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:

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

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.

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

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.

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

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:

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

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:

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

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:

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

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.

> **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 $p_1, p_2, ..., p_k$, 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:

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

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:

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

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:

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

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:

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

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:

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

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 $\text{geom}(x, n, N, M)$, then we can de fine its expected value and variance as follows:

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

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:

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

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

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

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:

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

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:

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

Definition 6 (Markov Process)

A sequence $X = \{X_n\}_{n=1}^{\infty}$ where each $X_n + 1$ is **conditionally independent** of $\{X_{n-1} \ldots X_1\}$ 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 $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$ 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:

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

- $X_3 \mid X_2 = 1, X_1 = 1 \sim \text{Bern}\left(\frac{M-2}{N-2}\right)$, in case the first two extractions were successes.
- $X_3 \mid X_2 = 1, X_1 = 0 \sim \text{Bern}\left(\frac{M-1}{N-2}\right)$, in case the first extraction was a failure and the second a success.
- $X_3 \mid X_2 = 0, X_1 = 1 \sim \text{Bern}\left(\frac{M-1}{N-2}\right)$, 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 $\{X_n, X_{n-1}, ..., X_1\}$ 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 $W_i \overset{\text{i.i.d.}}{\sim} \text{Geom}(p)$. The random variable

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

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:

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

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:

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

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:

- `dnbinom(x - k, k, p)` $= \mathbb{P}[\![X = x]\!]$, is the probability mass function (p.m.f.).
- `pnbinom(x - k, k, p)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function (c.d.f.).
- `qnbinom(q, k, p) + k` $= x$ "if" $\mathbb{P}[\![X \leq x]\!] = q$, is the quantile function.
- `rnbinom(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 if $[0, 1]$ 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:

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

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 buy 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:

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

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:

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

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:

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

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

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:

$$Z = \frac{X - \mu}{\sigma^2} \tag{0.22}$$

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:

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

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 μ (the mean) and σ^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 μ and σ^2. The probability density function (p.d.f.) of X is defined as follows:

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

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 μ serves to control the **location** of the distribution's peak, whilst the value of σ^2 serves to control the **spread** of the distribution around the mean. This is illustrated in Figure 0.1.

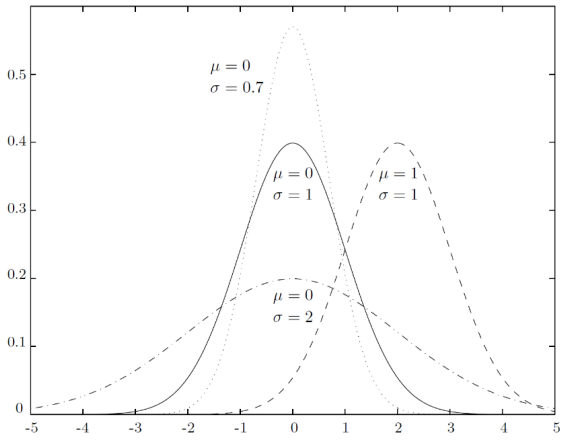

Figure 0.1: Normal distributions with different parameters

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 $\mathrm{Var}\{X\} = \sigma^2$.

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:

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

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:

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

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

Let's now take a look at what is probably the most important distribution for this course: the **Poisson distribution**. Let's first take a look at its definition.

Definition 11 (Poisson Distribution)

The numbero of "**rare**" events occurring within a fixed interval of time has **Poisson Distribution**.

♣

This definition looks a bit vague in that we still need to clarify what we mean by "rare" events. Before doing so, let's first take a look at its probability mass function.

Probability Mass Function

Let $X \sim \text{Poisson}(\lambda)$ be a Poisson random variable with parameter $\lambda > 0$. The probability mass function (p.m.f.) of X is defined as follows:

$$p_X(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \forall x \in \{0, 1, 2, ...\} \tag{0.26}$$

Though this formulation may look strange, it is indeed a probability mass function. Indeed it is both positive for all x and it sums to 1.

Positivity

To understand why it is positive there is not much to say, all the components of the product in Equation 0.26 are positive for any $\lambda > 0$ and any $x \in \{0, 1, 2, ...\}$.

Normalization

To understand why it sums to 1 we can consider the definition of $f(\lambda) = e^{\lambda}$ as the limit of an infinite series:

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \iff \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} = 1$$

where the series on the right-hand side is exactly the formulation of the p.m.f. in Equation 0.26. If we try to see this the other way around, we may wonder which is the constant factor k that makes the function $\frac{\lambda^x}{x!}$ be a proper p.m.f. We can find such a constant by solving the following equation:

$$1 = k \cdot \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \iff k = \frac{1}{\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}} = e^{-\lambda}$$

Expected Value and Variance

We can try to compute the **expected value** of the Poisson distribution by leveraging again the Taylor series expansion of the exponential function:

$$\begin{aligned} \mathbb{E}[\![X]\!] &= \sum_{x=0}^{\infty} x \; p_{X(x)} = \sum_{x=0}^{\infty} x \; e^{-\lambda} \frac{\lambda^x}{x!} \\ &= 0 + \sum_{x=1}^{\infty} x \; e^{-\lambda} \frac{\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda} \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} \lambda e^{\lambda} = \lambda \end{aligned} \tag{0.27}$$

The parameter λ of the Poisson distribution is called the **rate** or **frequency** parameter, since it represents the expected (mean) number of events per fixed amount of time.

