Lecture notes of Mathematical Statistics 2018-19
part 1 - Probability

Antonelli Fabio

2018

Capitolo 1

PROBABILITY SPACES

The main tool that Statistics exploits is Probability theory, which is the mathematical branch that focuses on the modeling, assessment and management of random phenomena, i.e. phenomena whose outcome is not certain, but might present several possible results. Examples as such are

Examples 1.0.1. *1. The throw of a coin or a die;*

2. Drawing 5 cards from a deck;

3. The number of calls in an hour received by a call center;

4. The waiting time at a bus stop.

From the examples it is clear that the set of possible outcomes may be finite, countable or an interval. In the first two cases we have discrete phenomena, otherwise they are said continuous.

To describe a random phenomenon we need three ingredients.

1. A set: Ω also said **sample or event space**, describing all the possible results
2. A family of subsets of Ω, $\mathcal{F}$ (called $\sigma-$**algebra**), listing the sets of results we are able to observe, i.e. what information is at our disposal. It should represent that we are able
   - to say whether nothing happened or everything happened;
   - to recognize the non occurrence of an event, if we can recognize its occurrence;
   - to say whether two events occurred at the same time or either one.
3. A rule, quantifying how possible each event is. Hence, this is a function that acts on the sets in $\mathcal{F}$.

Mathematically we have the following

Definition 1.0.1. *A **probability space** is a triple $(\Omega, \mathcal{F}, P)$, where*

1. *Ω is a set.*

2. *$\mathcal{F}$ is a $\sigma-$algebra on Ω, i.e. it verifies*

   (a) *$\emptyset, \Omega \in \mathcal{F}$;*

   (b) *if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;*

   (c) *if $\{A_i\}_{i\geq 1}$ is a countable family of sets in $\mathcal{F}$, then*

$$A_1 \cup A_2 \cup \cdots \cup A_n \cup \cdots = \bigcup_{i\geq 1}^{\infty} A_i \in \mathcal{F}.$$

3. *$P : \mathcal{F} \longrightarrow [0,1]$ a probability, that is a function such that*

   (a) *$P(\Omega) = 1$;*

   (b) *if $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$, $A_i \cap A_j = \emptyset$ for all $i \neq j$, then*

$$P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} P(A_i) = P(A_1) + \cdots + P(A_n) \ldots$$

If Ω is a set with a finite number of elements, $|\Omega| = n < +\infty$, then we can use as $\sigma-$algebra the family of all subsets of Ω, $\mathcal{P}(\Omega)$ (power set).

Example 1.0.1. *Toss of a die.*
To represent this random phenomenon we choose $\Omega = \{1,2,3,4,5,6\}$, listing all the possible results of a single throw. Also, by using subsets of Ω we can describe any combination of results. For instance

$$\{\textit{an even number is obtained}\} = \{2,4,6\}, \text{ or}$$
$$\{\textit{ a number less than or equal to 4 is obtained}\} = \{1,2,3,4\}.$$

Finally, assuming that the die is fair, we have a natural way to attribute the likelihood of any event to happen. Indeed we expect

$$P(\{i\}) = \frac{1}{6}, \qquad i = 1, \ldots, 6,$$

since we have no reason to think a number more likely than the others. Consequently, we may attribute probability also to more complex events, for instance

$$P(\{\textit{an even number is obtained}\}) = P(\{\textit{either 2 or 4 or 6 is obtained}\}) = P(\{2,4,6\}$$
$$= \quad P(\{2\} \cup \{4\} \cup \{6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2},$$

by property 3 (b).

Recall that

$$\begin{aligned}(A\cap B)\cup C &= (A\cup C)\cap(B\cup C),\quad (A\cup B)\cap C=(A\cap C)\cup(B\cap C)\\ (A\cap B)^c &= A^c\cup B^c,\quad (A\cup B)^c=A^c\cap B^c,\quad (A^c)^c=A\end{aligned}$$

Remark 1.0.1. *By property 3b, we have for $A,B\in\mathcal{F},\{A_n\}_n\subseteq\mathcal{F}$*

1. *$P(A^c)=1-P(A)$, since $\Omega=A\cup A^c$, $\emptyset=A\cap A^c$*
2. *$P(A\cup B)=P(A)+P(B)-P(A\cap B)$, since $A\cup B=(A\cap B^c)\cup(A\cap B)\cup(B\cap A^c)$, so*

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

   $A=(A\cap B^c)\cup(A\cap B)$ e $B=(B\cap A^c)\cup(B\cap A)$ as disjoint unions.
3. $P\Big(\bigcap\limits_{n\in\mathbb{N}}A_n\Big)=1-P\Big(\bigcup\limits_{n\in\mathbb{N}}A_n^c\Big)$

It also holds

Proposition 1.0.1. *Let $(\Omega,\mathcal{F},P)$ be a probability space, then*

1. *For any$A,B\in\mathcal{F}$ so that $A\subseteq B$ we have $P(A)\le P(B)$;*
2. *for any sequence os sets in $\mathcal{F}$ such that $A_1\subseteq A_2\subseteq\ldots$, $\bigcup\limits_{i=1}^{+\infty}A_i\in\mathcal{F}$, we have*

   $$\lim_{k\to+\infty}P(A_k)=P(\bigcup_{n=1}^{+\infty}A_n),$$

   (alternatively $A_1\supseteq A_2\supseteq\ldots$, $\bigcap\limits_{n=1}^{+\infty}A_n\in\mathcal{F}$ implies $\lim\limits_{k\to+\infty}P(A_k)=P(\bigcap\limits_{n=1}^{+\infty}A_n)$).
3. *for any$\{A_n\}_n\subseteq\mathcal{F}$, $\bigcup_{n=1}^{+\infty}A_n\in\mathcal{F}$ and $P(\bigcup\limits_{n=1}^{+\infty}A_n)\le\sum\limits_{n=1}^{+\infty}P(A_n)$.*

1.1 Uniform sample spaces

As in the case of the die in the previous example, there are many situations, when there is a family of possible results where we have no reason to think one more likely than the others:

- tossing of a coin or a die;
- drawing a number at bingo.

In general, if Ω has m elements $\Omega = \{\omega_1, \ldots, \omega_m\}$ and they are equally likely, necessarily

$$P(\omega_i) = \frac{1}{m} = \frac{1}{|\Omega|}, \qquad \forall \quad i = 1, \ldots, m,$$

and we say that $(\Omega, \mathcal{P}(\Omega), P)$ is a **uniform sample space**. Consequently for any $A \subseteq \Omega$, we have

$$P(A) = \frac{|A|}{|\Omega|},$$

hence it is important to recognize correctly the space Ω and to count correctly its cardinality and the cardinality of its subsets.

Example 1.1.1. *Two dice are tossed and we want to evaluate the probability to obtain 8 with the sum of the two results. We might (incorrectly) think that* $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$*, giving all the possible results of the sum. In this case, assuming that each sum is equally likely, we should have*

$$P(sum = 8) = \frac{1}{11}.$$

This is not true, because the sum comes from the contribution of the values of the TWO dice and we need to be able to distinguish between those. Thus

$$\Omega = \{(i, j),\, i, j = 1, \ldots, 6\} \Rightarrow |\Omega| = 36, \Rightarrow P((i, j)) = \frac{1}{36}, \forall i, j, = 1, \ldots, 6$$

and all the possible cases that give 8 as a sum are

$$A = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} \Rightarrow P(A) = \frac{5}{36}.$$

1.1.1 Some combinatorics

We here recall some combinatorics to help us count the cardinality of sets

1. Given two sets A e B their **Cartesian product** is

   $$A \times B = \{(i, j), :\ i \in A,\ j \in b\},$$

   and we have (provided the two sets are finite)

   $$|A \times B| = |A| \times |B|.$$

   The Cartesian product can be extended to n sets obtaining all the ordered $n-$ples.

2. **Factorial**.

   It represents the number of ways to put n distinguishable (numbered) balls into n boxes, one per box.

   $$n! = n \cdot (n - 1) \cdots 2 \cdot 1, \qquad n \in \mathbb{N}.$$

   Properties

(a) $0! = 1$;

(b) $n! = n \cdot (n-1)!$

If the balls are indistinguishable (so the order does not count), for $m \geq n$ boxes, we will obtain that the number of ways to put one ball per box is given by

$$\binom{m}{n} = \frac{m!}{n!(m-n)!}, \qquad m, n \in \mathbb{N},$$

called **binomial coefficient**.

Properties

(a) $\binom{m}{0} = 1$, $\binom{m}{m} = 1$, $\binom{m}{1} = m$, $\binom{m}{m-1} = m$;

(b) $\binom{m}{n} = \binom{m}{m-n}$;

(c) $\binom{m}{n} = \binom{m-1}{n-1} + \binom{m-1}{n}$;

(d)

$$(a+b)^n = \sum_{n=0}^{m} \binom{m}{n} a^n b^{(m-n)}, \quad a, b \in \mathbb{R}$$

whence $\sum_{n=0}^{m} \binom{m}{n} = 2^m, \quad \Rightarrow \quad \sum_{n=0}^{m} \binom{m}{n} \frac{1}{2^m} = 1.$

3. If we can put n balls into m boxes with no constraint as to the number of balls in each box, then we have always m choices for each ball and the number of possible ways becomes m^n.

1.2 Exercises

1. Licence plates are formed with two letters out of an alphabet with 26 symbols and 5 numbers. How many licence plates can be formed if

   (a) letters and numbers can be repeated;

   (b) letters and numbers cannot be repeated and the order counts;

   Solution: In the first case we have 26 options for each letter and 10 for each number, therefore the total number of possible licence plates will, be

   $$26^2 10^5.$$

   In the second case we have 26 choices for the first letter, 25 for the second, while we have 10 choices for the first number , 9 for the second, 8 for the third, 7 for the fourth and 6 for the fifth, hence the total number of plates we can form is $26 \cdot 25 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6$.

2. Two dice are tossed, compute the probability of

   (a) $E = \{$ sum of the results is odd $\}$;
   (b) $F = \{$ at least one of the two results $= 1\}$;
   (c) $G = \{$ sum of the results$= 5\}$;
   (d) $E \cap F$, $E \cup F$, $F \cap G$, $E \cap F^c$, $E \cap F \cap G$

   Solution: The sample space is given by the ordered pairs of the results of the dice $\Omega = \{(i, j : i, j = 1, \ldots, 6\}$ that has 36 elements.

   The pairs in E must have an even and an odd number, which are half the total number of cases, i.e. 18, hence $P(E) = \dfrac{18}{36} = \dfrac{1}{2}$. To count the possible cases for F, if we fix 1 as the result of the first die, then we have 6 possibilities for the second one, viceversa fixing 1 as the result of the second die, we have other 5 distinct possibilities since we do not want to count the pair (1,1) twice, thus $P(F) = \dfrac{11}{36}$. As for the others we have

   $G = \{(1,4),(2,3),(3,2),(4,1)\} \Rightarrow P(G) = \dfrac{4}{36}$.

   $E \cap F = \{(1,2),(2,1),(1,4),(4,1),(1,6),(6,1)\} \Rightarrow P(E \cap F) = \dfrac{6}{36}$.

   $E \cup F = E \cup \{(1,1),(3,1),(1,3),(1,5),(5,1)\} \Rightarrow P(E \cap F) = \dfrac{23}{36}$.

   $E \cap F^c = E \setminus (E \cap F) \Rightarrow P(E \cap F^c) = \frac{12}{36}$.

   $E \cap F \cap G = F \cap G = \{(1,4),(4,1)\} \Rightarrow P(E \cap F \cap G) = P(F \cap G) = \dfrac{2}{36}$.

3. Show that

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

4. Say which one is true

   (a) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
   (b) $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
   (c) $(A \cup B) \cup C = A \cup (B \cup C)$
   (d) $(A \cup B) \cap C = A \cup (B \cap C)$
   (e) $(A \cup B) \cap C = (A \cup C) \cap (B \cup C)$

5. At a round table there are 10 seats. The guests chose where to sit randomly. Compute the probability the guests A and B sit next to each other.

6. Two dice are tossed. Compute

   (a) The probability of $A = \{$ the difference of the results $= 2\}$;
   (b) $P(A \cap B)$ and $P(A \cap C)$, where $B = \{$exactly one of the results is odd $\}$, $C = \{$the result of the first die is odd $\}$.

7. A password is formed by 8 characters chosen from 26 letters, that might be either uppercase or lowercase. Compute the probability to generate a password

   (a) with only lowercase letters and with repetitions;

   (b) with only lowercase letters, without repetitions and with order.

Capitolo 2

CONDITIONAL PROBABILITY AND INDEPENDENCE

2.1 Conditional Probability

When we know that an event happened, we may recalibrate our computation of probabilities for all the other events on the basis of this knowledge.

Definition 2.1.1. *Given $A, B \in \mathcal{F}$ with $P(B) > 0$, we call* ***probability of*** *A* ***given*** *B*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{2.1}$$

Remarks 2.1.1.

1. *Under this new probability $P(B|B) = 1$, that is to say it is certain;*
2. *$P(\cdot|B)$ is a probability, in the sense that it verifies the definition;*
3. *formula (2.1) may be alternatively written*

$$P(A \cap B) = P(A|B)P(B), \tag{2.2}$$

thus the conditional probability may be a tool to compute the probability of the intersection of two events.

Examples 2.1.1.

1. *If we have a box with 6 black balls and 4 white ones and we draw two balls without replacement, to compute the probability to obtain a white ball at the second draw, we have to use the conditional probability. Indeed, let us denote by $B_i = \{$ white ball at the $i-th$ draw$\}$, $N_i = \{$ black ball at the $i-th$ draw$\}$, $i = 1, 2$, then*

$$P(B_2) = P(B_2 \cap B_1) + P(B_2 \cap N_1) = P(B_2|B_1)P(B_1) + P(B_2|N_1)P(N_1) = \frac{3}{9}\frac{4}{10} + \frac{4}{9}\frac{6}{10} = \frac{2}{5}$$

2. *A certain illness affects* 3% *of a population. We denote by* $A = \{$ *affected* $\}$, $A^c = \{$ *not affected* $\}$, *so* $P(A) = 3\%$ *and* $P(A^c) = 97\%$.

   A new test to detect the illness is administered to the population obtaining a positive for 98% affected people and a negative for 96% of unaffected individuals, that is to say we have the following conditional probabilities

$$P(+|A) = 98\%, \quad P(-|A) = 2\%, \quad P(-|A^c) = 96\%, \quad P(+|A^c) = 4\%.$$

   We want to compute the probability of having a positive when giving the test to a new individual and the probability that the individual is actually ill, knowing the test marked positive: $P(+)$ *e* $P(A|+)$.

   We may proceed as follows

$$P(+) = P(+ \cap A) + P(+ \cap A^c) = P(+|A)P(A) + P(+|A^c)P(A^c) = \frac{98 \cdot 3 + 4 \cdot 97}{10^4}$$

   and

$$P(A|+) = \frac{P(+ \cap A)}{P(+)} = \frac{P(+|A)P(A)}{P(+)} = \frac{98 \cdot 3}{98 \cdot 3 + 4 \cdot 97} = \frac{294}{682},$$

   less than 50%, implying that the test is not reliable.

In the second example we exploited the following two formulas.

Proposition 2.1.1. ***Bayes' Formula.***
Let $A, B \in \mathcal{F}$ *with* $P(A) > 0, P(B) > 0$, *then*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

And if we have

Definition 2.1.2. *A family of sets* $\{E_i\}_{i=1}^n$ *is said a* ***finite partition*** *of* Ω *if*

1. $E_i \cap E_j = \emptyset$ *for all* $i \neq j$, $i, j = 1, \ldots, n$;
2. $\Omega = \bigcup_{i=1}^{n} E_i$.

Then we have

Proposition 2.1.2. ***Total probabilities formula***
Let $\{E_i\}_{i=1}^n \subseteq \mathcal{F}$ *be a finite partition of* Ω, *then for any* $A \in \mathcal{F}$ *it holds*

$$P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i).$$

2.2 Independence

Another fundamental concept in probability is independence, which means that the occurrence of an event does not influence the occurrence of another, we have the following

Definition 2.2.1. *Two events A e B are* ***independent*** *if*

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

Given $I \subseteq \mathbb{N}$, a family of events $\{A_i\}_{i \in I}$ is ***independent*** *if for any integer $k \leq |I|$ and $A_{i_1}, \dots A_{i_k}$ distinct sets, it holds*

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}).$$

Example 2.2.1. *For instance the tosses of three dice are clearly independent, therefore if we want to compute the probability to obtain the same number on all the three dice, we may proceed as follows*

$$\begin{aligned} P(\textit{same number}) &= \sum_{i=1}^{6} P(\{\textit{result die } 1 = i\} \cap \{\textit{result die } 2 = i\} \cap \{\textit{result die } 3 = i\}) \\ &= \sum_{i=1}^{6} P(\{\textit{result die } 1 = i\}) P(\{\textit{result die } 2 = i\}) P(\{\textit{result die } 3 = i\}) \\ &= \frac{6}{216} = \frac{1}{36}. \end{aligned}$$

If A and B are independent then $P(A|B) = P(A)$. Also, independence is a much stronger notion, indeed if A, B are independent, then so are A^c and B, A and B^c, A^c and B^c, since (for instance)

$$P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c),$$

We remark that independence can be realized also with respect to a conditional probability, in this case we talk about conditional independence.

Definition 2.2.2. *Given events $A_1, \dots, A_n$ they are* ***conditionally independent*** *given B with$P(B) > 0$ if for any $k \leq n$ and $i_1 \neq \cdots \neq i_k \in \{1, \dots, n\}$ we have*

$$P(A_{i_1} \cap \cdots \cap A_{i_k}|B) = P(A_{i_1}|B) \cdots P(A_{i_k}|B).$$

2.3 Exercises

1. Let A and B be two independent events with $P(A) = \frac{1}{4}$, $P(B) = \frac{1}{5}$, $P(A \cup B) = \frac{1}{2}$. Compute $P(A^c|B^c)$.

2. Let A, B, C be three events such that $P(A|B \cap C) = \dfrac{2}{5}$ and $P(B|C) = \dfrac{1}{3}$. Compute $P(A \cap B|C)$.

3. If $P(B) = \frac{1}{3}, P(A|B) = \frac{1}{4}, P(C|B \cap A) = \frac{1}{5}$, how much is $P(A \cap B \cap C)$?

4. Two players, A and B, draw two cards each without replacement, only once, in turn, from a deck with two J's, two K's and two Q's. A starts. A player wins if he draws at least a J and the adversary draws no J's. Otherwise they are even. Compute the probability A wins, B wins and of a tight?

5. In a tennis tournament games are played at the same time on three fields. The game on each field happens independently of the games on the other fields. On each field the game is won with three sets out of 5.

   We denote by $(A_i, B_i), i = 1, 2, 3$ the couple of players on each field and we have

   $$P(A_1 \text{ wins}) = \frac{1}{2}, \qquad P(A_2 \text{ wins}) = \frac{2}{3}, \qquad P(A_3 \text{ wins}) = \frac{1}{4}$$

   (a) Compute the probability that the game is won in at most 4 sets respectively on each of the three fields.

   (b) Compute the probability that the game is won in at most 4 sets on all the fields.

6. A tv set is produced in two factories, A and B, covering respectively 40% and 60% of the production. In factory A a certain defect occurs with probability 0.05, while in factory B with probability 0.15. Shops are stacked randomly with tv's from both factories.

   (a) When buying a tv set, what is the probability to get a defective one?

   (b) Knowing to have bought a defective tv, what is the probability it came from factory B?

7. There are two coins, A which is fair and B that gives head with probability $\frac{1}{4}$. A coin is chosen randomly and it is tossed repeatedly.

   (a) Compute the probability to have head at the first toss.

   (b) If two heads are obtained in two tosses, what is the probability that is coin A that is being tossed?

8. A certain item produced by a factory can present two types of defects, with respective percentages of 3% and 7%. The defects happen independently.

   (a) Compute the probability an item presents both defects.

   (b) Compute the probability an item presents at least one of the defects.

   (c) What is the probability to have defect 1, knowing that the item is defective?

Capitolo 3

RANDOM VARIABLES

3.1 Distribution function and density function

To describe the random phenomena, it is often necessary to eliminate any ambiguity, hence we need to mathematically describe any quantity that comes into play. The notion of **random variable** is the proper tool to do so.

Definition 3.1.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A map $X : \Omega \longrightarrow \mathbb{R}$ is a **random variable (r.v.)** if for any $t \in \mathbb{R}$ we have that the set $\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}$*

This definition tells that we are always able to observe, on the basis of our information ($\mathcal{F}$), whether the r.v. happens to be below or above a given level t. From now on, the argument of the function will always be omitted for the sake of simplicity, i.e. the set $\{\omega \in \Omega : X(\omega) \leq t\}$ will be written as $\{X \leq t\}$.

To describe the probabilistic behavior of a r.v. we introduce

Definition 3.1.2. *Let X be a r.v. on $(\Omega, \mathcal{F}, P)$, then we call **(cumulative) distribution function** of X, the function*

$$F_X(t) = P(X \leq t) \qquad t \in \mathbb{R}.$$

This means that we can always attribute a probability to the possibility that X remains below or above a given level t. Based on the definition, the distribution function verifies the following properties

1. $\lim_{t \to +\infty} F_X(t) = 1\,; \quad \lim_{t \to -\infty} F_X(t) = 0;$
2. it is non decreasing, i.e. $x < y \Rightarrow F_X(x) \leq F_X(y)$;
3. it is right continuous with left limits, for any $t_0 \in \mathbb{R}$, $\lim_{t \to t_0^+} F_X(t) = F_X(t_0), \quad \exists \lim_{t \to t_0^-} F_X(t)$

Conversely a function F that satisfies the above three properties can be viewed as the distribution function of some r.v. X.

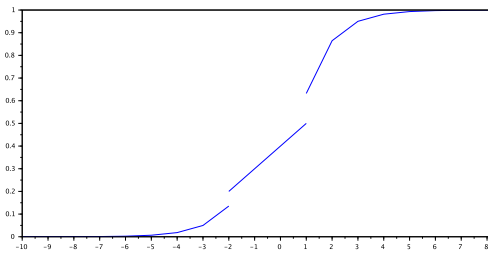

Figura 3.1: The typical shape of a distribution function.

From the definition, we obtain for any $a < b$ that $P(a < X \leq b) = F_X(b) - F_X(a)$. Basically a distribution function describes the probabilistic behavior of a r.v. representing some random phenomenon. If F is a distribution function for some X we write $X \sim F$ and we also say that F is the **law** of X.

Remark 3.1.1. *The above three properties imply that a distribution function can have at most countably many jump discontinuities.*

Last remark implies that a distribution function is differentiable everywhere except at the jump discontinuities and at most countably many other points. From the definition, we obtain for any $a < b$ that $P(a < X \leq b) = F_X(b) - F_X(a)$, consequently, whenever a distribution function has a jump, say at t_0, we have that the corresponding r.v. has a probability concentrated at that point, that is to say $P(X = t_0) = F_X(t_0) - F_X(t_0-)$.

If a distribution function is piecewise constant, this means that for any interval $P(a < X \leq b) = F_X(b) - F_X(a) = 0$, unless it contains a jump, hence all the probability is concentrated at the jumps and the r.v. can take at most countably many values with positive probability. R.v.'s of this type are called **discrete**, and we have

Definition 3.1.3. *For a discrete r.v. X we call (discrete) probability density function the map*

$$\begin{aligned} p: \quad & \mathbb{R} \longrightarrow [0,1] \\ & x \longrightarrow P(X = x) \end{aligned}$$

such that

a. $p(x) = 0$ *except at most countably many values* $\{x_i\}_{i \in I} \subseteq \mathbb{R}$, *where* $I \subseteq \mathbb{N}$.

b. $\sum_{x \in \mathbb{R}} p(x) = \sum_{i \in I} p(x_i) = 1$.

Conversely a function that verifies the above definition can be considered the probability density of some discrete r.v. Sometimes this function is called **probability mass function (pmf)** to point out that the probability is concentrated on points.

Example 3.1.1. *The following function*

$$p(1)=\frac{1}{4},\, p(2)=\frac{2}{5},\, p(3)=\frac{1}{10},\, p(4)=\frac{3}{10}, \qquad p(x)=0,\, forallx\neq 1,2,3,4$$

cannot be considered the probability density function of a discrete r.v. taking values 1,2,3,4, since

$$\frac{1}{4}+\frac{2}{5}+\frac{1}{10}+\frac{3}{10}=\frac{21}{20}>1.$$

Similarly if we set

$$p(1)=\frac{1}{4},\, p(2)=\frac{3}{4},\, p(3)=\frac{1}{10},\, p(4)=-\frac{1}{10}, \qquad p(x)=0,\, \forall\, x\neq 1,2,3,4$$

is not a density, since a probability cannot be negative.

Consequently any probability relative to X can be computed by using the density function, indeed for any $A\subseteq\mathbb{R}$, we have

$$P(X\in A)=\sum_{x_i\in A}p(x_i), \quad \Rightarrow \quad P(a<X\leq b)=\sum_{x_i\in(a,b]}p(x_i), \qquad \forall\, a<b.$$

A first example of a discrete r.v. is given by the **indicator function of a set** $A\in\mathcal{F}$, defined as

$$\mathbf{1}_A(\omega)=\begin{cases} 1 & \omega\in A \\ 0 & \omega\in A^c, \end{cases}$$

its density is then $P(\mathbf{1}_A=1)=P(A)$ and$P(\mathbf{1}_A=0)=P(A^c)=1-P(A)$.

This type of r.v.'s are said Bernoullian

$$X=\begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

where $0<p<1$. We write $X\sim$Bin$(1,p)$.

Another immediate example of discrete r.v. is the **uniform** one, when X takes finitely many values with equal probability

$$P(X=x_i)=\frac{1}{n}, \qquad i=1,\ldots,n.$$

Definition 3.1.4. *If a distribution function F_X is differentiable in $\mathbb{R}$, then its derivative, f_X, is said* ***probability density function*** *of X and it verifies:*

1. $f_X(t)\geq 0$, *for all* $t\in\mathbb{R}$;
2. $\displaystyle\int_{\mathbb{R}} f_X(t)dt=1$.

In this case X is said an **(absolutely) continuous** r.v. and we remark that, since there are no jumps, no single point is charged with positive probability, $P(X = t_0) = 0$, $\forall\, t_0 \in \mathbb{R}$. The r.v. has positive probability to be in intervals and this notion corresponds to the notion of mass density in physics.

From the fundamental theorem of Calculus it follows that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \quad F_X(b) - F_X(a) = \int_a^b f_X(t)dt, \qquad \forall\, x, a \leq b \in \mathbb{R}.$$

Example 3.1.2. *Similarly to the discrete case, we may define a uniform r.v. on an interval of the real line $[a, b]$. In this case we want that the probability charges uniformly only this interval. Hence its probability density must be*

$$f_X(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x).$$

This function verifies the properties that define a density and its antiderivative (i.e. the distribution function) is

$$F_X(x) = x\mathbf{1}_{[0,1)}(x) + \mathbf{1}_{[1,+\infty)}(x).$$

In this case we write $X \sim U([a, b])$.

We remark that to define a probability density it s enough to have a non negative integrable function, which can always be normalized to obtain a density.

Example 3.1.3. *Let $f(x) = C(4x - 2x^2)\mathbf{1}_{(0,2)}(x)$, with $C \in \mathbb{R}_+$. We want to determine the constant C that makes f a density.*

The function $f(x)$ is equal to $C(4x - 2x^2)$ in the interval $(0, 2)$ and 0 otherwise, thus it is non negative. To verify the second property we set

$$\int_{\mathbb{R}} f(x)dx = 1,$$

which implies

$$C\int_0^2 (4x - 2x^2)dx = 1 \Rightarrow C(2x^2 - \frac{2}{3}x^3)\Big|_0^2 \Rightarrow C\frac{8}{3} = 1 \Rightarrow C = \frac{3}{8}.$$

3.2 Main discrete r.v.'s

In this section we describe briefly the main distributions for discrete random variables. The first two can be constructed from a repeated trials scheme, whether this regards tosses of a coin or draws from a box containing two types of balls.

The idea is the following: there is a box with m balls, whose r are white and the other $m-r$ are black and n repeated draws of one ball at a time are done from this box, $r \leq m, n$ are all integers.

Every time we draw a white ball, we mark it as a success, hence we define for $i = 1, \ldots n$,

$$X_i = \begin{cases} 1 \text{ white at the i-th draw} \\ 0 \text{ black at the i-th draw.} \end{cases}$$

Draws may happen with replacement o without.

3.2.1 Draws with replacement

In this case each draw is repeated in the same conditions as the others and clearly they do not influence one another. So the above r.v.'s are all distributed in the same way, denoting by $p = \frac{r}{m}$ and $1 - p = \frac{m-r}{m}$, namely we have for all $i = 1, \dots n$

$$X_i = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1-p. \end{cases}$$

If we want to count the number of successes in n draws, we may define the r.v.

$$S_n = \sum_{i=1}^{n} X_i,$$

then this r.v. can take all the values $k = 0, 1, \dots n$ and its density will be

$$\begin{aligned} & P(S_n = k) \\ = \ & (\#\text{ ways to position the k objects in n boxes}) P(\text{ first k successes and last n-k unsuccesses}) \\ = \ & \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

Definition 3.2.1. *A r.v. X that takes values $k = 0, 1, \dots n$ such that*

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.1}$$

is called ***binomial of parameters*** **n** ***and*** **p** *and we write $X \sim Bin(n, p)$.*

A binomial r.v. is appropriate whenever the repeated trials are independent.

Example 3.2.1. *A book has 400 pages and each page, INDEPENDENTLY OF ALL THE OTHERS, has 2% probability to present at least one misprint. Denote by S_{400} the number of pages that contain at least a misprint. What is the probability there are at most 4 pages with at least one misprint?*

In this case we may consider the trials independent, hence $S_{400} \sim Bin(400, 0.02)$, then we have

$$P(S_{400} \leq 4) = \sum_{k=0}^{4} P(S_{400} = k) = \sum_{k=0}^{4} \binom{400}{k} 0.02^k 0.98^{400-k}.$$

3.2.2 Draws without replacement

Exactly with the same composition of white and black balls in the box, we now consider a sequence of draws without replacement, we are assuming $n \geq r \leq m$ and again we want to count the number of success in n trials, that is computing the density of

$$S_n = \sum_{i=1}^{n} X_i.$$

In this case S_n can assume all the values between 0 and r. The trials are no longer independent, because what happened in the previous draws determines the composition of balls in the box at the following draw.

If we want to find $P(S_n = k)$, $k = 0, \dots, r$, we may proceed as follows: the total number of cases is $\binom{m}{n}$, because it corresponds to all the possible ways to select n objects from a set of m, while all the cases favorable to the chosen outcome are those that see exactly k objects selected from the set of r white balls and exactly $n-k$ from the set of $m-r$ black balls. We may therefore conclude

$$P(S_n = k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}.$$

Definition 3.2.2. *A r.v. X has a **hypergeometric distribution of parameters** (r, m, n), if it takes values $k = 0, \dots, \min(r, n)$ and*

$$P(X = k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}. \tag{3.2}$$

Example 3.2.2. *At poker with 52 cards, 4 suits, 13 values, we want to compute the probability to obtain a poker at the first hand.*

The total number of cases is $\binom{52}{5}$. Obtaining a poker means that we have 13 different possibilities of poker, (one per value) then we have to consider the probability to get a specific one, say a poker of aces.

To do so, we divide the population of cards in two types: ace - no ace. There are 4 individuals of the first type and 48 of the second, hence we want to draw exactly all four individuals of the first type and one of the second, thus

$$P(poker) = 13\frac{\binom{4}{4}\binom{48}{1}}{\binom{52}{5}}.$$

3.2.3 Distribution with countably many values

So far we have been considering discrete r.v.'s that could assume only finitely many values with positive probability, in this section we present other two discrete r.v.'s that instead may take countably many values.

1. Suppose that we are performing draws with replacement from a box with white and black balls, such that at each trial the probability to get a white ball is $p := \frac{r}{m}$. We repeat the draws until we fish the first white ball. If $\{X_i\}$ is the sequence of the independent trials, we may denote by

$$\tau = \inf\{n : X_n = 1\},$$

the time of first success, which can assume all the integers. For $k = 1, 2, \dots$ to draw the first white ball exactly at time k, this means that the first $k-1$ draws have seen only black balls (with probability $1-p$ each time) and the last a white ball, hence

$$P(\tau = k) = (1-p)^{k-1}p. \tag{3.3}$$

A r.v. with this density is said to follow a (modified) **geometric** distribution with parameter $0 < p < 1$.

