | MUg(Y) \le f(Y)) = F(x).
$$

So, given that we accepted the distribution of Y, is the distribution of X, that random variable we want to sample/simulate.
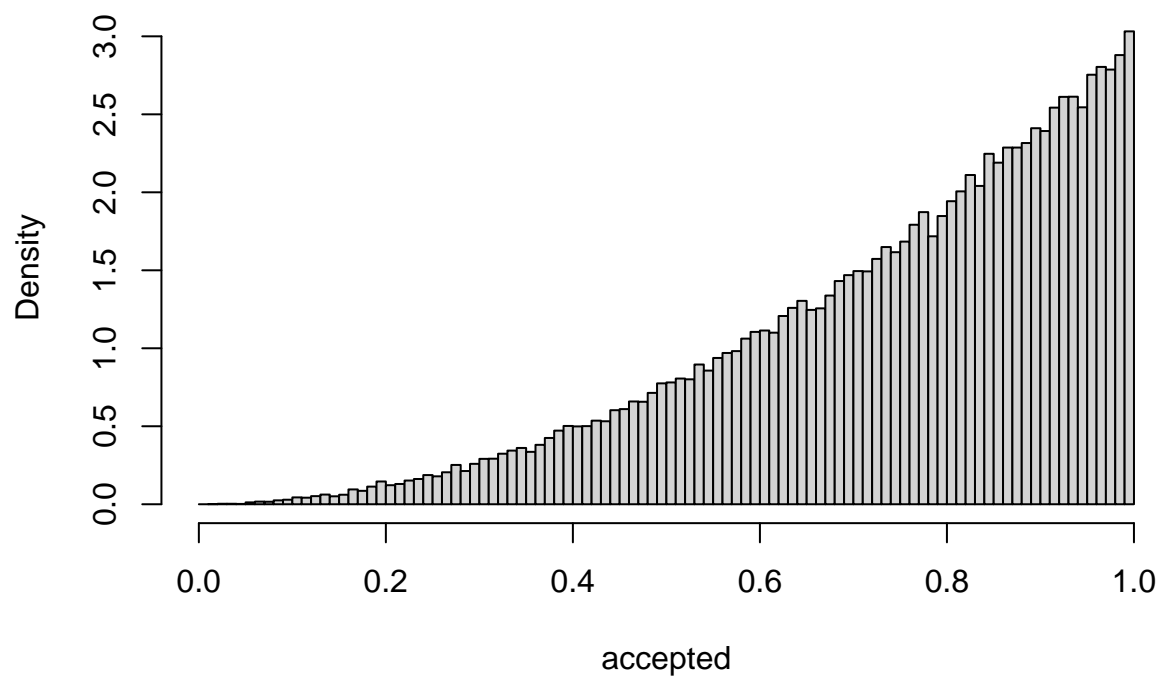
8.4 A simple example with code

- Consider the pdf $f(x) = 3x^2$ on the unit interval $[0, 1]$.
- Consider the pdf $g(x) = 1$ on the unit interval.
- Of course, $f(x) \le 3g(x)$.

```
samplef <- function(){
x=-1
while(x==-1){
y = runif(1)
u = runif(1)
if( u*3 < 3*(y^2)){ x<- y}
}
x
}
```

```
accepted=replicate(100000, samplef())
B = seq(0, 1, by=0.01)
hist(accepted, prob=TRUE, B)
```

Histogram of accepted

9 Summary

- We rediscovered probability spaces
- The central role of uniform random variables in theory and practice
- Sampling methods such as acceptance/rejection

10 Endnotes

- Kolmogorov, 1933 laid the foundations of modern probability theory be showing that probability theory could be completed in mathematical apparatus of measure theory.

* Rmd Source * Version: 19 April 2024