

## 0 Common Random Variable Distributions

In the last chapters we introduced the concept of random variables, and how to compute important quantities such as the expectation, variance and other important characteristics for those random variables, such as their c.d.f and p.d.f.

Even though from a theoretical point of view that is enough to compute everything we need about a random variable, in practice it is useful to note that there are some random variables that behave in a very specific way and that we can use as building blocks to model more complex phenomena. In this chapter we will introduce some of the most common **families of random variables**, that is, groups of random variables that share some common characteristics and that can be used to model specific types of phenomena.

### 0 Bernoulli and Binomial Distributions

The simplest random variable distribution we can think about is the **Bernoulli distribution**.

#### Definition 1 (Bernoulli Distribution)

A random variable with two possible outcomes, 0 and 1 (usually representing *failure* and *success* respectively), is called a **Bernoulli random variable**, its distribution is a **Bernoulli distribution** and any experiment with a *binary outcome* is a Bernoulli **trial**.

The **sample space** of the random variable is given by  $\Omega_X = \{0, 1\}$ . The distribution is modeled by a *single parameter*  $p$  which represents the *probability of success* for the trial. Therefore the probability of a failure is  $1 - p$ .

#### Probability Mass Function

Let  $X$  be a Bernoulli random variable with parameter  $p$ . The probability mass function (p.m.f.) of  $X$  is defined as follows:

$$p_X(x) = \begin{cases} q = 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases} \quad 0.1$$

#### Expected Value and Variance

Let  $X$  be a Bernoulli random variable with parameter  $p$ . Considering its probability mass function in Equation 0.1, we can compute its expected value.

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x \cdot p_X(x) = 0 \cdot (1 - p) + 1 \cdot (p) = p \quad 0.2$$

Given the expected value computed previously, we can also plug it into [ciao!eq:variance\\_random\\_variable\\_expanded](#) to compute the variance of a Bernoulli random variable.

$$\text{Var}\{X\} = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p) \quad 0.3$$

Now that we have defined all the important characteristics of a Bernoulli random variable, we can try to use it to model some more complex experiments. Suppose for example that we want to **replicate** a Bernoulli trial multiple times, say  $n$  and each of those trials is independent, this is how we get a **Binomial distribution**.

#### Definition 2 (Binomial Distribution)

A variable described as the number of successes  $Y$  in a sequence of independent Bernoulli trials  $X_i \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$ , has **binomial distribution**. Its parameters are  $n$ , the number of trials, and  $p$ , the probability of success in each trial.

Given  $n$  independent Bernoulli trials  $X_i \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$ , we can define the random variable  $Y$  as the number of successes in those trials as  $Y = \sum_{i=1}^n X_i$ .

### Probability Mass Function

Let  $X$  be a Binomial random variable with parameters  $n$  and  $p$ . The probability mass function (p.m.f.) of  $X$  is defined as follows:

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad 0.4$$

To get a better understanding of why the p.m.f. as defined as in Equation 0.4, we can think about the following:

- We need to have exactly  $x$  successes, which happens with probability  $p^x$ .
- We need to have exactly  $n - x$  failures, which happens with probability  $(1 - p)^{n-x}$ .
- The successes and failures can be arranged in any order, and there are  $\binom{n}{x}$  ways to choose which  $x$  trials are successes out of  $n$  total trials.

### Expected Value

The **expected value** of a Binomial random variable can be computed using the linearity of expectation as follows:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i] = n \cdot p \quad 0.5$$

where we used the fact that each  $X_i$  is a Bernoulli random variable with parameter  $p$ , therefore its expected value is  $p$  as shown in Equation 0.2. As far as the **variance** is concerned, we can compute it in the following way:

$$\text{Var}\{X\} = \text{Var}\left\{\sum_{i=1}^n X_i\right\} = \sum_{i=1}^n \text{Var}\{X_i\} = n \cdot p \cdot (1 - p) = npq \quad 0.6$$

Notice that we could use the fact that the  $X_i$  are independent to compute the variance of their sum as the sum of their variances, as the last property of the last chapter states.

### R Implementation

In R we have the following functions to work with Binomial random variables at our disposal:

- `dbinom(x, n, p) = P[X = x]`, that is the probability mass function (p.m.f.).
- `pbinom(x, n, p) = P[X ≤ x]`, that is the cumulative distribution function (c.d.f.).
- `qbinom(q, n, p) = x` if  $P[X ≤ x] = q$ , that is the quantile function.
- `rbinom(r, n, p) = {x_1, x_2, …, x_r}`, that is a vector of  $r$  random samples drawn from the distribution.

### Remark

All R functions that allow us to work with any common random variable distribution follow the same naming convention, where the first letter indicates the type of function (d for p.m.f./p.d.f., p for c.d.f., q for quantile function and r for random sampling), followed by the name of the distribution.

