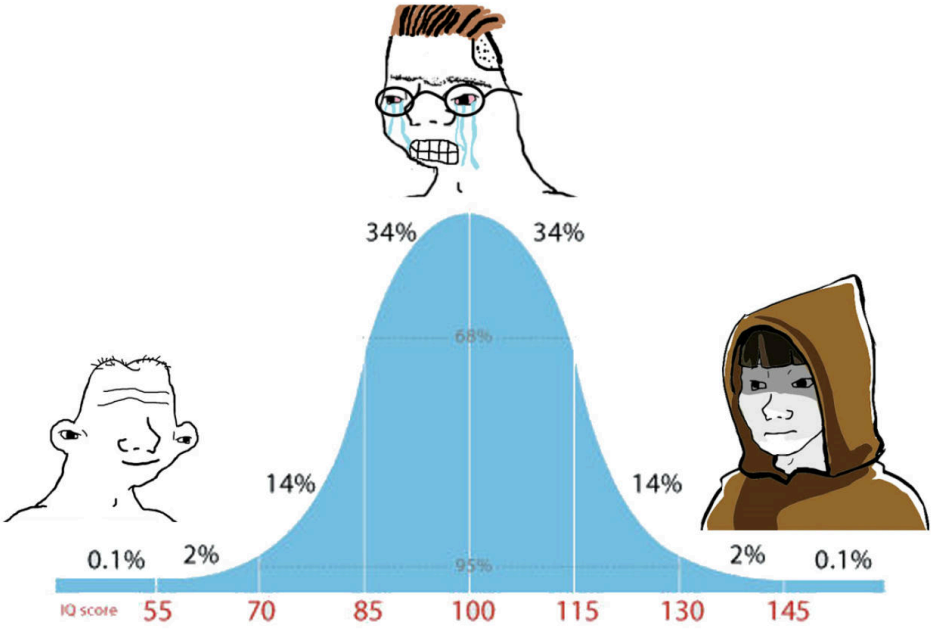
Contents


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 on their own represent an aspect of the data as a whole.

A statistical enquiry will typically focus on a well-defined collection of objects constituting a **population**. When desired information is available for all objects in the population, a **census** is available.

In general, such a situation is hardly possible, either because it would be too expensive or too time consuming to do so or simply because the population has an infinite amount of members. A more reasonable approach is to extract a subset of the population, called **sample** that is both sufficiently small to be able to work with and sufficiently large to capture all the nuances of the population as a whole.

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. □

Each object of the population possesses many features, some of which may or may not be of interest. Any feature whose value might change from object to object in the population and that has relevance with respect to a statistical enquiry is called a **variable**.

Variables are generally distinct in **numerical** variables and **categorical** variables. Numerical variables are distinct in **discrete** and **continuous**. Numerical variables are discrete if the set of its possible values is either finite or countably infinite. Numerical variables are continuous if the set of its possible values is uncountably infinite. Categorical variables are distinct in **ordinal** and **nominal**. Categorical variables are ordinal if the set of its possible values obeys an objective hierarchy or ordering of some sort, otherwhise are called **nominal**.

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

Solution:

- A numerical discrete variable could be the number of items sold in a store, since such number is necessarely an integer (it's not possible to sell, say, half an item, or three quarters of an item). Another example is the number of attempts necessary to win the lottery: it could be infinite, but it's still countable;
- A numerical continuous variable could be the temperature measured in a certain meteorological station, since such value is a real number (it could be approximated to an integer, but it would entail losing much informaton);
- A categorical ordinal variable could be the ranks in an army, such as general, private, captain, etcetera. Such ranks can be arranged in a (very) strict hierarchy, for example corporal is lower than general while corporal is higher than private;
- A categorical nominal variable could be the colors of a dress. It would make little sense to say that, for example, red scores higher than green or that pink scores lower than blue, at least in an objective way.

□

1.2. Frequencies for a single variable

Given a variable, let $V = \{v_1, v_2, ..., v_N\}$ be the set of all possible values for that variable. If said variable is discrete, to obtain a firsthand impression of the values in the sample it is useful to compute the **absolute frequency** of each possible value $v_j \in V$. The absolute frequency f_j is defined as the number of occurrences of v_j in the sample. A more general quantity is the **absolute frequency distribution**, defined as the function f that, for any v_j, outputs the corresponding absolute frequency.

From the absolute frequency it is possible to define what is called the **cumulative absolute frequency** F_j, given by the sum of the absolute frequencies of all possible values $v_k \in V$ such that $v_k \leq v_j$. As for the absolute frequency, it is possible to define the **cumulative absolute frequency distribution** as the function F that, for any v_j, outputs the corresponding cumulative absolute frequency.

$$F_j = \sum_{k:v_k \leq v_j}^{|V|} f_k$$

The relative frequency p_j is given by the ratio between the absolute frequency f_j and the sample size n. As for the absolute frequency, it is possible to define the **relative frequency distribution** as the function p that, for any v_j, outputs the corresponding relative frequency.

$$p_j = \frac{f_j}{n}$$

From the relative frequency it is possible to define what is called the **cumulative relative frequency** P_j, given by the sum of the relative frequencies of all possible values $v_k \in V$ such that $v_k \leq v_j$. As for the relative frequency, it is possible to define the **cumulative relative frequency distribution** as the function P that, for any v_j, outputs the corresponding cumulative relative frequency.

$$P_j = \sum_{k:v_k \leq v_j}^{|V|} p_k$$

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	

Compute the absolute frequency, the relative frequency, the cumulative absolute frequency and the cumulative relative frequency of this variable.

Solution:

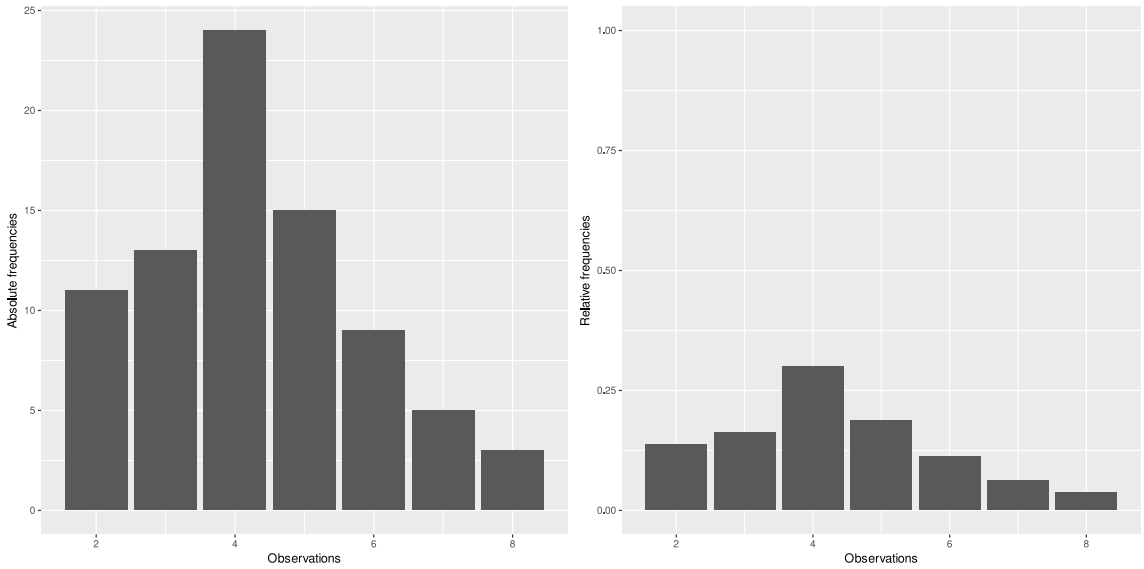
Number of rooms	Absolute frequency	Relative frequency	Cumulative absolute frequency	Relative absolute frequency
2	11	0.138	11	0.138
3	13	0.162	24	0.3
4	24	0.3	48	0.6
5	15	0.188	63	0.787
6	9	0.112	72	0.9
7	5	0.062	77	0.963
8	3	0.038	80	1

□

Sometimes, computing frequencies and arranging them in a frequency table is not particularly representative. An alternative representation to the tabular representation of the absolute and relative frequencies is the **bar plot**. A bar plot is a graphical representation constitued by a cartesian plane, where on the x axis the possible values of the variable are arranged whereas on the y axis the absolute or relative frequencies of its possible values are arranged. For each possible value v_i, a rectangle is drawn above it, having size f_i (for the absolute frequency) or p_i (for the relative frequency).

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

Solution:



□

When the variable under analysis is continuous, it is impractical to compute the frequency for each of its possible values, since it is expected that each possible value will appear no more than a few times. A better approach is to partition the values of the variable into **classes**: each of these disjointed sets represents an interval and contains the number of observations that lie in such interval. As long as the union of all classes reconstructs the entire set of values, the choice of the intervals and their size is arbitrary. Two intervals may or may not have the same size.

A class may or may not contain their extremes: if it contains its highest value (the leftmost value of the interval), said class is said to be **closed on the right**, whereas if it contains its lowest value (the rightmost value of the interval) said class is said to be **closed on the left**. In general, a class is closed on the right but not on the left.

Once a class is defined, it is possible to compute the frequencies with respect to the sizes of the classes, which are representative of the frequencies of its values.

Exercise 1.2.3: 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 absolute frequency, the relative frequency, the cumulative absolute frequency and the cumulative relative frequency of this variable.

Solution: This variable is not discrete, but continuous. The lowest value is 171, whereas the highest is 227. A possible approach would be to divide the values into the following six classes:

$$[171-180) \quad [180-190) \quad [190-199) \quad [199-208) \quad [208-218) \quad [218-227)$$

From said construction, it is possible to compute the four frequencies:

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
(218,227]	4	0.1	40	1

□

It is possible to draw a bar plot relative to a continuous variable, referring to the frequencies of the classes instead of the frequencies of each single value of the sample.

If the intervals that define each class are not of equal size, a bar plot becomes less representative, because class sizes are not taken into account. An alternative plot for this scenario is the **histogram**: an histogram is a bar plot where the width of each rectangle depends on the size of the interval of the related class, whereas the height is given by the ratio between the frequency of the class and its size. This quantity, called **frequency density**, can be thought of as the frequency of a class "normalized" with respect to its size.

1.3. Frequencies for two variables

In most cases, there is interest in analyzing more than one variable at a time for the same sample. For the sake of simplicity, consider a situation where only two variables are under analysis at the same time (it is trivial to generalize to m variables), either both discrete or both continuous. The set of observations will therefore be in the form $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where x_i is the value coming from the first variable and y_i is the value coming from the second variable.

Given two variables X and Y, let $V = \{(v_j, v_i)\}$ with $1 \leq i \leq N$ and $1 \leq j \leq M$ be the set of all possible couples that can be constructed by choosing a value for X and a value for Y. In the same fashion as the single variable case, it is possible to compute the **absolute frequency** f_{ji} as the number of observations in the sample having v_j as the value for the first variable and v_i as the value for the second variable. From the absolute frequency it is possible to compute the **double absolute frequency distribution** as the function f that, for any (v_j, v_i), outputs the corresponding absolute frequency.

From the absolute frequency it is possible to define what is called the **cumulative absolute frequency** $F_{j,i}$, given by the sum of the absolute frequencies of all possible values $(v_a, v_b) \in V$ such that $v_a \leq v_j$ and $v_b \leq v_i$. As for the absolute frequency, it is possible to define the **cumulative double absolute frequency distribution** as the function F that, for any (v_j, v_i), outputs the corresponding cumulative absolute frequency.

$$F_{j,i} = \sum_{a:v_a \leq v_j, b:v_b \leq v_i}^{|V|} f_{a,b}$$

The relative frequency $p_{j,i}$ is given by the ratio between the absolute frequency $f_{j,i}$ and the sample size n. As for the absolute frequency, it is possible to define the **double relative frequency distribution** as the function p that, for any (v_j, v_i), outputs the corresponding relative frequency.

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

From the relative frequency it is possible to define what is called the **cumulative relative frequency** $P_{j,i}$, given by the sum of the relative frequencies of all possible values $(v_a, v_b) \in V$ such that $v_a \leq v_j$ and $v_b \leq v_i$. As for the relative frequency, it is possible to define the **cumulative double relative frequency distribution** as the function P that, for any (v_j, v_i), outputs the corresponding cumulative relative frequency.

$$P_{j,i} = \sum_{a:v_a \leq v_j, b:v_b \leq v_i}^{|V|} p_{a,b}$$

One way to obtain a graphical representation for two variables is simply to extend bar plots to three dimensions (the values for the first character on the x axis, the values for the second character on the y axis and the frequencies on the z axis). An alternative representation is the **heat map**: a table where the possible values for the two variables are arranged on the sides and each (v_j, v_i) cell is coloured: said colour is as intense as the value of $f_{j,i}$.

