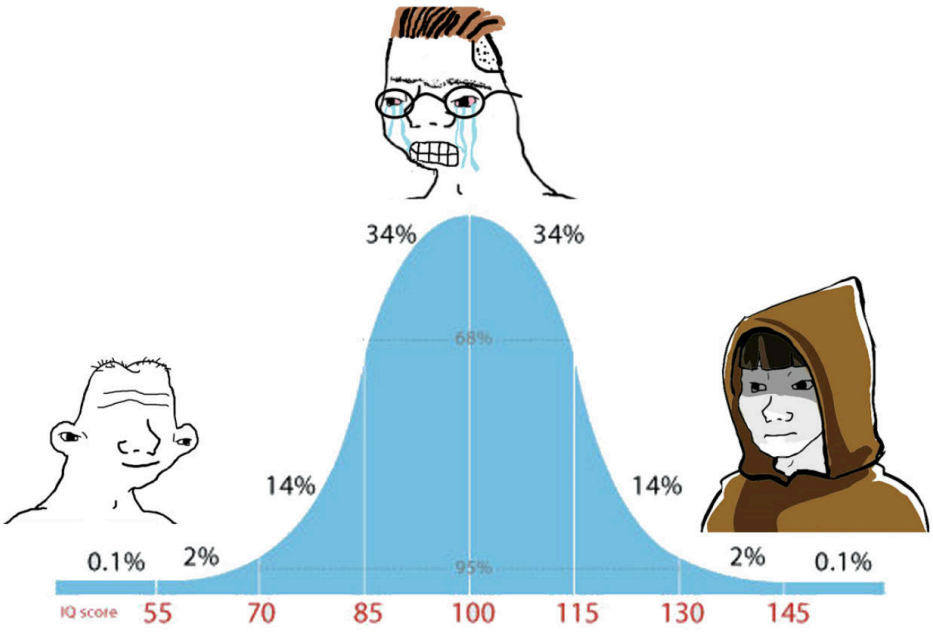
Contents


34%
34%
68%
14%
14%
0.1%
2%
95%
2%
0.1%
IQ score
55
70
85
100
115
130
145

1. Descriptive statistics

1.1. Introduction

Descriptive statistics instructs how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.

Collections of facts are called **data**: descriptive statistics provides methods for organizing, summarizing and drawing conclusions based on information contained in the data. This is done through graphical representations, called **plots**, or through **summary measures**, numbers that can capture on their own an aspect of the entire data.

A statistical enquiry will typically focus on a well-defined collection of objects, a **population**. A population can be constituted by a finite or infinite number of objects. Each object in the population possesses many features, which may or may not be of interest.

Any feature whose value can change from object to object in the population and that is relevant with respect to the statistical enquiry is called a **variable**. Variables are generally denoted with uppercase letters. For a variable X, the set $D(X)$ that contains all possible values that X can take is called the **support** of X.

When the value of all variables is known for all objects in the population, this is referred to as **census**. In most situations, a census is impossible to obtain, either because it would be too expensive or too time consuming to do so or because the population is made of infinitely many objects. If this is the case, a solution is to extract a subset of the population, called **sample**, and operate on such sample. A sample should not be too small, otherwise it might fail to capture all the nuances of the population, but not too big, or it would be hard to manipulate.

Exercise 1.1.1: Suppose there's interest in analyzing an aspect of the population that lives in a certain town. Being impractical to ask each person, the idea is to extract a reasonably sized sample. Which of the two approaches here presented is preferable?

- Picking each person entering an elementary school in a day;
- Picking each person entering a supermarket in a day.

Solution: The second one, because it is more likely to capture as many different people as possible. □

The different values that are attained by each element are called **observations**. Observations are denoted with the lowercase counterpart of the symbol used for the variable. That is, if X is a variable, then $x_1, x_2, ..., x_n$ are the n observations in a sample, ordered with a certain criteria: x_1 is the value of X observed for the first element of the sample, x_2 is the value of X for the second element, ecc...

Variables are classified into **numerical** variables and **categorical** variables. Numerical variables can either be **discrete** or **continuous**. Numerical variables are discrete if their support is either a finite set or a countably infinite set. Numerical variables are continuous if their support is a uncountably infinite set.

Categorical variables can either be **ordinal** or **nominal**. Categorical variables are ordinal if the set of its possible values obeys an objective hierarchy or ordering of sort, otherwhise are called nominal. Categorical variables can exist in a continuum, but are much more likely to take finite values.

Exercise 1.1.2: Provide an example for each of the four types of variables.

Solution:
- A numerical discrete variable is the number of items sold in a store, since such number is necessarely an integer (it's not possible to sell half an item, or three quarters of an item);
- A numerical continuous variable is the temperature measured with a thermometer, since this value is a real number;
- A categorical ordinal variable is the level of satisfaction of a customer ("very good", "good", "ok", "bad", ecc...). This is because, despite not being numbers, it's still possible to arrange them in an hierarchy ("very good" is higher than "good", "ok" is higher than "bad", ecc...);
- A categorical nominal variable is the color of a dress, since it's impossible to state (objectively) that a certain color is "better" or "greater" or "higher" than another.

□

The most straightforward way to organize the collected observations is to arrange them into a table, with each row representing the observation for a single element and each column representing one of the variables. The i,j-th entry of this table represents the observation of the j-th variable of the i-th individual.

This tabulated structure, the starting point for any statistical enquiry, is also referred to as a **dataset**. A **univariate** data set consists of a (tabulated) collection of observations on a single variable of a sample or population. If the variables are two, the dataset is **bivariate**. If the variables are three or more, the dataset is **multivariate**.

Petal Length	Petal Width	Species
1.4	0.2	Iris Setosa
1.3	0.2	Iris Setosa
1.5	0.2	Iris Setosa
1.4	0.2	Iris Setosa
1.7	0.4	Iris Setosa
4.7	1.4	Iris Versicolor
4.5	1.5	Iris Versicolor
4.9	1.5	Iris Versicolor
4.6	1.5	Iris Versicolor
4.5	1.3	Iris Versicolor
6	2.5	Iris Virginica
5.1	1.9	Iris Virginica
5.9	2.1	Iris Virginica

Petal Length	Petal Width	Species
5.6	1.8	Iris Virginica
6.6	2.1	Iris Virginica

Table 1: Small portion of the known `Iris` dataset by Ronald Fisher. This dataset contains information regarding three species of flowers, *Iris Setosa*, *Iris Versicolor* and *Iris Virginica*. The first and second column contain, respectively, the length and the width of the petal of the flower (in centimetres), while the third column contains the species of the flower. This is an example of multivariate dataset, with two numerical variables and a categorical variable.

A known graphical representation of bivariate numerical datasets is the **scatter plot**. A scatter plot is made up by of dots drawn in a two-dimensional Cartesian plane; each dot represents an object in the sample and its x, y coordinates are the observations of the respective variables for that element. A scatter plot gives an insight on how the data is "spread out" in space.

A scatter plot is also possible when the number of variables is three. If all three variables are numerical, it entails drawing a three-dimensional Cartesian plane and each point would have x, y, z coordinates, one for each variable. A three-dimensional scatter plot is very hard to read however, which is why it is not so common.

If instead the three variables are two numerical and one categorical, a scatter plot is effective. The idea is to draw a two-dimensional Cartesian plane, assigning the x, y coordinates to the two numerical variables, and representing the categorical variable either by drawing dots of different colors or dots of different shapes.

Exercise 1.1.3: Draw a scatter plot of Table 1.

Solution:

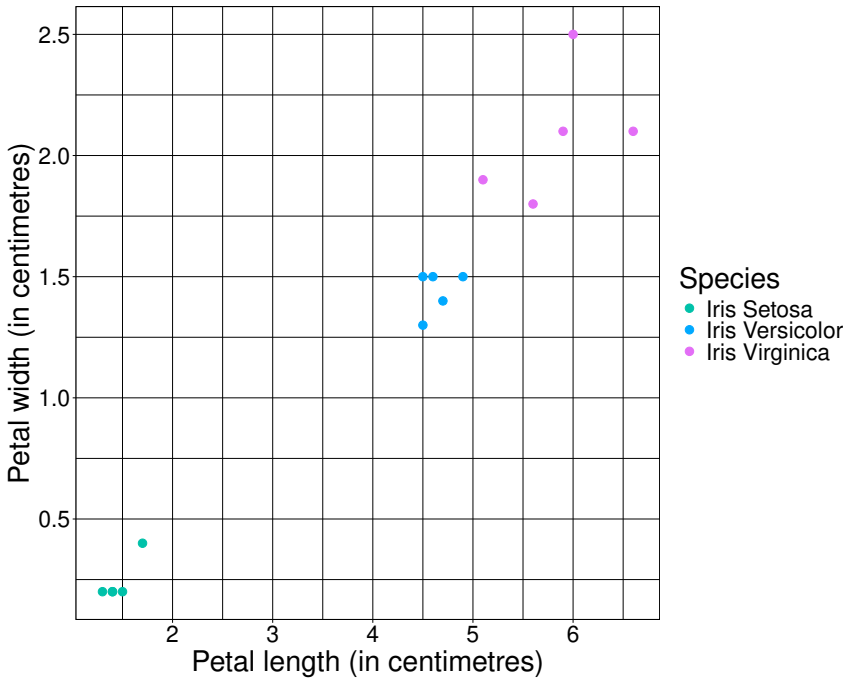


Figure 2: Scatterplot of Table 1. The two numerical variables occupy the x and y axes, while the categorical variable (the species) is mapped to the color of the dots.

□

1.2. Frequencies for a single variable

Consider a single discrete variable X. Suppose that a univariate dataset is constructed from a sample of size n of observations $O = (x_1, x_2, ..., x_n)$ of X.

Any statistical enquiry should start from the **absolute frequencies** of the values in the support of X. For a given $d_i \in D(X)$, the absolute frequency f_i is defined as the number of observations in the sample whose value is d_i.

$$f_i = |\{x \mid x \in O, x = d_i\}|$$

From the absolute frequency it is possible to define what is called the **cumulative absolute frequency** F_i of $d_i \in D(X)$ as the sum of the absolute frequencies of all $d_j \in D(X)$ such that $d_j \leq d_i$.

$$F_i = \sum_{j:d_j \leq d_i}^{|D(X)|} f_j$$

The **relative frequency** p_i of $d_i \in D(X)$ is given by the ratio between the absolute frequency f_i and the sample size n.

$$p_i = \frac{f_i}{n}$$

Relative frequencies are often preferred to absolute frequencies, because by definition they all lie in the $[0, 1]$ interval, no matter the order of magnitude of the observations. This gives "fairer" comparisons between the frequencies of different variables, especially if their support varies greatly.

From the relative frequency it is possible to define what is called the **cumulative relative frequency** P_i of $d_i \in D(X)$, given by the sum of the relative frequencies of all $d_j \in D(X)$ such that $d_j \leq d_i$.

$$P_i = \sum_{j:d_j \leq d_i}^{|D(X)|} p_j$$

An element in the support of X could never appear among the observations. That is, X could attain a certain value *in theory*, but never appear in the sample. It is therefore perfectly valid for a member of $D(X)$ to have an absolute/relative frequency equal to 0. Taking into account the absolute/relative frequency of absent data might make sense if $D(X)$ is a finite set. If $D(X)$ is a discrete but infinite set (the set of natural numbers, for example), members of the support with frequency 0 give no meaningful insights.

It is common to compute all four frequencies for all elements of the sample and then arrange them into a **frequency table**.

Exercise 1.2.1: Suppose that the number of rooms in a sample of 80 flats has been counted, and reported in the following table:

3	4	2	6	5	2	4	4	2	5	4	4	5	7	5	4	5	7	8	4	3	6	2	3	5	2	7
2	4	8	4	2	6	5	4	4	6	5	3	3	8	5	2	5	6	5	5	4	2	6	4	5	5	7
3	4	3	3	3	4	4	3	4	6	4	3	7	4	4	6	4	2	4	4	6	3	2	3	5	4	

Construct the frequency table.

Solution:

Number of rooms	Absolute frequency	Relative frequency	Cumulative absolute frequency	Relative absolute frequency
2	11	0.138	11	0.138
3	13	0.162	24	0.3
4	24	0.3	48	0.6
5	15	0.188	63	0.787
6	9	0.112	72	0.9
7	5	0.062	77	0.963
8	3	0.038	80	1

Table 3: The frequency table for the given dataset

□

Sometimes, computing frequencies and arranging them in a frequency table is not particularly informative. A visual and more direct representation of the absolute/relative frequency is the **bar plot**, or **histogram**. A bar plot is constituted by a Cartesian plane where on the x axis lie the members of the support of the variable. For each, a rectangle is drawn with said member as base and whose height is proportional to its absolute/relative frequency.

Exercise 1.2.2: Consider Exercise 1.2.1. Draw a box plot of the relative frequencies.

Solution:

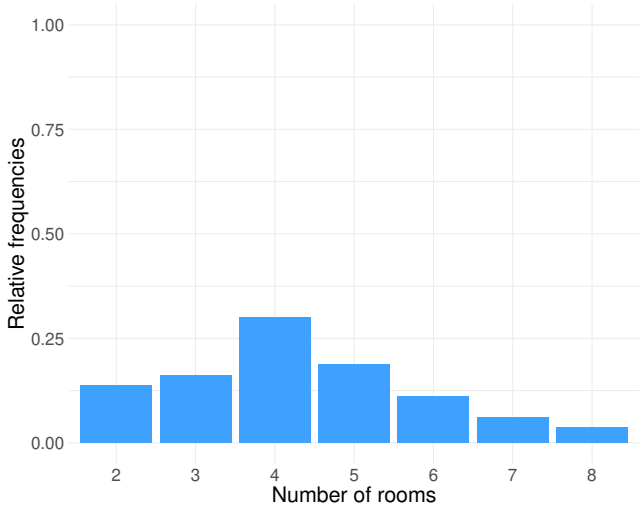


Figure 3: The bar plot of the relative frequencies for the given dataset.

□

An alternative graphical representation is the **pie chart**, where a circle is drawn and then partitioned into slices that are proportional to the absolute/relative frequencies.

Exercise 1.2.3: Consider Exercise 1.2.1. Draw a pie chart of the absolute frequencies.

Solution:

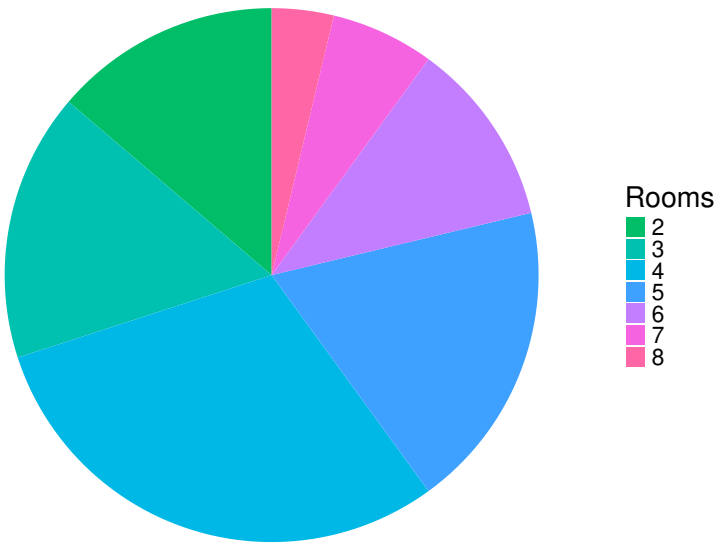


Figure 4: The pie chart of the absolute frequencies for the given dataset.

□

If the variable at hand is categorical, it is still possible to compute a frequency table as long as the support of the variable is a finite or countable set. If the variable is continuous, it is clearly impossible to compute the frequencies "as is", because its support is an uncountable set.

To circumvent the problem, the observations are partitioned into **classes**: each of these disjointed sets contains the observations that fall in a given range. The frequencies are then computed with respect to the size of these classes, which are in a finite number. The frequency of a class is an estimate of the frequency of all the elements that lie inside said class.

The number of classes in which to partition the data can be chosen arbitrarely, as well as the size of the classes. However, these are choices that are to be taken sensibly, since a poor partition choice can give misleading results. In general, a greater number of observations requires a greater number of classes. As a rule of thumb:

$$\text{number of classes} \approx \sqrt{\text{number of observations}}$$

In the simplest scenario, all classes are of the same size. However, it is also perfectly valid to have classes of different sizes. This is a sensible choice when the data is unevenly distributed, with some subsets of the support much more present in the observations than others. Ideally, the classes should be smaller for overrepresented subsets of the support and larger for underrepresented subsets.

A class may or may not contain its own extremes. If a class contains its rightmost value (the highest) it is said to be **closed on the right**. If a class contains its leftmost value (the lowest), it is said to be **closed on the left**. A class can be closed only on the right, only on the left, both on the right and on the left or neither. It is common practice to have classes that are closed on the left but not on the right.

Exercise 1.2.4: Suppose that the level of cholesterol in a sample of 40 patients has been measured, and reported in the following table:

213	174	193	196	220	183	194	200	192	200	200	199	204	191	227	183	178	183	221	204
188	193	221	212	187	181	193	205	196	211	202	213	216	206	195	191	171	194	184	191

Compute the frequency table.

Solution: This variable is not discrete, but continuous. The lowest value is 171, whereas the highest is 227. Since $\sqrt{40} \approx 6.324$, the dataset can be partitioned into the following classes:

[171-180) [180-190) [190-199) [199-208) [208-218) [218-227)

The frequency table with respect to these classes is as follows:

Class	Absolute frequency	Relative frequency	Cumulative absolute frequency	Relative absolute frequency
(171,180]	3	0.075	3	0.075
(180,190]	7	0.175	10	0.25
(190,199]	13	0.325	23	0.575
(199,208]	8	0.2	31	0.775
(208,218]	5	0.125	36	0.9

Class	Absolute frequency	Relative frequency	Cumulative absolute frequency	Relative absolute frequency
(218,227]	4	0.1	40	1

Table 6: The frequency table for the given dataset

□

A bar plot can be drawn for continuous variables as well, where the height of the bars is proportional to the frequency of the classes. If the classes are of different sizes, the width and the height of the bars ought to be adjusted in order to take into account the size of the class. This is done by drawing bars whose width is proportional to the size of the class and whose height is proportional to the **frequency density**, the frequency divided by the size.

The same could be done for a pie chart, scaling the radius of the slices with respect to the size of the classes. However, this would result in a hardly readable plot. This is one of the reasons why bar plots are to be preferred, at least for continuous data, to pie charts.

1.3. Central tendency indices

Visual summaries of data can give useful insights on its overall structure, but for a statistical enquiry this is not enough. A more formal analysis of data often requires the calculation and interpretation of numerical **summary measures**. A summary measure is a single value computed from the data that captures, on its own, one or more aspects of the data.

Summary measures that provide information regarding the "center" of the sample are called **central tendency indices**.

1.3.1. Sample mean

Given a discrete numerical variable X, let $x_1, x_2, ..., x_n$ be the observations of X collected from a sample of size n. The **sample mean** $\overline{x}$ of the variable X is a summary measure that describes its average value:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The sample mean may or may not be itself an element of the sample.

Exercise 1.3.1.1: What is the sample mean of the variable $X = \text{Number of rooms}$ in Exercise 1.2.1?

Solution:

$$\overline{x} = \frac{3 + 4 + 2 + 6 + 5 + 2 + 4 + ... + 3 + 5 + 4}{80} \approx 4.312$$

□

The sample mean is a simple and effective summary measure, but has its fair share of problems. For example, it is very sensitive to **outliers**, observations that significantly deviate from the other

observations. A single observation that is very different from the others can skew the sample mean considerably.

Lemma 1.3.1.1: If $\overline{x}$ is the sample mean of a given sample, a linear transformation $f(x) = ax + b$ with $a, b \in \mathbb{R}$ applied to all elements of the sample gives a new sample whose sample mean is $a\overline{x} + b$. That is, if an entire sample is linearly transformed, the sample mean is transformed in accord.

Proof: Let $y_i = ax_i + b$ be the i-th element of the transformed sample. This gives:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} ax_i + b}{n} = \frac{a\sum_{i=1}^{n} x_i}{n} + \frac{\sum_{i=1}^{n} b}{n} = a\overline{x} + \frac{b\cancel{n}}{\cancel{n}} = a\overline{x} + b$$

□

The sample mean for a continuous variable can be computed using the exact same formula as above. An alternative definition is the "averages of the averages": computing the sample mean for each class and then averaging the result among classes. This is faster, but less accurate.

If the variable is categorical, it makes no sense to define a mean. Even if there's an order for the support of the variable, an average value would still be meaningless (what is the average between "good" and "very good"?).

1.3.2. Sample median

Given a numerical variable X (discrete or continuous), let $x_1, x_2, ..., x_n$ be the observations of X collected from a sample of size n, arranged from lowest to highest (including duplicates). The **sample median** $\tilde{x}$ is a summary measure that describes the value that lies in the middle of the ordered sequence of observations.

$$\tilde{x} = \begin{cases} \text{The } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ value if } n \text{ is odd} \\ \text{The average of the } \left(\frac{n}{2}\right)^{\text{th}} \text{ and the } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ value if } n \text{ is even} \end{cases}$$

Exercise 1.3.2.1: R built-in dataset `penguins` contains bodily measurements of different species of penguins. The length of the flippers (in centimetres) of the first 10 Gentoo penguins are presented below:

21.1	23.0	21.0	21.8	21.5	21.0	21.1	21.9	20.9	21.5

Compute the sample median.

Solution: By sorting the sample, the sample median is given by $(21.1 + 21.5)/2 = 21.3$. □

Sample mean and sample median might appear similar, but are actually quite different, and often do not coincide. For example, outliers do not affect the sample median, because what matters for the sample median are the values that lie at the center of the sample. Moreover, it is possible to define a sample median for ordinal attributes.

1.3.3. Sample percentiles

Given a numerical variable X (discrete or continuous), let $x_1, x_2, ..., x_n$ be the observations of X collected from a sample of size n, arranged from lowest to highest (including duplicates). Having fixed a real value $p \in [0, 1]$, the **100-p sample percentile** is the value q such that at least the $100p\%$ of the sample has a value greater or equal than q and the $100(1-p)\%$ of the sample has a value less than q.

$$q = \begin{cases} x_{\lfloor np \rfloor + 1} & \text{if } np \notin \mathbb{Z} \\ \frac{1}{2}\left(x_{np} + x_{np+1}\right) & \text{if } np \in \mathbb{Z} \end{cases}$$

Sample percentiles are just a generalization of the sample mean. Indeed, the 50-th percentile ($p = 0.5$) is precisely the sample mean, that splits the data into two equally-sized subsets. Percentiles split the data in the same way, but unevenly. Common choices of p are:

1. $p = 0.25$, the 25-th percentile, also called **first quartile**: $1/4$ of the sample are on the left and $3/4$ of the sample are on the right. Denoted as Q_1;
2. $p = 0.50$, the 50-th percentile or sample mean, also called **second quartile**. Denoted as Q_2;
3. $p = 0.75$, the 75-th percentile, also called **third quartile**: $3/4$ of the sample are on the left and $1/4$ of the sample are on the right. Denoted as Q_3.

Exercise 1.3.3.1: Compute the three quartiles of Exercise 1.3.2.1

Solution:

$$Q_1 = x_3 = 21.025 \qquad Q_2 = (x_5 + x_6)/2 = 21.300 \qquad Q_3 = x_8 = 21.725$$

□

An interesting graphical representation of a sample and its three main percentiles is the **box plot**. The plot is composed by a straight line that goes from the lowest to the highest value in the sample, above of which lies a rectangle whose sides are drawn where the first and third quartiles are and a vertical line inside of said rectangle where the second quartile are.

Exercise 1.3.3.2: Draw a boxplot of the dataset in Exercise 1.3.2.1.

Solution:

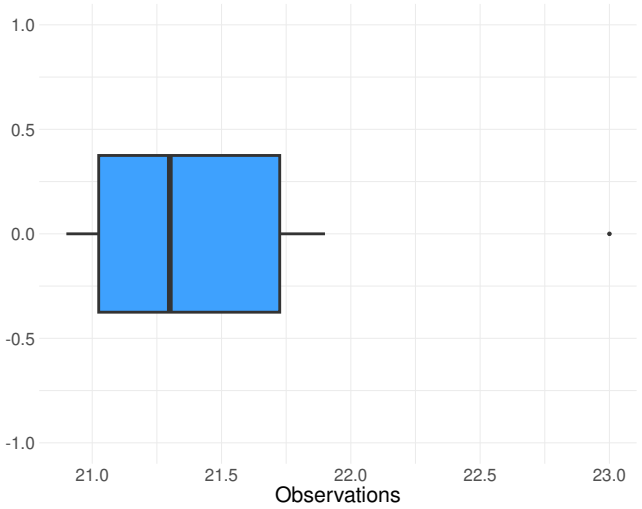


Figure 5: A boxplot of the dataset in Exercise 1.3.2.1, showing the three quartiles.

□

1.3.4. Sample mode

Given a sample for any non-continuous variable, the **sample mode** is the observation having the highest frequency. A sample can have a single mode (**monomodal**), two modes (**bimodal**) or many modes (**multimodal**). The sample mode of a continuous variable is computed with respect to the frequency of the classes. The sample mode is easily read from a bar plot as the highest rectangle.

1.4. Variability indices

Summary measures that provide information on how the sample is "dispersed", that is, how its values are detached from their centre, are called **variability indices**.

1.4.1. Range

Given a numerical variable X, let $x_1, x_2, ..., x_n$ be the observations of X collected from a sample of size n. The **range** is the difference between the observation with the highest value and the observation with the lowest value. This summary measure is very simple, but depends exclusively on the extremes of the sample.

A slightly better alternative is the **interquartile range** (**IQR** for short), given by the difference between the third quantile and the first quantile. The IQR represents the interval where most of the observations lie, without being skewed by the extremes.

$$\text{IQR} = Q_3 - Q_1$$

1.4.2. Sample variance

Given a numerical variable X, let $x_1, x_2, ..., x_n$ be the observations of X collected from a sample of size n. The difference between an observation x_i and the sample mean $\overline{x}$ is called **residue**, or **deviation from the mean**.

A residue will be positive if the observation is larger than the mean and negative if smaller than the mean. If all the residues are small in magnitude, then all observations are close to the mean and there is little variability. On the other hand, if some of the deviations are large in magnitude, then some observations lie far from the mean, suggesting a greater amount of variability.

To obtain a summary measure from the residues, they have to be combined in some way into a single value. The simplest way would be to sum all residues, but Lemma 1.4.2.1 shows that this approach leads to nowhere.

Lemma 1.4.2.1: The sum of all residues is always zero.

Proof:

$$\sum_{i=1}^{n} \overline{x} - x_i = \sum_{i=1}^{n} \overline{x} - \sum_{i=1}^{n} x_i = n\overline{x} - \sum_{i=1}^{n} x_i = n\left(\overline{x} - \frac{\sum_{i=1}^{n} x_i}{n}\right) = n(\overline{x} - \overline{x}) = n \cdot 0 = 0$$

Separating $\sum_{i=1}^{n} \overline{x} - x_i$ into two sums is justified because both sums are finite. □

There are a number of way to circumvent the issue arisen in Lemma 1.4.2.1. For example, instead of computing the sum of all residues, one could compute the sum of the absolute values of the residues. Since the absolute value operator can introduce some technical difficulties, a simpler approach is to work with the square of the residues. This leads to the **sample variance**:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$$

Exercise 1.4.2.1: R built-in dataset `ToothGrow` contains measurements concerning Guinea pigs in a lab experiment, where they were fed either orange juice or ascorbic acid ("pure" vitamin C) to investigate a link between teeth grow and vitamin C intake. The length of the teeth of the first 10 Guinea pigs are presented below:

4.2	11.5	7.3	5.8	6.4	10.0	11.2	11.2	5.2	7.0

Knowing that the sample mean is 7.98, compute the sample variance.

Solution:

$$s^2 = \frac{(4.2 - 7.98)^2 + (11.5 - 7.98)^2 + ... + (5.2 - 7.98)^2 + (7.0 - 7.98)^2}{10 - 1} = 7.54$$

□

The reason why the denominator has $n - 1$ instead of n will become clear when introducing parameter estimation.

Lemma 1.4.2.2: If s_x^2 is the sample variance of a given sample, a linear transformation $f(x) = ax + b$ with $a, b \in \mathbb{R}$ applied to all elements of the sample gives a new sample whose sample variance is $s_y = a^2 s_x^2$.

