Probability and Statistics

Data Science Engineering

Chapter 2: Random variables

Random Variables. Cumulative distribution function. Discrete Random variables. Probability function. Probability models. Expectation and Moments. Law of averages. Continuous random variables. Probability density. Continuous distributions. Normal distribution. Simulation of random variables.

---

1. Random variables

The language of events is sometimes cumbersome. The usual description of events uses a numerical description and events are identified by numbers. For example, in dice rolling the events are identified by the numerical output $\Omega = \{1, 2, 3, 4, 5, 6\}$. This leads to the definition of random variables.

Definition 1.1. A random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a function

$$X : \Omega \to \mathbb{R}$$

with the property that, for every $x \in \mathbb{R}$, the preimage $X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\}$ is an event in $\mathcal{A}$.

A random variable simply maps elements in the sample space to numbers, with the condition that intervals of the form $(-\infty, x]$ have preimages in the σ–algebra $\mathcal{A}$ of the probability space, where probabilities are defined.

Example 1.2. Let $\mathcal{A} = \{\emptyset, A, \bar{A}, \Omega\}$ be a Bernouilli algebra on a sample space Ω. The map

$$X : \Omega \to \mathbb{R}$$

defined by

$$X(\omega) = \begin{cases} 0 & x \notin A \\ 1 & x \in A \end{cases}$$

is a random variable: we have

$$X^{-1}(-\infty, x] = \begin{cases} \emptyset & x < 0 \\ \bar{A} & 0 \leq x < 1 \\ \Omega & x \geq 1 \end{cases}$$

Such random variables are called indicator random variables for the set A and are usually written as $\mathbb{1}_A$.

Remark 1.3. Usually capital letters like $X, Y, Z, \ldots$ are used to denote random variables in the probability setting. For a subset $A \subset \mathbb{R}$ we use the shorthand $X \in A$ for $X^{-1}(A)$. Thus, we write $\mathbb{P}(X < 0), \mathbb{P}(0 < X \leq 1)$ or $\mathbb{P}(X = 2)$ instead of $\mathbb{P}(X^{-1}(-\infty, 0)), \mathbb{P}(X^{-1}((0, 1])$, $\mathbb{P}(X^{-1}(2))$ or $\mathbb{P}(\{\omega \in \Omega : X(\omega) = 2\})$. □

The purpose of the definition of random variables is to translate the probability function from the probability space to $\mathbb{R}$. This is generally accomplished by the cumulative distribution function.

Definition 1.4. Let X be a random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The *(cumulative) distribution function (CDF)* of X is

$$F_X : \mathbb{R} \to \mathbb{R}$$

defined as

$$F_X(x) = \mathbb{P}(X^{-1}((-\infty, x]) = \mathbb{P}(X \leq x).$$

Example 1.5. Let $\mathbb{1}_A$ be the indicator function of the event A in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(A) = p$. The cumulative function of $\mathbb{1}_A$ is

$$F_{\mathbb{1}_A}(x) = \begin{cases} 0 & x < 0 \\ 1 - p & 0 \leq x < 1 \\ 1 & x \geq 1. \end{cases}$$

□

Distribution functions are identified by the following properties:

Proposition 1.6. *Let F_X be the distribution function of a random variable X. Then*

(1) $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to \infty} F_X(x) = 1$*;*

(2) F_X is non decreasing: if $x < y$, then $F_X(x) \leq F_X(y)$;

(3) F_X is right–continuous: $\lim_{x \downarrow a} F_X(x) = F_X(a)$.

The above properties identify the class of real functions which are distribution functions of some random variable.

Probabilities of general events $X \in A$ can be expressed in terms of the distribution function (sometimes in a not simple way). Some examples are

Proposition 1.7. *Let X be a random variable in a probability space with distribution function F_X. Then*

(i) For $a, b \in \mathbb{R}$, $a < b$, we have $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.

(ii) For $a \in \mathbb{R}$, we have $\mathbb{P}(X = a) = F_X(a) - \lim_{x \uparrow a} F_X(x)$.

2. Discrete Random variables

A random variable is *discrete* if it takes a countable number of values.

Definition 2.1. A random variable X on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is discrete if $X(\Omega)$ is countable.

An equivalent definition is that the distribution function F_X of X is a right-continuous step function: there is a countable set $\{x_1, x_2, \ldots\} \subset \mathbb{R}$ of real numbers such that F_X is constant outside this set, and F_X has jump discontinuities at the points in this set.

For discrete random variables it is usually preferred to identify the probability distribution by the so–called *probability function* instead of the cumulative distribution function.

Definition 2.2. The probability function of a discrete random variable is

$$P_X : \mathbb{R} \to [0, 1]$$

defined as $P_X(x) = \mathbb{P}(X = x)$.

Proposition 2.3. *Let P_X be the probability function of a random variable X. Then*

(1) $P_X(x) \geq 0$ for all $x \in X(\Omega)$;

(2) $\sum_{x \in X(\Omega)} P_X(x) = 1$.

Example 2.4. By throwing a dice we naturally identify the outcomes with a discrete random variable X which takes values in $\{1, 2, 3, 4, 5, 6\}$. Its probability function is $\mathbb{P}(X = i) = 1/6$ for $1 \leq i \leq 6$. This shows how natural is the use of random variables in many situations. □

Example 2.5. An indicator function $\mathbb{1}_A$ is obviously a discrete random variable taking values in $\{0, 1\}$ with probability function $\mathbb{P}(\mathbb{1}_A = 0) = 1 - \mathbb{P}(A)$ and $\mathbb{P}(\mathbb{1}_A = 1) = \mathbb{P}(A)$.