Before actually computing the variance of the Poisson distribution it is necessary to compute another quantity. Specifically by recalling ciaoneeq_expected_value_function_random_variable we can notice the following:

$$\begin{aligned} \mathbb{E}[\![X(X-1)]\!] &= \sum_{x=0}^{\infty} x(x-1) \; p_X(x) = \sum_{x=0}^{\infty} x(x-1) \; e^{-\lambda} \frac{\lambda^x}{x!} \\ &= 0 + 0 + e^{-\lambda} \lambda^2 \sum_{x=2}^{\infty} x(x-1) \frac{\lambda^{x-2}}{x(x-1)(x-2)!} = \lambda^2 \end{aligned}$$

Now, if we notice that by linearity of expectation we can also write:

$$\mathbb{E}[\![X(X-1)]\!] = \mathbb{E}[\![X^2 - X]\!] = \mathbb{E}[\![X^2]\!] - \mathbb{E}[\![X]\!]$$

Therefore we can combine the results above and conclude that $\mathbb{E}[\![X^2]\!] = \lambda^2 + \lambda$. Now we are finally ready to give an expression for the **variance**:

$$\text{Var}\{X\} = \mathbb{E}[\![X^2]\!] - (\mathbb{E}[\![X]\!])^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda \tag{0.28}$$

We can also generalize what we have seen before when we were computing $\mathbb{E}[\![X(X-1)]\!]$:

$$\mathbb{E}\left[\left[\prod_{i=0}^{k-1}(X-i)\right]\right] = \lambda^k$$

R Implementation

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

- `dpois(x, lambda)` $= p_X(x)$, is the probability mass function (p.m.f.).
- `ppois(x, lambda)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function (c.d.f.).
- `qpois(q, lambda)` $= x = F^{-1}(q)$, i.e., $\mathbb{P}[\![X \leq x]\!] = q$, is the quantile function.
- `rpois(r, lambda)` simulates r realizations of X.

Properties of Poisson Random Variables

In this section we are going to observe some important properties of Poisson random variables, which will come in very handy in the next few chapters.

Poisson Approximation of Binomial Distribution

In this section we are going to see how the Poisson distribution can be used to approximate a Binomial distribution when the number of trials considered is large and the probability of success p of those Bernoulli trials is small. This approximation is adequate say, for $n \geq 30$ and $p \leq 0.05$ and becomes more and more accurate as n increases and p decreases.

Theorem 1 (Law of Rare Events)

$$\lim_{\substack{n\to\infty, p\to 0 \\ np=\lambda}} \binom{n}{x} p^x (1-p)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!} \tag{0.29}$$

The convergence presented in Theorem 1 is called **convergence in distribution**, which is not the same as the usual convergence we are used to.

In our case, this means that, as the number of Bernoulli trials n increases and the probability of success p decreases in such a way that their product np remains constant, say equal to λ, the distribution of the Binomial random variable $X \sim \text{Binom}(n, p)$ approaches the distribution of the Poisson random variable $Y \sim \text{Poisson}(\lambda)$.

To make this more concrete, consider a sequence of random variables $X_n \sim \text{Binom}\left(n, \frac{\lambda}{n}\right)$, where $\frac{\lambda}{n} = p$, for some adequate value of λ, that is, $p = \frac{\lambda}{n} < 1$. We can notice that $np = \lambda$ but if $n \to \infty$ then $p = \frac{\lambda}{n} \to 0$. This means that the distribution of X_n approaches the distribution of $Y \sim \text{Poisson}(\lambda)$ as n increases. If $n \to \infty$ then $X_n \xrightarrow{\text{d}} Xl \sim \text{Poisson}(\lambda)$, where "$\xrightarrow{d}$" indicates *convergence in distribution*. To better understand this, it is useful to remember that each X_n can be seen as a function of ω: $X_n(\omega)$.

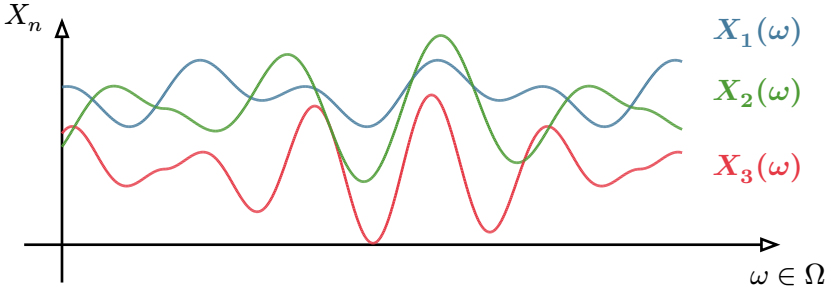

Figure 0.2: Different random variables X_n plots for the same outcome ω

Figure 0.2 shows that every time we perform the experiment, we get one outcome $\omega \in \Omega$ and each $X_n(\omega)$ gets its individual value x_n.