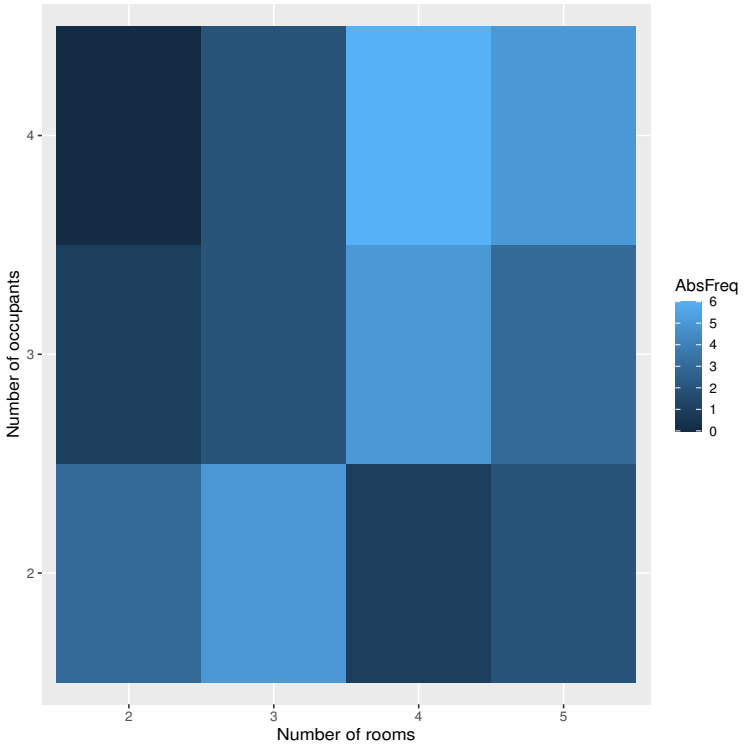
Exercise 1.3.1: Suppose that a sample has been collected regarding the number of rooms in a flat and the number of people inhabiting such flat:

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

Compute the four frequencies for the sample and draw a heat map for the absolute frequency.

Solution:

Number of rooms	Number of occupants	Absolute frequency	Relative frequency	Cumulative absolute frequency	Relative absolute frequency
2	2	3	0.086	3	0.086
2	3	1	0.029	4	0.114
2	4	0	0	4	0.114
3	2	5	0.143	9	0.257
3	3	2	0.057	11	0.314
3	4	2	0.057	13	0.371
4	2	1	0.029	14	0.4
4	3	5	0.143	19	0.543
4	4	6	0.171	25	0.714
5	2	2	0.057	27	0.771
5	3	3	0.086	30	0.857
5	4	5	0.143	35	1



□

Aside from said frequencies, that are analogous to single variable frequencies, other frequencies can be considered when analyzing samples with two variables. One such example are **marginal frequencies**, frequencies computed on exclusively one of the two variables without taking into account the value of the other.

1.4. Central tendency indices

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

1.4.1. Sample mean

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

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

Exercise 1.4.1.1: Consider the sample $\{1, 0, 7, 4, 4\}$. What is its sample mean?

Solution:

$$\overline{x} = \frac{1 + 0 + 7 + 4 + 4}{5} = \frac{16}{5} = 3.2$$

□

If the relative or absolute frequency of the sample is known, it can be used to compute its sample mean:

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{N} v_j f_j = \frac{v_1 f_1 + v_2 f_2 + ... + v_N f_N}{n} \qquad \overline{x} = \sum_{j=1}^{N} v_j p_j = v_1 p_1 + v_2 p_2 + ... + v_N p_N$$

Exercise 1.4.1.2: Let $v_1 = 4, v_2 = 5, v_3 = 8$ and $f_1 = 2, f_2 = 3, f_3 = 1$. What is the sample mean?

Solution:

$$\overline{x} = \frac{4 \cdot 2 + 5 \cdot 3 + 8 \cdot 1}{2 + 3 + 1} = \frac{31}{6} = 5.16$$

□

The difference between an observation x_i and the sample mean $\overline{x}$ is called **residue**.

Lemma 1.4.1.1: The sum of all residues is always zero:

$$\sum_{i=1}^{n} \overline{x} - x_i = (\overline{x} - x_1) + (\overline{x} - x_2) + ... + (\overline{x} - x_n) = 0$$

Lemma 1.4.1.2: Consider the sum of the squares of all residues:

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

Its minimum is the sample mean.

If a variable is continuous and not discrete, it is still possible to compute its sample mean using the formula. An alternative approach is to use the formula but substituting x_i with the average value from each i-th class. This is faster than computing the sample mean using every single value, but it is less precise.

Theorem 1.4.1.1: Let $\overline{x}$ be the sample mean of some sample, and let $T = ax + b$ with $a, b \in \mathbb{R}$ a linear transformation. Suppose that T is applied to each element of the sample: the sample mean of the transformed sample is $a\overline{x} + b$.

1.4.2. Sample median

Given a discrete numerical variable x, let $x_1, x_2, ..., x_n$, be the observations collected from the sample of such variable, arranged from lowest to highest (including duplicates). The **sample median** $\tilde{x}$ is a summary measure that describes the central value, and is calculated as either the middle value of such sequence if n is odd or the average of the two middle values if n is even:

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

Exercise 1.4.2.1: Consider the sample $\{1, 0, 7, 4, 4\}$. What is its sample median?

Solution: Rearranging the sample in increasing order gives $\{0, 1, 4, 4, 7\}$. The sample size is 5, which is odd, therefore:

$$\tilde{x} = x_{\frac{n+1}{2}} = x_3 = 4$$

□

Sample mean and sample median are not the same, and in general do not coincide. In an ordered sample, the extreme values do not influence the sample median, because it depends exclusively on the center of the sample, whereas the sample mean is influenced by the sample as a whole.

1.4.3. Sample mode

Given a discrete numerical variable x, let $x_1, x_2, ..., x_n$, be the observations collected from the sample of such variable. The **sample mode** is the value that appears the most, that is the value having the highest frequency. Note that a sample can have a single mode (**monomodal**) or many modes (**multimodal**).

If a variable is continuous, it is still possible to compute the sample mode of a sample as the class whose frequency is the highest.

1.4.4. Sample percentiles

Given a discrete numerical variable x, let $x_1, x_2, ..., x_n$ be the observations collected from the sample of such variable, with n being the cardinality of the sample. 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. Even though p can be any number, the most useful values 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, also called **second quartile**: $1/2$ of the sample are on the left and $1/2$ of the sample are on the right (this is equivalent to the sample median). 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.

Let np be the product between the sample size and the chosen real value p. The $100p$-th percentile q is the value that, arranging the sample in increasing order, is greater or equal than at least np values and less than or equal than $n(1-p)$ values.

The value of q depends on the value of np: if np is not an integer, then q is the value holding the position $\lfloor np \rfloor + 1$ in the arranged sample. If np is an integer, the value q is given by the mean between the np-th element and the $n(p+1)$-th element of the arranged sample:

$$q = \begin{cases} x_{\lfloor np \rfloor + 1} & \text{se } np \text{ non é intero} \\ \frac{1}{2}\left(x_{np} + x_{np+1}\right) & \text{se } np \text{ é intero} \end{cases}$$

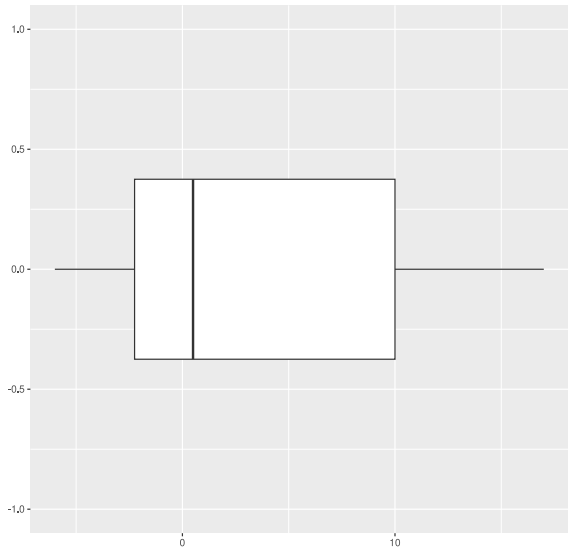
An interesting graphical representation of a sample and its three main percentiles is the **box plot**. The plot is composed by a straight line that goes from the lowest to the highest value in the sample, above of which lies a rectangle whose sides are drawn where the first and third quartiles are and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

Exercise 1.4.4.1: Given the following sample, compute the first, second and third quantile and draw the corresponding box plot:

-6 -3 -3 0 0 1 10 10 10 17

Solution:

$$Q_1 = x_3 = -3 \qquad Q_2 = \frac{x_5 + x_6}{2} = 0.5 \qquad Q_3 = x_8 = 10$$



□

1.5. Variability indices

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

1.5.1. Sample variance

The **sample variance** s^2 is a summary measure that describes how "spread out" are the values of the sample, or equivalently how close its values are to the sample mean, and is defined as:

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

Exercise 1.5.1.1: Consider the sample $\{0, 3, 7, 14\}$, whose sample mean is 6. Compute its sample variance.

Solution:

$$s^2 = \frac{\sum_{i=1}^{4} (x_i - \overline{x})^2}{3} = \frac{(0-6)^2 + (3-6)^2 + (7-6)^2 + (14-6)^2}{3} = \frac{36+9+1+64}{3} = \frac{110}{3} \approx 36.7$$

□

When the observations are distant from the sample mean, the sample variance of the sample is big, whereas it is small when the observations are closer to the sample mean.

As for the sample mean, the sample variance can be computed from the relative or absolute frequency:

$$s^2 = \frac{1}{n} \sum_{j=1}^{n} \left(v_j - \overline{x}\right)^2 f_j \qquad\qquad s^2 = \sum_{j=1}^{n} \left(v_j - \overline{x}\right)^2 p_j$$

Theorem 1.5.1.1: Given a discrete numerical variable x, let $x_1, x_2, ..., x_n$, be the observations collected from the sample of such variable, and let c be a numerical constant. Then:

1. If, for each $1 \leq i \leq n$, the y variable is constructed as $y_i = x_i + c$, it is true that $s_y^2 = s_x^2$;
2. If, for each $1 \leq i \leq n$, the y variable is constructed as $y_i = cx_i$, it is true that $s_y^2 = c^2 s_x^2$;

Where s_x^2 is the sample variance of the "original" variable x and s_y^2 is the sample variance of the "transformed" variable y.

1.5.2. 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 when there's interest in having an index having the same unit of measurement of the sample, since the sample variance is a squared quantity, and therefore it's unit of measurement is also squared.

1.5.3. Interquartile range

The **interquartile range** (**IQR** for short) is given by the average between the third quantile and the first quantile:

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

Represents the interval where most of the observations lie.

1.5.4. Coefficient of variation

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

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

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

2. Probability theory

2.1. Probability

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 as $\mathcal{S}$, is called **sample space**. The sample space can be either discrete or continuous.

Exercise 2.1.1: Provide some examples of experiments.

Solution:

- The roll of a six-sided dice is an experiment, since the resulting value of the dice is unknown until the dice is rolled. The sample space $\mathcal{S}$ contains 6 elements:

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

- The drawing of a card from a (standard) deck is an experiment, since the value of the card is unknown until the card is drawn. The sample space $\mathcal{S}$ contains 52 elements:

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

- The gender assigned to the offspring of a couple is an experiment, since their gender is unknown until (roughly) 4 months since conception. The sample space $\mathcal{S}$ contains 8 elements:

$$\mathcal{S} = \{MMM, MMF, MFM, FMM, MMF, FFM, FMF, FFF\}$$

□

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** otherwise (it contains multiple outcomes). An event can either occur or not occur, depending on the outcome of the experiment.

Exercise 2.1.2: Provide some examples of events.

Solution:

- Consider the roll of a six-sided dice. The subset $A = \{1, 3, 5\}$ of the sample space $\mathcal{S}$ corresponds to the event "an even number". It is a compound event;
- Consider the drawing of a card from a deck. The subset $B = \{A\heartsuit, A\diamondsuit, A\clubsuit, A\spadesuit, K\heartsuit, K\diamondsuit, K\clubsuit, K\spadesuit\}$ of the sample space $\mathcal{S}$ corresponds to the event "either an ace or a king of any set". It is a compound event;
- Consider the gender assigned to the offspring of a couple. The subset $C = \{FFF\}$ of the sample space $\mathcal{S}$ corresponds to the event "exclusively female offspring". It is a simple event.

