Lectures on stochastic simulation

Matti Vihola

January 9, 2020

Preface

This is a summary of the lectures of the Spring 2020 course "MATS442 Stochastic simulation" at Department of Mathematics and Statistics, University of Jyväskylä. These notes are inspired by the lecture notes of Antti Penttinen [19], Geoff Nicholls [16] Adam Johansen, Ludger Evers and Nick Whiteley [11], and by the textbooks [22, 4, 18].

The purpose of these notes is to support the lectures, so they may not be well suited for self study. Some important methods (and examples) are covered also in the problems classes. The "Monte Carlo Statistical Methods" book by Christian P. Robert and George Casella [21, 22] is a recommended supporting material. Other references to the literature are given during the course regarding more specific topics.

1 Introduction

Simulation of stochastic systems provides powerful tools to inspect complex models. Monte Carlo methods use simulations in order to approximate expectations and probabilities related to (nearly) arbitrary models. The methods have been used (in the modern sense) already from the 1950s, and by the increase of computational power and the methodological advances over the years, they have become central tools in many applications. The analysis, efficient implementation and development of simulation methods are all active research areas. The simulation methods tend to rely on a handful of elegant key ideas, many of which are touched within this course.

Prerequisities

The course requires background in

- basic (vector) calculus (differentiation, integration),
- basic probability (probabilities, expectation, conditioning, joint distributions...), and
- knowledge of standard limit theorems in probability (law of large numbers, central limit theorem).

Basic programming skills are also useful.

Learning outcomes

After taking this course, you will be able to:

- *apply* simulation methods in practice,
- *understand* why the methods work, (and why they sometimes work poorly),
- *modify* the methods and *combine* them for your needs (for some specific application).

Theory of stochastic simulation may be categorised as *applied probability*, and the application of the methods in practice as *computational statistics*.

The programming environment

In the lectures, we focus on methods (algorithms) and theory behind them, and in the exercises we also implement the algorithms and experiment them in practice.

We are using primarily the Julia programming language in the course. If you are not familiar with Julia, that is not a problem, as basic use is simple and similar to R/Matlab/Python, and no advanced programming skills will be needed. There is also plenty of online introductory material available for self-study. Solutions to implementation problems may be returned also using another programming language, such as R or Python.

Why Julia? Because it is fast, which is essential because many simulation-based methods are computationally intensive...

1.1 Conventions

We will generally use the symbols $\mathbb{P}$ and $\mathbb{E}$ for probability and expectation, denote random variables[1] with capital letters, and $\mathbf{1}(\cdot)$ stands for the characteristic function (e.g. $\mathbf{1}(X \in A) = 1$ if $X \in A$ and $\mathbf{1}(X \in A) = 0$ otherwise).

In this course, we focus on the two common types of distributions p:

continuous distribution defined by a probability density function (p.d.f.) p on a space $\mathbb{X} = \mathbb{R}^d$.

discrete distribution defined by a probability mass function (p.m.f.) p on a finite space $\mathbb{X} = \{x_1, \ldots, x_m\}$ or a countably infinite space $\mathbb{X} = \{x_1, x_2, \ldots, \}$.

The notation $X \sim p$ means that X is a random variable has distribution p and $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$ or $(X_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim} p$ means $(X_k)_{k\geq 1}$ are independent and each $X_k \sim p$.

Remark 1.1. The $\mathbb{X} = \mathbb{R}^d$ then $X \sim p$ is a *random vector* with distribution p, that is, the coordinates $X^{(1)}, \ldots, X^{(d)}$ are *random numbers* with joint density $p(x^{(1)}, \ldots, x^{(d)})$.