Proof: Let $y_i = ax_i + b$ be the i-th element of the transformed sample. Using Lemma 1.3.1.1:

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n-1} = \frac{\sum_{i=1}^n (ax_i + \cancel{b} - a\overline{x} - \cancel{b})^2}{n-1} = \frac{\sum_{i=1}^n (a(x_i - \overline{x}))^2}{n-1} =$$
$$= \frac{\sum_{i=1}^n a^2(x_i - \overline{x})^2}{n-1} = a^2 \left(\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1} \right) = a^2 s_x^2$$

□

1.4.3. Sample standard deviation

The **sample standard deviation** is defined as the square root of the sample variance:

$$s = \sqrt{s^2}$$

It is sometimes preferred to the sample variace because it has the same unit of measurement as the data. Being the sum of squared quantities, the unit of measurement of the variance is also squared.

1.4.4. Coefficient of variation

The **coefficient of variation** (**CV** for short) is given by the ratio between the sample standard deviation and the sample mean:

$$\mathrm{CV} = \frac{s}{\overline{x}}$$

It describes how the observations are "spread out" while taking into account how they differ from the sample mean.

1.5. Frequencies for two variables

Most statistical enqiries are interested in analyzing more than one variable at the same time.

Consider two discrete variables X and Y. Suppose that a bivariate dataset is constructed from a sample of size n of observations $O = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ of X and Y. The support of both variables considered at the same time is the Cartesian product of the two supports, that is $D(X, Y) = D(X) \times D(Y)$.

The **double absolute frequency** $f_{i,j}$ of $(d_i, d_j) \in D(X, Y)$ is defined as the number of observations in the sample that have d_i as the value attained by X and d_j as the value attained by Y.

$$f_{i,j} = \left| \{(x, y) \mid (x, y) \in O, x = d_i, y = d_j\} \right|$$

From the double absolute frequency it is possible to define what is called the **double cumulative absolute frequency** $F_{i,j}$ of $(d_i, d_j) \in D(X, Y)$ as the sum of the double absolute frequencies of all $(d_a, d_b) \in D(X, Y)$ such that $d_a \leq d_i$ and $d_b \leq d_j$.

$$F_{i,j} = \sum_{a:d_a \leq d_i, b:d_b \leq d_j}^{|D(X,Y)|} f_{a,b}$$

The **double relative frequency** $p_{i,j}$ of $(d_i, d_j) \in D(X, Y)$ is given by the ratio between the double absolute frequency $f_{i,j}$ and the sample size n.

$$p_{i,j} = \frac{f_{i,j}}{n}$$

From the double absolute frequency it is possible to define what is called the **double cumulative relative frequency** $P_{i,j}$ of $(d_i, d_j) \in D(X,Y)$ as the sum of the double relative frequencies of all $(d_a, d_b) \in D(X,Y)$ such that $d_a \leq d_i$ and $d_b \leq d_j$.

$$P_{i,j} = \sum_{a:d_a \leq d_i, b:d_b \leq d_j}^{|D(X,Y)|} p_{a,b}$$

Exercise 1.5.1: Suppose that a sample has been collected regarding the number of rooms in a flat and the number of people inhabiting such flat. Compute the four frequencies for the sample.

4,3 2,2 5,4 4,4 3,2 4,4 5,2 4,4 3,4 3,2 2,2 3,3 3,2 5,4 4,3 5,3 4,3 4,4

4,3 2,3 5,4 3,3 4,4 4,2 4,3 5,3 5,3 2,2 5,4 3,2 5,2 5,4 4,4 3,4 3,2

Solution:

Number of rooms	Number of occupants	Double absolute frequency	Double relative frequency	Double cumulative absolute frequency	Double cumulative relative frequency
2	2	3	0.086	3	0.086
3	2	5	0.143	8	0.229
4	2	1	0.029	9	0.257
5	2	2	0.057	11	0.314
2	3	1	0.029	12	0.343
3	3	2	0.057	14	0.4
4	3	5	0.143	19	0.543
5	3	3	0.086	22	0.629
2	4	0	0	22	0.629
3	4	2	0.057	24	0.686
4	4	6	0.171	30	0.857
5	4	5	0.143	35	1

□

Frequencies for three or more variables can be defined following the same approach. Also, if one or more of the variables are continuous, it is still possible to assign them a frequency by partitioning the observations into classes.

A known graphical representation for the double absolute frequency or double relative frequency is the **heat map**, or **color map**. A heat map is a rectangle partitioned into cells, where to each cell is assigned one member of the support. Each cell is colored with a different shade of the same color: the brightness of the color is proportional to the double absolute/relative frequency of the support element assigned to that cell.

Exercise 1.5.2: Draw a color map for the absolute frequency in Exercise 1.5.1.

Solution:

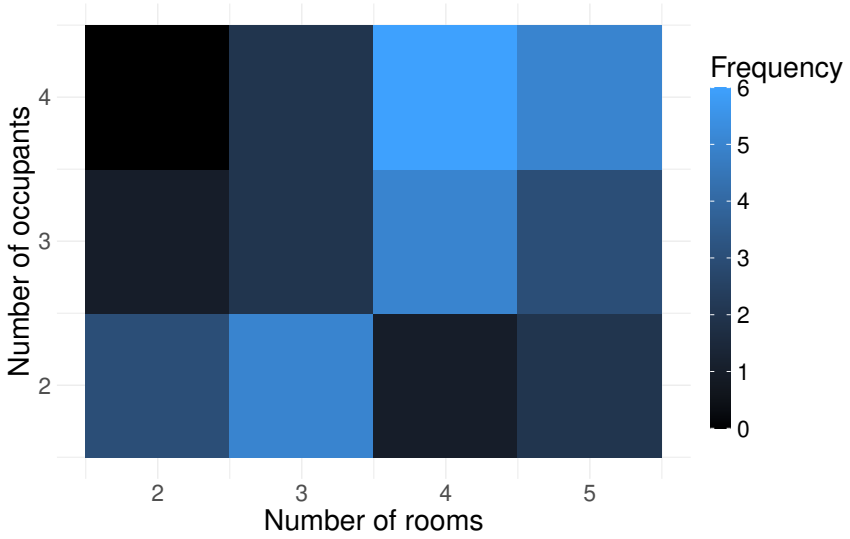


Figure 6: A color map for Exercise 1.5.1. The brightness of the color blue is proportional to the frequency of a particular combination of rooms and occupants.

□

When dealing with bivariate or multivariate data, it can still be interesting to know the frequency of the values of a single variable, without considering the value of the other(s). Frequencies that take into account only one variable in each observation neglecting the others are called **marginal frequencies**.

For example, consider two variables X and Y and a sample of size n of observations $O = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$. For a $d_i \in D(X)$, the marginal absolute frequency with respect to X is defined as the number of observations in the sample that have d_i as the value attained by X and any value as the value attained by Y. On the other hand, for a $d_j \in D(Y)$ the marginal absolute frequency with respect to Y is defined as the number of observations in the sample that have any value as the value attained by X and d_j as the value attained by Y.

$$f_{i,} = |\{(x, y) \mid (x, y) \in O, x = d_i\}| \qquad f_{,j} = |\{(x, y) \mid (x, y) \in O, y = d_j\}|$$

2. Probability theory

2.1. Sample spaces and events

Probability theory is a mathematical framework providing methods that describe situations and events having an unforseeable outcome, quantifying chance and randomness related to said results.

Any activity or process having at least one (unnkowable in advance) outcome is called an **experiment**. The set containing all possible outcomes of an experiment, denoted $\mathcal{S}$ or Ω, is called **sample space**. The sample space can be either discrete or continuous.

Exercise 2.1.1: Provide some examples of experiments.

Solution:

- Rolling a six-sided die is an experiment; the value read on the top is unkowable until the die is rolled. The sample space $\mathcal{S}$ contains 6 elements:

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$

- Drawing a card from a (standard) deck is an experiment; the value on the card is unknowable until the card is turned face up. The sample space $\mathcal{S}$ contains 52 elements:

$$\mathcal{S} = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit, A\diamondsuit, 2\diamondsuit, ..., K\diamondsuit, A\clubsuit, 2\clubsuit, ..., K\clubsuit, A\spadesuit, 2\spadesuit, ..., K\spadesuit\}$$

- Tossing three coins (or any number of coins) is an experiment; the side of the coin facing up (heads or tails) is unkowable until the coin is flipped. The sample space $\mathcal{S}$ contains 8 elements:

$$\mathcal{S} = \{TTT, TTH, THT, HTT, THH, HHT, HTH, HHH\}$$

□

Any subset of the sample space is called an **event**. An event can be either **simple** if it's a singleton (it contains a single outcome of the experiment) or **compound** (it contains multiple outcomes). An event can either occur or not occur, depending on the outcome of the experiment.

When an experiment is performed, a particular event A is said to occur if the outcome of the experiment is either A itself or is a subset of A. In general, exactly one simple event will occur, but many compound events will occur simultaneously. This is why "simple event" and "outcome" are used interchangeably.

Exercise 2.1.2: Provide some examples of events, referring to Exercise 2.1.1.

Solution:

- Consider the roll of a six-sided die. The subset $A = \{1, 3, 5\}$ of the sample space $\mathcal{S}$ is the event "an even number". It is a compound event;
- Consider the drawing of a card from a deck. The subset $B = \{A\heartsuit, A\diamondsuit, A\clubsuit, A\spadesuit, K\clubsuit\}$ of the sample space $\mathcal{S}$ is the event "either an ace of any suit or the king of clubs". It is a compound event;

- Consider the side facing up of three coins after being tossed. The subset $C = \{HHH\}$ of the sample space $\mathcal{S}$ is the event "all three heads". It is a simple event.

□

Events are nothing more than (sub)sets. They can therefore be given a graphical representation through **Euler-Venn diagrams**.

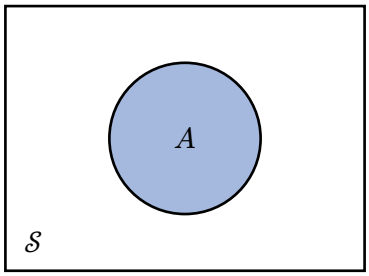


Figure 7: Drawing an event as an Euler-Venn diagram. The event is the circle labeled A, while the sample space is the rectangle labeled $\mathcal{S}$ enclosing the circle.

Moreover, events can be manipulated using the algebra of set theory to construct new events. Given a sample space $\mathcal{S}$ and two events A and B, subsets of $\mathcal{S}$:

- The **intersection** of A and B, denoted as $A \cap B$, corresponds to the event containing all outcomes contained both in A and in B. That is, $A \cap B$ occurs if and only if both A and B occur at the same time;
- The **union** of A and B, denoted as $A \cup B$, corresponds to the event containing all outcomes contained either in A, in B or in both. That is, $A \cup B$ occurs if at most A or B occurs.
- The **complement** of A, denoted as A^c, corresponds to the event containing all outcomes not contained in A. That is, A^c occurs if and only if A does not occur;

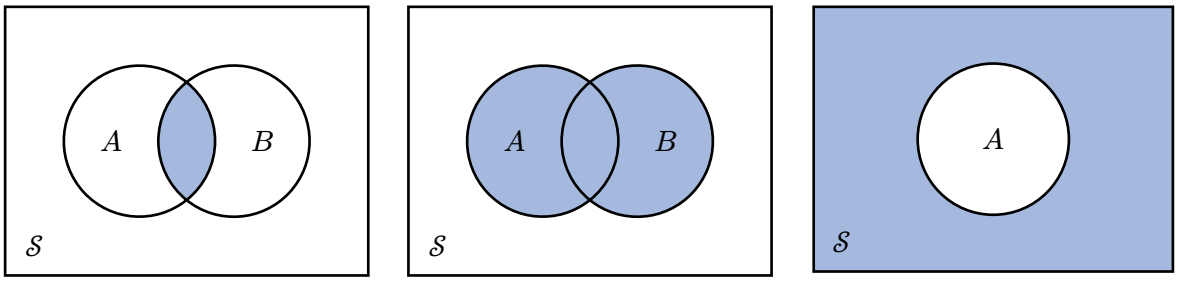


Figure 8: On the left, the intersection of two events A and B, subsets of the sample space $\mathcal{S}$. In the middle, their union. On the right, the complement of A.

The union and intersection of three or more events are computed from extending the definition for two events in the standard way.

Exercise 2.1.3: Provide some examples of complemented, intersected and unified events, referring to Exercise 2.1.1.

Solution:

- Consider the roll of a six-sided die. The subset $A = \{6\}$ of the sample size $\mathcal{S}$ is the event "six", while the subset $A^c = \{1, 2, 3, 4, 5\}$ is the event "any number but six";
- Consider the drawing of a card from a deck. Let $B_1 = \{A\heartsuit, A\spadesuit, K\clubsuit\}$ be the event "an ace of hearts, an ace of spades or the king of clubs", and let $B_2 = \{A\diamondsuit, A\clubsuit, K\clubsuit\}$ be the

event "an ace of diamonds, an ace of clubs or the king of clubs". The set $B = B_1 \cup B_2 = \{A\heartsuit, A\diamondsuit, A\clubsuit, A\spadesuit, K\clubsuit\}$ is the event "either an ace of any suit or the king of clubs";
- Consider the toss of three coins. Let $C_1 = \{HTT, HHT, HTH, HHH\}$ be the event "first coin heads", and let $C_2 = \{THT, HHT, THH, HHH\}$ be the event "second coin heads". The set $C = C_1 \cap C_2 = \{HHT, HHH\}$ is the event "first and second coin heads".

□

The empty set $\emptyset$ denotes the event of having no outcome whatsoever, also called the **null event**. If the intersection of two events is the null event, such events are said to be **mutually exclusive** events, or **disjoint** events. In other words, two events are said to be mutually exclusive if they can't happen at the same time. By definition, an event and its complement are mutually exclusive.

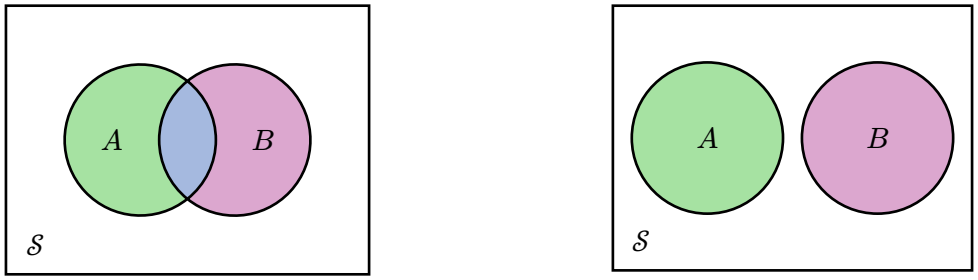


Figure 9: The events on the left are not mutually exclusive, because their intersection is not the empty set. The events on the right are mutually exclusive.

Exercise 2.1.4: Provide an example of disjoint events, referring to Exercise 2.1.1.

Solution: Consider the drawing of a card from a deck, with sample space $\mathcal{S}$. Consider the following events:

$$B_1 = \{7\heartsuit, 7\diamondsuit, 7\clubsuit, 7\spadesuit\} = \text{A 7 of any suit}$$
$$B_2 = \{Q\heartsuit, Q\diamondsuit, Q\clubsuit, Q\spadesuit\} = \text{A queen of any suit}$$
$$B_3 = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\} = \text{Any card of the hearts suit}$$

Intersecting them gives:

$$B_1 \cap B_2 = \{7\heartsuit, 7\diamondsuit, 7\clubsuit, 7\spadesuit\} \cap \{Q\heartsuit, Q\diamondsuit, Q\clubsuit, Q\spadesuit\} = \emptyset$$
$$B_2 \cap B_3 = \{Q\heartsuit, Q\diamondsuit, Q\clubsuit, Q\spadesuit\} \cap \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\} = \{Q\heartsuit\}$$
$$B_1 \cap B_3 = \{7\heartsuit, 7\diamondsuit, 7\clubsuit, 7\spadesuit\} \cap \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\} = \{7\heartsuit\}$$

Which means that B_1 and B_2 are mutually exclusive, while both B_1 and B_3 and B_2 and B_3 are not. □

2.2. Axiomatic definition of probability

To an event A, it is possible to associate a value, its **probability**, denoted as $P(A)$. The value of $P(A)$ represents a measure of likelihood, certainty or confidence of the event A to occur. Intuitively, an higher value of probability corresponds to an higher likelihood.

Modern probability theory, like set theory, is defined axiomatically. Such axioms are also called **Kolmogorov axioms**, and are the smallest amount of axioms from which to derive a theory of probability free of contradictions.

Axiom 2.2.1 (First Kolmogorov axiom): For any sample space $\mathcal{S}$ and for any event $A \subseteq (\mathcal{S})$, $P(A) \geq 0$.

Axiom 2.2.2 (Second Kolmogorov axiom): For any sample space $\mathcal{S}$, $P(\mathcal{S}) = 1$.

Axiom 2.2.3 (Third Kolmogorov axiom): Let $\mathcal{S}$ be any sample space. If $A_1, A_2, ... \subseteq \mathcal{S}$ is a countably infinite collection disjoint events:

$$P(A_1 \cup A_2 \cup ...) = \sum_{i=1}^{\infty} P(A_i)$$

Axiom 2.2.1 states that probabilities are always non-negative. Axiom 2.2.2 states that the probability of the sample space, which is the probability of any event to happen, is 1. Axiom 2.2.3 states that the probability of the union of disjoint events is the sum of the individual probabilities.

The Kolmogorov axioms are sufficient to derive the entire probability theory, and set some constrains on which probability assignments are valid and which are not. However, they do not define how this assignment is supposed to be made.

At least for discrete sample spaces, the simplest way to do so is applying the so-called **Principle of Indifference**. The Principle is as follows: when there isn't sufficient information on how to distribute probabilities to the possible simple outcomes of an experiment, the "fairest" way to assign probabilities is to equally distribute them among the outcomes.

More formally: let $\mathcal{S}$ be a sample space constituted by n simple events $\{A_1, ..., A_n\}$. If there isn't enough information on what the values of $P(A_1), ..., P(A_n)$ should be, then it has to be assumed $P(A_1) = ... = P(A_n) = 1/n$.

The Principle is highly contentious, and if applied incorrectly can lead to irrational results. Moreover, it can't be applied to continuous sample spaces. However, it is often invoked (even without mention) when no other option is possible, and an assignment done invoking the Principle is always compliant with the Kolmogorov axioms.

Exercise 2.2.1: Assign probabilities to the events of Exercise 2.1.1, applying the Principle of Indifference.

Solution:

- Rolling a six-sided die is an experiment with $|\mathcal{S}| = 6$. Applying the Principle of Indifference:

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$$

- Drawing a card from a (standard) deck is an experiment with $|\mathcal{S}| = 52$. Applying the Principle of Indifference:

$$P(\{A\heartsuit\}) = P(\{2\heartsuit\}) = ... = P(\{K\heartsuit\}) = P(\{A\diamondsuit\}) = P(\{2\diamondsuit\}) = ... = P(\{K\diamondsuit\}) =$$
$$P(\{A\clubsuit\}) = P(\{2\clubsuit\}) = ... = P(\{K\clubsuit\}) = P(\{A\spadesuit\}) = P(\{2\spadesuit\}) = ... = P(\{K\spadesuit\}) = \frac{1}{52}$$

- Tossing three coins is an experiment with $|\mathcal{S}| = 6$. Applying the Principle of Indifference:

$$P(\{TTT\}) = P(\{TTH\}) = P(\{THT\}) = P(\{HTT\}) =$$
$$P(\{THH\}) = P(\{HHT\}) = P(\{HTH\}) = P(\{HHH\}) = \frac{1}{8}$$

□

From the Kolmogorov axioms, many more useful properties can be derived.

Lemma 2.2.1: For any sample space $\mathcal{S}$, $P(\emptyset) = 0$. That is, the null event cannot occur.

Proof: Consider the countably infinite collection of events $\emptyset, \emptyset, ...$ Set algebra states that $\emptyset \cup \emptyset \cup ... = \emptyset$, therefore $P(\emptyset \cup \emptyset \cup ...) = P(\emptyset)$. Set algebra also states that $\emptyset \cap \emptyset = \emptyset$, which means that $\emptyset, \emptyset, ...$ is a countably infinite collection disjoint events. Applying Axiom 2.2.3:

$$P(\emptyset) = P(\emptyset \cup \emptyset \cup ...) = \sum_{i=1}^{\infty} P(\emptyset)$$

Axiom 2.2.1 states that $P(\emptyset)$ has to be non-negative, and the only way for $P(\emptyset)$ to be greater than or equal to 0 while at the same time being equal to $\sum_{i=1}^{\infty} P(\emptyset)$ is for $P(\emptyset)$ to be 0. □

Lemma 2.2.2: Let $\mathcal{S}$ be a sample space. If $A_1, A_2, ..., A_n \subseteq \mathcal{S}$ is a finite collection of disjoint events, the following equality holds:

$$P(A_1 \cup A_2 \cup ... \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

Which means that Axiom 2.2.3 can be extended to finite collections as well.

Proof: Consider the countably infinite collection of events:

$$A_1, A_2, ..., A_n, A_{n+1} = \emptyset, A_{n+2} = \emptyset, ...$$

Which is the original collection but infinitely extended with empty sets. By definition, $A \cap \emptyset = \emptyset$ for any set A, which means that this collection is a countably infinite collection of disjoint events. Applying Axiom 2.2.3:

$$P(A_1 \cup A_2 \cup ... \cup A_n \cup \emptyset \cup \emptyset \cup ...) = \sum_{i=1}^{\infty} P(A_i)$$

It is possible to split the summation in two like so:

$$P(A_1 \cup A_2 \cup ... \cup A_n) + P(\emptyset \cup \emptyset \cup ...) = \sum_{i=1}^{n} P(A_i) + \sum_{i=n+1}^{\infty} P(\emptyset)$$

But Lemma 2.2.1 states that $P(\emptyset) = 0$, therefore:

$$P(A_1 \cup A_2 \cup ... \cup A_n) + P(\emptyset \cup \emptyset \cup ...) = P(A_1 \cup A_2 \cup ... \cup A_n) + 0 = \sum_{i=1}^{n} P(A_i)$$

□

Exercise 2.2.2: Consider the three events:

$$A = \{7\heartsuit, 7\diamondsuit, 7\clubsuit, 7\spadesuit\} \qquad B = \{K\heartsuit, K\diamondsuit, K\clubsuit, K\spadesuit\} \qquad C = \{A\diamondsuit, 2\diamondsuit, ..., K\diamondsuit\}$$

What are $P(A)$, $P(B)$ and $P(C)$?

Proof: The three compound events are constituted by single events that are all disjoint with each other. By Principle of Indifference, each of these single events has probability $1/52$. Lemma 2.2.2 can then be applied:

$$\begin{aligned}
P(A) &= P(\{7\heartsuit, 7\diamondsuit, 7\clubsuit, 7\spadesuit\}) = P(\{7\heartsuit\}) + P(\{7\diamondsuit\}) + P(\{7\clubsuit\}) + P(\{7\spadesuit\}) = \\
&= \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{4}{52} = \frac{1}{13} \\
P(B) &= P(\{K\heartsuit, K\diamondsuit, K\clubsuit, K\spadesuit\}) = P(\{K\heartsuit\}) + P(\{K\diamondsuit\}) + P(\{K\clubsuit\}) + P(\{K\spadesuit\}) = \\
&= \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{4}{52} = \frac{1}{13} \\
P(C) &= P(\{A\diamondsuit, 2\diamondsuit, ..., K\diamondsuit\}) = P(\{A\diamondsuit\}) + P(\{2\diamondsuit\}) + ... + P(\{K\diamondsuit\}) = \\
&= \frac{1}{52} + \frac{1}{52} + ... + \frac{1}{52} = \frac{13}{52} = \frac{1}{4}
\end{aligned}$$

□

Theorem 2.2.1 (Complement rule): For any sample space $\mathcal{S}$ and any event $A \subseteq \mathcal{S}$:

$$P(A) + P(A^c) = 1$$

Proof: By set algebra, $A \cup A^c = \mathcal{S}$. Since Axiom 2.2.2 states that $P(\mathcal{S}) = 1$, then $P(A \cup A^c) = 1$. Also, set algebra dictates that $A \cap A^c = \emptyset$, which means that A and A^c are disjoint events. Applying Lemma 2.2.2:

$$\sum_{i=1}^{2} P(A_i) = P(A) + P(A^c) = P(A \cup A^c) = 1$$

□

Lemma 2.2.3: For any sample space $\mathcal{S}$ and any event $A \subseteq \mathcal{S}$, $0 \leq P(A) \leq 1$.

Proof: By Theorem 2.2.1, $P(A) + P(A^c) = 1$. Since Axiom 2.2.2 states that both $P(A)$ and $P(A^c)$ can't be negative, the only way for $P(A) + P(A^c) = 1$ to be true is if both $P(A)$ and $P(A^c)$ are less than or equal to 1. Combining both boundaries, $0 \leq P(A) \leq 1$. □

Theorem 2.2.2 (Addition law): For any sample space $\mathcal{S}$ and any two events $A, B \subseteq \mathcal{S}$:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof: From set algebra, it is possible to rewrite both $A \cup B$ and B as the union of two disjoint events:

$$A \cup B = A \cup (A^c \cap B) \qquad B = (A^c \cap B) \cup (A \cap B)$$

Applying Lemma 2.2.2 to both expression:

$$P(A \cup B) = P(A \cup (A^c \cap B)) = P(A) + P(A^c \cap B)$$
$$P(B) = P((A^c \cap B) \cup (A \cap B)) = P(A^c \cap B) + P(A \cap B)$$

Moving $P(A \cap B)$ to the left hand side of the second expression gives $P(B) - P(A \cap B) = P(A^c \cap B)$. Substituting into the first equation gives $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. □

Exercise 2.2.3: Consider Exercise 2.2.2. What is $P(A \cup B \cup C)$?

Proof: $P(A \cup B \cup C)$ is equivalent to $P((A \cup B) \cup C)$. $A \cup B$ is the union of two disjoint events, since $A \cap B = \emptyset$, which means that Axiom 2.2.3 can be applied. This gives:

$$P(A \cup B) = P(A) + P(B) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

On the other hand, $A \cup B$ and C are not disjoint:

$$(A \cup B) \cap C = \{7\heartsuit, 7\diamondsuit, 7\clubsuit, 7\spadesuit, K\heartsuit, K\diamondsuit, K\clubsuit, K\spadesuit\} \cap \{A\diamondsuit, 2\diamondsuit, ..., K\diamondsuit\} = \{7\diamondsuit, K\diamondsuit\}$$

By Principle of Indifference:

$$P((A \cup B) \cap C) = P(\{7\diamondsuit, K\diamondsuit\}) = P(\{7\diamondsuit\}) + P(\{K\diamondsuit\}) = \frac{1}{52} + \frac{1}{52} = \frac{1}{26}$$

Applying Theorem 2.2.2:

$$P(A \cup B \cup C) = P(A \cup B) + P(C) - P((A \cup B) \cap C) = \frac{2}{13} + \frac{1}{4} - \frac{1}{26} = \frac{19}{52}$$

□

Theorem 2.2.3 (Boole's inequality): For any sample space $\mathcal{S}$ any countable (possibly infinite) collection of events $A_1, A_2, ..., A_n \subseteq \mathcal{S}$:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$$

Proof: The statement can be proven invoking the Principle of Induction. For the base case $n = 1$:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i) \Rightarrow P\left(\bigcup_{i=1}^{1} A_i\right) \leq \sum_{i=1}^{1} P(A_i) \Rightarrow P(A_1) \leq P(A_1)$$

Which is trivially true. As for the inductive step, Theorem 2.2.2 states that:

$$P\left(\bigcup_{i=1}^{n+1} A_i\right) = P\left(\left(\bigcup_{i=1}^{n} A_i\right) \cup A_{n+1}\right) = P\left(\bigcup_{i=1}^{n} A_i\right) + P(A_{n+1}) - P\left(\left(\bigcup_{i=1}^{n} A_i\right) \cap A_{n+1}\right)$$

Substituting:

$$P\left(\bigcup_{i=1}^{n+1} A_i\right) \leq \sum_{i=1}^{n+1} P(A_i) \Rightarrow$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) + \cancel{P(A_{n+1})} - P\left(\left(\bigcup_{i=1}^{n} A_i\right) \cap A_{n+1}\right) \leq \cancel{P(A_{n+1})} + \sum_{i=1}^{n} P(A_i) \Rightarrow$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) + P\left(\left(\bigcup_{i=1}^{n} A_i\right) \cap A_{n+1}\right)$$

$P\left(\bigcup_{i=1}^{n} A_i\right)$ is smaller than or equal to $\sum_{i=1}^{n} P(A_i)$ by the inductive hypothesis. This means that the inequality is true if and only if the second term on the right hand side is non-negative. Since this is guaranteed by Axiom 2.2.1, the result is proven. □

Probability theory is powerful, but doesn't explain what a probability value is exactly supposed to be. That is, what is the epistemological meaning of probability? How is probability "mapped" to reality? As a matter of fact, probabilities can be assigned to events in any possible way that is constrained by the axioms even if the result is not applicable to real life, even if such an assignment defies known physical laws.