> ⚠ **Warning**
>
> Convergence in distribution **does not mean** that for each ω we have:
>
> $$X_n(\omega) = x_n \quad \text{and} \quad X(\omega) = x$$
>
> rather, it is interested in the **probability** of X_n taking values in certain intervals converging to the probability of X taking values in the same intervals as n goes to infinity.

Additivity of Poisson Random Variables

Another very important property of Poisson random variables is their **additivity**. Let's look at the following theorem:

Theorem 2 (Additivity of Poisson Random Variables)

If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are two **independent** Poisson random variables, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

To view this under a more practical light, consider two disjoint periods of time π_1 and π_2, say $\pi_1 = [0, t_1), \pi_2 = [t_1, t_2]$. Suppose for each of these periods we define a Poisson random variable, for instance $X \sim \text{Pois}(\lambda)$ counts the number of rare events occurring during π_1 and $Y \sim \text{Pois}(\mu)$ counts number of rare events occurring in π_2; where X and Y represent respectively the fact that we are expecting to observe λ events during π_1 and μ events to happen during the span of π_2: $\mathbb{E}[\![X]\!] = \lambda, \mathbb{E}[\![Y]\!] = \mu$.

Let's now consider the period $\Pi = [0, t_2]$ and define the random variable W as the number or rare events occurring during Π, intuitively also this variable is a Poisson. If we also remember that the parameter of a Poisson random variable models the expected number of events occurring during the time period, we can intuitively say that it makes sense to expect to observe $\lambda + \mu$ events during the period Π. Therefore we can conclude that $W \sim \text{Pois}(\lambda + \mu)$.

> ⚠ **Warning**
>
> The additivity property of Poisson random variables **only holds** when the random variables considered are **independent**.

Let's now try to prove the theorem in a formal way. By the notion of independence we know that the following equation holds:

$$\begin{aligned} \mathbb{P}[\![X = r \wedge Y = s]\!] &= \mathbb{P}[\![X = r]\!]\mathbb{P}[\![Y = s]\!] \\ &= \lambda^r \frac{e^{-\lambda}}{r!}\ \mu^s \frac{e^{-\mu}}{s!} \end{aligned}$$

Now we actually need to consider all possible ways of obtaining $W = n$, that is, we need to consider all the pairs (r, s) such that $r + s = n$. Therefore we can write:

$$\begin{aligned}\mathbb{P}[\![X+Y=n]\!] &= \sum_{r=0}^{n} \mathbb{P}[\![X=r \wedge Y=n-r]\!] \\ &= \sum_{r=0}^{n} \frac{\lambda^r e^{-\lambda}}{r!} \frac{\mu^{n-r} e^{-\mu}}{(n-r)!}\end{aligned}$$

We can multiply and divide everything inside the sum by $n!$, this is useful since now we can bring out of the summation all the terms that do not depend on r and obtain the following:

$$\begin{aligned}\mathbb{P}[\![X+Y=n]\!] &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{r=0}^{n} \binom{n}{r} \lambda^r \mu^{n-r} \\ &= \frac{(\lambda+\mu)^n \; e^{-(\lambda+\mu)}}{n!}\end{aligned}$$

which is exactly the p.m.f. of a Poisson random variable with parameter $\lambda+\mu$. This can be easily generalized to the sum of k independent Poisson random variables by *mathematical induction*. It is actually possible to generalize this property even further: we can even consider the case in which there is a **infinite countable** number of independent Poisson random variables.

Theorem 3 (Generalized Additivity of Poisson Random Variables)

Let $X_j \sim \text{Pois}(\lambda_j)$ for $j = 1, 2, \ldots$ be a sequence of independent random variables. If we have that $\sum_{j=1}^{\infty} \lambda_j = \lambda < \infty$, i.e., the series converges, then we have that:

$$\mathbb{P}\left[\!\!\left[S = \sum_{j=2}^{\infty} X_j < \infty \right]\!\!\right] = 1 \quad \text{and} \quad S \sim \text{Pois}(\lambda)$$

that is, the infinite sum of independent Poisson r.v.'s is still a Poisson r.v. If, on the other hand, the series $\sum \lambda_j = \infty$ then also the probability that the infinite sum diverges is equal to 1.

The type of convergence used by this theorem is called **almost sure convergence**, which is stronger than convergence in distribution. To better understand this, suppose we have a sequence $X_i \sim \text{Pois}(\lambda_i)$. We can define **partial sums**: $S_1 = X_1$, $S_2 = X_1 + X_2$, $S_3 = X_1 + X_2 + X_3$, and so on until S_n. Along with these random variables we can also define the partial sums of their parameters: $\mu_1 = \lambda_1$, $\mu_2 = \lambda_1 + \lambda_2$, $\mu_3 = \lambda_1 + \lambda_2 + \lambda_3$, and so on until μ_n, so that $S_n \sim \text{Pois}(\mu_n)$. If we have that $\mu_n \underset{\infty}{\to} \mu < \infty$, then we have that $S_n \xrightarrow{\mathbb{P}} S \sim \text{Pois}(\mu)$, where "$\xrightarrow{\mathbb{P}}$" indicates **convergence in probability**.