For $f : \mathbb{X} \to \mathbb{R}$, we write $\mathbb{E}_p[f(X)]$ meaning the expectation of $f(X)$ when $X \sim p$. That is,

$$\mathbb{E}_p[f(X)] = \begin{cases} \int_{\mathbb{X}} f(x)p(x)\mathrm{d}x, & \text{if } X \text{ is continuous (and } p \text{ is a p.d.f.)} \\ \sum_{x\in\mathbb{X}} f(x)p(x), & \text{if } X \text{ is discrete (and } p \text{ is a p.m.f.)} \end{cases}$$

We also write similarly $\mathrm{Var}_p\big(f(X)\big) = \mathbb{E}_p\big[(f(X) - \mathbb{E}_p[f(X)])^2\big] = \mathbb{E}_p\big[f^2(X)\big] - (\mathbb{E}_p[f(X)])^2$.

Remark 1.2. If $\mathbb{X} = \mathbb{R}^d$, we mean that

$$\int_{\mathbb{X}} f(x)p(x)\mathrm{d}x = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x^{(1)}, \ldots, x^{(d)})p(x^{(1)}, \ldots, x^{(d)})\mathrm{d}x^{(1)} \cdots \mathrm{d}x^{(d)}.$$

We shall omit the domain of integration in most cases if we integrate over the whole space $\mathbb{X}$. For instance, if $\mathbb{X} = \mathbb{R}$, we may write

$$\int f(x)p(x)\mathrm{d}x = \int_{\mathbb{X}} f(x)p(x)\mathrm{d}x = \int_{-\infty}^{\infty} f(x)p(x)\mathrm{d}x.$$

---

1. We use the term 'random variable' regardless of dimension and nature (vector, numbers, integers etc.).

(NB: Not to be confused with the indefinite integral!)

Remark 1.3 (*, starred sections, remarks etc. such as this are *non-examinable* extra material, which may be safely skipped). For those who are familiar with general probability theory, the integral above can also be taken with respect to an arbitrary (sigma-finite) measure "dx" on a general measurable space $\mathbb{X}$ (instead of the Lebesgue measure on $\mathbb{R}^d$ equipped with the Borel sets, or the counting measure on countable $\mathbb{X}$ equipped with the power set). Then, p is the density (Radon-Nikodym derivative) of the distribution of interest with respect to dx. Note also that if μ is a probability measure of interest on $\mathbb{X}$, we may also take itself as the dominating measure d$x = \mu(\mathrm{d}x)$, in which case $p \equiv 1$. Most of the techniques presented in the course generalise into such a setting directly.

The functions $f : \mathbb{X} \to \mathbb{R}$ for which expectations are defined must, of course, be measurable. We shall implicitly assume the required measurability of any such test function (or set), without explicit notification.

We do not explicitly state the probaiblity space where the random variables etc. are defined. Instead, we either work with countable sequences of independent and identically distributed random variables, or more generally, dependent sequences defined by conditional probabilities, such as discrete-time Markov chains. For such countable 'algorithmic' definitions, the existence of the underlying probability space is standard (using the Ionescu-Tulcea extension theorem).

1.2 The Monte Carlo method

Let us start by a very simple but illustrative example.

Example 1.4 (Finding approximation of π by simulation). Suppose 'rain drops' fall uniformly in a 2×2 metre square. Let us check how they could be used to determine an approximation of π.

Probability of one drop hitting a unit radius disc inside the square

$$\beta = \frac{\text{area of disc}}{\text{area of square}} = \frac{\pi}{4}$$

If $H_k = 1$ if the drop hit the unit disc and $H_k = 0$ otherwise[2], then

$$4\Big(\frac{1}{n}\sum_{k=1}^{n} H_k\Big) \xrightarrow{n\to\infty} \pi, \qquad \text{(almost surely)}$$

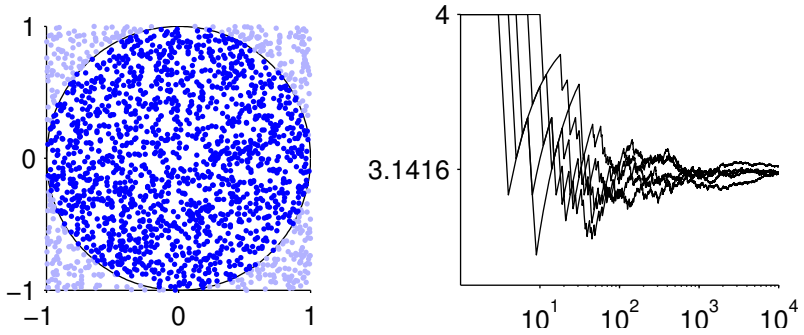
by the (strong) law of large numbers.