Actually every discrete random variable can be written as a linear combination of indicator variables: if A_i is the event that $X = x_i$ for each $x_i \in X(\Omega)$ then $X = \sum_i x_i \mathbb{1}_{A_i}$. □

Example 2.6. We toss a coin till Heads shows up. The random variable X which counts the number of tosses takes values in $\mathbb{N}$, a countable infinite set. Its probability function is $\mathbb{P}(X = i) = (1/2)^i$ (and certainly, $\sum_{i \geq 1} \mathbb{P}(X = i) = 1$). □

Given a random variable, we can obtain new ones by using functions.

Proposition 2.7 (Functions of random variables)**.** *Let X be a discrete random variable on a probability space and let $g : \mathbb{R} \to \mathbb{R}$ be a function such that the preimage of each interval $(-\infty, x]$, $x \in \mathbb{R}$ belongs to the Borel σ–algebra (the class of such functions is called 'measurable' and it includes continuous functions).*

Then the composition $Y = g(X)$ is a discrete random variable on the same probability space.

3. Discrete Probability models

We now come to one important part of the course. A large part of probability theory is developed upon simple models which are used once an again to build more complex ones. We next describe some of the most important ones. They are identified both as a model and give the name to their probability distributions and, by abuse of language, to the random variables.

3.1. **Bernoulli model.**

The Bernoulli model is the simplest one, associated to the Bernoulli algebra and to the indicator functions. In this model we are only interested in a single event $A \subset \Omega$ and want to determine if this event occurs. Formally,

Definition 3.1. A random variable X has Bernoulli distribution with parameter $p \in [0, 1]$ if $X(\Omega) = \{0, 1\}$ and

$$\mathbb{P}(X = 0) = 1 - p$$
$$\mathbb{P}(X = 1) = p.$$

Equivalently, $X = \mathbb{1}_A$ for some $A \subset \Omega$ with $\mathbb{P}(A) = p$. We write $X \sim \mathrm{Be}(p)$.

There is not much to say about the Bernoulli distribution except that it is the building block of many more complicated models.

3.2. **Binomial model.**

We again focus on a single event A in the sample space, but now we make n *independent* repetitions of the experience associated to the probability space, and we are interested in counting how many times the event occurs in these n repetitions. By the independence, any possible output of the n experiments in which there are k occurrences of A has probability $p^k q^{n-k}$, where $p = \mathbb{P}(A)$ and $q = 1 - p$. There $\binom{n}{k}$ outputs with precisely k occurrences of A. This gives the probability distribution of our random variable.

Definition 3.2. A random variable X has the Binomial distribution with parameters n and p if

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k},\ k = 0, 1, \ldots, n.$$

We write $X \sim \mathrm{Bin}(n, p)$.

Using the Newton Binomial, one can check it is indeed a probability function:

$$\sum_{k=0}^{n} \mathbb{P}(X=k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p+(1-p))^n = 1\ .$$

Proposition 3.3. *Let $X \sim Bin(n,p)$. Then*

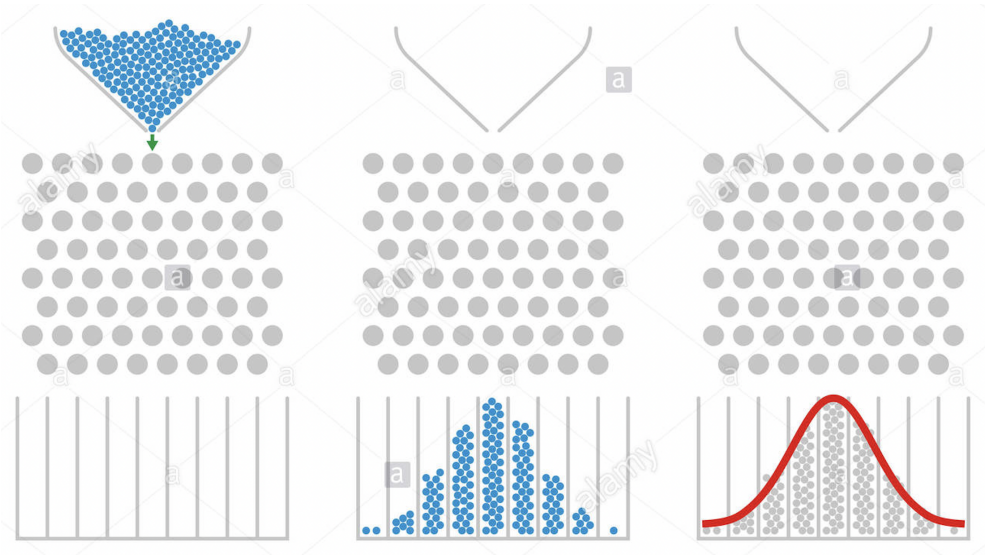
$$\mathbb{P}(X=k) \leq \mathbb{P}(X=k+1),\ 0 \leq k \leq (n+1)p-1$$

and

$$\mathbb{P}(X=k) \geq \mathbb{P}(X=k+1),\ (n+1)p-1 < k \leq n.$$

In other words, the sequence $\mathbb{P}(X=0), \mathbb{P}(X=1), \ldots, \mathbb{P}(X=n)$ is unimodal with a maximum at $\lfloor (n+1)/p \rfloor$.

Remark 3.4 (Galton Board)**.** The Galton Board is a physical experiment which illustrates the shape of the probability function of the Binomial. A number of marbles are thrown (each one independently) into a grid of posts. Everytime a marbles hits a post, it bounces right or left, each with probability $1/2$. The random variable X that counts the number of tight bounces, which is the final position of the marble, is distributed as a Binomial $\text{Bin}(n,p)$ where n is the number of rows of the grid.



The Binomial model is a quite fundamental one. Among other interesting properties, we have.