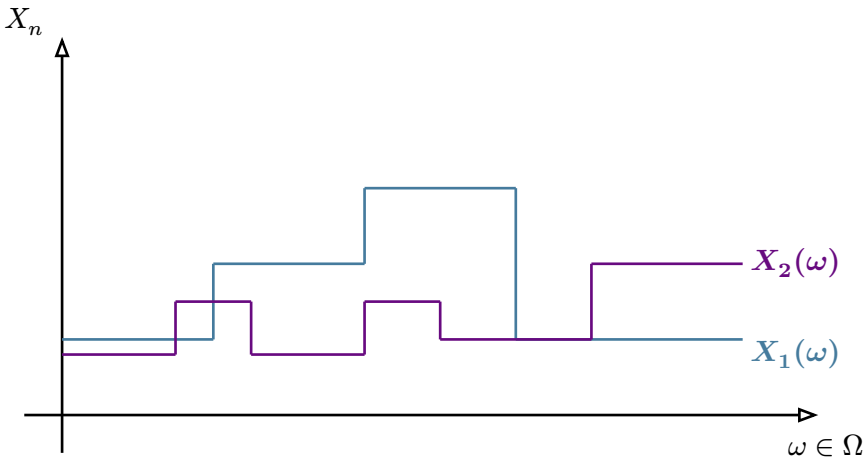

Figure 0.3: Two random variables X_1 and X_2 for the same outcome ω

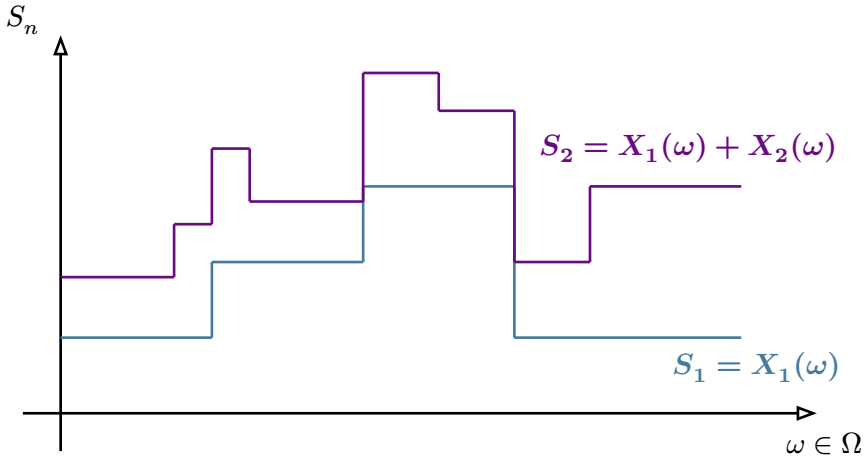

Figure 0.4: Partial sums S_1 and S_2 for the same outcome ω

Looking at the figures above and trying to consider a fixed value of ω, in Figure 0.3 the observed values of the sequence may not converge, but in Figure 0.4, for each ω the observed values $X_n(\omega) = x_n$ always converge, if n is large enough. The convergence works every time we perform the experiment. This happens thanks to the strong dependence that exists between the partial sums.

Poisson - Multinomial Relationship

Up to this point we have talked about sums, basically we have seen that if we know the values of the X's then we can easily compute the value of their sums. Now we would like to invert this relation. Suppose we know the value of a sum of Poisson random variables, we'd like to be able to infer something about the values of the individual Poisson r.v.'s that are being summed up. Let's look at the following theorem.

Theorem 4 (Poisson - Multinomial Relationship)

Let $S_n = X_1 + ... + X_n$ be the sum of n independent Poisson random variables each with parameter λ_i and let $\lambda = \lambda_1 + ... + \lambda_n$. The **conditional distribution** of the vector $\boldsymbol{X} = (X_1, ..., X_n)$ given the value of S_n is **multinomial** with its parameter being $\boldsymbol{p} = (\lambda_1/\lambda, ..., \lambda_n/\lambda)$.

Intuitively, this makes sense, because if we know that the total number of events observed is k, then the only uncertainty the remains is about how these k events are distributed among the different X_i's. This is exactly what a multinomial distribution models.

To see this in a more formal way, suppose we have $r_1 + r_2 + ... + r_n = s$, then we can write:

$$\begin{aligned}\mathbb{P}[\![X_1 = r_1, ..., \ X_n = r_n \mid S_n = s]\!] &= \frac{\mathbb{P}[\![X_1 = r_1, ..., \ X_n = r_n \ , S_n = s]\!]}{\mathbb{P}[\![S_n = s]\!]} \\ &= \frac{\Pi_{j=1}^n \left(\lambda_j^{r_j} \frac{e^{-\lambda_j}}{r_j!}\right)}{\lambda^s \frac{e^\lambda}{s!}} = \frac{s!}{\pi_{j=1}^n r_j!} \left(\frac{\lambda_1}{\lambda}\right)^{r_1} ... \left(\frac{\lambda_n}{\lambda}\right)^{r_n}\end{aligned}$$

where the first equality comes from the definition of conditional probability in ciaoneeq:conditional_cdf and the second equality is obtained by noticing that $S_n = s$ is a redundant condition once we have all the values of the X_i's and by multiplying the p.m.f.'s of the individual Poisson random variables. As far as the third equality is concerned the λ^s has been replaced by a product of $\lambda^{r_1} \cdot ... \cdot \lambda^{r_n} = \lambda$, and the other simplifications are highligted in green.