□

Being sets, events can be manipulated using set algebra. In particular, given two events A and B:

- 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. A^c is also called the **complementary event** of A;
- 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.

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

Solution:

- Consider the roll of a six-sided dice. The subset $A = \{1, 2, 3, 4, 5\}$ of the sample space $\mathcal{S}$ corresponds to the event "any number but 6". It is the complement of the event "exactly six";
- Consider the drawing of a card from a deck. The subset $B = \{A\heartsuit, A\diamondsuit, A\clubsuit, A\spadesuit, K\heartsuit, K\diamondsuit, K\clubsuit, K\spadesuit\}$ of the sample space $\mathcal{S}$ is actually a union of two smaller events, the first being "an ace of any set" and the second being "a king of any set";
- Consider the gender assigned to the offspring of a couple. Consider the two events "a male as first born" and "a female as third born". Their intersection, representing the event "a male as first born and a female as third born" is given by:

$$\{MMM, MMF, MFM, MFF\} \cap \{MMF, MFF, FMF, FFF\} = \{MMF, MFF\}$$

□

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 esclusive** events, or **disjoint** events. In other words, two events are said to be mutually esclusive if they have no way of happening at the same time.

Modern probability theory, like set theory, is defined axiomatically. Such axioms are also called **Kolmogorov axioms**, and are (supposed to be) the minimum amount of axioms that are needed to construct a theory of probability free of contradictions.

To an event A, it is possible to associate a value called its **probability**, denoted as $P(A)$, that represents a measure of likelihood, certainty or confidence of such event to occur (intuitively, the higher the value of probability, the higher the likelihood). Probabilities obey three axioms, here stated:

1. For any event A, $P(A) \geq 0$. That is, the probability of an event happening is non negative;
2. $P(\mathcal{S}) = 1$. That is, the probability of any even happening at all is fixed as 1;
3. If $A_1, A_2, ...$ is a collection of countably infinite disjoint events, the following equality holds:

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

That is, given a set of events where no event can occur if at most another one of them occurs, the probability of any such event to occur is the sum of the individual probabilities.

From such axioms, it is possible to derive many useful consequences.

Theorem 2.1.1: $P(\emptyset) = 0$. That is, the null event cannot occur.

Proof: Consider the countably infinite collection of events $\emptyset, \emptyset, ...$ By definition, the null event is disjoint with itself, since set algebra gives $\emptyset \cap \emptyset = \emptyset$. The collection $\emptyset, \emptyset, ...$ is therefore made up of disjoint events, and by set algebra $\emptyset \cup \emptyset \cup ... = \emptyset$, therefore $P(\emptyset \cup \emptyset \cup ...) = P(\emptyset)$. Since by axiom 3 $P(\emptyset \cup \emptyset \cup ...) = \sum_{i=1}^{\infty} P(\emptyset)$, by transitive property $\sum_{i=1}^{\infty} P(\emptyset) = P(\emptyset)$. Since by axiom 1 the value of $P(\emptyset)$ has to be non negative, such equality can hold exclusively if $P(\emptyset) = 0$. □

Theorem 2.1.2: If $A_1, A_2, ..., A_n$ is a collection of finitely many disjoint events, the following equality holds:

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

Proof: Consider the countably infinite collection of events $A_1, A_2, ..., A_n, A_{n+1} = \emptyset, A_{n+2} = \emptyset, ..., \emptyset$, that is, a collection constructed by encoding countably infinitely many null events to the original collection. Applying axiom 3 to such collection gives:

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

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

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

But by Theorem 2.1.1, $P(\emptyset) = 0$. Therefore:

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

□

Theorem 2.1.3: For any event A, $P(A) + P(A^c) = 1$.

Proof: By definition of complementary event, $A \cup A^c = \mathcal{S}$. They are also disjoint events, since one cannot happen it the other one happened. It is therefore possible to apply Theorem 2.1.2 and state that $\sum_{i=1}^{2} P(A_i) = P(A) + P(A^c) = P(A \cup A^c)$. But, as stated, $A \cup A^c = \mathcal{S}$, and by axiom 2 $P(\mathcal{S}) = 1$. Therefore, by transitive property, $P(A) + P(A^c) = 1$. □

Theorem 2.1.4: For any event A, $0 \leq P(A) \leq 1$.

Proof: By Theorem 2.1.3, $P(A) + P(A^c) = 1$. By axiom 1, both probabilities are greater or equal than 0, therefore, for the equality to hold, both probabilities have to be lower or equal than 1. Combining the two boundaries, $0 \leq P(A) \leq 1$. □

Theorem 2.1.5: For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: By set algebra, the event $A \cup B$ can itself be seen as the union of two disjoint events, A and $A^c \cap B$. It is therefore possible to apply Theorem 2.1.2, resulting in:

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

In the same fashion, the event B can be seen as the union of the disjoint events $A \cap B$ and $A^c \cap B$. Applying Theorem 2.1.2 gives:

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

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

Theorem 2.1.6: For any three events A, B and C:

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

Proof: Works similarly as Theorem 2.1.5. □

Theorem 2.1.7 (Boole's inequality): Given any countable set of events $A_1, A_2, ..., A_n$:

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

It should be stressed that the Kolmogorov axioms simply describe the rules by which probability works, but do not define the probability of any event itself. Infact probabilities can be assigned to any event in any possible way that is constrained by the axioms, but such value can have no bare on reality or on intuition and yet construct a model that is consistent.

Exercise 2.1.4: Provide an example of a probability model that constrasts with reality but obeys Kolmogorov's axioms.

Solution: Consider the toss of a coin. Such action can be conceived as an experiment, since whose side the coin is gonna land when tossed is unknown until the coin lands. Only two events are possible, heads or tails; since a coin cannot land on both sides at the same time, such events are disjoint.

It is a known fact that the probability of both events is 0.5, and indeed such assignment respects all of three axioms. But by choosing the assignment, say, 0.2 to the landing on heads and 0.8 to the landing on tails, no axiom is violated, even though such an assignment has very little resonance with experience or common sense.

This does not mean that probabilities can be assigned at libitum, since they still ought to comply with the axioms. For example, assigning 0.4 to the probability of the coin to land on heads and 0.3 to the probability of the coin to land on tails won't do, since axiom 2 would be violated. As another example, assigning 1.5 to the probability of the coin to land on heads and -0.5 to the probability of the coin to land on tails would violate axiom 1, and therefore invalid. □

The appropriate or correct assignment depends on how one *interprets* probability, that is to say how one intends the link between the mathematical treatment of probability and the physical world. This quest is just as philosophical as mathematical.

The oldest definition of probability, also called **classical probability**, states that the probability of an event A is the ratio of the number of favorable events and the entire number of outcomes, assuming all outcomes are equally likely.

Said definition is considered outdated mainly for two reasons. On the one hand, the definition lies on a circular argument, since it presupposes the notion of "equally likely outcomes". On the other hand, an event does not necessarely have many outcomes all of them equally likely.

One possible and often invoked interpretation of probability is the **objective** interpretation, also called **frequentist** interpretation. Consider an experiment that can be repeatedly performed in an identical and independent fashion, and let A be an event consisting of a fixed set of outcomes of the experiment. If the experiment is performed n times, the event A will occur $n(A)$ times (with $0 \leq n(A) \leq n$) and will not occur $n - n(A)$ times. The ratio $n(A)/n$ is called the **relative frequency** of occurrence of the event 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 objective interpretation of probability states that this limiting relative frequency is indeed the probability of A to occurr.

This interpretation of probability is said to be objective in the sense that it rests on a property of the experiment 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, and therefore the same probability).

This interpretation has limited appliability, since not all events can be performed n number of times to draw similar conclusions. In situations such as these, it makes more sense to interpret probability in a **subjective** way, which can be thought of as the "degree of confidence" with which an agent believes an event to occur.

The simplest situation to model is the one where to each simple event $E_1, E_2, ..., E_N$ is assigned the same value of probability $P(E_i)$:

$$1 = \sum_{i=1}^{N} P(E_i) \Rightarrow P(E_i) = \frac{1}{N}$$

That is, if there are N equally likely outcomes, the probability of one of such outcomes to happen is $1/N$.

More generally, consider an event A containing $N(A)$ number of outcomes. Then the task of computing probabilities reduces itself to **counting**:

$$P(A) = \sum_{E_i \in A} P(E_i) = \sum_{E_i \in A} \frac{1}{N} = \frac{N(A)}{N}$$

Given two events A and B with $P(B) > 0$, the probability of A to occur given that B occurred is called the **conditional probability** of A given B, and is given as:

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

Theorem 2.1.8 (Law of total probability): Let $A_1, A_2, ..., A_n$ be a finite partition of a sample space $\mathcal{S}$ such that no event has assigned zero probability, and let B be any event in $\mathcal{S}$. Then:

$$P(B) = P(B \mid A_1)P(A_1) + ... + P(B \mid A_n)P(A_n) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

Theorem 2.1.9 (Bayes' theorem): Let $A_1, A_2, ..., A_n$ be a finite partition of a sample space $\mathcal{S}$. Each event A_j has a probability $P(A_j)$, also called its **prior probability**, that is non zero. Let B be any event in $\mathcal{S}$ whose probability is non zero. The probability $P(A_j \mid B)$, also called the **posterior probability**, is given as:

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

Exercise 2.1.5: 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?
3. If a customer returns to the store with a DVD player that needs warranty repair work, what is the probability that it is a brand 1 DVD player?

Consider two events, A and B, the second happening after the first. The fact that B occurred may or may not influence the probability of A to occur. If the probability of A to happen is the same whether or not B happened,

that is to say if $P(A)$ and $P(A \mid B)$ are equal, The event A is said to be **independent** of B. Otherwise, it's said to be **dependent** of B.

Theorem 2.1.10: Event independence is symmetric. In other words, given two events A and B, if A is independent of B, then B is independent of A.

Proof: If A is independent of B, then $P(A \mid B) = P(A)$. Applying Theorem 2.1.9 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, by definition, means that B is independent of A as well. □

An equivalent definition of independent events is as follows. Given two independent events A and B, by the previous definition $P(A) = P(A \mid B)$, so:

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

Event independence can be extended to a situation with more than two events. Given a collection of n events $A_1, A_2, ..., A_n$, such events are said to be **mutually independent** if for every $k = 2, 3, ..., n$ and for every subset of indices $i_1, i_2, ..., i_k$:

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

Kolmogorov axioms define the properties of probability but do not offer a method for assigning them to events. The simplest approaches, such as assigning the same probability to each event, are far to weak to model reality. Also, the sample space and the events are different from experiment to experiment, making it hard to generalize a coherent theory of probability. A more powerful concept to be introduced which can help model probability is the **random variable**.

A random variable can be conceived as a mapping from the sample space to the real line. In other words, a random variable is a function that assigns a probability to any possible event of the sample space. Given a sample space $\mathcal{S}$, a random variable X for such sample space is defined as $X : \mathcal{S} \mapsto \mathbb{R}$, and the probability of such variable to assume a certain value x of the sample space, called its **realization**, is denoted as $P(X = x)$.

Exercise 2.1.6: Suppose a real estate investment has been carried out, and there are three apartments of different value ready to be sold. Assuming that the probability of selling one, two or three of those is the same, model this situation using a random variable.

Solution: It is possible to model this scenario using a random variable X, whose realizations correspond to which and how many apartments were sold. Each realization can be conceived as a triple (a_1, a_2, a_3), where each a_i has value 1 if the i-th apartment was sold and 0 otherwise:

$$\Omega = \{(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)\}$$

The realizations of X correspond to the events with a matching number of 1s. To each value of X is then possible to assign a value of probability by equally dividing the total probability among the events, such that $P(\Omega) = 1$.

$X(0,0,0) = 0$	$P(X = 0) = P(\{(0,0,0)\}) = 12.5\%$
$X(0,0,1) = X(1,0,0) = X(0,1,0) = 1$	$P(X = 1) = P(\{(0,0,1), (1,0,0), (0,1,0)\}) = 37.5\%$
$X(1,1,0) = X(0,1,1) = X(1,0,1) = 2$	$P(X = 2) = P(\{(1,1,0), (0,1,1), (1,0,1)\}) = 37.5\%$
$X(1,1,1) = 3$	$P(X = 3) = P(\{(1,1,1)\}) = 12.5\%$

□

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.

The set of values that a random variable can assume is called its **support**.

2.2. Combinatorics

Combinatorics is an area of mathematics primarily concerned with counting, that is the enumeration of the possible arrangements or configurations of specified structures.