3.3. **Poisson Model.** The values of the probability function of a Binomial variable are somewhat cumbersome to compute. A useful simplification is the one that can be obtained in the limit as n grows but np converges to a constant. More precisely, if $X_n \sim \text{Bin}(n, p_n)$

with $\lim_{n\to\infty} np_n = \lambda$, then the probability function of X_n has a limit which can be written in a simpler way:

$$\begin{aligned}\mathbb{P}(X_n = k) &= \binom{n}{k} p_n^k q_n^{n-k} \\ &\sim \frac{1}{k!}(np_n)^k(1-p_n)^{-k}(1-p_n)^n \\ &\sim \frac{1}{k!}\lambda^k e^{-\lambda},\end{aligned}$$

where in the last line we have used that

$$\lim_{n\to\infty}(1-\lambda/n)^n = e^{-\lambda}.$$

This leads to the following definition.

Definition 3.5. A random variable X has the Poisson distribution with parameter $\lambda > 0$ if

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, 2, \cdots$$

and we write $X \sim \mathrm{Po}(\lambda)$.

Using the taylor expansion of the exponential function, one can check it is indeed a probability function:

$$\sum_{k=\geq 0} \mathbb{P}(X = k) = e^{-\lambda}\sum_{k\geq 0}\frac{\lambda^k}{k!} = 1\ .$$

In fact, the expression of the Poisson probability function approximates quite well the Binomial one even for small values of n. Here we present an example for $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Po}(np)$ when $n = 30$ and $p = 0.1$:

k	$\mathbb{P}(X = k)$	$\mathbb{P}(Y = k)$
0	0.042	0.049
1	0.141	0.149
2	0.227	0.224
3	0.236	0.224
4	0.177	0.168

Examples of random quantities that follow a Poisson distribution are the number of radioactive particles, the number of phone calls, the number of impacts of drop rain. All these are associated to large number of trials with small probability of success which can be imagined to be independent.

3.4. **Geometric model.** We again repeat independently an experience associated to a Bernoulli algebra but now we count the number of repetitions till the first occurrence of A (the waiting times till the first 'success'). By independence, the probability of that the first occurrence appears at the k-th trial is $q^{k-1}p$, where $q = 1 - p$.

Definition 3.6. A random variable X has the geometric distribution with parameter $p \in (0,1]$ if

$$\mathbb{P}(X = k) = q^{k-1}p,\ k = 1, 2, 3, \dots$$

We write $X \sim \text{Geom}(p)$.

The name of the distribution comes from the fact that the sum of probabilities, which must add up to one, is the geometric series: as $p \in (0,1]$,

$$\sum_{k\geq 1} \mathbb{P}(X = k) = p\sum_{k\geq 1}(1-p)^{k-1} = p\sum_{k\geq 0}(1-p)^k = \frac{p}{1-(1-p)} = 1.$$

It is important to note that the upper tail of the geometric has a nice expression:

$$\mathbb{P}(X > k) = q^k,\ k = 1, 2, 3, \dots \tag{1}$$

Be careful: a slight variation of the geometric distribution $Y = X - 1$, which counts the number of 'failures' before the first 'success' is usually also called geometric, with probability distribution

$$\mathbb{P}(Y = k) = q^k p,\ k = 0, 1, 2, \dots$$

The geometric distribution has a characteristic property, the 'lack of memory'. If we know that up to the k–th repetition we have no success, then the probability of having to wait at least s additional repetitions is the same as from the start:

$$\mathbb{P}(X > r + s | X > r) = \frac{\mathbb{P}(X > r + s, X > s)}{\mathbb{P}(X > r)} = \frac{q^{r+s}}{q^r} = \mathbb{P}(X > s).$$

where we have used (1). In other words, if we are waiting for Heads in coin tossing, the fact that we have waited 10^3 tosses without seeing the event does not mean that Heads are approaching faster in the future.

Proposition 3.7. *Let X be a discrete random variable taking values in the positive integers with $\mathbb{P}(X = 1) = p$. If X has the 'lack of memory' property then $X \sim Geom(p)$.*

3.5. **Negative Binomial model.** If instead of waiting for the first 'success' as in the geometric distribution, one waits till the appearance of the r–th 'success', then the probability of waiting up to k repetitions is that of having $r-1$ successes in the first $k-1$ trials (a binomial distribution) and then having the r–th one in the k–th trial.

Definition 3.8. A random variable X has the negative binomial distribution with parameter p and r if

$$\mathbb{P}(X = k) = \binom{k-1}{r-1} p^r q^{k-r},\ k = r, r+1, \dots$$

We write $X \sim \text{NegBin}(p, r)$.

Of course, when $r = 1$ we obtain the geometric distribution.

3.6. **Hypergeometric model.** In most of the above examples we consider independent repetition of trials of a simple Bernoulli experience. When sampling without replacement from a population, the iteration of the sampling is not bound to independence, because the result in the k–trial affects the probability distribution in the $(k+1)$–th trial. The typical example is drawing balls from an urn without replacement (as opposite to drawing them with replacement, when the trials are independent).

In the hypergeometric model, we extract samples of size r out of a population of size n which has m individuals of type 1 and $n-m$ individuals of type 2. We are then interested in counting the number of individuals of type 1 in the sample. This leads to the following definition.

Definition 3.9. A random variable X has the hypergeometric distribution with parameters n, m and r if

$$\mathbb{P}(X=k) = \frac{\binom{m}{k}\binom{n-m}{r-k}}{\binom{n}{r}},\ k=0,1,2,\dots$$

We write $X \sim \mathrm{HypGeom}(n,m,r)$.

The range of values of k for which the above definition is meaningful is

$$\max\{0, r-(n-m)\} \le k \le \min\{r,m\}.$$

In order to not bother about these boundary values we may adopt the (reasonable) convention that a binomial coefficient $\binom{a}{b}$ equals zero whenever $b>a$ or $b<0$.

The name of the distribution comes from the fact that the sum of probabilities is an hypergeometric (finite) series.

3.7. **Uniform model.** The basic distribution we have repeatedly seen in a finite sample space $\Omega = \{1,2,\dots,n\}$ is the uniform one, where each experiment gets the same probability.