> **Remark**
>
> Notice that, in case $n = 2$, the multinomial distribution reduces to a *Binomial distribution*. Given $S_2 = s$, if $X_1 = r$ and $X_2 = s - r$, we have that:
>
> $$\begin{aligned}\mathbb{P}[\![X_1 = r, \ X_2 = s - r \mid S_2 = s]\!] &= \mathbb{P}[\![X_1 = r \mid S_2 = s]\!] \\ &= \binom{s}{r} p^r (1-p)^{s-r}\end{aligned}$$
>
> where $p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$.

> **Remark**
>
> In a very similar fashion it is possible to do the opposite, that is, let $S \sim \text{Pois}(\lambda)$ and assume that, conditionally on S, X has a $\text{Binom}(S, p)$ distribution. Then X and $Y = S - X$ are **independent** Poisson random variables with parameters $\lambda_1 = \lambda p$ and $\lambda_2 = \lambda(1-p)$.

To see why this last remark is true, we can produce the following derivation:

$$\begin{aligned}\mathbb{P}[\![X = r, \ S - X = k]\!] &= \mathbb{P}[\![S = k + r]\!] \mathbb{P}[\![X = r \mid S = k + r]\!] \\ &= \frac{\lambda^{k+r} e^{-\lambda}}{(k+r)!} \binom{k+r}{r} p^r (1-p)^k \\ &= \frac{(\lambda p)^r e^{-\lambda p}}{r!} \ \frac{(\lambda(1-p))^k e^{-\lambda(1-p)}}{k!}\end{aligned}$$

Here, instead of starting from the conditional, and writing it as the ratio between the joint and the marginal, we have started from the joint of X, Y being equal to r, k writing it as the product of the marginal of S and the conditional of X given S. The marginal of S can be found by noticing that $S \sim \text{Pois}(\lambda)$. The conditional of X given S is a Binomial distribution, indeed we want to estimate the probability of having exactly r successes out of $k + r$ trials, where the probability of success is p.

Notice how $\mathbb{P}[\![X = r \mid S = n]\!] = \binom{n}{r} p^r (1-p)^{n-r} \overset{\substack{n=k+r \\ n-r=k}}{\Longrightarrow} \binom{k+r}{r} p^r (1-p)^{k+r-r}$. And we have written $e^{-\lambda}$ as $e^{-\lambda(p+1-p)} = e^{-\lambda p} \ e^{-\lambda(1-p)}$ which has been later split into the two partes in the last equality. We can notice that the two factors are exactly the p.m.f.'s of two independent Poisson random variables with parameters λp and $\lambda(1-p)$ respectively.

In the end, the important takeaway is that, with this type of random variables, if we have the marginals, we can also derive the conditionals and vice-versa. This property may look trivial but it is actually quite unique in the whole world of probability distributions.

0 Exponential Distribution

Another very important distribution that is often used to model 'natural' phenomena is the **Exponential distribution**. Specifically it is often used to model **time**: waiting times, inter-arrival times, hardware lifetime, failure times and so on.

We are not going to spend time on giving the definition of this distribution, since it is pretty much all contained in the few lines above. Similarly to Poisson random variables, exponential random variables are also characterized by a **rate** parameter λ which models the expected number of events occurring per unit of time.

Probability Density Function

Since this distribution models time, it is only possible to define it in a continuous fashion. Let $X \sim \text{Exp}(\lambda)$ be an exponential random variable with parameter $\lambda > 0$. The **probability density function** of X is defined as:

$$f_X(x) = \lambda e^{-\lambda x} \quad \forall x > 0 \tag{0.30}$$

By means of this p.d.f. we can also write the **cumulative distribution function** as follows:

$$F_X(x) = \mathbb{P}[\![X \leq x]\!] = \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x} \tag{0.31}$$

Sometimes it may be really useful to deal with the **survival function**, which is defined as in ciaoneeq_02_survival_function, therefore in this specific case we have:

$$S_X(x) = \mathbb{P}[\![X > x]\!] = 1 - F_X(x) = e^{-\lambda x} \tag{0.32}$$

Expected Value and Variance

Since we know the expected value of a Poisson random variable with parameter λ is exactly λ, we can leverage this, and obtain that, since λ here models the amout of 'rare' events occurring per unit of time, the **expected** waiting time for the occurrence of one such event is:

$$\mathbb{E}[\![X]\!] = \frac{1}{\lambda} \tag{0.33}$$

To see this in a more formal way, we can compute the expected value as follows:

$$\mathbb{E}[\![X]\!] = \int_0^\infty x \ \lambda e^{-\lambda x} \ dx \quad \text{integrating by parts: } \begin{cases} u = x, \ du = dx \\ dv = \lambda e^{-\lambda x}, \ v = -e^{-\lambda x} \end{cases}$$

$$= \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} \ dx = 0 - \left[\left(\frac{e^{-\lambda x}}{\lambda}\right)\right]_0^\infty = \frac{1}{\lambda}$$

