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 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:

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

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.

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

Given the expected value compute previously, we can also plug it into ciaoneeq:variance_random_variable_expanded to compute the variance of a Bernoulli random variable.

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

Now that we have defined all the important characteristics of a Bernoulli random variable, we can try to use is to model some more complex experiment. 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 $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, 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 $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, 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:

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

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:

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

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:

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

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)` $= \mathbb{P}[\![X = x]\!]$, that is the probability mass function (p.m.f.).

- `pbinom(x, n, p)` $= \mathbb{P}[\![X \leq x]\!]$, that is the cumulative distribution function (c.d.f.).

- `qbinom(q, n, p)` $= x$ if $\mathbb{P}[\![X \leq x]\!] = q$, 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.

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 $p_1, p_2, ..., p_k$, 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 $(X_1, X_2, ..., X_k)$ has **joint multinomial distribution** with probability mass function (p.m.f.) defined as follows:

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

where we implicitly have the constraints that $\sum_{i=1}^{k} x_i = n$ and we have introduced the **multinomial coefficient** $\frac{n!}{x_1! ... x_k!}$ 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:

$$\mathbb{E}[\![X_i]\!] = p_1 \qquad \text{Var}\{X_i\} = n \cdot p_i \cdot (1 - p_i)$$

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:

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

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:

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

where x can take any positive integer value, that is $x \in \{1, 2, 3, ...\}$. The rational behind this formulation is that the 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:

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

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:

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

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:

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

> 💬 **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:

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

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 $\Omega_X = \{1, 2, ...\}$, whilst in the second case we have $\Omega_X = \{0, 1, 2, ...\}$, 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:

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

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 hyper geom(x, n, N, M), then we can de fine its expected value and variance as follows:

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

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 $\{X_1, X_2, ..., X_n\}$ 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 $\forall i_1, i_2, ..., i_k$ and for $k \geq 1$ we can compute the joint probability of $(X_{i1}, X_{i2}, ..., X_{ik})$ 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* $\forall n \in \{1, 2, 3, ...\}$. Given any subset of indices $i_1, i_2, ..., i_k$ with $k \geq 1$, the finite-dimensional distribution is defined as the joint distribution of the random variables $(X_{i1}, X_{i2}, ..., X_{ik})$.

The simplest stochastic process we can think about is a collection $X_i \overset{\text{i.i.d.}}{\sim} F_X \quad i = 1, 2, ...$ 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:

$$\forall n, \forall i_1, i_2, ..., i_n : \mathbb{P}\left[\!\!\left[\left(X_{i_1}, ..., X_{i_n}\right)\right]\!\!\right] = F_{X_{i1},...X_{in}}\left(x_{i1}, ..., x_{in}\right) = \prod_{j=1}^{n} F_X\left(x_{ij}\right),$$

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 $X_i \overset{\text{ind}}{\sim} F_{X_i}$ then we have that the joint probability of the sample is given by:
>
> $$F_{X_{i1},...X_{in}}\left(x_{i1}, ..., x_{in}\right) = \prod_{j=1}^{n} F_{X_{ij}}\left(x_{ij}\right)$$
>
> 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, ..., 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:

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

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:

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

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

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

<div align="right">0.15</div>

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:

$$\mathbb{P}[\![Y_1 = 1 \wedge Y_2 = 1 \wedge ... \wedge Y_7 = 1]\!] = p_{Y_1}(y_1) \cdot p_{Y_2 \mid Y_1}(1 \mid 1)$$
$$\cdot ... \cdot p_{Y_7 \mid Y_6}(1 \mid 1)$$
$$= p \cdot \prod_{i=1}^{7-1} p_{Y_{i+1} \mid Y_i}(1 \mid 1)$$
$$= p \cdot \prod_{i=1}^{6}(1-p)^2 = p(1-p)^{12} = \frac{1}{2}\left(\frac{1}{2}\right)^{12} = \frac{1}{2^{13}}$$

> **Definition 6 (Markov Process)**
>
> A sequence $X = \{X_n\}_{n=1}^{\infty}$ where each $X_n + 1$ is **conditionally independent** of $\{X_{n-1}...X_1\}$ 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. ♣