## 0 Multinomial Distribution

After introducing the Bernoulli and Binomial distributions, we can now generalize those concepts to the **Multinomial distribution**. If the experiments we are modeling are binary there are only two possible outcomes and we can model a repetition of them by means of the binomial. In case the experiments have *more than two possible outcomes*, say  $k$  we need to use the multinomial distribution.

### Definition 3 (Multinomial Distribution)

A random variable described as the counts of each outcome in a sequence of independent trials with  $k$  possible outcomes, has **multinomial distribution**. Its parameters are  $n$ , the number of trials, and  $p_1, p_2, \dots, p_k$ , the probabilities of each outcome in each trial, such that  $\sum_{i=1}^k p_i = 1$ .



### Probability Mass Function

Let  $X_i$  be the number of times outcome  $i$  occurs in  $n$  independent trials, each with  $k$  possible outcomes. The random vector  $(X_1, X_2, \dots, X_k)$  has **joint multinomial distribution** with probability mass function (p.m.f.) defined as follows:

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \dots, X_k = x_k] = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \quad 0.7$$

where we implicitly have the constraints that  $\sum_{i=1}^k x_i = n$  and we have introduced the **multinomial coefficient**  $\frac{n!}{x_1! \dots x_k!}$  which counts the number of ways to arrange  $n$  trials with  $x_i$  occurrences of outcome  $i$  for each  $i$ . Of course we also need to have that the values  $x_i \geq 0$ .

It is always possible to transform a multinomial distribution into a bunch of binomial distributions by considering each outcome separately. To do so, we just need to focus on one of the  $k$  outcomes at a time, where the success is getting that specific outcome, and the failure is getting any of the other  $k - 1$  outcomes.

### Expected Value, Variance and Covariance

By focusing solely on the outcome  $i$ , since the experiments are Bernoulli trials with success probability  $p_i$ , we can use the results we obtained for the Binomial distribution to compute the expected value and variance of the random variable  $X_i$  as follows:

$$\mathbb{E}[X_i] = p_i \quad \text{Var}\{X_i\} = n \cdot p_i \cdot (1 - p_i)$$

Since we are dealing with a vector of random variables, we can also compute the **covariance** between any two random variables  $X_i$  and  $X_j$  with  $i \neq j$  as follows:

$$\text{Cov}\{X_i, X_j\} = -n \cdot p_i \cdot p_j \quad 0.8$$

Intuitively this negative covariance makes sense, since if the count of outcome  $i$  increases, the count of outcome  $j$  must decrease, given that the total number of trials  $n$  is fixed.

## R Implementation

Since the multinomial distribution is a joint distribution over multiple random variables, in R we only have two functions to work with it:

- `dmultinom`: the joint probability density function (p.m.f.)
- `rmultinom`: the function to generate random samples from the distribution.

We don't have a specific function for the cumulative distribution function (c.d.f.) or the quantile function; they are indeed very hard to define and manage for joint distributions.

## 0 Geometric Distribution

Another common random variable distribution is the **Geometric distribution**, it is again very much related to the Bernoulli distribution.

### Definition 4 (Geometric Distribution)

A random variable that models the number of Bernoulli trials needed to get the first success, has **Geometric distribution**. Its parameter is  $p$ , the probability of success in each trial.



### Probability Mass Function

Let  $X$  be a Geometric random variable with parameter  $p$ . The probability mass function (p.m.f.) of  $X$  is defined as follows:

$$\mathbb{P}[X = x] = (1 - p)^{x-1} \cdot p \quad 0.9$$

where  $x$  can take any positive integer value, that is  $x \in \{1, 2, 3, \dots\}$ . The rational behind this formulation is that to have the first success at trial  $x$  we need to have  $x - 1$  failures, each of which happens with probability  $1 - p$ , and a success, that happens with probability  $p$ .

### Expected Value and Variance

Let  $X$  be a Geometric random variable with parameter  $p$ . Considering its probability mass function in Equation 0.9. To compute its **expected value**, suppose we can write the random variable  $X$  as follows:

$$X = \sum_{i=1}^{\infty} I_i$$

Where  $I_i$  is an indicator that **at least**  $i$  trials are needed to get the first success. We can compute the expected value of  $X$  as follows:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{E}[I_i]\right] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i]$$

But  $\mathbb{P}[X \geq i]$  is the probability that the first  $i - 1$  trials are failures, so  $\mathbb{P}[X \geq i] = (1 - p)^{i-1}$  therefore we have:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} (1 - p)^{i-1} = \sum_{j=0}^{\infty} (1 - p)^j = \frac{1}{1 - (1 - p)} = \frac{1}{p} \quad 0.10$$

Notice how we use the formula for the convergence of a **geometric series** to compute the final result, this is why this distribution is called **geometric**. As far as the **variance** is concerned, we can compute it in the following way:

$$\text{Var}\{X\} = \frac{1-p}{p^2} \quad 0.11$$

### Remark

Notice how the expected value of a Geometric random variable has the formulation in Equation 0.10 by no surprise: the more the probability of success  $p$  increases, the less trials we expect to need in order to get the first success.