Definition 1.5 (Monte Carlo)**.** Assume $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$. Then, for $f : \mathbb{X} \to \mathbb{R}$,

$$I_p^{(n)}(f) := \frac{1}{n}\sum_{k=1}^{n} f(X_k) \tag{1}$$

is the *Monte Carlo approximation* of $\mathbb{E}_p[f(X)]$.

2. $H_k \sim \text{Bernoulli}(\beta)$: $\mathbb{P}(H_k = 1) = \beta$, $\mathbb{P}(H_k = 0) = 1 - \beta$.



(a) (b)

Figure 1: The rain drops example: (a) 'rain drops' falling inside the unit disc highlighted (b) five realisations of the experiment: estimates of π converge as $n \to \infty$ (note the log scale on n).

Example 1.6. Example 1.4 corresponds to

- $\mathbb{X} = \mathbb{R}^2$
- $(X_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1]^2)$,
- $f(x) = 4 \cdot \mathbf{1}\left(\|x\| \leq 1\right)$ or $f\big(x^{(1)}, x^{(2)}\big) = 4 \cdot \mathbf{1}\left((x^{(1)})^2 + (x^{(2)})^2 \leq 1\right)$,

or, equivalently, to simulated Bernoulli random variables,

- $\mathbb{X} = \{0, 1\}$
- $(H_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim}$ Bernoulli$(\pi/4)$ (NB: We simulate H_k above as $H_k = f(X_k)$, which does not require us to evaluate $\pi/4$!)
- $f(h) = 4h$.

In both cases, $\mathbb{E}_p[f(X)] = \pi$.

Remark 1.7. If $\mathbb{X}$ is finite, then $\mathbb{E}_p[f(X)]$ is a finite sum and can, in principle, be computed exactly. However, we might not be able to calculate $p(x)$ exactly (cf. Example 1.6), or the space $\mathbb{X}$ may have a huge number of elements, for example if $\mathbb{X}$ is the set of all 100×100 binary images $\mathbb{X} = \{0,1\}^{100\times 100}$, in which case the Monte Carlo method may still be relevant...

Example 1.8. Definition 1.5 allows for calculating:

- Probabilities: $\mathbb{P}(X \in A)$ for $X \sim p$, by choosing $f(x) = \mathbf{1}\left(x \in A\right)$ (cf. Example 1.4).
- Multiple expectations (or probabilities) simultanously: $E_p[f_k(X)]$ for a number of test functions $f_1, \ldots, f_m$. For example, the mean of random vector $X = \big(X^{(1)}, \ldots, X^{(d)}\big) \sim p$ is a vector of means of individual coordinates,

$$\mathbb{E}_p[X] = \big(\mathbb{E}_p[f_1(X)], \ldots, \mathbb{E}_p[f_d(X)]\big),$$

  where $f_k\big(x^{(1)}, \ldots, x^{(d)}\big) = x^{(k)}$.

  Note that we may simulate $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$ and construct all $I_p^{(n)}(f_1), \ldots, I_p^{(n)}(f_d)$ using the *same* samples $X_1, \ldots, X_n$.

1.3 Properties of Monte Carlo estimators

We need the strong law of large numbers and the central limit theorem frequently, so we shall restate them here without proof.

