Inferential Statistics **First semester, 2021**

Lecture 0: Probability and random variables

Instructor: Erlis Ruli (ruli@stat.unipd.it), Department of Statistical Sciences

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

0.1 Preface

All relevant phenomena are subject to uncertainty. Whatever is the phenomenon we are studying, to support our theory we need to:

(i) collect data through experiments;

(ii) analyse the data.

Inferential Statistics offers principled methods for both, collecting and analysing data. This course focuses on (ii) with the aim of extracting evidence, i.e. making *inference*, about or against a scientific theory (supposition, hypothesis, etc.) from the data. To accomplish this inference, the scientific theory must be suitably translated into a probabilistic model. The data are then used in order to measure the support to this probabilistic model, i.e. support to the scientific theory.

Here are three practical inferential problems.

Example 0.1 *Every washing machine (WM) sold in the UE market must be accompanied by a technical documentation which describes, among other things, the energy consumed during a typical washing cycle. In order to measure the consumed energy, the WM is tested in laboratory several times, under the same conditions. With the observed values of energy consumption the manufacturer wants to estimate or learn the most "representative value". This is an estimation problem, the subject of Lecture 4.*

Example 0.2 (Example 0.1 cont'd.) *The manufacturer must also declare a measure of variability for this estimate, i.e. it has to declare tolerance limits within which the consumed energy is expected to vary with high probability. This is an confidence set problem, the subject Lecture 5.*

Example 0.3 (Example 0.2 cont'd.) *It is claimed that to reduce energy consumption current WM motors must be replaced by a newer model. WMs with the current motor and WMs with the new motor are tested and their energy consumption is measured. Based on the observed data we must tell if the claim is true or not. This is an hypothesis testing problem, the subject of Lecture 6.*

0.1.1 Content and course planning

Lecture 0 focuses on probability theory for random variables and probability models. Lecture 1 focuses on random vectors and convergence of random variables. Summary of the data is discussed in Lectures 2 and 3. Lectures 4, 5 and 6, discuss estimation, confidence intervals and hypothesis testing respectively. These are the three pillars of classical (or frequentist) statistical inference. Lecture 7 introduces Bayesian statistics, a popular alternative to classical statistical inference. Finally, Lecture 8 introduces nonparametric statistical inference.

0.2 Set theory

A set S is a well-defined collection of distinct objects. When s *belongs to* S we write $s \in S$. The negation of this statement is $s \notin S$, i.e. s doesn't belong to S. We say that A is a subset of S and write $A \subseteq S$, if for every $s \in A$, it follows that $s \in S$. If also $S \subseteq A$ then $A = S$.

The following definitions are useful for creating new sets from old ones. The complement of the set A with respect to S is the set $A^c = \{s \in S : s \notin A\}$[1].

The union of the sets $A_1, \ldots, A_n$ is denoted by $\cup_{j=1}^n A_j = A_1 \cup A_2 \cup \cdots \cup A_n$, and is defined by

$$\cup_{j=1}^n A_j = \{s \in S : s \in A_j, \text{for at least one } j = 1, 2, \ldots, n\}.$$

[1] This is the usual set builder notation to be read "the set of all s in S such that s is not in A"

The intersection of the sets $A_1, \ldots, A_n$ is denoted by $\cap_{j=1}^{n} A_j = A_1 \cap A_2 \cap \cdots \cap A_n$, and is defined by

$$\cap_{j=1}^{n} A_j = \{s \in S : s \in A_j, \text{for all } j = 1, 2, \ldots, n\}.$$

Set intersection and set union can be extended to an infinite (i.e. countable or uncountable) collections of sets $A_1, A_2, \ldots$. The set which contains no elements is called the *empty set*, denoted by $\emptyset$.

Sets A_1 and A_2 are called *disjoint* if $A_1 \cap A_2 = \emptyset$. Sometimes, when dealing with disjoint sets, we will use the shorthand notation $A + B$ in place of $A \cup B$ especially when the union involves more than two disjoint sets. At times we will use the identity $A \cap B^c = A - B$.