We remark that being (3.3) a probability density we have

$$\sum_{k=1}^{+\infty} p(1-p)^{k-1} = 1. \tag{3.4}$$

By using (3.4) it is straightforward to compute the probability of the tail of the geometric distribution

$$\begin{aligned} P(T \geq k) &= \sum_{h=k}^{\infty} p(1-p)^{h-1} = \sum_{h=k}^{\infty} p(1-p)^{h-k+k-1} \\ &= (1-p)^{k-1} \sum_{h=k}^{\infty} p(1-p)^{h-k} = (1-p)^{k-1} \sum_{j=0}^{\infty} p(1-p)^{j} = (1-p)^{k-1}. \end{aligned}$$

The geometric density has also the property of the **loss of memory**: if we want to compute the probability to obtain the first success at a given time, knowing that no success occurred before, it is equivalent to start our clock afresh, indeed for $h \geq 1$ we have

$$\begin{aligned} P(T = h+k | T > k) &= \frac{P(T = h+k, T > k)}{P(T > k)} = \frac{P(T = h+k)}{P(T \geq k+1)} = \frac{p(1-p)^{h+k-1}}{(1-p)^{k}} \\ &= p(1-p)^{h-1} = P(T = h). \end{aligned}$$

Example 3.2.3. *Two identical bags, A and B, are presented, the first contains numbered marbles from 1 to 50, the second marbles numbered from 1 to 45. A bag is chosen randomly and then draws with replacement are performed until a number greater than 40 is drawn.*

Let us denote by τ the time of first success, what is its distribution?

If the bag A is chosen then $\tau = \tau_A \sim geom(\frac{1}{5})$, otherwise $\tau = \tau_B \sim geom(\frac{1}{9})$, therefore we have

$$P(\tau = k) = P(\tau = k|A)P(A) + P(\tau = k|B)P(B) = \frac{1}{2}\Big[\frac{1}{5}(\frac{4}{5})^{k-1} + \frac{1}{9}(\frac{8}{9})^{k-1}\Big]$$

2. Another interesting distribution is the **Poisson of parameter** $\lambda > 0$. We say that $X \sim \text{Poisson}(\lambda)$, if X takes values $k = 0, 1, 2, \dots$ and

$$P(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}.$$

Being a probability density we have that

$$1 = \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} \Rightarrow \mathrm{e}^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}.$$

It is particularly useful to model the number of arrivals at a service in a given period, indeed it attributes a rather high probability to having a few arrivals in the period and exponentially small probability to having a large number of arrivals.

It also has the following interesting approximating property

Proposition 3.2.1. *If $X \sim Bin(n, p_n)$ such that $np_n \longrightarrow \lambda > 0$ as $n \to +\infty$, then for $k = 0, \ldots, n$ it holds*

$$P(X = k) \longrightarrow \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}$$

3.3 The main absolutely continuous r.v.'s

As we already mentioned, many random phenomena cannot be represented by discrete r.v.'s, since they take values in real intervals:

1. the waiting time at a bus stop;
2. the lifetime of an electrical component;
3. the temperature of a certain geographical area.

In this section we present the main distributions of absolutely continuous r.v.'s. In this case the distribution function is differentiable and therefore the probability laws of the r.v.'s are characterized by means of the probability density functions.

3.3.1 The exponential density

A r.v. has an exponential density, $T \sim \exp(\lambda)$, $\lambda > 0$, if its density is

$$f_T(t) = \lambda \mathrm{e}^{-\lambda t} \mathbf{1}_{(0,+\infty)}(t),$$

whence we obtain by integration the cumulative distribution function

$$P(T \leq t) = F_T(t) = \int_{-\infty}^{t} \lambda \mathrm{e}^{-\lambda s} \mathbf{1}_{(0,+\infty)}(x) dx = (1 - \mathrm{e}^{-\lambda t}) \mathbf{1}_{(0,+\infty)}(t). \tag{3.5}$$

The exponential density is often used to describe a waiting time or a survival time, because it is concentrated on the positive real values and it gives high probability to small time intervals, while it is extremely unlikely to exceed large reals, indeed it is the complementary probability of (3.5)

$$P(T > t) = \mathrm{e}^{-\lambda t},$$

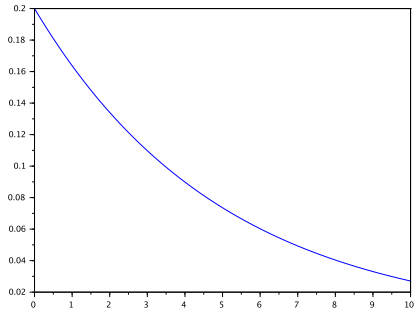

Figura 3.2: The exponential density function with parameter 0.2

hence it is exponentially small.

Similarly to the geometric distribution, it shows loss of memory

$$\begin{aligned} P(T > s+t | T > t) &= \frac{P(T > s+t, T > t)}{P(T > t)} = \frac{P(T > s+t)}{P(T > t)} \\ &= \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda t}} = \mathrm{e}^{-\lambda s} = P(T > s), \end{aligned}$$

in a certain sense, if something survived until a given time t, the probability to survive further is equivalent to the initial one (as if the clock had been resetted). Therefore, this type of distribution is not the best to represent the survival time of components that wear with usage.

3.3.2 The Standard Normal

We say that X has a Standard Normal distribution, $X \sim \mathcal{N}(0,1)$, if its density is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}.$$

Let us remark that this implies, being a density, that

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} dx = 1 \quad and \quad \forall\, a < b \in \mathbb{R}, \quad P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} dx.$$

Unfortunately the last integral cannot be computed with elementary integration for all values of a and b, but it can be computed numerically. The cumulative distribution function of a Standard Normal is usually denoted by $\phi(x)$

$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} dy, \qquad x \in \mathbb{R}.$$

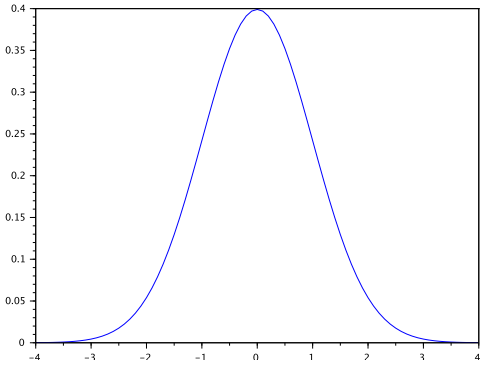

Figura 3.3: The standard normal density function.

and its values are tabulated for values of x between -3 e 3 with intervals of 0,01. They are not usually tabulated for larger values since the integration shows that the interval $[-3.3]$ absorbs 99% of the probability.

Nevertheless some consideration about ϕ help in the calculations. This function is even and the total underlying area is 1, hence we have the following two results

$$\phi(0) = \frac{1}{2}, \qquad \phi(-x) = 1 - \phi(x).$$

The standard Normal is often used to represent a random error, which might be either positive or negative with equal probability and very likely to be small and more than exponentially unlikely to be large.

Remark 3.3.1. *Let us remark that the standard Normal is characterized by two parameters that here take the specific values 0 and 1. Later on, the role of these two parameters will become clearer.*

3.3.3 The Gamma density

Here we introduce a family of density functions, particularly useful for its flexibility.

We say that X follows a $\Gamma(\alpha, \lambda)$ distribution for α and $\lambda > 0$, if its density function is given by

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} \mathrm{e}^{-\lambda x} \mathbf{1}_{\{x>0\}},$$

where $\Gamma(\alpha)$ is the constant that makes the function a density. More precisely, this constant can be expressed by means of an integral

$$\Gamma(\alpha) = \int_0^{+\infty} y^{\alpha-1} \mathrm{e}^{-y} dy,$$

that unfortunately cannot be always computed with elementary methods for any value of α.

Even though $\Gamma(\alpha)$ is not always easily computable, we may deduce some properties and compute it for some particular values of α.

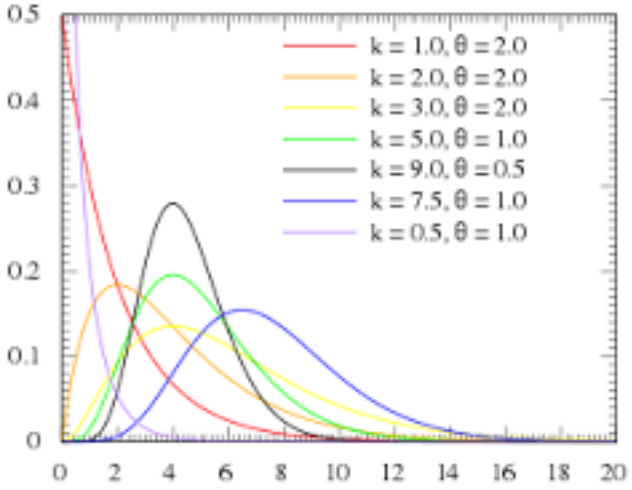

Figura 3.4: gamma density functions.

1.
$$\Gamma(1) = \int_0^{+\infty} \mathrm{e}^{-y} dy = -\mathrm{e}^{-y}\Big|_0^{+\infty} = 1.$$

2. For any $\alpha > 0$, it holds
$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha).$$

3. Hence for $n \in \mathbb{N}$,
$$\Gamma(n+1) = n\Gamma(n) = n(n-1)\Gamma(n-1) = \cdots = n(n-1)(n-2)\ldots 1\Gamma(1) = n!$$

The family of Gamma densities includes also the exponential of parameter λ which can be seen as a $\Gamma(1, \lambda)$. The density $\Gamma(n, \lambda)$ is also called Erlang-n distribution.

3.4 Transformations of random variables

An extremely important tool to generate random variables with different distributions is to apply a function to an originally given r.v.

It is clear that if $X : \Omega \longrightarrow \mathbb{R}$ is a r.v. and $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ is a real function, then also $Y = \psi(X)$ is a r.v.

The question is what is the distribution of this new r.v.?

The idea is to exploit the original density to identify the new one. For discrete r.v.'s this is readily done

Example 3.4.1. *Let us consider the r.v. X with density*

$$P(X = -3) = \frac{1}{8}, \quad P(X = -2) = \frac{1}{8}, \quad P(X = -1) = \frac{1}{4}, \quad P(X = 1) = \frac{1}{4}, \quad P(X = 2) = \frac{1}{4},$$

then the r.v. $Z = X^4$ *(*$\psi(x) = x^4$*) can take values 1,16, and 81 and it has the following density*

$$\begin{aligned} P(Z=1) &= P(X^4=1) = P(X=1)+P(X=-1) = \frac{1}{2}, \\ P(Z=16) &= P(X^4=16) = P(X=2)+P(X=-2) = \frac{3}{8}, \\ P(Z=81) &= P(X^4=81) = P(X=-3) = \frac{1}{8} \end{aligned}$$

When dealing with absolutely continuous r.v., we start from the distribution function and then, if possible, we differentiate to obtain the density. Let X be a r.v. and $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ a function, then the r.v. $Y = \phi(X)$ has distribution function

$$F_Y(y) = P(Y \leq y) = P(\phi(X) \leq y) = P(X \in \phi^{-1}((-\infty, y])),$$

where $\phi^{-1}((-\infty, y]) = \{x \in \mathbb{R} : \phi(x) \leq y\}$. If ϕ is an invertible function that goes to $-\infty$ for $y \to -\infty$, this set becomes the interval $(-\infty, \phi^{-1}(y)]$ and we obtain

$$F_Y(y) = P(Y \leq y) = P(\phi(X) \leq y) = P(X \in (-\infty, \phi^{-1}(y)]) = F_X(\phi^{-1}(y)).$$

Besides if F_X and ϕ are differentiable, differentiating both sides of the above (applying the chain rule to the right one) we find the density

$$f_Y(y) = \frac{dF_X}{dy}(\phi^{-1})(y) = f_X(\phi^{-1}(y))\frac{d}{dy}(\phi^{-1}(y)) = f_X(\phi^{-1}(y))\frac{1}{\phi'(\phi^{-1}(y))}.$$

This method is often used to simulate the r.v.'s on a computer

Example 3.4.2. *Let* X *be a uniform r.v. on* $(0, 1]$*, and let us define* $Y = -\frac{1}{\lambda}\ln X$*, with* $\lambda > 0$*. Then the r.v.* Y *takes values in* $[0, +\infty)$*. and for* $y \geq 0$*, its distribution function is given by*

$$F_Y(y) = P(Y \leq y) = P(-\frac{1}{\lambda}\ln X \leq y) = P(\ln X \geq -\lambda y) = P(X \geq \mathrm{e}^{-\lambda y}) = 1 - \mathrm{e}^{-\lambda y}$$

whence, differentiating

$$f_Y(y) = \lambda \mathrm{e}^{-\lambda y}\mathbf{1}_{[0,+\infty)}(y),$$

i.e. $Y \sim \exp(\lambda)$*).*

An important class is given by the linear transformations.

Remark 3.4.1. *Let* X *be a r.v. with density* f_X *and let us set* $Y = aX + b$*,* $a \neq 0, b \in \mathbb{R}$*, then*

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P(X \leq \frac{y-b}{a}) = F_X(\frac{y-b}{a}),$$

and differentiating we have

$$f_Y(y) = \frac{1}{a}f_X(\frac{y-b}{a}).$$

As a consequence of this remark, we discover that the Normal distribution comprehends actually a family of parametrized distributions. Given $X \sim \mathcal{N}(0;1)$, i.e. with density

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}, \qquad x \in \mathbb{R},$$

if we set $Y = \sigma X + \mu$ with $\sigma > 0$ and $\mu \in \mathbb{R}$, then

$$f_Y(y) = \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

and we say that $Y \sim \mathcal{N}(\mu; \sigma^2)$ follows a Gaussian (or Normal) with parameters μ and σ^2. We remark that the corresponding density is a dilation/or contraction plus a shift of the density of the Standard Normal.

This gives a way to compute the probabilities of a Normal by using the tables of the standard Normal. Indeed,solving the linear relation for X we have that if we start from a Normal $Y \sim \mathcal{N}(\mu; \sigma^2)$, then

$$\frac{Y - \mu}{\sigma} \sim \mathcal{N}(0;1)$$

which is called the standardized of Y. So if we want to compute

$$P(a \leq Y \leq b) = P(\frac{a-\mu}{\sigma} \leq \frac{Y-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}) = \phi(\frac{b-\mu}{\sigma}) - \phi(\frac{a-\mu}{\sigma})$$

Example 3.4.3. *A temperature in Celsius is a Gaussian* $\sim \mathcal{N}(30, 4)$. *Since we want a stable temperature, it is important to compute the probability* $P(T > 34)$. *Then we may standardize* T *getting*

$$P(T > 34) = P\left(\frac{T-30}{2} > \frac{34-30}{2}\right) = P(\mathcal{N}(0,1) > 2) = 1 - \phi(2) = 1 - 0.9772 = 0.0228.$$

Remarks 3.4.1.

1. *The function* ϕ *might be also piecewise invertible, the importance is to be able to write its counterimage in terms of intervals. If* X *has density* f_X *and* $Y = X^2$, *to find* f_Y, *we remark immediately that* Y *is concentrated for* $y \geq 0$ *(being a square) and its distribution function is*

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= P(X \leq \sqrt{y}) - P(X \leq -\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}), \end{aligned}$$

whence differentiating we have

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}) = \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})).$$

For instance if $X \sim \mathcal{N}(0; \frac{1}{2\lambda})$, *for* $\lambda > 0$, *then* $Y = X^2$ *has density for* $y > 0$

$$f_Y(y) = \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})) = \frac{(\lambda)^{\frac{1}{2}}}{\sqrt{\pi}} y^{\frac{1}{2}-1} \mathrm{e}^{-\lambda y},$$

which is a $\Gamma(\frac{1}{2}, \lambda)$, *whence we also deduce that* $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

2. *In case a r.v. X is concentrated on some interval, we have to keep this in mind. For instance, for $X \sim \exp(\lambda)$ $Y = X^4$, we know that X is concentrated on $\mathbb{R}^+$, therefore*

$$F_Y(y) = P(Y \leq y) = P(0 \leq X^4 \leq y) = P(0 \leq X \leq \sqrt[4]{y}) = F_X(\sqrt[4]{y}),$$

so differentiating we obtain

$$f_Y(y) = \frac{1}{4} y^{\frac{1}{4}-1} \lambda \mathrm{e}^{-\lambda \sqrt[4]{y}}.$$

3.5 Exercises

1. Say whether the following is a probability density or not

$$P(X = 1) = \frac{1}{4},\ P(X = 2) = \frac{2}{5},\ P(X = 3) = \frac{1}{10},\ P(X = 4) = \frac{3}{10}.$$

2. Let $T \sim$geom(p). Evaluate

$$P(T = k + n | T \geq n)$$

3. There are three cards numbered 5,6 and 7 and a box with 6 white balls. A card is drawn and as many black balls as the value of the card are put in the box, then draws with replacement are performed until the first black ball is drawn. If T denotes the first time a black ball is drawn, compute the law of T.

4. A deck of cards contains 4 kings, 4 queens, 4 jacks and 4 aces, A and B draw in turn a card from the deck. A wins if he draws a king, while B wins if he draws an ace. A starts and at each turn the player, if he does not win, replaces the card in the deck, shuffles it and passes it to the other player to play.

   (a) Compute the probability that each player wins.

   (b) If D denotes the length of the game, what density does it have?

5. A fair die is tossed two times and every time if the value appeared on the die is even as many balls are put in a box, otherwise no ball is put. We denote by N_2 the number of balls in the box after two tosses of the die, compute the law of N_2.

Capitolo 4

MULTIDIMENSIONAL RANDOM VARIABLES

In many situations we might need more than a single r.v. to describe a phenomenon:

1. repeated tosses of a coin or a die;
2. a sequence of draws;
3. the geographical position of an erratic animal in a region;
4. the characteristics of a pediatric population (weight and height for instance).

In all these cases we expect to attribute probability to events that are described by a random vector or sequence rather than by a single r.v.

We therefore need to extend the previous concepts of probability density and distribution to more r.v.'s.

4.1 Joint Distributions and Independence

If we have two r.v.'s X and Y, from now on the intersection of events described by the two r.v.'s will be written by using a coma

$$P(\{X \in A\} \cap \{Y \in B\} = P(X \in A, Y \in B).$$

Definition 4.1.1. *: Let $(X_1, \dots X_k)$ be a random vector. Then we call **joint distribution function** of $X_1, \dots, X_k$*

$$F_{X_1,..,X_k}(x_1, .., x_k) = P((X_1, .., X_k) \in (-\infty, x_1] \times .. \times (-\infty, x_k]) = P(X_1 \leq x_1, .., X_k \leq x_k)$$

for $x_1, .., x_k \in \mathbb{R}$.

In this case we may say that

1. F is non decreasing in each variable;

2. F is right continuous in the following sense

$$\lim_{h\downarrow 0} F_{X_1,\dots,X_k}(x_1+h,\dots,x_k+h) = F_{X_1,\dots,X_k}(x_1,\dots,x_k)$$

3. $F_{X_1,\dots,X_k}(x_1,\dots,x_k) \to 1$ if each $x_i \to +\infty$ and $F_{X_1,\dots,X_k}(x_1,\dots,x_k) \to 0$ if $x_i \to -\infty$ for any $i = 1,\dots,k$

When the components of the random vector are discrete r.v.'s, then the probability is concentrated in points of $\mathbb{R}^n$ and we may define

Definition 4.1.2. *let $(X_1,\dots,X_k)$ be a vector of discrete r.v.'s, then we call **joint probability density** $X_1,\dots X_k$, the family of probabilities given by*

$$p(x_1,\dots,x_k) := P(X_1 = x_1,\dots,X_k = x_k),$$

for $x_1,\dots x_k \in \mathbb{R}$.

If, instead, the joint distribution function is k times differentiable with continuity then we may define the

Definition 4.1.3. *Let $(X_1,\dots,X_k)$ be a vector of absolutely continuous r.v.'s such that $F_{X_1,..,X_k}(x_1,..,x_k)$ is k times differentiable with continuity then we define **joint probability density** of $(X_1,\dots,X_k)$*

$$f_{X_1,\dots,X_k}(x_1,\dots,x_k) = \frac{\partial^k}{\partial x_1 \dots \partial x_k} F_{X_1,\dots,X_k}(x_1,\dots,x_k)$$

Similarly to the one-dimensional case, the following hold

1. $p(x_1,\dots,x_k)$, $f_{X_1,\dots,X_k}(x_1,\dots,x_k) \geq 0$, $\forall\,(x_1,\dots,x_k) \in \mathbb{R}$, for each $(x_1,\dots,x_k) \in \mathbb{R}$;
2. $\sum_{x_1}..\sum_{x_k} p_{X_1,..,X_k}(x_1,..,x_k)=1$, $\underbrace{\int_{\mathbb{R}}...\int_{\mathbb{R}}}_{k \text{ times}} f_{X_1,..,X_k}(x_1,..,x_k)dx_1..dx_k = 1$.

Whenever we have a random vector $(X_1,\dots,X_k)$ we call marginal densities, the probability density functions of each component

$$p_{X_1}(\cdot),\dots,p_{X_k}(\cdot)$$

which can be obtained from the joint density by saturating all the other components

$$p_{X_i}(x_i) = \sum_{x_1}\cdots\sum_{x_{i-1}}\sum_{x_{i+1}}\cdots\sum_{x_k} p(x_1,\dots,x_k).$$

By saturating only a group of components, we may obtain all the joint densities of any subgroup of r.v.'s. We may conclude that given a joint density, we can always deduce the marginal ones. Instead the same marginal densities can, in general, correspond to different joint densities.

The same happens in the absolutely continuous case

$$f_{X_i}(x_i) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_{i-1}}\int_{-\infty}^{x_{i+1}}\cdots\int_{-\infty}^{x_k} f_{X_1,\dots,X_k}(x_1,\dots,x_k)dx_k\dots dx_{i+1}dx_{i-1}\dots dx_1.$$

Examples 4.1.1.

1. *Let X and Y be two discrete r.v.'s taking respectively values in $\{-1,0,1,2\}$ andin $\{-2,-1,1\}$. Their joint density $p_{X,Y}(x,y)$ is*

Y/X	$X=-1$	$X=0$	$X=1$	$X=2$
$Y=-2$	$\frac{1}{8}$	$\frac{1}{12}$	α	0
$Y=-1$	0	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{8}$
$Y=1$	$\frac{1}{6}$	0	0	$\frac{1}{12}$

with $\alpha \in \mathbb{R}$. We want to compute α the marginal densities p_X e p_Y.

To find α it is enough to keep in mind that also a joint density must sum to 1, whence we deduce that $\alpha = \frac{1}{6}$.

Since there are only two r.v.'s, it was possible to represent the joint density by a table and we obtain the marginals just by summing by the rows or by the columns:

$$\begin{aligned} P(X=-1) &= \frac{1}{8}+\frac{1}{6}=\frac{7}{24},\ P(X=0)=\frac{1}{12}+\frac{1}{6}=\frac{1}{4}, \\ P(X=1) &= \frac{1}{6}+\frac{1}{12}=\frac{1}{4},\ P(X=2)=\frac{1}{8}+\frac{1}{12}=\frac{5}{24}, \\ P(Y=-2) &= \frac{1}{8}+\frac{1}{12}+\frac{1}{6}=\frac{9}{24},\ P(Y=-1)=\frac{1}{6}+\frac{1}{12}+\frac{1}{8}=\frac{9}{24}, \\ P(Y=1) &= \frac{1}{6}+\frac{1}{12}=\frac{1}{4}. \end{aligned}$$

2. *Let us consider the following joint densities given by the tables*

Y/X	$X=-1$	$X=0$	$X=2$
$Y=-2$	$\frac{1}{4}$	$\frac{1}{8}$	0
$Y=1$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$Y=0$	$\frac{1}{8}$	0	$\frac{1}{8}$

Y/X	$X=-1$	$X=0$	$X=2$
$Y=-2$	$\frac{1}{4}$	$\frac{1}{8}$	0
$Y=1$	$\frac{1}{4}$	0	$\frac{1}{8}$
$Y=0$	0	$\frac{1}{8}$	$\frac{1}{8}$

In both cases, summing rows and columns, we have

$$P(X=-1)=\frac{1}{2},\quad P(X=0)=\frac{1}{4},\quad P(X=2)=\frac{1}{4}$$

e

$$P(Y=-2)=\frac{3}{8},\quad P(Y=1)=\frac{3}{8},\quad P(Y=0)=\frac{1}{4}$$

hence we cannot establish to which joint density the marginals correspond.

The only case when the marginals identify uniquely the joint density is when there is independence.

Definition 4.1.4. *R.v.'s $X_1,\dots X_n$ are **independent** if for any choice of sets $B_1,\dots B_n \subseteq \mathbb{R}$, it holds*

$$P(X_1 \in B_1,\dots,X_n \in B_n) = P(X_1 \in B_1)\dots P(X_n \in B_n). \tag{4.1}$$

A sequence of r.v.'s $\{X_i\}_{i\geq 1}$ is independent if (4.1) is verified for any fixed $n \in \mathbb{N}$.

By taking any $x_1, \dots x_n \in \mathbb{R}$ and $B_1 = (-\infty, x_1], \dots, B_n = (-\infty, x_n]$, from (4.1)) we have

$$\begin{aligned} F_{X_1,\dots,X_n}(x_1,\dots,x_n) &= P(X_1 \le x_1, \dots, X_n \le x_n) \\ &= P(X_1 \le x_1) \dots P(X_n \le x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n), \end{aligned} \tag{4.2}$$

which implies for discrete r.v.'s

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \dots P(X_n = x_n) \tag{4.3}$$

and for absolutely continuous r.v.'s

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n). \tag{4.4}$$

Clearly, when independence occurs, by using (4.2), (4.3) or (4.4), from the marginals we may deduce the joint density.

Examples of independent r.v.'s are Bernoulli's associated to the repeated tosses of a coin or a die.

Remark 4.1.1. *When new r.v.'s are generated by applying functions to the given r.v.'s, independence is maintained. For instance if X and Y are independent and we take two functions $\phi, \psi : \mathbb{R} \longrightarrow \mathbb{R}$ then also $Z = \phi(X)$ and $W = \psi(Y)$ are independent.*

Example 4.1.1. *Let us consider the independent r.v.'s X and Y with densities*

$$\begin{aligned} &P(X=-2) = \frac{1}{4}, \quad P(X=-1) = \frac{1}{4}, \quad P(X=1) = \frac{1}{4}, \quad P(X=2) = \frac{1}{4} \\ &P(Y=-4) = \frac{1}{2}, \quad P(Y=-1) = \frac{1}{8}, \quad P(Y=1) = \frac{1}{8}, \quad P(Y=3) = \frac{1}{4}, \end{aligned}$$

then also X^2 and Y^2 are independent, indeed

$$P(X^2=1) = P(X^2=4) = \frac{1}{2}, \quad P(Y^2=1) = \frac{1}{4},\ P(Y^2=9) = \frac{1}{4},\ P(Y^2=16) = \frac{1}{2}$$

and

$$\begin{aligned} &P(X^2=1, Y^2=1) \\ =\ &P(X=1,Y=1) + P(X=-1,Y=1) + P(X=1,Y=-1) + P(X=-1,Y=-1) \\ =\ &P(X=1)P(Y=1)+P(X=-1)P(Y=1)+P(X=1)P(Y=-1)+P(X=-1)P(Y=-1) \\ =\ &\frac{1}{4}\cdot\frac{1}{8} + \frac{1}{4}\cdot\frac{1}{8} + \frac{1}{4}\cdot\frac{1}{8} + \frac{1}{4}\cdot\frac{1}{8} = \frac{1}{8} = \frac{1}{4}\cdot\frac{1}{2} = P(X^2=1)P(Y^2=1). \end{aligned}$$

Verifying the relation for the other possible values we get to the conclusion.

In the absolutely continuous case, condition (4.4) gives also a criterion to verify independence, indeed it implies that the joint density has to factorize into separate functions of the single components.

Example 4.1.2. *Let us consider the joint density*

$$f(x,y) = \frac{32\mathrm{e}^{\frac{9}{2}}}{75} xy^2 \mathrm{e}^{-\frac{1}{2}x-4y} \mathbf{1}_{\{x\geq 1, y\geq 1\}},$$

then we can immediately say that it comes from independent r.v.'s, since it can be written as $\frac{32\mathrm{e}^{\frac{9}{2}}}{75} f_1(x) f_2(y)$*, where*

$$f_1(x) = x\mathrm{e}^{-\frac{1}{2}x} \mathbf{1}_{\{x\geq 1\}}, \qquad f_2(y) = y^2 \mathrm{e}^{-4y} \mathbf{1}_{\{y\geq 1\}},$$

It remains to establish the constants for each marginal, but integrating we have

$$\int_1^{+\infty} x\mathrm{e}^{-\frac{1}{2}x} dx = \frac{6}{\sqrt{\mathrm{e}}},$$

whence

$$f_X(x) = \frac{\sqrt{\mathrm{e}}}{6} x\mathrm{e}^{-\frac{1}{2}x} \mathbf{1}_{\{x\geq 1\}}, \qquad f_Y(y) = \frac{64\mathrm{e}^4}{25} y^2 \mathrm{e}^{-4y} \mathbf{1}_{\{y\geq 1\}}.$$

4.2 Conditional distribution and density

Not in all the situations independence occurs, for instance when we consider draws with replacement, the results at each time are linked to one another.

Example 4.2.1. *In a box there are 6 black marbles and 4 white ones. A ball is drawn from the box and it is replaced together with an additional ball of the same color, then a second draw is performed. If we denote by*

$$X_i = \begin{cases} 1 \text{ if white at the } i\text{-th draw} \\ 0 \text{ black at the } i\text{-th draw} \end{cases}$$

with $i = 1, 2$*, then those r.v.'s are not independent and we want to compute the joint density, for instance the case* $P(X_1 = 1, X_2 = 1)$*.*

To answer the question, the only way is by conditioning

$$P(X_1 = 1, X_2 = 1) = P(X_2 = 1 | X_1 = 1) P(X_1 = 1) = P(X_2 = 1 | X_1 = 1) \frac{4}{10} = \frac{5}{11} \frac{4}{10} = \frac{2}{11}$$

and similarly we may compute all the other values.

In general we have

Definition 4.2.1. *Let* X *and* Y *be two discrete r.v.'s with joint density* $p_{X,Y}(x,y)$ *and marginals* $p_X(x)$ *,* $p_Y(y)$*, we define conditional density of* Y *given* X

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}, \qquad \text{where } p_X(x) \neq 0$$

Similarly in the absolutely continuous case we have

Definition 4.2.2. *Let X and Y be two absolutely continuous r.v.'s with joint density $f_{X,Y}$ and respective marginals f_X and f_Y, we define the conditional density of Y given X as*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \qquad \text{where } f_X(x) \neq 0 \tag{4.5}$$

Examples 4.2.1.

1. *The number of jobs arriving at a server depends upon the number of terminals connecting to the server, which can be at most 6. If X, denoting the number of terminals connecting to the server, is a $Bin(6, \frac{1}{2})$, when $X = k$ then the number of jobs is a r.v. following a Poisson distribution of parameter $100k$. Let us denote by N the number of jobs arriving at the server, we want to know its density.*

   What we know is that $N|X = k \sim Poisson(100k)$ and $X \sim Bin(6, \frac{1}{2})$ and we can exploit these two facts to compute the joint density for any $k = 0, \ldots, 6$, $h \in \mathbb{N}$

   $$P(X = k, N = h) = P(N = h|X = k)P(X = k) = \binom{6}{k}\left(\frac{1}{2}\right)^6 \frac{(100k)^h}{h!}\mathrm{e}^{-100k}.$$

   Consequently, to find the density of N we have to saturate the joint density with respect to k

   $$P(N = h) = \sum_{k=0}^{6} P(X = k, N = h) = \sum_{k=0}^{6} \binom{6}{k}\left(\frac{1}{2}\right)^6 \frac{(100k)^h}{h!}\mathrm{e}^{-100k}.$$

2. *If we have a set $A \subseteq \mathbb{R}^2$, we may define a bidimensional r.v. (X, Y) with uniform density on A, by setting*

   $$f_{X,Y}(x,y) = \frac{1}{area(A)}\mathbf{1}_A(x,y).$$

   In what follows we take a bidimensional r.v. (X, Y) uniform on the triangle Δ with vertices in $(0,0)$, $(1,0)$ e $(1,2)$:

   $$f_{X,Y}(x,y) = \begin{cases} 1 & (x,y) \in \Delta \\ 0 & (x,y) \in \Delta^c. \end{cases}$$

   We want to find the density of Y given $X = \frac{1}{3}$?

   By definition

   $$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}\mathbf{1}_{\{f_X(x)>0\}} \tag{4.6}$$

   in our case then for $x \in [0,1]$

   $$\begin{aligned} f_X(x) &= \int_0^{2x} dy = 2x \\ f_{Y|X}(y|x) &= \frac{1}{2x}\mathbf{1}_{\{0\leq y\leq 2x\}} \end{aligned}$$

obtaining a uniform distribution on $[0, 2x]$*, for each fixed* $x \in (0, 1]$*. By taking* $x = \frac{1}{3}$ *we get the conclusion.*

We remark that the conditional density verifies all the properties of a density whenever we fix a value x for the r.v. with respect to which we are conditioning.