Definition 3.10. A random variable X has the uniform distribution with parameter n if

$$\mathbb{P}(X=k) = \frac{1}{n},\ k=1,2,\dots,n.$$

We write $X \sim \mathrm{U}(n)$.

We simply note that a discrete random variable can not have the uniform distribution on an infinite countable set, say $\Omega = \mathbb{N}$. This would lead to $\mathbb{P}(X=k)=0$ for all k and then, by σ–additivity, $\mathbb{P}(\mathbb{N}) = \sum_{x\in\mathbb{N}} \mathbb{P}(X=x) = 0$, contradicting the first axiom of a probability measure.

4. Expectation and moments

Expectation is a central concept in probability and statistics. The mean of a sequence $x = (x_1, \dots, x_n)$ of real numbers is

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}.$$

If there are repetitions in the sequence then we can collect the values y_i which are repeated n_i times and rewrite the mean as

$$\bar{x} = \sum_i \frac{n_i}{n} y_i.$$

The expectation of a random variable mimics the definition and the spirit of the mean of a sequence of numbers, by substituting the relative frequencies n_i/n by the probabilities:

Definition 4.1 (Expectation)**.** The *expectation* (or *expected value*, *mean*) of a discrete random variable X is

$$\mu_X = \mathbb{E}(X) = \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(X = x),$$

whenever the sum is absolutely convergent.

In other words, the expectation is the sum of values of the random variable weighted by their probabilities. We will see later on that $\mathbb{E}(X)$ is the ''best constant approximation' to X, a clear intuitive fact. Of course the expectation can be large because X takes very large values even with small probabilities, so the expectation may be a misleading representative of a random variable. A large amount of probability and statistics is devoted to clarify the above statement.

The caution in the definition about the convergence of the sum is not superfluous: there are random variables which do not have expectation, although sometimes the value ∞ is accepted.

Example 4.2. Let X be a random variable taking values on the positive integers with probability

$$\mathbb{P}(X = k) = \frac{6}{\pi k^2}.$$

One can check (is a famous problem in the history of mathematics) that $\sum_{k \geq 1} \mathbb{P}(X = k) = 1$. However $\sum_k k\mathbb{P}(X = k) = (6/\pi^2) \sum_{k \geq 1} 1/k$ is the harmonic series which diverges. □

The expectation of the basic distributions we have seen so far is as follows.

Proposition 4.3. *We have*

Distribution	*Expectation*
$X \sim Be(p)$	p
$X \sim Bin(n,p)$	np
$X \sim Po(\lambda)$	λ
$X \sim Geom(p)$	$1/p$
$X \sim NegBin(r,p)$	r/p
$X \sim HypGeom(n,m,r)$	$rm/n.$

One important property of expectation is *linearity*. For this it is meaningful to have a look on the distribution of the sum of two discrete random variables.

Definition 4.4 (Sum of random variables)**.** Let X, Y be two discrete random variables on the same probability space. The random variable $Z = X+Y$, defined as $Z(\omega) = X(\omega)+Y(\omega)$ for each $\omega \in \Omega$ has probability function

$$\mathbb{P}(Z = k) = \sum_i \mathbb{P}(X = i, Y = k - i),$$

where $(X = i, Y = k - i)$ is shorthand for $\{X = i\} \cap \{Y = k - i\}$.

Proposition 4.5. *Let X, Y be two random variables on the same probability space with $|\mathbb{E}(X)|, |\mathbb{E}(Y)| < \infty$. Then*

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Moreover, for each $\lambda \in \mathbb{R}$,

$$\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X).$$

Proposition 4.5 is particularly useful. For example, it provides a simple way to compute the expectation of Binomial and Negative Binomial distributions, as both are sums of simple random variables.

Example 4.6. Recall that if $X \sim \text{NegBin}(r, p)$, then

$$X = Z_1 + \cdots + Z_r$$

where $Z_i \sim \text{Geom}(p)$. By linearity and Proposition 4.3

$$\mathbb{E}(X) = \mathbb{E}(Z_1 + \cdots + Z_r) = \mathbb{E}(Z_1) + \cdots + \mathbb{E}(Z_r) = r/p.$$

A second property that must be highlighted is that the expectation can be used to bound the probability a random variable is 'too large'.

Theorem 4.7 (Markov inequality)**.** *Let X be a discrete non-negative random variable with $|\mathbb{E}(X)| < \infty$. Then, for each $a \geq 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

The last important property of the expectation is related to functions of random variables. Given g a measurable function and X a random variable, one can sometimes obtain an explicit expression for the distribution of $Y = g(X)$. However the following result, usually called the theorem of expectation or the formula of change of variables for expectation, is often useful.

Theorem 4.8. *Let X be a discrete random variable on a probability space and let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function. Then the expectation of $Y = g(X)$ satisfies*

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \sum_{x \in X(\Omega)} g(x)\mathbb{P}(X = x).$$

As commented, the expectation is the best constant that approximates X, but fails short to capture many properties of their probability function.

Example 4.9. Let $X \sim U(3)$ and $Y \sim \text{Geom}(1/2)$. Then

$$\mathbb{E}(X) = 2 = \mathbb{E}(Y)$$

but their probability functions look very different.

To refine the information given by the expectation, we generalize it to the notion of moments.

Definition 4.10 (Moments)**.** Let X be a discrete random variable. The *k–th moment* of X is

$$\mathbb{E}(X^k) = \sum_{x \in X(\Omega)} x^k \mathbb{P}(X = x).$$

whenever the sum is absolutely convergent.

The *k–th central moment* of X is

$$\mathbb{E}((X - \mathbb{E}(X))^k) = \sum_{x \in X(\Omega)} (x - \mathbb{E}(X))^k \mathbb{P}(X = x).$$

The *variance* of a discrete random variable X is the second central moment, that is

$$\sigma_X^2 = \text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

The *standard deviation* of X is

$$\sigma_X = +\sqrt{Var(X)}.$$

The variance of X measures the typical deviation of X with respect to the expected value. The smaller the variance, the more concentrated are the values of X around its mean (less probability that it takes values far from its mean).