Theorem 1.9 (Strong law of large numbers)**.** *Assume $Y_1, Y_2, \ldots$ are i.i.d. random numbers such that $\mu = \mathbb{E}[Y_1]$ is finite. Then,*

$$\frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow{n\to\infty} \mu \qquad \text{a.s. (almost surely).} \tag{2}$$

Remark 1.10. Recall that $Z_n \to Z$ a.s. $\implies$ $Z_n \to Z$ in probability $\implies$ $Z_n \to Z$ in distribution;

$$\begin{aligned}
Z_n \to Z \text{ a.s.} \quad &\iff \quad \mathbb{P}(Z_n \xrightarrow{n\to\infty} Z) = 1 \\
Z_n \to Z \text{ in probability} \quad &\iff \quad \text{For all } \epsilon > 0,\ \mathbb{P}(|Z_n - Z| \le \epsilon) \xrightarrow{n\to\infty} 1 \\
Z_n \to Z \text{ in distribution} \quad &\iff \quad \text{For all continuity points } t \text{ of the mapping } t \mapsto \mathbb{P}(Z \le t), \\
&\qquad\qquad \mathbb{P}(Z_n \le t) \xrightarrow{n\to\infty} \mathbb{P}(Z \le t),
\end{aligned}$$

In particular, you may always replace "almost surely" in (2) by "in probability," but not vice versa. (Theorem 1.9 with "in probability" instead of "a.s." is known as the *weak* law of large numbers.)

Theorem 1.11 (Central limit theorem)**.** *Assume $Y_1, Y_2, \ldots$ are i.i.d. random numbers with $\sigma^2 := \mathrm{Var}\big(Y_1\big) \in (0, \infty)$, then with $\mu = \mathbb{E}[Y_k]$,*

$$\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(Y_k - \mu) \xrightarrow{n\to\infty} N(0,1) \qquad \text{in distribution,}$$

in other words,

$$\mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(Y_k - \mu) \le t\right) \xrightarrow{n\to\infty} \Phi(t) \qquad \text{for all } t \in \mathbb{R},$$

where Φ is the standard normal c.d.f., that is, $\Phi(t) := \mathbb{P}(Z \le t)$ with $Z \sim N(0,1)$.

Proposition 1.12. *The Monte Carlo estimators satisfy the following properties:*
Unbiasedness *If $\mathbb{E}_p[f(X)]$ is finite, then $\mathbb{E}[I_p^{(n)}(f)] = \mathbb{E}_p[f(X)]$ for all $n \ge 1$.*
Strong consistency *If $\mathbb{E}_p[f(X)]$ is finite, then $I_p^{(n)}(f) \xrightarrow{n\to\infty} \mathbb{E}_p[f(X)]$ almost surely.*
Variance *If $\mathrm{Var}_p[f(X)] < \infty$, then $\mathrm{Var}[I_p^{(n)}(f)] = \frac{1}{n}\mathrm{Var}_p[f(X)]$.*

Proof. Let $Y_k = f(X_k)$, then $\mathbb{E}[Y_k] = \mathbb{E}_p[f(X)]$. Now,

$$\mathbb{E}[I_p^{(n)}(f)] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[Y_k] = \mathbb{E}_p[f(X)].$$

Strong consistency follows from application of the strong law of large numbers, because $Y_k := f(X_k)$ are i.i.d. random variables with expectation $\mathbb{E}_p[f(X)]$. Finally,

$$\mathrm{Var}[I_p^{(n)}(f)] = \frac{1}{n^2}\mathrm{Var}\Big(\sum_{k=1}^{n} Y_k\Big) = \frac{1}{n}\mathrm{Var}(Y_1). \qquad \square$$

Proposition 1.13 (Asymptotic Monte Carlo error)**.** *Assume* $(X_k)_{k\geq 1} \overset{i.i.d.}{\sim} p$ *and* $f : \mathbb{X} \to \mathbb{R}$ *is such that with* $\sigma^2 := \mathrm{Var}_p\big(f(X_1)\big) \in (0, \infty)$*,*

(i) $\sqrt{n}\big[I_p^{(n)}(f) - \mathbb{E}_p[f(X)]\big] \xrightarrow{n\to\infty} N(0, \sigma^2)$ *in distribution.*