We also remark that before we deduced the Binomial distribution by applying the independence of the sequence of the trials in the case of draws with replacement, while we obtained the hypergeometric distribution by using conditioning in the case of draws without replacement.

4.3 Transformations of multidimensional random variables

Similarly to what we did for a single r.v., we may generate new r.v.'s by applying real valued or vector valued functions to multidimensional r.v.'s.

For the sake of simplicity, we will present transformations involving only two r.v.'s, but many of them can be generalized to a higher number of r.v.'s.

Let X and Y be two r.v.'s with joint density $f_{X,Y}(x, y)$ and let $\phi : \mathrm{R}^2 \longrightarrow \mathrm{R}^2$ be an invertible and differentiable vector function. Let $(Z, W) = \phi(X, Y) = (\phi_1(X, Y), \phi_2(X, Y))$, then

$$\begin{aligned} P(Z \leq z, W \leq w) &= P\Big((X, Y) \in \phi^{-1}((-\infty, z] \times (-\infty, w])\Big) \\ &= \int\int_{\phi^{-1}((-\infty,z]\times(-\infty,w])} f_{X,Y}(x, y)dxdy \\ &= \int_{-\infty}^{z}\int_{-\infty}^{w} f_{X,Y}(\phi^{-1}(u, v))|J_{\phi^{-1}}(u, v))|dudv, \end{aligned}$$

where we denoted by J the Jacobian matrix

$$J_\phi = \begin{pmatrix} \dfrac{\partial\phi_1}{\partial x} & \dfrac{\partial\phi_1}{\partial y} \\ \dfrac{\partial\phi_2}{\partial x} & \dfrac{\partial\phi_2}{\partial y} \end{pmatrix}.$$

Then the vector $(Z, W) = \phi(X, Y)$ has density

$$f_{Z,W}(z, w) = \frac{1}{|J_\phi(\phi^{-1}(z, w))|} f_{X,Y}(\phi^{-1}(z, w)) \tag{4.7}$$

Example 4.3.1. *A famous example is given by the Box Muller transformation that allows to generate a Standard Normal from a uniform r.v.*

Let (R, Θ) *a random vector uniformly distributed on the rectangle* $[0, 1] \times [0, 2\pi]$ *with density*

$$f_{R,\Theta}(r, \theta) = \frac{1}{2\pi}\mathbf{1}_{[0,1]}(r)\mathbf{1}_{(0,2\pi)}(\theta).$$

We consider the following transformation $(r,\theta) \longrightarrow (z,w)$ *on the two subsets* $[0,1]\times(0,\pi]$ *and* $[0,1]\times[\pi,2\pi)$

$$(z,w) = \phi(r,\theta) = (\cos\theta\sqrt{-2\ln r}, \sin\theta\sqrt{-2\ln r}),$$

with inverse function

$$\phi^{-1}(z,w) = (\mathrm{e}^{-\frac{z^2+w^2}{2}}, \mathrm{arccotan}(\frac{z}{w})).$$

The Jacobian is hence computed

$$J_\phi = \begin{pmatrix} \cos\theta & \dfrac{-2}{2r\sqrt{-2\ln r}} & -\sin\theta\sqrt{-2\ln r} \\ \sin\theta & \dfrac{-2}{2r\sqrt{-2\ln r}} & \cos\theta\sqrt{-2\ln r} \end{pmatrix},$$

which gives

$$|detJ_\phi(\phi^{-1}(z,w))| = |-\frac{1}{r}\cos^2\theta - \frac{1}{r}\sin^2\theta| = \frac{1}{r} = \frac{1}{\mathrm{e}^{-\frac{z^2+w^2}{2}}}.$$

From formula (4.7) *we derive the density ofZ, W*

$$\begin{aligned} f_{Z,W}(z,w) &= f_{R,\Theta}(\phi^{-1}(z,w)) = \frac{1}{2\pi}\mathbf{1}_{[0,1]}(r)\mathbf{1}_{(0,2\pi)}(\theta)r \\ &= \frac{\mathrm{e}^{-\frac{z^2+w^2}{2}}}{2\pi}\mathbf{1}_{[0,1]}(\mathrm{e}^{-\frac{z^2+w^2}{2}})\{\mathbf{1}_{(0,\pi]}(\mathrm{arccotan}(\frac{z}{w})) + \mathbf{1}_{[\pi,2\pi)}(\mathrm{arccotan}(\frac{z}{w}))\} \\ &= \frac{1}{2\pi}\mathrm{e}^{-\frac{z^2+w^2}{2}}\mathbf{1}_{\mathbb{R}}(z)\{\mathbf{1}_{(0,+\infty)}(w) + \mathbf{1}_{(-\infty,0)}(w)\} \\ &= \frac{1}{2\pi}\mathrm{e}^{-\frac{z^2+w^2}{2}}, \qquad z,w\in\mathbb{R}. \end{aligned}$$

Since this is a joint density we have

$$\int_{\mathbb{R}}\int_{\mathbb{R}} f_{Z,W}(z,w)dzdw = 1 \Rightarrow \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{2\pi}\mathrm{e}^{-\frac{z^2+w^2}{2}}dzdw = 1,$$

whence

$$\begin{aligned} 1 &= \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{2\pi}\mathrm{e}^{-\frac{z^2+w^2}{2}}dzdw = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\mathrm{e}^{-\frac{z^2}{2}}dz\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\mathrm{e}^{-\frac{w^2}{2}}dw \\ &= (\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\mathrm{e}^{-\frac{z^2}{2}}dz)^2, \end{aligned}$$

that is

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\mathrm{e}^{-\frac{z^2}{2}}dz = 1.$$

i.e. we found the marginal is a standard Normal.

Actually in the above example the two r.v.'s Z, W are independent (the joint density factors out). In presence of independence often one can run more precise computations. In what follows we are going to consider some important transformations that become very manageable under independence.

Given X and Y we want to consider the following transformations

$$Z = \min(X, Y), \qquad \max(X, Y), \qquad X + Y$$

and determine their densities.

1. Let $Z = \min(X, Y)$. If X and Y are both discrete, so is Z and we have

$$P(Z = z) = P(\min(X, Y) = z) = P(X = z, Y \geq z) + P(X > z, Y = z),$$

which under independence becomes

$$P(Z = z) = P(X = z)P(Y \geq z) + P(X > z)P(Y = z).$$

Example 4.3.2. *A fair die and a fair tetrahedron are tossed until the first 4 appears on one of them. If T_1 represents the number of times the die is tossed to obtain the first 4 and T_2 the number of times the tetrahedron is tossed to obtain the first 4, then the number of tosses necessary to obtain the first 4 on either one is given by $T = \min(T_1, T_2)$ and the two r.v.'s are independent.*

In general, if $T_1 \sim$geom(p) and $T_2 \sim$geom(q) are independent r.v.'s, the density of $T = \min(T_1, T_2)$ is

$$\begin{aligned} P(T = k) &= P(T_1 = k)P(T_2 \geq k) + P(T_1 > k)P(T_2 = k) \\ &= p(1-p)^{k-1}(1-q)^{k-1} + (1-p)^k q(1-q)^{k-1} \\ &= (1-p)^{k-1}(1-q)^{k-1}[p + q(1-p)] \\ &= [p + q - pq][(1-p)(1-q)]^{k-1} = [p + q - pq][1 - (p + q - pq)]^{k-1} \end{aligned}$$

that is $T \sim$geom$(p + q - pq)$

If instead X and Y are two independent absolutely continuous r.v.'s with marginals f_X ed f_Y and distribution functions F_X and F_Y, we may look at the survival function of Z, $G(z) = P(Z > z)$, indeed we have

$$\begin{aligned} P(Z > z) &= P(\min(X, Y) > z) = P(X > z, Y > z) = P(X > z)P(Y > z) \\ &= (1 - F_X(z))(1 - F_Y(z)), \end{aligned}$$

by taking the complements

$$F_Z(z) = F_X(z) + F_Y(z) - F_X(z)F_Y(z) \tag{4.8}$$

and differentiating

$$f_Z(z) = f_X(z)(1 - F_Y(z)) + f_Y(z)(1 - F_X(z)). \tag{4.9}$$

If $X \sim \exp(\lambda)$ and $Y \sim \exp(\mu)$ we have the suggestive formula

$$f_Z(z) = \lambda \mathrm{e}^{-\lambda x}\mathrm{e}^{-\mu x} + \mu \mathrm{e}^{-\mu x}\mathrm{e}^{-\lambda x} = (\lambda + \mu)\mathrm{e}^{-(\lambda+\mu)x}$$

i.e. an exponential of parameter $\lambda + \mu$.

2. Let $Z = \max(X, Y)$. If X and Y are both discrete, so is Z and we have

$$\begin{aligned} P(Z = z) &= P(\max(X, Y) = z) = P(X = z, Y \leq z) + P(X < z, Y = z) \\ &= \sum_{y \leq z} P(X = z, Y = y) + \sum_{x < z} P(X = x, Y = z). \end{aligned}$$

If independence occurs, the previous becomes

$$P(Z = z) = \sum_{y \leq z} P(X = z)P(Y = y) + \sum_{x < z} P(X = x)P(Y = z)$$

When we have $T_1 \sim \mathrm{geom}(p)$ and $T_2 \sim \mathrm{geom}(q)$ independent r.v's, then $T = \max(T_1, T_2)$ has the following density

$$\begin{aligned} P(T = k) &= P(\max(T_1, T_2) = k) = P(T_1 = k, T_2 \leq k) + P(T_1 < k, T_2 = k) \\ &= P(T_1 = k)P(T_2 \leq k) + P(T_1 < k)P(T_2 = k) \\ &= P(T_1 = k)[1 - P(T_2 \geq k + 1)] + [1 - P(T_1 \geq k)]P(T_2 = k) \\ &= p(1-p)^{k-1}(1 - (1-q)^k) + (1 - (1-p)^{k-1})q(1-q)^{k-1} \\ &= p(1-p)^{k-1} + q(1-q)^{k-1} - (p + q - pq)(1 - (p + q - pq))^{k-1} \end{aligned}$$

When X and Y are two independent absolutely continuous r.v.'s with marginals f_X ed f_Y and distribution functions F_X and F_Y we have

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(\max(X, Y) \leq z) = P(X \leq z, Y \leq z) \\ &= P(X \leq z)P(Y \leq z) = F_X(z)F_Y(z), \end{aligned}$$

whence differentiating

$$f_Z(z) = f_X(z)F_Y(z) + f_Y(z)F_X(z). \tag{4.10}$$

Again the formula specialized for the exponential distributions is very suggestive, so let $X \sim \exp(\lambda)$ and $Y \sim \exp(\mu)$, then

$$f_Z(z) = \lambda \mathrm{e}^{-\lambda x}(1 - \mathrm{e}^{-\mu x}) + \mu \mathrm{e}^{-\mu x}(1 - \mathrm{e}^{-\lambda x}) = \lambda \mathrm{e}^{-\lambda x} + \mu \mathrm{e}^{-\mu x} - (\lambda + \mu)\mathrm{e}^{-(\lambda+\mu)x},$$

that is to say

$$f_{\max(X,Y)}(z) = f_X(x) + f_Y(y) - f_{\min(X,Y)}(z).$$

3. We are now concerned with $Z = X + Y$. Again for discrete r.v.'s we can compute the density directly obtaining

$$P(Z = z) = P(X + Y = z) = \sum_x P(X + Y = z, X = x) = \sum_x P(Y = z - x, X = x)$$

and under independence we may go one step further and write

$$P(Z = z) = \sum_x P(Y = z - x)P(X = x).$$

We are going to examine two particular cases

Let $X \sim$Poisson(λ) and $Y \sim$Poisson(μ) be two independent r.v.'s. Then $X + Y$ still takes values in $k = 0, 1, 2, \ldots$ and we have

$$P(X + Y = k) = \sum_{h=0}^{\infty} P(Y = k - h)P(X = h) = \sum_{h=0}^{k} P(Y = k - h)P(X = h),$$

the last passage due to the fact that a Poisson distribution is null for negative values. Therefore

$$\begin{aligned} P(X + Y = k) &= \sum_{h=0}^{k} P(Y = k - h)P(X = h) = \sum_{h=0}^{k} \mathrm{e}^{-\lambda}\frac{\lambda^h}{h!}\mathrm{e}^{-\mu}\frac{\mu^{k-h}}{(k-h)!} \\ &= \frac{\mathrm{e}^{-(\lambda+\mu)}}{k!}\sum_{h=0}^{k}\frac{k!}{h!(k-h)!}\mu^{k-h}\lambda^h \\ &= \frac{\mathrm{e}^{-(\lambda+\mu)}}{k!}\sum_{h=0}^{k}\binom{k}{h}\mu^{k-h}\lambda^h = \mathrm{e}^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!}, \end{aligned}$$

i.e. $X + Y$ is a Poisson($\lambda + \mu$).

The same property holds for the binomial distribution. Let $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$, $m, n \in \mathbb{N}$ be two independent r.v.'s, then $Z = X + Y \sim Bin(m + n, p)$, indeed X can be seen as the sum of n independent Bin$(1, p)$ and similarly Y can be seen as the sum of m independent Bin$(1, p)$, then

$$Z = X_1 + \cdots + X_n + Y_1 + \cdots + Y_m = Z_1 + \ldots Z_{m+n}$$

is the sum of $m + n$ independent Bin$(1, p)$.

For absolutely continuous r.v.'s we have to find first the distribution function of Z by means of the joint density. Considering immediately the independent case

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

we have

$$\begin{aligned} F_Z(z) &= P(Z \le z) = P(X+Y \le z) = \int\int_{x+y\le z} f_X(x) f_Y(y) dy dx \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y) dy dx = \int_{-\infty}^{+\infty} f_X(x) F_Y(z-x) dx \end{aligned}$$

and differentiating under the integral sign we get to

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx. \tag{4.11}$$

In the special case of independent $X \sim \mathcal{N}(0,\sigma^2)$,$Y \sim \mathcal{N}(0,\eta^2)$, applying (4.11) we can write

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\eta^2}} \mathrm{e}^{-\frac{(z-x)^2}{2\eta^2}} dx \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\eta^2}} \mathrm{e}^{-\frac{z^2-2zx+x^2}{2\eta^2}} dx \\ &= \frac{1}{2\pi\sqrt{\sigma^2\eta^2}} \mathrm{e}^{-\frac{z^2}{2\eta^2}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2\sigma^2} - \frac{x^2-2zx}{2\eta^2}} dx. \end{aligned}$$

The exponent $-\frac{x^2}{2\sigma^2} - \frac{x^2-2zx}{2\eta^2}$, completing the squares, can be rewritten as follows

$$\begin{aligned} -\frac{x^2}{2\sigma^2} - \frac{x^2-2zx}{2\eta^2} &= -\frac{(\eta^2+\sigma^2)x^2 - 2\sigma^2 zx}{2\sigma^2\eta^2} = -\frac{(\eta^2+\sigma^2)[x^2 - 2\frac{\sigma^2}{\eta^2+\sigma^2} zx]}{2\sigma^2\eta^2} \\ &= -\frac{x^2 - 2\frac{\sigma^2}{\eta^2+\sigma^2} zx + \frac{\sigma^4}{(\eta^2+\sigma^2)^2} z^2 - \frac{\sigma^4}{(\eta^2+\sigma^2)^2} z^2}{2\frac{\sigma^2\eta^2}{\eta^2+\sigma^2}} \\ &= -\frac{(x - \frac{\sigma^2}{\eta^2+\sigma^2} z)^2}{2\frac{\sigma^2\eta^2}{\eta^2+\sigma^2}} + \frac{\sigma^2}{\eta^2(\eta^2+\sigma^2)} \frac{z^2}{2} \end{aligned}$$

and consequently we have

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi\sqrt{\sigma^2\eta^2}} \mathrm{e}^{-\frac{z^2}{2\eta^2} + \frac{\sigma^2}{\eta^2(\eta^2+\sigma^2)}\frac{z^2}{2}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{(x-\frac{\sigma^2}{\eta^2+\sigma^2}z)^2}{2\frac{\sigma^2\eta^2}{\eta^2+\sigma^2}}} dx \\ &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2(\eta^2+\sigma^2)}} \frac{1}{\sqrt{\eta^2+\sigma^2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\frac{\sigma^2\eta^2}{\eta^2+\sigma^2}}} \mathrm{e}^{-\frac{(x-\frac{\sigma^2}{\eta^2+\sigma^2}z)^2}{2\frac{\sigma^2\eta^2}{\eta^2+\sigma^2}}} dx \\ &= \frac{1}{\sqrt{2\pi(\eta^2+\sigma^2)}} \mathrm{e}^{-\frac{z^2}{2(\eta^2+\sigma^2)}}, \end{aligned}$$

being the integral equal to 1 because of a density $\mathcal{N}(\frac{\sigma^2}{\eta^2+\sigma^2}z, \frac{\sigma^2\eta^2}{\eta^2+\sigma^2})$ and we can conclude that $Z = X + Y \sim \mathcal{N}(0, \eta^2 + \sigma^2)$.

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\nu, \eta^2)$, then we can write $X = W + \mu$ and $Y = U + \nu$, where $W \sim \mathcal{N}(0, \sigma^2)$ and $U \sim \mathcal{N}(0, \eta^2)$. So $Z = X + Y = W + U + \mu + \nu$ is a $\mathcal{N}(\mu + \nu, \eta^2 + \sigma^2)$. We summarize with the

Proposition 4.3.1. *Let X_i, $i = i, \dots, n$ independent r.v.'s $\mathcal{N}(\mu_i, \sigma_i^2)$, then $Z = \sum_{i=1}^{n} X_i \sim \mathcal{N}(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$.*

Example 4.3.3. *The height of the 18 year old male population is distributed as a r.v. $X \sim \mathcal{N}(177\,cm; 9\,cm^2)$. During a screening the height of a large sample of this population is measured with a toll that makes a Gaussian error $W \sim \mathcal{N}(-1\,cm; 0,4\,cm^2)$. We would like to compute the probability that the measured height is greater than or equal to 175 cm.*

The actually measured height is given by $X + W$ and we can reasonably assume that the two r.v.'s are independent, hence we know that $X + W \sim \mathcal{N}(176\,cm; 9,4\,cm^2)$, that we only have to standardize in order to compute the probability

$$P(X + W \geq 175) = P(\frac{X + W - 176}{\sqrt{9,4}} \geq \frac{175 - 176}{\sqrt{9,4}}) = 1 - \Phi(-\frac{1}{\sqrt{9,4}}) = \Phi(\frac{1}{\sqrt{9,4}}).$$

The same property is shared by the Gamma distributions, provided the second parameter remains fixed.

Proposition 4.3.2. *Let $X_i \sim \Gamma(\alpha_i, \lambda)$, $i = 1, \dots, n$, $\alpha_i, \lambda > 0$, independent r.v.'s, then*

$$\sum_{i=1}^{n} X_i \sim \Gamma(\sum_{i=1}^{n} \alpha_i, \lambda)$$

Consequently, if $X_1, \dots, X_n$ are i.i.d. $\exp(\lambda) = \Gamma(1, \lambda)$, then $X_1 + \dots + X_n \sim \Gamma(n, \lambda)$.

4. We already saw that if $X \sim \mathcal{N}(0; \frac{1}{2\lambda})$ then $Y = X^2 \sim \Gamma(\frac{1}{2}, \lambda)$, consequently if we take $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(0; \frac{1}{2\lambda})$ r.v.'s, then by proposition 4.3.2 we immediately have that

$$Z = X_1^2 + \dots + X_n^2 \sim \Gamma(\frac{n}{2}, \lambda).$$

If $n = 2k$ (even, we already know this distribution as an Erlang with n phases, when $n = 2k + 1$ (odd) this distribution is also known as $\chi^2(n)$ with n degrees of freedom. By iteration we also find

$$\Gamma(\frac{2n+1}{2}) = \frac{2n-1}{2}\Gamma(\frac{2n-1}{2}) = \dots = \frac{2n-1}{2}\frac{2n-3}{2}\dots\frac{1}{2}\Gamma(\frac{1}{2}) = \frac{(2n-1)!!}{2^n}\sqrt{\pi},$$

where we defined

$$(2n-1)!! = (2n-1)(2n-3)\dots 3 \cdot 1, \qquad (2n)!! = (2n)(2n-2)\dots 4 \cdot 2.$$

4.4 Exercises

1. If $X \sim \mathcal{N}(3;4)$ what is the density of $Y = 3X - 2$?
2. If $X \sim \mathcal{N}(3;3)$ compute $P(X \leq -1)$.
3. Let X be a r.v. with distribution function

$$F(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{50}t^2 & 0 \leq t \leq 5 \\ -\frac{1}{50}t^2 + \frac{2}{5}t - 1 & 5 \leq t \leq 10 \\ 1 & t \geq 10 \end{cases}$$

   (a) Say where is concentrated X.

   (b) Say whether X has a density and, if so, compute it.

4. For $\lambda > 1$, let us consider the function

$$f(x) = C\frac{1}{x^\lambda + 1}\mathbf{1}_{[1,+\infty)}(x)$$

   (a) Determine C so that f is a probability density.

   (b) Compute the density of $Y = \ln(X)$.

5. Let $X \sim \mathcal{N}(2,1)$, compute the density of $Y = \mathrm{e}^X$.
6. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, compute the density of $Y = X^2$.
7. Let X be a uniform r.v. on $[-1,1]$. Determine $P(|X| > \frac{1}{2})$.
8. The lifetime of a computer in years has density $\exp(\frac{1}{4})$. Buying a used computer of this type today, what is the probability it will work for other 4 years?
9. A test result of a student population is a r.v. $X \sim \mathcal{N}(100, 100)$. The result of each student is independent of all the others. We choose 5 students randomly. Compute the probability that

   (a) all their results are less or equal to 80;

   (b) exactly 3 out of 5 have results higher or equal to 84.

10. For fixed $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be i.i.d. r.v.'s so that $X_k \sim \exp(\frac{\lambda}{n})$ for $\lambda > 0$ and each $k = 1, \ldots, n$. Determine the laws of

$$Y_n = \max(X_1, \ldots, X_n) \quad \text{e} \quad Z_n = \min(X_1, \ldots, X_n).$$

Capitolo 5

EXPECTATION AND MOMENTS

The knowledge of a r.v. is well summarized by its probability density, which sometimes might be not quite complex to describe and evaluate. It is therefore quite useful to have some quantities that are able to summarize easily the behavior of a r.v.

5.1 The Expectation

If we think that the probability density is the analogous of the mass density in physics, it is quite natural to think that there must exist a concept analogous to that notion of barycenter.

Definition 5.1.1. *Let X be a r.v., we define* ***mathematical mean or expectation*** *of X the quantity*

$$\mathbb{E}(X) = \begin{cases} \sum_{i\in I} x_i P(X = x_i), & \text{if } \sum_{i\in I} p(X = x_i) = 1, \text{ for } \{x_i\}_{i\in I\subseteq\mathbb{N}} \text{ (discrete)} \\ \int_{-\infty}^{+\infty} x f_X(x)dx, & \text{if } \int_{-\infty}^{+\infty} f_X(x)dx = 1 \text{ (abs. cont.)}, \end{cases} \tag{5.1}$$

provided that

$$\sum_{i\in I} |x_i| P(X = x_i) < +\infty, \quad \int_{-\infty}^{+\infty} |x| f_X(x)dx < +\infty.$$

Other names: expected value, average.
If $\mathbb{E}(X) = 0$ then we say that the r.v. is **centered**.

Properties of the mean

1. Triangular inequality: $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.
2. Linearity: if X and Y are two r.v. with finite mean, $\alpha, \beta \in \mathbb{R}$ then

$$\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y).$$

3. $\mathbb{E}(\alpha) = \alpha$ for any constant α.

4. Monotonicity: if $X \geq 0$ then $\mathbb{E}(X) \geq 0$.

   As a consequence, if X and Y are such that $P(X \leq Y) = 1$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$, applying the monotonicity and the linearity to the mean of $Y - X \geq 0$.

   Thus, if X is bounded $|X| \leq M$, then $|\mathbb{E}(X)| \leq \mathbb{E}(|X|) \leq M$.

Proposition 5.1.1. *Let X be a r.v. and $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ a function, such that*

$$\begin{aligned} &(\textit{discrete}) \quad \sum_{i\in I} |x_i| P(X = x_i) < +\infty, \qquad \sum_{i\in I} |\phi(x_i)| P(X = x_i) < +\infty \\ &(\textit{abs.cont.}) \quad \int_{-\infty}^{+\infty} |x| f_X(x)dx < +\infty, \qquad \int_{-\infty}^{+\infty} |\phi(x)| f_X(x)dx < +\infty \end{aligned}$$

then

$$\mathbb{E}(\phi(X)) = \begin{cases} \displaystyle\sum_{i\in I} \phi(x_i) P(X = x_i) \\ \displaystyle\int_{-\infty}^{+\infty} \phi(x) f_X(x)dx \end{cases}$$

This proposition is quite important because it says that it is not necessary to know the density of $Y = \phi(X)$ to compute its mean, but we may use the density of the original r.v. X.

This proposition holds also when $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector, using the joint density of the components, for instance in the absolutely continuous case, for $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ such that

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} |\phi(x_1, \ldots, x_n)| f_{\mathbf{X}}(x_1, \ldots, x_n)dx_1, \ldots dx_n < +\infty$$

we have that

$$\mathbb{E}(\phi(X_1, \ldots, X_n)) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \phi(x_1, \ldots, x_n) f_{\mathbf{X}}(x_1, \ldots, x_n)dx_1, \ldots dx_n.$$

Applying the proposition 5.1.1 to $\phi(x) = |x|$, we may say that the mean exists whenever $\mathbb{E}(|X|) < +\infty$.

Moreover, choosing $\phi(x) = x^k, \phi(x) = (x - \mathbb{E}(X))^k$ we have

Definition 5.1.2. *Let $k \in \mathbb{N}$ and X be a r.v. such that $\mathbb{E}(|X|^k) < \infty$, then we call **moment of order** k of X the quantity $\mathbb{E}(X^k)$ e **centered moment of order** k the quantity $\mathbb{E}[(X - \mathbb{E}(X))^k]$.*

Remarks 5.1.1.

1. *We speak of centered moments because $Y = X - \mathbb{E}(X)$ has mean 0.*

2. *If X has a moment order k, then it has all the moments of order $0 \leq r \leq k$.*

3. If X and Y have each the moment of order k, then $X+Y$ has moment of order k.

When $\mathbb{E}(X^2) < +\infty$ the centered moment of order 2 is particularly important and it is called **Variance** and we write $\mathrm{Var}(X)$. Explicitly written it is

$$\mathrm{Var}(X) = \begin{cases} \sum_{i\in I}[x_i - \mathbb{E}(X)]^2 P(X = x_i), & \text{(discrete)} \\ \int_{-\infty}^{+\infty} [x - \mathbb{E}(X)]^2 f_X(x)dx, & \text{(abs. cont.)}, \end{cases} \tag{5.2}$$

hence it represents the mean quadratic error (weighted against the density) that one makes substituting the r.v. with its mean. It can be considered as an index of the dispersion of the values of the r.v. with respect to its mean. More disperse values will give a higher variance, while more concentrated one a lower variance.

Properties of the variance

1. $\mathrm{Var}(X) \geq 0$ for all X and $\mathrm{Var}(X) = 0$ if and only if X is a constant.
2. $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$, indeed
$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}(X))^2] &= \mathbb{E}[X^2 - 2X\mathbb{E}(X) + \mathbb{E}[(\mathbb{E}(X))^2] \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \end{aligned}$$
3. $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, for any $a \in \mathbb{R}$ as a matter of fact
$$\begin{aligned} \mathrm{Var}(aX) &= E[((aX) - \mathbb{E}(aX))^2] = \mathbb{E}[(a^2X^2 - (aE(X))^2] \\ &= \mathbb{E}[a^2(X - \mathbb{E}(X))^2] = a^2\mathbb{E}[(X^2 - \mathbb{E}(X))^2] = a^2\mathrm{Var}(X) \end{aligned}$$
4. $\mathrm{Var}(X + c) = \mathrm{Var}(X)$, $c \in \mathbb{R}$, since from the definition we obtain
$$Var(X + c) = \mathbb{E}([(X + c) - \mathbb{E}(X + c)]^2) = \mathbb{E}[(X + c - \mathbb{E}(X) - c)^2] = Var(X).$$

Therefore the dispersion does not change if a r.v. is shifted, but it is sensitive to changes of measure.

Example 5.1.1. *If L is a r.v. expressing the length in centimeters of the cut of a material and it has been noted that its variance is $\sigma^2_{cm} = 49cm^2$ then switching to inches, as 1 cm = 0.39 inches we have that $\sigma^2_{in} = 49 \cdot 0.39^2 in^2 = 7.477in^2$. Of course we had not a reduction of the error, but the effect is due only to having used a larger unit of measure.*

The variance is not linear, indeed if X and Y have finite variances, then we have

$$\begin{aligned} Var(X + Y) &= \mathbb{E}((X + Y)^2) - (\mathbb{E}(X + Y))^2 = \mathbb{E}(X^2 + 2XY + Y^2) - [\mathbb{E}(X) + \mathbb{E}(Y)]^2 \\ &= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - [(\mathbb{E}(X))^2 + 2\mathbb{E}(X)E(Y) + (\mathbb{E}(Y))^2] \\ &= E(X^2) - (\mathbb{E}(X))^2 + \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 + 2[\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)] \\ &= Var(X) + Var(Y) + 2[\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)], \end{aligned}$$

where

$$\mathbb{E}(XY) = \sum_x \sum_y xyP(X = x, Y = y), \quad \text{or } \mathbb{E}(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf_{X,Y}(x, y)dxdy$$

the quantity $[\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)]$ has a meaning and a name

Definition 5.1.3. *Given two r.v.'s X and Y such that $Var(X)$, $Var(Y) < +\infty$ we call* ***covariance*** *of X and Y*

$$cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \tag{5.3}$$

Explicitly, for discrete r.v.'s (5.3) becomes

$$\text{cov}(X, Y) = \sum_{i \in I} \sum_{j \in J} (x_i - \mathbb{E}(X))(y_j - \mathbb{E}(Y))P(X = x_i, Y = y_j) \quad \text{or}$$

$$\text{cov}(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))(y - \mathbb{E}(Y))f_{X,Y}(x, y)dxdy,$$

whence the covariance is giving a measure of how the two r.v.'s vary together. With a few calculations we may rewrite

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

The covariance may be positive and in this case we say that the r.v.'s are **positively correlated**, negative and we say they are **negatively correlated** or $\text{cov}(X, Y) = 0$ and we say that X and$Y$ are **uncorrelated**.

When the r.v.'s are independent it is easy to verify that

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y),$$

which implies they are also uncorrelated. The opposite is not true as we can see from the following

Example 5.1.2. *Take*

$$X = \begin{cases} 1 & p \\ -1 & 1 - p \end{cases}$$

and $Y = X^2 \equiv 1$. Then X and Y are dependent, but they are uncorrelated, indeed $0 = \mathbb{E}(X) = \mathbb{E}(X^3) = \mathbb{E}(X \cdot X^2)$, so $\mathbb{E}(XY) = 0 = \mathbb{E}(X)\mathbb{E}(Y)$.

Summarizing we may write

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2cov(X, Y).$$

and if X and Y are uncorrelated (or even better independent), we have

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y),$$

a relation we can extend to any number of independent r.v.'s $X_1, \ldots, X_n$ by iteration

$$\text{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i)$$

Remark 5.1.1. *Covariance is a bilinear form, i.e.* $\forall\, \alpha, \beta \in \mathbb{R}$ *it verifies*

1. $cov(X, Y) = cov(Y, X)$ *(simmetry)*
2. $cov\ (\alpha X + \beta Y, Z) = \alpha cov(X, Z) + \beta cov(Y, Z)$ *(linearity in each component)*
3. *the previous implies* $cov(\alpha X, \beta Y) = \alpha\beta cov(X, Y)$*, i.e. it is not invariant with respect to changes of units of measure.*

By exploiting the Cauchy-Schwarz inequality

$$|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2) \tag{5.4}$$

for X, Y with $\mathrm{Var}(X), \mathrm{Var}(Y) < +\infty$, we can define the **correlation coefficient**

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}},$$

which verifies

1. $\rho_{X,Y} = 0$ if and only if $\mathrm{cov}(X, Y) = 0$;
2. $\rho_{aX,bY} = \rho_{X,Y}$, invariant with respect to scale changes;
3. From 5.4, $-1 \leq \rho_{X,Y} \leq 1$.