The following form is often more useful to compute the variance.

Lemma 4.11. *$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.*

The variance of the basic distributions we have seen so far is as follows:

Proposition 4.12. *We have*

Distribution	*Variance*
$X \sim Be(p)$	pq
$X \sim Bin(n,p)$	npq
$X \sim Po(\lambda)$	λ
$X \sim Geom(p)$	q/p^2
$X \sim NegBin(r,p)$	rq/p^2
$X \sim HypGeom(n,m,r)$	$rm(n-m)(n-r)/n^2(n-1)$
$X \sim U(n)$	$(n^2-1)/12$.

The variance is a quadratic operator, this means that for any $\lambda \in \mathbb{R}$

$$\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X).$$

However it is not true that the variance of a sum of random variables is the sum of their respective variances. This is true if the variables are independent, as we will see in next chapter.

Using the variance one can refine Markov inequality as follows:

Theorem 4.13 (Chebyshev inequality)**.** *Let X be a discrete random variable with $\mathbb{E}(X^2) < \infty$. Then, for each $a \geq 0$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{Var(X)}{a^2}.$$

One particular consequence is that, if $\mathrm{Var}(X) = 0$ then $X = \mathbb{E}(X)$ with probability one, that is, X is a constant.

Example 4.14. Let X be a random variable with uniform distribution $U(n)$. Then

$$\mathbb{E}(X) = \frac{n+1}{2} \qquad \mathrm{Var}(X) = \frac{n^2-1}{12}.$$

Chebyshev inequality gives

$$\mathbb{P}(|X - (n+1)/2| \geq k) \leq \frac{n^2-1}{12k^2}$$

while the actual value is

$$\mathbb{P}(|X-(n+1)/2| \geq k) = 1-\mathbb{P}\left(\frac{n-2k+1}{2} < X < \frac{n+2k+1}{2}\right) = 1-\frac{2k-1}{n} = \frac{n-2k+1}{n}.$$

For example, for $n = 13$ the two values are

k	1	2	3	4	5	6
Chebyshev	14	7/2	14/9	7/8	14/25	7/18
Actual value	12/13	10/13	8/13	6/13	4/13	2/13

which shows that the bounds can be rather poor. However, the Chebyshev estimation is valid for any probability distribution and it can be tight. □

Finally, we introduce the exponential moments.

Definition 4.15 (MGF)**.** The *moment generating function* of a discrete random variable X is

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{x \in X(\Omega)} e^{tx} \mathbb{P}(X = x)$$

if the sum is convergent in some interval around $t = 0$.

By using the Taylor expansion of the exponential and the linearity of the expectation, we can write

$$M_X(t) = \sum_{k \geq 0} \frac{\mathbb{E}(X^k) t^k}{k!}, \tag{2}$$

and the MGF is only defined if *all* moments are finite.

Example 4.16. Let $X \sim \mathrm{Be}(p)$, then

$$M_X(t) = e^{t \cdot 0} \mathbb{P}(X = 0) + e^{t \cdot 1} \mathbb{P}(X = 1) = 1 - p - pe^t.$$

Let $Y \sim \mathrm{Po}(\lambda)$, then

$$M_Y(t) = \sum_{k \geq 0} e^{tk} \mathbb{P}(Y = k) = e^{-\lambda} \sum_{k \geq 0} \frac{(\lambda e^t)^k}{k!} = e^{\lambda(e^t - 1)}.$$

We give a table with the MGF of the random variables we have introduced

Distribution	MGF
$X \sim \mathrm{Be}(p)$	$1 - p + pe^t$
$X \sim \mathrm{Bin}(n, p)$	$(1 - p + pe^t)^n$
$X \sim \mathrm{Po}(\lambda)$	$e^{\lambda(e^t - 1)}$
$X \sim \mathrm{Geom}(p)$	$pe^t/(1 - (1 - p)e^t)$
$X \sim \mathrm{NegBin}(r, p)$	$(pe^t/(1 - (1 - p)e^t))^r$.

From the previous table, we observe that the MGF of a sum of independent random variables from the original MGF of the random variables. We will make this explicit in coming sections.

A natural question is probability is whether two random variables X and Y with the same moments, have the same distribution. As the MGF encodes all moments by (2), it is equivalent to say whether the MGF uniquely determined the distribution.

Theorem 4.17. *Let X and Y random variables. Suppose that there exists $\delta > 0$ such that for all $t \in (-\delta, \delta)$, (i) $|M_X(t)|, |M_Y(t)| < \infty$ and $M_X(t) = M_Y(t)$, then X and Y have the same distribution.*

The proof of this theorem is out of the scope of this course, but we can use it as a tool to prove Proposition **??**.

5. Continuous random variables

Continuous random variables are roughly identified by the fact that the distribution function is continuous. As it happens, this requirement is not enough to identify the class of continuous random variables. Instead we ask the stronger requirement that the distribution function can be obtained by integration of a density function.

Definition 5.1 (Continuous random variable)**.** A random variable is continuous if there is a function $f : \mathbb{R} \to \mathbb{R}$ such that, for each $x \in \mathbb{R}$,

$$F_X(x) = \int_{-\infty}^{x} f(t)dt.$$

The function f is called the *probability density function* of X and usually denoted by f_X.

By the Fundamental Theorem of Calculus (FTC), we have

$$f_X(x) = F'_X(x),$$

at each point x where F_X has a derivative. It is a result from Calculus that a continuous random variable has a continuous distribution function. This in particular shows that

$$\mathbb{P}(X = x) = F_X(x) - \lim_{t \uparrow x} F_X(t) = 0$$

for all x. In particular, $\mathbb{P}(X \in A) = 0$ for every countable set $A \subset \mathbb{R}$.