### Memoryless Property

Until now we have not mentioned the cumulative distribution function (c.d.f.) of this random variable. However, it is interesting to notice that the c.d.f. of a Geometric random variable has a very special property, called the **memoryless property**.

Imagine that we have already performed at least  $y$  trials of an Bernoulli experiment without getting a success. The probability that we are going to *keep going* for at least another  $x$  trials without getting a success can be modeled with in the following way:

$$\mathbb{P}[X > x + y \mid X - y] = \mathbb{P}[X > x] \quad 0.12$$

In other words, the probability of needing more than  $x + y$  trials given that we have already performed  $y$  trials without success is equal to the probability of needing more than  $x$  trials from scratch. This property is called **memoryless** because the process does not care about what happened in the past, it only cares about the present situation.

### R Implementation

Before understanding how R provides us with functions to work with this kind of random variable, it is crucial to understand that there are two different conventions to define this random variable.

Previously we defined a geometric random variable as the number of trials needed in order to observe a success. However, it is also common to define it as the number of **failures before** the success. In the first case we have  $\Omega_X = \{1, 2, \dots\}$ , whilst in the second case we have  $\Omega_X = \{0, 1, 2, \dots\}$ , since we can have zero failures before the first success.

The second one is exactly the convention that R uses, therefore all the functions we are going to introduce now are based on that definition. To switch from the second definition to the first one, it is necessary to first transform the random variable  $X$  into the random variable  $X = Y + 1$ . Similarly we'll have that:

$$\mathbb{P}[X = x] = \mathbb{P}[Y = x - 1]$$

In R we have the following functions to work with Geometric random variables:

- `dgeom(x-1, p) =  $\mathbb{P}[X = x]$`
- `pgeom(x-1, p) =  $\mathbb{P}[X \leq x]$`
- `qgeom(q, p) = x - 1 if  $\mathbb{P}[X \leq x] = q$`
- `rgeom(r, p)` simulates  $r$  realizations of  $X - 1$

## 0 Hyper-geometric Distribution

Another important random variable distribution is the **hypergeometric distribution**, which is used to model experiments where we draw samples without replacement from a finite population.

### Definition 5 (Hypergeometric Distribution)

A random variable that models the number of successes in a sample of size  $n$  drawn **without replacement** from a population of size  $N$  containing  $M$  successes and  $N - M$  failures has **hypergeometric distribution**.

### Probability Mass Function

Let  $X$  be a hypergeometric random variable with parameters  $N$  (population size),  $M$  (number of successes in the population),  $M$  (number of failures),  $n$  (the sample size). The probability mass function (p.m.f.) of  $X$  is defined as in the following equation:

$$\mathbb{P}[X = x] = \text{hyper geom}(x, n, M, N) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \quad 0.13$$

where  $x$  is an integer such that  $\max(0, n - N + M) \leq x \leq \min(n, M)$

### Expected Value and Variance

Let  $X$  be a hypergeometric random variable with p.m.f. given by  $\text{hyper geom}(x, n, N, M)$ , then we can define its expected value and variance as follows:

$$\mathbb{E}[X] = n \cdot \frac{M}{N} \quad \text{Var}\{X\} = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \left(1 - \frac{M}{N}\right) \quad 0.14$$

## 0 Introduction to Stochastic Processes

In this section we will try to build a link between everything we have seen so far about random variables and basic probability theory, and the core concept of this course: **stochastic processes**. A sequence  $\{X_n\}$  of random variables is a **stochastic process**. With the term “sequence” we refer to an *infinite random vector*.

If we consider a **finite collection** of random variables  $\{X_1, X_2, \dots, X_n\}$  we can characterize all we need to know about such random variables and their relationships by means of their **joint probability distribution**. Indeed starting from it we can compute all the marginal probabilities; furthermore we can also notice that  $\forall i_1, i_2, \dots, i_k$  and for  $k \geq 1$  we can compute the joint probability of  $(X_{i1}, X_{i2}, \dots, X_{ik})$  by integrating (or summing) out all the other variables from the joint distribution.

If we can do this for every possible finite subset of r.v.’s from our infinite collection, that means we know the **law** (which is the distribution in this context of random processes) of the random sequence. Informally speaking, we can say that if  $X = \{X_n\}_{n=1}^{\infty}$  the **law of  $X$**  is defined as the collection of all the *finite-dimensional distributions*  $\forall n \in \{1, 2, 3, \dots\}$ . Given any subset of indices  $i_1, i_2, \dots, i_k$  with  $k \geq 1$ , the finite-dimensional distribution is defined as the joint distribution of the random variables  $(X_{i1}, X_{i2}, \dots, X_{ik})$ .