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 2.2.1: 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. Therefore, the experiment as a whole has $6 \cdot 2 = 12$ possible outcomes. Indeed:

$$\Omega_1 = \{H, T\}$$

$$\Omega_2 = \{1, 2, 3, 4, 5, 6\}$$

$$\Omega = \{\{H,1\},\{T,1\},\{H,2\},\{T,2\},\{H,3\},\{T,3\},\{H,4\},\{T,4\},\{H,5\},\{T,5\},\{H,6\},\{T,6\}\}$$

□

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:

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

Exercise 2.2.2: 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 2.2.3: 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 2.2.4: 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:

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

Exercise 2.2.5: 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:

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

□

3. Random variables

3.1. Discrete random variables

The **probability mass function** (abbreviated as pmf) of a discrete random variable X, denoted as $p(X)$, is a function that assigns a probability to each possible value that such random variable can assume. More formally, given a random variable X, for each value x of its sample space the pmf of X is defined as:

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

The **cumulative distribution function** (abbreviated as cdf) of a discrete random variable X, denoted as $F(X)$, is defined as the probability of such random variable to assume a value less than or equal to a threshold. More formally, given a random variable X, for each value x of its sample space the cdf of X is defined as:

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

Let X be a discrete random variable with support D and probability mass function $p(X)$. The **expected value** (or **mean value**) of X, denoted as $E(X)$ or μ_X is given by:

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

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

The expected value of a random variable is the equivalent of the mean with respect to populations.

Exercise 3.1.1: Let X be the Apgar score of a randomly selected child born at a certain hospital during the next year, with pmf as follows:

X	0	1	2	3	4	5	6	7	8	9	10
$p(X)$	0.002	0.001	0.002	0.005	0.02	0.04	0.18	0.37	0.25	0.12	0.01

What is the expected value of X?

Solution:

$$E(X) = \sum_{x \in D} x \cdot p(x) = 0 \cdot 0.002 + 1 \cdot 0.001 + 2 \cdot 0.002 + 3 \cdot 0.005 + 4 \cdot 0.02 + 5 \cdot 0.04 + 6 \cdot 0.18 + 7 \cdot 0.37 + 8 \cdot 0.25 + 9 \cdot 0.12 + 10 \cdot 0.01 = 7.15$$

□

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 and probability mass function $p(X)$, and let $h(X)$ be a function whose argument is (the realization of) the random variable X. The expected value of $h(X)$ is still defined as:

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

Theorem 3.1.1: Let X be a discrete random variable with support D and probability mass function $p(X)$. Given two coefficients a and b, the following equality holds:

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

Proof: Let D be the support of X and $p(X)$ its probability mass function. Let Y be the random variable $Y = aX + b$. The probability of Y to assume a particular value $ax + b$ does not depend on a and b, but only on x. Therefore, $P(Y = ax + b) = P(X = x)$. Computing the expected value for Y gives:

$$E(Y) = \sum_{x \in D} (ax + b) \cdot P(Y = ax + b) = \sum_{x \in D} (ax + b) \cdot P(X = x) = \sum_{x \in D} ax \cdot P(X = x) + b \cdot P(X = x) =$$
$$a \sum_{x \in D} x \cdot P(X = x) + b \sum_{x \in D} P(X = x) = aE(X) + b \cdot 1 = aE(X) + b$$

Where $\sum_{x \in D} P(X = x) = 1$ stems from the fact that said summation is the sum of the probabilities of the entire sample space, which is 1 by definition. □

Theorem 3.1.2: Let X and Y be two random variables. $E[X + Y] = E[X] + E[Y]$.

Let X be a discrete random variable with support D and probability mass function $p(X)$. The **variance** of X, denoted as $V(X)$ or σ_X^2 is given by:

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

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

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

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

The variance and the standard deviation measure how a random variable is "spread out", in the sense of how much the values of the support of said variable are detached from its expected value.

Lemma 3.1.1: Let X be a discrete random variable with support D and probability mass function $p(X)$. The following equality holds:

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

Theorem 3.1.3: Let X be a discrete random variable with support D and probability mass function $p(X)$. Given two coefficients a and b, the following equality holds:

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

Proof: Let Y be the random variable $Y = aX + b$. From Theorem 3.1.1, $E(Y) = E(aX + b) = aE(X) + b$. Substituting this expression in the variance one gives:

$$V(Y) = E((Y - E(Y))^2) = E((ax + \cancel{b} - aE(X) - \cancel{b})^2) = \sum_{x \in D} (ax - aE(X))^2 P(Y = ax + b) =$$
$$\sum_{x \in D} a^2 (x - E(X))^2 P(X = x) = a^2 \sum_{x \in D} (x - E(X))^2 P(X = x) = a^2 V(X)$$

□

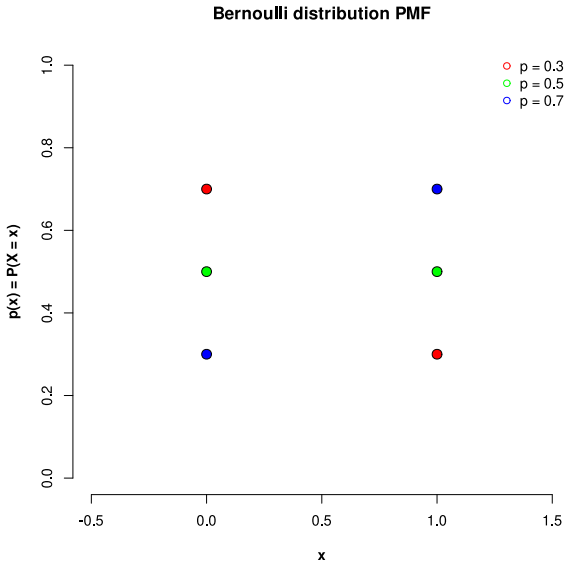
Theorem 3.1.4: Let X and Y be two random variables. $V[X+Y] = V[X] + V[Y]$ if X and Y are independent.

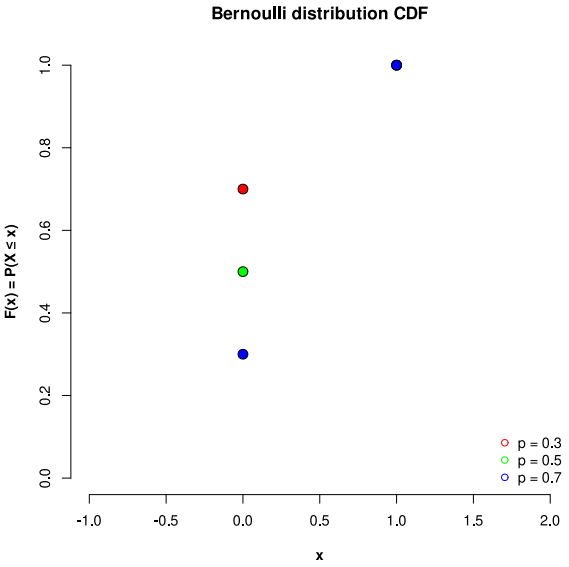
3.2. Known discrete random variables

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.

3.2.1. 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 3.2.1.1: 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:

$$E(X) = 0 \cdot (1-p) + 1 \cdot p = 0 + p = p$$

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

□

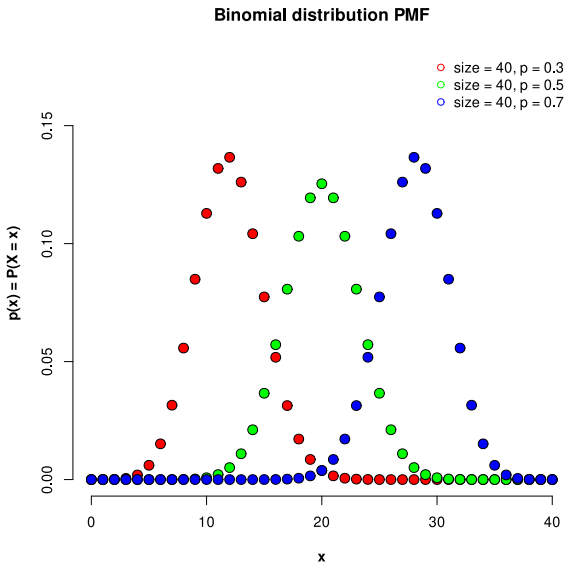
3.2.2. Binomial random variable

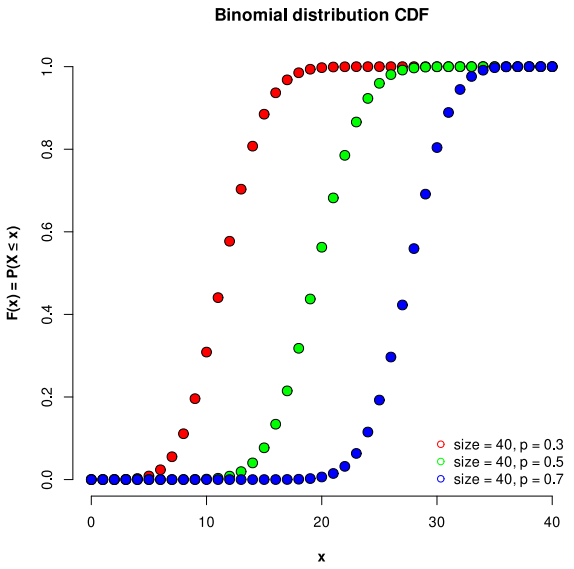
Let $Y_1, Y_2, ..., Y_n$ be n independent and identically distributed Bernoulli random variables (all having the same parameter p). Let X be the random variable defined as the sum of all said variables:

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

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:





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

Theorem 3.2.2.1: 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 3.1.2 and Theorem 3.1.4 (the latter can be applied since the Bernoulli random variables that constitue X are independent).

$$E(X) = E(Y_1 + Y_2 + ... + Y_n) = E(Y_1) + E(Y_2) + ... + E(Y_n) = nE(Y_i) = np$$

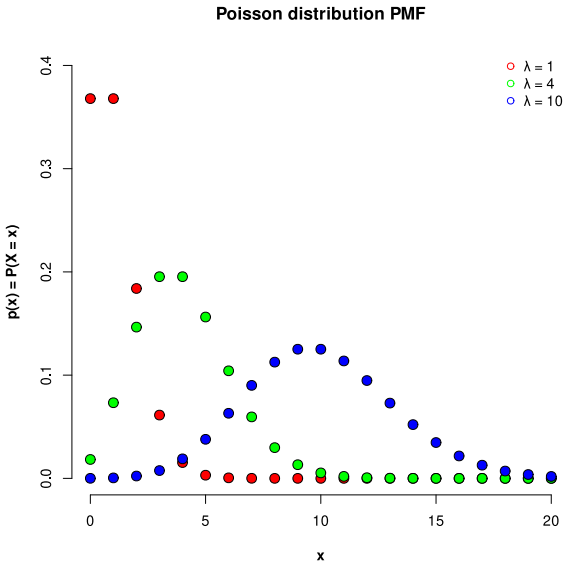
$$V(X) = V(Y_1 + Y_2 + ... + Y_n) = V(Y_1) + V(Y_2) + ... + V(Y_n) = nV(Y_i) = np(1-p)$$

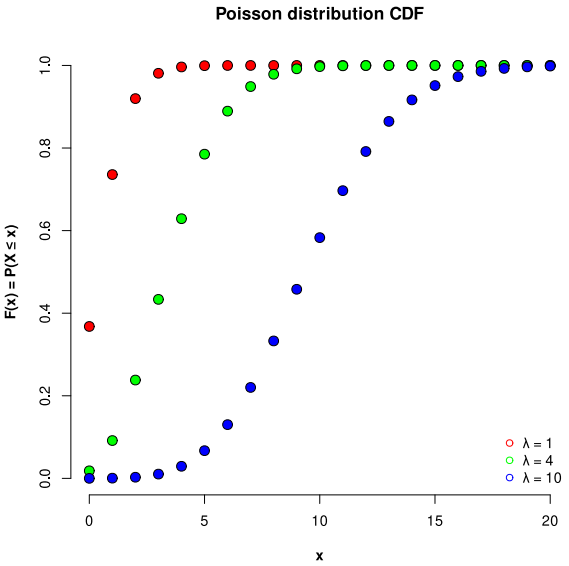
Where $E(Y_i)$ and $V(Y_i)$ are retrieved from Theorem 3.2.1.1. □

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

3.2.3. 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 3.2.3.1: 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:

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

□

Exercise 3.2.3.1: 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.