By the Taylor expansion, $F_X(a+h) = F_X(a) + F'_X(a)h + O(h^2) = F_X(a) + f_X(a)h + O(h^2)$, thus

$$\mathbb{P}(a < X < a + h) = F_X(a + h) - F_X(a) = f_X(a)h + O(h^2),$$

which can be written, for small h, as

$$\frac{\mathbb{P}(a < X < a + h)}{h} \approx f_X(a),$$

which explains the name 'probability density' for f_X. So, large values of f_X indicate large probability of being locally around the argument. In this sense one can interpret continuous random variables as limit versions of discrete ones.

The probability that X lies in a set A can be obtained from the density function as

$$\mathbb{P}(X \in A) = \int_A f_X(t)dt.$$

In particular,

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b) = \int_a^b f_X(t)dt.$$

The density function of a continuous random variable has the following properties:

Proposition 5.2. *Let f_X be the density function of continuous random variable X. The following holds:*

(1) $f_X(x) \geq 0$ for all $x \in \mathbb{R}$, and

(2) $\int_{-\infty}^{\infty} f_X(t)dt = 1$.

The above properties characterize the class of functions which are density functions of some continuous random variable.

6. Probability models

For continuous models it will be useful the following notation, for any set $S \subseteq \mathbb{R}$, we write

$$\mathbb{1}_S(x) = \mathbb{1}_{x \in S}$$

for the indicator function that $x \in S$.

The following are some of the most important continuous distributions.

6.1. (Continuous) Uniform distribution.

We choose a random point in an interval $[a, b]$. Note that this is the same as choosing a random point in (a, b), as point probabilities are zero. All intervals with the same length have the same probability. This leads to:

Definition 6.1. A random variable X has the *(continuous) uniform distribution* on a finite interval (a, b) if its density function is

$$f_X(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x) = \begin{cases} 1/(b-a) & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases}.$$

The probability distribution function of X is,

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a}x & a \leq x < b \\ 1 & x \geq b. \end{cases}$$

we write $X \sim \mathrm{U}([a, b])$.

The (continuous) uniform model is the continuous version of the (discrete) uniform model. This is described as follows Let $Y_n \sim \mathrm{U}(n)$, consider $X_n = Y_n/n$ and let $X \sim \mathrm{U}([0, 1])$. For $x \in [0, 1]$, F_{X_n} satisfies

$$\lim_{n\to\infty} F_{X_n}(x) = \mathbb{P}(X_n \leq x) = \mathbb{P}(Y_n \leq xn) \approx \frac{xn}{n} = x = F_X(x)$$

Thus, F_{X_n} 'tends' to F_X. We need to careful with what 'tends' mean for a sequence of functions, here it is pointwise convergence.

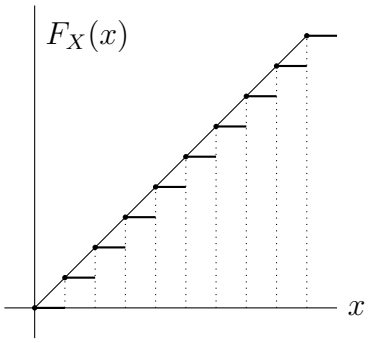


FIGURE 1. The continuous uniform distribution as a limit of the discrete one.

6.2. **Exponential distribution.** The exponential distribution can be seen as the limiting distribution of a geometric one. The model corresponds to the time a random event occurs when the probability of occurring in a small interval of length ℓ is a Bernoulli variable with probability proportional to ℓ, independent of the occurrence in other disjoint intervals.

Definition 6.2. A random variable X has the *exponential distribution* with parameter $\lambda > 0$ if its density function is

$$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{(0,\infty)}(x)$$

The probability distribution function of X is, $F_X(x) = 0$ for $x \leq 0$, and

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \quad \text{for } x > 0.$$

we write $X \sim \text{Exp}(\lambda)$.

If $X_n \sim \text{Geom}(\lambda/n)$, then

$$\mathbb{P}(X_n \leq kn) = 1 - (1 - \lambda/n)^{kn} \to 1 - e^{-\lambda k},$$

This illustrates how the exponential distribution can be seen as the limit of a Geometric distribution. In particular, the exponential distribution also has the memoryless property:

Proposition 6.3. *Let X be a random variable with the exponential distribution. Then, for $s, t \geq 0$*

$$\mathbb{P}(X > t + s | X > s) = \mathbb{P}(X > t).$$

It can be shown that a continuous random variable taking values in $(0, \infty)$ with the memoryless property has an exponential distribution.

Remark 6.4. An interesting connection with the Poisson distribution is worth mentioning. Suppose that the number of events in a time interval $[0, t]$ is a random variable X_t with a Poisson law $X_t \sim \text{Po}(\lambda t)$ and moreover, for every $t_1 \leq t_2 \leq t_3 \leq t_4$, the number of events happening in $[t_1, t_2)$ is independent from the number of events happening in $[t_3, t_4)$. This is

known as a *Poisson Point Process (PPP)* and there are many models that are based on them. The waiting time for the first event to happen is a random variable T with distribution

$$\mathbb{P}(T \leq t) = \mathbb{P}(X_t \geq 1) = 1 - \mathbb{P}(X_t = 0) = 1 - e^{-\lambda t},$$

so that T follows an Exponential distribution $T \sim \text{Exp}(\lambda)$.

6.3. **Normal distribution.** The Normal distribution is among the most important ones in Probability and Statistics. One of the reasons is that it can be seen as the limiting distribution of the Binomial distribution. As such, it models the sum of (infinitely many) independent Bernoulli variables with. It occurs in random phenomena which are the sum of many independent inputs. It was observed by Gauss as the law of errors in measurements, and for that reason it is also known as the *Gaussian distribution*.

Definition 6.5. A random variable X has the *normal distribution* with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ if its density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We write $X \sim \text{N}(\mu, \sigma^2)$.