As far as the variance is concerned we need to first compute the value of $\mathbb{E}[\![X^2]\!]$:

$$\mathbb{E}[\![X^2]\!] = \int_0^\infty x^2 \ \lambda e^{-\lambda x} \ dx \quad \text{integrating by parts: } \begin{cases} u = x^2, \ du = 2x \\ dv = \lambda e^{-\lambda x}, \ v = -e^{-\lambda x} \end{cases}$$

$$= \left[-x^2 e^{-\lambda x}\right]_0^\infty + \int_0^\infty 2xe^{-\lambda x} \ dx = 0 + \frac{2}{\lambda} \int_0^\infty \lambda x e^{-\lambda x} \ dx = \frac{2}{\lambda^2}$$

Now we can compute the **variance** as follows:

$$\text{Var}\{x\} = \mathbb{E}[\![X^2]\!] - \mathbb{E}[\![X]\!]^2 = \left(\frac{2}{\lambda^2}\right) - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2} \tag{0.34}$$

R Implementation

As usual, in `R` we have the following functions to work with exponential random variables:

- `dexp(x, lambda)` $= f_X(x)$, is the probability density function (p.d.f.).
- `pexp(x, lambda)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function (c.d.f.).
- `qexp(q, lambda)` $= x = F^{-1}(q)$, i.e., $\mathbb{P}[\![X \leq x]\!] = q$, is the quantile function.
- `rexp(r, lambda)` simulates r realizations of X.

Poisson - Exponential Relationship

As we have already hinted, there is a very strong relationship between Poisson and Exponential random variables. This section is dedicated to exploring it in detail.

Consider a sequence or *rare* events, where the number N_t of occurrences during a period of time of length t is modeled as a Poisson random variable with parameter λ proportional to t. In other words we can write $N_1 \sim \text{Pois}(\lambda), N_t \sim \text{Pois}(\lambda t)$.

Consider now the event $A =$ "the time T until the next event (arrival) is greater than t". This is basically equivalent to saying that "during a period of time of length t no events occur": we can write the event as $A = \{N_t = 0\}$. If we try to *compute the probability of* A we get the following:

$$\mathbb{P}[\![A]\!] = \mathbb{P}[\![T > t]\!] = \mathbb{P}[\![N_t = 0]\!] = e^{-\lambda t}$$

This is because, if we take a look at the p.m.f. of a Poisson random variable in Equation 0.26, we can see that our parameter when considering the time period of length t is exactly λt, therefore if we set $x = 0$ we get exactly the following equality:

$$p_X(x) = e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda t}$$

But this is exactly equal to the **survival function** of an Exponential random variable with parameter λt as shown in Equation 0.32. This property is also known as the **inevitability of exponential distribution**.

Properties of Exponential Random Variables

We have defined quite a bit of properties about Poisson random variables; not surprisingly, since they are so tightly related to Exponential ones, many of these properties can be translated to Exponential random variables as well.

Memoryless Property

One of the most important properties of Exponential random variables is their **memoryless property**. We are going to present it in the following theorem.

Theorem 1 (Memoryless Property for Exponential)

Suppose that an exponential random variable T represents a waiting time. Regardless of the event $T > t$, when the total waiting time exceeds t, the remaining waiting time still has exponential distribution with the same parameter λ.

$$\mathbb{P}[\![T > t + x \mid T > t]\!] = \mathbb{P}[\![T > x]\!] \quad \text{for } t, x > 0$$

where t represents the portion of waiting time that has already elapsed, and x represents the additional remaining time.

This can be proved by recognizing the survival function of the exponential distribution:

$$\begin{aligned}\mathbb{P}[\![T > t + x \mid T > t]\!] &= \frac{\mathbb{P}[\![T > t + x \cap T > t]\!]}{\mathbb{P}[\![T > t]\!]} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} \\ &= e^{-\lambda x} = \mathbb{P}[\![T > x]\!]\end{aligned}$$

Remark

Just like the **geometric distribution** is the only discrete memoryless distribution, the **exponential distribution** is the only continuous memoryless distribution.

This is by no surprise, since we have already seen that the Binomial distribution is related to the Geometric distribution in a very similar way as the Poisson distribution is related to the Exponential distribution.

Minimization Property

When we talked about the Poisson distribution, we have mentioned the case in which we may have different independent Poisson random variables and we were interested in their sum, in which case the sum was still a Poisson random variable, and on the other hand we have also seen that if we had the sum of independent Poisson random variables, conditionally on the value of the sum, the individual random variables were multinomially distributed.

Since exponential random variables do not consider the number of events occurring but rather times until the occurrence of such events, it makes sense to consider the **minimum** of a set of independent exponential random variables.

Theorem 2 (Minimization Property)

Let $X_j \sim \text{Exp}(\lambda_j)$ for $j = 1, 2, ..., n$ be a collection of **independent** exponential random variables, then we have that:

$$L_n = \min\{X_1, X_2, ..., X_n\} \sim \text{Exp}(\lambda)$$

where $\lambda = \sum_{j=1}^{n} \lambda_j$ is the parameter of the resulting exponential random variable.