The simplest stochastic process we can think about is a collection  $X_i \stackrel{\text{i.i.d.}}{\sim} F_X \quad i = 1, 2, \dots$ . A finite subset of them is called a **sample** from distribution  $F_X$ . The reason why it is the simplest is given in the following equation:

$$\forall n, \forall i_1, i_2, \dots, i_n : \mathbb{P}[\!(X_{i_1}, \dots, X_{i_n})\!] = F_{X_{i_1}, \dots, X_{i_n}}(x_{i_1}, \dots, x_{i_n}) = \prod_{j=1}^n F_X(x_{ij}),$$

That is, the joint distribution of any finite subset of them can be computed as the product of their marginal distributions, since they are all **independent** and **identically distributed**.

### Remark

If the random variables are independent but not identically distributed we need to know the **marginal distribution** for each one of the random variables. Namely, if  $X_i \stackrel{\text{ind}}{\sim} F_{X_i}$  then we have that the joint probability of the sample is given by:

$$F_{X_{i_1}, \dots, X_{i_n}}(x_{i_1}, \dots, x_{i_n}) = \prod_{j=1}^n F_{X_{ij}}(x_{ij})$$

Namely, we need to have knowledge about a countable number of marginal distributions.

### Example: Sequence of independent non identically distributed random variables

Suppose we are dealing with a sequence of independent random variables which are not **identically distributed**. To keep the matter simple, let's suppose that the distribution changes according to the index of the random variable in the sequence and the basic distribution is always a Bernoulli distribution, that is:  $X_i \sim \text{Bern}\left(\frac{1}{i}\right)$ .

Of course, considering everything we have said so far, we can say that  $\{X_n\}_{n=1}^\infty$  is a stochastic process.

Consider now the following object:

$$Y_n = \sum_{i=1}^n X_i \text{ Bin}(n, p))$$

And consider the collection  $\{Y_n\}_{n=1}^\infty$ ; that one is also a **stochastic process**. Again, in this case  $Y_i$ 's are surely **not identically distributed**, indeed if  $n \neq m$ ,  $Y_m$  and  $Y_n$  have a different distribution while both being Binomial random variable. As far as independency is concerned we can take a look at the following equation:

$$Y_{n+1} = \sum_{i=1}^{n+1} X_i = Y_n + X_{n+1}$$

if we try to study the value of  $Y_{n+1}$  alone we can correctly conclude that it may take any value in  $\{0, \dots, n+1\}$ ; however if we consider  $Y_{n+1} \mid Y_n = n$  we can easily see that  $Y_{n+1}$  can only take the values in  $\{n, n+1\}$ , thus  $Y_{n+1}$  and  $Y_n$  are **not independent**. Indeed the conditional distribution of  $Y_{n+1} \mid Y_n$  is given by:

$$p_{Y_{n+1} \mid Y_n}(y_{n+1} \mid y_n) = \begin{cases} 1-p & \text{if } y_{n+1} = y_n \\ p & \text{if } y_{n+1} = y_n + 1 \\ 0 & \text{otherwise} \end{cases}$$

That is actually quite trivial to compute since we are dealing with Binomial random variables built from independent Bernoulli trials. In **general**, supposing we are working with Binomial random variables, we have that:

$$p_{Y_1, Y_2, \dots, Y_n}(y_1, y_2, \dots, y_n) = p_{Y_1}(y_1) p_{Y_2 | Y_1}(y_2 | y_1) \dots p_{Y_n | Y_{n-1}}(y_n | y_{n-1})$$

Suppose that we know  $Y_n = y_n$  and  $Y_{n-1} = y_{n-1}$ , let's see how we can use this information:

$$Y_{n+1} = Y_n + X_{n+1}$$

Basically, the first information is very useful since it tells us how many successes we had up to trial  $n$ , whilst the second information is just telling us that we can write  $Y_n = y_{n-1} + X_n$ , but we already know the value of  $Y_n$  so that second piece of information is not really adding anything new in case we already know  $Y_n$ .

### Warning

This does not mean, by any means, that  $Y_{n+1}$  and  $Y_{n-1}$  are independent. Indeed the value of  $Y_{n+1}$  is very much dependent on the value of  $Y_{n-1}$ :  $Y_{n+1} = Y_{n-1} + X_n + X_{n+1}$ . To be more precise, we can also write the conditional probability of  $Y_{n+1} | Y_{n-1}$  as follows:

$$p_{Y_{n+1} | Y_{n-1}}(y_{n+1} | y_{n-1}) = \begin{cases} (1-p)^2 & \text{if } y_{n+1} = y_{n-1} \\ 2(1-p)p & \text{if } y_{n+1} = y_{n-1} + 1 \\ p^2 & \text{if } y_{n+1} = y_{n-1} + 2 \\ 0 & \text{otherwise} \end{cases}$$

What we can say about  $Y_{n+1}$  and  $Y_{n-1}$  is that they are **conditionally independent** given  $Y_n$ , this is very useful because it allows us to simplify the computation of joint probabilities.