Furthermore, letting $\hat{\sigma}_n^2$ *stand for the sample variance:*

$$\hat{\sigma}_n^2 := \frac{1}{n-1}\sum_{k=1}^{n}\big(f(X_k) - I_p^{(n)}(f)\big)^2;$$

(ii) for any $\beta \in \mathbb{R}$*,*

$$\mathbb{P}\big(\sqrt{n}\big[I_p^{(n)}(f) - \mathbb{E}_p[f(X)]\big] \leq \beta\hat{\sigma}_n\big) \xrightarrow{n\to\infty} \Phi(\beta), \text{ and}$$

(iii) for any $\alpha \in (0, \infty)$*, the following confidence interval is consistent:*

$$\mathbb{P}\left(\mathbb{E}_p[f(X)] \in \left[I_p^{(n)}(f) \pm \alpha\frac{\hat{\sigma}_n}{\sqrt{n}}\right]\right) \xrightarrow{n\to\infty} 1 - 2\Phi(-\alpha).$$

Recall the following lemma for the proof:

Lemma 1.14 (Slutsky)**.** *Suppose the random numbers* $X_n \to X$ *in distribution and* $Y_n \to y$ *in probability, where* $y \in \mathbb{R}$ *is a constant, then:*

(i) $X_nY_n \to Xy$ *in distribution.*

(ii) If $y \neq 0$*, then* $X_n/Y_n \to X/y$ *in distribution.*

Proof of Proposition 1.13. (i) is an application of the CLT with $Y_k := f(X_k)$, and because $\hat{\sigma}^2 \to \sigma^2$ almost surely (and in probability), (ii) follows by Lemma 1.14.

Consider then (iii), and observe that

$$\begin{aligned}&\mathbb{P}\left(\mathbb{E}_p[f(X)] \in \left[I_p^{(n)}(f) \pm \alpha\frac{\hat{\sigma}_n}{\sqrt{n}}\right]\right)\\ &\quad= \mathbb{P}\left(I_p^{(n)}(f) - \mathbb{E}_p[f(X)] \leq \alpha\frac{\hat{\sigma}_n}{\sqrt{n}}\right) - \mathbb{P}\left(I_p^{(n)}(f) - \mathbb{E}_p[f(X)] < -\alpha\frac{\hat{\sigma}_n}{\sqrt{n}}\right).\end{aligned}$$

The first term converges to $\Phi(\alpha) = 1 - \Phi(-\alpha)$ by (ii). The second can be sandwiched between $\Phi(-\alpha - \epsilon)$ and $\Phi(-\alpha)$ for arbitrary $\epsilon > 0$. □

Remark 1.15. Proposition 1.13 is an asymptotic result, so it does not give any guarantees for a finite n. In practice, the approximation is often informative for moderate α and large n.

Remark 1.16. The variance expression of Proposition 1.12 can be used directly to build non-asymptotic upper bounds by Chebychev's inequality,

$$\mathbb{P}(|I_p^{(n)}(f) - \mathbb{E}_p[f(X)]| \geq \epsilon) \leq \frac{\mathrm{Var}[I_p^{(n)}(f)]}{\epsilon^2} = \frac{\mathrm{Var}_p[f(X)]}{n\epsilon^2} \qquad \text{for all } \epsilon > 0.$$

Note that we need to know $\mathrm{Var}_p[f(X)]$, or we need to be able to upper bound $\mathrm{Var}_p[f(X)]$, in order to use this bound.

Remark 1.17 (*). The Chebychev bound is rather pessimistic for the tails: if we set $\epsilon = t/\sqrt{n}$, then the bound is $O(t^{-2})$ for large t. If more is known about $f(X)$, tighter tail bounds are possible. For instance, in the bounded case $|f(X) - \mathbb{E}_p[f(X)]| \leq c$, a Hoeffding inequality implies

$$\mathbb{P}(|I_p^{(n)}(f) - \mathbb{E}_p[f(X)]| \geq \epsilon) \leq 2\exp\big(-2\epsilon^2 n/c^2\big),$$

and therefore for $\epsilon = t/\sqrt{n}$, we get $O(e^{-2t^2/c^2})$ bound.