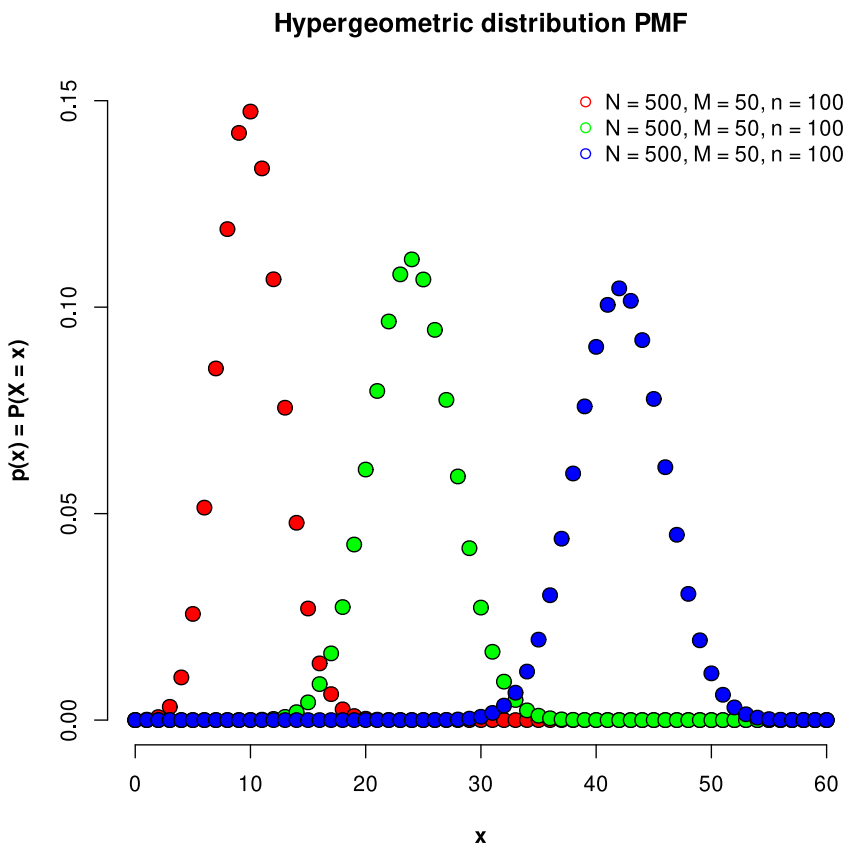
The probability mass function of a Poisson distribution can be adapted in this sense if, instead of the expected value λ, one is given α, the expected number of events occurring in a unitary time interval, and a time interval Δt. The probability of k events to occur in a time slice Δt is then as follows:

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

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

3.2.4. 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 3.2.4.1: 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:

$$p(2) = P(X = 2) = \frac{\binom{12}{2}\binom{20-12}{5-2}}{\binom{20}{5}} = \frac{\frac{12!}{2!(12-2)!}\frac{8!}{3!(8-3)!}}{\frac{20!}{5!(20-5)!}} = \frac{\frac{12 \cdot 11 \cdot \cancel{10!}}{2 \cdot \cancel{10!}}\frac{8 \cdot 7 \cdot 6 \cdot \cancel{5!}}{6 \cdot \cancel{5!}}}{\frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot \cancel{15!}}{120 \cdot \cancel{15!}}} =$$

$$\frac{\cancel{12} \cdot 11 \cdot 8 \cdot 7 \cdot 6}{\cancel{12}} \cdot \frac{\cancel{20} \cdot 6}{\cancel{20} \cdot 19 \cdot 18 \cdot 17 \cdot 16} = \frac{22176}{93024} \approx 0.238$$

□

Theorem 3.2.4.1: 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.

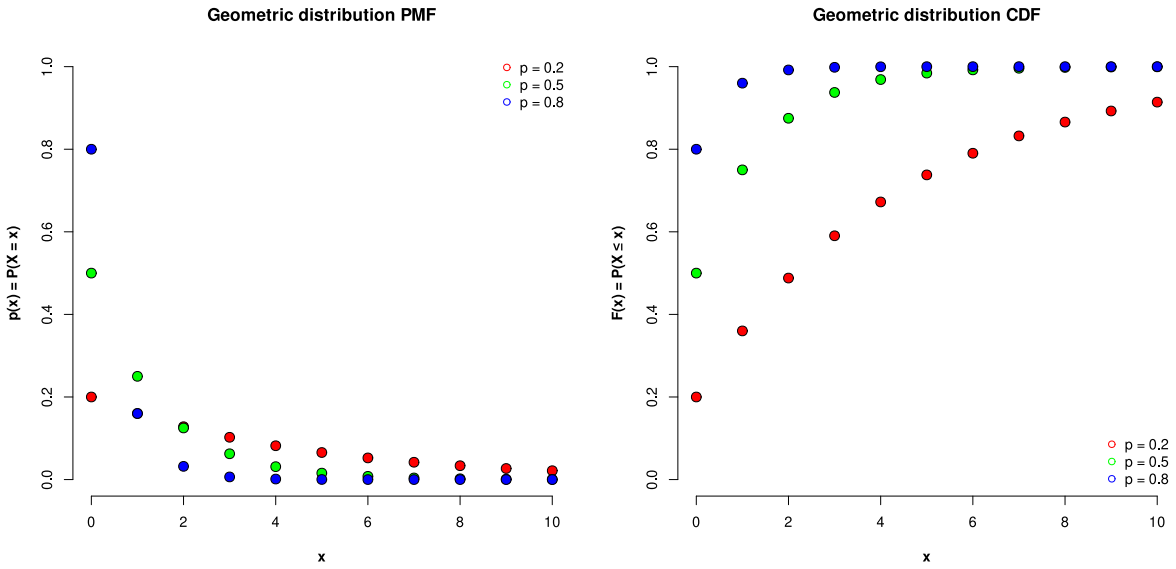
Another similarity between the two comes from observing the equations is Theorem 3.2.4.1. The ratio M/N is the proportion of successes in the population, meaning that it's the probability of picking an element of the entire population that has a value 1. This ratio has the same role that the parameter 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 3.2.2.1), 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 $p = M/N$.

3.2.5. 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 3.2.5.1: 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:

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

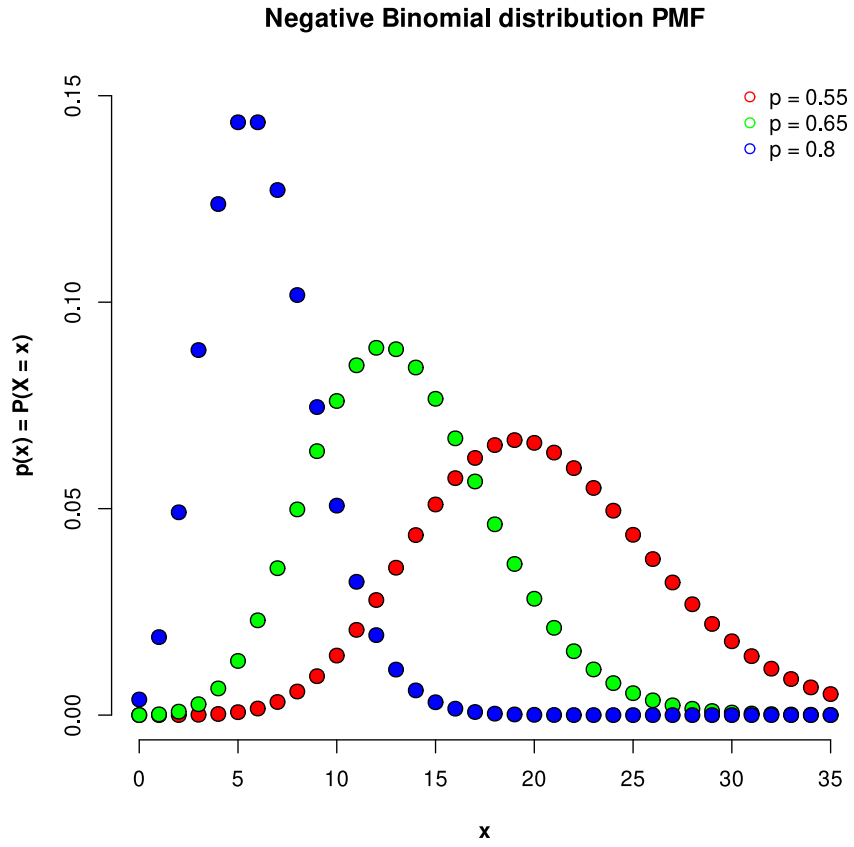
Then, applying Lemma 3.1.1:

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

□

3.2.6. 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:

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

3.3. Continuous random variables

The **probability density function** (abbreviated as pdf) of a continuous random variable X is a function $f(x)$ such that, for any pair of numbers a and b:

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

That is, the probability of the random variable X to assume a value that lies in the $[a, b]$ interval is equal to the integral of $f(x)$ with said points as extremes, which is also the area under the curve of $f(x)$ in the interval $[a, b]$.

Any function can be a pdf as long as it satisfies the following two conditions:

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

Exercise 3.3.1: Suppose that the chanche of having imperfection on the surface of a tire can be described as a random variable. Let X be the angle induced by any vector that goes from the centre of the tire to its surface. Suppose the probability distribution function of X is given by:

$$f(x) = \begin{cases} \frac{1}{360} \text{ if } 0 \leq x \leq 360 \\ 0 \qquad \text{otherwise} \end{cases}$$

Is this probability distribution function well-defined? If it is, what is the probability of finding an imperfection between the angles 90 and 180?

Solution: This pdf can assume either 0 or $\frac{1}{360}$ as its values, therefore the first condition is satisfied. As for the second condition:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{360} \frac{1}{360}dx + \int_{0}^{+\infty} 0dx = \frac{1}{360}\int_{0}^{360} 1dx = \frac{1}{360}(360 - 0) = \frac{360}{360} = 1$$

The requested probability is:

$$P(90 \leq X \leq 180) = \int_{90}^{180} \frac{1}{360}dx = \frac{1}{360}\int_{90}^{180} 1dx = \frac{1}{360}(180 - 90) = \frac{90}{360} = 0.25$$

□

The **cumulative distribution function** (abbreviated as cdf) of a continuous random variable X, denoted as $F(X)$, is defined as the probability of such random variable to assume a value less than or equal to a threshold. Given a random variable X, its cdf is defined as:

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

Given $F(a)$, the probability of a random variable to assume a value less than or equal to a, it is possible to compute $P(X > a)$, the probability of said random variable to assume a value greater than a, since the two events are the complement of each other:

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a)$$

Lemma 3.3.1: Let X be a random variabile, and let a and b be two real numbers (suppose $a < b$). The following holds:

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

Proof: By definition, the events A = "the realization of X lies in $(-\infty, a]$" and B = "the realization of X lies in $(a, b]$" are incompatible. Therefore:

$$P(-\infty < X \leq a) + P(a < X \leq b) - P(\{-\infty < X \leq a\} \cap \{a < X \leq b\}) =$$
$$P(-\infty < X \leq a) + P(a < X \leq b) - 0 = P(\{-\infty < X \leq a\} \cup \{a < X \leq b\})$$

From which stems:

$$F(b) = P(-\infty < X \leq b) = P(\{-\infty < X \leq a\} \cup \{a < X \leq b\}) =$$
$$P(-\infty < X \leq a) + P(a < X \leq b) = F(a) + P(a < X \leq b) \Rightarrow F(b) - F(a) = P(a < X \leq b)$$

□

Corollary 3.3.1: The probability of a continuous random variable X to assume a specific value c is always 0.

Proof: $P(X = c)$ is equivalent to $P(c < X \leq c)$. Therefore Lemma 3.3.1 applies:

$$P(c < X \leq c) = F(c) - F(c) = 0$$

□

Corollary 3.3.1 is sound even from a logical perspective: the probability of choosing a specific value in (a subset of) $\mathbb{R}$ is necessarely infinitesimal, since $\mathbb{R}$ is not a countable set. Therefore, when dealing with the realization of a continuous random variables, it only makes sense to think in terms of intervals.

Let X be a continuous random variable probability distribution function $f(x)$. The **expected value** (or **mean value**) of X, denoted as $E(X)$ or μ_X is given by:

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

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

Let X be a continuous random variable probability distribution function $f(x)$. The **variance** of X, denoted as $V(X)$ or σ_X^2 is given by:

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

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

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

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

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 cumulative distribution function of X evaluated at $\eta(p)$.

Exercise 3.3.2: 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:

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

What is the cdf of X? What are the expected value and the variance of X? What is the 50-th percentile?