5.2 Expectation and Variance of the main distributions

5.2.1 Discrete distributions

1. $X \sim \mathrm{Bin}(1, p)$, then

$$\mathbb{E}(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = 1 \cdot p = p.$$

In this case $X^2 \equiv X$ and $\mathbb{E}(X^2) = p$, so

$$\mathrm{Var}(X) = p - p^2 = p(1 - p).$$

2. X uniform on $\{x_1, \ldots, x_m\}$, then

$$\mathbb{E}(X) = \sum_{i=1}^{m} x_i P(X = x_i) = \frac{1}{m}\sum_{i=1}^{m} x_i,$$

that is the arithmetic mean of the values. Consequently

$$\mathrm{Var}(X) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mathbb{E}(X))^2.$$

3. $X \sim \text{Bin}(n,p)$. Then X can be written as $X = X_1 + \cdots + X_n$, for i.i.d. $\text{Bin}(1,p)$ i.i.d. X_i. Then by linearity

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np.$$

Being the r.v.'s independent, the same applies to the variance

$$\text{Var}(X) = \text{Var}(X_1) + \ldots \text{Var}(X_n) = np(1-p).$$

4. $X \sim$ hypergeometric(r,m,n), then

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{n} k \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}} = \sum_{k=1}^{n} k \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}} = \frac{r}{m} n \sum_{k=1}^{n} \frac{\binom{r-1}{k-1}\binom{m-1-(r-1)}{n-1-(k-1)}}{\binom{m-1}{n-1}} \\
&= \frac{r}{m} n \sum_{h=0}^{n-1} \frac{\binom{r-1}{h}\binom{m-1-(r-1)}{n-1-h}}{\binom{m-1}{n-1}} = \frac{r}{m} n
\end{aligned}$$

For the variance we first have to compute

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{k=0}^{n} k^2 \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}} = \sum_{k=0}^{n} k(k-1+1) \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}} \\
&= \sum_{k=2}^{n} k(k-1) \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}} + \sum_{k=1}^{n} k \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}} \\
&= \frac{r(r-1)}{m(m-1)} n(n-1) \sum_{k=2}^{n} \frac{\binom{r-2}{k-2}\binom{m-2-(r-2)}{n-2-(k-2)}}{\binom{m-2}{n-2}} + \frac{r}{m} n \\
&= \frac{r(r-1)}{m(m-1)} n(n-1) \sum_{h=0}^{n-2} \frac{\binom{r-2}{h}\binom{m-2-(r-2)}{n-2-h}}{\binom{m-2}{n-2}} + \frac{r}{m} n \\
&= \frac{r(r-1)}{m(m-1)} n(n-1) + \frac{r}{m} n,
\end{aligned}$$

hence

$$\text{Var}(X) = \frac{r}{m} n \Big[\frac{(r-1)}{(m-1)} (n-1) - \frac{r}{m} n + 1 \Big].$$

5. $X \sim \text{geom}(p)$, then

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=1}^{+\infty} kp(1-p)^{k-1} = \sum_{k=1}^{+\infty} (k-1+1)p(1-p)^{k-1} \\
&= \sum_{k=1}^{+\infty} (k-1)p(1-p)^{k-1} + \sum_{k=1}^{+\infty} p(1-p)^{k-1} \\
&= \sum_{k=2}^{+\infty} (k-1)p(1-p)^{k-1-1}(1-p) + 1 = \sum_{h=1}^{+\infty} hp(1-p)^{h-1}(1-p) + 1 \\
&= (1-p)\mathbb{E}(X) + 1
\end{aligned}$$

and solving the equation $\mathbb{E}(X) = \dfrac{1}{p}$.

Similarly we proceed for the variance

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = \sum_{k=1}^{\infty} (k-1+1)^2 p(1-p)^{k-1} \\
&= \sum_{k=1}^{\infty} [(k-1)^2 + 2(k-1) + 1 + 1] p(1-p)^{k-1} \\
&= (1-p) \sum_{k=2}^{\infty} (k-1)^2 p(1-p)^{k-2} + 2(1-p) \sum_{k=2} (k-1) p(1-p)^{k-2} \\
&+ \sum_{k=1}^{\infty} p(1-p)^{k-1},
\end{aligned}$$

setting $h = k - 1$ we get

$$\begin{aligned}
\mathbb{E}(X^2) &= (1-p) \sum_{h=1}^{\infty} h^2 p(1-p)^{h-1} + 2(1-p) \sum_{h=1}^{\infty} h p(1-p)^{h-1} + 1 \\
&= (1-p)\mathbb{E}(X^2) + 2(1-p)\mathbb{E}(X) + 1 = (1-p)\mathbb{E}(X^2) + 2(1-p)\frac{1}{p} + 1
\end{aligned}$$

and solving the equation we obtain

$$E(X^2) = \frac{2}{p^2} - \frac{1}{p} \Rightarrow \mathrm{Var}(X) = E(X^2) - E(X)^2) = \frac{1}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}.$$

6. $X \sim \mathrm{Poisson}(\lambda)$, then

$$\mathbb{E}(X) = \sum_{k=0}^{+\infty} k \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=1}^{+\infty} k \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} = \lambda \sum_{k=1}^{+\infty} \mathrm{e}^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \sum_{h=0}^{+\infty} \mathrm{e}^{-\lambda} \frac{\lambda^h}{h!} = \lambda.$$

For the variance

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{k=0}^{\infty} k^2 \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=0}^{\infty} k(k-1+1) \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} \\
&= \sum_{k=0}^{\infty} k(k-1) \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} + \sum_{k=0}^{\infty} k \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} \\
&= \lambda^2 \sum_{k=2}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^{k-2}}{(k-2)!} + \lambda =^{h=k-2} \lambda^2 \sum_{h=0}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^h}{h!} + \lambda
\end{aligned}$$

hence $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

5.2.2 Absolutely continuous distributions

1. (Exponential) If $T \sim \exp(\lambda)$, $\lambda > 0$, then

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^\infty x\lambda \mathrm{e}^{-\lambda x} dx = -x\mathrm{e}^{-\lambda x}\Big|_0^\infty + \int_0^\infty \mathrm{e}^{-\lambda x} dx = 0 + \frac{1}{\lambda}\int_0^\infty \mathrm{e}^{-\lambda x} dx = \frac{1}{\lambda} \\
\mathrm{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \int_0^\infty x^2 \lambda \mathrm{e}^{-\lambda x} dx - \frac{1}{\lambda^2} \\
&= -x^2\mathrm{e}^{-\lambda x}\Big|_0^\infty + \int_0^\infty 2x\mathrm{e}^{-\lambda x} dx = 0 + \frac{1}{\lambda}\int_0^\infty 2x\lambda \mathrm{e}^{-\lambda x} dx - \frac{1}{\lambda^2} \\
&= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}
\end{aligned}$$

2. (Gaussian) If $X \sim \mathcal{N}(0,1)$, then

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} dx = 0$$

since the integrand function is odd.

For the variance, the integrand is instead even and by integration by parts we have

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X^2) = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} dx = 2\int_0^{+\infty} x^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} dx \\
&= -2x\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}\Big|_0^{+\infty} + 2\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} dx = 2\frac{1}{2} = 1.
\end{aligned}$$

Thus 0 and 1, the parameters characterizing the standard Normal are the mean and the variance. The same is true for any Normal, indeed if $Y \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma \neq 0$, then it can be written as $Y = \sigma X + \mu$, thus

$$\mathbb{E}(Y) = \mathbb{E}(\sigma X + \mu) = \sigma \cdot 0 + \mu = \mu, \qquad \mathrm{Var}(Y) = \mathrm{Var}(\sigma X) = \sigma^2 \mathrm{Var}(X) = \sigma^2$$

3. (Gamma) If $X \sim \Gamma(\alpha, \lambda)\alpha, \lambda > 0$, it is possible to compute all the moments of this r.v.

$$\begin{aligned}
\mathbb{E}(X^n) &= \int_0^{+\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{n+\alpha-1} \mathrm{e}^{-\lambda x} dx = \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)\lambda^n} \int_0^{+\infty} \frac{\lambda^{n+\alpha}}{\Gamma(n+\alpha)} x^{n+\alpha-1} \mathrm{e}^{-\lambda x} dx \\
&= \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)\lambda^n} \cdot 1 = \frac{(n-1+\alpha)(n-2+\alpha)\ldots\alpha\Gamma(\alpha)}{\Gamma(\alpha)\lambda^n} = \frac{\overbrace{(n-1+\alpha)(n-2+\alpha)\ldots\alpha}^{n\,factors}}{\lambda^n},
\end{aligned}$$

whence

$$\mathbb{E}(X) = \frac{\alpha}{\lambda}, \quad \mathbb{E}(X^2) = \frac{\alpha(\alpha+1)}{\lambda^2} \quad \Rightarrow \quad Var(X) = \frac{\alpha}{\lambda^2}$$

Remark 5.2.1. *The previous formula allows also to compute all the moments of the standard Normal. As a matter of fact, for n odd*

$$\mathbb{E}(X^{(2n-1)}) = 0, \qquad n \in \mathbb{N},$$

because the integrands result odd, instead for n even we may think that $Y = X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$*, therefore*

$$\mathbb{E}(X^{2n}) = \frac{1}{\sqrt{\pi}} 2^n \frac{(2n-1)!!}{2^n} \sqrt{\pi} = (2n-1)!!$$

We remark that in all the examples shown above, the parameters characterizing the distributions are always linked to the mean and the variance.

5.3 Exercises

1. If $X \sim \text{geom}(\frac{1}{3})$, compute the variance of $2X - 4$.

2. Let X and Y be two r.v.'s with $\mathbb{E}(X) = \mu_X$, $\mathbb{E}(Y) = \mu_Y$, correlation coefficient ρ and variances $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$. Compute $\mathbb{E}((X+Y)^2)$.

3. Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s following a Poisson (λ). If $Y = X_1 + \cdots + X_n$, what is $\mathbf{Var}(Y)$?

4. Let $X_1, \ldots, X_n$ be independent r.v.'s. $\text{Bin}(1, p)$ and $Y = X_1 + \cdots + X_n$, what is the density of Y? $\mathbb{E}(Y)$, $\mathbf{Var}(Y)$?

5. If $\mathbb{E}(X) = \mu, Var(X) = \sigma$, $\mathbb{E}(Y) = \lambda$ and X , Y are independent, compute $\mathbb{E}(X^2 Y)$.

6. Let
$$f_T(t) = \lambda\beta t^{\beta-1} \mathrm{e}^{-\lambda t^\beta}, \qquad t > 0$$
which is called a Weibull distribution with parameters λ,β.

   Show that it can be obtained by the transformation $T = S^{\frac{1}{\beta}}$, where $S \sim \exp(\lambda)$ and compute $\mathbb{E}(T)$ and $\text{Var}(T)$.

7. A deck of cards contains 4 kings, 4 queens, 4 jacks and 4 aces, A and B draw in turn a card from the deck. A wins if he draws a king, while B wins if he draws an ace. A starts and at each turn the player, if he does not win, replaces the card in the deck, shuffles it and he passes it to the other player to play.

   If D denotes the length of the game, compute $\mathbb{E}(D)$and $\text{Var}(D)$.

8. A fair die is tossed two times and every time if the value appearing on the die is even as many balls are put in a box, otherwise no ball is put. We denote by N_2 the number of balls in the box after two tosses of the die, compute the law of N_2. Compute $E(N_2)$ and $\text{Var}(N_2)$.

9. In a class $\frac{2}{5}$ of the women got a high score in a test, $\frac{2}{5}$ an average score and the rest a low score. The men taking the test instead got: $\frac{1}{4}$ a high score , $\frac{1}{2}$ an average one and the rest a a low score. $\frac{3}{5}$ of the students of the class are women. Let X be a r.v. taking 1 if the student is a woman and 0 if it is a man, while Y takes 1 if a student got a low score, 2 for an average one and 3 for a high score.

   (a) Write the joint density of X, Y.

   (b) Compute the marginal densities, $\mathbb{E}(X), \mathbb{E}(Y)$, $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$.

   (c) Compute $\mathrm{cov}(X, Y)$.

10. Let X and Y be two discrete r.v.'s taking respectively values in $\{-2, 0, 1\}$ and in $\{0, 2\}$. Their joint density $p_{X,Y}(x, y)$ is given by

Y/X	$X=-2$	$X=0$	$X=1$
$Y=0$	$\frac{1}{4}$	$\frac{1}{8}$	α
$Y=2$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

    with $\alpha \in \mathbb{R}$.

   (a) Determine α and compute the marginal densities p_X, p_Y.

   (b) Compute$\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{E}(XY)$.

11. Two fair dice are tossed repeatedly. X is the number of tosses of the first die to obtain the first 2, while Y is the number of tosses of the second die to obtain the first 5 or 6.

   (a) What are the densities of X and Y?

   (b) If Z is the number of tosses to obtain both the first 2 on the first die and the first 5 or 6 on the second die , determine the density of Z.

12. A WWF group is formed by 9 Spanish kids - 3 boys and 6 girls, 3 French kids - 2 boys and 1 girl, 6 Greek kids - 4 boys and 2 girls. A kid is chosen at random If $X \in \{1, 2, 3\}$ denotes whether the selected kid i Spanish, French or Greek and $Y \in \{0, 1\}$ if it is a boy or a girl

   (a) determine the joint density of X, Y;

   (b) say if X, Y are independent;

   (c) compute the probability density of $Z = X + Y$, $\mathbb{E}(Z)$ and $\mathrm{Var}(Z)$;

   (d) compute $P(X = 1|Y = 1)$.

13. Let the two r.v.'s X and Y take respectively values in $\{-1, 0, 1\}$ and in $\{-1, 1\}$. Their joint density $p_{X,Y}(x, y)$ is given by

Y/X	$X=-1$	$X=0$	$X=1$
$Y=-1$	$\frac{1}{4}$	$\frac{1}{8}$	α
$Y=1$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

    with $\alpha \in \mathbb{R}$.

(a) Compute α and the marginal densitiesp_X e p_Y.

(b) Say if X and Y are independent, uncorrelated or correlated.

14. A is a fair coin and B a coin giving head with probability $\frac{1}{4}$. Let N_1 the number of tosses for the first head on the coin A and N_2 the number of tosses for the first head on the coin B

    (a) Compute the probability to get the first head on the coin A before than on the coin B.

    (b) Compute the mean of the number of tosses to get the first head either on A or on B.

15. (optional) Let X and Y be two uniform r.v.'s on $[0, 1]$, compute

    (a) the joint density of$X$ ed $X - Y$;

    (b) the density of $X - Y$;

    (c) $\mathrm{Var}(X - Y)$.

16. Let $X \sim \exp(\frac{1}{9})$ and $Y \sim \Gamma(5.4, 3)$ be two independent r.v.'s. let $Z = X - Y$, $W = \frac{1}{2}XY$

    (a) Compute $Var(Z)$ e $Var(W)$.

    (b) Compute $\mathrm{cov}(Z, W)$.

17. For fixed $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be i.i.d. r.v.'s all $X_k \sim \exp(\frac{\lambda}{n})$ for $\lambda > 0$, $k = 1, \ldots, n$. Find the distributions of

$$Y_n = \max(X_1, \ldots, X_n) \quad \text{e} \quad Z_n = \min(X_1, \ldots, X_n).$$

Capitolo 6

CONDITIONAL DENSITY AND CONDITIONAL EXPECTATION

When a r.v. Z depends on given r.v.'s $(X_1, \dots, X_n)$ then its expectation is computed exploiting the joint density of the vector. As we saw before, often this joint density might be expressed by means of the conditional densities.

Given two r.v.'s X and Y we recall that the conditional density of Y given X is defined as

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{P(X=x, Y=y)}{P(X=x)}, \qquad \text{where } p_X(x) \neq 0$$

if the r.v.'s are discrete with joint density $p_{X,Y}(x,y)$ with joint density $p_{X,Y}(x,y)$ and marginals $p_X(x)$, $p_Y(y)$

or

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \qquad \text{where } f_X(x) \neq 0 \tag{6.1}$$

if the r.v.'s are absolutely continuous with joint density $f_{X,Y}$ and marginals f_X and f_Y.

One might exploit relation (6.1) to deduce the joint density and the other marginal.

Example 6.0.1. *Let Y be an exponential r.v.of parameter $\lambda > 0$. If $Y = y$, then $X|Y = y$ is an exponential of parameter $\frac{1}{y}$. Hence the joint density is*

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) = \lambda \mathrm{e}^{-\lambda y} \mathbf{1}_{(0,+\infty)}(y) \frac{1}{y} \mathrm{e}^{-\frac{1}{y}x} \mathbf{1}_{(0,+\infty)}(x),$$

whence the other marginal density is given by

$$f_X(x) = \int_0^{+\infty} \lambda \mathrm{e}^{-\lambda y} \mathbf{1}_{(0,+\infty)}(y) \frac{1}{y} \mathrm{e}^{-\frac{1}{y}x} \mathbf{1}_{(0,+\infty)}(x) dy.$$

For each fixed x, the conditional density verifies the definition of density

1. $f_{Y|X}(y|x) \geq 0$;
2. $\displaystyle\int_{\mathbb{R}} f_{Y|X}(y|x)dy = \int_{\mathbb{R}} \frac{f_{X,Y}(x,y)}{f_X(x)} dy = \frac{1}{f_X(x)} \int_{\mathbb{R}} f_{X,Y}(x,y)dy = \frac{1}{f_X(x)} f_X(x) = 1.$

Consequently, it is possible to define an expectation with respect to the conditional density.

Definition 6.0.1. *We call* ***conditional expectation of*** Y ***given*** $X = x$*, which we write* $\mathbb{E}(Y|X = x)$*, the quantity*

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} y f_{Y|X}(y|x)dy, \qquad \Big(\mathbb{E}(Y) = \sum_y y p_{Y|X}(y|x)\Big) \tag{6.2}$$

For fixed x, (6.2) defines a real value, thus we are defining a function, that we denote for the moment by $\phi(x)$. If we apply this function to the r.v. X we construct a new r.v. $Y = \phi(X)$ that we call **conditional expectation of** Y **given** X and that we denote by $\mathbb{E}(Y|X)$.

Henceforth the conditional expectation itself is a r.v. and the following theorem explains it can be extremely useful to compute expectations without using the joint and the marginal densities.

Proposition 6.0.1. *Let* X *and* Y *be two r.v.'s with joint density* $f_{X,Y}(x, y)$ *and so that* $\mathbb{E}(X), \mathbb{E}(Y) < +\infty$*. Then*

$$\mathbb{E}(Y) = \int_{\mathbb{R}} \mathbb{E}(Y|X = x) f_X(x)dx, \qquad \Big(\mathbb{E}(Y) = \sum_x \mathbb{E}(Y|X = x) p_X(x)\Big)$$

This proposition might be summarized by writing that $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$.

Example 6.0.2. *Depending on the outside temperature given by a r.v.* $\Theta \sim \Gamma(\alpha, \lambda)$*, the queue at a museum follows a Poisson r.v., that is given* $\Theta = \theta > 0$ *then* $N \sim Poisson(\theta)$*. We want to compute the average length of the queue.*

Theoretically to compute the expectation, we should first write the joint density of the two r.v.'s and then obtain the marginal density

$$\begin{aligned} f_{N|\Theta}(k, \theta) &= \frac{\theta^k}{k!}\mathrm{e}^{-\theta k}, \quad f_\Theta(\theta) = \frac{\theta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\lambda\theta}\mathbf{1}_{\{\theta>0\}} \\ f_N(k) &= \int_{\mathbb{R}} \frac{\theta^k}{k!}\mathrm{e}^{-\theta k}\frac{\theta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\lambda\theta}\mathbf{1}_{\{\theta>0\}}d\theta \quad \Rightarrow \mathbb{E}(N) = \sum_{k=0}^{+\infty} k f_N(k). \end{aligned}$$

Viceversa is is much simpler to use the previous proposition, indeed given $\Theta = \theta$*,* $N|\Theta = \theta \sim Poisson(\theta)$*, hence* $\mathbb{E}(N|\Theta = \theta) = \theta$*, which defines the function* $\phi(\theta) = \theta$*. Consequently* $\mathbb{E}(N|\Theta) = \Theta$ *and*

$$\mathbb{E}(N) = \mathbb{E}(\mathbb{E}(N|\Theta)) = \mathbb{E}(\Theta) = \frac{\alpha}{\lambda}$$

We may proceed similarly ot compute the variance. Since $\mathbb{E}(Y|X)$ is a r.v.. (as function of X) we can compute its variance

$$\mathrm{V}ar(\mathbb{E}(Y|X)) = \mathbb{E}[(\mathbb{E}(Y|X))^2] - [\mathbb{E}(\mathbb{E}(Y|X)]^2 = \mathbb{E}[(\mathbb{E}(Y|X))^2] - [\mathbb{E}(Y)]^2,$$

where we applied the previous proposition in the last step.

Furthermore we define conditional variance of Y given X as the variance computed with respect to the conditional density, i.e.

$$\mathrm{V}ar(Y|X=x) = \int_{\mathbb{R}} y^2 f_{Y|X}(y|x)dy - [\mathbb{E}(Y|X=x)]^2 =: \psi(x)$$
$$\mathrm{V}ar(Y|X) = \psi(X) =: \mathrm{V}ar(Y|X) = \mathbb{E}(Y^2|X) - [\mathbb{E}(Y|X)]^2.$$

The first relation has defined a new function of x, which was applied to the r.v. X. Then the following holds

Proposition 6.0.2. *Let X and Y be two r.v.'s with finite second moments and with joint density $f_{X,Y}(x,y)$, then*

$$Var(Y) = \mathbb{E}(Var(Y|X)) + Var(\mathbb{E}(Y|X)) \tag{6.3}$$

Example 6.0.3. *Going back to the previous example we may easily compute also the variance of N, indeed applying the above proposition, we have $Var(N|\Theta=\theta) = \theta$, whence the conditional variance defines the function $\psi(\theta) = \theta$, so $\psi(\Theta) = \Theta$ and we may conclude*

$$Var(Y) = \mathbb{E}(Var(Y|X)) + Var(\mathbb{E}(Y|X)) = \mathbb{E}(\Theta) + Var(\Theta) = \frac{\alpha}{\lambda} + \frac{\alpha}{\lambda^2}$$

6.1 Exercises

1. The daily average temperature of a certain area in August is given by a r.v. $T \sim \Gamma(99;3)$. If the temperature T takes the value t, then the needed power supply in KW x 1000 for that area is distributed as a r.v. $\Gamma(t^2, 4t)$.

   (a) Denoting by X needed power supply, write the joint density of T, X.

   (b) Compute the expectation and the variance of the needed power supply X.

2. Consider the bidimensional r.v. (X,Y) with density

   $$f_{XY}(x,y) = \frac{1}{8}(6-x-y)\mathbf{1}_{(0,2)}(x)\mathbf{1}_{(2,4)}(y),$$

   compute

   (a) $\mathbb{E}(Y|X=x)$

   (b) $\mathbb{E}(Y^2|X=x)$ e $\mathrm{Var}(Y|X=x)$

   (c) $\mathbb{E}(Y)$

3. The joint density function of X and Y is given by $f_{X;Y}(x;y) = 8xy$ per $0 < x < y < 1$; find$\mathbb{E}(Y|X=x)$ and $\mathrm{Var}(Y|X=x)$.

4. Let Y be an exponential r.v. with parameter $\lambda > 0$. If $Y = y$ then $X|Y = y$ is a r.v. with distribution $\exp(\frac{1}{y})$.

   Compute $\mathbb{E}(X)$ e $\mathrm{Var}(X)$.

5. Let X and Y be two r.v.'s with joint density

$$f_{X,Y}(x,y) = \frac{1}{y}\mathrm{e}^{-y-\frac{x}{y}}\mathbf{1}_{\{x>0,y>0\}}$$

   Compute the marginal densities, $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathrm{cov}(X,Y)$, $\mathbb{E}(X^3|Y)$.

Capitolo 7

SUMMARY EXERCISES

1. Let $X_1, \dots, X_{20}$ be independent $\Gamma(0.5, 2)$ r.v.'s, what is the expectation of $X_1 + \cdots + X_{20}$?

2. Two services A, B, with separate queues, have two tellers each, A_1, A_2, B_1, B_2 with respective service times $T_{A_1} \sim \exp(\frac{1}{5})$, $T_{A_2} \sim \exp(\frac{1}{10})$, $T_{B_1} \sim \exp(\frac{1}{6})$, $T_{B_2} \sim \exp(\frac{1}{8})$ all independent.

   The first free teller at service A serves the next customer, the same for service B.

   (a) A customer arrives at the service A, whose both tellers have been just occupied. In average how long will he have to wait? The same for server B?

   (b) A customer needs both services, whose tellers have been all just occupied, so he chooses at random the queue where to wait. What is the probability he will have to wait less than 7 minutes to be served?

   (c) If the customer has waited less than 7 minutes to be served, what is probability he was stand ing in the queue for service B?

3. An athlete runs 400 m in a random time with distribution $\Gamma(460, 11.5)$. The chronometer he uses for the training sessions has a measuring error distributed as a $\mathcal{N}(0, 2)$. If we denote by T the athlete's running time and by R the measuring error, compute the expectation and the variance of the athlete's measured time.

4. Let A, B be two independent events $P(A) = \frac{1}{3}$ and $P(B) = \frac{1}{5}$, then is it true $P(A|B) = \frac{8}{15}$? Justify the answer.

5. If $P(A|B) = \frac{1}{3}$ and $P(B) = \frac{1}{4}$, how much is $P(B \cap A^c)$?

6. Say whether the following is a probability density, justifying the answer.

$$P(X = 0) = \frac{1}{4}, \quad P(X = -1) = \frac{1}{3}, \quad P(X = 1) = \frac{1}{6}, \quad P(X = 2) = \frac{1}{6}$$

7. Let X, Y, Z three independent geometric r.v.'s with respective parameters p_1, p_2, p_3, compute $E(\min(X, Y)Z^2)$.

8. If X is a Poisson (λ), what is $E(e^X)$?

9. Let X and Y be two r.v.'s with joint density $p_{X,Y}(x,y)$ given by

Y/X	$X=1$	$X=2$	$X=3$
$Y=1$	15%	5%	0
$Y=2$	5%	35%	10%
$Y=3$	0	20%	10%

   (a) Compute p_X and p_Y, $E(X), E(Y)$, $Var(X)$, $Var(Y)$.

   (b) Say whether X and Y are uncorrelated or not.

10. We have to guess three digits of a combination. The numbers can be repeated on different digits. One begins trying to guess the first digit and then one proceeds to the next one. For each digit three attempts are allowed. At each attempt it is communicated whether the digit was right or wrong.

    (a) Compute the probability to guess the first digit.

    (b) There is a win of 400 euros for guessing three digits, 200 for guessing 2, 100 for 1 and 0 if no digits was guessed. Let V be the r.v. expressing the win, compute the density of$V$.

    (c) Compute $\mathbb{E}(V)$.

11. An untidy boy has a chest of three drawers where 12 pairs of socks, 12 t-shirts and 12 shorts are mixed. Each drawer contains 12 pieces of garments (apair of socks is counted as one piece). In the first drawer there are 6 shorts, in the second 2 and in the third the rest. The boy chooses a drawer at random and takes at random a garment.

    (a) What is the probability he will not get a short?

    (b) Knowing he did not get a short, what is the probability he chose the second drawer?

12. A box has 6 white balls, a die is tossed and black balls are added to the box as many as the nearest even number greater than or equal to the result of the toss. When this operation is completed, draws with replacement are performed until the first black ball is drawn. If T is the number of draws to get the first black ball,

    (a) compute the density of T;

    (b) compute $E(T)$.

13. A bank is open from 9 a.m. to 11 a.m. Each hour a number of arriving customers is distributed as a Poisson (10) independent of the arrivals in the other hour

    (a) What is the probability that the total number of customers arrived in the two hours is greater than or equal to 15?

(b) In average compute how many people will arrive during the two hours.

14. Two fair tetrahedrons with the faces numbered from 1 to 4 are tossed. Let X the r.v. that gives the maximum value between the two results.

    (a) compute the density of X?

    (b) Compute $E(X)$ and $\mathrm{Var}(X)$?

15. Three fair dice are tossed and a player wins i euros if the number 6 appears i times and he loses 1 euro if 6 does not appear on any die. If we deniote by V lthe r.v. representing the win/loss of the player

    (a) Compute the density of V.

    (b) Compute the $\mathbb{E}(V)$.

16. At a supermarket there are three cashiers A, B, C, they have independent service times $\tau_A \sim \exp(\frac{1}{5}), \tau_B \sim \exp(\frac{1}{7}), \tau_C \sim \exp(\frac{1}{8})$. Time is measured in minutes.

    Four customers arrive at the cashiers. The first three go to one cashier each.

    (a) How long will have the fourth customer to wait if there is only one waiting line?

    (b) How long will have the fourth customer to wait if there is a waiting line in front of each cashier and the customer chooses one randomly?

17. Let X be a r.v. with density

$$f(x) = 4x\mathrm{e}^{-2x^2}\mathbf{1}_{\{x>0\}}.$$

    Compute the expectation and the variance of X.

18. The weight in grams of a certain food product is distributed as a r.v. $\mathcal{N}(500, 30)$.

    Compute the probability to have a weight less than 485 grams.

    Taking 10 packets of the same product, all produced independently, compute the probability

    (a) that at least two of the packets weighs less that 485 grams;

    (b) that at most two of the packets weighs less that 485 grams.

19. Let X be a r.v. with density

$$f(x) = \frac{3}{4}(1-x^2)\mathbf{1}_{[-1,1]}(x).$$

    Compute $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

20. A WWF group is formed by 9 Spanish kids - 3 boys and 6 girls, 3 French kids - 2 boys and 1 girl, 6 Greek kids - 4 boys and 2 girls. A kid is chosen at random If $X \in \{1, 2, 3\}$ denotes whether the selected kid is respectively Spanish, French or Greek and $Y \in \{0, 1\}$ if it is respectively a boy or a girl

    (a) determine the joint density of X, Y;

    (b) say if X, Y are independent;

    (c) compute the probability density of $Z = X + Y$, $\mathbb{E}(Z)$ and $\mathrm{Var}(Z)$;

    (d) compute $P(X = 1|Y = 1)$.

Indice

1 PROBABILITY SPACES **1**
- 1.1 Uniform sample spaces . . . 3
  - 1.1.1 Some combinatorics . . . 4
- 1.2 Exercises . . . 5

2 CONDITIONAL PROBABILITY AND INDEPENDENCE **8**
- 2.1 Conditional Probability . . . 8
- 2.2 Independence . . . 10
- 2.3 Exercises . . . 10

3 RANDOM VARIABLES **12**
- 3.1 Distribution function and density function . . . 12
- 3.2 Main discrete r.v.'s . . . 15
  - 3.2.1 Draws with replacement . . . 16
  - 3.2.2 Draws without replacement . . . 16
  - 3.2.3 Distribution with countably many values . . . 17
- 3.3 The main absolutely continuous r.v.'s . . . 19
  - 3.3.1 The exponential density . . . 19
  - 3.3.2 The Standard Normal . . . 20
  - 3.3.3 The Gamma density . . . 21
- 3.4 Transformations of random variables . . . 22
- 3.5 Exercises . . . 25

4 MULTIDIMENSIONAL RANDOM VARIABLES **26**
- 4.1 Joint Distributions and Independence . . . 26
- 4.2 Conditional distribution and density . . . 30
- 4.3 Transformations of multidimensional random variables . . . 32
- 4.4 Exercises . . . 39

5 EXPECTATION AND MOMENTS **40**
- 5.1 The Expectation . . . 40
- 5.2 Expectation and Variance of the main distributions . . . 44
  - 5.2.1 Discrete distributions . . . 44
  - 5.2.2 Absolutely continuous distributions . . . 47

5.3 Exercises 48

6 CONDITIONAL DENSITY AND CONDITIONAL EXPECTATION 51

6.1 Exercises 53

7 SUMMARY EXERCISES 55

8 APPENDIX A - FUNDAMENTALS OF SET THEORY 61

Capitolo 8

APPENDIX A - FUNDAMENTALS OF SET THEORY

Sets and Operations on sets

A set is a collection of objects. Usually a set is denoted by a capital letter, $A, B, C, ...$, the objects in the sets are called **elements** and they are usually indicated with small case letters $a, b, c....$ If the element a belongs to the set A we write $a \in A$, we write instead $a \notin A$ if a is not in in A. The set with no elements is said **empty set** and it is denoted by $\emptyset$.

Sets can be described by:

- the list of elements in the set $A = \{a, b, c, \dots\}$;
- a property

$$M = \{\text{University students attending the probability course}\}.$$

The number of elements in a set is said **cardinality** of the set.

Definition 8.0.1.

1. *If all the elements of A belong also to B, we say that A is a **subset** of B and we write $A \subseteq B$.*
$$A \subseteq B, \qquad \text{if } \forall x \in A \Rightarrow x \in B$$
2. *We call **intersection** of two sets, $A \cap B$, the set of elements that belong to both*
$$A \cap B = \{x \text{ such that } x \in A \text{ and } x \in B\},$$
*and two sets are said **disjoint** if $A \cap B = \emptyset$ (no common elements).*
3. *We call **union** of two sets, $A \cup B$ the set of all elements that are either A or in B (or in both)*
$$A \cup B = \{x \text{ such that } x \in A \text{ and/or } x \in B\},$$

4. *If we have a universal reference set, Ω, where all sets live then, given a set $A \subseteq \Omega$, we define the* ***complement*** *of A, the set of all the elements that are not in A*

$$A^c = \{x \in \Omega,\, x \notin A\}.$$

In general the difference between two sets A, B, $A \setminus B$, is the set of elements that are in A but not in B.