1.4 About uniformly distributed pseudo-random numbers

During this course, we shall assume that we can access $(U_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$, a sequence of independent random variables uniformly distributed on the interval $(0,1)$. All algorithms are based on these random variables, and all theoretical results given below rely on this (rather strong) assumption.

In practice, when the algorithms are implemented on a computer, the sequence $(U_k)_{k\geq 1}$ are not going to be random, but *pseudo-random*. That is, $(U_k)_{k\geq 1}$ are in fact produced by a *deterministic* recursive algorithm with a finite state, a pseudo-random number generator (PRNG). Setting a *seed* of the algorithm means that we set the state variables of the algorithm to given initial values. The sequence $(U_k)_{k\geq 1}$ is entirely determined by the seed. However, a good PRNG approximates 'true randomness' rather well (is indistinguishable by a wide range of statistical tests).

It is essential to use a good PRNG for stochastic simulation, such as the *Mersenne twister* [13], which is the default PRNG for many environments, including Julia, Matlab, R and Python, and there are free implementations for most other environments. Remember also to seed your algorithm, if your implementation does not do that automatically.

1.5 Monte Carlo vs. other numerical integration methods

There are several other numerical integration methods, which may be used to calculate expectations instead of the Monte Carlo method. It is not straightforward to say which method works the best for a given problem, but here are some thoughts about the strengths and weaknesses of the Monte Carlo method:

+ Monte Carlo methods are generally applicable. For instance, the functions f and p need not be continuous, differentiable etc.
+ Monte Carlo is often easy to implement.
+ Monte Carlo can work well in multiple dimensions, where grid-based methods can be inefficient/inapplicable. This is supported by the "$O(n^{-1/2})$ rate of convergence" which is independent of the dimension.
− Even though the MC rate is usually $O(n^{-1/2}$, the constants involved may grow exponentially in dimension. (That is, MC does not generally 'beat the curse of dimensionality')
− Deterministic methods may have better rate of convergence than the Monte Carlo rate $n^{-1/2}$ (but may also deteriorate faster when dimension increases).
− Monte Carlo estimate is always random, so we never have guaranteed tolerance, but only statistical evidence (consistent confidence intervals at best).

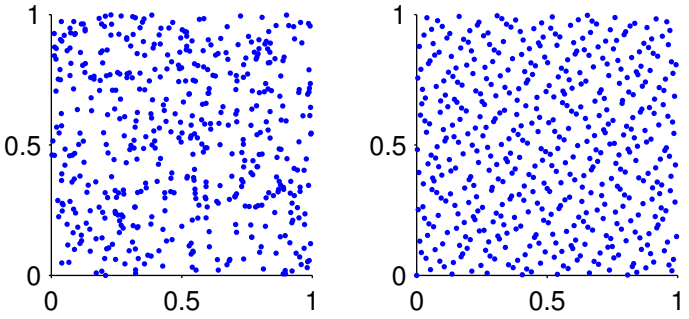


Figure 2: 500 points on $[0,1]^2$ which are (a) i.i.d. pseudo-random (b) from a low-discrepancy sequence (Halton).

Remark 1.18 (*). It may be good to know that there are also so-called *quasi Monte Carlo* methods, which may behave better in some applications (they often have a better rate of convergence). They are similar to Monte Carlo (based on averages), but instead of using i.i.d. (pseudo-)random variables $(U_k)_{k\geq 1}$, they use specifically designed 'low-discrepancy sequences' which 'fill' up the unit interval (or unit hypercube) in a deterministic way so that the points are scattered in a 'uniform' manner; see Figure 2.

We do not consider QMC methods further in the course, but note that QMC is also active in research, and succesful combinations of (randomised) QMC and MC have been suggested recently...