If we now try to look at the joint probabilities we may be interested in, we can use what we have just observed to write the following:

$$\begin{aligned} p_{Y_1, \dots, Y_n}(y_1, \dots, y_n) &= p_{Y_1}(y_1) \cdot p_{Y_2 | Y_1}(y_2 | y_1) \cdot p_{Y_3 | Y_2}(y_3 | y_2) \dots \\ &= p_{Y_1}(y_1) \prod_{i=1}^{n-1} p_{Y_{i+1} | Y_i}(y_{i+1} | y_i) \end{aligned} \quad 0.15$$

If each  $X_i$  is the result of a coin toss we can model a  $Y_n$  as the *number of wins* in the first  $n$  throws we can model a  $Y_n$  as the *number of wins* in the first  $n$  throws. For the first toss we are going to have the following:

$$p_{Y_1}(y_1) = \begin{cases} 1-p & \text{if } y_1 = 0 \\ p & \text{if } y_1 = 1 \end{cases}$$

This is straightforward since  $Y_1$  is just a Bernoulli random variable. For the second toss we have:

$$p_{Y_2}(y_2) = \begin{cases} (1-p)^2 & \text{if } y_2 = 0 \\ 2(1-p)p & \text{if } y_2 = 1 \\ p^2 & \text{if } y_2 = 2 \end{cases}$$

We can derive this result by noticing that  $Y_2 \sim \text{Binom}(2, p)$ . Given these pieces of information we can compute several different probabilities. For instance  $\mathbb{P}[Y_1 = 1]$ ,  $\mathbb{P}[Y_n = 1]$ . But what about the probability of getting a win in the first throw and only lose in the next 6? This can be modeled by the following equation which leverages Equation 0.15:

$$\begin{aligned}
\mathbb{P}[Y_1 = 1 \wedge Y_2 = 1 \wedge \dots \wedge Y_7 = 1] &= p_{Y_1}(y_1) \cdot p_{Y_2 | Y_1}(1 | 1) \\
&\quad \cdot \dots \cdot p_{Y_7 | Y_6}(1 | 1) \\
&= p \cdot \prod_{i=1}^{7-1} p_{Y_{i+1} | Y_i}(1 | 1) \\
&= p \cdot \prod_{i=1}^6 (1-p)^2 = p(1-p)^{12} = \frac{1}{2} \left(\frac{1}{2}\right)^{12} = \frac{1}{2^{13}}
\end{aligned}$$

### Definition 6 (Markov Process)

A sequence  $X = \{X_n\}_{n=1}^\infty$  where each  $X_n + 1$  is **conditionally independent** of  $\{X_{n-1} \dots X_1\}$  given  $X_n$  is called a **Markov process** or **Markov chain**, which is a stochastic process with some interesting properties that make it easier to study and analyze.



## Stochastic Processes and Common Random Variables

Let's now try to review the concept of **geometric distribution** we have already seen before but in a stochastic process fashion. Consider the random variables  $X_i \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$  and consider the following:

$$W_1 = \min\{n : X_n = 1\}$$

Clearly we can see  $W_1 \sim \text{Geom}(p)$ . And we can sort of consider its value as the 'expected waiting time' until the first success. We can also consider the random variable  $Y_n$  as the sum of all the  $X_i$  up to time  $n$ .

We can define two random sequences as follows:

- $W = \{W_n\}$  as the sequence of **waiting times** between changes in the process (or in the value of  $Y_n$ ). We also refer to them as **inter-arrival times**.
- $Y = \{Y_n\}$  as the sequence that counts the number of successes up to time  $n$ , basically it is a **counting process**.

We can actually do the same for **hyper-geometric** random variables. Suppose we have for instance  $N$  balls in an urn, of which  $M$  are successes (say red) and  $N - M$  are failures. The probability of success when drawing a ball at random from the urn is  $p = \frac{M}{N}$ . So if we consider the first ball extraction we obtain the following:

$$X_1 \sim \text{Bern}\left(p_1 = \frac{M}{N}\right)$$

Now, if we move to the second extraction, we still have a Bernoulli random variable, but we don't know its parameter, since it depends on the result of the first extraction. We can consider two cases:

- $X_2 | X_1 = 1 \sim \text{Bern}\left(\frac{M-1}{N-1}\right)$ , in case the first extraction was a success.
- $X_2 | X_1 = 0 \sim \text{Bern}\left(\frac{M}{N-1}\right)$ , in case the first extraction was a failure.

If we now take a look at  $X_3$ , we can notice that this gets more complicated:

- $X_3 | X_2 = 1, X_1 = 1 \sim \text{Bern}\left(\frac{M-2}{N-2}\right)$ , in case the first two extractions were successes.
- $X_3 | X_2 = 1, X_1 = 0 \sim \text{Bern}\left(\frac{M-1}{N-2}\right)$ , in case the first extraction was a failure and the second a success.
- $X_3 | X_2 = 0, X_1 = 1 \sim \text{Bern}\left(\frac{M-1}{N-2}\right)$ , in case the first extraction was a success and the second a failure.