Indeed, we can take a look at the cumulative distribution function of L_n:

$$
\begin{aligned}
F_{L_n}(x) &= \mathbb{P}[\![L_n \leq x]\!] = 1 - \mathbb{P}[\![L_n > x]\!] \\
&= 1 - \mathbb{P}[\![X_1 > x,\, X_2 > x,\, ...,\, X_n > x]\!] \\
&= 1 - \prod_{j=1}^{n} \mathbb{P}[\![X_j > x]\!] \quad \text{by independence} \\
&= 1 - \prod_{j=1}^{n} e^{-\lambda_j x} = 1 - e^{\sum_{j=1}^{n} -\lambda_j x} = 1 - e^{-\lambda x}
\end{aligned}
$$

where we went from the first to the second line by noticing that the minimum of n r.v.'s is greater than x if and only if all the individual r.v.'s are greater than x. In the end we have obtained exactly the survival function of an exponential random variable with parameter λ, as we were supposed to.

> **Remark**
>
> It is important to notice that the minimum L_n of independent exponential random variables is **not independent** of $\{X_1, X_2, ..., X_n\}$. Indeed $L_n = X_k$ for some $k \in \{1, ..., n\}$ and:
>
> $$\mathbb{P}[\![L_n = X_k]\!] = \frac{\lambda_k}{\lambda} = \frac{\lambda_k}{\sum_{j=1}^{n} \lambda_j}$$
>
> Indeed, by the law of total probability for continuous random variables we have that:
>
> $$
> \begin{aligned}
> \mathbb{P}[\![L_n = X_k]\!] &= \int_0^{\infty} \mathbb{P}\left[\!\!\left[\bigcap_{j \neq k} (X_k < X_j) \mid X_k = x \right]\!\!\right] f_{X_k}(x)\, dx \\
> &= \int_0^{\infty} \mathbb{P}\left[\!\!\left[\bigcap_{j \neq k} X_j > x \right]\!\!\right] f_{X_k}(x)\, dx = \int_0^{\infty} \left(\prod_{j \neq k} e^{-\lambda_j x} \right) \left(\lambda_k e^{-\lambda_k x} \right) dx \\
> &= \lambda_k \int_0^{\infty} e^{-(\lambda_1, ..., \lambda_n) x}\, dx = \frac{\lambda_k}{\sum_{j=1}^{n} \lambda_j}
> \end{aligned}
> $$
>
> where we switched from the first to the second equality by noticing that the all the X_j's are independent from the conditioning event $X_k = x$. It is clear that $\sum_{k=1}^{n} \mathbb{P}[\![L_n = X_k]\!] = 1$.

The above result can actually be generalized to the case in which we have a countably infinite number of independent exponential random variables. If the same of the parameters converges to a finite value, say λ, when we have that:

$$\mathbb{P}[\![L_n = x]\!] = \frac{\lambda_k}{\lambda} \quad \text{where } \lambda = \sum_{j=1}^{\infty} \lambda_j < \infty$$

> **Warning**
>
> The same reasoning **cannot** be applied to the **maximum** of two or more independent exponential random variables. Indeed:
>
> $$
> \begin{aligned}
> \mathbb{P}[\![\max\{X_1, X_2\} \leq x]\!] &= \mathbb{P}[\![X_1 \leq x,\, X_2 \leq x]\!] = \mathbb{P}[\![X_1 \leq x]\!]\, \mathbb{P}[\![X_2 \leq x]\!] \\
> &= \left(1 - e^{-\lambda_1 x}\right) \left(1 - e^{-\lambda_2 x}\right)
> \end{aligned}
> $$

which cannot be simplified to the c.d.f. of an exponential random variable. Intuitively, this happens because we cannot reduce the c.d.f to a survival function as we did in the original case with the minimum.

0 Gamma Distribution

When dealing with discrete random variables, we have seen how, given a Bernoulli trial and a series of independent repetitions of it, we can define more and more complex distributions, arriving at the negative binomial distribution, which models the number of trials until a fixed number of successes is observed, which in practice is the sum of independent geometric random variables.

Now we are going to se how, starting from a bunch of independent exponential random variables, we can define a more general distribution, called the **Gamma distribution**, which models the behavior of their sum.

Definition 12 (Gamma Distribution)

Let $X_1, X_2, ..., X_n$ be independent exponential random variables with parameter λ. The **Gamma distribution** with parameters $\alpha = n$, $\lambda = \lambda$ serves to model the distribution of the sum $Y = X_1 + X_2 + ... + X_n$. We write $Y \sim \text{Gamma}(\alpha, \lambda)$.

♣

The parameter α is often called the **shape parameter**, while λ is called the **rate parameter**. We can notice that when $\alpha = 1$ we have that the Gamma distribution reduces to the Exponential distribution.

Probability Density Function

The **probability density function** of a Gamma random variable with parameters α, λ is defined as follows:

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \; x^{\alpha-1} e^{-\lambda x} \quad \forall x > 0 \tag{0.35}$$

Intuitively the factor $x^{\alpha-1}$ is used to gradually decrease the speed of the exponential decay, this is mainly the reason why, for the shape parameter $\alpha = 1$ the distribution reduces to the exponential one. Figure 0.5 shows how the α parameter affects the shape of the p.d.f. of a Gamma random variable.