For $\mu = 0$ and $\sigma^2 = 1$ the corresponding normal distribution $\text{N}(0, 1)$ is called the *standard normal distribution* and has a more transparent density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

This is a symmetric function with respect to the origin and has particularly small tails.

The density function of a normal distribution does not have a primitive which can be expressed as a finite combination of elementary functions. For $X \sim \text{N}(0, 1)$ we denote

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

The function $\Phi(x)$ is closely related to the *error function*, that will appear often in these studies, which is defined as a renormalized version of it, only integrating in non-negative values:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\, dt.$$

As no explicit expression exists, the values of the normal distribution were historically recorded in tables. These values are accessible through most standard mathematical softwares, particularly in Python and R.

Nevertheless, we can obtain good bounds on the tails of the distribution. Let $X \sim \text{N}(0, 1)$. Then, for $x > 0$,

$$1 - \Phi(x) = \mathbb{P}(X > x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^2/2} dt < \frac{1}{x\sqrt{2\pi}} e^{-x^2}.$$

One can reduce a normal distribution $\mathrm{N}(\mu, \sigma^2)$ to a standard one by a linear change of variables.

Proposition 6.6. *If $X \sim N(\mu, \sigma^2)$ then $Z = (X - \mu)/\sigma$ is a standard normal, that is $Z \sim N(0, 1)$.*

In the celebrated treatise by Laplace on probability one can already find what is known as the De Moivre-Laplace theorem which states that the binomial distribution tends to the Normal one, as n grows large.

Theorem 6.7 (De Moivre-Laplace)**.** *For n large, let $X \sim Bin(n, p)$ and $Y \sim N(\mu = np, \sigma^2 = npq)$. Then for any k 'close to' np we have*

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k} \sim \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(k-\mu)^2}{2\sigma^2}} = f_Y(k).$$

The de Moivre-Laplace Theorem is the first form of the celebrated Central Limit Theorem, one of the central results in Probability and Statistics. The general form of this basic result will be discussed later on in this course.

Example 6.8. We toss a coin 400 times. The probability that we obtain more than 220 heads is

$$\mathbb{P}(X \geq 220) \approx \mathbb{P}(Y \geq 220) = \mathbb{P}(Z \geq 2) \approx 1 - 0.9772 = 0.0228$$

where $X \sim \mathrm{Bin}(400, 1/2)$ counts the number of heads in 400 tosses, $Y \sim \mathbb{N}(200, 100)$ and $Z = (Y - 200)/10$ is a standard normal.

6.4. **Gamma distribution.** The *Euler Gamma function* $\Gamma(x)$ is defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \text{ for } x > 0.$$

The integral is convergent for every positive x and $\Gamma(1) = 1$. By integrating by parts it is readily seen that

$$\Gamma(x + 1) = x\Gamma(x),$$

so that for integer values we have

$$\Gamma(n + 1) = n!$$

The Gamma function can thus be seen as a continuous interpolation of the factorial. The function is used to define a family of probability distributions which extend the exponential distribution (adding a polynomial correction to it) and has many applications in probability, statistics and engineering.

Definition 6.9. A random variable X has the *Gamma distribution* with parameters $\alpha > 0$ (the *shape*) and $\lambda > 0$ (the *rate*) if its density function is $f_X(x) = 0$ for $x \leq 0$ and

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \text{ for } x > 0.$$

We write $X \sim \text{Gamma}(\alpha, \lambda)$.

Sometimes the parametrization used is $\theta = 1/\lambda$ instead of λ, and we call θ the *scale*.

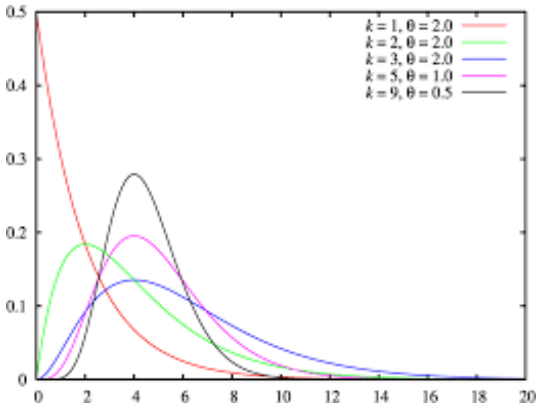


FIGURE 2. Some examples of the density function of a gamma distribution (Wikipedia), here $k = \alpha$ is the shape and θ is the scale.

The exponential distribution is $\text{Gamma}(1, \lambda)$. For integer values of α we will later show that $\text{Gamma}(k, \lambda)$ can be seen to be the independent sum of k exponential random variables $\text{Exp}(\lambda)$. As such, it models the waiting time for the k-th appearance of an event, when the events follow a Poisson Point Process (see end of Section 6.2). Therefore, $\text{Gamma}(k, \lambda)$ can be seen as the continuous version of the negative binomial distribution.

6.5. **Beta distribution.** The *Euler Beta function* is defined as

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt, \text{ for } x, y > 0.$$

The integral is convergent for positive values of x and y and it can be shown that

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Therefore, for integer values k and l,

$$B(k+1, l+1) = \frac{k!\, l!}{(k+l)!} = \frac{1}{\binom{k+l}{k}}.$$

The Beta function gives raise to the Beta distribution.

Definition 6.10. A random variable X has the *Beta distribution* with parameters $\alpha > 0$ and $\beta > 0$ if its density function is

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \text{ for } x \in (0, 1).$$

We write $X \sim \text{Beta}(\alpha, \beta)$.

7. Expectation and Moments

Expectation and general moments can also be defined for continuous random variables, by replacing sums by integrals, and having the same properties as the ones proved in Section 4.