- $X_3 \mid X_2 = 0, X_1 = 0 \sim \text{Bern}\left(\frac{M}{N-2}\right)$ , in case the first two extractions were failures.

We can notice that, if we take in consideration the **sum** of the successes up to time  $n$ , we can actually define the conditional distribution based solely on the value of that sum, not the individual outcomes of each previous extraction. We say that, in general,  $X_{n+1}$  is **conditionally independent** on  $\{X_n, X_{n-1}, \dots, X_1\}$  given  $Y_n = \sum_{i=1}^n X_i$ .

## 0 Negative Binomial Distribution

Earlier in this chapter we have been talking about binomial distributions and how they are related to the geometric distribution, which measures the ‘waiting time’ until the first success in a sequence of Bernoulli trials.

### Definition 7 (Negative Binomial Distribution)

Given a sequence of independent Bernoulli trials with success probability  $p$ , we can model the **number of trials** to obtain  **$k$  successes** with a **Negative Binomial distribution**. Suppose we have  $W_i \stackrel{\text{i.i.d.}}{\sim} \text{Geom}(p)$ . The random variable

$$N_k = \sum_{i=1}^k W_i \quad 0.16$$

follows a **Negative Binomial distribution** with parameters  $k$  and  $p$ .



### Probability Mass Function

Let  $N_k$  be a Negative Binomial random variable with parameters  $k$  and  $p$ . The probability mass function (p.m.f.) of  $N_k$  is defined as follows:

$$p_X(x) = \binom{x-1}{k-1} (1-p)^{x-k} p^k \quad \forall x \geq k \quad 0.17$$

intuitively, this formulation makes sense: to have the  $k$ -th success at trial  $x$  we need to have  $k-1$  successes in the first  $x-1$  trial (which can happen in  $\binom{x-1}{k-1}$  ways). Every combination  $k-1$  successes and  $x-k$  failures happens with probability  $p^{k-1}(1-p)^{x-k}$ ; finally we need to have a success at trial  $x$ , which happens with probability  $p$ .

### Expected Value and Variance

Let  $N_k$  be a Negative Binomial random variable with parameters  $k$  and  $p$ . Considering its probability mass function in Equation 0.17, we can compute its expected value and variance as follows:

$$\mathbb{E}[N_k] = \frac{k}{p} \quad \text{Var}\{N_k\} = k \frac{1-p}{p^2} \quad 0.18$$

### R Implementation

Before introducing the R functions to work with Negative Binomial random variables, it is important to notice that there are two different conventions to define this random variable, R actually uses a different one with respect to the one we have just introduced, similarly to what happened with the Geometric distribution, we will need to adjust the parameters accordingly:

- `dnb(x - k, k, p) = P[X = x]`, is the probability mass function (p.m.f.).
- `pnb(x - k, k, p) = P[X ≤ x]`, is the cumulative distribution function (c.d.f.).
- `qnb(q, k, p) + k = x` “if”  $P[X ≤ x] = q$ , is the quantile function.
- `rnb(r, k, p)` simulates  $r$  realizations of  $X - k$ .

To switch from the R definition to the one we have introduced, it is necessary to first transform the random variable  $X$  into the random variable  $X = Y + k$ .

## 0 Uniform Distribution

In the past few sections we have been focusing our attention on **discrete random variables**, but as we know, there are also **continuous random variables**.

### Definition 8 (Uniform Distribution)

A random variable that has an equal probability of taking any value within a given interval  $[a, b]$  has **Uniform distribution**. Its parameters are  $a$  and  $b$ , the endpoints of the interval. If the interval of values is  $[0, 1]$  we say that the random variable has a **standard uniform distribution**.



### Probability Density Function

Let  $X$  be a Uniform random variable with parameters  $a$  and  $b$ . The probability density function (p.d.f.) of  $X$  is defined as follows:

$$f_X(x) = \frac{1}{b - a} \quad \forall x \in [a, b] \quad 0.19$$

Actually the above definition is not really precise, indeed when we talk about continuous random variables we cannot really talk about probability, rather we need to talk about **density**. The reason because uniform distributions are so important is that they are the building blocks for all other random variable distributions.

### Remark

In order for Equation 0.19 to be valid, it is necessary that the value  $|b - a|$  is a finite positive number so that there is no chance of dividing by zero or by infinity. The rational behind this is quite simple, if we try to choose a random number in an interval of infinite length, we cannot do it with uniform probability since the density would be zero everywhere.