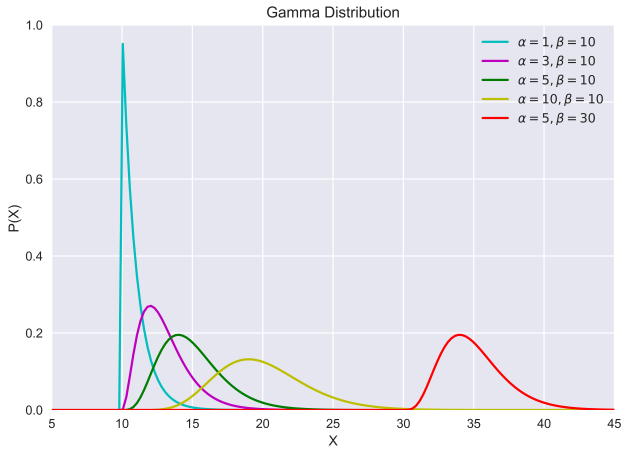

Figure 0.5: Probability density function of a Gamma random variable for different values of the shape parameter

Gamma Function

In the above definition we have used the **Gamma function** $\Gamma(\alpha)$ which can be seen as a generalization of the factorial function to real numbers. Suppose we are to write an algorithm that computes the factorial of a number n: in this case we must proceed **recursively**, since the factorial is defined by induction.

For the Gamma function we are going to proceed similarly, the only difference is that we are going to define it for any $\alpha \in \mathbb{R}$. One may therefore think that we can define it as follows:

$$\Gamma(\alpha) = \alpha \cdot \Gamma(\alpha - 1)$$

The problem is the **starting point** (the base case of induction). This was easy in case of the factorial in that we stopped when reaching $0! = 1$. Here we need to proceed differently, before defining the function formally it is important to start from the mathematica (probabilistic, actually) motivation behind it. Consider a random variable with the following probability density function:

$$f_X(x) = k\ x^{\alpha-1} e^{-\lambda x} \quad \forall \alpha, x > 0$$

where k is a normalizing constant that makes the area under the curve equal to 1. To find the value of k we need to solve the following integral.

$$1 = k \int_0^\infty x^{\alpha-1}\ e^{-\lambda x}\ dx$$

To solve this we proceed by parts, identifying the following: $\begin{cases} u = x^{\alpha-1},\ du = (\alpha-1)x^{\alpha-2} \\ dv = e^{-\lambda x} dx,\ v = -\frac{e^{-\lambda x}}{\lambda} \end{cases}$. If $\alpha = n$, integration by parts takes us to the following:

$$\frac{1}{k} = \int_0^\infty x^{n-1}\ e^{-\lambda x}\ dx = \left[x^{n-1}\frac{e^{-\lambda x}}{-\lambda}\right]_0^\infty - \int_0^\infty (n-1)x^{n-2}\ \frac{e^{-\lambda x}}{(-\lambda)}$$

$$= 0 + \frac{n-1}{\lambda}\int_0^\infty x^{n-2}\ e^{-\lambda x}\ dx$$

$$= (n-1)\frac{n-2}{\lambda^2}\int_0^\infty x^{n-3}\ e^{-\lambda x}\ dx = \ldots = \frac{(n-1)!}{\lambda^n}$$

Therefore we can conclude that $k = \frac{\lambda^n}{(n-1)!}$. This is not the normalization factor we find in Equation 0.35, that's because this is only valid in case α is an integer.

When lambda is not an integer, the exponent of the x's that get derived inside the integration by parts will not reach 0 and we will not be able to stop the process, instead they may reach negative values. In general we have that the equation

$$\frac{1}{k} = \int_0^\infty x^{\alpha-1}\ e^{-\lambda x}\ dx$$

has not a closed (analytical) form solution for any α. To solve this, a named function has been defined, called the **Gamma function**, which has actually been computed by setting $\boldsymbol{\lambda = 1}$ and is defined as $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\ dt$.

Expected Value and Variance

Following we are going to provide the formulas to compute the expected value and the variance of a Gamma random variable with parameters α, λ.

$$\mathbb{E}[\![X]\!] = \frac{\alpha}{\lambda} \qquad \mathrm{Var}\{X\} = \frac{\alpha}{\lambda^2} \tag{0.36}$$

Additivity of Exponential Random Variables

Now that we have defined the Gamma distribution it is important to mention that, if we have n independent random variables $X_j \sim \mathrm{Exp}(\lambda)$ for $j = 1, 2, \ldots, n$, then their sum $S_n = X_1 + X_2 + \ldots + X_n$ is a Gamma random variable with parameters $\alpha = n$ and $\lambda = \lambda$.

R Implementation

As usual, in `R` we have the following functions to work with Gamma random variables:

- `dgamma(x, shape, rate)` $= f_X(x)$, is the probability density function
- `pgamma(x, shape, rate)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function
- `qgamma(q, shape, rate)` $= x = F^{-1}(q)$, i.e., $\mathbb{P}[\![X \leq x]\!] = q$, the quantile function
- `rgamma(r, shape, rate)` simulates r realizations of X.