5. *Given two A and B, we call* ***Cartesian Product***

$$A \times B = \{(a, b) : a \in A, b \in B\},$$

the set of all the ordered pairs, where the first component is an element in A and the second an element in B.

Intersection, Union, Cartesian product can be extended to any number of sets, or even to a countable quantity of sets

$$A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^{n} A_i$$
$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^{n} A_i$$
$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n),\, a_1 \in A_1, \ldots, a_n \in A_n\}.$$

Examples 8.0.1.

1. *The previous set, M, is a subset of $L = \{$University students$\}$.*
2. *If $N = \{$University students attending the Physics course$\}$, then*

   $M \cap N = \{$University students attending both the probability and the physics courses$\}$
   $M \cup N\{$University students attending either one or both the probability and the physics courses
3. *For any two sets it holds*

$$A \cap B \subseteq A,\, B \subseteq A \cup B.$$

4. *With the previous notation*

   $M^c = \{$University students NOT attending the probability course$\}$
   $(M \cap N)^c = \{$University students attending neither the probability nor the physics course$\}$
5. *If we have the three sets*

$$A = \{\text{values of the domestic interest rate}\},$$
$$B = \{\text{values of the unemployment rate}\},$$
$$C = \{\text{values of the ratio between deficit and GDP}\},$$

   then the triple (3,10,1.2) might describe the economical state of a country

Properties of the set operations

1. **Intersection**

$$A \cap A = A$$
$$A \cap B = B \cap A$$
$$A \cap B \subseteq A; A \cap B \subseteq B$$
$$A \cap \emptyset = \emptyset$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

2. **Union**

$$A \cup A = A$$
$$A \cup B = B \cup A$$
$$A \subseteq A \cup B; B \subseteq A \cup B$$
$$A \cup \emptyset = A$$
$$A \cup (B \cup C) = (A \cup B) \cup C$$

3. **Distributive properties**

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

4. **Complement**

$$A \cup A^c = \Omega, \quad A \cap A^c = \emptyset$$
$$(A^c)^c = A, \quad A \subseteq B \Rightarrow B^c \subseteq A^c$$
$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c$$

proof:

$$(A \cup B)^c = \{x \in \Omega : x \notin A \cup B\} = \{x \in \Omega : x \notin A \text{ and } x \notin B\}$$
$$A^c \cap B^c = \{x \in \Omega : x \in A^c \text{ and } x \in B^c\} = \{x \in \Omega : x \notin A \text{ and } x \notin B\}$$
$$(A \cap B)^c = \{x \in \Omega : x \notin A \cap B\} = \{x \in \Omega : x \notin A \text{ and/or } x \notin B\},$$
$$A^c \cup B^c = \{x \in \Omega : x \notin A \text{ and/or } x \notin B\}$$

Example 8.0.1. *: Let* $A = \{1, 2, 3\}$ *and* $B = \{2, 3, 5\}$*, then*

$$A \cap B = \{2, 3\}, \quad A \cup B = \{1, 2, 3, 5\}, \quad A \setminus B = \{1\}.$$

Lecture notes of Mathematical Statistics 2018-19
part 2 - Convergence, LLN, TLC

Antonelli Fabio

2018

Capitolo 1

SEQUENCES OF R.V.'S, MOMENT GENERATING FUNCTION

In what follows the aim will be to understand the behavior of numerical samples observed as an outcome of some phenomenon. The underlying idea is that they are determined by some probabilistic law that we have to recognize.

Samples can be very large or they might be observed several times, generating long series of data.

If we think that those observations are the realizations of r.v.'s, we need mathematical tools to handle large sequences of r.v.'s $\{X_n\}_{n\in\mathbb{N}}$. In other words we need to speak about the asymptotic behavior of sequences of r.v.'s, that is we need to study their convergence properties.

In probability theory, the two main theorems study the convergence of sequences of r.v.'s: the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) and they are fundamental for Mathematical Statistics.

1.1 What convergence?

When dealing with r.v.'s, we often describe them by evaluating probabilitites and more specifically by identifying their probability distributions.

As a matter of fact, the only thing usually at our disposal is a long sequence of numbers, without being able to observe directly the phenomenon that generates these results. Therefore we rearrange the results in a frequency table that groups together "close" values and we plot the corresponding histogram to have an idea of the potential probability density determining this result.

Given a numerical sample $x_1, x_2, \dots, x_N$ of size N with only $r \leq N$ distinct values, for $i = 1, \dots, r$ we call **absolute frequency** of the value x_i

$$n_i = \#\text{times the value } x_i \text{ appears in the sample of size } N$$

and **relative frequency** of x_i

$$f_i = \frac{n_i}{N}$$

Clearly $n_1 + \cdots + n_r = N$ and $f_1 + \cdots + f_r = 1$

Individuals	Days of sick leave
a1	14
a2	10
a3	3
a4	14
a5	1
a6	21
a7	12
a8	7
a9	8
a10	10
a11	14
a12	7
a13	8
a14	10
a15	4
a16	11
a17	5
a18	7
a19	4
a20	12
a21	10
a22	3
a23	4
a24	6
a25	0
a26	2
a27	4
a28	4
a29	7
a30	3

Values	Absolute frequency	Relative frequency
0	1	1/30
1	1	1/30
2	1	1/30
3	3	1/10
4	5	1/6
5	1	1/30
6	1	1/30
7	4	2/15
8	2	1/15
9	0	0
10	4	2/15
11	1	1/30
12	2	1/15
13	0	0
14	3	1/10
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0
21	1	1/30

SAMPLE SIZE 30

Figura 1.1: Frequency table of number of days of sick leave

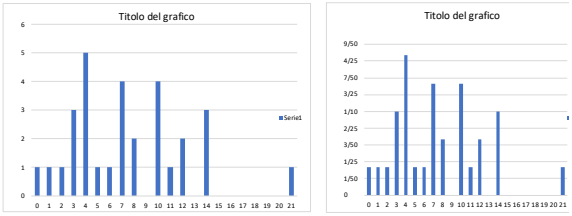

Figura 1.2: Graphs of Frequency Table

A data table might result to be very large and hence quite difficult to handle. It is then useful to have graphical methods to summarize data. They can be graphically pictured by a line or a bar graph, plotting the ordered values on the horizontal axis and correspondigly lines/bars as high as the frequencies.

This graph takes the name of **histogram**, which is usually represented with adjacent bars. If the variable takes finitely (not too) many values, it is straightforward to draw the histogram (see figure 2).

If the sample is generated by a random phenomenon, the histogram gives an idea of the underlying probability density. Thus, one hypothesizes the sample follows a density, that quite seems to match the frequency and runs some test to check whether this hypothesis can be accepted or not.

To have theoretical results about the behavior of a sequence (i.e. a very large number) of r.v.'s can be fundamental in order to understand what test we have to run, that is with which distribution we should compare our data.

Since we deal with probabilities, to understand the asymptotic behavior of a sequence of random variables it might be sufficient to check that there is a small probability that convergence might not work. This type of convergence is called **convergence in probability** and it will be used in the main theorem of next section. To prove it, we need to introduce first a very important inequality that allows to control the deviation from the expectation

by means of the variance.

Proposition 1.1.1. *(Chebyshev's inequality) Let X be a r.v. with $Var(X) < +\infty$. Then, if we set $\mu := \mathbb{E}(X)$ and $\sigma^2 := Var(X)$, for any $i\ \epsilon > 0$ it holds*

$$P(|X - \mathbb{E}(X)| > \epsilon) \leq \frac{Var(X)}{\epsilon^2} \tag{1.1}$$

Dimostrazione. Given $\epsilon > 0$, consider the set $A = \{|X - \mathbb{E}(X)| > \epsilon\}$,whence $A^c = \{|X - \mathbb{E}(X)| \leq \epsilon\}$. We may write

$$Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[(X - \mathbb{E}(X))^2(\mathbf{1}_A + \mathbf{1}_{A^c})] > \epsilon^2 P(A) + 0$$

obtaining the statement. □

Example 1.1.1. *In repeated tosses of a coin, we do not know whether it is fair or not. This means that the success/insuccess scheme*

$$X_i = \begin{cases} 1 & \text{if } H \text{ with prob. } p \\ 0 & \text{if } T \text{ with prob. } 1-p, \end{cases}$$

p is unknown. To find out if the coin is fair, we toss it 100 times. If the coin is fair then $S_{100} = X_1 + \cdots + X_{100}$ representing the number of successes is a $Bin(100, \frac{1}{2})$ r.v., whence $\mathbb{E}(S_{100}) = 100 \cdot \frac{1}{2} = 50$ and $Var(S_{100}) = 100 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25$. Applying Chebyshev's inequality, we should have

$$P(|S_{100} - 50| > 10) \leq \frac{25}{100} = 0.25,$$

In other words, on 100 tosses we expect the number of heads to be between 40 and 60 with at least 75% of probability.

Definition 1.1.1. *Let $\{X_n\}_n$ and X r.v.'s, then we say that $X_n \longrightarrow X$ in probability if $\forall \eta, \epsilon > 0$ there exists an $n(\epsilon, \eta) \in \mathbb{N}$ such that*

$$P(|X_n - X| > \epsilon) \leq \eta, \qquad \text{for } n > n(\epsilon, \eta)$$

or equivalently

$$\lim_{n \to +\infty} P(|X_n - X| > \epsilon) = 0, \qquad \forall\ \epsilon > 0$$

and we write $X_n \longrightarrow_P X$.

Chebyshev's inquality implies this kind of convergence.

Example 1.1.2. *For $n \in \mathbb{N}$, let $X_n \sim \mathcal{N}(5, \frac{1}{n})$. By Chebyshev's inequality, for any $\epsilon > 0$ we have*

$$P(|X_n - \mathbb{E}(X_n)| > \epsilon) \leq \frac{Var(X_n)}{n\epsilon^2} \longrightarrow 0, \quad \text{as } n \to 0$$

that is

$$P(|X_n - 5| > \epsilon) \leq \frac{1}{\epsilon^2}, \qquad \forall\ \epsilon > 0$$

i.e. $X_n \longrightarrow_P 5$.

Often r.v.'s are identified solely by means of their distribution (or density) functions, it is therefore useful to have a notion of convergence involving directly the distribution functions of the sequence of r.v.'s.

Definition 1.1.2. *Let $\{X_n\}_n$ and X be r.v.'s with respective distribution functions $\{F_n\}_n$ e F. We say that X_n converges in distribution (or in law) to X if*

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

at any continuity point of *F and we write*

$$X_n \Rightarrow X, \quad or \quad X_n \longrightarrow_{\mathcal{L}} X, \quad or \quad X_n \longrightarrow_d X.$$

The first notion of convergence is stronger than the second, but they coincide when the limit r.v. X is a constant.

1.2 The Law of Large Numbers

From now on, when we refer to a **random sample of size n from the population $F(x)$ (or $f(x)$)** we mean a sequence of independent r.v.'s $X_1, \ldots, X_n$ with the same distribution function $F(X)$ (or density function $f(x)$). In this case we say that the sample is independent and identically distributed (i.i.d.). Often the dependence of the density on a set of parameters $\theta \in \mathbb{R}^k$ will be explicitly written by saying that the sample follows a density $f(x|\theta)$.

Example 1.2.1. *Let $X_1, \ldots, X_n$ be a random sample from a density exp(λ), then we write that the joint density is given by*

$$f(x_1, \ldots, x_n|\lambda) = \lambda^n \mathrm{e}^{-\lambda \sum_{i=1}^n x_i}.$$

When evaluating the expectation and the variance of the most common distributions, we noticed that in all cases they were related to the parameters characterizing them.

Hence it might be extremely useful to have a tool that allows to estimate the expectation of the random sample from the observed data, so that also the parameters of the law producing the data could be identified.

Theorem 1.2.1. *(Weak Law of Large Numbers) Let $\{X_n\}_n$ be a sequence of i.i.d. r.v.'s with expectation $\mathbb{E}(X_1) = \mu$ and variance $Var(X_1) = \sigma^2 < +\infty$. Let*

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k. \tag{1.2}$$

Then for any $\epsilon > 0$ it holds

$$\lim_{n\to+\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0. \tag{1.3}$$

Dimostrazione. The proof is a direct applicaton of Chebyshev's inequality to the r.v. $\overline{X}_n$. Indeed

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}(\frac{1}{n}\sum_{k=1}^{n} X_k) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k) = \frac{1}{n}n\mu = \mu$$

and, by the independence of the r.v.'s,

$$Var(\overline{X}_n) = Var(\frac{1}{n}\sum_{k=1}^{n} X_k) = \frac{1}{n^2}\sum_{k=1}^{n} Var(X_k) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

By Chebyshev's inequality we then have

$$\lim_{n\to+\infty} P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}, \longrightarrow 0 \quad as \quad n \to +\infty \quad \forall\, \epsilon > 0.$$

□

The sequence of r.v.'s $\{\overline{X}_n\}_n$ is called the sequence of **sample means** .

This theorem affirms that if we have a sequence of independent trials, even ignoring the probability law generating them, we may always estimate the sample's expectation by computing the arithmetic mean of the realized data set. Eventually all realizations should lead to the same result.

Definition 1.2.1. *Let $X_1, \ldots, X_n$ be a random sample of size n and $t : \mathbb{R}^n \longrightarrow \mathbb{R}^d$, for some $d \in \mathbb{N}$, then $Y = t(X_1, \ldots, X_n)$ is called a **statistic**. The probability distribution of a statistic is called **sampling distribution**.*

Statistics may be used to estimate parameters on the basis of observations. From the Law of Large numbers we may conclude that the sample mean $\overline{X}_n$ is a statistic and that estimates the expectation of the sample. In this case we say that the statistic is a **point estimator**.

When speaking of a sample coming from a Poisson distribution, the sample mean $\overline{X}_n$ is a point estimator directly of the parameter λ characterizing the density, but if we a drawing a sample from a $\Gamma(\alpha, \lambda)$, then $\overline{X}_n$ actually estimates $\frac{\alpha}{\lambda}$, that is a function $h(\alpha, \lambda)$ of the two parameters characterizing the density. In general we have

Definition 1.2.2. *Given a sample $X_1, \ldots, X_n, \ldots$ from a population $f(x|\theta)$, where $\theta \in \mathbb{R}^k$ for some $k \in \mathbb{N}$, we call **estimator of** $h(\theta)$ the sequence of statistics*

$$T_n = t_n(X_1, \ldots, X_n)$$

*estimating $h(\theta)$. An estimator is called **unbiased** if $\mathbb{E}(T_n) = h(\theta)$, for all $n \in \mathbb{N}$.*

From the LLN we may conclude that

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \Rightarrow t_n(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n)$$

is an unbiased estimator of the expectation μ if the sample is i.i.d with this expectation.

Example 1.2.2. *Suppose we want to establish whether a coin is fair or not. Then we may toss repeatedly the coin generating the Bernoulli sequence of tosses*

$$X_i = \begin{cases} 1 & p \\ 0 & 1-p, \end{cases}$$

where p is the unknown probability of Head. We know that $\mathbb{E}(X_i) = p$ for all i and the LLN affirms that $\overline{X}_n \longrightarrow \mathbb{E}(X_1) = p$, hence we may conclude

$$\frac{number\ of\ heads\ in\ n\ tosses}{total\ number\ of\ tosses} = \overline{X}_n \longrightarrow p.$$

In other words the relative frequency gives an approximation of p, the parameter that tells us if the coin is fair. If, after quite many tosses, we realize that the resulting $\overline{X}_n$ is stable around $\frac{1}{2}$ (up to any decimal digit we choose), we might conclude that the coin is fair, otherwise it is not.

A natural question that arises from the previous example is: how many tosses are necessary to draw our conclusion?

Chebyshev 's inequality provides a tool to answer this question.

Example 1.2.3. *Referring again to the multiple tosses of a coin, suppose we consider admissible an error of 1% with a probability at most of $\frac{1}{100}$. From Chebyshev's inequality we have*

$$P(|\overline{X}_n - p| > \frac{1}{100}) \leq \frac{Var(\overline{X}_n)}{(\frac{1}{100})^2} = \frac{p(1-p)}{n} 100^2$$

so we want set this quantity less than $\dfrac{1}{100}$. But the estimate depends, through the variance, on the unknown parameter p (that we want to estimate). To make it not depend upon p, we notice that $p(1-p) \leq \dfrac{1}{4}$ for any $0 < p < 1$, hence we may impose

$$\frac{p(1-p)}{n} 100^2 \leq \frac{2500}{n} \leq \frac{1}{100}$$

whence $n \geq 250000$.

Actually the empirical trials to perform are much fewer, justified by another theoretical result that will improve our estimates.

As a matter of fact the LLN has another "stronger version" called the **Strong Law of Large Numbers**, because the convergence will not depend on ϵ any longer, this type of convergence is called $P-$ **almost surely**

Theorem 1.2.2. *(Strong Law of Large Numbers) Let $\{X_n\}_n$ be a sequence of i.i.d. r.v.'s with expectation $\mathbb{E}(X_1) = \mu$ and variance $Var(X_1) = \sigma^2 < +\infty$ and $\overline{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$.*

Then

$$P(\lim_{n \to +\infty} \overline{X}_n = \mu) = 1 \tag{1.4}$$

and we write $X_n \longrightarrow X$ $P-a.s.$, as $n \to +\infty$.

By this version of the LLN we know we may identify, with certainty, the mean of the sample with any precision we wish, provided to run a large enough number of trials.

Remark 1.2.1. *Most important, both types of convergence, exhibit a continuity property, that is the convergence is tranferrred to any continuous trasformation of the sample*

If $X_n \longrightarrow_P X$ (or $X_n \longrightarrow X$ $P-a.s.$) as $n \to +\infty$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$ is a continuous function, then

$$f(X_n) \longrightarrow_P f(X) \quad (or\ f(X_n) \longrightarrow f(X) \quad P-a.s.) \qquad as\ n \to +\infty. \tag{1.5}$$

Frow the LLN and Remark 1.2.1 we may find the estimator for the parameters of many distributions.

1. If the sample is drawn from a $\mathcal{N}(\mu, \sigma^2)$ or from a Poisson(λ), then $(\overline{X}_n)$ is an estimator of the expectation, that is the parameter μ in the first case and the parameter λ in the second.

2. If the sample is drawn from a geom(p) or from an exp(λ), then $(\overline{X}_n)$ is an estimator of $\dfrac{1}{p}$ in the first case or of $\dfrac{1}{\lambda}$ in the second. Consequently, by Remark 1.2.1, in both cases $(\overline{X}_n)^{-1}$is an estimator of the parameter, as $g(x) = \frac{1}{x}$ is continuous for $x > 0$.

3. If we are dealing with a e $\Gamma(\alpha, \lambda)$, then $(\overline{X}_n)$ estimates the ratio $\dfrac{\alpha}{\lambda}$, but none separately.

Remark 1.2.1 has also an important consequence that allows to provide an estimator for the variance.

Let us assume that the random sample $\{X_n\}_n$, with unknown expectation μ and variance σ^2, has finite moments up to the fourth order. Assuming for the moment to know μ, also the sequence $\{(X_n - \mu)^2\}_n$ is formed by i.i.d. r.v.'s with finite mean and variance, thus applying the LLN we have

$$\Sigma^2 := \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2 \longrightarrow \ \mathrm{Var}(X) = \sigma^2.$$

This estimator is even unbiased since

$$\mathbb{E}(\Sigma^2) = \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2\right) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}((X_k - \mu)^2) = \frac{1}{n}n\sigma^2 = \sigma^2.$$

This estimator cannot be exploited if μ is unknown as it is often the case, but it is still possible to construct an estimator of the variance. Indeed if we define

$$S_n^2 := \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2, \tag{1.6}$$

we have

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu + \mu - \overline{X}_n)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2 + \frac{2}{n}\sum_{k=1}^{n}(X_k - \mu)(\mu - \overline{X}_n) + (\mu - \overline{X}_n)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2 + 2(\mu - \overline{X}_n)(\overline{X}_n - \mu) + \ (\mu - \overline{X}_n)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2 - (\mu - \overline{X}_n)^2 \end{aligned}$$

The first term goes to σ^2 as we just showed, while the second goes to 0, since the function $g(x) = (x-\mu)^2$ is continuous.

Unfortunately the estimator S^2 is not unbiased, since from above we have

$$\begin{aligned} \mathbb{E}(S_n^2) &= \mathbb{E}\Big(\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2\Big) - \mathbb{E}((\mu - \overline{X}_n)^2) \\ &= \frac{1}{n}\sum_{k=1}^{n} Var(X_k) - Var(\overline{X}_n) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2, \end{aligned}$$

therefore if we define

$$s^2 = \frac{n}{n-1}S_n^2 := \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 \longrightarrow \sigma^2$$

this is an unbiased estimator.

The LLN can be exploited to estimate the probability density of a discrete r.v. that takes distinct values $\{x_1, \ldots, x_m\}$. Indeed as for the Bernouilli r.v.'s, we denote by $p_i = P(X = x_i)$ the unknown density and we take a sample of size n, $X_1, \ldots, X_n$, and we count the number of times the value x_i appeared

$$\frac{\#\text{ times } x_i \text{ appeared}}{n} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{\{X_k = x_i\}}$$

which is the **sample frequency**. The LLN implies that this sum converges to the unknown value p_i since the r.v.'s $\mathbf{1}_{\{X_k=x_i\}} \sim Bin(1, p_i)$ are independent.

This explains why in a numerical sample often the sample frequecy substitutes the density.

1.3 The Moment Generating Function

In order to deal with convergence in distribution we need some specific mathematical tools that allow to reconstruct the probability density of a r.v. by means of computationally

easier quantities. These are called generating functions and we here analyze a specific one: the **moment generating function**.

Definition 1.3.1. *Let X be a r.v. (or a random vector $\mathbf{X} = (X_1, \ldots, X_d)$), we call **moment generating function** of X*

$$M_X(t) = E(\mathrm{e}^{\mathbf{t}\cdot\mathbf{X}}), \qquad \mathbf{t} \in \mathbb{R}^d,$$

where $\mathbf{t} \cdot \mathbf{X}$ denotes the scalar product of two vectors in $\mathbb{R}^d$.

The function M_X makes sense only for those $\mathbf{t}$ that make the expectation finite. From now one we will consider only the one-dimensional case.

We remark that for any r.v. X, $M_X(0) = 1$ and since it can be proven that this is a continuous function, this implies that $M_X(t)$ is defined in some neighborhood of 0.

MGF's provide a tool to identify uniquely the distribution function. We give the following theorem without a proof.

Proposition 1.3.1. *Let X and Y be two r.v.'s with respective mgf's M_X and M_Y so that $M_X(t) = M_Y(t)$ for $t \in (a, b)$ for some open interval. Then X and Y have the same distribution function.*

Often, it is possible to compute the moment generating function of a combination of r.v.'s more easily than computing directly its distribution function. If this is the case, then one can employ Theorem 1.3.1 to obtain the resulting distribution function.

Properties of the moment generating function

1. $M_X(0) = 1$ for any r.v. X.
2. If X is a r.v. with mgf M_X defined in a neighborhood of 0, then for any $a, b \in \mathbb{R}$ we have
$$M_{aX+b}(t) = E(\mathrm{e}^{t(aX+b)}) = E(\mathrm{e}^{taX}\mathrm{e}^{tb)}) = \mathrm{e}^{tb}E(\mathrm{e}^{(ta)X}) = \mathrm{e}^{tb}M_X(at).$$
3. If X and Y are independent r.v.'s with respective mgf's M_X and M_Y defined on a common neighborhood of 0, then we have
$$M_{X+Y}(t) = E(\mathrm{e}^{t(X+Y)}) = E(\mathrm{e}^{tX}\mathrm{e}^{tY)}) = E(\mathrm{e}^{tX})E(\mathrm{e}^{tY)}) = M_X(t)M_Y(t).$$
4. If X is a r.v. with mgf M_X n times differentiable with continuity in a neighborhood of 0, then it holds
$$M_X^{(k)}(0) = E(X^k), \qquad \forall\, k \leq n$$
since
$$M_X^{(k)}(t) = \frac{d^k}{dt^k}E(\mathrm{e}^{tX}) = E(\frac{d^k}{dt^k}\mathrm{e}^{tX}) = E(X^k\mathrm{e}^{tX})$$
and evaluating the expression at 0 we have the previous equality.

Moment generating function of the main distributions

1. $X \sim$ Bin $(1,p)$, then

$$M_X(t) = \mathrm{e}^t p + 1 - p = p(\mathrm{e}^t - 1) + 1, \qquad t \in \mathbb{R}.$$

2. $X \sim$ Bin (n,p), then $X = X_1 + \cdots + X_n$ with independent $X_i \sim$ Bin $(1,p)$, hence

$$M_X(t) = M_{X_1}(t) \ldots M_{X_n}(t) = (\mathrm{e}^t p + 1 - p)^n, \qquad t \in \mathbb{R}.$$

3. $X \sim$ Poisson(λ), then

$$M_X(t) = \sum_{k=0}^{\infty} \mathrm{e}^{tk} \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} \frac{(\mathrm{e}^t \lambda)^k}{k!} = \mathrm{e}^{-\lambda(1-\mathrm{e}^t)}, \quad t \in \mathbb{R}.$$

4. $X \sim$ geom(p), then

$$M_X(t) = \sum_{k=1}^{\infty} \mathrm{e}^{tk} p(1-p)^{k-1} = p\mathrm{e}^t \sum_{k=1}^{\infty} [\mathrm{e}^t(1-p)]^{k-1} = \frac{p\mathrm{e}^t}{1 - e^t(1-p)}, \quad t < \ln(\frac{1}{1-p}).$$

5. $X \sim \exp(\lambda)$, then

$$M_X(t) = \int_0^{\infty} \mathrm{e}^{tx} \lambda \mathrm{e}^{-\lambda x} dx = \frac{\lambda}{\lambda - t}, \qquad t < \lambda.$$

6. $X \sim \mathcal{N}(0;1)$. The computation is, as usual, performed by completing the squares

$$\begin{aligned} M_X(t) &= \int_{\mathbb{R}} \mathrm{e}^{tx} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2 - 2tx}{2}} dx \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2-2tx}{2}} dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2-2tx+t^2}{2}} \mathrm{e}^{\frac{t^2}{2}} dx \\ &= \mathrm{e}^{\frac{t^2}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-t)^2}{2}} = \mathrm{e}^{\frac{t^2}{2}}. \end{aligned}$$

Therefore we may compute the mgf of any $Y \sim \mathcal{N}(\mu;\sigma^2)$, by property 2., obtaining

$$M_Y(t) = \mathrm{e}^{t\mu} \mathrm{e}^{\sigma^2 \frac{t^2}{2}} \qquad t \in \mathbb{R},$$

7. $X \sim \Gamma(\alpha, \lambda)$, then

$$\begin{aligned} M_X(t) &= \int_0^{\infty} \mathrm{e}^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\lambda x} dx = \int_0^{\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-(\lambda-t)x} dx \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \int_0^{\infty} \frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-(\lambda-t)x} dx = \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \qquad t < \lambda. \end{aligned}$$

1.4 The Central Limit Theorem

The moment generating function is a tool to verify the convergence of sequences of the distribution functions of r.v.'s.

Theorem 1.4.1. *(Continuity theorem) Let $\{X_n\}_n$ and X v.a. Then as $n \to +\infty$, $X_n \Rightarrow X$ if and only if $M_{X_n}(t) \longrightarrow M_X(t)$ for all t in some neighborhood of 0.*

Thus convergence in distribution gets translated into the convergence of the moment generating function and once we find the limit of the mgf sequence we know this is in a 1:1 correspondence with a distribution function.

We can now prove another result for the sum of r.v.'s , which is far more accurate than the LLN.

Theorem 1.4.2. *(Central Limit Theorem)*
Let X_n a sequence of i.i.d. r.v.'s with finite mean μ and variance σ^2.
Let us denote by $S_n = X_1 + \cdots + X_n$, then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow X \sim \mathcal{N}(0;1), \quad \textit{per } n \to +\infty$$

Dimostrazione. Without loss of generality we may consider $\mu = 0$ (otherwise we repeat the proof using the r.v.'s $\{X_n - \mu\}$).

The moment generating function of $\dfrac{S_n}{\sigma\sqrt{n}}$ is

$$M_{\frac{S_n}{\sigma\sqrt{n}}}(t) = M_{S_n}(\frac{t}{\sigma\sqrt{n}}) = M_{X_1+\cdots+X_n}(\frac{t}{\sigma\sqrt{n}}) =^{\text{ind.}} \prod_{i=1}^{n} M_{X_i}(\frac{t}{\sigma\sqrt{n}}) =^{\text{i.d.}} \Big(M_{X_1}(\frac{t}{\sigma\sqrt{n}})\Big)^n.$$

But X_1 has finite first and second moments, so its mgf is twice differentiable with continuity and we can approximate by Taylor expansion around 0 up to the second order getting to

$$\begin{aligned} M_{X_1}(\frac{t}{\sigma\sqrt{n}}) &= M_{X_1}(0) + M'_{X_1}(0)\frac{t}{\sigma\sqrt{n}} + M''_{X_1}(0)\frac{t^2}{2\sigma^2 n} + o(\frac{1}{n}) \\ &= 1 + \frac{t^2}{2n} + o(\frac{1}{n}) \end{aligned}$$

and we may conclude that

$$\Big(M_{X_1}(\frac{t}{\sigma\sqrt{n}})\Big)^n = \Big(1 + \frac{t^2}{2n} + o(\frac{1}{n})\Big)^n \longrightarrow \mathrm{e}^{\frac{t^2}{2}}$$

that is to say the mgf of a standard Normal. □

Remark 1.4.1. *The CLT could be stated directly in terms of the sample mean* $\bar{X}_n = \dfrac{S_n}{n}$ *rather than of the sum. Indeed, for any* i*, we have* $E(X_i) = \mu$ *and* $Var(X_i) = \sigma^2$*, then*

$$E(S_n) = n\mu, \quad Var(S_n) = n\sigma^2, \quad E(\bar{X}_n) = \mu, \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

whence

$$\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \frac{S_n - n\mu}{\sqrt{\sigma^2 n}} = n\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2 n}} = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}}$$

The CLT provides quite accurate estimates already with a sample of size $n = 30$ or 50.

Examples 1.4.1.

1. *As we said before, to understand whether a coin is fair, it might be reasonable to estimate the distance between the sample mean and 0.5. Let us say that an error of 0.01 is admissible for us, i.e.*

   $$P(|\bar{X}_n - \frac{1}{2}| > 0.01)$$

   with a probability at most equal to 0.003.

   By LLN, recalling that $p(1-p) \leq \frac{1}{4}$ *for all* $0 < p < 1$*, we have*

   $$P(|\bar{X}_n - \frac{1}{2}| > 0.01) \leq \frac{10000}{4n} \leq \frac{3}{1000}$$

   if $n > 833334$*, hence we should observe about 800000 tosses to say if a sample mean giving the value* 0.49 *is acceptable*

   By the CLT instead we have

   $$\begin{aligned} P(|\bar{X}_n - p| > 0.01) &= P(\bar{X}_n - p > 0.01) + P(\bar{X}_n - p < -0.01) \\ &= P(\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} > \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}) + P(\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} < \frac{-0.01\sqrt{n}}{\sqrt{p(1-p)}}) \\ &\simeq (1 - \Phi(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}})) + \Phi(\frac{-0.01\sqrt{n}}{\sqrt{p(1-p)}}) \\ &= 2(1 - \Phi(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}})) = 2(1 - \Phi(2 \cdot 0.01\sqrt{n})), \end{aligned}$$

   where in the last passage we substituted $p = \frac{1}{2}$*. By the simmetry of the standard Normal we know that this probability is less than* $2 \cdot 0.0014 = 0.0028$ *as soon as*

   $$2 \cdot 0.01\sqrt{n} \geq 3 \quad \Rightarrow n \geq 22500.$$

2. *The concentration of a polluting agent in a specimen of water from a certain area follows an exponential distribution of parameter 2. Let* $X_1, \ldots X_{100}$ *be 100 independently collected specimens. Compute*

(a) The probability the concentration in a specimen is greater than or equal to 1;

(b) approximately the probability total number of specimens where the concentration was over 1 is bigger than or equal to 20.

The answer to the first question comes directly from the properties of the exponential distribution, indeed $P(X_1 > 1) = \mathrm{e}^{-2} \simeq 0.14$.

For the second question we use the Normal approximation. Infact we constructed new independent Bernoullian r.v.'s given by

$$Y_i = \begin{cases} 1 & \text{se } X_i > 1 \\ 0 & \text{se } X_i \leq 1 \end{cases} \qquad i = 1, ..., 100$$

that mark 1 anytime the concentration goes over 1.