2 Variable transformation methods

Obviously, many interesting Monte Carlo problems assume that $(X_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim} p$, where p is not $\mathcal{U}(0,1)$. We need methods to transform $(U_k)_{k\geq 1} \sim U(0,1)$ into $(X_k)_{k\geq 1}$. In this section, we consider methods that

- Transform single $U \sim U(0,1)$ into a single $X \sim p$.
- Transform multiple $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} U(0,1)$ into single or multiple $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} p$, where $1 \leq m \leq n$.

2.1 Inverse distribution function method

Recall that the (cumulative) distribution function (c.d.f.) F of a random variable X is defined as $F(x) := \mathbb{P}(X \leq x)$ for all $x \in \mathbb{R}$. Recall also that if X has density p, then

$$F(x) = \int_{-\infty}^{x} p(t)\mathrm{d}t.$$

Theorem 2.1. *Assume $U \sim \mathcal{U}(0,1)$ and let $F : A \to (0,1)$ be a c.d.f. on an open interval[3] $A \subset \mathbb{R}$, which is continuous and strictly increasing, with inverse $F^{-1} : (0,1) \to A$. Then, $X := F^{-1}(U) \sim F$, that is, X has the c.d.f. F.*

3. May be infinite: (a,b), (a,∞), $(-\infty,b)$ or $\mathbb{R}$.

Proof. A direct calculation shows that $\mathbb{P}(X \leq x) = F(x)$ for all $x \in A$:

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(F^{-1}(U) \leq x) \\ &= \int_0^1 \mathbf{1}\left(F^{-1}(u) \leq x\right) \mathrm{d}u \\ &= \int_0^1 \mathbf{1}\left(u \leq F(x)\right) \mathrm{d}u = F(x). \qquad \square\end{aligned}$$

Example 2.2. If we want $X \sim \mathrm{Exp}(r)$, that is, $X \sim p(x)$ with

$$p(x) = r \exp(-rx)\mathbf{1}\left(x \geq 0\right),$$

then the c.d.f. is for $x > 0$

$$F(x) = \int_0^x r \exp(-rt)\mathrm{d}t = 1 - \exp(-rx),$$

with inverse $F^{-1}(u) = -\log(1-u)/r$. The algorithm is

(i) $U \sim \mathcal{U}(0,1)$

(ii) $X := -\log(U)/r$,

because if $U \sim \mathcal{U}(0,1)$, then also $1 - U \sim \mathcal{U}(0,1)$.

```
n = 1000          # Number of samples to simulate
u = rand(n)       # Vector of n independent U(0,1)
x = -log.(u)/2    # Vector of n indepedent Exp(2)
```

Theorem 2.3. *Assume p is a p.m.f. on $\mathbb{X} = \{x_1, x_2, \ldots\}$. Suppose $U \sim \mathcal{U}(0,1)$ and define the random variable*

$$K := \min\left\{k \geq 1 \,:\, \sum_{j=1}^{k} p(x_j) \geq U\right\}.$$

Then, $X := x_K$ has distribution p.

Proof. Define $F(k) := \sum_{j=1}^{k} p(x_j)$ with $F(0) := 0$, and note that

$$\mathbb{P}(K = k) = \mathbb{P}\big(F(k-1) < U \leq F(k)\big) = F(k) - F(k-1) = p(x_k),$$

and therefore $\mathbb{P}(X = x_k) = \mathbb{P}(K = k) = p(x_k)$. $\square$

Example 2.4. If $0 < \tilde{p} < 1$ and $\tilde{q} = 1 - \tilde{p}$, and we want to simulate $X \sim$ Geometric$(\tilde{p})$ then

$$p(k) = \tilde{p}\tilde{q}^{k-1}, \qquad k \in \mathbb{N} = \{1, 2, \ldots\}$$

with

$$F(k) = \sum_{i=1}^{k} p(i) = 1 - \tilde{q}^k.$$

Smallest k giving $1 - \tilde{q}^k \geq u$ is

$$k = \left\lceil \frac{\log(1-u)}{\log(\tilde{q})} \right\rceil$$

where $\lceil x \rceil$ rounds up (smallest integer not less than x).