Solution: The cdf of X is clearly 0 for any x lower than 0 and 1 for any x greater than 1. As for $x \in [0, 1]$:

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

The expected value is given by computing the integral:

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

As for the variance:

$$V(X) = E(X^2) - (E(X))^2 = -\left(\frac{3}{8}\right)^2 + \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx = -\frac{9}{64} + \int_{-\infty}^{0} x^2 \cdot f(x)dx +$$
$$\int_{0}^{1} x^2 \cdot f(x)dx + \int_{1}^{+\infty} x^2 \cdot f(x)dx = -\frac{9}{64} + 0 + \int_{0}^{1} x^2 \cdot \frac{3}{2}(1-x^2)dx + 0 = -\frac{9}{64} + \frac{3}{2}\int_{0}^{1} x^2 - x^4 dx =$$
$$-\frac{9}{64} + \frac{3}{2}\left(\int_{0}^{1} x^2 dx - \int_{0}^{1} x^4 dx\right) = -\frac{9}{64} + \frac{3}{2}\left(\left(\frac{1^3}{3} - \frac{0^3}{3}\right) - \left(\frac{1^5}{5} - \frac{0^5}{5}\right)\right) = -\frac{9}{64} + \frac{3}{2}\left(\frac{1}{3} - \frac{1}{5}\right) = \frac{19}{320}$$

The 50-th percentile is given when $p = 0.5$. Applying the formula gives:

$$0.5 = F(\eta(p)) = \frac{1}{2}\eta(p)(3 - (\eta(p))^2) \Rightarrow (\eta(p))^3 - 3\eta(p) + 1 = 0 \Rightarrow \eta(p) \approx 0.347$$

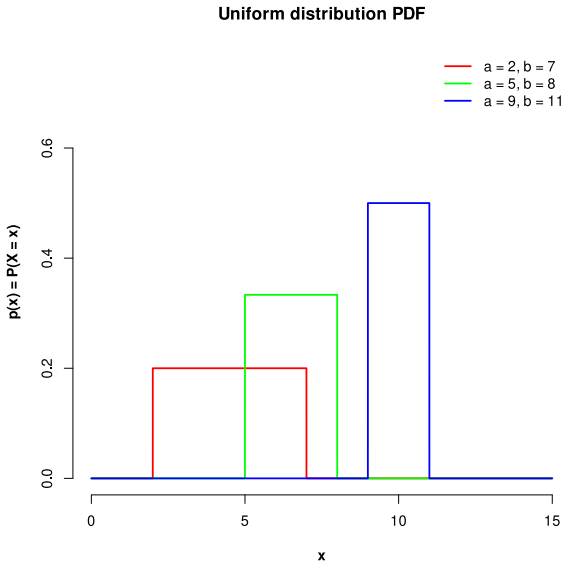
□

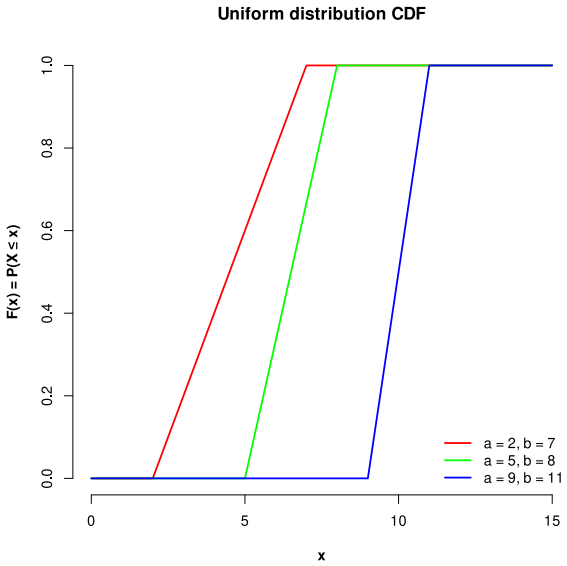
3.4. Known continuous random variables

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.

3.4.1. 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 3.4.1.1: The expected value and variance of a random variable $X \sim U(a,b)$ are as follows:

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

Proof:

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

$$0 + \int_{a}^{b} x\left(\frac{1}{b-a}\right)dx + 0 = \frac{1}{b-a}\int_{a}^{b} xdx = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{1}{\cancel{(b-a)}}\frac{\cancel{(b-a)}(b+a)}{2} = \frac{b+a}{2}$$

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

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

$$\frac{1}{b-a}\int_{a}^{b} x^2 + \frac{(b+a)^2}{4} - (b+a)xdx = \frac{1}{b-a}\left(\int_{a}^{b} x^2dx + \int_{a}^{b} \frac{(b+a)^2}{4}dx - \int_{a}^{b} (b+a)xdx\right) =$$

$$\frac{1}{b-a}\int_{a}^{b} x^2dx + \frac{(b+a)^2}{4(b-a)}\int_{a}^{b} 1dx - \frac{b+a}{b-a}\int_{a}^{b} xdx = \frac{1}{b-a}\left(\frac{b^3}{3} - \frac{a^3}{3}\right) + \frac{(b+a)^2}{4\cancel{(b-a)}}\cancel{(b-a)} -$$

$$\frac{b+a}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{1}{\cancel{(b-a)}}\frac{\cancel{(b-a)}(b^2+ba+a^2)}{3} + \frac{(b+a)^2}{4} - \frac{b+a}{\cancel{(b-a)}}\frac{\cancel{(b-a)}(b+a)}{2} =$$

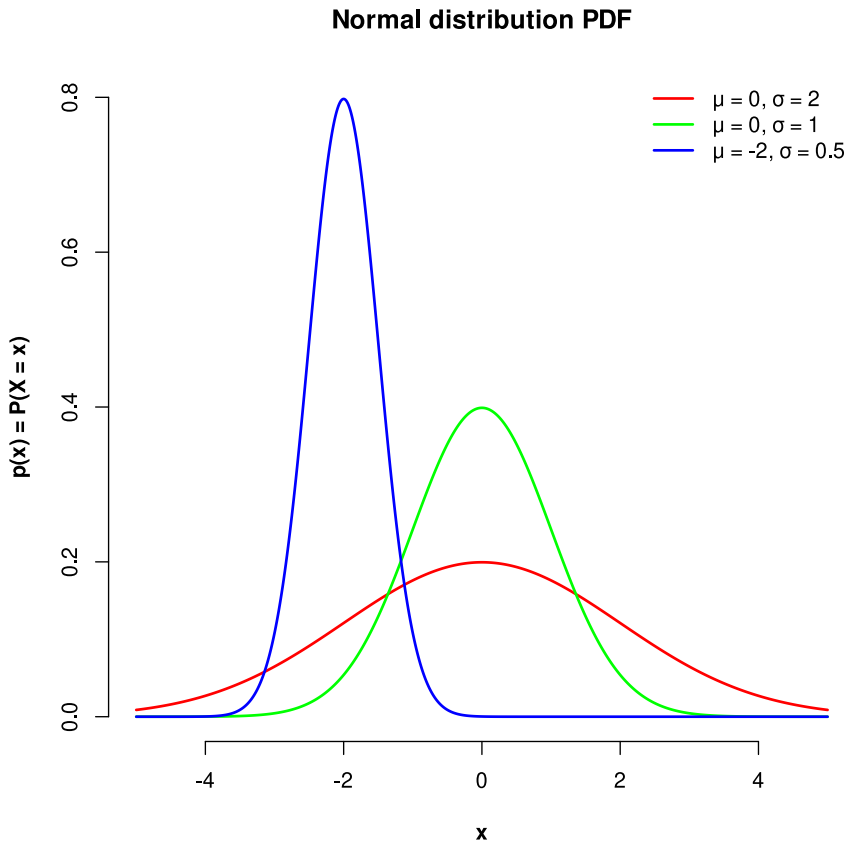
$$\frac{(b+a)^2 - ba}{3} + \frac{(b+a)^2}{4} - \frac{(b+a)^2}{2} = \frac{4(b+a)^2 - 4ba + 3(b+a)^2 - 6(b+a)^2}{12} =$$

$$\frac{(b+a)^2 - 4ba}{12} = \frac{b^2 + a^2 + 2ba - 4ba}{12} = \frac{b^2 + a^2 - 2ba}{12} = \frac{(b-a)^2}{12}$$

□

3.4.2. 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 3.4.2.1: 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:

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

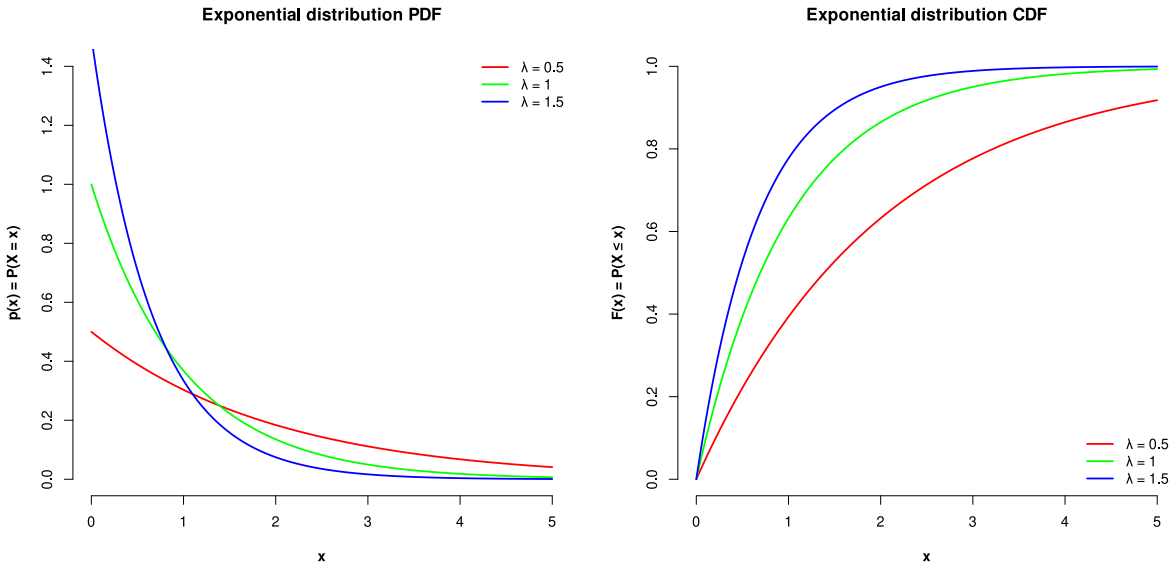
$$\Phi(1.09) - \Phi(-0.54) = \Phi(1.09) - (1 - \Phi(0.54)) = 0.8621 - 0.2946 = 0.5675$$

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

□

3.4.3. 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:



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

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

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

Proof:

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

$$-\int_{0}^{+\infty} x\left(-\lambda e^{-\lambda x}\right)dx = -\left[x e^{-\lambda x}\right]_{0}^{\infty} + \int_{0}^{+\infty} e^{-\lambda x}dx = -\left(\infty \cdot e^{-\lambda \cdot \infty} - 0 \cdot e^{-\lambda \cdot 0}\right) + \int_{0}^{-\infty} \frac{e^{u}}{-\lambda}du =$$

$$-0 - \frac{1}{\lambda}\int_{0}^{-\infty} e^{u}du = \frac{1}{\lambda}\int_{-\infty}^{0} e^{u}du = \frac{1}{\lambda}(e^{0} - e^{-\infty}) = \frac{1}{\lambda}(1 - 0) = \frac{1}{\lambda}$$

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

$$\int_{0}^{+\infty} x^2 \cdot f(x)dx = -\frac{1}{\lambda^2} + 0 + \int_{0}^{+\infty} \lambda x^2 e^{-\lambda x}dx = -\frac{1}{\lambda^2} - \int_{0}^{+\infty} x^2\big(-\lambda e^{-\lambda x}\big)dx =$$

$$-\frac{1}{\lambda^2} - \left[x^2 e^{-\lambda x}\right]_0^{\infty} + \int_{0}^{+\infty} 2xe^{-\lambda x}dx = -\frac{1}{\lambda^2} - \left(\infty^2 \cdot e^{-\lambda \cdot \infty^2} - 0^2 \cdot e^{-\lambda \cdot 0^2}\right) -$$

$$\frac{2}{\lambda}\int_{0}^{+\infty} x\big(-\lambda e^{-\lambda x}\big)dx = -\frac{1}{\lambda^2} - 0 - \frac{2}{\lambda}\int_{0}^{+\infty} x\big(-\lambda e^{-\lambda x}\big)dx = -\frac{1}{\lambda^2} - \frac{2}{\lambda}\left[xe^{-\lambda x}\right]_0^{\infty}$$

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

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

□

Exercise 3.4.3.1: 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 3.4.3.2: The geometric distribution function has the memorylessness property.