Hence $S_{100} = \sum_{i=1}^{100} Y_i$ *is the number of times the threshold 1 was passe in the 100 observations and this must be a* $Bin(100, 0.14)$*, whence* $E(S_{100}) = 14$ *and* $Var(S_{100}) = 11.7$*, so applying the CLT we get*

$$P(S_{100} > 20) = P(\frac{S_{100} - 14}{\sqrt{11.7}} > \frac{20 - 14}{\sqrt{11.7}}) \simeq 1 - \Phi(1.75) = 1 - 0.95994 = 0.04006.$$

1.5 Exercises

1. Give the definition of moment generating function and show its properties.
2. State and prove the Central Limit Theorem.
3. State and prove the Law of Large Numbers.
4. If$X \sim \mathcal{N}(3; 2)$ and $Y \sim \exp(2)$ are independent r.v.'s , what is the moment generating function of $2X + 3Y$?
5. The moment generating function of i.d. independent X_1, X_2 is given by
$$M(t) = \Big(\frac{0.5}{0.5 - t}\Big)^2, \qquad t < \frac{1}{2},$$
what is the mgf of $Y = -X_1 + 4X_2 + 1$ and say where it is defined.
6. Let X be a r.v. with density $f_X(x) = \frac{1}{2}\mathrm{e}^{-|x|}$, compute its mgf.
7. In a production line a dangerous gas is employed. An alarm rings every time the level of that gas in the air is higher than a given threshold. In that case production is interrupted and resumed the next day. The alarm system is defective and with probability $\frac{15}{16}$ it rings because of an actual danger, while with probability $\frac{1}{10}$ the alarm rings even though the threshold was not exceeded. The threshold is exceeded every day with probability $\frac{1}{25}$.

(a) Compute the probability that the alarm might ring in any given day.

(b) Every day works independently of all the others. Compute approximately the probability thatin 100 days the alarm rings more than 20 times

8. Given $X \sim \mathcal{N}(\mu; \sigma^2)$ compute its fourth moment.

9. If we have a sequence of i.i.d. r.v.'s $\{X_n\}_n \sim \mathcal{N}(0; 4)$ and consider

$$Y_n = \frac{X_1^2 + \dots X_n^2}{n}.$$

(a) Compute the density of Y_n

(b) Compute the limit of Y_n.

10. Let $T \sim$exp(0.2)

(a) Compute the density, the expectation and the variance of T^2.

(b) Let us assume to have a sequence of $\{T_k^2\}$ of independent r.v.'s all with the same density as of T^2. Compute approximatelly $P(\frac{S_{100}}{10} \leq 2)$, where $S_n = T_1^2 + \dots + T_n^2$.

11. Two r.v.'s independent $X, Y \sim \Gamma(2, 2)$.

(a) Say what is the density of $X + Y$.

(b) If $(X_1, Y_1), (X_2, Y_2) \dots$ are independent copies of X and Y and

$$Z_n = \sum_{k=1}^{n} (X_k + Y_k)$$

compute approximately $P(Z_{200} - 20 \geq 0)$.

12. Let $\{X_n\}_{\{n \in \mathbb{N}\}}$ be a sequence of i.i.d. exp(λ) r.v.'s and let $Y_n = \dfrac{X_n}{\log n}$, $n \geq 2$.

(a) Say what is the density of Y_n.

(b) Say if Y_n converges in distribution.

13. Let $X_1, \dots X_{500} \sim geom(\frac{1}{3})$ be independent r.v.'s. Estimate the probability that the sample mean $\bar{X}_{500} = \sum_{i=1}^{500} X_i$ may differ from the sample's expectation more than $\frac{1}{10}$.

14. The price X of a certain asset is a r.v. with expectation 10 and variance 2. The price is observed with a rounding error $Y \sim \mathcal{N}(0, 0.5)$ independent of X. The price is observed for 36 working days

Estimate the probability that the sample mean is less than 9.

15. Let $X \sim \mathcal{N}(0, 1)$ and $Y = \exp(X)$.

(a) Compute $E(Y)$ e $\mathrm{Var}(Y)$.

(b) Let $Y_1, \ldots, Y_{40}$ be independent copies of Y. Compute approximately $P(\bar{Y}_{40} \leq 1)$, where

$$\bar{Y}_{40} = \frac{Y_1 + \cdots + Y_{40}}{40}.$$

16. The lifetime (in hours) of a certain type of light bulbs is a r.v. $T \sim \Gamma(28000, 10)$.

(a) Compute expectation and variance of T.

(b) Compute approximately the probability a randomly selected lightbulb works after 2850 hours.

Indice

1 SEQUENCES OF R.V.'S, MOMENT GENERATING FUNCTION 1
1.1 What convergence? 1
1.2 The Law of Large Numbers 4
1.3 The Moment Generating Function 8
1.4 The Central Limit Theorem 11
1.5 Exercises 13

Lecture notes of Mathematical Statistics 2018/19 - part 3
Estimators, Hypothesis Testing
Linear Regression

Fabio Antonelli

2018

Capitolo 1

POINT ESTIMATORS

We already introduced estimators for the expectation and the variance of a random sample and we intuitively constructed estimators for the parameters of any of the presented distributions by employing the continuity property of the convergence in probability or almost surely.

But do we have general methods to construct point estimators?

Given different choices of estimators for the same parameter, what criterion could we apply to select one rather than another?

We start by partially answering the second question.

1.1 Best unbiased estimators

Surely a criterion to select an estimator of a parameter is if the estimator is unbiased or not.

Namely when T is an unbiased estimator of θ, we have that $\mathbb{E}(T) = \theta$ and consequently the mean square error that we commit by substituting θ with T is

$$\mathbb{E}\Big[(T-\theta)^2\Big] = \mathbb{E}\Big[(T-E(T))^2\Big] = Var(T).$$

But we might have two unbiased estimators for the same parameter, for instance, given a sample $X_1, \dots, X_n$ with mean μ and variance σ^2, assuming to know μ, both

$$\Sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2, \qquad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$$

are unbiased estimators of σ^2.

A very reasonable criterion to select an estimators among the unbiased ones is to choose the one that realizes the minimal variance.

Definition 1.1.1. *Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample of size n from a population $f(\mathbf{x}|\theta)$. An unbiased estimator $T(\mathbf{X})$ of $h(\theta)$ (for some function h) is called the* ***a best unbiased estimator*** *or* ***an estimator of uniform minimal variance (UMV)*** *if*

$$Var(T(\mathbf{X})) \leq Var(V(\mathbf{X}))$$

for any other unbiased estimator of $h(\theta)$.

Unfortunately, in general it is quite difficult to compute UMV estimators. Here we report a criterion

Theorem 1.1.1. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n from a population $f(\mathbf{x}|\theta)$. and let $T(\mathbf{X})$ be an unbiased estimator of $h(\theta)$.*

If for any r.v. $Y = V(\mathbf{X})$ such that $\mathbb{E}_\theta(V(\mathbf{X})) = 0$ for any value of θ, for some function $V : \mathbb{R}^n \longrightarrow \mathbb{R}$, we have

$$\mathbb{E}_\theta(T(\mathbf{X})V(\mathbf{X})) = cov_\theta(T(\mathbf{X}), V(\mathbf{X})) = 0 \tag{1.1}$$

then $T(\mathbf{X})$ is a UMV estimator for $h(\theta)$).

We do not give the proof of this result here, because it is beyond our purposes. In any case (1.1) is usually quite hard to be verified and it usually done by differentiating w.r.t. the parameter.

It is possible to give a lower bound of the variance by the Cramer Rao theorem.

Nevertheless there are many estimators that are not unbiased, but that are easy to construct and quite meaningful.

1.2 The method of moments

As we already pointed out, the more moments we are able to compute the more information we have about the distribution of a random variable, as the moment generating function shows.

As a matter of fact it is possible to construct quite easily estimators of the moments that might be used to estimate the parameters characterizing a density.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n from a population $f(\mathbf{x}|\theta)$, with $\theta = (\theta_1, \ldots, \theta_k)$ and let us assume that the r.v. X_1 has finite moments up to the order $m \geq k$. Then we may estimate each moment in the following way

$$\begin{aligned}
M_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \mu_1(\theta_1, \ldots, \theta_k) = \mathbb{E}(X_1) \\
M_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \qquad \mu_2(\theta_1, \ldots, \theta_k) = \mathbb{E}(X_1^2) \\
&\vdots \\
M_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad \mu_k(\theta_1, \ldots, \theta_k) = \mathbb{E}(X_1^k)
\end{aligned}$$

The idea is to match the estimators and and the expression of the moments as functions of the parameters and to try to invert the relations.

Example 1.2.1. *Let us assume to have a random sample $(X_1, \ldots, X_n)$ from a population $\Gamma(\alpha, \lambda)$, then we know that the first and second moments have the theoretical expression*

$$\mathbb{E}(X) = \frac{\alpha}{\lambda}, \qquad \mathbb{E}(X^2) = \frac{\alpha(\alpha+1)}{\lambda^2}$$

consequently we set

$$M_1 = \frac{\alpha}{\lambda}, \qquad M_2 = \frac{\alpha(\alpha+1)}{\lambda^2}$$
$$\Downarrow$$
$$\alpha = \frac{M_2}{M_1} - 1, \qquad M_2 = \frac{M_2 - 1}{M_1^2}$$

1.3 Maximum Likelihood estimators

We here introduce a second manner to construct estimators. The underlying idea is to find an estimator that is the parameter point that makes the observations the most likely.

Definition 1.3.1. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n from a population $f(\mathbf{x}|\theta)$, the function that to each value of the parameter θ associates the joint density of the sample*

$$\theta \longrightarrow f(x_1|\theta) \cdots f(x_n|\theta) = L(\theta|\mathbf{x}), \forall \mathbf{x} = (x_1, \ldots, x_n)$$

*is called the **likelihood function**. Similarly we may define the **log likelihood function** as*

$$\ln(L(\theta|\mathbf{x})) = \sum_{i=1}^{n} \ln f(x_i|\theta).$$

Therefore we have

Definition 1.3.2. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n from a population $f(\mathbf{x}|\theta)$, the **Maximum Likelihood Estimator (MLE)** of θ is the statistic defined as*

$$T(X_1, \ldots, X_n) = argmax_\theta L(\theta|X_1, \ldots, X_n)$$

that is to say the statistic that realizes the maximum of the Likelihood function.

As a function of a possibly multidimensional parameter, the candidate estimators are obtained as critical points of the likelihood function, for each fixed realized value of the sample $(x_1, \ldots, x_n)$, that is to say

$$\hat{\theta}(\mathbf{x}) \quad \text{such that } \frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0$$

Due to the monotonicity of the logarithm function, it is equivalent to find the critical points of the log likelihood function.

Example 1.3.1. *Given* $\mathbf{X} = (X_1, \ldots, X_n)$ *a random sample of size* n*, with* $X_i \sim \mathcal{N}(\mu, \sigma^2)$*, consequently the likelihood function is given by*

$$L(\theta, \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\}, \qquad \theta = (\mu, \sigma)$$

and therefore

$$\ln(L(\theta, \mathbf{x})) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Differentiating with respect ot μ *and* σ *and setting the derivatives equal to zero we obtain*

$$\begin{aligned} \frac{\partial}{\partial\mu}\ln(L(\theta|\mathbf{x})) &= \frac{1}{n\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \\ \frac{\partial}{\partial\sigma}\ln(L(\theta|\mathbf{x})) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0, \end{aligned}$$

whence

$$\begin{aligned} \hat{\mu}(X_1, \ldots, X_n) &= \overline{X}_n \\ \hat{\sigma}^2(X_1, \ldots, X_n) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2, \end{aligned}$$

that is to say that $\overline{X}_n$ *and* S_n^2 *are the maximum likelihood estimators.*

The MLE estimators have an important invariance property

Proposition 1.3.1. *If* $\hat{\theta}(X_1, \ldots, X_n)$ *is a MLE estimator for the parameter* θ*, then a MLE estimator for* $h(\theta)$ *for any function* h *is given by* $h(\hat{\theta}(X_1, \ldots, X_n))$*.*

Example 1.3.2. *This implies that if* S_n^2 *is the MLE for the variance* σ^2 *then* $\sqrt{S_n^2}$ *is a MLE for the standard deviation. This is a result that we partially already obtained by the continuity of the convergence in probability or a.s., the proposition implies that it verifies also the maximum likelihood property.*

1.4 Exercises

1. Find the maximum likelihood estimator for a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a Bin $(1, p)$.
2. Casella- Berger : # 7.1, 7.2(a), 7.6 (b) and (c), 7.9, 7.14

Capitolo 2

HYPOTHESES TESTING

We introduce another method of statistical inference that will lead to interval estimators for the parameters.

2.1 Definition and construction of a test

Definition 2.1.1. *A **hypothesis** is any statement about a population parameter. An **hypotheses testing** is a comparison between two hypotheses in order to understand which one can be accepted. The two complementary hypotheses are called **the null hypothesis**, usually denoted by H_0 and the **alternative hypothesis**, usually denoted by H_1*

In general the two hypotheses are represented by a partition of the space Θ where the parameter we want to estimate takes values:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_0^c.$$

For instance we could want to run a test to understand if, in average, a given random sample exceeds a prefixed level. In this case the test would be formulated as

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0.$$

After observing the sample the experimenter has to decide either to accept H_0 as true or reject H_0 as false and accept H_1.

Definition 2.1.2. *A **hypothesis test** is a rule that specifies*

1. *for which sample values the decision is made to accept H_0 as true.*
2. *For which sample values H_0 is rejected and H_1 is accepted as true.*

The set of the sample space for which $H_0 =$ will be rejected is called **the rejection region**, the complement is called **the acceptance region**.

TYpically a hypothesis test is specified by means of a **test statistic** $T(X_1, \ldots, X_n)$ and the decision to accept or reject the null hypothesis is determined by the fact that the estimating statistic falls or not in the rejection region.

Example 2.1.1. *Suppose we have a random sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ and for a given value μ_0 we want to run the following test*

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0.$$

Then once we decide what level of error is admissible, say $\epsilon > 0$, we may describe the rejection region of H_0 in terms of distance from μ_0 exceeding this prefixed ϵ. The random sample follows a $\mathcal{N}(\mu, 1)$, consequently the sample mean, estimating μ, follows a $\mathcal{N}(\mu, \frac{1}{n})$. Hence we may describe belonging to the rejection region R as

$$\{\overline{X}_n \in R\} = \{|\overline{X}_n - \mu_0| > \epsilon\} = \{|\overline{X}_n - \mu_0|\sqrt{n} > \epsilon\sqrt{n}\}.$$

We remark that $P(\{\overline{X}_n \in R\})$ is easy to evaluate, since $(\overline{X}_n - \mu_0)\sqrt{n} \sim \mathcal{N}(0, 1)$.

A test of common use is the Likelihood Ratio test

Definition 2.1.3. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n from a population $f(\mathbf{x}|\theta)$, with maximum likelihood function $L(\theta|\mathbf{x})$, to test*

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_0^c.$$

*we define the **maximum likelihood statistic***

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\mathbf{x})}{\sup_{\theta \in \Theta} L(\theta|\mathbf{x})}$$

*and a **maximum likelihood statistic test** is any test that has a rejection region R of the form $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$ for some $c \in [0, 1]$.*

The idea of this test is that the ratio captures the likelihood of the null hypotesis with respect to all possibilities. If the sample has a high probability to fall into the rejection region with a test with a "small" c, this means that the alternative hypothesis, given the sample, is much more likely to happen than the null one, thus H_0 should be rejected (its weight is bigger). If instead c is close to 1, the null hypothesis appears to be very likely.

Example 2.1.2. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n from a population $\mathcal{N}(\mu, 1)$ and we want to test*

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

hence $\sup_{\mu\in\Theta_0} L(\mu|\mathbf{x}) = L(\mu_0|\mathbf{x})$ *and we know that the MLE estimator for the mean of a Gaussian sample, for any value of the parameter is given by the sample mean* $\overline{X}_n$*, so the maximum likelihood statistic is*

$$\lambda(\mathbf{x}) = \frac{(2\pi)^{-\frac{n}{2}}\exp\{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2\}}{(2\pi)^{-\frac{n}{2}}\exp\{-\frac{1}{2}\sum_{i=1}^n (x_i-\overline{x}_n)^2\}} = \exp\Big\{-\frac{1}{2}\Big(\sum_{i=1}^n (x_i-\mu_0)^2 - \sum_{i=1}^n (x_i-\overline{x}_n)^2\Big)\Big\}$$

But the exponent can be simplified

$$\begin{aligned}
&\sum_{i=1}^n (x_i-\mu_0)^2 - \sum_{i=1}^n (x_i-\overline{x}_n)^2 = \sum_{i=1}^n [x_i^2 - 2\mu_0 x_i + \mu_0^2 - x_i^2 + 2\bar{x}_n x_i + \bar{x}_n^2] \\
= &\sum_{i=1}^n [-2(\mu_0-\bar{x}_n)x_i + \mu_0^2 - \bar{x}_n^2] = -2(\mu_0-\bar{x}_n)\sum_{i=1}^n x_i + n(\mu_0^2-\bar{x}_n^2) \\
= &-2(\mu_0-\bar{x}_n)n\bar{x}_n + n(\mu_0^2-\bar{x}_n^2) = -2n\mu_0\bar{x}_n + 2n\bar{x}_n^2 + n\mu_0^2 - n\bar{x}_n^2 = n(\bar{x}_n-\mu_0)^2
\end{aligned}$$

thus

$$\lambda(\mathbf{x}) = \mathrm{e}^{-\frac{1}{2}n(\bar{x}_n-\mu_0)^2}$$

and, for any given $c\in[0,1]$*, we may describe the rejection region as*

$$\begin{aligned}
\{\lambda(\mathbf{X})\le c\} &= \{\mathrm{e}^{-\frac{1}{2}n(\overline{X}_n-\mu_0)^2}\le c\} = \{(\overline{X}_n-\mu_0)^2 \ge -\frac{2}{n}\ln c\} \\
&= \{|\overline{X}_n-\mu_0| \ge \sqrt{-\frac{2}{n}\ln c}\}
\end{aligned}$$

As an exercise, construct a maximum likelihood statistic test for a random sample $\mathbf{X} = (X_1,\ldots,X_n)$ of size n from a population $\exp(\lambda)$.

2.1.1 Error probabilities and Power of a Test

Since tests are based on statistics, that are random variables, when accepting or rejecting a hypothesis, we take a decision "in probability", hence with the possibility of making mistakes. It is then important to understand what kind of mistake we risk to make and evaluate the error probabilities of making it.

Traditionally errors are classified into two types:

- If $\theta\in\Theta_0$ but the test incorrectly leads to a rejection of H_0, an error of **type I** has been made.
- If $\theta\in\Theta_0^c$ but the test incorrectly leads to an acceptance of H_0, an error of **type II** has been made.

If R is the rejection region of a hypothesis test using the random sample $\mathbf{X} = (X_1,\ldots,X_n)$ and the statistic $T(\mathbf{X})$, the test will make an error of type I if for $\theta\in\Theta_0$, it happens that $T(\mathbf{X})\in R$, consequently the probability to make an error of type I is given by

$$P_\theta(T(\mathbf{X})\in R) \qquad \theta\in\Theta_0$$

while the probability to make an error of type II will be given by

$$P_\theta(T(\mathbf{X}) \in R^c) \qquad \theta \in \Theta_0^c$$

In general making an error of type I is less severe than making an error of type II (one accepts a hypotheses that is not true). It naturally follows the

Definition 2.1.4. *We call **power function** of a test with test statistic $T(\mathbf{X})$ and rejection region R the function*

$$\beta(\theta) = P_\theta(T(\mathbf{X}) \in R).$$

*The **level of a test** is defined to be*

$$\sup_{\theta \in \Theta_0} P_\theta(T(\mathbf{X}) \in R).$$

If this number is equal to $\alpha \in (0,1)$ we say that the test is a level α test.

In general a good test has power function close to 0 when $\theta \in \Theta_0$ and close to 1 when $\theta \in \Theta_0^c$.

Examples 2.1.1.

1. *(Binomial power function) In this example we understand that the power function depends on the type of test we are running.*

   Assume that we have a binomial statistic $T(\mathbf{X}) \sim Bin(5,p)$ and we consider the test

   $$H_0 : p \leq \frac{1}{2}$$
   $$H_1 : p > \frac{1}{2}.$$

   If we consider as rejection region the set when all successes occur, the power function is then given by

   $$\beta(p) = P(T(\mathbf{X}) \in R) = P(T(\mathbf{X}) = 5) = p^5.$$

   Under H_0 the level of the test is

   $$\sup_{p \leq \frac{1}{2}} P(T(\mathbf{X}) = 5) = \sup_{p \leq \frac{1}{2}} p^5 = (\frac{1}{2})^5 = 0.0312$$

   so it seems to be good enough as the probability of making an error of type I is reasonably and uniformly small. On the other hand, the probability of an error of type II is instead not so small, indeed

   $$\beta(p) = 1 - P(T(\mathbf{X}) = 5) = 1 - p^5, \qquad p > \frac{1}{2}$$

   is always too large.

But we may correct our test by saying that our rejection region is described by 3,4 or 5 successes and in this case the power function becomes

$$\beta(p) = P(T(\mathbf{X}) \in R) = P(T(\mathbf{X}) = 3, 4, 5) = p^5 + 5p^4(1-p) + 10p^2(1-p)^2$$

and the probability of an error of type II has decreased, but the probability of an error of type I has increased.

Something in between?

2. *(Normal power function) Assume to have a Normal random sample* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ *and we want to run the test*

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0.$$

As before, once we decided what level of error is admissible, say $\epsilon > 0$*, then the rejection region is of the type* $[\epsilon, +\infty)$*, hence* $\{\overline{X}_n \in R\} = \{|\overline{X}_n - \mu_0| > \epsilon\}$ *and we may evaluate the power function as*

$$\begin{aligned}
\beta(\mu) &= P_\mu(|\overline{X}_n - \mu_0| > \epsilon) = 1 - P_\mu(|\overline{X}_n - \mu_0| \leq \epsilon) \\
&= 1 - P_\mu(\mu_0 - \mu - \epsilon \leq \overline{X}_n - \mu \leq \mu_0 - \mu + \epsilon) \\
&= 1 - P_\mu((\mu_0 - \mu - \epsilon)\sqrt{n} \leq (\overline{X}_n - \mu)\sqrt{n} \leq (\mu_0 - \mu + \epsilon\sqrt{n}) \\
&= 1 - P([\mu_0 - \mu + \epsilon]\sqrt{n} \leq \mathcal{N}(0,1) \leq [\mu_0 - \mu - \epsilon]\sqrt{n}) \\
&= 1 - \Phi([\mu_0 - \mu + \epsilon]\sqrt{n}) + \Phi([\mu_0 - \mu - \epsilon]\sqrt{n})
\end{aligned}$$

and if we want a test at level α *we just have to solve the equation* ($\mu = \mu_0$)

$$\alpha = 2(1 - \Phi(\epsilon\sqrt{n}))$$

For instance assume that the sample size is 100 and $\alpha = 0,05$*, then we have that the test allows for an error given by*

$$\Phi(10\epsilon) = 0.975 \quad \Rightarrow 10\epsilon = 1.645 \Rightarrow \epsilon = 0.1645.$$

If we wanted to allow for a smaller error we should use a larger sample.

2.2 Sampling and Testing from Gaussian samples

In this section we will assume that the random sample of size n is always drawn from a Gaussian population $\mathcal{N}(\mu, \sigma^2)$.

Given $\mathbf{X} = (X_1, \ldots, X_n)$ we already know that the sample mean $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ and consequently

$$n\frac{\Sigma_n^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \Gamma(\frac{n}{2}, \frac{1}{2}) = \chi^2(n)$$

whenever we have a knowledge of the two parameters.

We would like to know if we can deduce distributional properties also for the other estimators of the variance and of the standardized estimators even when the parameters are not known.

The above remarks are summarized in the following theorem together with a quite surprising result

Theorem 2.2.1. *(CB - Theorem 5.3.1) Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *a random sample from a population* $\mathcal{N}(\mu, \sigma^2)$ *and let*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

then

1. $\overline{X}_n \sim \mathcal{N}(\mu, \dfrac{\sigma^2}{n})$.
2. $(n-1)\dfrac{s_n^2}{\sigma^2} \sim \chi^2(n-1)$.
3. $\overline{X}_n$ *and* s_n^2 *are independent random variables.*

We already proved the first result. The other two follow from the Cochran theorem that we are going to present in the next subsection

2.2.1 The Cochran theorem

For $m \in \mathbb{N}$ on the vector space $\mathbb{R}^m$, it is usually defined the Euclidean scalar product given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{m} x_i y_i \qquad \mathbf{x} = (x_1, \ldots, x_m),\ \mathbf{y} = (y_1, \ldots, y_m)$$

and we say that two vectors are orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Let us now take an $m \times m$ square matrix A and let us denote by A^* its transpose.

We have that

$$\begin{aligned} \langle A \cdot \mathbf{x}, \mathbf{y} \rangle &= \sum_{i=1}^{m} (A \cdot \mathbf{x})_i y_i = \sum_{i=1}^{m}\sum_{k=1}^{m} a_{ik} x_k y_i \\ &= \sum_{i=1}^{m}\sum_{k=1}^{m} a^*_{ki} x_k y_i = \sum_{k=1}^{m} x_k \sum_{i=1}^{m} a^*_{ki} y_i = \sum_{k=1}^{m} x_k (A^* \cdot \mathbf{y})_k = \langle \mathbf{x}, A^* \cdot \mathbf{y} \rangle \end{aligned}$$

We call **orthogonal** matrix, any matrix O such that $O^* = O^{-1}$. The matrix O represents an orthogonal change of basis

By the previous remark we have that

$$\langle O \cdot \mathbf{x}, O \cdot \mathbf{y} \rangle = \langle \mathbf{x}, O^* O \cdot \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle$$

hence $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if and only if so are $O \cdot \mathbf{x}$ and $O \cdot \mathbf{y}$.

Taken any two linear subspaces E and F of $\mathbb{R}^m$, we say that they are orthogonal if any vector in E is orthogonal to any other vector in F and we define the orthogonal projection of a vector $\mathbf{x} \in \mathbb{R}^m$ on a linear subspace E, the mapping that associates to $\mathbf{x}$ the vector $\mathbf{y} \in E$ that has minimal distance from $\mathbf{x}$ and that we are going to denote by $\mathbf{y} = P_E\mathbf{x}$

Hence any vector $\mathbf{x}$ can be written as

$$\mathbf{x} = P_E\mathbf{x} + (\mathbf{x} - P_E\mathbf{x}) = P_E\mathbf{x} + (I - P_E)\mathbf{x} =: \mathbf{x}_1 + \mathbf{x}_2$$

where the two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal. We denote by $E^\perp$ the set of vectors orthogonal to any vector in E. This is a linear subspace as well.

Example 2.2.1. *Let E be the linear subspace of $\mathbb{R}^m$ of the vectors that have all equal components. This is a one dimensional subspace since a basis is given by the vector $\mathbf{1} = (1, \ldots, 1)$. In order to find the orthogonal projection on E we have to find the scalar t that minimizes the square distance*

$$\psi(t) = |\mathbf{x} - t\mathbf{1}|^2 = \sum_{i=1}^{m} (x_i - t)^2.$$

By taking the derivative and setting it equal to zero we obtain

$$\psi'(t) = -2\sum_{i=1}^{m} (x_i - t) = 0 \quad \Leftrightarrow \quad t = \frac{1}{m}\sum_{i=1}^{m} x_i =: \bar{x},$$

therefore $P_E\mathbf{x} = (\bar{x}, \ldots, \bar{x})$.

Theorem 2.2.2. *Let $\mathbf{X} = (X_1, \ldots, X_m)$ a random vector of i.i.d. $\mathcal{N}(0,1)$ and let $E_1, \ldots, E_k$ be orthogonal linear subspaces of R^m. For $i = 1, \ldots, k$, we denote by $n_i = dim(E_i)$ and by P_{E_i} the othrogonal projection mapping on the ith subspace. Then $Y_i = P_{E_i}\mathbf{X}$ are independent r.v.'s and their norms $||Y_i||^2 \sim \chi^2(n_i)$.*

Dimostrazione. We prove the case $k = 2$, since it is possible to iterate the proof to more subspaces. Hence we have two subspaces E_1, E_2, with $\dim(E_1) = n_1$, $\dim(E_2) = n_2$ and $n_1 + n_2 \leq m$. Using a basis for E_1, a basis for E_2 and a basis for the complementary linear subspace, we may imagine that the first n_1 coordinates in any vector are relative to the space E_1, the second n_2 to the space E_2 and the rest to the orthogonal space to both completing R^m. (we are employing an orthogonal change of basis that will leave the independence unaltered)

Therefore, applying the projections to the sample, we have that

$$\begin{aligned} Y_1 = P_{E_1}\mathbf{X} &= (X_1, \ldots, X_{n_1}, 0, \ldots, 0) \\ Y_2 = P_{E_2}\mathbf{X} &= (0, \ldots, 0, X_{n_1+1}, \ldots, X_{n_1+n_2}, 0, \ldots, 0) \end{aligned}$$

are independent vectors, as functions of independent random variables, with independent components.

Taking the norms we have

$$\begin{aligned} ||Y_1||^2 &= |X_1|^2 + \dots |X_{n_1}|^2 \sim \chi^2(n_1) \\ ||Y_2||^2 &= |X_{n_1+1}|^2 + \dots + |X_{n_1+n_2}|^2 \sim \chi^2(n_2), \end{aligned}$$

recalling that the sum of the squares of n independent $\mathcal{N}(0,1)$ r.v.'s is distributed as a $\Gamma(\frac{n}{2}, \frac{1}{2}) = \chi^2(n)$. □

The above theorem implies that, starting from a Gaussian random sample $\mathbf{X} = (X_1, \dots, X_m) \sim \mathcal{N}(0,1)$, if we take as $E_1 = \{t\mathbf{1} = t(1, \dots, 1),\ t \in \mathbb{R}\}$ and $E_2 = E_1^{\perp}$, we have

$$\begin{aligned} P_{E_1}\mathbf{X} &= (\overline{X}_m, \dots, \overline{X}_m) \\ P_{E_2}\mathbf{X} = (I - P_{E_1})\mathbf{X} &= (X_1 - \overline{X}_m, \dots, X_m - \overline{X}_m) \end{aligned}$$

define two independent vectors and the components of the second are all independent of each other. We remark that the second subspace has $\dim(E_2) = m - 1$.

Consequently

$$||P_{E_2}\mathbf{X}||^2 = \sum_{i=1}^{m} |X_i - \overline{X}_m|^2 = (m-1)s_m^2$$

is distributed as a $\chi^2(m-1)$ and that it is independent of $\overline{X}_m$, so concluding Theorem 2.2.1.

Definition 2.2.1. *Let $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2(n)$ be two independent random variables then $T = \sqrt{n}\dfrac{Z}{\sqrt{W}}$ has a Student's t-distribution with n degrees of freedom and it has density given by*

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\Big(1 + \frac{t^2}{n}\Big)^{-\frac{n+1}{2}}, \qquad t \in \mathbb{R}$$

and we write $T \sim t(n)$.

We remark that also this density is symmetric, but with a polynomial decay rather than an exponential one as for the Gaussian densities.

Consequences of Theorem 2.2.1 and of the previous definition are

1. for $i = 1, \dots, n$, $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. $\Rightarrow$ $Y_i = \dfrac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$ i.i.d.

2. $\overline{X}_n \sim \mathcal{N}(\mu, \dfrac{\sigma^2}{n}) \Rightarrow \dfrac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$ and

$$\frac{\overline{X}_n - \mu}{\sigma} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}_n \Rightarrow \overline{Y}_n\sqrt{n} \sim \mathcal{N}(0,1).$$

3.

$$\begin{aligned}\frac{s_n^2}{\sigma^2}(n-1) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2=\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma}-\frac{\overline{X}_n-\mu}{\sigma})^2\\ &= \sum_{i=1}^{n}(Y_i-\overline{Y}_n)^2=:(s_n^Y)^2(n-1)\sim\chi^2(n-1).\end{aligned}$$

4. $\overline{X}_n$ and s_n^2 are independent random variables as functions of independent r.v.'s.

5. $\frac{\overline{Y}_n\sqrt{n}}{s_n^Y\sqrt{n-1}}\sqrt{n-1}=\frac{\frac{\overline{X}_n-\mu}{\sigma}\sqrt{n}}{\frac{s_n\sqrt{n-1}}{\sigma}}\sqrt{n-1}=\frac{\overline{X}_n-\mu}{s_n}\sqrt{n}\sim t(n-1)$

Hence if we want ot run a test for the mean for a Gaussian sample $\mathcal{N}(\mu,\sigma^2)$ with unknown μ,σ,

$$\begin{aligned}H_0 &: \mu\le\mu_0\\ H_1 &: \mu>\mu_0\end{aligned}$$

we may choose the rejection region to be described by $[c,+\infty)$ for some $c>0$ and we may exploit the statistic $\frac{\overline{X}_n-\mu}{s_n}\sqrt{n}$ to evaluate the probabilities.