Exercise 2.2.4: Consider the toss of a six-sided die. Are these probability assignments valid?

	$P(\{1\})$	$P(\{2\})$	$P(\{3\})$	$P(\{4\})$	$P(\{5\})$	$P(\{6\})$
Assignment 1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Assignment 2	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$
Assignment 3	0	0	0	0	0	1
Assignment 4	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$-\frac{1}{6}$
Assignment 5	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{8}$

Solution:

1. Yes, since it complies with Axiom 2.2.1, Axiom 2.2.2 and Axiom 2.2.3. This is the assignment derived from invoking the Principle of Indifference. Since there is no other information on how the die is constructed, it is reasonable to assume that the die can land to any side;

2. Yes, since it complies with Axiom 2.2.1, Axiom 2.2.2 and Axiom 2.2.3. If the die is loaded, meaning that it has been crafted in such a way that the center of mass is not in its geometrical center, this assignment is actually possible;
3. Yes, since it complies with Axiom 2.2.1, Axiom 2.2.2 and Axiom 2.2.3. However, there is no real world die that can be described by this assignment: it's impossible to construct a die, no matter how precisely, that will always land on the same side;
4. No, because it defies Axiom 2.2.1, since $P(\{6\}) \leq 0$;
5. No, because it defies Axiom 2.2.2:

$$P(\mathcal{S}) = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\}) = \\ = \frac{1}{3} + \frac{1}{6} + \frac{1}{8} + \frac{1}{3} + \frac{1}{6} + \frac{1}{8} = \frac{5}{4}$$

□

There are multiple schools of thought in regards to interpreting probability in an epistemological sense. Most interpretations fall into two broad categories: **objective** interpretations, such as the **frequentist** interpretation, and **subjective** interpretations, such as the **Bayesian** interpretation.

The frequentist interpretation relies on empirical data for assigning probabilities. Consider an experiment that can be repeatedly performed in the exact same way (tossing a coin, rolling a die, drawing a card, ecc...). Let A be any event; if the experiment is performed n times, the event A will occur $n(A)$ times (with $0 \leq n(A) \leq n$). The ratio $n(A)/n$ is called the **relative frequency** of occurrence of A in the sequence of n attempts.

Empirical data suggests that the relative frequency fluctuates considerably if n is a small number, while tends to stabilize itself as n grows. Ideally, repeating such experiment infinitely many times, it would be possible to obtain a "perfect" frequency, called **limiting relative frequency**. The frequentist interpretation of probability states that this limiting relative frequency is precisely the probability that should be assigned to A.

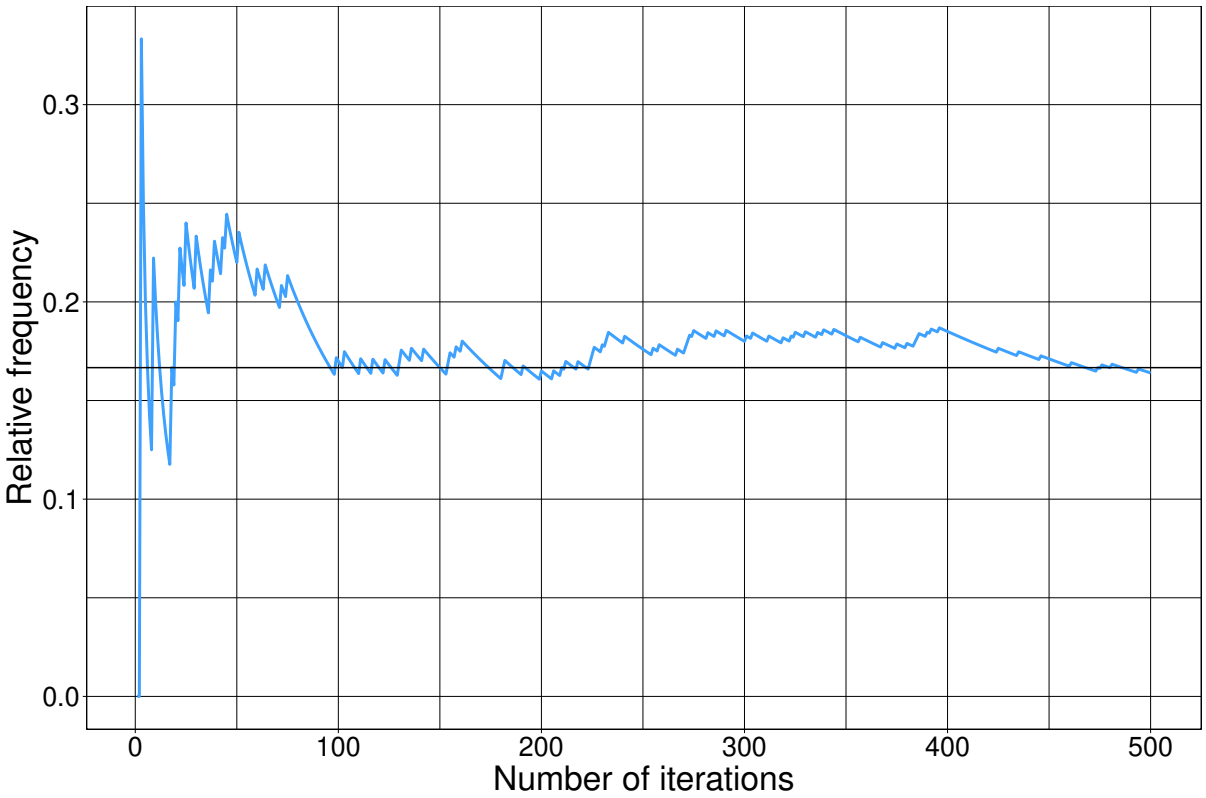


Figure 10: Result of simulating a dice roll using the `sample` function in R. The horizontal black line has as intercept $1/6$, the probability of rolling any face, understood to be the limiting relative frequency. The blue line interpolates the relative frequencies computed at each iteration. Even though in the earlier iterations the two frequencies don't match, they get closer as the iterations increase.

This interpretation of probability is said to be objective in the sense that it rests on a fact of nature and not on the concerns of the agent performing it. Ideally, two agents performing the same experiment the same number of times would obtain the same relative limiting frequency. Note that not all real-world phenomena can be tested multiple times, which means that its applicability is limited[1].

The Bayesian interpretation relies instead on a "degree of confidence" with which an agent believes an event to occur, meaning that different agents will have different descriptions of the same phenomena. The probability assigned to an event is precisely this degree of confidence, that can change if the agent acquires more knowledge on the generation of the event.

2.3. Combinatorics

As long as the sample space is small and has a simple structure (a deck of cards, the faces on a die, ecc...), counting its elements one by one is sufficient to know its cardinality. However, for large sample spaces and for sample spaces with a non obvious structure, counting is unfeasible. The mathematical discipline concerned with counting, as in the enumeration of the possible arrangements or configurations of specified structures, is **combinatorics**.

The most important building block of combinatorics is the **Fundamental Principle of Counting**. Consider an experiment having n components, with each i-th component having x_i possible outcomes. The number of outcomes of the experiment as a whole is given by:

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot (...) \cdot x_n$$

Exercise 2.3.1: Consider an experiment consisting in the toss of a coin followed by the roll of a die. How many outcomes does this experiment have?

Solution: A coin toss has two possible outcomes, while the roll of a die has six. The number of outcomes of the experiment as a whole can be computed applying the Fundamental Principle of Counting, giving $6 \cdot 2 = 12$. This is confirmed by counting its elements one by one.

$$\Omega_1 = \{H, T\} \qquad \Omega = \Omega_1 \times \Omega_2 = \{\{H,1\},\{T,1\},\{H,2\},\{T,2\},\{H,3\},\{T,3\},$$
$$\Omega_2 = \{1,2,3,4,5,6\} \qquad \{H,4\},\{T,4\},\{H,5\},\{T,5\},\{H,6\},\{T,6\}\}$$

□

Starting from the Fundamental Principle of Counting, it is possible to describe many common counting situations.

A **sequence with repetition** is a situation dealing with ordered sequences of k elements (possibly repeated) chosen among n, such that:

$$n \cdot n \cdot n \cdot (...) \cdot n = n^k$$

[1] It could be objected that, even though a physical experiment might not be possible, a simulated experiment on a computer could. However, this is only true to some extent, since there are phenomena that are impossibly hard to model, or too computationally expensive.

Exercise 2.3.2: What are the possible arrangements of birthdays of three people?

Solution: A year is consituted of 365 days, so the birthdays of three people can be arranged in $365 \cdot 365 \cdot 365 = 365^3 = 48627125$ possible ways. □

A **sequence without repetition** is a situation dealing with ordered sequences of k elements (none repeated) chosen among n with $k \leq n$, such that:

$$n \cdot (n-1) \cdot (n-2) \cdot (...) \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

Exercise 2.3.3: In how many ways is it possible to arrange the birthdays of 23 people such that no two people have birthday the same day?

Solution:

$$\frac{n!}{(n-k)!} = \frac{365!}{(365-23)!} = \frac{365!}{342!} = \frac{365 \cdot 364 \cdot (...) \cdot 344 \cdot 343 \cdot \cancel{342!}}{\cancel{342!}} \approx 4.22 \times 10^{58}$$

□

A **permutation** is a situation dealing with ordered sequences of k elements (none repeated) chosen among n with $k = n$, such that:

$$n \cdot (n-1) \cdot (n-2) \cdot (...) \cdot 2 \cdot 1 = n!$$

Exercise 2.3.4: In how many ways is it possible to arrange a deck of playing cards?

Solution:

$$52! = 52 \cdot 51 \cdot 50 \cdot (...) \cdot 2 \cdot 1 \approx 8.06 \times 10^{67}$$

□

A **combination** is a situation dealing with unordered sequences of k elements (none repeated) chosen among n with $k \leq n$, such that:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdot (n-2) \cdot (...) \cdot (n-k+1)}{k!}$$

The expression $\binom{n}{k}$ is called **binomial** and is read "n choose k".

Exercise 2.3.5: In how many ways is it possible to arrange 20 people in groups of 4?

Solution: This is a combination, since the ways each group is arranged is irrelevant:

$$\binom{20}{4} = \frac{20!}{4!(20-4)!} = \frac{20!}{4! \cdot 16!} = \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot \cancel{16!}}{4! \cdot \cancel{16!}} = \frac{20 \cdot 19 \cdot 18 \cdot 17}{24} = \frac{116280}{24} = 4845$$

□

2.4. Conditional probability

When performing experiments, events can influence one another: knowing that an event has occurred can change the probability of another event occurring. Let $\mathcal{S}$ be a sample space, and let A and B bet two events of $\mathcal{S}$, with $P(B) > 0$. The probability of A to occur given that B occurred (that is, "knowing in advance it occurred") is called the **conditional probability** of A given B, and is given as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Where A is the **conditioned event** and B is the **conditioning event**. The formula, when written in the following form:

$$P(A \cap B) = P(A \mid B)P(B)$$

Is referred to as the **multiplication rule**.

Exercise 2.4.1: Consider rolling a six-sided die. What is the probablility of rolling a 6, knowing that an even number has been rolled?

Solution: Let A be the event "a six" and let B be the event "an even number". This gives:

$$P(A \mid B) = \frac{P(\{6\} \cap \{2,4,6\})}{P(\{2,4,6\})} = \frac{P(\{6\})}{P(\{2\}) + P(\{4\}) + P(\{6\})} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{6} + \frac{1}{6}} = \frac{1}{3}$$

This is reasonable, because knowing that the rolled side is an even number restricts the possible outcomes to $\mathcal{S}' = \{2,4,6\}$, all equally probable by Principle of Indifference. □

The rationale behind the formula for conditional probability is as follows[2]. $P(A \mid B)$ assumes that B has occurred, hence the sample space in which $A \mid B$ lies is not $\mathcal{S}$, but in the subsets of $\mathcal{S}$ in which B occurs. This is done by intersecting A with B, so that only the subevents of A that are "compatible" with B can occur. The proportionality constant $1/P(B)$ ensures that an event conditioned with itself, $P(B \mid B)$, has probability 1, since an event will obviously happen if it is known in advance that it will.

Theorem 2.4.1 (Law of total probability): Let $\mathcal{S}$ be a sample space. Let $A_1, A_2, ..., A_n \subseteq \mathcal{S}$ be a finite collection of mutually exclusive events partitioning $\mathcal{S}$, and let $B \subseteq \mathcal{S}$ be any event. The following holds:

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

[2]Some authors sidestep the heuristic argument and directly include the definition of conditional probability as a fourth axiom of probability.

Proof: First, note how $P(E) = P(E \cap \mathcal{S})$ for any $E \subseteq \mathcal{S}$, since by definition $E \cap \mathcal{S} = E$. Moreover, if $A_1, ..., A_n$ partitions $\mathcal{S}$, then $\bigcup_i A_i = \mathcal{S}$. Then:

$$P(B) = P(B \cap \mathcal{S}) = P\left(B \cap \bigcup_i A_i\right) = P\left(\bigcup_i (B \cap A_i)\right)$$

$A_1, ..., A_n$ are assumed to be mutually exclusive, which means that $(B \cap A_1), ..., (B \cap A_n)$ are also mutually exclusive. Applying Lemma 2.2.2:

$$P\left(\bigcup_i (B \cap A_i)\right) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B \mid A_i) P(A_i)$$

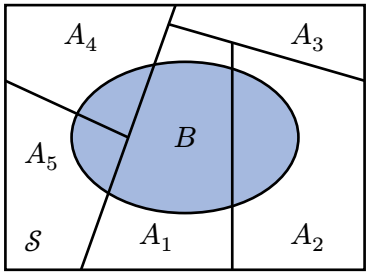


Figure 11: Geometric interpretation of the Law. Five mutually exclusive events $A_1, ..., A_5$ partition the sample space, possibly including B.

□

To better visualize chains of events, a **probability tree** can be drawn. The nodes correspond to a branching, while the edges correspond to the different branches. Each branching corresponds to a collection of mutually exclusive events that partition the sample space, the edges are labeled with the probability of such branching taking place. Summing probabilities in any subtree should add up to 1.

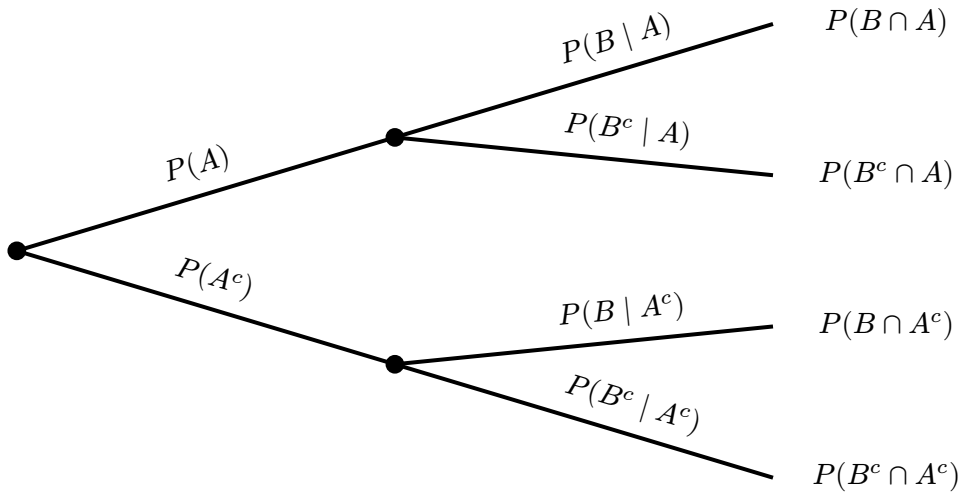


Figure 12: Probability tree for two events, A and B. By definition $A \cup A^c = \mathcal{S}$ and $A \cap A^c = \emptyset$, making it a partition. Same goes for B.

Exercise 2.4.2: An electronics store sells three different brands of DVD players. Of its DVD player sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3. Each manufacturer offers a 1-year warranty on parts and labor. It is known that 25% of brand 1's DVD players require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.

1. What is the probability that a randomly selected purchaser has bought a brand 1 DVD player that will need repair while under warranty?
2. What is the probability that a randomly selected purchaser has a DVD player that will need repair while under warranty?

Solution: Let A_i with $i \in \{1, 2, 3\}$ be the event "the DVD player comes from the i-th brand", and let B be the event "the DVD player needs to be repaired". The experiment is performed in two steps, first a DVD player is chosen, then knowing its brand it is determined whether it needs to be repaired. The probabilities are then distributed as follows:

$$P(A_1) = \frac{1}{2} \qquad P(A_2) = \frac{3}{10} \qquad P(A_3) = \frac{1}{5}$$
$$P(B \mid A_1) = \frac{1}{4} \qquad P(B \mid A_2) = \frac{1}{5} \qquad P(B \mid A_3) = \frac{1}{10}$$

Notice how *"It is known that 25% of..."* denotes a conditional probability, not an intersection of probabilities, because it represents a knowledge update. This is better understood by rephrasing *"It is known that 25% of brand 1's DVD players require warranty repair work"* as the less misleading *"Knowing that the DVD player at hand comes from brand 1, there's a 25% probability of it requiring warranty repair work"*.

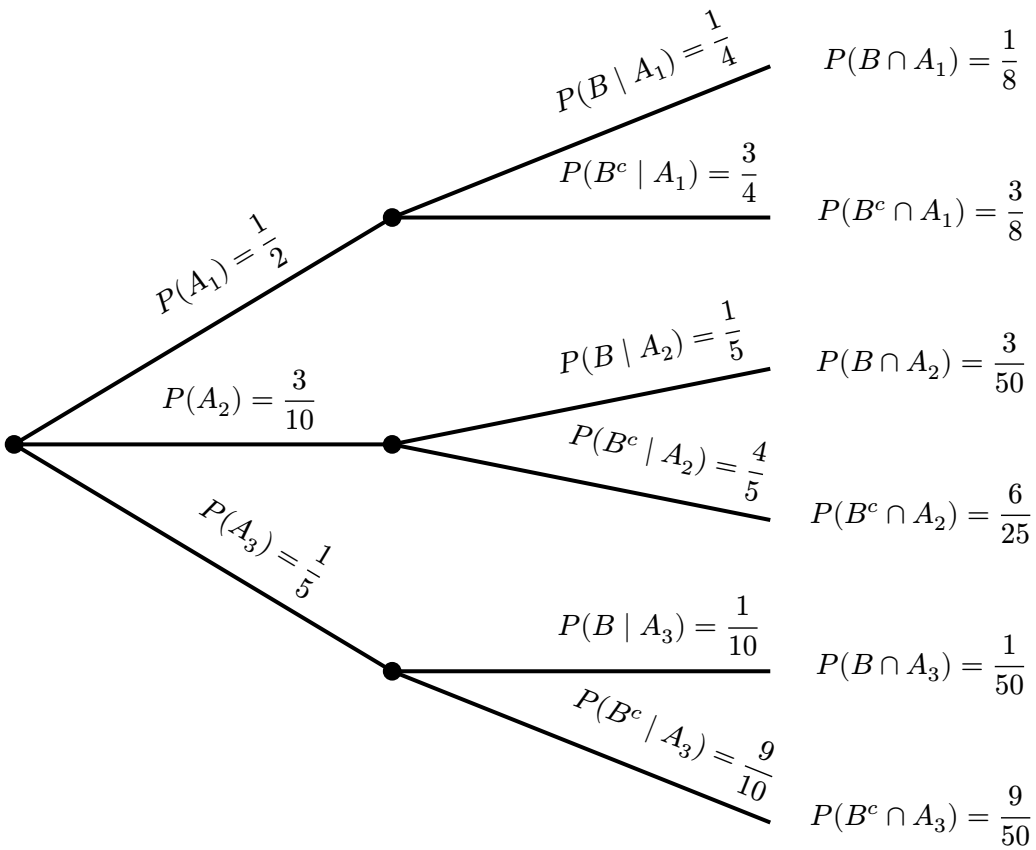


Figure 13: The probability tree in its entirety.

The probability of a DVD player coming from brand i will need repair is $P(B \cap A_i)$. By multiplication rule:

$$P(B \cap A_1) = P(B \mid A_1)P(A_1) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$$
$$P(B \cap A_2) = P(B \mid A_2)P(A_2) = \frac{1}{5} \cdot \frac{3}{10} = \frac{3}{50}$$
$$P(B \cap A_3) = P(B \mid A_3)P(A_3) = \frac{1}{10} \cdot \frac{1}{5} = \frac{1}{50}$$

The probability of any DVD player to need repair is $P(B)$. Applying Theorem 2.4.1:

$$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3) = \frac{1}{8} + \frac{3}{50} + \frac{1}{50} = \frac{41}{200}$$

□

Conditional probability plays a crucial role in Bayesian statistics. In this interpretations, probability assignments are performed by an agent based on the information that it possesses on the phenomena under analysis. The more information is collected, the more "refined" the assignment becomes.

In this sense $P(A)$ represents the initial assignment, an "absolute probability" based on the information the agent has *a priori*, while $P(A \mid B)$ is an update on the assignment, incorporating the knowledge carried by B. This is why in Bayesian statistics $P(A)$ and $P(A \mid B)$ are respectively referred to as **prior probability** and **posterior probability**. If A is independent of B ($P(A) = P(A \mid B)$), it means that B provided no contribution to the knowledge of the phenomena.

Theorem 2.4.2 (Bayes' theorem): Let $\mathcal{S}$ be a sample space. Let $A_1, A_2, ..., A_n \subseteq \mathcal{S}$ be a finite collection of mutually exclusive events partitioning $\mathcal{S}$, and let $B \subseteq \mathcal{S}$ be any event such that $P(B) > 0$. For any A_i with $i \in \{1, ..., n\}$, the following holds:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

Proof: By multiplication rule:

$$P(A_i \cap B) = P(A_i \mid B)P(B) \qquad\qquad P(B \cap A_i) = P(B \mid A_i)P(A_i)$$

Since $A_i \cap B = B \cap A_i$, the first expression can be substituted in the second:

$$P(B \cap A_i) = P(B \mid A_i)P(A_i) = P(A_i \mid B)P(B) \Rightarrow P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}$$

Since $A_1, ..., A_n$ is a collection of mutually exclusive events that partition $\mathcal{S}$, it is possible to apply Theorem 2.4.1 at the denominator:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

□

The fact that inverting conditional probabilities using Theorem 2.4.2 can lead to completely different values breeds apparent paradoxes.

Exercise 2.4.3 (Medical test paradox): Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time (*true positive*), whereas an individual without the disease will show a positive test result only (*false positive*) of the time. If a randomly selected individual is tested and the result is positive, what is the probability of the individual to really have the disease?

Solution: Let A be the event "the individual has the disease" and let B be the event "the test returns positive". The probability of randomly selecting an individual with the disease is $P(A) = 0.001$. A true positive corresponds to $B \mid A$, having probability 0.99, while a false positive corresponds to $B \mid A^c$ with probability 0.02. The event "knowing that the test is positive, the individual has the disease" would correspond to $A \mid B$. Applying Theorem 2.4.2:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{\sum_{j=1}^{2} P(B \mid A_j)P(A_j)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)} =$$

$$= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.02 \cdot (1 - 0.001)} \approx 0.05$$

The apparent paradox stems from the fact that the diagnostic test is very accurate (99% chance of a true positive) and yet a very tiny fraction of people tested positive actually have the disease. The explanation is that the disease is so rare that most of the positive test results come from errors rather than from the individuals actually having the disease. If the disease were to be more common, this discrepancy wouldn't arise. □

Let $\mathcal{S}$ be a sample space, and let $A, B \subseteq \mathcal{S}$ two events. If $P(A) = P(A \mid B)$, meaning that the occurrence of B does not influence the chances of A to occur, A is said to be **independent** of B. If $P(A) \neq P(A \mid B)$, A is said to be **dependent** on B.

Dependence and mutual exclusiveness are two distinct concepts. If $A, B \in \mathcal{S}$ are mutually exclusive, then they are certainly also dependent, because knowing that one has happened implies that the other cannot happen. Indeed, if A and B are mutually exclusive then $P(A \cap B) = 0$, which gives $P(A \mid B) = P(A \cap B)/B = 0$.

Corollary 2.4.1: Given a sample space $\mathcal{S}$ and two events $A, B \subseteq \mathcal{S}$, if A is independent of B then B is independent of A. That is, event independence is symmetric.

Proof: If A is independent of B, then $P(A \mid B) = P(A)$. Applying Theorem 2.4.2 gives:

$$P(A \mid B) = \cancel{P(A)} = \frac{P(B \mid A)\cancel{P(A)}}{P(B)} \Rightarrow P(B \mid A) = P(B)$$

Which is the definition of independence. □

Independence can be extended to three or more events. Given a collection of n events $A_1, A_2, ..., A_n$ over a sample space $\mathcal{S}$, such events are said to be **mutually independent** if for every $k = 2, 3, ..., n$ and for every subset of indices $i_1, i_2, ..., i_k$:

$$P\big(A_{i,1} \cap A_{i,2} \cap ... \cap A_{i,k}\big) = P\big(A_{i,1}\big) \cdot P\big(A_{i,2}\big) \cdot ... \cdot P\big(A_{i,k}\big)$$

3. Random variables

3.1. Random variables

Approaching experiments by constructing a model that operates case by case (tossing a coin, rolling a dice, drawing a card, ecc...) is insufficient if the goal is to put probability theory on solid grounds. What is needed is to *abstract* the modeling aspects in such a way that they are applicable everywhere.

Instead of referring to the possible outcomes of an experiment as they are, a better way is to map each outcome (each simple event) to a number with a fixed mapping rule. For a given sample space $\mathcal{S}$, any mapping from $\mathcal{S}$ to $\mathbb{R}$ is called a **random variable**. The name "random variable" highlights two aspects: it's a "variable" because different numerical values are possible, but it's also "random" because the observed value depends on which of the possible experimental outcomes results.

Random variables are generally denoted with uppercase letters. For a given sample space $\mathcal{S}$ and an event $s \in \mathcal{S}$, let $X : \mathcal{S} \mapsto \mathbb{R}$ be a random variable. The real number that is mapped by X to s is $X(s)$. X may not be injective, meaning that more than one event can be mapped to the same number. For any random variable X, the set of possible values that X can return (its image) is also called its **support**, denoted as $D(X)$.

Like events, it's possible to assign probabilities to random variables: the probability assigned to a certain mapping $X(s)$ with $s \in \mathcal{S}$ is just the probability assigned to s, meaning that $P(X(s)) = P(s)$. With a slight abuse of notation, it is much more common to write $P(X(s))$ as $P(X = X(s))$, emphasising the value returned by the variable rather than the outcome of the experiment which resulted in said value.

Exercise 3.1.1: Consider an experiment consisting in tossing three coins. Construct a random variable that maps each outcome to the number of heads appearing in that outcome.

Solution: The sample space of the experiment is (referring to Exercise 2.1.1):

$$\mathcal{S} = \{TTT, TTH, THT, HTT, THH, HHT, HTH, HHH\}$$

The number of heads in each possible outcome is $0, 1, 2$ or 3. These are captured by the following events:

$$\text{Zero heads} = \{TTT\} \qquad \text{Two heads} = \{THH, HHT, HTH\}$$
$$\text{One head} = \{TTH, THT, HTT\} \qquad \text{Three heads} = \{HHH\}$$

Let X be the random variable that maps each possible event to a number, counting the number of heads in such outcome. The support of X is $D(X) = \{0, 1, 2, 3\}$. Since $\mathcal{S}$ is discrete, the mappings described by X can be enumerated:

$$X(TTT) = 0 \qquad X(THH) = X(HHT) = X(HTH) = 2$$
$$X(TTH) = X(THT) = X(HTT) = 1 \qquad X(HHH) = 3$$

Probabilities are assigned by invoking the Principle of Indifference:

$$P(X=0) = P(\{TTT\}) = \frac{1}{8} \qquad P(X=2) = P(\{THH, HHT, HTH\}) = \frac{3}{8}$$
$$P(X=1) = P(\{TTH, THT, HTT\}) = \frac{3}{8} \quad P(X=3) = P(\{HHH\}) = \frac{1}{8}$$

□

Random variables fall in two broader categories: **discrete** and **continuous**. A random variable is said to be discrete if the set of values it can assume is either finite or countably infinite. A random variable is said to be continuous if the two following properties apply:

1. Its set of possible values consists either of all numbers in a single (possibly infinite) interval on the real line or all numbers in a disjoint union of such intervals;
2. The probability of the random variable to assume a specific value is always zero.

3.2. Discrete random variables

Instead of referring to the probability values assigned to the elements of the support of a random variable one by one, it is more convenient to group all those probabilities into a single function.