The sets $A_1, A_2, \ldots$ are said to be *pairwise disjoint* if $A_i \cap A_j = \emptyset$ for all $i \neq j$. If $A_1, A_2, \ldots$ are pairwise disjoint sets and $\cup_j A_j = \mathcal{S}$, then such a collection of sets is called a *partition* of S.

Sets have many interesting properties amongst which we recall **De Morgan's laws**:

$$(\cup_j A_j)^c = \cap_j A_j^c \quad \text{and} \quad (\cap_j A_j)^c = \cup_j A_j^c.$$

The sequence of sets $A_1, A_2, \ldots$, is said to be a *monotone sequence* of sets if either

(i) $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, that is A_n is *non decreasing*, to be denoted by $A_n\uparrow$ or

(ii) $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, that is A_n is *non increasing*, to be denoted by $A_n\downarrow$.

The *limit* of a monotone sequence of sets is defined as follows:

(i) $\lim_{n\to\infty} A_n = \cup_{n=1}^{\infty} A_n$ if $A_n\uparrow$;

(ii) $\lim_{n\to\infty} A_n = \cap_{n=1}^{\infty} A_n$ if $A_n\downarrow$.

More generally, for *any* sequence $A_1, A_2, \ldots$, we define the sets

$$\underline{A} = \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{j=n}^{\infty} A_j \quad \text{and} \quad \overline{A} = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{j=n}^{\infty} A_j.$$

The sets $\underline{A}$ and $\overline{A}$ are called the *inferior limit* and *superior limit*, respectively. The sequence $A_1, A_2, \ldots$ has a *limit* if $\overline{A} = \underline{A}$.

0.3 Events, classes of sets and probability

An *experiment* is a procedure performed under certain conditions whereby the procedure can be repeated any number of times under the same conditions, and upon the completion of the procedure certain results are observed. An experiment for which the outcome is random, i.e. it cannot be determined in advance, except that it is known to be one of a set of possible outcomes, is called a *random experiment*. Some random experiments are:

- telephone calls arriving at a telephone exchange,
- defective items in a shipment,
- milliseconds taken by your PC to reboot,
- energy consumption of a WM during a wash cycle, etc..

The set of all possible outcomes of an experiment is called the *sample space* and is denoted by $\mathcal{S}$. Sample spaces mostly fall into one of the following categories:

- **finite set**: such as the toss of a coin for which $\mathcal{S} = \{H, T\}$ with H, T replaced by numbers if they make more sense. More generally, falls in this category any random experiment with n possible outcomes.
- **countable set**: the sample space for an experiment with countably many possible outcomes is ordinarily the set $\mathbb{Z}_{\geq 0} = \{0, 1, \dots, \}$ or the set $\mathbb{Z} = \{\dots, -1, 0, 1, \dots, \}$ of all integers.
- **real line**: $\mathbb{R} = (-\infty, \infty)$ is the most common sample space and is used for virtually all numerical phenomena that are not inherently integer-valued, i.e. measurement errors in scientific observations. Intervals in $\mathbb{R}$ such as the unit interval $[0, 1]$ and the positive half line $\mathbb{R}_{\geq 0} = [0, \infty)$ are also common sample spaces.
- **finite product sample space**: these are random experiments consisting of n replications of an underlying experiment with sample space $\mathcal{S}_0$ (which might be $\{H, T\}$, $\mathbb{N}$ or $\mathbb{R}$). The sample space for the repeated experiment is

$$\mathcal{S} = \mathcal{S}_0 \times \mathcal{S}_0 \times \cdots \times \mathcal{S}_0 = \mathcal{S}_0^n = \{(s_1, \dots, s_n) : s_i \in \mathcal{S}_0 \text{ for all } i\}.$$

- **infinite product sample space**: If a basic experiment is repeated infinitely many times the sample space is the set $\mathcal{S} = \mathcal{S}_0^\infty$ of all infinite sequences $s = (s_1, s_2, \ldots)$ of elements of $\mathcal{S}_0$. An example is the random walk, where $\mathcal{S}_0 = \{-1, 1\}$ and an outcome is the entire sequence s of steps, each of which is ± 1.