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

$$P(X \geq t + t_0 \mid X \geq t_0) = \frac{P[(X \geq t + t_0) \cap (X \geq t_0)]}{P(X \geq t_0)} = \frac{P(X \geq t + t_0)}{P(X \geq t_0)} =$$

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

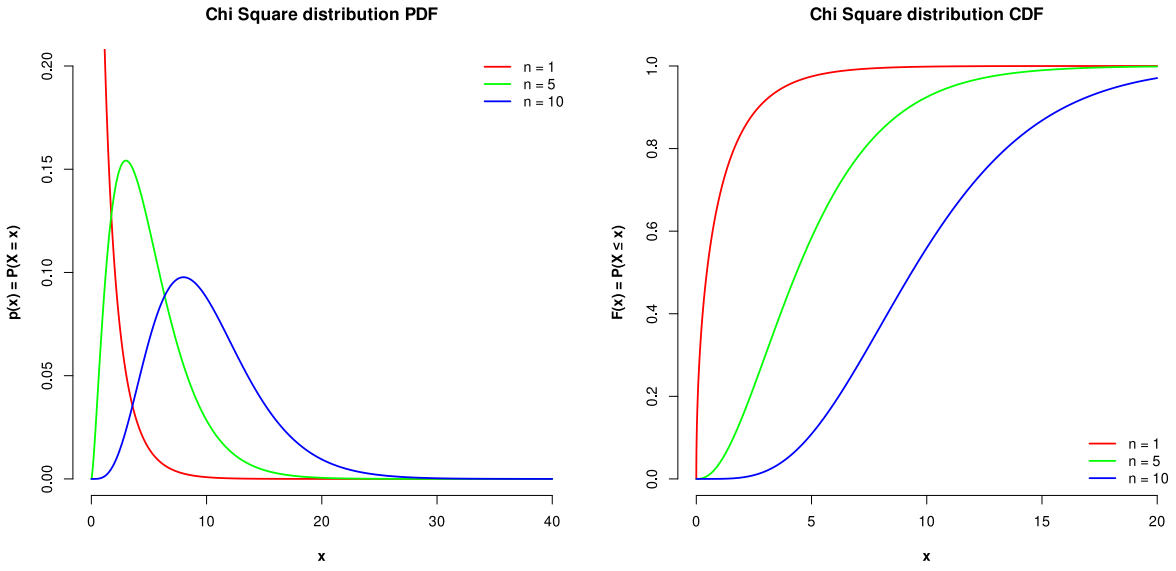
□

3.4.4. Chi-square 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 3.4.4.1: The expected value and variance of a random variable $X \sim \chi^2(n)$ are as follows:

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

Proof: The expected value of X can be computed by applying Theorem 3.1.2 and Lemma 3.1.1:

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

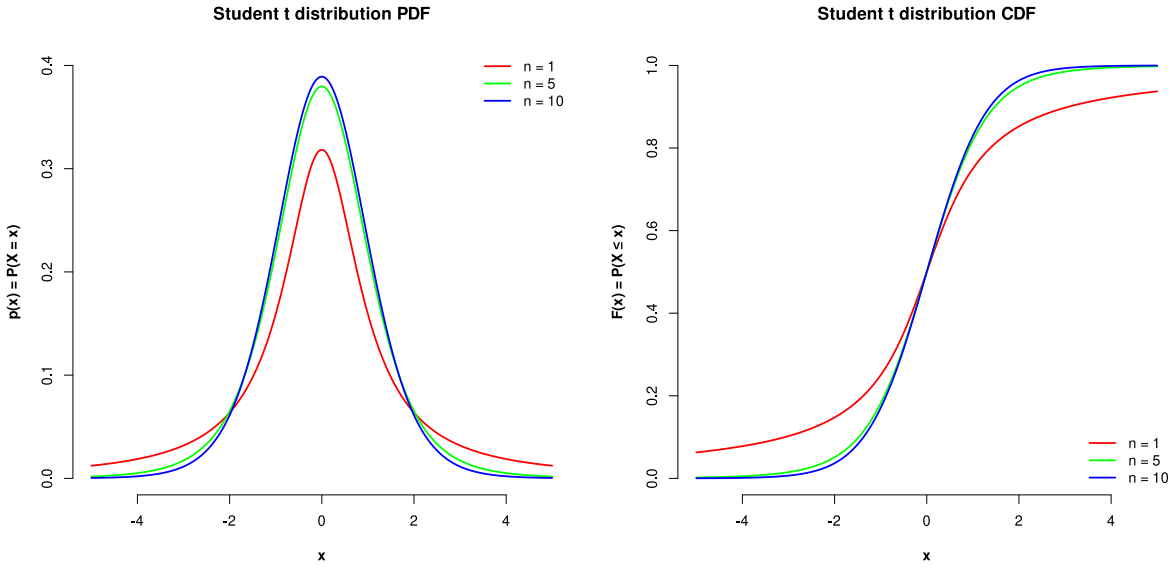
□

3.4.5. Student t random variable

Given an integer n, let Z be a standard normal distribution and Y_n a Chi-square 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 3.4.5.1: 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.

3.5. 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 3.5.1: 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:

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

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

$$\sum_{y \in D(Y)} p(X = 1000, Y) = P(X = 1000, Y = 500) + P(X = 1000, Y = 1000) + P(X = 1000, Y = 5000) = 0.1 + 0.1 + 0.05 = 0.25$$

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

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

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

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

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

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:

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

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

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

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 3.5.2: Consider Exercise 3.5.1. 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. □

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 3.5.3: Consider Exercise 3.5.1. 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 3.5.1: Let X and Y be two random variables, and let a and b be two real numbers. The following equality holds:

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

Theorem 3.5.2: 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 3.5.3: 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 3.5.3, it is bounded, and therefore is indicative of the scale of the values of the variables.

Theorem 3.5.4: 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 3.5.4 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 3.5.1: If X and Y are two independent random variables, $\mathrm{Corr}(X, Y) = 0$.

Theorem 3.5.5: 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 3.1.1 and Theorem 3.1.2 gives:

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

Applying Theorem 3.1.3 and Theorem 3.1.4 gives:

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

□

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:

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

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

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:

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

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

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

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

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

Sample variance:

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

$$= P(\{X_1 = 80\} \wedge \{X_2 = 100\} \vee \{X_1 = 100\} \wedge \{X_2 = 80\} \vee \{X_1 = 100\} \wedge \{X_2 = 120\} \vee X_1 = 100\} \wedge \{X_2 = 120\}) = 0.06 + 0.06 + 0.15 -$$

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

□

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:

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

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[1] 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:

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

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

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:

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

□

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:

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

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:

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

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.

[1] 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.

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 $X_1=x_1\wedge X_2=x_2\wedge\ldots\wedge X_n=x_n$ 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:

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

$$\frac{d}{d\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)=0\Rightarrow0+0-\cancel{\frac{1}{2\sigma^2}}\frac{d}{d\mu}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)=0\Rightarrow\sum_{i=1}^{n}\frac{d}{d\mu}(x_i-\mu)^2=0\Rightarrow$$

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

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[2]:

$$P\left(\overline{x} - z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{x} + z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

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.

[2]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:

$$P(\text{Type I Error}) = P(\text{p-value} \leq \alpha \mid H_0) = P(F(y) \leq \alpha) = P\big[F^{-1}(F(y)) \leq F^{-1}(\alpha)\big] = P\big[Y \leq F^{-1}(\alpha)\big] = P\big[Y \leq F^{-1}(\alpha)\big] = F\big[F^{-1}(\alpha)\big] = \alpha$$

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[3].

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 $Z = \big(\overline{X} - \mu_0/(\sigma/\sqrt{n})\big)$.

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:

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{2 - 2.06}{1.41/\sqrt{51}} = 3.04$$

[3] This is analogous to a trial: it is better to absolve a criminal than to condemn an innocent.

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:

$$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2$$

In the case of $H_1 : \mu \neq \mu_0$, it is still possible to retrieve an approximate solution:

$$n = \left[\frac{\sigma\big(z_{\alpha/2} + z_\beta\big)}{\mu_0 - \mu'}\right]^2$$

5. Regression model

5.1. Simple linear regression

The simplest deterministic mathematical relationship between two variables x and y is a linear relationship of the form $y = \beta_0 + \beta_1 x$. The set of pairs $(x, y) = (x, \beta_0 + \beta_1 x)$ determines a straight line in the cartesian plane having slope β_1 and y-intercept β_0. The assumption that said relationship holds is called the **linear model**; the evaluation of the attendibility of the linear model is called **regression analysis**.

If the two variables are not deterministically related, then for a fixed value of x, there is some uncertainty in the value of y. Suppose that x_i and y_i are the i-th pair drawn; the value observed for y_i will most likely be different from $Y_i = \beta_0 + \beta_1 x_i$, the value for y_i retrieved from the equation. This means that each variable Y_i can be conceived as a random variable, whose value depends on the value of x_i.

More generally, the variable whose value is fixed by the experimenter, in this case x, is called the **independent variable**, **predictor variable**, or **explanatory variable**. The variable whose value depends on x is denoted with y or Y depending on whether the value retrieved from the data or the random variable is being referred to. This variable is called the **dependent variable** or the **response variable**.

Suppose that β_0 and β_1 are known. Denote with ε the **random deviation**, or **random error**, that is the difference between the observed and the predicted value for the second variable. The value of Y is given by the following **model equation**:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

If ε were a null term, each observed pair (x, y) would precisely fall on the line $y = \beta_0 + \beta_1 x$, call the **true regression line** or **population regression line**. Since ε is in general not null, the observed pair (x, y) will most likely fall either above ($\varepsilon > 0$) or below ($\varepsilon < 0$) the true regression line. The smaller the ε, the better the model.

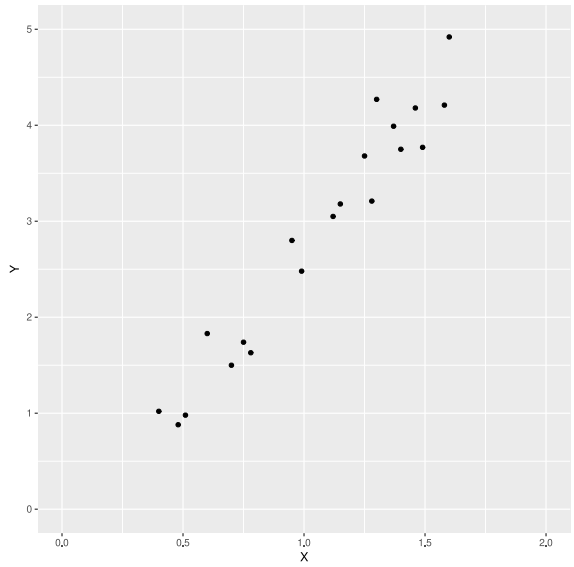
A powerful insight into the linear dependence of two variables is given by the **scatter plot**. This plot denotes each couple (x_i, y_i) as points on a cartesian plane. If the points are distributed roughly as a straight line, it means that the value of y_i can be predicted with reasonable prediction by the value of x_i.

Exercise 5.1.1: Consider the following dataset:

0.40	1.02	0.48	0.88	0.51	0.98	0.60	1.83	0.70	1.50	0.75	1.74	0.78	1.63	0.95	2.80	0.99	2.48	1.12	3.05
1.15	3.18	1.25	3.68	1.28	3.21	1.30	4.27	1.37	3.99	1.40	3.75	1.46	4.18	1.49	3.77	1.58	4.21	1.60	4.92

Draw a scatter plot and draw some conclusions.

Solution:



The distribution of the points on the plane is roughly line-shaped. Therefore, the predictions Y will be really close to the retrieved values y. □

In the context of inferential statistics, the pairs $(x_1, y_1), ..., (x_n, y_n)$ are values extracted from two populations (assuming with random sampling), and therefore the values of β_0 and β_1 are not known a priori, and have instead to be estimated with respect to the values of the sample.

Since the values of Y and ε depend on the value of x, each Y_i and ε_i can be conceived as random variables. To be able to estimate the values of β_0 and β_1 it is reasonable to take the following hypotheses for granted:

- Each ε_i is an i.i.d normal random variable (this also means that each Y_i is i.i.d, but not necessarely normal);
- $E(\varepsilon_i) = 0$ for each $i \in \{1, 2, ..., n\}$;
- $V(\varepsilon_i) = \sigma^2$ for each $i \in \{1, 2, ..., n\}$.

Let b_0 and b_1 be any estimate (no matter how reasonable they are) for β_0 and β_1 respectively. Denote by $b_0 + b_1 x$ the value of the Y variable retrieved by using said estimates for the equation. Out of all (b_0, b_1) couples, the estimates that are of interest are the ones that best represent the actual (unknown) values of β_0 and β_1.