The **probability mass function** (**pmf** for short) of a discrete random variable X with support $D(X)$ is a function that assigns a probability to all possible values that X can take. The pmf of X is denoted as $p(X)$, and is formally defined as:

$$p(x) = P(X = x) = P(s : s \in \mathcal{S}, X(s) = x)$$

Note that, again, there could be more than one s such that $X = X(s)$; any choice is valid. Also, $p(x)$ must be 0 for any $x \notin D(X)$, because if there is no $s \in \mathcal{S}$ such that $X(s) = x$, then $p(x) = p(\emptyset) = 0$.

To ensure compliance with the Kolmogorov axioms, any pmf $p(X)$ must satisfy two conditions:

$$p(x) \geq 0 \ \forall x \in D(X) \qquad \sum_{x \in D(X)} p(x) = 1$$

Exercise 3.2.1: What is the pmf of the variable X in Exercise 3.1.1?

Solution:

$$p(x) = P(X = x) = \begin{cases} \frac{1}{8} \text{ if } x = 0 \vee x = 3 \\ \frac{3}{8} \text{ if } x = 1 \vee x = 2 \\ 0 \text{ otherwise} \end{cases}$$

This is a valid pmf because it's always non-negative and the sum of the probabilities over the entire support is 1:

$$\sum_{x \in \{0,1,2,3\}} p(x) = p(0) + p(1) + p(2) + p(3) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$$

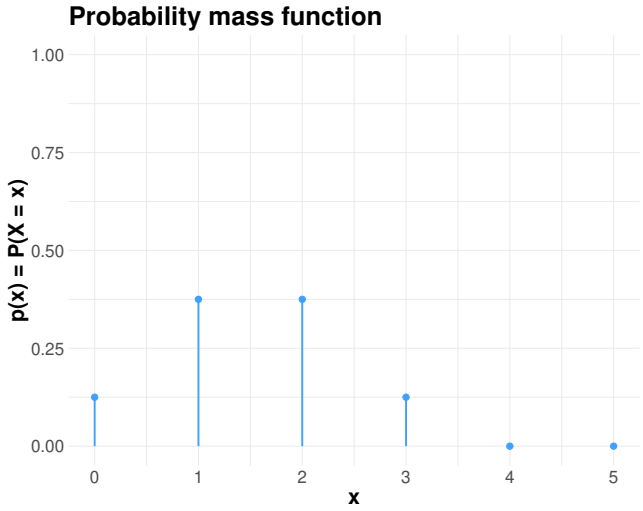


Figure 14: Plot of the probability mass function $p(x) = P(X = x)$

□

The **cumulative distribution function (cdf** for short) of a discrete random variable X is defined as the probability of such random variable to take a value less than or equal to a threshold. The cdf of X is denoted as $F(X)$, and is formally defined as:

$$F(x) = P(X \leq x) = \sum_{y: y \in D(X), y \leq x} p(y)$$

For any $x \notin D(X)$, $F(x)$ returns the cdf of closest possible value of X to the left of x. In particular, if there is no $y \in D(X)$ such that $y \leq x$, then $F(x) = 0$.

Exercise 3.2.2: A store carries flash drives with either 1 GB, 2 GB, 4 GB, 8 GB, or 16 GB of memory. Let X be the random variable "amount of memory in a purchased drive". The pmf of X is as follows:

$$p(x) = \begin{cases} 0.05 \text{ if } x = 1 \\ 0.10 \text{ if } x = 2 \\ 0.35 \text{ if } x = 4 \\ 0.40 \text{ if } x = 8 \\ 0.10 \text{ if } x = 16 \\ 0 \quad \text{otherwise} \end{cases}$$

What is the cdf of X?

Solution:

$$F(1) = \sum_{y:y\in D(X),y\leq 1} p(y) = p(1) = 0.05$$
$$F(2) = \sum_{y:y\in D(X),y\leq 2} p(y) = p(1) + p(2) = 0.05 + 0.10 = 0.15$$
$$F(4) = \sum_{y:y\in D(X),y\leq 4} p(y) = p(1) + p(2) + p(4) = 0.15 + 0.35 = 0.50$$
$$F(8) = \sum_{y:y\in D(X),y\leq 8} p(y) = p(1) + p(2) + p(4) + p(8) = 0.50 + 0.40 = 0.90$$
$$F(16) = \sum_{y:y\in D(X),y\leq 16} p(y) = p(1) + p(2) + p(4) + p(8) + p(16) = 0.90 + 0.10 = 1.00$$

To sum up:

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.05 & \text{if } 1 \leq x < 2 \\ 0.15 & \text{if } 2 \leq x < 4 \\ 0.5 & \text{if } 4 \leq x < 8 \\ 0.9 & \text{if } 8 \leq x < 16 \\ 1 & \text{if } x \geq 16 \end{cases}$$

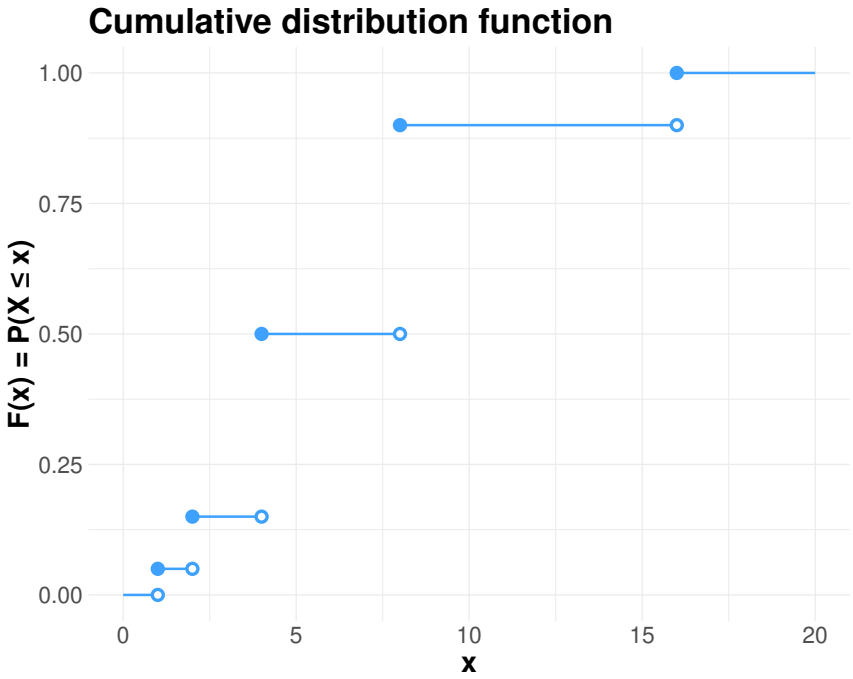


Figure 15: Plot of the cumulative distribution function $F(x) = P(X \leq x)$

□

The cdf can be obtained from the pmf, but it's also possible to go the other way around: obtaining the pmf from the cdf.

For a given sample space $\mathcal{S}$, let $X : \mathcal{S} \mapsto \mathbb{R}$ be a discrete random variable with known cdf $F(x)$ and known support $D(X) = x_0, x_1, \ldots$. Suppose that $a \in D(X)$ is a value for which there's interest in computing $p(a)$; let $a^- \in D(X)$ be highest value in the support of X being smaller than a. That is, $a^- = \max_{x\in D(X)}\{x \mid x < a\}$. To obtain $p(a)$, one computes:

$$\begin{aligned}F(a) - F(a^-) = &\sum_{y:y\in D(X), y\leq a} p(y) - \sum_{y:y\in D(X), y\leq a^-} p(y) = \\ &= (p(x_0) + p(x_1) + ... + p(a^-) + p(a)) - (p(x_0) + p(x_1) + ... + p(a^-)) = \\ &= \cancel{p(x_0)} + \cancel{p(x_1)} + ... + \cancel{p(a^-)} + p(a) - \cancel{p(x_0)} - \cancel{p(x_1)} - ... - \cancel{p(a^-)} = \\ &= p(a)\end{aligned}$$

In general, to obtain the probability of a random variable to take any value that lies in the interval $[a, b]$, one computes:

$$\begin{aligned}F(b) - F(a^-) = &\sum_{y:y\in D(X), y\leq b} p(y) - \sum_{y:y\in D(X), y\leq a^-} p(y) = \\ &= (p(x_0) + ... + p(a^-) + p(a) + ... + p(b)) - (p(x_0) + ... + p(a^-)) = \\ &= \cancel{p(x_0)} + ... + \cancel{p(a^-)} + p(a) + ... + p(b) - \cancel{p(x_0)} - ... - \cancel{p(a^-)} = \\ &= p(a) + ... + p(b) = P(X = a \vee ...X = b) = \\ &= P(a \leq x \leq b)\end{aligned}$$

Notice how it's necessary to use a^- and not a, since otherwise the leftmost extreme would be left out.

Exercise 3.2.3: Consider Exercise 3.2.2. What is $P(2 \leq x \leq 8)$?

Solution: Being $D(X) = \{1, 2, 4, 8, 16\}$, one has $2^- = 1$. Therefore:

$$P(2 \leq x \leq 8) = F(8) - F(2^-) = F(8) - F(1) = 0.9 - 0.05 = 0.85$$

□

Let X be a discrete random variable with support $D(X)$ and probability mass function $p(X)$. The **expected value** (or **mean value**) of X, denoted as $E(X)$ or μ_X, is the sum of all possible values in $D(X)$ weighted by their probabilities:

$$E(X) = \mu_X = \sum_{x\in D(X)} x \cdot p(x)$$

When the variable X is known from context, the pedix X in μ_X is omitted.

The expected value is, in some sense, the counterpart of the sample mean, but with different weights to each term of the sum. Like the sample mean, it is perfectly valid for the expected value to be a value that is not present in the support of the related variable. Interestingly, the expected value is not guaranteed to exist.

Exercise 3.2.4: What is the expected value of the variable X in Exercise 3.2.2?

Solution:

$$\begin{aligned}E(X) = &\sum_{x\in D(X)} x \cdot p(x) = 1 \cdot 0.05 + 2 \cdot 0.10 + 4 \cdot 0.35 + 8 \cdot 0.40 + 16 \cdot 0.10 = \\ &= 0.05 + 0.20 + 1.40 + 3.20 + 1.60 = 6.45\end{aligned}$$

□

The expected value is oblivious to whether its argument is a random variable or a function whose input is a random variable. In other words, let X be a discrete random variable with support $D(X)$ and pmf $p(X)$, and let $h(X)$ be a function whose argument is X itself. The expected value of $h(X)$ is still defined as:

$$E(h(X)) = \mu_{h(X)} = \sum_{x \in D(X)} h(x) \cdot p(x)$$

Theorem 3.2.1: Let X be a discrete random variable with support $D(X)$ and pmf $p(X)$. For any $a, b \in \mathbb{R}$:

$$E(aX + b) = aE(X) + b$$

Proof: Let $h(x) = aX + b$. Then:

$$\begin{aligned} E(h(x)) = E(aX + b) = \sum_{x \in D(X)} h(x) \cdot p(x) = \sum_{x \in D(X)} (ax + b) \cdot p(x) = \\ = \sum_{x \in D(X)} ax \cdot p(x) + b \cdot p(x) = \sum_{x \in D(X)} ax \cdot p(x) + \sum_{x \in D(X)} b \cdot p(x) = \\ = a \sum_{x \in D(X)} x \cdot p(x) + b \sum_{x \in D(X)} p(x) = aE(X) + b \cdot 1 = aE(X) + b \end{aligned}$$

Since $\sum_{x \in D(X)} p(x) = 1$ by definition. □

Let X be a discrete random variable with support $D(X)$ and probability mass function $p(X)$. The **variance** of X, denoted as $V(X)$ or σ_X^2, is a measure of how the values in $D(X)$ differ from the expected value:

$$V(X) = \sigma_X^2 = \sum_{x \in D(X)} (x - E(X))^2 \cdot p(x) = E((X - E(X))^2)$$

When the variable X is known from context, the pedix X in σ_X^2 is omitted. If the expected value is the counterpart to the sample mean, the variance is the counterpart to the sample variance.

Exercise 3.2.5: What is the variance of the variable X in Exercise 3.2.2?

Solution:

$$\begin{aligned} V(X) = \sum_{x \in D(X)} (x - E(X))^2 \cdot p(x) = (1 - 6.45)^2 \cdot 0.05 + ... + (16 - 6.45)^2 \cdot 0.10 = \\ = 29.7 \cdot 0.05 + 19.8 \cdot 0.10 + 6.0 \cdot 0.35 + 2.4 \cdot 0.40 + 91.2 \cdot 0.10 \approx \\ \approx 1.49 + 1.98 + 2.10 + 0.96 + 9.12 \approx 15.65 \end{aligned}$$

□

The square root of the variance is called the **standard deviation**, denoted as $SD(X)$ or σ_X:

$$SD(X) = \sigma_X = \sqrt{V(X)}$$

Lemma 3.2.1 gives a simpler way to compute the variance of a discrete random variable.

Lemma 3.2.1: Let X be a discrete random variable with support D and probability mass function $p(X)$. The following equality holds:

$$V(X) = \left(\sum_{x \in D} x^2 \cdot p(x)\right) - (E(X))^2 = E(X^2) - (E(X))^2$$

Proof: Expanding the square in the formula:

$$\begin{aligned}
V(X) &= \sum_{x \in D(X)} (x - E(X))^2 \cdot p(x) = \sum_{x \in D(X)} (x^2 + (E(X))^2 - 2xE(X)) \cdot p(x) = \\
&= \sum_{x \in D(X)} x^2 \cdot p(x) + \sum_{x \in D(X)} (E(X))^2 \cdot p(x) - \sum_{x \in D(X)} 2xE(X) \cdot p(x) = \\
&= \left(\sum_{x \in D(X)} x^2 \cdot p(x)\right) + (E(X))^2 \left(\sum_{x \in D(X)} p(x)\right) - 2E(X)\left(\sum_{x \in D(X)} x \cdot p(x)\right) = \\
&= E(X^2) + (E(X))^2 \cdot 1 - 2E(X) \cdot E(X) = E(X^2) + (E(X))^2 - 2(E(X))^2 = \\
&= E(X^2) - (E(X))^2
\end{aligned}$$

□

Theorem 3.2.2: Let X be a discrete random variable with support $D(X)$ and pmf $p(X)$. For any $a, b \in \mathbb{R}$:

$$V(aX + b) = a^2 V(X)$$

Proof: Applying Theorem 3.2.1:

$$\begin{aligned}
V(aX + b) &= E(((aX + b) - E(aX + b))^2) = E((aX + \cancel{b} - aE(X) - \cancel{b})^2) = \\
&\quad -E((a(X - E(X)))^2) = E(a^2(X - E(X))^2) = a^2 E((X - E(X))^2) = \\
&\quad = a^2 V(X)
\end{aligned}$$

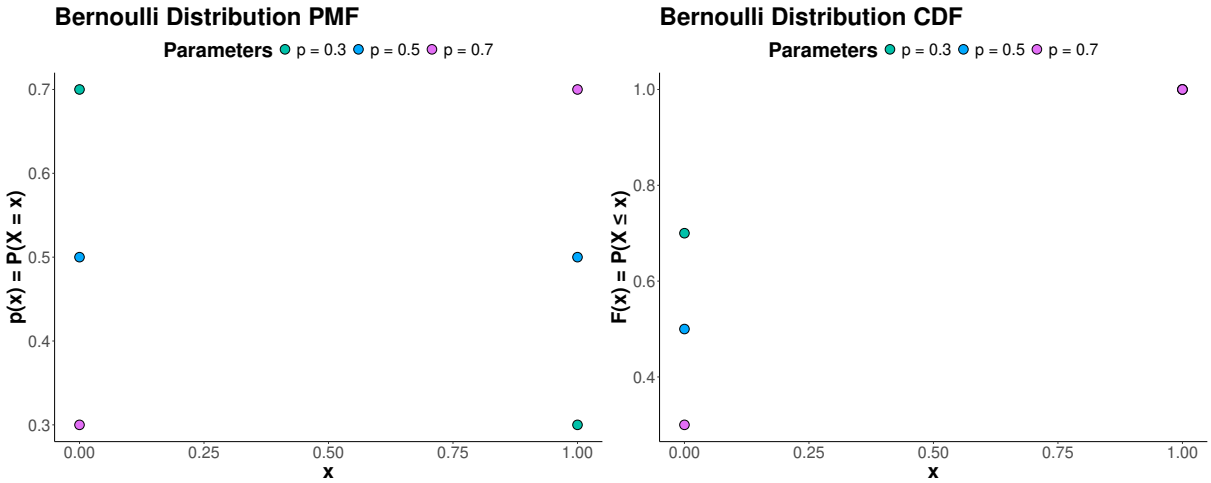
□

3.3. Known discrete random variables

Some specific discrete random variables have been studied extensively, mostly because they model very well many phenomena in the real world. For this reason, such random variables have proper names. To denote that a random variable X has the same distribution as a known random variable F, the notation $X \sim F$ is used.

3.3.1. Bernoulli random variable

A discrete random variable X is distributed as a **Bernoulli random variable** of parameter $p \in [0, 1]$ (denoted as $X \sim B(p)$) if it can assume exclusively the values 1 and 0, with probabilities p and $1 - p$ respectively. The pdf and cdf of a Bernoulli random variable X of parameter p are therefore as follows:



$$p(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \qquad F(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Bernoulli random variables model experiments that have two mutually exclusive results: success ($X = 1$) or failure ($X = 0$), with nothing in between.

Theorem 3.3.1.1: The expected value and variance of a random variable $X \sim B(p)$ are as follows:

$$E(X) = p \qquad V(X) = p(1-p)$$

Proof:

$$\begin{aligned} E(X) &= 0 \cdot (1-p) + 1 \cdot p = \\ &= 0 + p = \\ &= p \end{aligned}$$

$$\begin{aligned} V(X) &= (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = \\ &= p^2(1-p) + p(1-p)^2 = \\ &= (p^2 + p(1-p))(1-p) = \\ &= (\cancel{p^2} + p - \cancel{p^2})(1-p) = \\ &= p(1-p) \end{aligned}$$

□

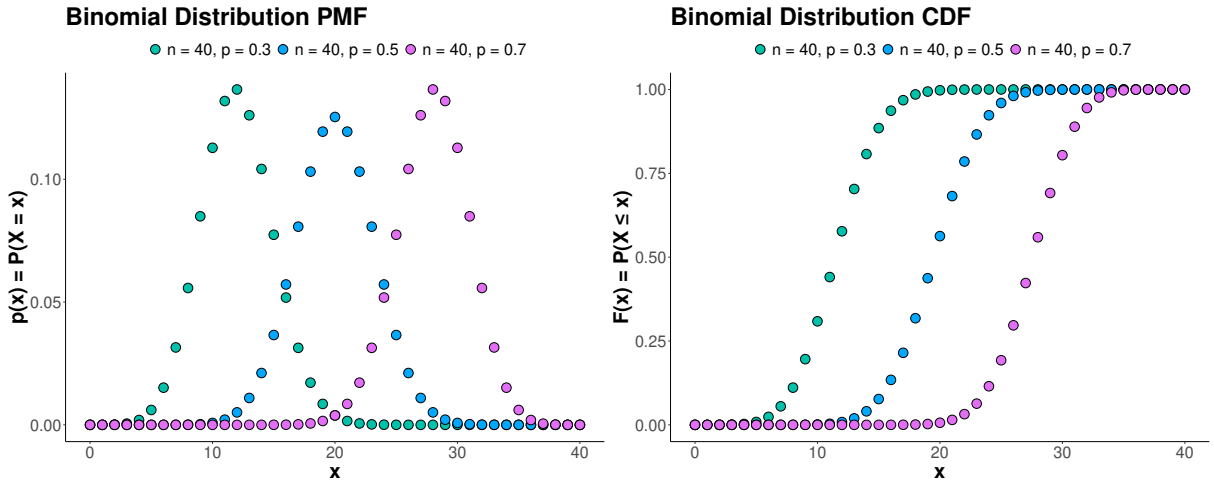
3.3.2. Binomial random variable

Let $Y_1, Y_2, ..., Y_n$ be n independent and identically distributed Bernoulli random variables (all having the same parameter p). Let X be the random variable defined as the sum of all said variables:

$$X = \sum_{i=1}^{n} Y_i = Y_1 + Y_2 + ... + Y_n$$

The random variable X defined as such is distributed as a **binomial random variable** of parameters p and n (denoted as $X \sim Bi(n, p)$).

Since a specific realization of X is a sum of 0s and 1s, a realization k is simply the number of Bernoulli variables that define X that had assumed value 1. The pdf and cdf of a binomial random variable X of parameters n and p are therefore as follows:



$$p(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } \begin{cases} x \in \mathbb{N} \\ x \leq n \end{cases} \\ 0 & \text{otherwise} \end{cases} \quad F(x) = P(X \leq x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$$

Theorem 3.3.2.1: The expected value and variance of a random variable $X \sim Bi(p, n)$ are as follows:

$$E(X) = np \qquad V(X) = np(1-p)$$

Proof: This result can be proved by applying Theorem 3.6.1 and Theorem 3.6.2 (the latter can be applied since the Bernoulli random variables that constitue X are independent).

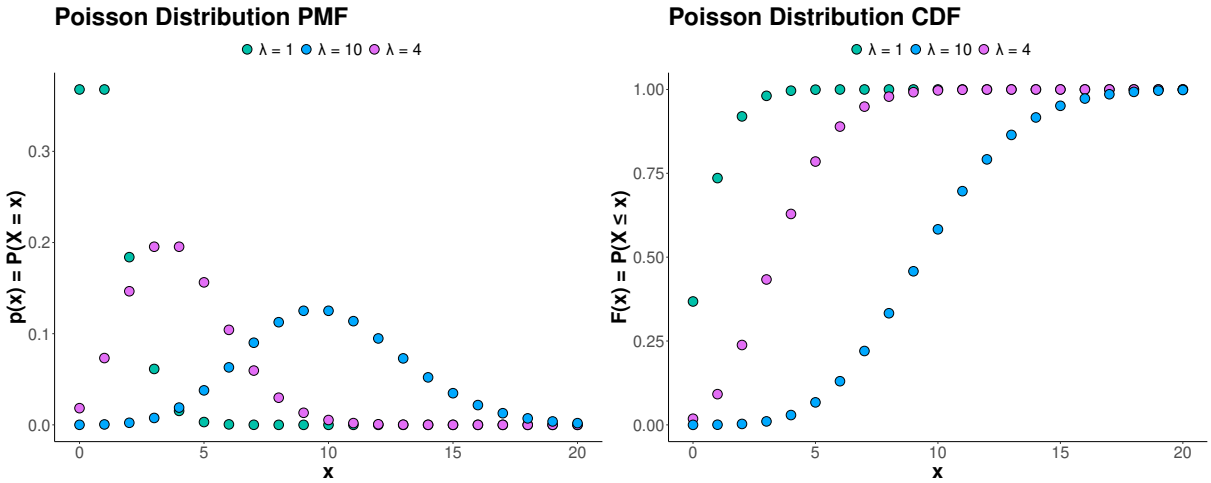
$$E(X) = E(Y_1 + Y_2 + ... + Y_n) = E(Y_1) + E(Y_2) + ... + E(Y_n) = nE(Y_i) = np$$
$$V(X) = V(Y_1 + Y_2 + ... + Y_n) = V(Y_1) + V(Y_2) + ... + V(Y_n) = nV(Y_i) = np(1-p)$$

Where $E(Y_i)$ and $V(Y_i)$ are retrieved from Theorem 3.3.1.1. □

Binomial random variables model experiments composed by many mutually exclusive results.

3.3.3. Poisson random variable

Let Y a binomial random variable, and let $\lambda \in \mathbb{R}^+$ be the product of its parameters n and p. By applying the double limit $n \to \infty, p \to 0$ while keeping their product constant a new random variable X is constructed, called a **Poisson random variable** (denoted as $X \sim \text{Pois}(\lambda)$). The pdf and cdf of a Poisson random variable X of parameter λ are therefore as follows:



$$p(x) = P(X = x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

$$F(x) = P(X \leq x) = \sum_{k \in \mathbb{N}, k \leq x} \frac{\lambda^k}{k!} e^{-\lambda}$$

Theorem 3.3.3.1: The expected value and variance of a random variable $X \sim \text{Pois}(\lambda)$ are as follows:

$$E(X) = \lambda \qquad V(X) = \lambda$$

Proof: Let $Y \sim Bi(n, p)$ be a random variable to which the double limit $n \to \infty, p \to 0$ is applied, and let $\lambda = np$. This results in:

$$E(X) = \lim_{\substack{n \to \infty \\ p \to 0}} E(Y) = \lim_{\substack{n \to \infty \\ p \to 0}} np = \lambda$$

$$V(X) = \lim_{\substack{n \to \infty \\ p \to 0}} V(Y) = \lim_{\substack{n \to \infty \\ p \to 0}} np(1-p) = \lambda(1-0) = \lambda$$

□

Exercise 3.3.3.1: Let X be the number of traps occurring in a particular type of transistor. Suppose $X \sim \text{Pois}(2)$; what is the probability of retrieving 3 traps? What is the probability of retrieving 3 or less traps?

Solution:

$$p(3) = P(X = 3) = \frac{2^3}{3!} e^{-2} = \frac{8}{6} e^{-2} \approx 0.18$$

$$F(3) = P(X \leq 3) = \sum_{k \in \mathbb{N}, k \leq 3} \frac{2^k}{k!} e^{-2} = \frac{2^0}{0!} e^{-2} + \frac{2^1}{1!} e^{-2} + \frac{2^2}{2!} e^{-2} + \frac{2^3}{3!} e^{-2} =$$

$$= e^{-2} \left(\frac{1}{1} + \frac{2}{1} + \frac{4}{2} + \frac{8}{6} \right) = e^{-2} \frac{19}{3} \approx 0.86$$

□

The Poisson distribution model events where the size of the population is very large and the probability of the event to occur is very small. This is why the Poisson distribution is used to model *rare events*, events that have a very slim, but still relevant, probability to occur in a certain span of time. More formally, a rare event can be modeled as such if the following properties hold:

1. There exist a parameter $\alpha > 0$ such that for any short time interval of length Δt, the probability that exactly one event occurs is $\alpha \Delta t \cdot o(\Delta t)$, where $o(\Delta t)$ is a little-o of Δt;
2. The probability of more than one event occurring during Δt is $o(\Delta t)$. In other words, it is much more likely that a single event happens during Δt than multiple events occur;
3. The number of events occurring during the time interval Δt is independent of the number that occur prior to this time interval.

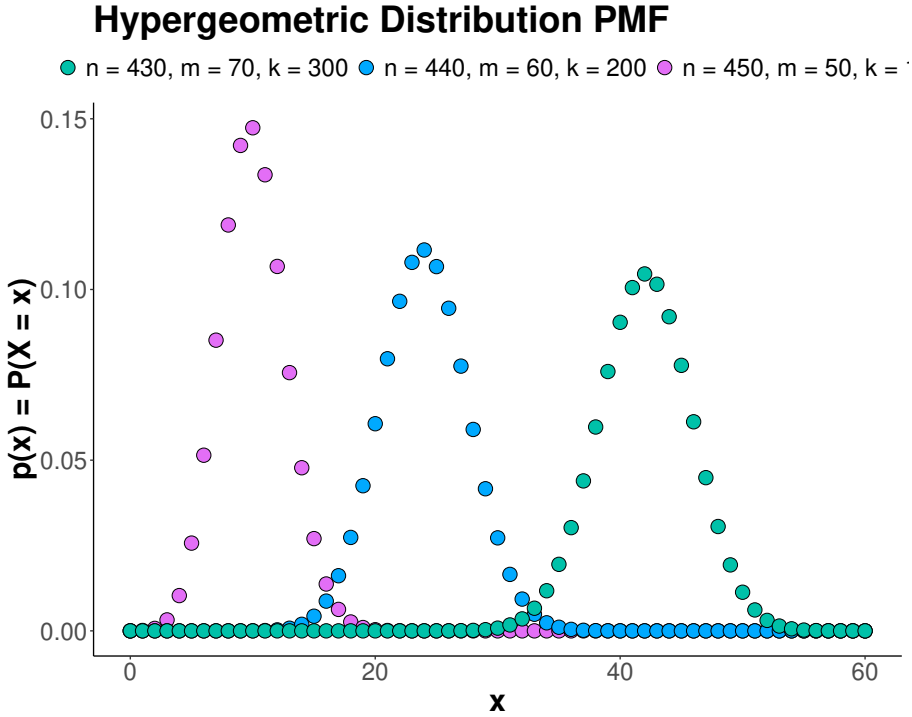
The probability mass function of a Poisson distribution can be adapted in this sense if, instead of the expected value λ, one is given α, the expected number of events occurring in a unitary time interval, and a time interval Δt. The probability of k events to occur in a time slice Δt is then as follows:

$$p_k(\Delta t) = \frac{(\alpha \Delta t)^k}{k!} e^{-\alpha \Delta t}$$

The occurrence of events over time as described is called a **Poisson process** and the parameter α specifies the *rate* of said process.