Namely for any $\mu\le\mu_0$ we would reject the hypothesis when $\overline{X}_n-\mu_0>c>0$. Since $P(\sqrt{n}s_n>0)=1$, to evaluate $P_\mu(\overline{X}_n-\mu_0>c)$ is the same as evaluating $P_\mu(\frac{\overline{X}_n-\mu_0}{s_n}\sqrt{n}>c)$ and $\frac{\overline{X}_n-\mu}{s_n}\sqrt{n}\sim t(n-1)$, so

$$P(\frac{\overline{X}_n-\mu_0}{s_n}\sqrt{n}>c)=P_\mu(\frac{\overline{X}_n-\mu}{s_n}\sqrt{n}+\frac{\mu-\mu_0}{s_n}\sqrt{n}>c)\le P_\mu(\frac{\overline{X}_n-\mu}{s_n}\sqrt{n}>c)\qquad\mu\le\mu_0$$

and if we decide that we want to have a test of level α it is sufficient to choose c so that $P(t(n-1)>c)\le\alpha$.

Definition 2.2.2. *Given a r.v. X with distribution function F_X, we call quantile or order α, with $0\le\alpha\le1$, the largest number x_α so that*

$$F_X(x_\alpha)=P(X\le x_\alpha)\le\alpha$$

Traditionally ϕ_α denotes the quantile of order α of a standard Normal, $\chi^2_\alpha(n)$ the quantile of a chi-square and $t_\alpha(n)$ the quantile of a Student's t-distribution with n degrees of freedom.

Therefore in the previous test we might choose $c=t_{1-\alpha}(n-1)$, since

$$P(t(n-1)>t_{1-\alpha}(n-1))=1-(1-\alpha)=\alpha.$$

2.3 Confidence intervals

We are now ready to discuss another way to estimate the parameters characterizing a random sample, that is by constructing random intervals that contain the parameters with high probability. This type of intervals are called confidence intervals and we have already implicitly used them to evaluate the error probabilities of a test.

Definition 2.3.1. *Given a random sample* $\mathbf{X} = (X_1, \dots, X_n)$*, an* ***interval estimator*** *or* ***confidence interval*** *of a parameter* θ *is any pair of statistics* $L(\mathbf{X}), U(\mathbf{X})$*, so that* $L(\mathbf{x}) \leq U(\mathbf{x})$ *for all* $\mathbf{x}$ *and* $L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})$ *with positive probability.*

Example 2.3.1. *Given a random sample* $X_1, \dots, X_8 \sim \mathcal{N}(\mu, 1)$ *then we have a positive probability that*

$$P(\overline{X}_8 - 1 \leq \mu \leq \overline{X}_8 + 1) = P(|\overline{X}_8 - \mu| \leq 1) = P(|\mathcal{N}(0,1)| \leq \sqrt{8}) = \Phi(2.82) - \Phi(-2.82) = 0.9962$$

We say we have a confidence interval of level α of the parameter θ if

$$P_\theta(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) \geq 1 - \alpha, \quad \forall\, \theta \in \Theta.$$

In the previous example we have a confidence interval at level 0.01.

As already outlined by the previous examples, a very common method to find confidence intervals is by inverting a test statistic, which usually boils down to finding the proper quantile for a theoretically known density.

Example 2.3.2.

1. *Let* $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ *and* $H_0 : \mu = \mu_0$ *versus* $H_1 : \mu \neq \mu_0$.
   If we want a test at level α *we want to find* $\epsilon > 0$ *so that*
   $$P_{\mu_0}(|\overline{X}_n - \mu_0| > \epsilon) = \alpha$$
   or equivalently
   $$P_{\mu_0}(|\overline{X}_n - \mu_0| \leq \epsilon) = 1 - \alpha.$$
   Since under H_0, $\overline{X}_n \sim \mathcal{N}(\mu_0, \dfrac{\sigma^2}{n})$*, we have that*
   $$P_{\mu_0}(|\overline{X}_n - \mu_0| \leq \epsilon) = P_{\mu_0}(|\overline{X}_n - \mu_0|\frac{\sqrt{n}}{\sigma} \leq \epsilon\frac{\sqrt{n}}{\sigma}) = P(\mathcal{N}(0,1) \leq \epsilon\frac{\sqrt{n}}{\sigma}))$$
   and it is therefore enough to set $\epsilon\frac{\sqrt{n}}{\sigma} = \Phi_{\alpha/2} \Rightarrow \epsilon = \frac{\sigma}{\sqrt{n}}\Phi_{\alpha/2}$ *to conclude that*
   $$[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi_{\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}\Phi_{\alpha/2}]$$
   is a confidence interval for μ *at level* $1 - \alpha$.

2. *Of course the previous procedure holds as long as σ is known, If it is not so,alternatively we may consider that* $\frac{\overline{X}_n - \mu}{s_n}\sqrt{n} \sim t(n-1)$ *and applying the same argument as before, exploiting a t− Student distribution (which is symmetric too) rather than the Normal one, thus a confidence interval at level* $1-\alpha$*will be given by*

$$[\overline{X}_n - \frac{s_n}{\sqrt{n}}t_{\alpha/2}(n-1), \overline{X}_n + \frac{s_n}{\sqrt{n}}t_{\alpha/2}(n-1)]$$

The following theorem says that the procedure we applied in the examples is indeed general. Given an acceptance region for a test, we can always construct a corresponding confidence interval. Given a confidence interval at level $1-\alpha$ for the test, we can always write the acceptance region in terms of the confidence interval, more precisely

Theorem 2.3.1. *For each value $\theta_0 \in \Theta$, if $\mathcal{A}(\theta_0)$ is the acceptance region of a level $\alpha \in (0,1)$ test of the type*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0 \tag{2.1}$$

then for the sample $\mathbf{X}$, the set

$$C(\mathbf{X}) = \{\theta_0 : \mathbf{X} \in \mathcal{A}(\theta_0)\}$$

is a $1-\alpha$ confidence interval.

Conversely if $C(\mathbf{X})$ is a $1-\alpha$ confidence interval, then the region defined by

$$\mathcal{A}(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}$$

is an acceptance region of a level α test of the type (2.1).

We now introduce a quantity that tells us right away how likely the null hypothesis is.

Definition 2.3.2. *A **p-value** is a test statistic $p(\mathbf{X})$ such that $0 \leq p(\mathbf{x}) \leq 1$ for all $\mathbf{x}$ in the sample space and we say that it is a **valid** p-value if, for every $\theta \in \Theta_0$ and any $0 \leq \alpha \leq 1$, we have*

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha. \tag{2.2}$$

The p-value is the smallest significance level at which the data lead to rejection of the null hypothesis. It gives the probability that data as unsupportive of H_0 as those observed will occur when H_0 is true. A small p-value (say, 0.05 or less) is a strong indicator that the null hypothesis is not true. The smaller the p-value, the greater the evidence for the falsity of H_0.

Indeed if $p(\mathbf{x})$ is small for most values $\mathbf{x}$ in the sample space, then (2.2) will stay true also for small α and the contrary will have high probability to happen

Example 2.3.3. *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *a population from a* $\mathcal{N}(\mu, \sigma^2)$*. We consider the two sided test* $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$.

Then the test rejects H_0 *if the statistic* $\frac{|\overline{X}_n - \mu_0|}{s_n}\sqrt{n}$ *, that follows a* $t(n-1)$ *takes large values, for any value of* σ*, so*

$$P_{\mu_0}(\frac{|\overline{X}_n - \mu_0|}{s_n}\sqrt{n} \geq t_{\alpha/2}(n-1)) = 2P_{\mu_0}(\frac{\overline{X}_n - \mu_0}{s_n}\sqrt{n} \geq t_{\alpha/2}(n-1)) \leq \alpha$$

then if $\bar{x}$ *and* s *are the realized value of* $\overline{X}_n$ *and* s_n*, then the* p *value will be given by*

$$p - value = 2P(T \geq \frac{\bar{x} - \mu_0}{s}),$$

where $T \sim t(n-1)$.

2.4 Exercises

1. Casella- Berger : # 8.1, 8.5 (a) and (b)

2. Recall that for X and Y two r.v.'s with joint density $f_{X,Y}(x, y)$ and $\phi : \mathrm{R}^2 \longrightarrow \mathrm{R}^2$ an invertible and differentiable vector function, we have that the random vector $(Z, W) = \phi(X, Y) = (\phi_1(X, Y), \phi_2(X, Y))$ has density

$$f_{Z,W}(z, w) = \frac{1}{|J_\phi(\phi^{-1}(z, w))|} f_{X,Y}(\phi^{-1}(z, w)) \tag{2.3}$$

   where J denotes the Jacobian matrix

$$J_\phi = \begin{pmatrix} \frac{\partial \phi_1}{\partial x} & \frac{\partial \phi_1}{\partial y} \\ \frac{\partial \phi_2}{\partial x} & \frac{\partial \phi_2}{\partial y} \end{pmatrix}.$$

   (a) Suppose that $X \sim \Gamma(\alpha, \lambda)$ and $Y \sim \Gamma(\delta, \lambda)$, with $\alpha, \delta, \lambda > 0$, are independent r.v.'s. Applying (2.3) find the joint density of

$$U = X, \qquad V = \frac{X}{X + Y}.$$

   (b) Where do x, y have to vary to have a positive joint density?

   (c) In the expression of the joint density $f_{U,V}(u, v)$ multiply by the appropriate power of $\frac{1}{v}$ to obtain a conditional density of $U|V = v$ to be a $\Gamma(\alpha + \delta, \frac{\lambda}{v})$.

   (d) By integrating with respect to u, find the marginal density $f_V(v)$, by multiplying and dividing by the appropriate constant to perform the integral.

The resulting density will be

$$f_V(v) = \frac{\Gamma(\alpha+\delta)}{\Gamma(\alpha)\Gamma(\delta)} v^{\alpha-1}(1-v)^{\delta-1}\mathbf{1}_{\{0<v<1\}} \tag{2.4}$$

and it is called density Beta of parameters α, δ and we write $U \sim \beta(\alpha, \delta)$.

3. Casella- Berger : # 8.6, 8.8, 8.12, 8.25, 8:37 (a) and (c),8.38.
4. Casella- Berger : # 9.2, 9.4, 9.13 (a).

2.5 Tests concerning more populations - ANOVA

Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are two Gaussian samples respectively from a $\mathcal{N}(\mu_x, \sigma_x^2)$ population of size n and from $\mathcal{N}(\mu_y, \sigma_y^2)$ population of size m. Assuming that the population variances σ_x^2 and σ_y^2 are known, let us consider a test of the null hypothesis that the two population means are equal; that is, let us consider a test

$$H_0 : \mu_x = \mu_y, \qquad H_1 : \mu_x \neq \mu_y$$

Since the estimators of μ_x and μ_y are the respective sample means $\overline{X}_n$ and $\overline{Y}_m$, it is reasonable to have a rejection region that is expressed in terms of the distance between those to be greater that an assigned value. Since $\overline{X}_n \sim \mathcal{N}(\mu_x, \frac{\sigma_x^2}{n})$ and $\overline{X}_m \sim \mathcal{N}(\mu_y, \frac{\sigma_y^2}{m})$ are two independent Gaussian r.v.'s, we have that

$$\overline{X}_n - \overline{Y}_m \sim \mathcal{N}(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m})$$

then if we want a level α test, we will have to reject H_0 when

$$\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \geq \phi_{\alpha/2}$$

Consequently if $v > 0$ is the realized value of the statistic

$$v = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

then we have

$$p - value = P(\mathcal{N}(0,1) \geq v).$$

If the variances are instead not known, but the sample size is the same for both $\mathbf{X}$ and $\mathbf{Y}$, then we may treat the two random samples as one. Indeed

$$Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n \sim \mathcal{N}(\mu_x - \mu_y, (\sigma_x + \sigma_y))$$

are still an indenpedent set of r.v.'s and we are in the same conditions as examples 2.3.2 and we may use the statistic $\frac{\overline{Z}_n}{s_n^z}\sqrt{n} \sim t(n-1)$.

Remarks 2.5.1.

1. *In the case of a unilateral test $H_0 : \mu_x \leq \mu_y$, $H_1 : \mu_x > \mu_y$, the same considerations hold and we will have to use ϕ_α rather than $\phi_{\alpha/2}$ and t_α rather than $t_{\alpha/2}$.*

2. *If the two samples do not have the same size, we may use the previous tests in an approximate manner. Namely, by the Law of Large Numbers, the estimators of the two variances, $(s_n^x)^2$ and $(s_m^y)^2$ converge respectively to σ_x^2 and σ_y^2 with probability 1. So for n and m sufficiently large we may think of substituting the variances with the estimators in the two tests above, taking the statistic*

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_x - \mu_y)}{\sqrt{\frac{(s_n^x)^2}{n} + \frac{(s_m^y)^2}{m}}}$$

still distributed as a $\mathcal{N}(0,1)$.

3. *If $\sigma_x = \sigma_y = \sigma$ is unknown, the previous test becomes exact. Indeed, both $(s_n^x)^2$ and $(s_m^y)^2$ are unbiased estimators for σ^2 and we also have that*

$$(n-1)\frac{(s_n^x)^2}{\sigma^2} + (m-1)\frac{(s_m^y)^2)}{\sigma^2} \sim \chi^2(n-1+m-1)$$

as sum of independent r.v.'s with gamma densities with the same second parameter. This statistic is also independent of $\overline{X}_n$ and $\overline{Y}_m$, being the samples independent. Consequently

$$\begin{aligned} & \frac{\frac{\overline{X}_n - \overline{Y}_m - (\mu_x-\mu_y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{(n-1)\frac{(s_n^x)^2}{\sigma^2}+(m-1)\frac{(s_m^y)^2}{\sigma^2}}}\sqrt{m+n-2} = \frac{\frac{\overline{X}_n - \overline{Y}_m - (\mu_x-\mu_y)}{\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{n-1}{m+n-2}(s_n^x)^2+\frac{m-1}{m+n-2}(s_m^y)^2}} \\ = \quad & \frac{\overline{X}_n - \overline{Y}_m - (\mu_x - \mu_y)}{\sqrt{s_p^2(\frac{1}{n}+\frac{1}{m})}} \sim t(m+n-2), \end{aligned}$$

which implies that the statistic

$$s_p^2 := \frac{n-1}{m+n-2}(s_n^x)^2 + \frac{m-1}{m+n-2}(s_m^y)^2$$

is an unbiased estimator of σ^2 and we may construct a test of level α by exploiting the Student's t-distribution as before.

2.5.1 The classical ANOVA Test

Generalizing what we did before, we obtain the ANOVA: Analysis Of Variance test.

The aim of this kind of test is to compare the means of several samples at the same time, samples that are usually assumed to be Normal. The ANOVA test has a very strict null hypothesis, hence the researcher's interest does not lie so much in accepting the null

hypothesis or not, rather than in understanding the deviations among means that lead to a rejection.

As a consequence the ANOVA test may be reformulated in a different ways, exploiting different estimators of cumulative variance of the samples, in order to point out different aspects.

In the classical Anova model we have k populations each of size n_i

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, \dots, k, \quad j = 1, \dots, n_i, \quad n := n_1 + \cdots + n_k$$

with unknown parameters μ_i and

1. the ϵ_{ij} are independent and Normally distributed (normal errors);
2. $\mathbb{E}(\epsilon_{ij}) = 0$ and $\mathrm{Var}(\epsilon_{ij}) = \sigma_i^2$;
3. $\sigma_i^2 = \sigma^2$, for all $i = 1, \dots, k$.

We will be interested in testing the classical ANOVA null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \quad versus \quad H_1 : \mu_i \neq \mu_j, \quad for\, some\, i \neq j.$$

Let us denote, for $i = 1, \dots, k$,

$$\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij},$$

then $Y_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n_i})$ and any linear combination of the sample means

$$\sum_{i=1}^{k} a_i \overline{Y}_i, \qquad a_1, \dots, a_n \in \mathbb{R}$$

is also Normal $\mathcal{N}\Big(\sum_{i=1}^{k} a_i \mu_i; \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i} \Big)$ and hence

$$\frac{\sum_{i=1}^{k} a_i \overline{Y}_i - \sum_{i=1}^{k} a_i \mu_i}{\sqrt{\sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}} \sim \mathcal{N}(0, 1).$$

As before, each

$$s_i^2 := \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$$

is an estimator of σ^2, independent of $\overline{Y}_i$ (for $i = 1, \dots, k$), and

$$(n_i - 1) \frac{s_i^2}{\sigma^2} \sim \chi^2(n_i - 1).$$

Consequently the pooled estimator

$$s_p^2 := \frac{1}{n_1 - 1 + \dots n_k - 1} \sum_{i=1}^{k} (n_i - 1) s_i^2 = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) s_i^2$$

is also an unbiased estimator of σ^2 and, since the s_i^2 are independent and independent of the $\overline{Y}_i$'s, it follows that

$$(n-k)\frac{s_p^2}{\sigma^2} \sim \chi^2(n-k), \qquad \frac{\sum_{i=1}^{k} a_i \overline{Y}_i - \sum_{i=1}^{k} a_i \mu_i}{\sqrt{s_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}} \sim t(n-k).$$

Remark 2.5.1. *We may give an equivalent expression of the test, namely the null hypothesis*

$$\mu_1 = \dots = \mu_k =: \mu$$

is equivalent to saying

$$\sum_{i=1}^{k} a_i \mu_i = \mu \sum_{i=1}^{k} a_i = 0,$$

for all choices of $a_1, \dots, a_k$ *such that* $\sum_{i=1}^{k} a_i = 0$.

Thus, we may reformulate the test as

$$H_0 : \sum_{i=1}^{k} a_i \mu_i = 0, \quad \text{for all } a_1, \dots, a_k \in \mathbb{R} \text{ such that } \sum_{i=1}^{k} a_i = 0$$

$$H_1 : \sum_{i=1}^{k} a_i \mu_i \neq 0 \quad \text{for some } a_1, \dots, a_k \in \mathbb{R} \text{ such that } \sum_{i=1}^{k} a_i = 0$$

and, to test at level α*, we would reject* H_0 *if*

$$\left| \frac{\sum_{i=1}^{k} a_i \overline{Y}_i}{\sqrt{s_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}} \right| \geq t_{\alpha/2}(n-k).$$

2.5.2 The ANOVA F-test

The ANOVA test may be rewritten by comparing the values of two estimators of the common variance.

With the notation introduced before, let us take $a_i = 1$ for all $i = 1, \dots, k$, then we know that the pooled estimator

$$s_p^2 = \sum_{i=1}^{k} s_i^2 \frac{n_i - 1}{n-k} = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=i}^{n_i} (Y_{ij} - \overline{Y}_i)^2$$

estimates σ^2 by averaging the single samples variances against the weight of the sample. It is a measure of the variances within groups.

Alternatively we may think of constructing a variance estimator in the following way: the r.v.'s $\overline{Y}_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n_i})$ are independent since they are formed on the basis of independent populations, consequently

$$\overline{\overline{Y}} := \sum_{i=1}^{k} \frac{n_i}{n} \overline{Y}_i \sim \mathcal{N}\Big(\sum_{i=1}^{k} \frac{n_i}{n} \mu_i, \frac{\sigma^2}{n} \Big)$$

and setting $\bar{\mu} := \sum_{i=1}^{k} \frac{n_i}{n} \mu_i$, we have that for $i = 1, \ldots, k$,

$$\sqrt{n_i} \frac{\overline{Y}_i - \mu_i}{\sigma} \sim \mathcal{N}(0,1), \qquad \sqrt{n} \frac{\overline{\overline{Y}} - \bar{\mu}}{\sigma} \sim \mathcal{N}(0,1)$$

$$(k-1)\frac{s^2}{\sigma^2} := \sum_{i=1}^{k} n_i \frac{(\overline{Y}_i - \mu_i)^2}{\sigma^2} - n \frac{(\overline{\overline{Y}} - \bar{\mu})^2}{\sigma^2} = \sum_{i=1}^{k} n_i \Big[\frac{(\overline{Y}_i - \mu_i)^2}{\sigma^2} - \frac{(\overline{\overline{Y}} - \bar{\mu})^2}{\sigma^2} \Big]$$

$$= \sum_{i=1}^{k} \frac{n_i}{\sigma^2} \Big[[(\overline{Y}_i - \mu_i) - (\overline{\overline{Y}} - \bar{\mu})][(\overline{Y}_i - \mu_i) + (\overline{\overline{Y}} - \bar{\mu})] \Big]$$

$$= \sum_{i=1}^{k} \frac{n_i}{\sigma^2} \Big[[(\overline{Y}_i - \mu_i) - (\overline{\overline{Y}} - \bar{\mu})]^2 + 2[(\overline{Y}_i - \mu_i) - (\overline{\overline{Y}} - \bar{\mu})](\overline{\overline{Y}} - \bar{\mu}) \Big]$$

$$= \sum_{i=1}^{k} \frac{n_i}{\sigma^2} [(\overline{Y}_i - \overline{\overline{Y}} - (\mu_i - \bar{\mu})]^2 + 2 \sum_{i=1}^{k} \frac{n_i}{\sigma^2} [(\overline{Y}_i - \mu_i) - (\overline{\overline{Y}} - \bar{\mu})](\overline{\overline{Y}} - \bar{\mu})$$

$$= \sum_{i=1}^{k} \frac{n_i}{\sigma^2} [(\overline{Y}_i - \overline{\overline{Y}} - (\mu_i - \bar{\mu})]^2 \sim \chi^2(k-1)$$

are independent r.v.'s, consequently

$$\sqrt{n} \frac{\overline{\overline{Y}} - \bar{\mu}}{s} \sim t(k-1)$$

is independent of $(n-k)\frac{s_p^2}{\sigma^2} \sim \chi^2(n-k)$, by Cochran's theorem.

Moreover

$$\mathbb{E}(s^2) = \sigma^2 \frac{1}{k-1} \Big[\sum_{i=1}^{k} Var(\sqrt{n_i} \frac{\overline{Y}_i - \mu_i}{\sigma}) - Var(\sqrt{n} \frac{\overline{\overline{Y}} - \bar{\mu})}{\sigma}) \Big] = \sigma^2$$

hence

$$s^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i [(\overline{Y}_i - \overline{\overline{Y}} - (\mu_i - \bar{\mu})]^2$$

is an unbiased estimator of σ^2.

Under $H_0 : \mu_1 = \cdots = \mu_k$, $\mu_i - \bar{\mu} = 0$ and s^2 becomes an unbiased estimator of σ^2, independent of the μ_i's, which is sensitive to the variation in the sizes of the subpopulations. This represents the the estimator of the variance among groups.

We may think of comparing the two variance estimators in order to accept or refuse the null hypothesis. Indeed we have the following

Remark 2.5.2. *Let $X \sim \Gamma(\alpha, \lambda)$ and $Y \sim \Gamma(\beta, \lambda)$ be two independent r.v.'s. and let us apply the transformation $\phi(x, y) = (x, \frac{x}{y})$, which immediately gives the inverse function $\phi^{-1}(u, v) = (u, \frac{u}{v})$. Applying the change (2.3), we have that $|detJ_{\phi^{-1}}(u, v)| = \dfrac{u}{v^2}$, whence*

$$\begin{aligned} f_{U,V} &= f_{X,Y}(\phi^{-1}(u,v))|detJ_{\phi^{-1}}(u,v)| \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} u^{\alpha-1} \mathrm{e}^{-\lambda u} \frac{\lambda^\beta}{\Gamma(\beta)} \Big(\frac{u}{v}\Big)^{\beta-1} \mathrm{e}^{-\frac{\lambda}{v}u} \frac{u}{v^2} \mathbf{1}_{\{u>0\}} \mathbf{1}_{\{v>0\}} \\ &= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} \frac{1}{v^{\beta+1}} u^{\alpha+\beta-1} \mathrm{e}^{-\lambda(1+\frac{1}{v})u} \mathbf{1}_{\{u>0\}} \mathbf{1}_{\{v>0\}} \end{aligned}$$

and we may obtain the marginal for $v > 0$

$$\begin{aligned} f_V(v) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{1}{v^{\beta+1}} (1+\frac{1}{v})^{-(\alpha+\beta)} \int_0^{+\infty} \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)} (1+\frac{1}{v})^{\alpha+\beta} u^{\alpha+\beta-1} \mathrm{e}^{-\lambda(1+\frac{1}{v})u} du \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{v^{\alpha-1}}{(v+1)^{\alpha+\beta}}. \end{aligned}$$

When $X \sim \chi^2(n) = \Gamma(\frac{n}{2}, \frac{1}{2})$ and $Y \sim \chi^2(m) = \Gamma(\frac{m}{2}, \frac{1}{2})$ then $\dfrac{X}{n} \sim \Gamma(\frac{n}{2}, \frac{n}{2})$ and $\dfrac{Y}{m} \sim \Gamma(\frac{m}{2}, \frac{m}{2})$ and the density of $U = \dfrac{X/n}{Y/m}$ is called a Fisher density with n, m degrees of freedom and it is explicitly obtained from the above formula. It is usually written

$$\frac{X}{Y}\frac{m}{n} \sim F(n, m).$$

The Fisher density with n, m degrees of freedom is tabulated for several choices of n, m and its quantiles of order α are usually denoted by $F_\alpha(n, m)$.

By the previous remark, we have that

$$(n-k)\frac{s_p^2}{\sigma^2} \sim \chi^2(n-k), \qquad (k-1)\frac{s^2}{\sigma^2} \sim \chi^2(k-1)$$

are independent and we may say that

$$\frac{\frac{(n-k)\frac{s_p^2}{\sigma^2}}{n-k}}{\frac{(k-1)\frac{s^2}{\sigma^2}}{k-1}} = \frac{s_p^2}{s^2} \sim F(n-k, k-1)$$

A test of order α may be run

$$H_0 : \mu_1 = \cdots = \mu_k \quad versus \quad \mu_i \neq \mu_j \quad for\, some \quad i \neq j$$

by rejecting H_0 if $\dfrac{s_p^2}{s^2} > F_\alpha(n-k, k-1)$.

2.6 Exercises

Section 11.4: # 11.9, 11.10

Capitolo 3

LINEAR REGRESSION

One of the main problems in statistics is to understand how the multidimensional statistical data are correlated. It is often easy to have an intuition about the correlation among different quantities, but it is much harder to estimate this relation with precision, so that it might be used also for predictive purposes.

The linear regression models try to give an answer to this kind of problem, assuming that the quantities at play are linked by a linear relation that is subject to a perturbation. In other words, a variable is viewed as an input biased by an error and the other one as the resulting output. In absence of any specific information it is quite natural to assume that the error should be represented by Gaussian r.v.'s with mean zero (equally likely errors by defect or excess).

3.1 The linear regression model

We introduce the model for two correlated samples and later we will extend to the multidimensional case.

Therefore we assume to have two data sets given by $x_1, \dots, x_n$ and $y_1, \dots, y_n$, where the latter is a realization of the random sample

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \dots, n, \tag{3.1}$$

for some constants α and β and ϵ_i are independent errors with $\mathbb{E}(\epsilon_i) = 0$ and unknown variance $\mathbb{E}(\epsilon_i^2) = \sigma^2$, for all $i = 1, \dots, n$.

Our aim is to use our observables, i.e. the input $x_1, \dots, x_n$ and the output $y_1, \dots, y_n$, to find the best relation of the type (3.1) linking the r.v.'s Y_i and $\alpha + \beta X_i$, where we know that $\mathbb{E}(Y_i | X_i = x_i) = \alpha + \beta x_i$ and $Var(Y_i | X_i = x_i) = \sigma^2$ and the r.v.'s $Y_i | X_i = x_i$ are independent.

The Y_i is sometimes called the **dependent or response variable** and $X_i = x_i$ as the **independent variable or predictor**.

Therefore we are looking for the best estimators for the three parameters α, β, σ to make the distance minimal between Y_i and X_i.