It is reasonable to assume that the best estimates for the two coefficients are the ones that minimize the distance between the each Y_i and each y_i. Keeping in mind how the (Euclidean) distance is computed and summing all of the n distances together, gives:

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, are called the **least square estimates**, and are the choices of β_0 and β_1 that minimize the function $f(b_0, b_1)$. The **estimated regression line**, or **least square line**, is the linear equation whose coefficients are $\hat{\beta}_0$ and $\hat{\beta}_1$.

Those coefficients can be retrieved by computing the partial derivatives of $f(b_0, b_1)$ with respect to both variables and setting them equal to 0:

$$\frac{\partial f(b_0, b_1)}{\partial b_0} \left(\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \right) = 0 \Rightarrow \sum_{i=1}^{n} \frac{\partial f(b_0, b_1)}{\partial b_0} \left((y_i - b_0 - b_1 x_i)^2 \right) = 0 \Rightarrow$$

$$\sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i) \frac{\partial f(b_0, b_1)}{\partial b_0} (y_i - b_0 - b_1 x_i) = 0 \Rightarrow \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i) \frac{\partial f(b_0, b_1)}{\partial b_0} (0 - 1 + 0) = 0 \Rightarrow$$

$$-2 \sum_{i=1}^{n} y_i - b_0 - b_1 x_i = 0 \Rightarrow \sum_{i=1}^{n} y_i - b_0 - b_1 x_i = 0$$

$$\frac{\partial f(b_0,b_1)}{\partial b_1}\left(\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2\right)=0\Rightarrow\sum_{i=1}^{n}\frac{\partial f(b_0,b_1)}{\partial b_1}\left((y_i-b_0-b_1x_i)^2\right)=0\Rightarrow$$

$$\sum_{i=1}^{n}2(y_i-b_0-b_1x_i)\frac{\partial f(b_0,b_1)}{\partial b_1}(y_i-b_0-b_1x_i)=0\Rightarrow\sum_{i=1}^{n}2(y_i-b_0-b_1x_i)\frac{\partial f(b_0,b_1)}{\partial b_1}(0+0-x_i)=0\Rightarrow$$

$$-2\sum_{i=1}^{n}x_i(y_i-b_0-b_1x_i)=0\Rightarrow\sum_{i=1}^{n}x_i(y_i-b_0-b_1x_i)=0$$

Extracting the two equations, called **normal equations**, and solving for b_0 and b_1, gives the two point estimators $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{S_{xy}}{S_{xx}}\qquad\qquad\hat{\beta}_0=\frac{\sum_{i=1}^{n}y_i-\hat{\beta}_1\sum_{i=1}^{n}x_i}{n}=\overline{y}-\hat{\beta}_1\overline{x}$$

Exercise 5.1.2: Consider Exercise 5.1.1. Compute the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ and draw the corresponding estimated regression line.

Solution: Computing $\overline{x}$ and $\overline{y}$ gives:

$$\overline{x}=\frac{0.40+0.48+...+1.60}{20}=\frac{21.16}{20}\approx 1.06\qquad\qquad\overline{y}=\frac{1.02+0.88+...+4.92}{20}=\frac{57.07}{20}\approx 2.86$$

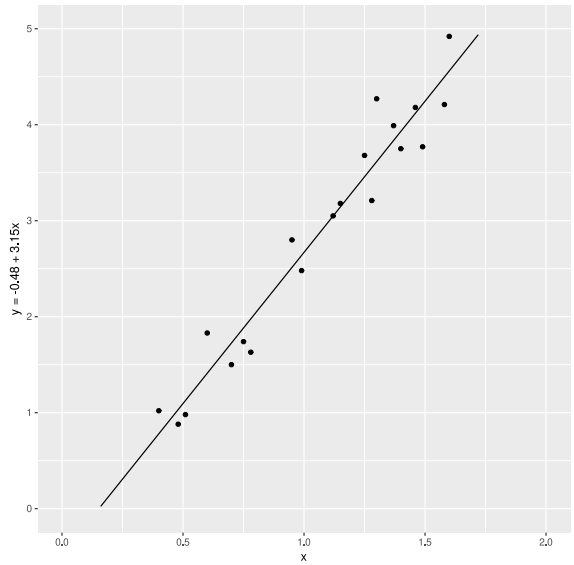
Computing S_{xy} and S_{xx} gives:

$$S_{xy}=\sum_{i=1}^{n}(x_i-1.06)(y_i-2.86)\approx 9.08\qquad\qquad S_{xx}=\sum_{i=1}^{n}(x_i-1.06)^2\approx 2.88$$

The estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$ is therefore:

$$\hat{\beta}_1=\frac{9.08}{2.88}\approx 3.15\qquad\qquad\hat{\beta}_0=2.86-\frac{9.08}{2.88}\cdot 1.06\approx -0.48$$

The estimated regression line is $y=-0.48+3.15x$. Plotting it gives:



□

The estimated regression line can be used to make predictions about values of y fixing a value of x, even if it's not present in the sample. Chosen a certain x', the value $\hat{\beta}_0+\hat{\beta}_1x'$ gives the point estimate of the expected value

of Y when $x = x'$. It should be noted that, since the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the values of the sample, the value $\hat{\beta}_0 + \hat{\beta}_1 x'$ will be a more and more inaccurate prediction for $E(Y)$ as the chosen x' is further and further away from the range of the sample.

Exercise 5.1.3: Consider the estimated regression line from Exercise 5.1.2. Compute the prediction for y' when $x' = 1.25$ and when $x' = 10$

Solution:

$$y' = -0.48 + 3.15 \cdot 1.25 = 3.46 \qquad\qquad y' = -0.48 + 3.15 \cdot 10 = 31.02$$

Even though the first prediction might be reasonable, the other one is much less reliable □

The parameter σ^2, the variance of the random error, determines the amount of variability inherent in the regression model. A large σ^2 will lead to observed pairs (x_i, y_i) that fall far from the true regression line, while a small σ^2 will lead to observed pairs close to the line.

An estimate of σ^2 will be used in confidence interval formulas and hypothesis-testing procedures. Because the equation of the true line is unknown, the estimate is based on the extent to which the sample observations deviate from the estimated line. Many large deviations from said line suggest a large σ^2, whereas small deviations suggest a small σ^2.

The **fitted values** of y, or **predicted values** of y, denoted as $\hat{y}_1$, $\hat{y}_2$, ..., $\hat{y}_n$, are obtained by substituting x_1, x_2, ..., x_n into the equation of the estimated regression line: $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$, $\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2$, ..., $\hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$. The **residuals** are the differences between the observed value and the predicted value of y: $y_1 - \hat{y}_1$, $y_2 - \hat{y}_2$, ..., $y_n - \hat{y}_n$.

The residual $y_i - \hat{y}_i$ corresponds to the vertical deviation between the point (x_i, y_i) and the least squares line, a positive number if the point lies above the line and a negative number if it lies below. If the residuals are all small (in absolute value), then most observations for y can be explained by the linear relationship between x and y (the ones unexplained can be conceived as noise), whereas many large residuals suggests that the linear relationship between x and y can explain very little observations for y.

Exercise 5.1.4: Consider the dataset shown in Exercise 5.1.1. Compute the fitted values and the residuals.

Solution:

Observation	Observed X	Observed Y	Fitted Y	Residue
1	0.4	1.02	0.78	0.24
2	0.48	0.88	1.03	-0.15
3	0.51	0.98	1.13	-0.15
4	0.6	1.83	1.41	0.42
5	0.7	1.5	1.72	-0.22
6	0.75	1.74	1.88	-0.14
7	0.78	1.63	1.98	-0.35
8	0.95	2.8	2.51	0.29
9	0.99	2.48	2.64	-0.16
10	1.12	3.05	3.05	0
11	1.15	3.18	3.14	0.04
12	1.25	3.68	3.46	0.22
13	1.28	3.21	3.55	-0.34

14	1.3	4.27	3.61	0.66
15	1.37	3.99	3.84	0.15
16	1.4	3.75	3.93	-0.18
17	1.46	4.18	4.12	0.06
18	1.49	3.77	4.21	-0.44
19	1.58	4.21	4.5	-0.29
20	1.6	4.92	4.56	0.36

□

Since σ^2 is itself a parameter of the linear model, and since it is necessarely unknown (it is estimated from the sample), it ought to be estimated. In regression analysis, the estimation of σ^2 is computed from summing and squaring the residuals. Define the **error sum of squares**, denoted as SSE, as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i = S_{yy} - \hat{\beta}_1 S_{xy}$$

Denoting with $\hat{\sigma}^2$ the estimator for σ^2:

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

The $n-2$ term comes from the fact that both $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators themselves, and therefore a degree of freedom is lost for each. Replacing each y_i in the formula for $\hat{\sigma}^2$ by the random variable Y_i gives the estimator S^2. Very roughly, it represents the size of a typical vertical deviation within the sample from the estimated regression line.

Lemma 5.1.1: S^2 is an unbiased estimator for σ^2.

SSE can be interpreted as a measure of how much variation in y is left unexplained by the model; that is, how much cannot be attributed to a linear relationship. If $\text{SSE} = 0$, then the model perfectly describes the relationship between the two variables in the data. A quantitative measure of the total amount of variation in observed y values is given by the **total sum of squares**, SST for short:

$$\text{SST} = S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \text{SSE} + \hat{\beta}_1 S_{xy}$$

SST is the sum of squared deviations between the observed y values and their mean (the sample mean), whereas SSE is the sum of squared deviations between the observed y values and the fitted y values. Whereas SSE represents the average vertical difference between the (x, y) pairs and the estimated regression line, SST represents the average horizontal difference between the (x, y) pairs and the straight line $f(x) = \overline{y}$.

Since SSE is the cumulative distance between the sample points and the estimated regression line, by definition there can be no line whose cumulative distance between itself and those points is smaller. Therefore, unless the estimated regression line and the line $f(x) = \overline{y}$ coincide, SSE is strictly smaller than SST.

For this reason, the ratio $\text{SSE} \ / \ \text{SST}$, is necessarely a number lying in the $(0, 1]$ interval. This ratio represents the proportion of total variation of the y variable that cannot be explained by the regression model, whereas 1 minus said ratio (a number still lying in the $(0, 1]$ interval) represents the proportion of total variation of the y variable that can be explained by the regression model. The value is also known as the **coefficient of determination**:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

The higher the value of r^2, the better the linear regression model is. If the value of r^2 is particularly low, one should investigate whether another regression model is more appropriate for the data at hand.

β_0 and β_1 are estimators, therefore their value is subject to some degree of variability depending on the values retrieved from the sample. Even if the values of x are fixed, the random variables ε introduce some inevitable uncertainty in the fitted value for y. It is therefore useful to consider y as a random variable $Y_i = \beta_0 + \beta_1 x + \varepsilon_i$ to make considerations on the nature of the estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\big(Y_i - \overline{Y}\big)}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 Y_i - \hat{\beta}_1 x_i Y_i}{n-2}$$

Theorem 5.1.1:
- $\hat{\beta}_1$ is an unbiased estimator for β_1. In other words, $E\big(\hat{\beta}_1\big) = \beta_1$;
- The variance and standard deviation of $\hat{\beta}_1$ are given by:

$$V\big(\hat{\beta}_1\big) = \frac{\sigma^2}{S_{xx}} \qquad \mathrm{SD}\big(\hat{\beta}_1\big) = \frac{\sigma}{\sqrt{S_{xx}}}$$

  Where σ^2 and σ are the variance and standard deviation of the random error(s), respectively. Replacing σ with $\hat{\sigma}$ (its estimator) in the second equation gives the expression to compute the estimated standard deviation of $\mathrm{SD}\big(\hat{\beta}_1\big)$:

$$s\big(\hat{\beta}_1\big) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

- $\hat{\beta}_1$ is normally distributed.

Note how in Theorem 5.1.1 the variance of $\hat{\beta}_1$ is obtained by dividing the variance of ε by S_{xx}, which is a measure of how the values of x are spread out with respect to their mean. Being S_{xx} the denominator means that an higher value of S_{xx}, (that is a greater variability in x), results in a smaller variance of $\hat{\beta}_1$, and therefore in a more precise estimation. Of course, if the values of x are way too far away from the mean, it most likely mean that the linear model is inappropriate.

Lemma 5.1.2: The following holds:

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t(n-2)$$

Theorem 5.1.2: The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are correct estimators for β_0 and β_1.

Theorem 5.1.3 (Gauss-Markov Theorem): The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimators for β_0 and β_1 having the smallest sample variance.