3.3.4. Hypergeometric random variable

Let N be the size of a population of individuals, each of them having associated either a value of 1 (success) or 0 (failure). Let M be the number of individuals whose value is 1, and therefore $N - M$ is the number of individuals whose number is 0. Let $n \leq N$ be the size of a sample extracted from the population. The random variable X whose values are the number of successes (of 1s) found in a sample of size n is said to distributed as an **hypergeometric random variable** (denoted $X \sim H(n, N, M)$). The pdf of an hypergeometric random variable X of parameters M, N and n is therefore as follows:



$$p(x) = P(X = x) = \begin{cases} \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{if } \max(0, n-N+M) \leq x \leq \min(n, M) \\ 0 & \text{otherwise} \end{cases}$$

The binomial $\binom{M}{x}$ is the number of ways it is possible to extract a sample where there are x individuals whose value is 1, while the binomial $\binom{N-M}{n-x}$ is the number of ways it is possible to extract a sample where there are $n-x$ individuals whose value is 0. The binomial $\binom{N}{n}$ is the number of combinations of elements of N of size n (without any requirement on the number of individuals having a particular value).

The constraint $x \leq \min(n, M)$ denotes that the number of observed successes cannot be greater than the size of the entire sample (and, of course, cannot be greater than the size of the entire population).

Exercise 3.3.4.1: A university IT office received 20 service orders for issues with printers, out of which 8 were laser printers and 12 were inkjet printers. A sample of 5 of these service orders were selected to perform a customer satisfaction survey. What is the probability that, out of those 5, 2 were inkjet printers?

Solution: It is possible to model this situation with an hypergeometric random variable. Since the outcome of interest is the one related to inkjet printers, the parameters of said variable X will be 5 for the sample size, 20 for the population size and 12 for the favorable population size. Therefore, $X \sim (5, 20, 12)$. Evaluating the pdf for $X = 2$ gives:

$$p(2) = P(X = 2) = \frac{\binom{12}{2}\binom{20-12}{5-2}}{\binom{20}{5}} = \frac{\frac{12!}{2!(12-2)!}\frac{8!}{3!(8-3)!}}{\frac{20!}{5!(20-5)!}} =$$

$$= \frac{\frac{12 \cdot 11 \cdot \cancel{10!}}{2 \cdot \cancel{10!}}\frac{8 \cdot 7 \cdot 6 \cdot \cancel{5!}}{6 \cdot \cancel{5!}}}{\frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot \cancel{15!}}{120 \cdot \cancel{15!}}} = \frac{22176}{93024} \approx 0.238$$

□

Theorem 3.3.4.1: The expected value and variance of a random variable $X \sim H(n, N, M)$ are as follows:

$$E(X) = n \cdot \frac{M}{N} \qquad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

The hypergeometric distribution is distinguished from the binomial distribution because the trials are not independent, since each time an individual is "inspected" it is removed from the sample, and therefore the subsequent probabilities are influenced by the outcome (since the number of individuals is decreased). By contrast, in the binomial distribution each trial is independent from the others.

Another similarity between the two comes from observing the equations is Theorem 3.3.4.1. The ratio M/N is the proportion of successes in the population, meaning that it's the probability of picking an

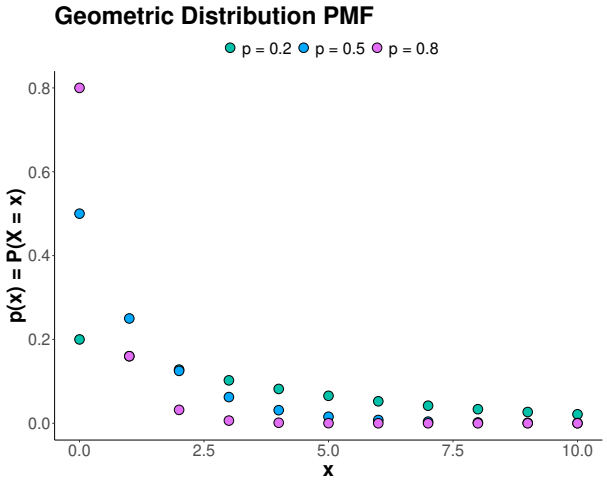
element of the entire population that has a value 1. This ratio has the same role that the parameter p has in the binomial distribution. Indeed, substituting M/N with p in said equations gives:

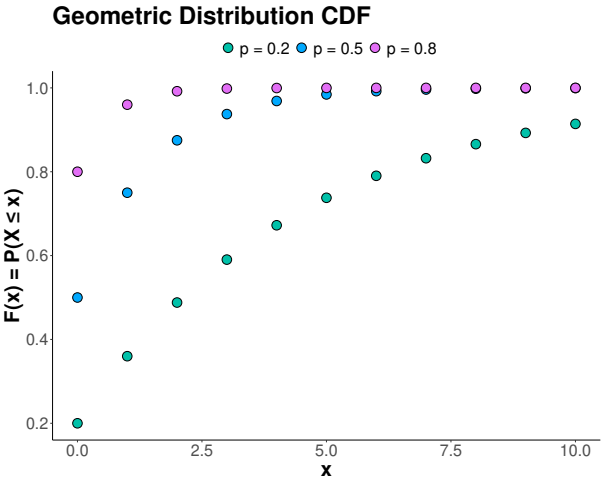
$$E(X) = np \qquad V(X) = \left(\frac{N-n}{N-1}\right) \cdot np(1-p)$$

Where the expected value is identical to the one of a binomial distributed random variable (Theorem 3.3.2.1), while the variance differs for a factor $(N-n)/(N-1)$. Since this factor, called **finite population correction factor**, is always less than 1, the variance of an hypergeometric random variable will always be smaller than a binomial random variable where $p = M/N$.

3.3.5. Geometric random variable

Let X be a random variable that represents the number of (failed) attempts necessary to have a Bernoulli random variable with parameter p to assume value 1. The random variable X is said to distributed as a **geometric random variable** (denoted $X \sim G(p)$). The pdf and cdf of a geometric random variable X of parameter p are therefore as follows:





$$p(x) = P(X = x) = \begin{cases} p(1-p)^x & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases} \qquad F(x) = P(X \leq x) = \sum_{k \in \mathbb{N}, k \leq x} p(1-p)^k$$

The factor $(1-p)^x$ represents the probability of obtaining a failure for exactly x times. This factor is then multiplied by p, which is the probability of obtaining a single success.

A geometric distribution $X \sim G(p)$ has a property called **memorylessness**, expressed mathematically as $P(X > x + y \mid X > y) = P(X > x)$ with x and y positive integers. In other words, the number of attempts necessary for an experiment to have a specific result does not depend on the previous ones.

Theorem 3.3.5.1: The expected value and variance of a random variable $X \sim G(p)$ are as follows:

$$E(X) = \frac{1-p}{p} \qquad V(X) = \frac{1-p}{p^2}$$

Proof: This result can be proven by applying known theorems concerning geometric functions:

$$
\begin{aligned}
E(X) &= p(1-p)^0 \cdot 0 + p(1-p)^1 \cdot 1 + p(1-p)^2 \cdot 2 + \ldots = \\
&= p(1-p) + 2p(1-p)^2 + 3p(1-p)^3 + \ldots = \\
&= \sum_{i=0}^{\infty} ip(1-p)^i = p\sum_{i=0}^{\infty} i(1-p)^i = p(1-p)\sum_{i=0}^{\infty} i(1-p)^{i-1} = \\
&= p(1-p)\left[\frac{d}{dp}\left(-\sum_{i=0}^{\infty}(1-p)^i\right)\right] = p(1-p)\frac{d}{dp}\left(-\frac{1}{p}\right) = \\
&= \cancel{p}(1-p)\left(\frac{1}{\cancel{p^2}}\right) = \frac{1-p}{p}
\end{aligned}
$$

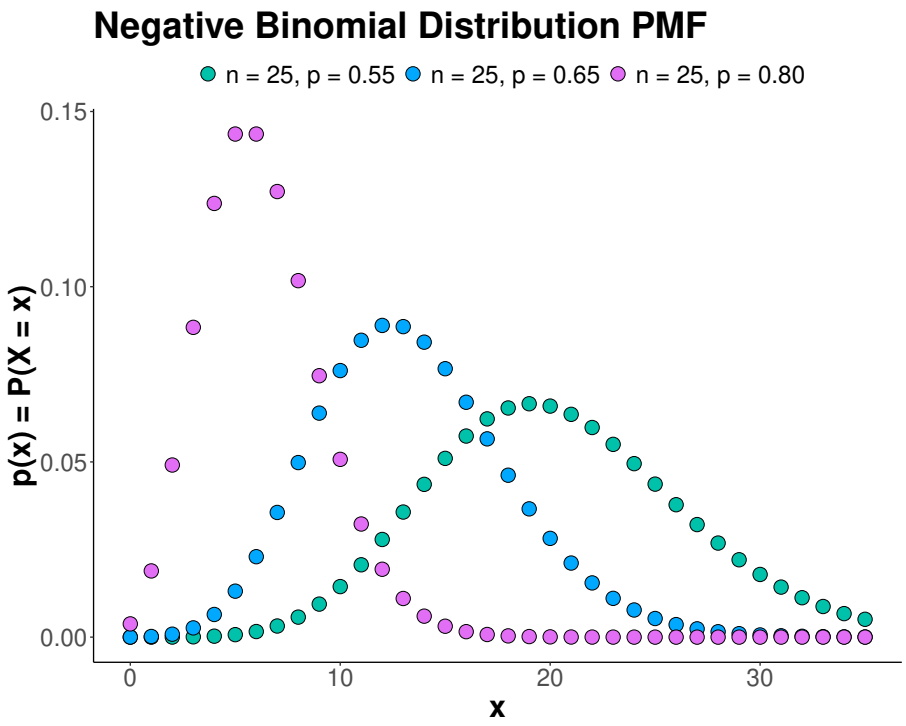
Then, applying Lemma 3.2.1:

$$
\begin{aligned}
V(X) &= E(X^2) - (E(X))^2 = \frac{(2-p)(1-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \\
&= \frac{2-2p-p+p^2}{p^2} - \frac{1+p^2-2p}{p^2} = \frac{2-3p+\cancel{p^2}-1-\cancel{p^2}+2p}{p^2} = \frac{1-p}{p^2}
\end{aligned}
$$

□

3.3.6. Negative binomial random variable

Let X be a random variable that represents the number of (failed) attempts necessary to have a Bernoulli random variable with parameter p to assume value 1 for r times. The random variable X is said to distributed as a **negative binomial random variable** (denoted $X \sim NB(r,p)$). The pdf of a negative binomial random variable X of parameters r and p is therefore as follows:



$$
p(x) = P(X = x) = \begin{cases} \binom{x+r-1}{r-1} p^r (1-p)^x & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}
$$

The factor $(1-p)^x$ represents the probability of obtaining a failure for exactly x times. The factor p^r represents the probability of obtaining a success r times. The factor $\binom{x+r-1}{r-1}$ represents the number of ways that $r-1$ successes out of $x+r-1$ attempts can be arranged.

Of course, if r is set to 1 said random variable reduces itself to a geometric random variable:

$$\binom{x+1-1}{1-1}p^1(1-p)^x = \binom{x}{0}p(1-p)^x = \left(\frac{x!}{0!(x-0)!}\right)p(1-p)^x = \left(\frac{\cancel{x!}}{\cancel{x!}}\right)p(1-p)^x = p(1-p)^x$$

3.4. Continuous random variables

Continuous random variables don't really have a concept of support, since a probability can be assigned to any real value. For this very reason, the only sensible way to describe the probability assignment for the values that it can take is through a function.

The **probability density function** (**pdf** for short) of a continuous random variable X is a function that, for any $a, b \in \mathbb{R}$, returns the probability of X to take on a value somewhere in $[a, b]$ when integrated in such interval. The pdf of X is denoted as $f(X)$, and is formally defined as:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

The graph of $f(x)$ is often referred to as the **density curve** of the random variable. Stated otherwise, $P(a \leq X \leq b)$ is given by the area under the section of the density curve of extremes a and b.

To ensure compliance with the Kolmogorov axioms, any pdf $f(X)$ must satisfy two conditions:

$$\forall x, f(x) \geq 0 \qquad\qquad \int_{-\infty}^{+\infty} f(x)dx = 1$$

Exercise 3.4.1: The amount of gravel (in tons) sold by a particular construction supply company in a given week can be modeled as a continuous random variable X with pdf:

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

If this pdf well-defined? If it is, what is the probability of the company having sold between 0.75 and 0.9 tons of gravel?

Solution: This is a valid pdf because it's always non-negative and when integrated over the entire number line gives 1:

$$\begin{aligned}\int_{-\infty}^{+\infty} f(x)dx &= \int_{-\infty}^{0} 0dx + \int_0^1 \frac{3}{2}(1-x^2)dx + \int_0^{+\infty} 0dx = 0 + \int_0^1 \frac{3}{2}(1-x^2)dx + 0 = \\ &= \frac{3}{2}\int_0^1 (1-x^2)dx = \frac{3}{2}\left(\int_0^1 1dx - \int_0^1 x^2dx\right) = \frac{3}{2}\left([x]_0^1 - \left[\frac{x^3}{3}\right]_0^1\right) = \\ &= \frac{3}{2}\left((1-0) - \left(\frac{1^3}{3} - \frac{0^3}{3}\right)\right) = \frac{3}{2}\left(1 - \frac{1}{3}\right) = \frac{3}{2}\cdot\frac{2}{3} = 1\end{aligned}$$

To get $P(0.75 \leq X \leq 0.9)$:

$$P(0.75 \leq X \leq 0.9) = \int_{0.75}^{0.9} f(x)dx = \frac{3}{2}\int_{0.75}^{0.9}(1-x^2)dx = \frac{3}{2}\left(\int_{0.75}^{0.9} 1dx - \int_{0.75}^{0.9} x^2 dx\right) =$$

$$= \frac{3}{2}\left([x]_{0.75}^{0.9} - \left[\frac{x^3}{3}\right]_{0.75}^{0.9}\right) = \frac{3}{2}\Big((0.9-0.75) - \frac{1}{3}((0.9)^3 - (0.75)^3)\Big) \approx$$

$$= \frac{3}{2}(0.15 - 0.1) = 1.5 \cdot 0.05 = 0.075$$

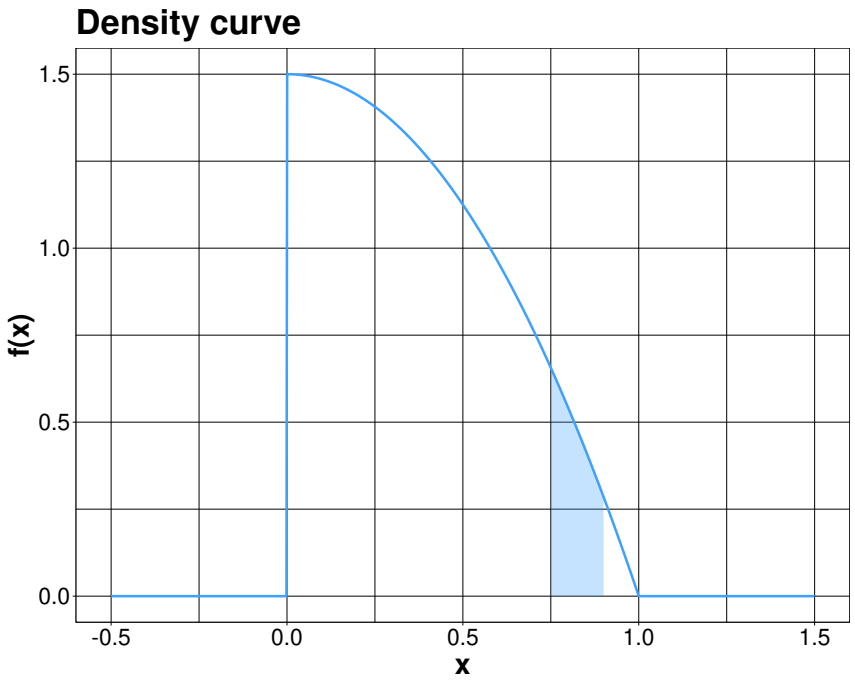


Figure 26: Plot of the density curve $f(x)$. The shaded area is $P(0.75 \leq X \leq 0.9)$.

□

Notice the difference between discrete and continuous variables: the pmf of discrete variables immediately gives probabilities, while the pdf of continuous variables has to be integrated. Also, as anticipated, the probability of a continuous random variable X to take on a specific value $c \in \mathbb{R}$ is always 0[3]:

$$P(X = c) = P(c \leq X \leq c) = \int_c^c f(x)dx = \lim_{\varepsilon \to 0}\int_{c-\varepsilon}^{c+\varepsilon} f(x)dx = 0$$

This also implies that taking into account the extremes of an interval when computing probabilities makes no difference:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

The **cumulative distribution function** (**cdf** for short) of a continuous random variable X is defined as the probability of such random variable to take a value less than or equal to a threshold. The cdf of X is denoted as $F(X)$, and is formally defined as:

[3]This is sound even from a logical perspective. The chance of picking a specific value in (a subset of) $\mathbb{R}$ is necessarely infinitesimal, since $\mathbb{R}$ is not a countable set.

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

The cdf for continuous random variables is the continuous counterpart to the cdf for discrete random variables, replacing a summation with an integral. This implies that to obtain the pdf from the cdf of a random variable, it suffices to take the derivative of the cdf.

Exercise 3.4.2: What is the cdf of the variable X in Exercise 3.4.1?

Solution: The cdf of X is clearly 0 for $x < 0$:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} 0dt = 0$$

For $x \in [0, 1]$:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} f(t)dt + \int_{0}^{x} f(t)dt = \int_{-\infty}^{0} 0dt + \int_{0}^{x} \frac{3}{2}(1-t^2)dt = \\ &= 0 + \frac{3}{2}\left(\int_{0}^{x} 1dt - \int_{0}^{x} t^2 dt\right) = \frac{3}{2}\left[(x-0) - \left(\frac{x^3}{3} - \frac{0^3}{3}\right)\right] = \\ &= \frac{3}{2}\left(x - \frac{x^3}{3}\right) = \frac{1}{2}x(3-x^2) \end{aligned}$$

For $x > 1$:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} f(t)dt + \int_{0}^{1} f(t)dt + \int_{1}^{x} f(t)dt = 0 + 1 + 0 = 1$$

This gives:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2}x(3-x^2) & \text{if } 0 \leq x \leq 1 \\ 1 & \text{otherwise} \end{cases}$$

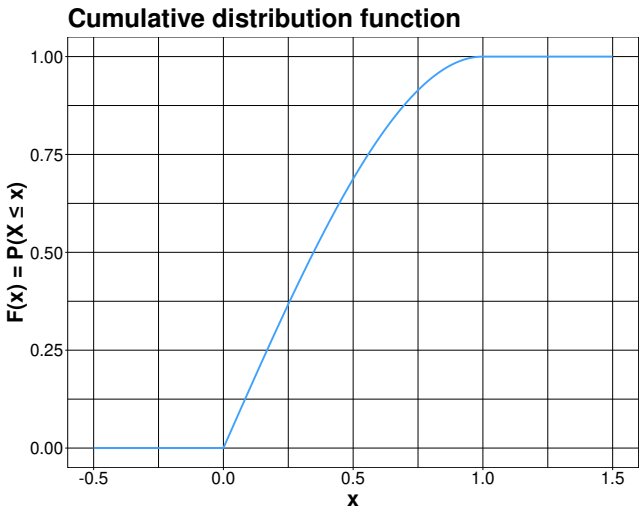


Figure 27: Plot of the cdf $F(x)$.

□

It's possible to connect the probabilities computed from the cdf of a continuous random variable with the probabilities computed from the pdf in the following way. For a given sample space $\mathcal{S}$, let $X : \mathcal{S} \mapsto \mathbb{R}$ be a continuous random variable with known cdf $F(x)$. For any $a, b \in \mathbb{R}$:

$$P(a \leq X \leq b) = \int_a^b f(x)dx = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx = F(b) - F(a)$$

It's also easy to show that:

$$P(X > x) = \int_x^{+\infty} f(t)dt = \int_{-\infty}^{+\infty} f(t)dt - \int_{-\infty}^x f(t)dt = 1 - F(x)$$

Let p be a real number between 0 and 1. The **(100p)-th percentile** of the distribution of a continuous random variable X, denoted as $\eta(p)$ is defined by:

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(t)dt$$

That is, the cdf of X evaluated at $\eta(p)$. In particular, the 50-th percentile is also called the **median**, denoted $\tilde{\mu}_X$ (or just $\tilde{\mu}$ if there's no ambiguity).

Exercise 3.4.3: What is the median of the variable X in Exercise 3.4.1?

Solution: The median is given by setting $p = 0.5$. Applying the formula gives:

$$0.5 = F(\eta(p)) \Rightarrow \frac{1}{2} = \frac{1}{2}\eta(p)(3 - (\eta(p))^2) \Rightarrow (\eta(p))^3 - 3\eta(p) + 1 = 0$$

The possible (approximated) solutions[4] are 0.3473, -1.8794 and 1.5321. The first solution is the only acceptable one, since the other two fall outside of $[0, 1]$. □

Let X be a continuous random variable and probability density function $f(X)$. The **expected value** (or **mean value**) of X, denoted as $E(X)$ or μ_X, is the continuous counterpart to the expected value of discrete random variables:

$$E(X) = \mu_X = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

When the variable X is known from context, the pedix X in μ_X is omitted.

Exercise 3.4.4: What is the expected value of the variable X in Exercise 3.4.1?

Solution:

[4] Computed with R issuing `polyroot(c(1, -3, 0, 1))`

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \int_{-\infty}^{0} x \cdot f(x)dx + \int_{0}^{1} x \cdot f(x)dx + \int_{1}^{+\infty} x \cdot f(x)dx =$$
$$= 0 + \int_{0}^{1} x \cdot \frac{3}{2}(1-x^2)dx + 0 = \frac{3}{2}\int_{0}^{1} x - x^3 dx = \frac{3}{2}\left(\int_{0}^{1} xdx - \int_{0}^{1} x^3 dx\right) =$$
$$= \frac{3}{2}\left(\left(\frac{1^2}{2} - \frac{0^2}{2}\right) - \left(\frac{1^4}{4} - \frac{0^4}{4}\right)\right) = \frac{3}{2}\left(\frac{1}{2} - \frac{1}{4}\right) = \frac{3}{8} \approx 0.375$$

□

The expected value is oblivious to whether its argument is a random variable or a function whose input is a random variable. In other words, let X be a continuous random variable and pdf $p(X)$, and let $h(X)$ be a function whose argument is X itself. The expected value of $h(X)$ is still defined as:

$$E(h(X)) = \mu_{h(X)} = \int_{-\infty}^{+\infty} h(x) \cdot f(x)dx$$

Theorem 3.4.1: Let X be a continuous random variable with pdf $f(x)$. For any $a, b \in \mathbb{R}$:

$$E(aX + b) = aE(X) + b$$

Proof: Let $h(x) = aX + b$. Then:

$$E(h(x)) = E(aX + b) = \int_{-\infty}^{+\infty} (ax + b) \cdot f(x)dx = \int_{-\infty}^{+\infty} ax \cdot f(x) + b \cdot f(x)dx =$$
$$= a\int_{-\infty}^{+\infty} x \cdot f(x)dx + b\int_{-\infty}^{+\infty} f(x)dx = aE(X) + b \cdot 1 = aE(X) + b$$

Since $\int_{-\infty}^{+\infty} f(x)dx = 1$ by definition. □

Let X be a continuous random variable with probability density function $f(x)$. The **variance** of X, denoted as $V(X)$ or σ_X^2, is a measure of how the values taken by the variable differ from the expected value:

$$V(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f(x)dx = E((X - E(X))^2)$$

When the variable X is known from context, the pedix X in σ_X^2 is omitted.

Exercise 3.4.5: What is the variance of the variable X in Exercise 3.4.1?

Solution:

$$
\begin{aligned}
V(X) &= \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f(x)dx = \\
&= \int_{-\infty}^{0} \left(x - \frac{3}{8}\right)^2 \cdot 0dx + \int_{0}^{1} \left(x - \frac{3}{8}\right)^2 \cdot \frac{3}{2}(1 - x^2)dx + \int_{1}^{+\infty} \left(x - \frac{3}{8}\right)^2 \cdot 0dx = \\
&= 0 + \frac{3}{2}\int_{0}^{1} \left(x - \frac{3}{8}\right)^2 (1 - x^2)dx + 0 = \frac{3}{2}\int_{0}^{1} \left(x^2 + \frac{9}{64} - \frac{3}{4}x\right)(1 - x^2)dx = \\
&= \frac{3}{2}\int_{0}^{1} x^2 + \frac{9}{64} - \frac{3}{4}x - x^4 - \frac{9}{64}x^2 + \frac{3}{4}x^3 dx = \\
&= \frac{3}{2}\left(\frac{9}{64}\int_{0}^{1} dx - \frac{3}{4}\int_{0}^{1} xdx - \int_{0}^{1} x^4 dx + \frac{55}{64}\int_{0}^{1} x^2 dx + \frac{3}{4}\int_{0}^{1} x^3 dx\right) = \\
&= \frac{3}{2}\left(\frac{9}{64} \cdot 1 - \frac{3}{4} \cdot \frac{1}{2} - \frac{1}{5} + \frac{55}{64} \cdot \frac{1}{3} + \frac{3}{4} \cdot \frac{1}{4}\right) = \frac{3}{2} \cdot \frac{19}{480} = \frac{19}{320} \approx 0.059
\end{aligned}
$$

□

The square root of the variance is called the **standard deviation**:

$$SD(X) = \sigma_X = \sqrt{V(X)}$$

Lemma 3.4.1 gives a simpler way to compute the variance of a continuous random variable.

Lemma 3.4.1: Let X be a continuous random variable with probability density function $f(x)$. The following equality holds:

$$V(X) = E(X^2) - (E(X))^2$$

Proof: Expanding the square in the formula:

$$
\begin{aligned}
V(X) &= \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f(x)dx = \int_{-\infty}^{+\infty} (x^2 + (E(X))^2 - 2xE(X)) \cdot f(x)dx = \\
&= \int_{-\infty}^{+\infty} x^2 \cdot f(x) + (E(X))^2 \cdot f(x) - 2xE(X) \cdot f(x)dx = \\
&= \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx + \int_{-\infty}^{+\infty} (E(X))^2 \cdot f(x)dx - \int_{-\infty}^{+\infty} 2xE(X) \cdot f(x)dx = \\
&= E(X^2) + (E(X))^2 \int_{-\infty}^{+\infty} f(x)dx - 2E(X)\int_{-\infty}^{+\infty} x \cdot f(x)dx = \\
&= E(X^2) + (E(X))^2 \cdot 1 - 2E(X) \cdot E(X) = \\
&= E(X^2) + (E(X))^2 - 2(E(X))^2 = E(X^2) - (E(X))^2
\end{aligned}
$$

□

Theorem 3.4.2: Let X be a continuous random variable and pdf $f(x)$. For any $a, b \in \mathbb{R}$:

$$V(aX + b) = a^2 V(X)$$

Proof: Applying Theorem 3.4.1:

$$V(aX+b) = E(((aX+b) - E(aX+b))^2) = E((aX + \not{b} - aE(X) - \not{b})^2) =$$
$$-E((a(X - E(X)))^2) = E(a^2(X - E(X))^2) = a^2 E((X - E(X))^2) =$$
$$= a^2 V(X)$$

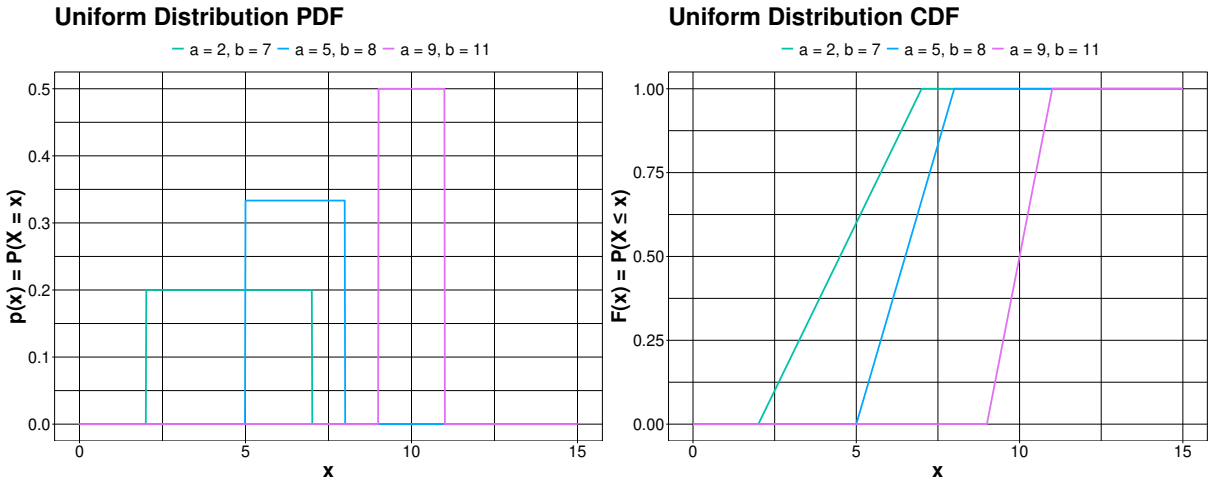
□

3.5. Known continuous random variables

Some specific continuous random variables have been studied extensively, mostly because they model very well many phenomena in the real world. For this reason, such random variables have proper names. To denote that a random variable X has the same distribution as a known random variable F, the notation $X \sim F$ is used.