- **function spaces**: in some random experiments the outcome is the path or trajectory followed by a system over an interval of time. Outcomes are then functions. For example, if the system is observed over $[0, 1]$ and its path is continuous, we may take $\mathcal{S} = \mathcal{F}[0, 1]$, the vector space of continuous, real-valued function on $[0, 1]$. The probability models for such systems are known as *stochastic processes*.

The elements of $\mathcal{S}$ are called *sample points* and subsets of $\mathcal{S}$ are called *events*, e.g. if $A \subseteq \mathcal{S}$ then A is called an *event*. More precisely, we are interested only on subsets $A \subseteq \mathcal{S}$ that form a *σ-field* of subsets of $\mathcal{S}$. This σ-filed of subsets, denoted by $\mathcal{A}$, is a collection of subsets (i.e. a set of sets) of $\mathcal{S}$ with the following properties:

(i) $\mathcal{A}$ is non empty (non trivial);

(ii) $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$ (closed under complementation);

(iii) $A_j \in \mathcal{A}$, for all $j = 1, 2, \ldots$ implies $\cup_{j=1}^\infty A_j \in \mathcal{A}$ (countably additive).

Thus, $A \subseteq \mathcal{S}$ is called an event provided $A \in \mathcal{A}$. The sets $\mathcal{S}$ and $\emptyset$ are always events and are called the *certain event* and the *impossible event*, respectively.

Definition 0.1 *A probability function $P : \mathcal{A} \to [0, 1]$ is a set function which assigns to each event A a real number denoted by $P(A)$, called the probability of A, and satisfies the following axioms:*

(1) $P(A)$ is non-negative, i.e. $P(A) \geq 0$ for every event A.

(2) P is normed, i.e. $P(\mathcal{S}) = 1$.

(3) P is σ-additive, i.e. for every collection of pairwise disjoint events $A_1, A_2, \ldots$, we have $P(\cup_j A_j) = \sum_j P(A_j)$.

The triple $(\mathcal{S}, \mathcal{A}, P)$ is known as a probability space.

This is the axiomatic definition of probability, which is due to A.N. Kolmogorov. Here are some properties that can be proved using this definition

Theorem 0.1 *On a probability space $(\mathcal{S}, \mathcal{A}, P)$ the following properties are true.*

(a) $P(A^c) = 1 - P(A)$, *for any event* A.

(b) $P(\emptyset) = 0$.

(c) For any events A, B, *if* $A \subseteq B$ *then* $P(A) \leq P(B)$.

(d) $P(A) \leq 1$ *for any event* A.

(e) For any events A, B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(f) For any event $A_1, A_2, \ldots$, $P(\cup_{j=i}^{\infty} A_j) \leq \sum_{j=1}^{\infty} P(A_j)$.

(g) Let $A_1, A_2, \ldots$ *be events such that* $A_n\!\uparrow$ *or* $A_n\!\downarrow$ *as* $n \to \infty$. *Then*

$$P\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} P(A_n).$$

Proof: We prove (a), (c) and part of (g) leaving the rest as exercises.
(a): The trick is to express A and A^c to form a partition. We have $\mathcal{S} = A \cup A^c$ and $A \cap A^c = \emptyset$. Thus from the first and second axioms of probability, i.e. axiom (1) and axiom (2) of Definition 0.1, it follows that

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

and the desired result follows.
(c): Again we need to build a partition. By assumption $A \subseteq B$, thus $B = A \cup (A^c \cap B)$. But $A \cap (A^c \cap B) = \emptyset$, thus by the third axiom of probability