### Uniform Property

For any  $h > 0$  and  $t \in [a, b - h]$  we have that:

$$\mathbb{P}[t < X < t + h] = \int_t^{t+h} \frac{1}{b - a} dx = \frac{h}{b - a}$$

is **independent of  $t$** . The probability is only determined by the length of the interval not by the location of the point in the interval.

### Expected Value and Variance

Let  $X$  be a Uniform random variable with parameters  $a$  and  $b$ . Considering its probability density function in Equation 0.19, we can compute its expected value and variance as follows:

$$\mathbb{E}[X] = \frac{a + b}{2} \quad \text{Var}\{X\} = \frac{(b - a)^2}{12} \quad 0.20$$

It is by no surprise that the expected value of a uniform random variable is the midpoint of the interval  $[a, b]$ . As far as the variance is concerned, we can notice that it increases quadratically with

the length of the interval. If we consider the standard uniform distribution, that is  $a = 0$  and  $b = 1$ , we have that  $\mathbb{E}[X] = \frac{1}{2}$  and  $\text{Var}\{X\} = \frac{1}{12}$ .

### Uniform Distribution Transformation and Standardization

One very common operation when working with this kind of random variable is to transform it into a standard uniform random variable and vice-versa. Consider two random variables  $X \sim \text{Uniform}(a, b)$  and  $Y \sim \text{Uniform}(0, 1)$ . We can transform  $X$  into  $Y$  and vice-versa as follows:

$$\begin{aligned} Y &= \frac{X - a}{b - a} \sim \text{Uniform}(0, 1) \\ X &= a + (b - a)Y \sim \text{Uniform}(a, b) \end{aligned} \quad 0.21$$

Let's now consider the following new random variable:

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}\{X\}}} \sim \text{Uniform}(z_l, z_u)$$

Abd suppose we want to compute its expected value and variance. To do so we need to compute the values of  $z_l$  and  $z_u$  first:

$$\begin{aligned} x = a \Rightarrow z_l &= \frac{a - \mathbb{E}[X]}{\sqrt{\text{Var}\{X\}}} = \frac{a - \frac{a+b}{2}}{\sqrt{\frac{(b-a)^2}{12}}} \\ x = b \Rightarrow z_u &= \frac{b - \mathbb{E}[X]}{\sqrt{\text{Var}\{X\}}} = \frac{b - \frac{a+b}{2}}{\sqrt{\frac{(b-a)^2}{12}}} \end{aligned}$$

Now that we have these values it is easy to notice that  $\mathbb{E}[Z] = 0$ ,  $\text{Var}\{Z\} = 1$ .

#### Definition 9 (Standardization)

Given **any** discrete or continuous random variable  $X$  with expected value  $\mathbb{E}[X] = \mu$  and variance  $\text{Var}\{X\} = \sigma^2$ , we can define the **standardized** random variable  $Z$  as follows:

$$Z = \frac{X - \mu}{\sigma} \quad 0.22$$

which satisfies  $\mathbb{E}[Z] = 0$  and  $\text{Var}\{Z\} = 1$ .

It is actually easy to see why the expected value and variance of  $Z$  are as we have just said:

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}\left[\frac{X - \mu}{\sigma}\right] = \frac{1}{\sigma}(\mathbb{E}[X] - \mu) = 0 \\ \text{Var}\{Z\} &= \text{Var}\left\{\frac{X - \mu}{\sigma}\right\} = \frac{1}{\sigma^2}(\text{Var}\{X\} - 0) = 1 \quad \blacksquare \end{aligned}$$

The set of possible values of  $Z$  and  $X$  are different. For instance, consider  $X \sim \text{Bern}(p)$  with  $\Omega_X = \{0, 1\}$ . The standardized random variable  $Z$  can now be built as follows:

$$Z = X - \frac{p}{\sqrt{p(1-p)}}$$

If we now take a look at the possible values of  $Z$  we have:  $\Omega_Z = \left\{ -\frac{p}{\sqrt{p(1-p)}}, \frac{1-p}{\sqrt{p(1-p)}} \right\}$  respectively when  $X = 0$  and  $X = 1$ . Clearly  $Z$  is **not a Bernoulli**. This tells us a very important fact about standardization.

### ⚠️ Warning

In general, a **standardized random variable** does not belong to the *same family* of the original random variable that was used to build the standardization

Following we have a theorem which tells us something very important about standardization and uniform random variables:

#### Theorem 1 (Standardization of Uniform Random Variables)

Given any uniform random variable  $X \sim \text{Uniform}(a, b)$ , it is **closed under linear transformation**, that is the uniformness of the random variable is preserved under any linear transformation, including **standardization**.

### ⚠️ Warning

Even though the name suggests it, the **standard uniform random variable**  $Y \sim \text{Uniform}(0, 1)$  is just a special case of uniform random variable. It is **not** the result of a standardization process.