Definition 7.1 (Expectation and Moments)**.** Let X be a continuous random variable with density f_X. In what follows we always assume that the corresponding integral is absolutely convergent. The *expectation* of X is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

The *k–th moment* of the random variable X is

$$\mathbb{E}(X^k) = \int_{-\infty}^{\infty} x^k f_X(x)\, dx,$$

and the *central k–th moment*

$$\mathbb{E}((X - \mathbb{E}(X))^k) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^k f_X(x)\, dx.$$

The second central moment is the *variance* of X,

$$\sigma_X^2 = \mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Finally, the *moment generating function* of X is

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx.$$

The following table summarizes the expectation and variance of the most common distributions.

Proposition 7.2. *We have*

Distribution	*Mean Value*	*Variance*	*MGF*
$X \sim U([a,b])$	$(a+b)/2$	$(b-a)^2/12$	$(e^{tb} - e^{ta})/t(b-a)$
$X \sim Exp(\lambda)$	$1/\lambda$	$1/\lambda^2$	$\lambda/(\lambda - t)$
$X \sim N(\mu, \sigma^2)$	μ	σ^2	$e^{\mu t + \sigma^2 t^2/2}$
$X \sim Gamma(\alpha, \lambda)$	α/λ	α/λ^2	$(1 - t/\lambda)^{-\alpha}$
$X \sim Beta(\alpha, \beta)$	$\alpha/(\alpha+\beta)$	$\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$	

As in the discrete case, the Markov and Chebyshev inequalities are valid for continuous random variables.

8. Functions of random variables

Functions of continuous random variables are also continuous random variables. The following is a useful result for computing density functions and expectations of such functions of random variables, and is analogue to the change of variable in integrals, that you are already familiar with.

Theorem 8.1 (Change of variable)**.** *Let X be a continuous random variable with density f_X. Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable strictly monotone function. The random variable $Y = g(X)$ has density*

$$f_Y(y) = \frac{1}{|g'(x)|} f_X(x), \text{ where } x = g^{-1}(y).$$

Example 8.2 (Affine Transformations)**.** One of the simplest transformations is the affine ones: $g(x) = ax + b$. So, if $Y = aX + b$, then

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

A particularly important case is standardizing random variables: that is transform the random variable into a mean zero, variance one, random variable while preserving the shape of the density. If $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$, the linear transformation

$$Z = \frac{X - \mu}{\sigma}$$

transforms X into a standardized random variables. Its density function is

$$f_Z(z) = \sigma f_X(\sigma z + \mu).$$

Example 8.3 (Lognormal distribution)**.** Let $Z \sim \mathrm{N}(0, 1)$ be a standard normal distribution. Consider the random variable $L = e^Z$. In this case $L = g(Z)$ with $g(z) = e^z$, a differentiable strictly increasing function. If $\ell = g(z)$ then $z = \ln(\ell)$. Therefore, $g'(z) = e^{\ln \ell} = \ell$ and Theorem 8.1 gives, for $\ell \geq 0$,

$$f_L(\ell) = \frac{1}{\ell} f_Z(\ln \ell) = \frac{1}{\ell} \frac{1}{\sqrt{2\pi}} e^{-(\ln \ell)^2/2}.$$

Since $\ln L \sim \mathrm{N}(0, 1)$, the random variable L is said to have the *lognormal distribution*. It is an important distribution with interesting properties which arises naturally in several random phenomena.

Theorem 8.1 can be extended to differentiable functions g which are not necessarily strictly monotone. The case $Y = X^2$ is an important example which illustrates the general situation.

Proposition 8.4. *Let X be a continuous random variable with density f_X. The density of $Y = X^2$ is, for $y > 0$,*

$$f_Y(y) = \frac{1}{2\sqrt{y}} (f_X(\sqrt{y}) + f_X(-\sqrt{y})).$$

When computing the expectation of $Y = g(X)$ it is often simpler to do it using f_X directly, without computing f_Y. This can be achieved as follows.

Theorem 8.5 (Expectation of a transformation)**.** *Let X be a continuous random variable with density f_X. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

if the integral is absolutely convergent.

The computation of the k-th moment of a random variable is an example of application of the above theorem, applied to the function $g(x) = x^k$.

9. Simulation of random variables

Most programming languages and software packages have implemented methods to generate a random number with the uniform distribution in the interval $[0, 1]$. It is invoked as `random()` in Python, or `rand()` in C++. Actually these numbers are produced by *deterministic* computational means that produce numbers which are called *Pseudo Random Number Generators*. They produce sequences of numbers which look random-like, that is, they have statistics close to what a truly random sequence would have. Among the most common devices are the *Linear Congruential Generators*, which are based on a congruence recurrence of the form

$$x_{n+1} = (ax_n + b) \pmod{m},$$

for suitable chosen a, b and m. The sequence starts with an initial number x_0, called the *seed*, which is often generated using the computer clock.

The *Mersenne Twister* is based on an analogous linear recurrence on a finite field by using a large Mersenne prime ($2^{19937} - 1$ is used) and it is the random generator used by Python or R among many other programming languages and mathematical software systems.

Remarkably, once the uniform distribution is at disposal then one can produce random samples for other distributions easily. The most common way is based in the following result.

Proposition 9.1. *Let X be a continuous random variable with distribution function F_X. Then*

$$U = F_X(X) \sim U([0, 1]).$$

Thus, if F_X is invertible (strictly monotone) then one can obtain the distribution of X from a uniform distribution simply by writing $X = F_X^{-1}(U)$.

Example 9.2. In order to sample the exponential distribution $X \sim \text{Exp}(\lambda)$ one can obtain a sample U with the uniform distribution in $[0, 1]$ and then apply the transformation

$$X = -\frac{1}{\lambda} \ln(1 - U).$$

□

There are other specific functions, particularly to sample the Normal distribution, whose distribution function is not expressible in simple analytic terms.

Discrete distributions can be also sampled with analogous methods. If X is a discrete random variable which takes integer values with distribution function F_X then we sample

$$X = k \text{ if } F_X(k-1) < U \leq F_X(k).$$

Sampling with R according to some distribution can be done directly by the sample call.

Simulation of random phenomena is a very common tool and it can become an art of many subtleties and technical difficulties.