3.5.1. Uniform random variable

A continuous random variable X is distributed as a **uniform random variable** of parameters a and b (denoted as $X \sim U(a, b)$) if the pdf and cdf of said variable are:



$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

Indeed, the relation between the two holds:

$$\int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{a} f(t)dt + \int_{a}^{x} f(t)dt = 0 + \int_{a}^{x} \frac{1}{b-a}dt =$$
$$= \frac{1}{b-a}\int_{a}^{x} 1dt = \frac{1}{b-a}(x-a) = \frac{x-a}{b-a}$$

Theorem 3.5.1.1: The expected value and variance of a random variable $X \sim U(a,b)$ are as follows:

$$E(X) = \frac{b+a}{2} \qquad\qquad V(X) = \frac{(b-a)^2}{12}$$

Proof: The expected value is given by:

$$\begin{aligned}
E(X) &= \int_{-\infty}^{+\infty} x \cdot f(x)dx = \int_{-\infty}^{a} x \cdot f(x)dx + \int_{a}^{b} x \cdot f(x)dx + \int_{b}^{+\infty} x \cdot f(x)dx = \\
&= 0 + \int_{a}^{b} x\left(\frac{1}{b-a}\right)dx + 0 = \frac{1}{b-a}\int_{a}^{b} xdx = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \\
&= \frac{1}{\cancel{(b-a)}}\frac{\cancel{(b-a)}(b+a)}{2} = \frac{b+a}{2}
\end{aligned}$$

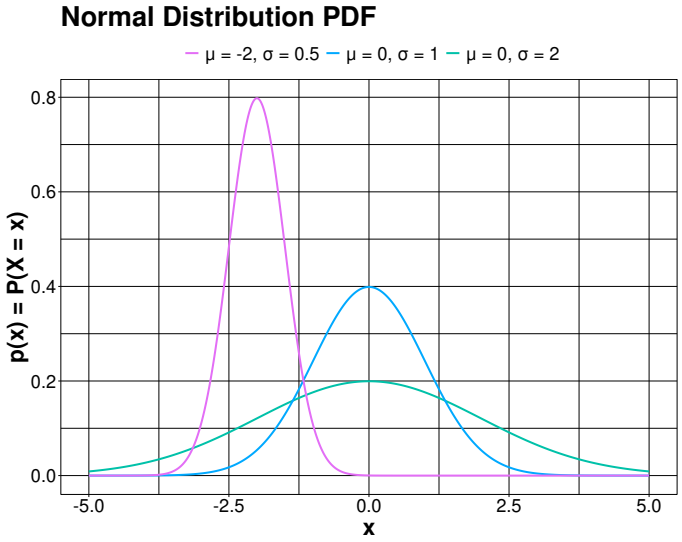
The variance is given by:

$$\begin{aligned}
V(X) &= \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f(x)dx = 0 + \left(\int_{a}^{b} (x - E(X))^2 \cdot f(x)dx\right) + 0 = \\
&= \int_{a}^{b} \left(x - \frac{b+a}{2}\right)^2 \left(\frac{1}{b-a}\right)dx = \frac{1}{b-a}\int_{a}^{b} x^2 + \frac{(b+a)^2}{4} - (b+a)xdx = \\
&= \frac{1}{b-a}\left(\int_{a}^{b} x^2dx + \int_{a}^{b} \frac{(b+a)^2}{4}dx - \int_{a}^{b} (b+a)xdx\right) = \\
&= \frac{1}{b-a}\int_{a}^{b} x^2dx + \frac{(b+a)^2}{4(b-a)}\int_{a}^{b} 1dx - \frac{b+a}{b-a}\int_{a}^{b} xdx = \\
&= \frac{1}{b-a}\left(\frac{b^3}{3} - \frac{a^3}{3}\right) + \frac{(b+a)^2}{\cancel{4(b-a)}}\cancel{(b-a)} - \frac{b+a}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \\
&= \frac{1}{\cancel{(b-a)}}\frac{\cancel{(b-a)}(b^2+ba+a^2)}{3} + \frac{(b+a)^2}{4} - \frac{b+a}{\cancel{(b-a)}}\frac{\cancel{(b-a)}(b+a)}{2} = \\
&= \frac{b^2+ba+a^2}{3} + \frac{a^2+b^2+2ab}{4} - \frac{a^2+b^2+2ab}{2} = \\
&= \frac{4b^2+4ba+4a^2+3a^2+3b^2+6ab-6a^2-6b^2-12ab}{12} = \\
&= \frac{a^2+b^2-2ab}{12} = \frac{(b-a)^2}{12}
\end{aligned}$$

□

3.5.2. Normal random variable

A continuous random variable X is distributed as a **normal random variable** (or **Gaussian random variable**) of parameters μ and σ with $-\infty < \mu < +\infty$ and $\sigma > 0$ (denoted as $X \sim N(\mu, \sigma)$) if the pdf of said variable is:



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

Where μ and σ are respectively the expected value and the standard deviation of the random variable. The normal random variable is the most important distribution in statistics, since it models many real world phenomena (IQ score, anthropometric measures, economic indicators, ecc...).

The distribution $Z \sim N(0,1)$, that is to say the normal distribution of parameters $\mu = 0$ and $\sigma = 1$, is called the **standard normal distribution**, having pdf:

$$f(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$$

Even though it is possible to compute the cdf of a normal random variable X by integrating its pdf, said computation is very hard since it's not approachable with standard integration techniques. Despite this, the values of the cdf of the standard normal distribution for most common values are known and provided in table form. The cdf of $Z \sim N(0,1)$ evaluated at z is denoted as $\Phi(z)$.

The cdf of the standard normal distribution is sufficient to also compute the cdf for any given random variable $X \sim N(\mu, \sigma)$. Infact, the **standardized** version of any X, which is given by subtracting the expected value of X from X and dividing the result by its standard deviation, is a standard normal distribution:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Infact, subtracting $E(X) = \mu$ from X sets its new expected value to 0 and dividing the result by $SD(X) = \sigma$ sets its new variance to 1:

$$E(Z) = E\left(\frac{X - E(X)}{SD(X)}\right) = \frac{1}{SD(X)} E(X - E(X)) = \frac{1}{SD(X)}(E(X) - E(X)) = \frac{0}{SD(X)} = 0$$

$$SD(Z) = V\left(\frac{X - E(X)}{SD(X)}\right) = \frac{1}{(SD(X))^2} V(X - E(X)) = \frac{V(X) - 0}{(SD(X))^2} = \frac{V(X)}{V(X)} = 1$$

Since the values of the cdf of Z are known, it is the possible to compute the values for the cdf (and pdf) of X by computing the ones for $(X - \mu)/\sigma$:

$$P(a < X \leq b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$F(a) = P(X \leq a) = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = P\left(Z \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Exercise 3.5.2.1: Suppose that the time a car driver takes to react to the brake lights on a decelerating vehicle can be modeled as a normal distribution of parameters $\mu = 1.25$ seconds and $\sigma = 0.46$ seconds. What is the probability that the reaction time lies between 1 second and 1.75 seconds? And the probability of it being greater than 2 seconds?

Solution:

$$P(1 < X \leq 1.75) = P\left(\frac{1-1.25}{0.46} < \frac{X-1.25}{0.46} \leq \frac{1.75-1.25}{0.46}\right) = P\left(\frac{-0.25}{0.46} < Z \leq \frac{0.5}{0.46}\right) \approx$$

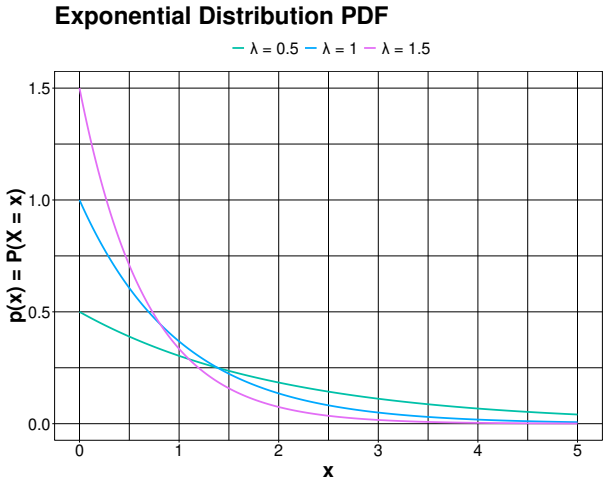
$$\Phi(1.09) - \Phi(-0.54) = \Phi(1.09) - (1 - \Phi(0.54)) = 0.8621 - 0.2946 = 0.5675$$

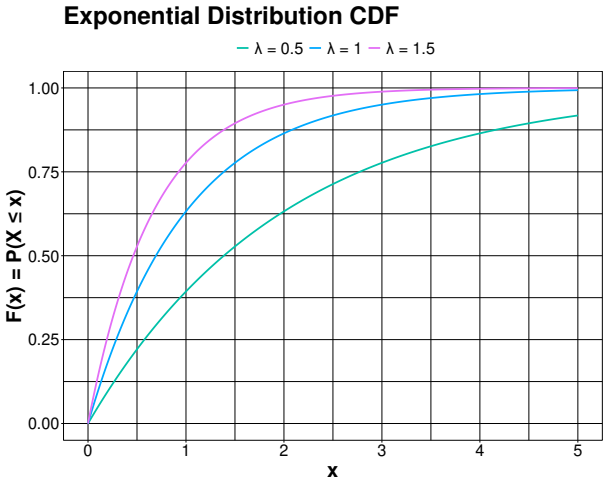
$$P(X > 2) = 1 - P(X \leq 2) = 1 - P\left(\frac{X-1.25}{0.46} \leq \frac{2-1.25}{0.46}\right) \approx 1 - \Phi(1.63) = 1 - 0.9484 = 0.0516$$

□

3.5.3. Exponential random variable

A continuous random variable X is distributed as an **exponential random variable** of parameter λ with $\lambda > 0$ (denoted as $X \sim E(\lambda)$) if the pdf and cdf of said variable are:





$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Indeed, the relation between the two holds:

$$\int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} f(t)dt + \int_{0}^{x} f(t)dt = 0 + \int_{0}^{x} \lambda e^{-\lambda t}dt =$$
$$= \int_{0}^{-\lambda x} \cancel{\lambda}\frac{-1}{\cancel{\lambda}} e^{u}du = -\left(e^{-\lambda x} - e^{0}\right) = 1 - e^{-\lambda x}$$

Theorem 3.5.3.1: The expected value and variance of a random variable $X \sim E(\lambda)$ are as follows:

$$E(X) = \frac{1}{\lambda} \qquad\qquad V(X) = \frac{1}{\lambda^2}$$

Proof: The expected value is given by:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \int_{-\infty}^{0} x \cdot f(x)dx + \int_{0}^{+\infty} x \cdot f(x)dx =$$
$$= 0 + \int_{0}^{+\infty} \lambda x e^{-\lambda x}dx = \int_{0}^{+\infty} x\lambda e^{-\lambda x}dx$$

Applying integration by parts:

$$E(X) = \int_{0}^{+\infty} x\lambda e^{-\lambda x}dx = -\int_{0}^{+\infty} x\left(-\lambda e^{-\lambda x}\right)dx = -\left[xe^{-\lambda x}\right]_{0}^{\infty} + \int_{0}^{+\infty} e^{-\lambda x}dx =$$
$$= -\left(\infty \cdot e^{-\lambda \cdot \infty} - 0 \cdot e^{-\lambda \cdot 0}\right) + \int_{0}^{+\infty} e^{-\lambda x}dx = -0 + \int_{0}^{+\infty} e^{-\lambda x}dx =$$
$$= \int_{0}^{+\infty} e^{-\lambda x}dx$$

Solving by substitution:

$$E(X) = \int_{0}^{+\infty} e^{-\lambda x}dx = \int_{0}^{-\infty} \frac{e^{u}}{-\lambda}du = \frac{1}{\lambda}\int_{-\infty}^{0} e^{u}du = \frac{1}{\lambda}(e^{0} - e^{-\infty}) = \frac{1-0}{\lambda} = \frac{1}{\lambda}$$

As for the variance:

$$V(X) = E(X^2) - (E(X))^2 = -\left(\frac{1}{\lambda}\right)^2 + \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx =$$
$$= -\frac{1}{\lambda^2} + \int_{-\infty}^{0} x^2 \cdot f(x)dx + \int_{0}^{+\infty} x^2 \cdot f(x)dx =$$
$$= -\frac{1}{\lambda^2} + 0 + \int_{0}^{+\infty} \lambda x^2 e^{-\lambda x}dx = -\frac{1}{\lambda^2} + \int_{0}^{+\infty} \lambda x^2 e^{-\lambda x}dx$$

Applying integration by parts:

$$V(X) = -\frac{1}{\lambda^2} + \int_0^{+\infty} \lambda x^2 e^{-\lambda x} dx = -\frac{1}{\lambda^2} - \int_0^{+\infty} x^2 \big(-\lambda e^{-\lambda x}\big) dx =$$
$$= -\frac{1}{\lambda^2} - \big[x^2 e^{-\lambda x}\big]_0^{\infty} + \int_0^{+\infty} 2x e^{-\lambda x} dx =$$
$$= -\frac{1}{\lambda^2} - \big(\infty^2 \cdot e^{-\lambda \cdot \infty^2} - 0^2 \cdot e^{-\lambda \cdot 0^2}\big) - 2\int_0^{+\infty} x e^{-\lambda x} dx =$$
$$= -\frac{1}{\lambda^2} - 0 - 2\int_0^{+\infty} x e^{-\lambda x} dx = -\frac{1}{\lambda^2} - 2\int_0^{+\infty} x e^{-\lambda x} dx$$

Applying integration by parts (again):

$$V(X) = -\frac{1}{\lambda^2} - 2\int_0^{+\infty} x e^{-\lambda x} dx = -\frac{1}{\lambda^2} - \frac{2}{\lambda}\int_0^{+\infty} x\big(-\lambda e^{-\lambda x}\big) dx =$$
$$= -\frac{1}{\lambda^2} - \frac{2}{\lambda}\big[x e^{-\lambda x}\big]_0^{\infty} + \frac{2}{\lambda}\int_0^{+\infty} e^{-\lambda x} dx =$$
$$= -\frac{1}{\lambda^2} - \frac{2}{\lambda}\big(\infty \cdot e^{-\lambda \cdot \infty} - 0 \cdot e^{-\lambda \cdot 0}\big) + \frac{2}{\lambda}\int_0^{+\infty} e^{-\lambda x} dx =$$
$$= -\frac{1}{\lambda^2} - \frac{2}{\lambda} \cdot 0 + \frac{2}{\lambda}\int_0^{+\infty} e^{-\lambda x} dx = -\frac{1}{\lambda^2} + \frac{2}{\lambda}\int_0^{+\infty} e^{-\lambda x} dx$$

Solving by substitution:

$$V(X) = -\frac{1}{\lambda^2} + \frac{2}{\lambda}\int_0^{+\infty} e^{-\lambda x} dx = -\frac{1}{\lambda^2} + \frac{2}{\lambda}\int_0^{-\infty} \frac{e^u}{-\lambda} du = -\frac{1}{\lambda^2} - \frac{2}{\lambda^2}\int_0^{-\infty} e^u du =$$
$$= -\frac{1}{\lambda^2} + \frac{2}{\lambda^2}\int_{-\infty}^{0} e^u du = -\frac{1}{\lambda^2} + \frac{2}{\lambda^2}(e^0 - e^{-\infty}) = -\frac{1}{\lambda^2} + \frac{2}{\lambda^2} = \frac{-1+2}{\lambda^2} = \frac{1}{\lambda^2}$$

□

Exercise 3.5.3.1: Suppose that the stress range of a certain bridge connection (measured in Megapascal) can be modeled as an exponential distribution X with expected value equal to 6. What is the probability of the stress to be less than or equal to 10 Megapascal?

Solution: If the expected value of X is 6, since the expected value of an exponential distribution is $1/\lambda$ the lambda parameter of X is $\lambda = 0.1667$.

$$P(X \leq 10) = F(10) = 1 - e^{-0.1667 \cdot 10} = 1 - e^{-1.667} \approx 1 - 0.189 = 0.811$$

□

The exponential distribution and the Poisson distribution are closely related. Indeed, suppose that the number of events occurring in any time interval of length Δt has a Poisson distribution with parameter $\alpha \Delta t$ (where alpha, the rate of the event process, is the expected number of events occurring in 1 unit of time) and that number of occurrences in nonoverlapping intervals are independent of one another. Then the distribution of elapsed time between the occurrence of two successive events is exponential with parameter $\lambda = \alpha$.

Theorem 3.5.3.2: The geometric distribution function has the memorylessness property.

Proof: Let $X \sim E(\lambda)$. Then:

$$P(X \geq t + t_0 \mid X \geq t_0) = \frac{P[(X \geq t + t_0) \cap (X \geq t_0)]}{P(X \geq t_0)} = \frac{P(X \geq t + t_0)}{P(X \geq t_0)} =$$
$$= \frac{1 - F(t + t_0)}{1 - F(t)} = e^{-\lambda t} = P(X \geq t)$$

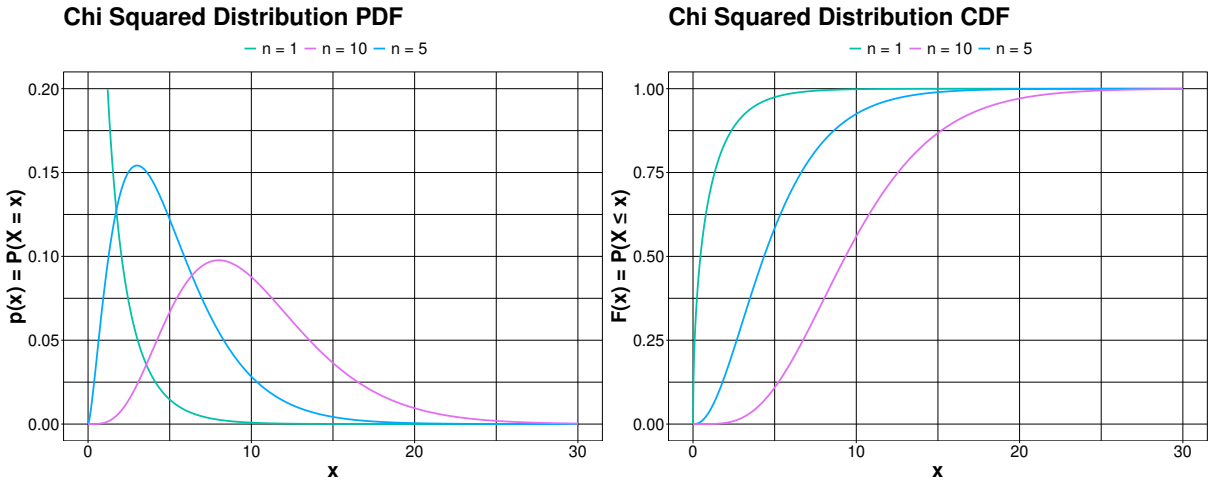
□

3.5.4. Chi-squared random variable

Given an integer n, let X be a random variable constructed by summing the squares of n independent standard normal random variables:

$$X = \sum_{i=1}^{n} Z_i^2 = Z_1^2 + Z_2^2 + ... + Z_n^2$$

The random variable X defined as such is called a **Chi-squared random variable** with n degrees of freedom (denoted as $X \sim \chi^2(n)$). Being the result of a sum of squared values, a Chi-squared random variable is always positive.



Theorem 3.5.4.1: The expected value and variance of a random variable $X \sim \chi^2(n)$ are as follows:

$$E(X) = n \qquad V(X) = 2n$$

Proof: The expected value of X can be computed by applying Theorem 3.6.1 and Lemma 3.2.1:

$$E(X) = E(Z_1^2 + Z_2^2 + ... + Z_n^2) = \sum_{i=1}^{n} E(Z_i^2) = \sum_{i=1}^{n} V(Z_i) + (E(Z_i))^2 = n(1 + 0^2) = n$$

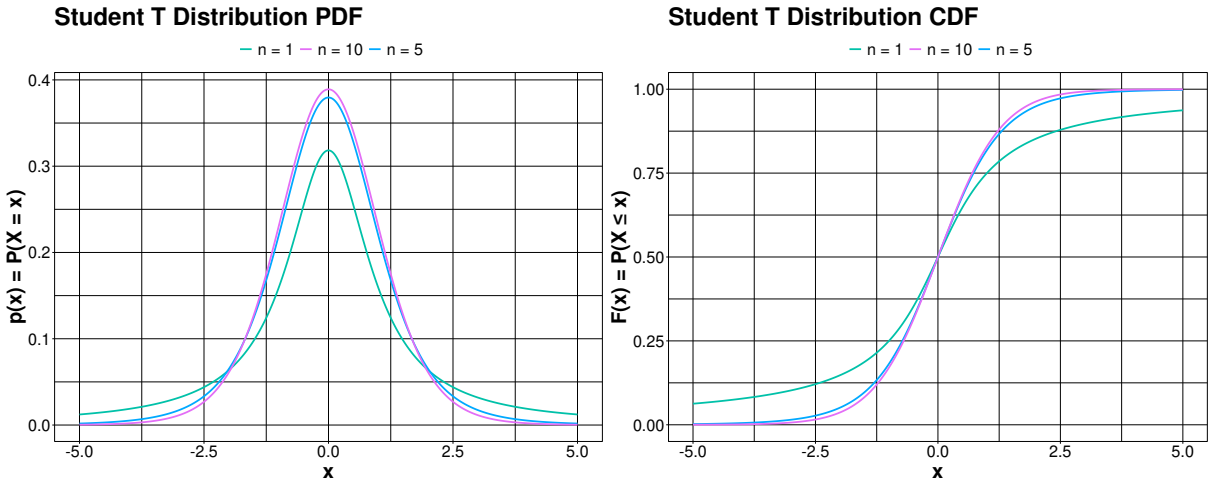
□

3.5.5. Student t random variable

Given an integer n, let Z be a standard normal distribution and Y_n a Chi-squared distribution with n degrees of freedom, independent of each other. Let X be a random variable constructed by computing the ratio of Z and the square root of Y_n:

$$X = \frac{Z}{\sqrt{Y_n}}$$

The random variable X defined as such is called a **Student t random variable** with n degrees of freedom (denoted as $X \sim T(n)$).



Theorem 3.5.5.1: The expected value and variance of a random variable $X \sim T(n)$ are as follows:

$$E(X) = \begin{cases} 0 & \text{if } n > 1 \\ \text{undefined} & \text{otherwise} \end{cases} \qquad V(X) = \begin{cases} \dfrac{n}{n-1} & \text{if } n > 2 \\ \text{undefined} & \text{otherwise} \end{cases}$$

As for the (standard) normal distribution, the values of its quantiles have been tabulated, since their calculations is generally unfeasible to be performed by hand.

The degrees of freedom of a Student t random variable are related to how "heavy" the tails of its pdf are. In particular, with n approaching ∞, the Student t distribution effectively becomes a normal distribution.

3.6. Joint probability distributions

Observing a single attribute is often too restrictive to analyze a problem. In general, a complex scenario is the result of an interplay between many phenomena, that ought to be observed simultaneously. In the realm of probability, this is equivalent to observing more than one random variable at once.

Let X and Y be two discrete random variables defined on the same sample space Ω of an experiment. The **joint probability mass function** $p(x, y)$ (joint pmf, for short) is defined for each couple (x, y) as:

$$p(x, y) = P(\{X = x\} \wedge \{Y = y\})$$

That is, the probability that x is the realization of X and at the same time that y is the realization of Y. Of course, just as the pmf of a discrete random variable:

$$p(x,y) \geq 0 \qquad \sum_{x \in D(X)} \sum_{y \in D(Y)} p(x,y) = 1$$

Let A be a set consisting of pairs of (x,y) values. The probability $P((X,Y) \in A)$ that the random pair (X,Y) lies in A is obtained by summing the joint pmf over pairs in A:

$$P((X,Y) \in A) = \sum_{(X,Y) \in A} \sum p(x,y)$$

The **marginal probability mass functions** of X and Y, denoted respectively by $p_X(x)$ and $p_Y(y)$, are given by:

$$p_X(x) = \sum_{y:p(x,y)>0} p(x,y) \forall x \in D(X) \qquad p_Y(y) = \sum_{x:p(x,y)>0} p(x,y) \forall y \in D(Y)$$

That is, the sum of all values of a probabilities "locking" one variable and "moving" along the other

Exercise 3.6.1: Suppose that a particular company offers insurance policies for home and automobile in such a way that there exist different possible detractions for both policies:

- For automobile insurance: 100€, 500€, 1000€;
- For home insurance: 500€, 1000€, 5000€.

Consider randomly selecting a customer having both home and automobile insurance. Let X be the amount of automobile insurance policy deductible and let Y be the amount of home insurance policy deductible. Consider the following joint pmf for said variables:

$p(x,y)$	500	1000	5000
100	0.3	0.05	0
500	0.15	0.2	0.05
1000	0.1	0.1	0.05

Is is well defined? What are the marginal probability mass functions of X and Y? What is the probability of X being greater or equal than 500? What is the probability of X and Y being equal?

Solution: The joint pmf is indeed well defined, because each probability is greater or equal than 0 and:

$$\sum_{x \in D(X)} \sum_{y \in D(Y)} p(x,y) = 0.3 + 0.05 + 0 + 0.15 + 0.2 + 0.05 + 0.1 + 0.1 + 0.05 = 1$$

The marginal probability mass function of X is given by:

$$\sum_{y \in D(Y)} p(X = 100, Y) = P(X = 100, Y = 500) + P(X = 100, Y = 1000) + P(X = 100, Y = 5000) = 0.3 + 0.05 + 0 = 0.35$$

$$\sum_{y \in D(Y)} p(X = 500, Y) = P(X = 500, Y = 500) + P(X = 500, Y = 1000) + P(X = 500, Y = 5000) = 0.15 + 0.2 + 0.05 = 0.4$$

$$\sum_{\in D(Y)} p(X = 1000, Y) = P(X = 1000, Y = 500) + P(X = 1000, Y = 1000) + P(X = 1000, Y = 5000) = 0.1 + 0.1 + 0.05 = 0.2$$

$$\sum_{x \in D(X)} p(X, Y = 500) = P(X = 100, Y = 500) + P(X = 500, Y = 500) + P(X = 1000, Y = 500) = 0.3 + 0.15 + 0.1 = 0.55$$

$$\sum_{\in D(X)} p(X, Y = 1000) = P(X = 100, Y = 1000) + P(X = 500, Y = 1000) + P(X = 1000, Y = 1000) = 0.05 + 0.2 + 0.1 = 0.35$$

$$\sum_{x \in D(X)} p(X, Y = 5000) = P(X = 100, Y = 5000) + P(X = 500, Y = 5000) + P(X = 1000, Y = 5000) = 0 + 0.05 + 0.05 = 0.1$$

$$p_X(x) = \begin{cases} 0.35 \text{ if } x = 100 \wedge (y = 500 \vee y = 1000 \vee y = 5000) \\ 0.4 \quad \text{if } x = 500 \wedge (y = 500 \vee y = 1000 \vee y = 5000) \\ 0.25 \text{ if } x = 1000 \wedge (y = 500 \vee y = 1000 \vee y = 5000) \end{cases}$$

$$p_Y(y) = \begin{cases} 0.55 \text{ if } y = 500 \wedge (x = 100 \vee x = 500 \vee x = 1000) \\ 0.35 \text{ if } y = 1000 \wedge (x = 100 \vee x = 500 \vee x = 1000) \\ 0.1 \quad \text{if } y = 5000 \wedge (x = 100 \vee x = 500 \vee x = 1000) \end{cases}$$

The probability of X being greater than 500 can be retrieved by focusing on the values of X and ignoring the ones of Y:

$$P(X \geq 500) = 0.15 + 0.2 + 0.05 + 0.1 + 0.1 + 0.05 = 0.65$$

The only values that X and Y can both assume are 500 and 1000. Therefore:

$$P(X = Y) = P(\{X = 500 \wedge Y = 500\} \cup \{X = 1000 \wedge Y = 1000\}) =$$
$$P(X = 500 \wedge Y = 500) + P(X = 1000 \wedge Y = 1000) = 0.15 + 0.1 = 0.25$$

□

In the more general case of having n discrete random variables $X_1, X_2, ..., X_n$, the joint pmf of said variables is given by the function:

$$p(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

Let X and Y be two continuous random variables defined on the same sample space Ω of an experiment. The **joint probability density function** $f(x, y)$ (joint pmf, for short) is a function such that, for any two-dimensional set A:

$$P((X, Y) \in A) = \int_A \int f(x, y) dx dy$$

Of course, just as the pdf of a continuous random variable:

$$f(x, y) \geq 0 \qquad \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$$

In particular, if A is the two-dimensional rectangle $\{(x, y) : a \leq x \leq b, c \leq y \leq d\}$, then:

$$P((X,Y) \in A) = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)dxdy$$

The **marginal probability density functions** of X and Y, denoted respectively by $f_X(x)$ and $f_Y(y)$, are given by:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy \text{ for } -\infty < x < +\infty \qquad f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx \text{ for } -\infty < y < +\infty$$

In the more general case of having n continuous random variables $X_1, X_2, ..., X_n$, the joint pdf of said variables is the function $f(x_1, x_2, ..., x_n)$ such that for any n intervals $[a_1, b_1], ..., [a_n, b_n]$:

$$P(a_1 \leq X_1 \leq b_1, ..., a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} f(x_1, ..., x_n)dx_1, ..., dx_n$$

The notion of dependence and independence of two random variables can be phrased in the language of joint probability mass/density functions. Two random variables X and Y are said to be independent if, for any pair (x, y) with x being a realization of X and y being a realization of Y, the following holds:

$$p(x,y) = p_X(x) \cdot p_Y(y) \text{ with } X, Y \text{ discrete} \qquad f(x,y) = f_X(x) \cdot f_Y(y) \text{ with } X, Y \text{ continuous}$$

Otherwise, X and Y are dependent (not independent).