Remark 3.1.1. *Linear regression specifies a linear dependence on the* ***parameters****. For instance*

$$\mathbb{E}(Y_i|X_i = x_i) = \alpha + \beta x_i^2.$$

still specifies a linear regression of Y_i *on*x_i^2*, while the model*

$$\mathbb{E}(Y_i|X_i = x_i) = \alpha + \beta^2 x_i.$$

specifies a quadratic regression between of Y_i *on* x_i*.*

So, given the data set $(x_1, y_1), \ldots, (x_n, y_n)$, realized by the random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, we want to minimize the distance between y_i and $\mathbb{E}(Y_i|X_i)$ for all $i = 1, \ldots, n$. Thus we want to find the functions $\hat{\alpha}(\mathbf{x}, \mathbf{y})$ and $\hat{\beta}(\mathbf{x}, \mathbf{y})$ that realize the the minimum of the quadratic function

$$f(\alpha, \beta) = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2.$$

As usual, being f a convex function, once we find a critical point, it has to be a minimum, thus differentiating f with respect to α and β and setting the partial derivatives equal to 0, we have

$$\begin{aligned} \frac{\partial f}{\partial \alpha} &= -2 \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)] = 0 \\ \frac{\partial f}{\partial \beta} &= -2 \sum_{i=1}^{n} x_i [y_i - (\alpha + \beta x_i)] = 0 \end{aligned}$$

whence

$$\begin{aligned} \bar{y} &:= \frac{1}{n} \sum_{i=1}^{n} y_i = \alpha + \beta \frac{1}{n} \sum_{i=1}^{n} x_i =: \alpha + \beta \bar{x} \\ \overline{xy} &:= \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \alpha \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{1}{n} \beta \sum_{i=1}^{n} x_i^2 =: \alpha \bar{x} + \beta \overline{x^2} \end{aligned}$$

and solving the system we obtain

$$\begin{aligned} \hat{\alpha} &:= \bar{y} - \beta \bar{x} \\ \overline{xy} &:= \hat{\alpha} \bar{x} + \beta \overline{x^2} = \bar{y} \bar{x} + \hat{\beta} [\overline{x^2} - \bar{x}^2] \quad \Rightarrow \quad \hat{\beta} = \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} \\ \Rightarrow \quad \hat{\alpha} &= \bar{y} - \hat{\beta} \bar{x} = \bar{y} - \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} \bar{x} \end{aligned}$$

and we denote the sample variances and sample covariance as

$$\overline{x^2} - \bar{x}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 =: \frac{S_{xx}^2}{n}, \qquad \overline{y^2} - \bar{y}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 =: \frac{S_{yy}^2}{n}$$

$$\overline{xy} - \bar{y}\bar{x} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) =: \frac{S_{xy}}{n}$$

In conclusion we have the following estimating functions for α and β

$$\hat{\beta}(\mathbf{x},\mathbf{y}) = \frac{S_{xy}}{S_{xx}^2}, \qquad \hat{\alpha}(\mathbf{x},\mathbf{y}) = \bar{y} - \hat{\beta}(\mathbf{x},\mathbf{y})\bar{x}.$$

We can easily prove that the two estimators $\alpha(\mathbf{X},\mathbf{Y})$ and $\beta(\mathbf{X},\mathbf{Y})$ are unbiased, namely

$$\begin{aligned}
\mathbb{E}(\hat{\beta}(\mathbf{X},\mathbf{Y})) &= \mathbb{E}\Big(\frac{\sum_{i=1}^n (Y_i - \overline{Y}_n)(X_i - \overline{X}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}\Big) \\
&= \mathbb{E}\Big(\mathbb{E}\Big(\frac{\sum_{i=1}^n (Y_i - \overline{Y}_n)(X_i - \overline{X}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}|\mathbf{X}=\mathbf{x}\Big)\Big) \\
&= \mathbb{E}\Big(\sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2}\mathbb{E}(Y_i - \overline{Y}_n|\mathbf{X}=\mathbf{x})\Big) \\
&= \mathbb{E}\Big(\sum_{i=1}^n \beta\frac{(x_i-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\Big) = \beta
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(\hat{\alpha}(\mathbf{X},\mathbf{Y})) &= \mathbb{E}(\overline{Y}) - \mathbb{E}(\hat{\beta}(\mathbf{X},\mathbf{Y})\overline{X}) = \alpha + \beta\bar{x} - \mathbb{E}(\mathbb{E}(\hat{\beta}(\mathbf{X},\mathbf{Y})\overline{X}|\mathbf{X}=\mathbf{x})) \\
&= \alpha + \beta\bar{x} - \mathbb{E}(\bar{x}\mathbb{E}(\hat{\beta}(\mathbf{X},\mathbf{Y})|\mathbf{X}=\mathbf{x})) = \alpha + \beta\bar{x} - \bar{x}\mathbb{E}(\beta) = \alpha.
\end{aligned}$$

From now on, to make our notation lighter, even though abusing it a little, we will keep the sample $\mathbf{X}$ fixed and the realized value $\mathbf{x}$ and we will omit the conditioning on the sample $\mathbf{Y}$ We may also easily compute the variances of the two estimators. First of all let us notice that $Y_1|X_1 = x_1, \ldots, Y_n|X_n = x_n$ are independent random variables. If we take a linear combination of those r.v.'s

$$\sum_{i=1}^n d_i Y_i, \quad \text{such that} \quad \sum_{i=1}^n d_i = 0 \tag{3.2}$$

then we have

$$\sum_{i=1}^n d_i(Y_i - \overline{Y}_n) = \sum_{i=1}^n d_i Y_i - \overline{Y}_n \sum_{i=1}^n d_i = \sum_{i=1}^n d_i Y_i$$

Let us take $d_i = \dfrac{(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2}$, which verify the relation (3.2), therefore we have

$$\begin{aligned}
Var(\hat{\beta}) &= Var\Big(\frac{\sum_{i=1}^n (Y_i - \overline{Y}_n)(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2}\Big) = Var\Big(\sum_{i=1}^n d_i(Y_i - \overline{Y}_n)\Big) \\
&= Var\Big(\sum_{i=1}^n d_i Y_i\Big) = \sum_{i=1}^n d_i^2 Var(Y_i) = \sigma^2 \sum_{i=1}^n d_i^2 = \frac{\sigma^2}{S_{xx}^2}.
\end{aligned}$$

By noticing that $\overline{Y}_n$ and $\hat{\beta}$ are uncorrelated we may compute also $\mathrm{Var}(\hat{\alpha})$, indeed

$$\begin{aligned} cov(\overline{Y}_n, \hat{\beta}) &= cov(\overline{Y}_n, \hat{\beta}) = cov(\overline{Y}_n, \sum_{i=1}^{n} d_i Y_i) = \sum_{i=1}^{n} d_i cov(\overline{Y}_n, Y_i) \\ &= \sum_{i=1}^{n} d_i cov(\frac{1}{n}\sum_{j=1}^{n} Y_j, Y_i) = \sum_{i=1}^{n} d_i \frac{1}{n} cov(Y_i, Y_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n} d_i = 0 \end{aligned}$$

since the r.v.'s Y_i are independent and $\mathrm{Var}(Y_i) = \sigma^2$ for all i's. Consequently,

$$\begin{aligned} Var(\hat{\alpha}) &= Var(\overline{Y}_n - \hat{\beta}\bar{x}) = Var(\overline{Y}_n) + Var(\hat{\beta}\bar{x}) + cov(\overline{Y}_n, \hat{\beta}\bar{x}) \\ &= \frac{\sigma^2}{n} + \bar{x}^2 Var(\hat{\beta}) + \bar{x} cov(\overline{Y}_n, \hat{\beta}) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}^2} = \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}\Big). \end{aligned}$$

To estimate the parameter σ, the natural choice falls on

$$\frac{S_{yy}^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2.$$

Unfortunately, this estimator is biased, indeed remarking that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}x_i + \hat{\alpha}) = \hat{\beta}\bar{x} + \hat{\alpha} = \bar{y}$$

we have

$$\begin{aligned} \mathbb{E}(S_{YY}^2) &= \sum_{i=1}^{n}\mathbb{E}(Y_i^2) - n\mathbb{E}(\overline{Y}^2) = \sum_{i=1}^{n}(\sigma^2 + (\alpha + \beta x_i)^2) - n[\frac{\sigma^2}{n} + (\alpha + \bar{x}\beta)^2] \\ &= (n-1)\sigma^2 + n\alpha^2 + 2\alpha\beta\sum_{i=1}^{n} x_i + \beta^2\sum_{i=1}^{n} x_i^2 - n\alpha^2 - 2n\bar{x}\alpha\beta - n\beta^2\bar{x}^2 \\ &= (n-1)\sigma^2 + \beta^2[\sum_{i=1}^{n} x_i^2 - \bar{x}^2] = (n-1)\sigma^2 + \beta^2 S_{xx}^2 \end{aligned}$$

On the other hand, setting $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ the estimated values, we have that

$$\begin{aligned} S_{yy}^2 - \beta^2 S_{xx}^2 &= \sum_{i=1}^{n}[(y_i - \bar{y})^2 - \beta^2(x_i - \bar{x})^2] = \sum_{i=1}^{n}[(y_i - \hat{\alpha} - \hat{\beta}\bar{x})^2 - \beta^2(x_i - \bar{x})^2] \\ &= \sum_{i=1}^{n}[(y_i - \hat{\alpha} - \hat{\beta}x_i + \hat{\beta}(x_i - \bar{x}))^2 - \beta^2(x_i - \bar{x})^2] \\ &= \sum_{i=1}^{n}[(y_i - \hat{y}_i + \hat{\beta}(x_i - \bar{x}))^2 - \beta^2(x_i - \bar{x})^2] \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\hat{\beta}(y_i - \hat{y}_i)(x_i - \bar{x}) + (x_i - \bar{x})^2(\hat{\beta}^2 - \beta^2)] \end{aligned}$$

So applying the above equalities of the model and taking expectations we get to

$$\begin{aligned}
(n-1)\sigma^2 &= \mathbb{E}(S_{YY}^2) - \beta^2 S_{xx}^2 \\
&= \mathbb{E}\Big[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2\Big] + \sum_{i=1}^n (x_i - \bar{x})^2 \mathbb{E}[\hat{\beta}^2 - \beta^2] + 2\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}(\hat{\beta}(Y_i - \hat{Y}_i)) \\
&= \sum_{i=1}^n \mathbb{E}(R_i^2) + S_{xx}^2 \frac{\sigma^2}{S_{xx}^2} + 2\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}(\hat{\beta}(\alpha - \hat{\alpha} + \beta x_i - \hat{\beta} x_i)) \\
&= \sum_{i=1}^n \mathbb{E}(R_i^2) + \sigma^2 + 2\sum_{i=1}^n (x_i - \bar{x}) x_i \mathbb{E}(\beta\hat{\beta} - (\hat{\beta})^2) \\
&= \sum_{i=1}^n \mathbb{E}(R_i^2) + \sigma^2 + 2\sum_{i=1}^n (x_i - \bar{x})^2 (\beta^2 - \beta^2 - \frac{\sigma^2}{S_{xx}^2}) = \sum_{i=1}^n \mathbb{E}(R_i^2) - \sigma^2
\end{aligned}$$

where we used the expression of Var $(\hat{\beta})$ and we set $R_i = Y_i - \hat{Y}_i$. Hence

$$\frac{1}{n-2}\sum_{i=1}^n \mathbb{E}(R_i^2) = \sigma^2$$

and we may conclude that $s^2 := \dfrac{1}{n-2}\displaystyle\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ is an unbiased estimator of σ^2. The r.v.'s R_i are called **residues**, that we will denote by r_i when referring to the realized values.

Let us remark that the residues r_i have the following property (verify as an exercise)

$$\sum_{i=1}^n r_i = 0, \qquad \sum_{i=1}^n r_i \hat{y}_i = 0 \tag{3.3}$$

By using property (3.3), with a few calculations (do it as an exercise), it is possible to show that

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

which implies that the total variance of the model is made up of two components, the first gives the error we make by substituting the actual values with the estimated values, the second represents the variance of the model. Therefore, if we take the ratio

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

that is necessarily less than or equal 1, we obtain the percentage of the variance that is explained by the model.

3.2 Hypothesis Testing

So far, we have not assumed any distributional property of the model. If we now take Gaussian errors $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, we consequently deduce the distributional properties for all the estimators (given $X_i = x_i$ for all $I = 1, \dots, n$)

$$\begin{aligned} Y_i &\sim \mathcal{N}(\alpha + \beta x_i, \sigma^2) \\ \overline{Y}_n &\sim \mathcal{N}(\alpha + \beta \bar{x}, \frac{\sigma^2}{n}) \\ \hat{\beta} &\sim \mathcal{N}(\beta, \frac{\sigma^2}{S_{xx}^2}) \quad \Rightarrow \frac{\hat{\beta} - \beta}{\frac{\sigma}{S_{xx}}} \sim \mathcal{N}(0,1) \\ \hat{\alpha} &\sim \mathcal{N}(\alpha, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2})) \quad \Rightarrow \frac{\hat{\alpha} - \alpha}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}}} \sim \mathcal{N}(0,1) \\ (n-2)\frac{s^2}{\sigma^2} &\sim \chi^2(n-2), \quad \text{independent of} \quad \hat{\beta} \text{ and } \hat{\alpha} \end{aligned}$$

The latter because of Cochran's theorem. Hence the statistics

$$\frac{\hat{\beta} - \beta}{s} S_{xx} \sim t(n-2), \quad \frac{\hat{\alpha} - \alpha}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}}} \sim t(n-2)$$

may be used to perform a bilateral or unilateral hypothesis test on the two parameters.

We might run the usual hypothesis testing for β either

$$H_0 : \beta = 0, \qquad H_1 : \beta \neq 0$$

comparing (under H_0) the statistic $\frac{\hat{\beta}}{s} S_{xx}$ with $t_{1-\alpha/2}(n-2)$ or a bilateral test

$$H_0 : \beta \geq 0, \qquad H_1 : \beta < 0$$

comparing the statistic $\frac{\hat{\beta}}{s} S_{xx}$ with $t_{1-\alpha}(n-2)$.

We point out that, under H_0, by squaring the above statistic, from the distribution of $\hat{\beta}$ and that of s^2 we may construct an ANOVA test given by the statistic

$$\frac{\hat{\beta}^2}{s^2} S_{xx}^2 \sim F(1, n-2) \tag{3.4}$$

as the ratio between two independent $\chi^2(1)$ and a $\chi^2(n-2)$.

The statistic (3.4) is indeed relative to an ANOVA test, since , under H_0 we may rewrite

$$\frac{\hat{\beta}^2}{s^2} S_{xx}^2 = \frac{\frac{S_{xY}^2}{S_{xx}^4} S_{xx}^2}{(n-2)\sum_{i=1}^n R_i^2} = \frac{\frac{S_{xY}^2}{S_{xx}^2}}{(n-2)\sum_{i=1}^n R_i^2} = \frac{\sum_{i=1}^n c_i (Y_i - \overline{Y}_n)^2}{(n-2)\sum_{i=1}^n R_i^2},$$

where the numerator is a weighted average of the quadratic distances, providing an estimator of the total variance, since

$$c_i = \frac{(x_i - \bar{x})^2}{S_{xx}^2} \geq 0 \quad \Rightarrow \quad \sum_{i=1}^{n} c_i = 1.$$

Remark 3.2.1. *When assuming a Gaussian distribution, the estimators we found for α and β are also the MLE estimators, indeed the likelihood function is given by*

$$L(\mathbf{y}|\alpha, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\{-\sum_{i=1}^{n} \frac{[y - (\alpha + \beta x_i)]^2}{2\sigma^2}\}$$

and consequently the loglikelihood function is

$$\ln L(\mathbf{y}|\alpha, \beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \sum_{i=1}^{n} \frac{[y - (\alpha + \beta x_i)]^2}{2\sigma^2}.$$

Maximizing the latter we find that $\hat{\alpha}$ and $\hat{\beta}$ are also the MLE estimators for α and β, while the MLE estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

3.3 Prediction

Linear regression may be used for predictive purposes to forecast what the result of an additional input will be. In the conditional Gaussian model we may give an accurate estimate, exploiting the previous results.

Namely we are hypothesizing that, given a new value of the predictor, say x_0, then the value of the output should be

$$(Y_0|X = x_0) = y_0 = \alpha + \beta x_0 + \epsilon_0,$$

where $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian error independent of all the other ones.

Exploiting the results of the previous section, we are able to provide an estimate of y by $\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$ and the approximation error is given by

$$y_0 - \hat{y}_0 = \alpha + \beta x_0 + \epsilon_0 - (\hat{\alpha} + \hat{\beta}x_0) = \epsilon_0 + (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})x_0$$

Since $\mathbb{E}(\hat{Y}_0) = \mathbb{E}(\hat{\alpha} + \hat{\beta}x_0) = \alpha + \beta x_0$, this is an unbiased estimator of $E(Y_0|X = x_0)$, moreover we know that $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$, $(\alpha - \hat{\alpha}) \sim \mathcal{N}(0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}))$, $(\beta - \hat{\beta}) \sim \mathcal{N}(0, \frac{\sigma^2}{S_{xx}^2})$ the latter two being independent of the first. This implies that $\hat{Y}_0$ has a Gaussian distribution also and it is enough to compute its variance.

$$\begin{aligned} & Var(\hat{\alpha} + \hat{\beta}x_0) = Var(\hat{\alpha}) + \bar{x}_0^2 Var(\hat{\beta}) + 2x_0 cov(\hat{\alpha}, \hat{\beta}) \\ = \; & Var(\hat{\alpha}) + \bar{x}_0^2 Var(\hat{\beta}) + 2x_0 cov(\overline{Y}_n - \hat{\beta}\bar{x}, \hat{\beta}) \\ = \; & Var(\hat{\alpha}) + \bar{x}_0^2 Var(\hat{\beta}) + 2x_0 cov(\overline{Y}_n,\ \hat{\beta}) - 2x_0\bar{x}Var(\hat{\beta}) = Var(\hat{\alpha}) + (x_0^2 - 2x_0\bar{x})Var(\hat{\beta}) \end{aligned}$$

whence

$$\begin{aligned} Var(Y_0 - \hat{Y}_0) &= Var(\epsilon_0) + Var(\hat{\alpha} + \hat{\beta}x_0) = \sigma^2 + \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}) + (\bar{x}_0^2 - 2\bar{x}_0\bar{x})\frac{\sigma^2}{S_{xx}^2} \\ &= \sigma^2\Big[(1 + \frac{1}{n}) + \frac{(x_0 - \bar{x})^2}{S_{xx}^2}\Big]. \end{aligned}$$

Lastly, $\hat{\alpha}, \hat{\beta}, \epsilon$ are all independent of $(n-2)\frac{s^2}{\sigma^2} \sim \chi^2(n-2)$, therefore so is $Y_0 - \hat{Y}_0$ and we may conclude that

$$\frac{Y_0 - \hat{Y}_0}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2}}} \sim t(n-2).$$

Consequently Y_0 falls in the (confidence) interval

$$\Big[\hat{Y}_0 - \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2}}t_{1-\alpha/2)}(n-2), \hat{Y}_0 - \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2}}t_{1-\alpha/2)}(n-2)\Big]$$

with probability greater than or equal to $1 - \alpha$.

3.4 Multidimensional Regression models

In this section we generalize the regression model to a multidimensional setting, where the input is given by more than one factor.

Before doing so, we need a little digression to define multivariate Gaussian random variables. If

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}$$

is a vector of m independent $\mathcal{N}(0,1)$ r.v.'s, then we know that their joint density is given by

$$f_{\mathbf{Z}}(z_1, \ldots, z_m) = \frac{1}{(2\pi)^{m/2}} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^m z_i^2}.$$

In vector and matrix notation we have that

$$\sum_{i=1}^m z_i^2 = ||\mathbf{z}||^2 = \mathbf{z}^* \cdot \mathbf{z} = \mathbf{z}^* I \mathbf{z},$$

where $\mathbf{z}^* = (z_1, \ldots, z_m)$ and I is the $m \times m$ identity matrix.

Consider a linear transformation $\mathbf{Y} = \phi(\mathbf{Y}) = \mathbf{C}|\mathbf{Z} + \mathbf{d}$, with $\mathbf{C}$ an invertible $m \times m$ matrix and $\mathbf{d} \in \mathbb{R}^m$, then $J_\phi \equiv \mathbf{C}|$, $|det\, \mathbf{C}|| \neq 0$ and

$$\mathbf{Z} = \mathbf{C}^{-1}(\mathbf{Y} - \mathbf{d}),$$

consequently the joint density of the vector $\mathbf{Y}$ will be given by

$$f_{\mathbf{Y}}(y_1,\dots,y_m)=\frac{1}{(2\pi)^{m/2}}\frac{1}{|det\,\mathbf{C}}\mathrm{e}^{-\frac{1}{2}||\mathbf{C}^{-1}(\mathbf{y}-\mathbf{d})||^2}$$

and we may also write

$$\begin{aligned}||\mathbf{C}^{-1}(\mathbf{y}-\mathbf{d})||^2 &= (\mathbf{C}^{-1}(\mathbf{y}-\mathbf{d}))^*\mathbf{C}^{-1}(\mathbf{y}-\mathbf{d})=(\mathbf{y}-\mathbf{d})^*(\mathbf{C}^{-1})^*\mathbf{C}^{-1}(\mathbf{y}-\mathbf{d})\\ &= (\mathbf{y}-\mathbf{d})^*(\mathbf{C}^*)^{-1}\mathbf{C}^{-1}(\mathbf{y}-\mathbf{d})=(\mathbf{y}-\mathbf{d})^*(\mathbf{C}\mathbf{C}^*)^{-1}(\mathbf{y}-\mathbf{d}).\end{aligned}$$

Let us denote by $\Sigma=\mathbf{C}\mathbf{C}^*$, then $(\mathbf{C}\mathbf{C}^*)_{ij}=\mathrm{cov}(Y_i,Y_j)$, so we call Σ the variance/covariance matrix. This matrix is symmetric and invertible since $det\,\Sigma=(det\,A)^2\neq 0$ and we may conclude that

$$f_{\mathbf{Y}}(y_1,\dots,y_m)=\frac{1}{(2\pi)^{m/2}}\frac{1}{\sqrt{det\Sigma}}\mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\mathbf{d})^*\Sigma^{-1}(\mathbf{y}-\mathbf{d})} \tag{3.5}$$

and we say that $\mathbf{Y}$ follows a multivariate Gaussian distribution $\mathcal{N}(\mathbf{d},\Sigma)$, while $\mathbf{Z}\sim\mathcal{N}(\mathbf{0},I)$ follows a standard multivariate Gaussian distribution.

As a matter of fact, we may repeat the previous argument for any matrix $\mathbf{C}$ $m\times k$ provided that $\Sigma:=(\mathbf{C}\mathbf{C}^*)$ is an invertible matrix and we may summarize our result by saying that given a standard multivariate Gaussian distribution $\mathbf{Z}\sim\mathcal{N}(\mathbf{0},I_{m\times m})$, any linear transformation

$$\mathbf{Y}=\mathbf{C}\mathbf{Z}+\mathbf{d},\qquad \mathbf{C}\quad m\times k\quad m\geq k$$

will follow a $\mathcal{N}(\mathbf{d},\mathbf{C}\mathbf{C}^*)$.

Remark 3.4.1. *We remark that given an invertible square symmetric matrix Σ, its eigenvalues will be all strictly greater than zero. If Λ denotes the diagonal matrix of its eigenvalues , then $\Sigma=O\Lambda O^{-1}$ will be an orthogonal transformation ($O^{-1}=O^*$) of Λ. Consequently, we may define the matrix $\sqrt{\Lambda}$ as the diagonal matrix that has the square roots of the eigenvalues on its main diagonal and*

$$\begin{aligned}A=\sqrt{\Sigma}:=&\ O\sqrt{\Lambda}O^{-1}\\ AA^* &= O\sqrt{\Lambda}O^{-1}(O\sqrt{\Lambda}O^*)^*=O\sqrt{\Lambda}O^*O\sqrt{\Lambda}O^*=O\Lambda O^*=\Sigma\\ A^{-1} &= (O\sqrt{\Lambda}O^{-1})^{-1}=O\sqrt{\Lambda}^{-1}O^{-1}\\ A^{-1}(A^{-1})^* &= O\sqrt{\Lambda}^{-1}O^{-1}(O\sqrt{\Lambda}^{-1}O^*)^*=O\Lambda^{-1}O^*=\Sigma^{-1}\\ \mathbf{w}^*\Sigma^{-1}\mathbf{w} &= \mathbf{w}^*A^{-1}(A^{-1})^*\mathbf{w}=\langle\mathbf{w}^*,A^{-1}(A^{-1})^*\mathbf{w}\rangle=\langle(A^{-1})^*\mathbf{w}^*,(A^{-1})^*\mathbf{w}\rangle=||(A^{-1})^*\mathbf{w}||^2\end{aligned}$$

In the multidimensional regression model we have $X_2,\dots X_k$ independent r.v.'s that take the values $x_i^2,\dots,x_i^k$ and r.v's $Y_1,\dots,Y_n$, representing the output determined by the k inputs. Knowing that

$$(Y_i|X_2=x_i^2,\dots X_k=x_i^k)=y_i,\qquad i=1,\dots n,$$

the suggested linear regression model is

$$Y_i=\beta_1\cdot 1+\beta_2x_i^2+\cdots+\beta_kx_i^k+\epsilon_i,\qquad \beta_1,\dots,\beta_k\in\mathbb{R}, \tag{3.6}$$

where the errors $\epsilon_i (\sim \mathcal{N}(0, \sigma^2))$ are taken to be independent, which immediately gives

$$\mathbb{E}(Y_i | X_2 = x_i^2, \dots X_k = x_i^k) = \beta_1 \cdot 1 + \beta_2 x_i^2 + \cdots + \beta_k x_i^k$$

and we are going to approximate those values by estimating $\beta_1, \dots, \beta_k, \sigma^2$. Abusing the notation a little bit, we are going to omit the conditioning in what follows.

In vector notation, (3.6) may be rewritten as

$$\mathbf{Y} = \beta_1 \mathbf{x}^1 + \beta_k \mathbf{x}^k = \mathbf{A} \cdot \mathbf{b} + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n \times n}), \tag{3.7}$$

where for $j = 2, \dots k$, we denoted

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{x}^1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \mathbf{x}^j = \begin{pmatrix} x_1^j \\ \vdots \\ x_n^j \end{pmatrix}, \mathbf{A} = \begin{pmatrix} x_1^1 & \dots & x_1^k \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^k \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \mathbf{e} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

We assume that $n \geq k$, since we may always add further observations to exceed the number of factors. Since the random vectors $\mathbf{X}^2, \dots, \mathbf{X}^k$ are assumed independent, this implies that the vectors determined by their realizations

$$\mathbf{x}^1 = \begin{pmatrix} x_1^1 \\ \vdots \\ x_n^1 \end{pmatrix}, \quad \dots \quad , \mathbf{x}^k = \begin{pmatrix} x_1^k \\ \vdots \\ x_n^k \end{pmatrix}$$

fix k linearly independent directions. In other words, (3.6) is equivalent to saying that the vector $\mathbb{E}(\mathbf{Y} | \mathbf{X}^1 = \mathbf{x}^1, \dots, \mathbf{X}^k = \mathbf{x}^k)$ lies in $E = \text{span}(\mathbf{x}^1, \dots, \mathbf{x}^k)$, which is a k-dimensional linear subspace of $\mathbf{R}^n$.

By Cochran's theorem and its consequences, we have that the orthogonal projection of $\mathbf{Y}$ on E, $P_E(\mathbf{Y}) = \mathbb{E}(\mathbf{Y} | \mathbf{X}^1 = \mathbf{x}^1, \dots, \mathbf{X}^k = \mathbf{x}^k)$, is independent of $(I - P_E)\mathbf{Y}$, which instead will lie in an $n - k$ dimensional linear subspace.

The orthogonal projection is the minimizer of the distance between $\mathbf{Y}$ and vectors in E, $||\mathbf{Y} - P_E \mathbf{Y}||^2 = \min_{\mathbf{z} \in E} ||\mathbf{Y} - \mathbf{Z}||^2$ and it is given by

$$P_E \mathbf{Y} = (\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* \mathbf{Y}.$$

Indeed we point out that, since $\mathbf{x}^1, \dots, \mathbf{x}^k$ are linearly independent vectors, the $k \times k$ matrix $\mathbf{A}^* \mathbf{A}$ must be invertible and

$$(\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* \mathbf{Y} = \mathbf{b} + (\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* \mathbf{e}$$

is a vector belonging to E, that coincides with $\mathbf{b}$ up to an error that belongs to the orthogonal space $E^\perp$. From (3.7), we obtain that an estimator for the vector $\mathbf{b}$ is given by

$$\hat{\mathbf{b}} = P_E \mathbf{Y} = (\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* \mathbf{Y}.$$

Remarks 3.4.1.

1. $\mathbf{A}^*\mathbf{A}$ *is the so called sample matrix of variance-covariance of the vectors* $\mathbf{x}^1, \dots \mathbf{x}^k$, *indeed it is a* $k \times k$ *symmetric matrix and for* $j, h = 1, \dots, k$

$$\begin{aligned}(\mathbf{A}^*\mathbf{A})_{11} &= 1, \quad (\mathbf{A}^*\mathbf{A})_{jj} = \sum_{i=1}^{n} (x_i^j)^2 \\ (\mathbf{A}^*\mathbf{A})_{1j} &= \sum_{i=1}^{n} x_i^j, \quad (\mathbf{A}^*\mathbf{A})_{hj} = \sum_{i=1}^{n} x_i^h x_i^j\end{aligned}$$

2. *As a linear transformation of* $\mathbf{e}$, *also* $(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{e}$ *is a Gaussian vector* $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C}\mathbf{C}^*)$, *where* $\mathbf{C} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$.

3. *Indeed we have that the* $k \times k$ *variance/covariance matrix of the vector* $\mathbf{Y}$

$$\begin{aligned}\Sigma &:= \mathbf{C}\mathbf{C}^* = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*((\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*)^* = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}((\mathbf{A}^*\mathbf{A})^{-1})^* \\ &= \mathbf{I}((\mathbf{A}^*\mathbf{A})^{-1})^* = ((\mathbf{A}^*\mathbf{A})^*)^{-1} = (\mathbf{A}^*\mathbf{A})^{-1}\end{aligned}$$

   is the inverse of the variance/covariance matrix of the original vector $\mathbf{X}$.

4. *The estimator* $\hat{\mathbf{b}}$ *is certainly unbiased, because*

$$\mathbb{E}(\hat{\mathbf{b}}) = \mathbb{E}((\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{Y}) = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbb{E}(\mathbf{Y}) = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}\mathbf{b} = \mathbf{b}$$

   since the vector of the expectations $\mathbb{E}(\mathbf{e})$ *has all the components equal to 0.*

5. $\hat{\mathbf{b}} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{e} + \mathbf{b}$ *is a linear transformation of the vector* $\mathbf{e}$ *from* $\mathbb{R}^n$ *in* $\mathbb{R}^k$, *hence it is a random vector with Gaussian density* $\mathcal{N}(\mathbf{b}, \sigma^2(\mathbf{A}^*\mathbf{A})^{-1}) = \mathcal{N}(\mathbf{b}, \Sigma)$*:*

$$f_{Z_1,\dots,Z_k}(z_1, \dots, z_k) = \frac{1}{(2\pi)^{n/2}\sqrt{det\,\Sigma}} \exp\{-\frac{1}{2}(\mathbf{z} - \mathbf{b})^*\Sigma^{-1}(\mathbf{z} - \mathbf{b})\}$$

   This implies that the standardized vector

$$\frac{1}{\sigma}\mathbf{A}^*\mathbf{A}(\mathbf{b} - \hat{\mathbf{b}}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{k\times k})$$

6. *The* n *dimensional vector of estimated values*

$$\hat{\mathbf{Y}} = \mathbf{A}\hat{\mathbf{b}} \quad \Rightarrow \mathbf{Y} = \mathbf{Y} - \hat{\mathbf{Y}} + \mathbf{Y} - \hat{\mathbf{Y}}$$

   is therefore the orthogonal projection of $\mathbf{Y}$ *on the space* E *and the vector of the residues* $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$ *is the orthogonal projection of* $\mathbf{Y}$ *on* $E^\perp$, *which is an* $n - k$ *dimensional linear subspace and, by Cochran's theorem the two vectors are independent. It follows that the variance estimator*

$$s^2 = \frac{||\mathbf{R}||^2}{n-k} = \frac{||\mathbf{Y} - \hat{\mathbf{Y}}||^2}{n-k} = \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{n-k}$$

is such that $(n-k)\frac{s^2}{\sigma^2} \sim \chi^2(n-k)$ *independent of* $\hat{\mathbf{Y}}, \hat{\mathbf{b}}$.

Each component of the estimating vector of the parameters is thus independent of s^2 *and we may standardize it and conclude*

$$\sigma_{ii}^2 := \sigma^2(\mathbf{A}^*\mathbf{A})^{-1})_{ii}, \qquad \Rightarrow \quad \frac{\hat{\beta}_i - \beta_i}{\sigma\sigma_{ii}} \sim \mathcal{N}(0,1), \quad \frac{\hat{\beta}_i - \beta_i}{s\,\sigma_{ii}} \sim t(n-k)$$

and we may perform all the usual hypothesis tests.

As an example, we may perform a global test of linear dependence, keeping in mind that the vector $\mathbf{x}^1$ corresponds to the constant term

$$H_0 : \beta_2 = \cdots = \beta_k = 0 \qquad \beta_i \neq 0 \quad \text{for some } i = 2, \ldots, k.$$

Indeed, the residue vector $\mathbf{R}$ is independent of $\hat{\mathbf{Y}}$, thus it is independent of $\hat{\mathbf{Y}} - \bar{y}$ also and the centered vector $\mathbf{Y} - \bar{y}$ gets decomposed into three independent vectors (since the first component will be always equal to 0)

$$\mathbf{Y} - \bar{y}\mathbf{1} = \mathbf{Y} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}} - \bar{y}\mathbf{1} + 0 \cdot (1, 0, \ldots, 0)$$

belonging to three orthogonal linear subspaces of respective dimensions $n-k, k-1, 1$. Consequently

$$\frac{(\mathbf{Y} - \bar{y}\mathbf{1})^2}{\sigma^2}(k-1) \sim \chi^2(k-1)$$

and we may perform an ANOVA F-test with the statistic

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \sim F(k-1, n-k)$$

and evaluate the determination coefficient

$$R = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

to establish what portion of the variance is explained by the linear model and to decide whether the model is acceptable or not.

As in the unidimensional case, we may use the multidimensional model for predictive purposes, assuming that a new output y_0 is generated by the model by a further input $\mathbf{x}_0$, i.e.

$$y_0 = \beta_1 x_0^1 + \beta_k x_0^k + \epsilon_0, \qquad \epsilon_0 \sim \mathcal{N}(0, \sigma^2) \quad \text{independent of } \epsilon_i,\ i = 1, \ldots, n$$

hence setting $\hat{y}_0 = \hat{\mathbf{b}}^* \cdot \mathbf{x}_0$, we have that

$$y_0 - \hat{y}_0 = (\mathbf{b} - \hat{\mathbf{b}})^* \cdot \mathbf{x}_0 + \epsilon_0 \sim \mathcal{N}(0, \sigma^2(1 + ((\mathbf{A}^*\mathbf{A})^{-1}\mathbf{x}_0)^* \cdot \mathbf{x}_0))$$

as linear transformation of a Gaussian. In order to compute the variance let us note that by independence

$$Var(y_0 - \hat{y}_0) = Var((\mathbf{b} - \hat{\mathbf{b}})^* \cdot \mathbf{x}_0 + \epsilon_0) = \sigma^2 + Var(\hat{\mathbf{b}})^* \cdot \mathbf{x}_0)$$

but

$$\hat{\mathbf{b}} \sim \mathcal{N}(\mathbf{b}, \sigma^2(\mathbf{A}^*\mathbf{A})^{-1}) \quad \Rightarrow \quad \hat{\mathbf{b}}^* \cdot \mathbf{x}_0 \sim \mathcal{N}(\mathbf{b} \cdot \mathbf{x}_0, \sigma^2((\mathbf{A}^*\mathbf{A})^{-1}\mathbf{x}_0)^* \cdot \mathbf{x}_0)$$

consequently we may use the statistic

$$\frac{y_0 - \hat{y}_0}{s\sqrt{1 + ((\mathbf{A}^*\mathbf{A})^{-1}\mathbf{x}_0)^* \cdot \mathbf{x}_0}} \sim t(n-k)$$

to perform a bilateral or unilateral test.

Indice

1 POINT ESTIMATORS **1**
- 1.1 Best unbiased estimators . . . 1
- 1.2 The method of moments . . . 2
- 1.3 Maximum Likelihood estimators . . . 3
- 1.4 Exercises . . . 4

2 HYPOTHESES TESTING **5**
- 2.1 Definition and construction of a test . . . 5
  - 2.1.1 Error probabilities and Power of a Test . . . 7
- 2.2 Sampling and Testing from Gaussian samples . . . 9
  - 2.2.1 The Cochran theorem . . . 10
- 2.3 Confidence intervals . . . 14
- 2.4 Exercises . . . 16
- 2.5 Tests concerning more populations - ANOVA . . . 17
  - 2.5.1 The classical ANOVA Test . . . 18
  - 2.5.2 The ANOVA F-test . . . 20
- 2.6 Exercises . . . 23

3 LINEAR REGRESSION **24**
- 3.1 The linear regression model . . . 24
- 3.2 Hypothesis Testing . . . 29
- 3.3 Prediction . . . 30
- 3.4 Multidimensional Regression models . . . 31