## R Implementation

In R we have the following functions to work with Uniform random variables:

- `dunif(x, a, b) =  $f_{X(x)}$` , is the probability density function (p.d.f.).
- `punif(x, a, b) =  $\mathbb{P}[X \leq x]$` , is the cumulative distribution function (c.d.f.).
- `qunif(q, a, b) =  $x = F^{-1}(q)$` , i.e.,  $\mathbb{P}[X \leq x] = q$ , is the quantile function.
- `runif(r, a, b)` simulates  $r$  realizations of  $X$ .

## 0 Normal (Gaussian) Distribution

Although it is not the distribution we are going to preponderantly use in this course, the **Normal distribution** is so common and important in all probability theory that it is worth spending some time on it. If the uniform distribution serves to express the idea of ‘equiprobability’, the normal distribution is often used to model ‘natural’ phenomena.

### Definition 10 (Normal Distribution)

A random variable that models phenomena where values tend to cluster around a central mean value with a certain variability has **Normal distribution**. Its parameters are  $\mu$  (the mean) and  $\sigma^2$  (the variance).

When dealing with this kind of random variables we often refer to the mean as **location parameter** and to the variance as **scale parameter**.

### Probability Density Function

Let  $X$  be a Normal random variable with parameters  $\mu$  and  $\sigma^2$ . The probability density function (p.d.f.) of  $X$  is defined as follows:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad 0.23$$

We can see how this formulation intuitively makes sense. The numerator of the fraction in the exponential is squared so that larger errors are more penalized (i.e., less likely) w.r.t. smaller errors. The numerator is then divided by the variance so that larger variances lead to less penalization for larger errors. Finally the whole expression is normalized by the factor  $\frac{1}{\sigma\sqrt{2\pi}}$  so that the total area under the curve is equal to 1.

We can see that the value of  $\mu$  serves to control the **location** of the distribution’s peak, whilst the value of  $\sigma^2$  serves to control the **spread** of the distribution around the mean. This is illustrated in Figure 0.1.

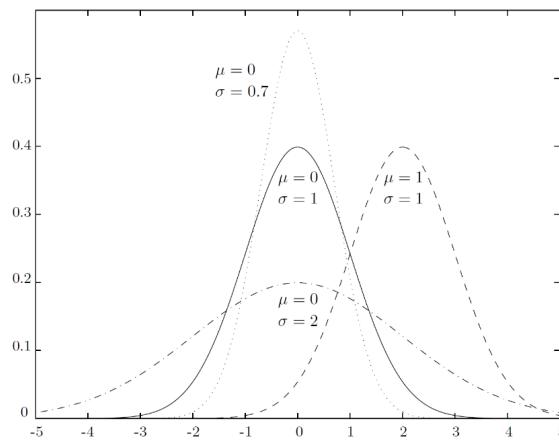


Figure 0.1: Normal distributions with different parameters

### Standardization

There is actually no point in computing the expected value and variance of a Normal random variable in that they are exactly equal to the parameters used to define the distribution:  $\mathbb{E}[X] = \mu$  and  $\text{Var}\{X\} = \sigma^2$ .

Nevertheless, it is possible to define a **standard normal random variable**  $Z$  such as it has expected value equal to 0 and variance equal to 1. Along with this fact, it is interesting to notice that Normal random variables are **closed under linear transformation**, that is if we take any Normal random variable and we apply a linear transformation to it, the resulting random variable is still Normal. In particular we can notice the following:

$$X = aZ + b \sim \text{Normal}(a\mathbb{E}[Z] + b, a^2 \text{Var}\{Z\}) \quad 0.24$$

and by simply plugging the knowledge that  $\mathbb{E}[Z] = 0$  and  $\text{Var}\{Z\} = 1$  in the above equation we have that  $X \sim \text{Normal}(b, a^2)$ .

### Transformation from and to Standard Normal

Given any Normal random variable  $X \sim \text{Normal}(\mu, \sigma^2)$  and a standard normal random variable  $Z \sim \text{Normal}(0, 1)$  we can transform  $X$  into  $Z$  and vice-versa as follows:

$$\begin{aligned} Z &= \frac{X - \mu}{\sigma} \sim \text{Normal}(0, 1) \\ X &= \mu + \sigma Z \sim \text{Normal}(\mu, \sigma^2) \end{aligned} \quad 0.25$$

This is indeed very similar to the standardization process we have already seen in the case of uniform random variables.

### R Implementation

In R we have the following functions to work with Normal random variables:

- `dnorm(x, mu, sigma) =  $f_{X(x)}$` , is the probability density function (p.d.f.).
- `pnorm(x, mu, sigma) =  $\mathbb{P}[X \leq x]$` , is the cumulative distribution function (c.d.f.).
- `qnorm(q, mu, sigma) =  $x = F^{-1}(q)$` , i.e.,  $\mathbb{P}[X \leq x] = q$ , is the quantile function.
- `rnorm(r, mu, sigma)` simulates  $r$  realizations of  $X$ .

## 0 Poisson Distribution