Exercise 3.6.2: Consider Exercise 3.6.1. Are X and Y dependent or independent?

Solution: X and Y are not independent, and it can be shown with a single counterexample. Consider $X = 1000$ and $Y = 5000$:

$$P(X = 1000 \wedge Y = 5000) = 0.05 \qquad P(X = 1000) \cdot P(Y = 5000) = 0.25 \cdot 0.1 = 0.025$$

Since the two values differ, X and Y ought to be dependent. □

Moreover, joint pmf/pdf can be used to prove the linearity of expected value and variance.

Theorem 3.6.1: For any two random variables X and Y with supports $D(X)$ and $D(Y)$ respectively, $E[X + Y] = E[X] + E[Y]$.

Unlike the expected value, the variance is not always a linear function, but only when the variables at hand are independent.

Theorem 3.6.2: For any two independent random variables X and Y with supports $D(X)$ and $D(Y)$ respectively, $V[X + Y] = V[X] + V[Y]$.

In the more general case of having n random variables $X_1, X_2, ..., X_n$, said variables are independent if for any subset $X_{i_1}, X_{i_2}, ..., X_{i_k}$ of size $k \in [2, n]$, the joint pmfs or pdfs is equal to the product of the marginal pmfs or pdfs.

Let X and Y be two discrete random variables with joint pmf $p(x, y)$ and marginal pmf of X $p_X(x)$. Then, for any X value x for which $p_X(x) > 0$, the **conditional probability mass function of Y given that X = x** is:

$$p_{Y \mid X}(y \mid x) = \frac{p(x, y)}{p_X(x)} \forall y \in D(Y)$$

Let X and Y be two continuous random variables with joint pdf $f(x, y)$ and marginal pdf of X $f_X(x)$. Then, for any X value x for which $f_X(x) > 0$, the **conditional probability density function of Y given that X = x** is:

$$f_{Y \mid X}(y \mid x) = \frac{f(x, y)}{f_X(x)} \text{ for } -\infty < y < +\infty$$

Let X and Y be two random variables having a joint pmf $p(x, y)$ or joint pdf $f(x, y)$ (according to whether they are discrete or continuous). Then the expected value of a function $h(X, Y)$, denoted as $E[h(X, Y)]$ or $\mu_{h(X,Y)}$, is given by:

$$E[h(X, Y)] = \begin{cases} \sum_{x \in D(X)} \sum_{y \in D(Y)} h(x, y) \cdot p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x, y) \cdot f(x, y) dx dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

The **covariance** between two random variables X and Y is given by:

$$\text{Cov}(X, Y) = \begin{cases} \sum_{x \in D(X)} \sum_{y \in D(Y)} (x - E(X))(y - E(Y)) p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E(X))(y - E(Y)) f(x, y) dx dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

The covariance is particularly useful for analyzing the linear similarity of two dependent random variables.

Exercise 3.6.3: Consider Exercise 3.6.1. Compute the covariance of X and Y.

Solution: First, the expected value of X and Y have to be computed:

$$E(X) = \sum_{x \in D(X)} x \cdot P(X = x) = 100 \cdot 0.35 + 500 \cdot 0.4 + 1000 \cdot 0.25 = 485$$

$$E(Y) = \sum_{y \in D(Y)} y \cdot P(Y = y) = 500 \cdot 0.55 + 1000 \cdot 0.35 + 5000 \cdot 0.1 = 1125$$

It is then possible to compute the covariance as:

$$\text{Cov}(X, Y) = \sum_{x \in D(X)} \sum_{y \in D(Y)} (x - E(X))(y - E(Y)) p(x, y) = (100 - 485)(500 - 1125)(0.3) +$$
$$(100 - 485)(1000 - 1125)(0.05) + (100 - 485)(5000 - 1125)(0) + (500 - 485)(500 - 1125)(0.15) +$$
$$(500 - 485)(1000 - 1125)(0.2) + (500 - 485)(5000 - 1125)(0.05) + (1000 - 485)(500 - 1125)(0.1) +$$
$$(1000 - 485)(1000 - 1125)(0.1) + (1000 - 485)(5000 - 1125)(0.05) = 136875$$

□

Theorem 3.6.3: Let X and Y be two random variables, and let a and b be two real numbers. The following equality holds:

$$\mathrm{Cov}(aX, bY) = ab\ \mathrm{Cov}(X, Y)$$

Theorem 3.6.4: Let X and Y be two random variables. The following equality holds:

$$\mathrm{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$$

Since the covariance has no minimum and no maximum, it isn't really indicative of the order of magnitude of the random variables.

A better measure of the relationship between two random variables X and Y is given by the **correlation coefficient**, denoted as $\mathrm{Corr}(X, Y)$ or $\rho_{X,Y}$ and given by:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{V(X)V(Y)}$$

Theorem 3.6.5: For any pair of random variables X and Y, $-1 \leq \mathrm{Corr}(X, Y) \leq 1$.

The correlation coefficient is more descriptive of the relationship between two random variables than their variance because, as stated in Theorem 3.6.5, it is bounded, and therefore is indicative of the scale of the values of the variables.

Theorem 3.6.6: Let X and Y be two random variables. If $\mathrm{Corr}(X, Y) = \pm 1$, there exist two real numbers a and b, with $a \neq 0$, such that $Y = aX + b$.

Theorem 3.6.6 can be generalized by stating that if $\mathrm{Corr}(X, Y)$ is close to ± 1, it means that the two random variables are almost linearly correlated, while if it is close to 0 it means that the two random variables are poorly linearly correlated, or not correlated at all.

Lemma 3.6.1: If X and Y are two independent random variables, $\mathrm{Corr}(X, Y) = 0$.

Theorem 3.6.7: Let X and Y be two random variables, and let a, b, c and d be four real numbers. If a and c have the same sign, the following equality holds:

$$\mathrm{Corr}(aX + b, cY + d) = \mathrm{Corr}(X, Y)$$

An example of a bivariate distribution is the **bivariate normal distribution**, an extension of the normal distribution in two dimensions:

$$f(x,y) = \frac{1}{\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right)$$

4. Inferential statistics

4.1. Random sampling

Statistics is a science concerned with deducing conclusions from experimental data. In most scenarios, it is either impossible or impractical to take into account every single member of the population. In situations like these, the best approach is to consider a subset of the population, called a **sample**, and analyze it hoping to draw conclusions that can be applied to the population as a whole.

To be able to extend the conclusions drawn from a sample to the entire population it is necessary to make some prior assumptions regarding the relationship between the two. A crucial (and often reasonale) hypothesis is to assume that the population has a probability distribution. Under this assumption, when a sample is drawn from the population, each element of the sample can be conceived as a random variable whose probability distribution is the sample distribution as the population one.

Suppose that the population has a probability distribution F. A set of random variables $X_1, X_2, ..., X_n$, each of them having a probability distribution F, constitutes a sample of F. The nature of F is known only to some extent: in some cases, the distribution of F is known while its parameters are not, in other cases the parameters of F are known but the distribution itself is unknown. There are even situations where almost everything regarding F is unknown.

Together with the hypothesis of each X_i having the same distribution, a second (much stronger) hypothesis is to assume that all of these random variables are independent of one another. A set of variables all having the same probability distribution and independent of one another are said to be **independent and identically distributed**, or **i.i.d.** for short. In turn, a sample constituted of i.i.d. variables is called a **random sample**.

Any value that can be calculated from a sample (that is, any function of the sample) is called a **statistic**. Prior to the act of sampling the value of any statistic is unknown, and can therefore be conceived as a random variable. For this reason, a statistic is often denoted with an uppercase letter while its specific realization (dependent on the sample drawn) with a lowercase letter. Being a random variable, it can be endowed with a probability distribution; the probability distribution of a statistic (interpreter as a random variable) is often referred to as a **statistic distribution**. The probability distribution of a statistic depends both on the probability distribution of the population from which the sample is drawn (normal, exponential, binomial, ecc...) and on the size n of the sample, but it also depends on *how* the sample is performed.

Let $X_1, X_2, ..., X_n$ be a random sample drawn from a certain population. Each of those variables, being distributed as the population itself, will all have the same mean and variance. Let μ and σ^2 be their respective values. It is possible to define the **sample mean** of said sample as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + ... + X_n}{n}$$

Theorem 4.1.1: Let $X_1, X_2, ..., X_n$ be a random sample from a (known) distribution with mean value μ and standard deviation σ. Then:

$$E\left(\overline{X}\right) = \mu \qquad\qquad V\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

Proof: Applying Theorem 3.2.1 and Theorem 3.6.1 gives:

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{E\left(\sum_{i=1}^{n} X_i\right)}{n} = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \frac{\cancel{n}\mu}{\cancel{n}} = \mu$$

Applying Theorem 3.2.2 and Theorem 3.6.2 gives:

$$V\left(\overline{X}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{V\left(\sum_{i=1}^{n} X_i\right)}{n^2} = \frac{\sum_{i=1}^{n} V(X_i)}{n^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n^2} = \frac{\cancel{n}\sigma^2}{n\cancel{^2}} = \frac{\sigma^2}{n}$$

□

Notice how Theorem 4.1.1 is completely independent of the nature of the population distribution.

Exercise 4.1.1: A certain brand of MP3 player comes in three sizes: the revenue in a given day is denoted by the random variable X. For each size, the percentage of customers buying said size has been computed:

Size	Price	% of customers
2GB	80$	20%
4GB	100$	30%
8GB	120$	50%

Suppose that 2 MP3s were sold today. X_1 and X_2 be the random variables denoting respectively the revenue of the first sale and the revenue of the second sale. Assume that X_1 and X_2 constitute a random sample, and are therefore independent and identically distributed. What is the joint probability distribution of X and Y? What are all the possible sample means and sample variances?

Solution: The expected value and variance of X are given by:

$$\mu = \sum xp(x) = 80 \cdot 0.2 + 100 \cdot 0.3 + 120 \cdot 0.5 = 106$$

$$\sigma^2 = \left(80^2 \cdot 0.2 - 106^2\right) + \left(100^2 \cdot 0.3 - 106^2\right) + \left(120^2 \cdot 0.5 - 106^2\right) = 244$$

X_1	X_2	$p(X_1, X_2)$	$\overline{X}$	s^2
80	80	$0.2 \cdot 0.2 = 0.04$	$\frac{80+80}{2} = 80$	$(80-80)^2 + (80-80)^2 = 0 + 0 = 0$
80	100	$0.2 \cdot 0.3 = 0.06$	$\frac{80+100}{2} = 90$	$(80-90)^2 + (100-90)^2 = 100 + 100 = 200$
80	120	$0.2 \cdot 0.5 = 0.1$	$\frac{80+120}{2} = 100$	$(80-100)^2 + (120-100)^2 = 400 + 400 = 800$
100	80	$0.3 \cdot 0.2 = 0.06$	$\frac{100+80}{2} = 90$	$(100-90)^2 + (80-90)^2 = 100 + 100 = 200$

100	100	$0.3 \cdot 0.3 = 0.09$	$\frac{100+100}{2} = 100$	$(100-100)^2 + (100-100)^2 = 0 + 0 = 0$
100	120	$0.3 \cdot 0.5 = 0.15$	$\frac{100+120}{2} = 110$	$(100-110)^2 + (120-110)^2 = 100 + 100 = 200$
120	80	$0.5 \cdot 0.2 = 0.1$	$\frac{120+80}{2} = 100$	$(120-100)^2 + (80-100)^2 = 400 + 400 = 800$
120	100	$0.5 \cdot 0.3 = 0.15$	$\frac{120+100}{2} = 110$	$(120-110)^2 + (100-110)^2 = 100 + 100 = 200$
120	120	$0.5 \cdot 0.5 = 0.25$	$\frac{120+120}{2} = 120$	$(120-120)^2 + (120-120)^2 = 0 + 0 = 0$

Sample mean:

$$P\left(\overline{X} = 80\right) = P(\{X_1 = 80\} \wedge \{X_2 = 80\}) = 0.04$$

$$P\left(\overline{X} = 90\right) = P(\{X_1 = 80\} \wedge \{X_2 = 100\} \vee \{X_1 = 100\} \wedge \{X_2 = 80\}) = 0.06 + 0.06 = 0.12$$

$$P\left(\overline{X} = 100\right) = P(\{X_1 = 100\} \wedge \{X_2 = 100\} \vee \{X_1 = 80\} \wedge \{X_2 = 120\} \vee \{X_1 = 120\} \wedge \{X_2 = 80\}) = 0.1 + 0.1 + 0.09 = 0.29$$

$$P\left(\overline{X} = 110\right) = P(\{X_1 = 100\} \wedge \{X_2 = 120\} \vee \{X_1 = 120\} \wedge \{X_2 = 100\}) = 0.15 + 0.15 = 0.3$$

$$P\left(\overline{X} = 120\right) = P(\{X_1 = 120\} \wedge \{X_2 = 120\}) = 0.25$$

Sample variance:

$$P(s^2 = 0) = P(\{X_1 = 80\} \wedge \{X_2 = 80\} \vee \{X_1 = 100\} \wedge \{X_2 = 100\} \vee \{X_1 = 120\} \wedge \{X_2 = 120\}) = 0.04 + 0.09 + 0.25 = 0.38$$

$$\{X_1 = 80\} \wedge \{X_2 = 100\} \vee \{X_1 = 100\} \wedge \{X_2 = 80\} \vee \{X_1 = 100\} \wedge \{X_2 = 120\} \vee X_1 = 100\} \wedge \{X_2 = 120\}) = 0.06 + 0.06 +$$

$$P(s^2 = 800) = P(\{X_1 = 120\} \wedge \{X_2 = 80\} \vee \{X_1 = 80\} \wedge \{X_2 = 120\}) = 0.1 + 0.1 = 0.2$$

□

Let $X_1, X_2, ..., X_n$ be a random sample drawn from a certain population, all having expected value μ and variance σ^2 It is possible to define the **sample variance** of said sample as:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

The square root of S^2, denoted as S, is called the **sample standard deviation**.

Theorem 4.1.2: Let $X_1, X_2, ..., X_n$ be a random sample from a (known) distribution with mean value μ and variance σ^2. Then:

$$E(S^2) = \sigma^2$$

4.2. Central Limit Theorem

Consider a sequence $\langle X_n \rangle = \{X_1, X_2, ..., X_n\}$ of identically distributed random variables such that their CDF depends on their index. That is, for each $i \in [1, n]$, the i-th random variable has as CDF a functions $F_i(x)$ that depends on the value of i. The sequence $\langle X_n \rangle$ is said to be **convergent in distribution** to a random variable X having CDF $F(x)$ if the following limit is valid for each $t \in \mathbb{R}$ such that F is continuous:

$$\lim_{n \to +\infty} F_n(t) = F(t)$$

To denote that a sequence of (identically distributed) random variables is convergent in distribution to a random variable X, the notation $\langle X_n \rangle \xrightarrow{d} X$ is used.

In simpler terms, if $\langle X_n \rangle \xrightarrow{d} X$ is true it means that $F_n(t)$ (dependent on n) is approximately equal to $F(t)$ (non dependent on n) and that the values of the indices of $\langle X_n \rangle$ (expected value, variance, ecc...) are approximately equal to those of X.

Exercise 4.2.1: Consider the following sequence $\langle X_n \rangle \in \mathbb{N}$ of random variables having the following CDF (dependent on n):

$$F_n(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ t^{\frac{n}{n+1}} & \text{if } 0 < t < 1 \\ 1 & \text{if } t \geq 1 \end{cases}$$

Study its convergence.

Solution: For $t \leq 0$ and $t \geq 1$, the CDF $F_n(t)$ is a constant, and therefore unproblematic. In the case of $0 < t < 1$, the limit is:

$$\lim_{n \to +\infty} t^{\frac{n}{n+1}} = t^{\lim_{n \to +\infty} \frac{n}{n+1}} = t^{\frac{1}{1+\frac{1}{n}}} = t^{\frac{1}{1+\frac{1}{+\infty}}} = t^{\frac{1}{1+0}} = t^1 = t$$

Therefore, $\langle X_n \rangle$ converges in distribution to a random variable having the following CDF (that does not depend on n):

$$F(t) = \begin{cases} 0 \text{ if } t \leq 0 \\ t \text{ if } 0 < t < 1 \\ 1 \text{ if } t \geq 1 \end{cases}$$

□

Consider a sequence $\langle X_n \rangle = \{X_1, X_2, ..., X_n\}$ of identically distributed random variables such that their CDF depends on their index. The sequence $\langle X_n \rangle$ is said to be **convergent in probability** to a random variable X if, for any $\varepsilon > 0$:

$$\lim_{n \to +\infty} P(|X_n - X| < \varepsilon) = 1$$

To denote that a sequence of (identically distributed) random variables is convergent in probability to a random variable X, the notation $\langle X_n \rangle \xrightarrow{p} X$ is used.

In simpler terms, a sequence $\langle X_n \rangle$ is convergent in probability the probability of an "unusual" outcome becomes smaller and smaller as the sequence progresses.

Theorem 4.2.1 (Weak Law of Large Numbers): Let $\langle X_n \rangle = \{X_1, X_2, ..., X_n\}$ be a sequence of identically distributed random variables, each having finite expected value μ and finite variance σ^2. Then:

$$\overline{X} \xrightarrow{p} \mu$$

In simpler terms, Theorem 4.2.1 states that the sample mean of a sequence of i.i.d. random variables gets closer and closer to their "true" expected value the longer of a sequence is considered.

Exercise 4.2.2: Consider the random variabile X, whose values are the possible outcomes of a 6-faced fair dice roll. Compare the expected value of X with the approximation retrieved by applying Theorem 4.2.1.

Solution: It is easy to see that:

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$$

Suppose that $n = 10$. Picking 10 random values for X (no matter how they are distributed) gives:

$$\frac{5+3+1+5+1+2+1+3+5+5}{10} = \frac{31}{10} = 3.1$$

Which is a reasonable approximation of $E(X)$. With $n = 20$:

$$\frac{6+4+1+1+3+4+2+2+4+1+2+6+5+2+6+2+4+5+6+3}{20} = \frac{69}{20} = 3.45$$

Which is an even better approximation. □

Note that, with respect to Theorem 4.2.1, the distribution of $\overline{X}$ is irrelevant. Also notice how the theorem does not give any indication on how "fast" the convergence of $\overline{X}$ to μ is. Intuitively, it is possible to relate the speed of convergence to the standard deviation of the X_i variables, since a smaller standard deviation entails that the values of $\overline{X}$ will be closer to their expected value, and therefore closer to μ.

Theorem 4.2.2 (Central Limit Theorem): Let $\langle X_n \rangle = \{X_1, X_2, ..., X_n\}$ be a sequence of i.i.d. random variables drawn from a population having finite mean μ and finite variance σ^2. Then:

$$X_1 + X_2 + ... + X_n \xrightarrow{d} A \sim N(n\mu, n\sigma^2)$$

Or equivalently, by normalizing:

$$\frac{X_1 + X_2 + ... + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim N(0, 1)$$

Note that Theorem 4.2.2 does not specify the nature of the distribution to which it is applied. This

means that, as long as it is possible to construct a sufficiently large[5] sequence of i.i.d. random variables, said sequence can always be treated as a standard normal random variable, even if the single variables have an unknown (yet equal among all of them) distribution.

In particular, Theorem 4.2.2 can be phrased with respect to the sample mean:

$$\frac{X_1 + X_2 + ... + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim N(0,1) \Rightarrow \frac{(X_1 + X_2 + ... + X_n - n\mu)/n}{(\sigma\sqrt{n})/n} \xrightarrow{d} Z \sim N(0,1) \Rightarrow$$

$$\frac{\overline{X} - \mu}{\sigma\sqrt{n}/n} \xrightarrow{d} Z \sim N(0,1) \Rightarrow \frac{n\left(\overline{X} - \mu\right)}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim N(0,1)$$

Exercise 4.2.3: Suppose that $\langle X_n \rangle$ is a sequence of 40 independent and randomly distributed random variables, each having $\mu = 14$ and $\sigma = 4.8$. What is the probability of $\overline{X}$ being less than or equal to 13?

Solution: Even though the distribution of X is unknown, it is still possible to apply Theorem 4.2.2:

$$P\left(\overline{X} \leq 13\right) = P\left(\frac{n\left(\overline{X} - \mu\right)}{\sigma\sqrt{n}} \leq \frac{n(13 - \mu)}{\sigma\sqrt{n}}\right) = P\left(Z \leq \frac{40(13 - 14)}{4.8 \cdot \sqrt{40}}\right) = \Phi\left(\frac{-40}{30.33}\right) \approx 0.09$$

□

On the other hand, it is not possible to apply Theorem 4.2.2 to the sample variance to retrieve its distribution. Also, if the sample size is too small, Theorem 4.2.2 does not apply, and therefore the distribution of the sample mean is unknown as well. Despite this, as long as the population is normally distributed, it is possible to infer something about the distribution of both.

Theorem 4.2.3: Let $X_1, X_2, ..., X_n$ be a random sample drawn from a population having mean μ and variance σ^2. If the population is normally distributed, the following holds:

$$\frac{n\left(\overline{X} - \mu\right)}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim N(0,1)$$

Theorem 4.2.4: Let $X_1, X_2, ..., X_n$ be a random sample drawn from a population having mean μ and variance σ^2. If the population is normally distributed, the following holds:

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

[5] Even though the theorem does not specify a minimum size of n such that the theorem holds for practical purposes, empirical data seems to suggest a value of 30 for most real-world applications.

Theorem 4.2.5: Let $X_1, X_2, ..., X_n$ be a random sample drawn from a population having mean μ and variance σ^2. If the population is normally distributed, $\overline{X}$ and S^2 are independent.

Corollary 4.2.1: Let $X_1, X_2, ..., X_n$ be a random sample drawn from a population having mean μ and variance σ^2. If the population is normally distributed, the following holds:

$$\frac{n\left(\overline{X} - \mu\right)}{S\sqrt{n}} \sim t_{n-1}$$

4.3. Point estimate

As stated, the interest of inferential statistics is to draw conclusions on the distribution of the population from a sample. In particular, even if the distribution of the population is known, its parameters (p for a Bernoulli, λ for an exponential, σ for a normal, ecc...) might not. It could therefore be interesting to approximate said parameters from the retrieved sample.

Any statistic defined with the intention of estimating a parameter θ is called an **estimator** of θ, and is therefore a random variable. Any particular value of an estimator of θ is called **esteem**, and is denoted as $\hat{\theta}$. The idea is to approximate the value of θ from the values of a sample in the form $X_1, X_2, ..., X_n$.

For most parameters of all distributions, there's a vast amount of estimators, each having pros and cons. In particular, there's a class of estimators called **Maximum Likelihood Estimators** (**MLE**) that are often employed in statistics. Estimators of this class are obtained from maximizing a specific function called *likelihood function*.

Let $X_1, X_2, ..., X_n$ be a random sample, and let $f(x_1, x_2, ..., x_n)$ be the joint probability mass function (or probability density function, if they are continuous) of the sample. The function $f(x_1, x_2, ..., x_n \mid \theta)$ can therefore be conceived as the *degree of certainty* associated with the events "The value of X_1 is x_1", "The value of X_2 is x_2", ..., "The value of X_n is x_n" happening all together knowing that the value of the parameter is indeed θ.

$$f(x_1, x_2, ..., x_n \mid \theta) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid \theta)$$

It is therefore reasonable to assume that a good value for θ is the value $\hat{\theta}$ that maximizes the function $f(x_1, x_2, ..., x_n \mid \theta)$ when the values of $X_1, ..., X_n$ are $x_1, ..., x_n$. In other words, taking the derivative of $f(x_1, x_2, ..., x_n \mid \theta)$ with respect to θ and equating it to 0. This function is called the **likelihood function**.

In maximizing the likelihood function, it is sometimes useful to make use of the property that $f(x_1, x_2, ..., x_n \mid \theta)$ and $\log(f(x_1, x_2, ..., x_n \mid \theta))$ have the same maxima. This means that it is possible to obtain $\hat{\theta}$ by maximizing this second function, called **log-likelihood function**.

Lemma 4.3.1: Let $X_1, X_2, ..., X_n$ be a random sample of a population that is Bernoulli distributed. The parameter p can be estimated with the following MLE:

$$d(X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Proof: By definition of Bernoulli variable of parameter p, $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. In a more compact form, it can be written as:

$$P(X_i = k) = p^k(1-p)^{1-k} \text{ with } k = 0, 1$$

Since the random variables in a random sample are independent:

$$f(x_1, x_2, ..., x_n \mid p) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid p) =$$
$$p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \cdot (...) \cdot p^{x_n}(1-p)^{1-x_n} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

Taking the logarithm gives:

$$\log(f(x_1, x_2, ..., x_n \mid p)) = \log\left(p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}\right) =$$
$$\log\left(p^{\sum_{i=1}^{n} x_i}\right) + \log\left((1-p)^{n-\sum_{i=1}^{n} x_i}\right) = \log(p)\left(\sum_{i=1}^{n} x_i\right) + \log(1-p)\left(n - \sum_{i=1}^{n} x_i\right)$$

Taking the derivative with respect to p and equating it to 0:

$$\frac{d}{dp}\left(\log(p)\left(\sum_{i=1}^{n} x_i\right) + \log(1-p)\left(n - \sum_{i=1}^{n} x_i\right)\right) = 0 \Rightarrow \frac{d}{dp}\left(\log(p)\left(\sum_{i=1}^{n} x_i\right)\right) +$$
$$\frac{d}{dp}\left(\log(1-p)\left(n - \sum_{i=1}^{n} x_i\right)\right) = 0 \Rightarrow \left(\sum_{i=1}^{n} x_i\right)\frac{d}{dp}(\log(p)) + \left(n - \sum_{i=1}^{n} x_i\right)\frac{d}{dp}(\log(1-p)) = 0 \Rightarrow$$
$$\left(\sum_{i=1}^{n} x_i\right)\frac{1}{p} - \left(n - \sum_{i=1}^{n} x_i\right)\frac{1}{1-p} = 0$$

Denoting with $\hat{p}$ the value for the specific realization of p when $X_1 = x_1 \wedge X_2 = x_2 \wedge ... \wedge X_n = x_n$ gives:

$$\left(\sum_{i=1}^{n} x_i\right)\frac{1}{\hat{p}} - \frac{n}{1-\hat{p}} + \left(\sum_{i=1}^{n} x_i\right)\frac{1}{1-\hat{p}} = 0 \Rightarrow \left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{\hat{p}} + \frac{1}{1-\hat{p}}\right) - \frac{n}{1-\hat{p}} = 0 \Rightarrow$$
$$\left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{\hat{p}(1-\hat{p})}\right) - \frac{n}{1-\hat{p}} = 0 \Rightarrow \left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{\hat{p}}\right) - n = 0 \Rightarrow \hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

□

Lemma 4.3.2: Let $X_1, X_2, ..., X_n$ be a random sample of a population that is Poisson distributed. The parameter λ can be estimated with the following MLE:

$$d(X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Proof: By definition of Bernoulli variable of parameter λ, $P(X_i = x_i) = \lambda^{x_i} e^{-\lambda}/x_i!$. Being all variables independent:

$$f(x_1, x_2, ..., x_n \mid \lambda) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid \lambda) =$$
$$\frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \cdot ... \cdot \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Taking the logarithm gives:

$$\log(f(x_1, x_2, ..., x_n \mid \lambda)) = \log\left(e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}\right) = \log\left(e^{-n\lambda}\right) + \log\left(\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}\right) =$$
$$\log\left(e^{-n\lambda}\right) + \log\left(\lambda^{\sum_{i=1}^{n} x_i}\right) - \log\left(\prod_{i=1}^{n} x_i!\right) = (-n\lambda)\log(e) + \left(\sum_{i=1}^{n} x_i\right)\log(\lambda) - \log\left(\prod_{i=1}^{n} x_i!\right) =$$
$$-n\lambda + \log(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\log(x_i!)$$

Taking the derivative with respect to λ and equating it to 0:

$$\frac{d}{d\lambda}\left(-n\lambda + \log(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\log(x_i!)\right) = 0 \Rightarrow \frac{d}{d\lambda}(-n\lambda) + \frac{d}{d\lambda}\left(\log(\lambda)\sum_{i=1}^{n} x_i\right) -$$
$$\frac{d}{d\lambda}\left(\sum_{i=1}^{n}\log(x_i!)\right) = 0 \Rightarrow -n\frac{d}{d\lambda}(\lambda) + \left(\sum_{i=1}^{n} x_i\right)\frac{d}{d\lambda}(\log(\lambda)) - 0 = 0 \Rightarrow -n + \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\lambda} = 0$$

Denoting with $\hat{\lambda}$ the value for the specific realization of λ when $X_1 = x_1 \wedge X_2 = x_2 \wedge ... \wedge X_n = x_n$ gives:

$$-n + \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\hat{\lambda}} = 0 \Rightarrow -n\hat{\lambda} + \sum_{i=1}^{n} x_i = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\hat{\lambda} \Rightarrow \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

□

Lemma 4.3.3: Let $X_1, X_2, ..., X_n$ be a random sample of a population that is normally distributed. The parameters μ and σ can be estimated with the following MLE:

$$d(X_1, X_2, ..., X_n) = \left\{\overline{X}, S\sqrt{\frac{n-1}{n}}\right\}$$

Proof: By definition of Normal variable of parameters μ and σ:

$$P(X_i = x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Being all variables independent:

$$f(x_1, x_2, ..., x_n \mid \mu, \sigma) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) =$$

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

Taking the logarithm gives:

$$\log(f(x_1, x_2, ..., x_n \mid \mu, \sigma)) = \log\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)\right) = \log\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\right) +$$

$$\log\left(\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)\right) = n\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)\log(e) =$$

$$-n\log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Taking the partial derivative with respect to μ and equating it to 0 gives:

$$\frac{d}{d\mu}\left(-\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = 0 \Rightarrow \frac{d}{d\mu}\left(-\frac{n}{2}\log(2\pi)\right) + \frac{d}{d\mu}(-n\log(\sigma)) +$$

$$\frac{d}{d\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = 0 \Rightarrow 0 + 0 - \frac{1}{2\sigma^2}\frac{d}{d\mu}\left(\sum_{i=1}^{n}(x_i - \mu)^2\right) = 0 \Rightarrow \sum_{i=1}^{n}\frac{d}{d\mu}(x_i - \mu)^2 = 0 \Rightarrow$$

$$\sum_{i=1}^{n} 2(x_i - \mu)\frac{d}{d\mu}(x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n} -2(x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n}(x_i - \mu) = 0$$

Denoting with $\hat{\mu}$ the value for the specific realization of μ when $X_1 = x_1 \wedge X_2 = x_2 \wedge ... \wedge X_n = x_n$ gives:

$$\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \Rightarrow (x_1 - \hat{\mu}) + +... + (x_n - \hat{\mu}) = 0 \Rightarrow -n\hat{\mu} + \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

Taking the partial derivative with respect to σ and equating it to 0 gives:

$$\frac{d}{d\sigma}\left(-\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = 0 \Rightarrow \frac{d}{d\sigma}\left(-\frac{n}{2}\log(2\pi)\right) + \frac{d}{d\sigma}(-n\log(\sigma)) +$$

$$\frac{d}{d\sigma}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = 0 \Rightarrow 0 - n\frac{d}{d\sigma}(\log(\sigma)) - \left(\sum_{i=1}^{n}(x_i - \mu)^2\right)\frac{d}{d\sigma}\left(\frac{1}{2\sigma^2}\right) = 0 \Rightarrow$$

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

Denoting with $\hat{\sigma}$ the value for the specific realization of σ when $X_1 = x_1 \wedge X_2 = x_2 \wedge ... \wedge X_n = x_n$ gives:

$$-\frac{n}{\hat{\sigma}}+\frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i-\hat{\mu})^2=0\Rightarrow n\hat{\sigma}^2=\sum_{i=1}^{n}(x_i-\overline{x})^2\Rightarrow\hat{\sigma}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}\Rightarrow$$

$$\hat{\sigma}=\sqrt{\frac{1}{n}}s\sqrt{n-1}\Rightarrow\hat{\sigma}=s\sqrt{\frac{n-1}{n}}$$

□

When choosing an estimator for a parameter θ, it is possible to inspect the candidates and "rank" them with respect to some properties that an estimator is expected to have.

Given a generic parameter θ, let $\hat{\theta}$ be one of its estimators, retrieved from the values of a random sample $X_1, X_2, ..., X_n$. One property that is favorable for $\hat{\theta}$ to have is **correctness**; an estimator is said to be correct if its expected value is the parameter itself. In other words, an estimator is correct if its distribution is "centered" on the true value of the parameter that it estimates:

$$E\left[\hat{\theta}\right]=\theta$$

Another favorable property is **consistency**. An estimator is said to be consistent if, the greater the size of the sample, the smaller is the difference between the estimation and the real value:

$$\lim_{n\to+\infty}P\left(\left|\hat{\theta}-\theta\right|\leq\varepsilon\right)=1\text{ for any }\varepsilon>0$$

Theorem 4.3.1: Let $X_1, X_2, ..., X_n$ be a random sample extracted from a population, and let $\hat{\theta}$ be a correct estimator for an unknown parameter θ of said population. If the following holds:

$$\lim_{n\to+\infty}V\left(\hat{\theta}\right)=0$$

then $\hat{\theta}$ is also consistent.

Lemma 4.3.4: Let $X_1, X_2, ..., X_n$ be a random sample extracted from a normally distributed population. The estimators $\hat{\mu}$ and $\hat{\sigma}$, that estimate μ and σ respectively, are both correct and consistent.

Proof: From Lemma 4.3.3, recall that:

$$\hat{\mu}=\overline{X}\qquad\qquad\hat{\sigma}^2=S^2\frac{n-1}{n}$$

As stated in Theorem 4.1.1, $E\left(\overline{X}\right)=\mu$. Since $\overline{X}=\hat{\mu}$, by transitive property $E(\hat{\mu})=\mu$, which means that $\hat{\mu}$ is correct. Applying Theorem 4.3.1 gives:

$$\lim_{n\to+\infty}V(\hat{\mu})=\lim_{n\to+\infty}V\left(\overline{X}\right)=\lim_{n\to+\infty}\frac{\sigma^2}{n}=\frac{\sigma^2}{+\infty}=0$$

Which means that $\hat{\mu}$ is also consistent.

□

4.4. Confidence intervals

A point estimate is not indicative of how reliable said estimation is. Also, since the value of a point estimate depends on the values of the sample drawn, there's no indication on why one value should be preferred over another, and no estimation will ever be exactly equal to the "true" value.

An alternative to reporting a single value for the parameter to be estimated is to report an *interval* of plausible values, an **interval estimate** or **confidence interval** (**CI** for short) endowed with a measure of its reliability.

To illustrate how a confidence interval is constructed, it is useful to start from a very simple (and unrealistic) situation and then introduce more and more complications.

The simplest scenario is the one where the parameter of interest is the population mean (the "true" mean), the population is known to be normally distributed and the value of the population standard deviation (the "true" standard deviation) is known.

Let $X_1, X_2, ..., X_n$ be the random variables denoting the observations and $x_1, x_2, ..., x_n$ be their realizations, resulting from a random sample having normal distribution with known mean μ and unknown standard deviation σ. Then, irrespective of the sample size n, the sample mean $\overline{X}$ is normally distributed with expected value μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Standardizing the sample mean allows to express it in terms of the population mean and standard deviation:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}$$

Suppose one requires the value of Z to possess a probability of 95%. The value for the standard normal random variable with said percentage is 1.96. Therefore:

$$P\left(-1.96 < \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} < 1.96\right) = 0.95 \Rightarrow P\left(-1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \overline{X} - \mu < 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 0.95 \Rightarrow$$

$$P\left(-1.96\left(\frac{\sigma}{\sqrt{n}}\right) - \overline{X} < -\mu < 1.96\left(\frac{\sigma}{\sqrt{n}}\right) - \overline{X}\right) = 0.95 \Rightarrow P\left(\overline{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 0.95$$

The interval $\left(\overline{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right), \overline{X} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right)$ is called a **confidence interval at** 95%. The value of 95% is called the **level of confidence**, while the value $2 \cdot 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ is the **width** of the interval. Such interval is a random interval because its endpoints depend on $\overline{X}$, which is a random variable.

After drawing the sample and collecting the realizations $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, it is possible to compute the (realization of the) sample mean $\overline{x}$, obtaining an *actual* interval.

Exercise 4.4.1: Industrial engineers study ergonomics to design keyboards that are comfortable to use and make people using them as productive as possible. Assume that a sample of $n = 31$ trained typists was selected and of them gave their evaluation of the best keyboard height. The sample average resulted in $\overline{x} = 80$ cm while the population mean, denoted by μ, is unknown. Assuming that the population standard deviation is known and equal to 2 cm, derive a confidence interval at 95%.

Solution:

$$\left(80 - 1.96\left(\frac{2}{\sqrt{31}}\right), 80 + 1.96\left(\frac{2}{\sqrt{31}}\right)\right) \approx (80 - 1.96 \cdot 0.36, 80 + 1.96 \cdot 0.36) \approx (79.3, 80.7)$$

□

It would be wrong to interpret a 95% confidence for μ as the probability that μ lies in $\left(\overline{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right), \overline{X} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right)$. The correct interpretation for a 95% confidence is that, by obtaining an interval from a drawn sample, there's a 95% chanche that said interval will contain μ somewhere.

The choice of 95% is, of course, arbitrary. In general, any percentage can be chosen by picking a quantile α. A $100(1-\alpha)\%$ confidence interval for the mean μ of a normal population when the value of σ is known is given by[6]:

$$P\left(\overline{x} - z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{x} + z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

Exercise 4.4.2: A production process for engine control housing units underwent a modification such that previously the hole diameter for bushing on the housing was distributed as a normal random variable with unknown mean and standard deviation 0.1 mm. It is believed that the modification did not alter the distribution of the hole diameter and the value of the standard deviation, while the mean might have changed. A sample of $n = 40$ housing units is selected and the hole diameter has been measured for each unit, obtaining a sample mean of $\overline{x} = 5.426$ mm. Construct a confidence interval for the average true diameter with confidence level of 90%.

Solution: If 90% is $1 - \alpha$, then α is 0.1. By looking at the tables it is possible to derive a value for $z_{0.95}$ of 1.645

$$\left(\overline{x} - z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{x} + z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = \left(5.426 - z_{0.95}\left(\frac{0.1}{\sqrt{40}}\right) < \mu < 5.426 + z_{0.95}\left(\frac{0.1}{\sqrt{40}}\right)\right) =$$
$$(5.426 - 1.645 \cdot 0.016 < \mu < 5.426 + 1.645 \cdot 0.016) = (5.4, 5.452)$$

□

The choice of a certain confidence level induces a certain interval size: if the confidence level is increased, the value of the quantile is also increased. This means that if the probability that the interval obtained from the sample contains the real value of the parameter is increased, the size of the interval is also increased. In other words, a gain in reliability entails a loss of precision, and it's not possible to be both precise and reliable.

If one is to explicitly pick both a confidence level and an interval size, it is the necessary to derive the sample size. If the size of the interval is denoted by w, it is possible to derive it by rearranging the expression of the width:

$$w = 2z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \Rightarrow w\sqrt{n} = 2z_{1-\alpha/2}\sigma \Rightarrow \sqrt{n} = \frac{2z_{1-\alpha/2}\sigma}{w} \Rightarrow n = \left(\frac{2z_{1-\alpha/2}\sigma}{w}\right)^2$$

To get a reasonable result, the value of n may or may not have to be ceiled.

[6]The notation for the quantile is very confusing. In some literature it is denoted as $z_{\alpha/2}$ and in others is $z_{1-\alpha/2}$.

Exercise 4.4.3: Suppose that the response time to a particular computer program is distributed as a normal random variable with unknown mean and standard deviation 25 μ m. When a new operating system is installed, there's interest in estimating the true average response time μ for such new environment. What sample size is needed to ensure that the confidence level is 95% and has width 10 μ s?.

Solution:

$$n = \left(\frac{2z_{0.995}\sigma}{w}\right)^2 = \left(\frac{2 \cdot 1.96 \cdot 25}{10}\right)^2 = (1.96 \cdot 5)^2 \approx 97$$

□

Consider a more general case where $X_1, X_2, ..., X_n$ are the random variables denoting the observations and $x_1, x_2, ..., x_n$ be their realizations, resulting from a random sample having unknown distribution with unknown mean μ and unknown standard deviation σ.

Even though the distribution of the X_i variable is unknown, if the sample size is sufficiently large, it is possible to apply the Central Limit Theorem and derive that their sum has a normal distribution (no matter the distribution of X_i). Also, the "real" standard deviation can still be substituted with its estimator s. Standardizing:

$$Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{s}$$

By picking a value of α and deriving a quantile $z_{1-\alpha/2}$, it is possible to construct a confidence interval with confidence level $100(1-\alpha)\%$ and the associated probability:

$$P\left(-z_{1-\alpha/2} < \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{s} < z_{1-\alpha/2}\right) = 1 - \alpha \Rightarrow P\left(-z_{1-\alpha/2} \cdot s < \sqrt{n}\left(\overline{X} - \mu\right) < z_{1-\alpha/2} \cdot s\right) = 1 - \alpha \Rightarrow$$

$$P\left(-z_{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \overline{X} - \mu < z_{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right) = 1 - \alpha \Rightarrow P\left(-z_{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right) - \overline{X} < -\mu < z_{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right) - \overline{X}\right) = 1 - \alpha \Rightarrow$$

$$P\left(\overline{X} - z_{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \overline{X} + z_{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right) = 1 - \alpha$$

The large sample intervals $\overline{X} \pm z_{1-\alpha/2} \cdot s/\sqrt{n}$ are a special case of a general large sample confidence interval for a parameter θ. Suppose that $\hat{\theta}$, an estimator for θ, has the three following properties:

1. It is normally distributed;
2. It's **unbiased**, meaning that $\mu_{\hat{\theta}} = \theta$;
3. $\sigma_{\hat{\theta}}$ the standard deviation of $\hat{\theta}$, is a known value.

For example, if θ is the (population) mean μ, the estimator $\hat{\theta} = \overline{X}$ possesses all three properties, since it is normally distributed, it is unbiased and the value of $\sigma_{\hat{\theta}}$ is known to be $\sigma/\sqrt{n}$. Standardizing θ gives:

$$Z = \frac{\hat{\theta} - \mu_{\hat{\theta}}}{\sigma_{\hat{\theta}}} = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

By picking a value of α and deriving a quantile $z_{1-\alpha/2}$, it is possible to construct a confidence interval with confidence level $100(1-\alpha)\%$ and the associated probability:

$$P\left(-z_{1-\alpha/2} < \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} < z_{1-\alpha/2}\right) \approx 1-\alpha$$

Exercise 4.4.4: Let $B \sim \text{Binom}(n,p)$, with both parameters unknown. Derive an estimator $\hat{p}$ for p and a confidence interval for $\hat{p}$ by performing a random sampling.

Solution: Let $X_1, X_2, ..., X_k$ be the random variables corresponding to the sampling. If $k \ll n$, each X_i is itself distributed as a binomial distribution. Recall that:

$$E(X) = np \qquad\qquad \text{SD}(X) = \sqrt{np(1-p)}$$

If $\overline{X}$ is an estimator of $E(X)$, and $E(X) = np$, then a natural choice for an estimator for p is $\overline{p} = \overline{X}/n$. Since $\overline{X}$ is normally distributed, then $\overline{X}/n$ is also normally distributed (dividing by n is a linear transformation). Also, $E(\hat{p}) = p$ and the expression for $\sigma_{\hat{\theta}}$ is known, because $\sigma_{\hat{\theta}} = \sqrt{p(1-p)/n}$. Standardizing, gives:

$$P\left(-z_{1-\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{1-\alpha/2}\right) \approx 1-\alpha$$

From which it is possible to derive a confidence interval:

$$-z_{1-\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{1-\alpha/2} \Rightarrow -z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \hat{p}-p < z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Rightarrow$$

$$\hat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Rightarrow \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

□

As long as the population size n is large, the CLT can be applied to aid the calculations. But if the size of the sample is small, it cannot be applied. In particular, with n being small, the random variable $\sqrt{n}\big(\overline{X}-\mu\big)/s$ is more "unstable" and spread out than a standard normal distribution.

When $\overline{X}$ is the mean of a random sample of (small) size n retrieved from a normal distribution with (unknown) mean μ and (unknown) standard deviation σ, the random variable defined as:

$$T = \frac{\overline{X}-\mu}{s/\sqrt{n}} = \frac{\sqrt{n}\big(\overline{X}-\mu\big)}{S}$$

Is distributed as a Student t distribution with $n-1$ degrees of freedom. A confidence interval with confidence level $100(1-\alpha)\%$ can then be constructed as:

$$\left(\overline{x} - t_{1-\alpha/2,t-1}\cdot\frac{s}{\sqrt{n}}, \overline{x} + t_{1-\alpha/2,t-1}\cdot\frac{s}{\sqrt{n}}\right)$$

4.5. Hypothesis testing

A parameter can be estimated either by a single number (point estimate) or by an interval of admissible values (a confidence interval). However, the objective of an investigation can also be putting a

claim concerning the parameter to the test, and see if such claim is plausible or not. The methods for accomplishing this comprise the part of statistical inference called **hypothesis testing**.

A **statistical hypothesis** is an assertion concerning the value of a single parameter (*"The value of this parameter is 0.5"*, *"The value of this parameter is lower than 3"*, ...), the value of several parameters (*"Parameter one is greater than parameter two"*, *"Parameters one and two are equal"*, ...), or about the form of a probability distribution (*"This sample was drawn from a normal distribution"*, *"This sample was drawn from a Poisson distribution"*, ...).

In any hypothesis testing problem, there are two mutually exclusive hypothesis under consideration: one is the one that is thought to be true, called the **null hypothesis**, and the other is its logical complement, called the **alternative hypothesis**. The null hypothesis is often denoted as H_0, while the alternative hypothesis as H_a or H_1. The objective is to decide, based on the information collected from a sample, which of the two is to be taken.

The alternative hypothesis should be taken into account if and only if the test says that the sample contradicts the null hypothesis with enough margin, and stick with the null hypothesis otherwise. If the sample is in line with the null hypothesis, this does not necessarely mean that the null hypothesis is true, it just means that there is not enough evidence to disprove it.

The simplest structure of a null hypothesis is $H_0 : \theta = \theta_0$. That is to say, the hypothesis is stating that the parameter θ is equal to the specific value θ_0. In this case, the alternative hypothesis is stating one of the following:

- θ is greater than θ_0, that is $H_a : \theta > \theta_0$
- θ is less than θ_0, that is $H_a : \theta < \theta_0$
- θ is different from θ_0, that is $H_a : \theta \neq \theta_0$

The first two are called **unilateral**, while the last one is called **bilateral**.

In an hypothesis testing problem, the rejection or confirmation of the null hypothesis is decided with respect to a **test statistic**. This statistic is a function of the sample data (a random variable) whose value obtained from the sample should be very different with respect to whether the null hypotesis is assumed to be true or to be false. If the value of the test statistic deviates decisively from the values expected from the null hypothesis, then the null hypothesis is rejected in favour of the alternative hypothesis. If the value of the test statistic is consistent with what is stated in the null hypothesis, then the null hypothesis is not rejected.

The issue is that a null hypothesis of the form $H_0 : \theta = \theta_0$ will never be confirmed with exact precision from the sample data, because the value of θ extracted from the sample will always be different from sample to sample. Therefore, the null hypothesis ought to be rejected when the value of θ retrieved from the sample deviates from θ_0 only within a small margin.

Such "closeness" to θ_0 is quantified in the **p-value**. Such value is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to what is stated in H_0 as the value calculated from the sample. A conclusion is reached by picking a number α called **significance level**, reasonably close to 0: H_0 is rejected in favour of H_a if the p-value is less than or equal to the level of significance, whereas H_0 will not be rejected if the p-value is greater than the level of significance. Even though α can be any value, it is customary to pick either 0.05, 0.01, 0.001 or 0.1.

The idea is that if the probability that the value of the test statistic computed from the sample is so extreme under the null hypothesis is very low, then such value cannot be justified by a fluctuation in the data, but ought to be interpreted as the null hypothesis poorly interpreting the scenario.

As stated before, if the data does not provide enough evidence to disprove the null hypothesis it does not necessarely mean that the null hypothesis is true: the data could be the result of a sample having very biased (unlikely, but still possible) outcome, that happened to agree with the null hypothesis. The values of the sample might also be biased in the other sense, appearing to favour the rejection of the null hypothesis simply because the sample was extremely favorable.

In both cases, a mistake is made. These two scenarios are summed up as:

- **Type I error**: rejecting H_0 even though it's true;
- **Type II error**: not rejecting H_0 even though it's false.

Theorem 4.5.1: In an hypothesis test, the probability of incurring in a type I error is equal to the level of significance of the test.

Proof: Let Y be the test statistic, with cdf given by F when H_0 is true. Suppose that Y has a continuos distribution over some interval, such that F is strictly increasing over such interval. If this is the case, F^{-1} is well defined.

Consider the case in which only values of Y are smaller than the computed value y are more contradictory to H_0 than y itself. Then:

$$\text{p-value} = P\left(\begin{array}{c} \text{A value for the test statistic at least as} \\ \text{contradictory to the null hypothesis is obtained} \end{array} \mid H_0 \right) = F(y)$$

Before having the sample data:

$$P(\text{Type I Error}) = P(\text{p-value} \leq \alpha \mid H_0) = P(F(y) \leq \alpha) = P[F^{-1}(F(y)) \leq F^{-1}(\alpha)] =$$
$$P[Y \leq F^{-1}(\alpha)] = P[Y \leq F^{-1}(\alpha)] = F[F^{-1}(\alpha)] = \alpha$$

The case in which only values of Y are greater than the computed value y are more contradictory to H_0 than y itself is analogous. The bilateral case is also analogous.

The theorem still holds for Y being a discrete distribution as long as it is possible to provide a well defined inverse function of the cdf. □

A formula for computing the probability of committing a type II error (often denoted with β) depends on the test statistic, and isn't always available.

If the probability of committing one of the two errors decreases, the probability of committing the other increases. Therefore, there's a tradeoff to be made. Out of the two errors, the type I error is generally considered to be more problematic than the type II error, because rejecting an hypothesis generally mean establishing an entire new framework, while not rejecting an hypothesis simply means keeping things as they are[7].

4.5.1. Z tests about μ, known σ

Let $X_1, ..., X_n$ be a random sample retrieved from a normal distribution with mean value μ and (known) standard deviation σ. Then, since the sum of normal distributions is itself normal, the sample mean $\overline{X}$ is normally distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$.

Let $H_0 : \mu = \mu_0$, where μ_0 is referred to as the **null value**. The alternative hypothesis can either be $H_1 : \mu > \mu_0$, $H_1 : \mu < \mu_0$ or $H_1 : \mu \neq \mu_0$. $\overline{X}$ can be standardized to get $Z = \left(\overline{X} - \mu_0/(\sigma/\sqrt{n})\right)$.

[7] This is analogous to a trial: it is better to absolve a criminal than to condemn an innocent.

When H_0 is true, $\mu_{\overline{X}} = \mu_0$. The statistic Z is a natural measure of the distance between $\overline{X}$, the estimator of μ, and its expected value when H_0 is true ($\mu_{\overline{X}}$). If the realization $\overline{x}$ of the sample mean $\overline{X}$ considerably exceeds μ_0 in a direction consistent with H_1, there is sufficient evidence to reject H_0.

Let $H_1 : \mu > \mu_0$. The null hypothesis ought to be rejected if a very large value $z = (\overline{x} - \mu_0/(\sigma/\sqrt{n}))$ from Z is retrieved. If this is the case, this implies that any value exceeding z is even more inconsistent with H_0 than z itself. The p-value of the test is therefore the probability of retrieving a value for Z greater or equal than z assuming H_0 to be true.

Exercise 4.5.1.1: In a certain city, it has been calculated 10 years ago that the amount of toxins in a battery of water is distributed with mean $\mu_0 = 2.0$ g. A random sample of 51 batteries gave a sample mean of 2.06 g and a sample standard deviation of 0.141 g. The two hypothesis are as follows:

$$H_0 : \mu = \mu_0 \qquad\qquad H_1 : \mu > \mu_0$$

Does this data provide compelling evidence that the claim $\mu = 2.0$ g still holds true to this day? Use a significance level of $\alpha = 0.01$.

Solution: Since the distribution of the population is not known and the standard deviation of the population is also not known, it is possible to apply the CLT and get:

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{2 - 2.06}{1.41/\sqrt{51}} = 3.04$$

Which is the standardized version of $\overline{X}$ assuming the null hypothesis to be true. This means that the evidence obtained from the data is roughly 3 standard deviation larger than the expected value under H_0. The p-value is given by:

$$p = P(Z \geq 3.04) = 1 - \Phi(3.04) = 1 - 0.9988 = 0.0012$$

Which is lower than the chosen α (and extremely low in general). Therefore, the null hypothesis ought to be rejected. □

Let $H_1 : \mu < \mu_0$. The null hypothesis ought to be rejected if a very small value $z = (\overline{x} - \mu_0/(\sigma/\sqrt{n}))$ from Z is retrieved. If this is the case, this implies that any value lower than z is even more inconsistent with H_0 than z itself. The p-value of the test is therefore the probability of retrieving a value for Z smaller or equal than z assuming H_0 to be true.

Let $H_1 : \mu \neq \mu_0$. The null hypothesis ought to be rejected if a very small or very large value $z = (\overline{x} - \mu_0/(\sigma/\sqrt{n}))$ from Z is retrieved. If this is the case, this implies that any value lower than z in the first case and a value greater than z in the second case is even more inconsistent with H_0 than z itself. The p-value of the test is therefore the probability of retrieving a value for $|Z|$ greater or equal than z assuming H_0 to be true.

The z tests with known σ are among the few for which there are simple formulas for computing β, the probability of a type II error to occur. Let $H_1 : \mu > \mu_0$, and let μ' be any value of μ that exceeds μ_0. If H_0 is not rejected when $\mu = \mu'$ then, by definition, a type II error occurred. Denote with $\beta(\mu')$ the probability of not rejecting H_0 when $\mu = \mu'$. This results in:

$$\beta(\mu') = P(H_0 \text{ is not rejected when } \mu = \mu') = P\Big(\overline{X} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n} \text{ when } \mu = \mu'\Big) =$$

$$P\left(\frac{\overline{X} - \mu'}{\sigma/\sqrt{n}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu = \mu'\right) = \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

In the case of $H_1 : \mu < \mu_0$, it is easy to see that:

$$\beta(\mu') = 1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

In the same fashion, if $H_1 : \mu \neq \mu_0$ then:

$$\beta(\mu') = \Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

When the alternative hypothesis is either $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$, the sample size n should be chosen to satisfy:

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \Phi(-z_\beta) = \beta$$

This is because $-z_\beta$ represents the z critical value that captures lower-tail area β. Solving for n gives:

$$n = \left[\frac{\sigma\big(z_\alpha + z_\beta\big)}{\mu_0 - \mu'}\right]^2$$

In the case of $H_1 : \mu \neq \mu_0$, it is still possible to retrieve an approximate solution:

$$n = \left[\frac{\sigma\big(z_{\alpha/2} + z_\beta\big)}{\mu_0 - \mu'}\right]^2$$