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

since by axiom (1) $P(A^c \cap B) \geq 0$.
(g): Suppose $A_n\!\uparrow$ and let's look again for a partition. After some thoughts we define the sets $B_1 = A_1$ and for $n > 1$, $B_n = A_n \cap A_{n-1}^c$. It follows that $\cup_n A_n = \cup_n D_n$ (proof ?) and

that $B_m \cap B_n = \emptyset$ for all $m \neq n$ (proof ?). Thus the sequence $B_1, B_2, \ldots$ is a sequence of pairwise disjoint sets and it is a nice one since by point (c) of the theorem

$$P(B_n) = P(A_n) - P(A_{n-1}).$$

Applying the third axiom of probability gives

$$\begin{aligned}
P\left(\lim_{n\to\infty} A_n\right) &= P(\cup_{n=1}^{\infty} A_n) = P(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} P(B_n) = P(B_1) + \lim_{n\to\infty} \sum_{j=2}^{n} P(B_j) \\
&= P(A_1) + \lim_{n\to\infty} \{P(A_2) - P(A_1) + P(A_3) - P(A_2) + \ldots + P(A_n) - P(A_{n-1})\} \quad (0.1) \\
&= P(A_1) + \lim_{n\to\infty} \{P(A_n) - P(A_1)\} \quad (0.2) \\
&= \lim_{n\to\infty} P(A_n).
\end{aligned}$$

In passing from equality (0.1) to (0.2) we have used the telescoping property: for any sequence of numbers $a_0, a_1, \ldots, a_n$, $\sum_{k=1}^{n}(a_k - a_{k-1}) = a_n - a_0$. ■

0.3.1 Conditional probability

Assuming that the event B has been observed, i.e. it is a fact that B is realised, and we may wish to calculate the probability of A, taking into account this fact. For instance, consider a fair die with sides ⚀, ⚃ and ⚅ painted red and the remaining sides printed black. The dice is rolled once and we are asked for the probability of that the upward side is ⚅, i.e. $A = \{⚅\}$. Clearly, $P(A) = \frac{1}{6}$. Now suppose that the die is rolled once and we are told that the colour of the upward side is red. The required probability is now $\frac{1}{3}$ because out of the read-coloured faces there is only one ⚅. This latter probability is called conditional probability of A, given the information that the uppermost side was painted red. If we set B for the event "the uppermost side is red", the above-mentioned conditional probability is denoted by $P(A|B)$.

Definition 0.2 *Let B be an event such that $P(B) > 0$. Then the conditional probability given AB is the set function denoted by $P(\cdot|B)$ and defined for every event A by*

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

$P(A|B)$ is called the conditional probability of A given B.

The following theorems state some useful properties related to the conditional probability function.

Theorem 0.2 *(Multiplicative Law) Let $A_1, \ldots, A_n$ be events such that $P(\cap_{j=1}^{n-1} A_j) > 0$. Then*

$$P(\cap_{j=1}^{n} A_j) = P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})P(A_{n-1}|A_1 \cap A_2 \cap \cdots \cap A_{n-2}) \cdots P(A_2|A_1)P(A_1).$$

Theorem 0.3 *(Law of Total Probability) Let $A_1, \ldots, A_n$ be a partition of $\mathcal{S}$, with $P(A_j) > 0$ for all j. Then for an event $B \in \mathcal{A}$, we have*

$$P(B) = \sum_j P(B|A_j)P(A_j).$$

The Bayes Theorem or the Law of Inverse Probability is founded on the Law of Total Probability but it has its own importance especially in Bayesian statistics as we will see in Lecture 7.

Theorem 0.4 *(Bayes Theorem) If $A_1, \ldots, A_n$ is a partition of $\mathcal{S}$ and $P(A_j) > 0$ for all j, then*

$$P(A_j|B) = \tfrac{P(B|A_j)P(A_j)}{\sum_j P(B|A_j)P(A_j)}.$$

The following two simple examples illustrate the properties shown above.

Example 0.4 *An urn contains 10 identical balls, except for the colour, of which five are black, three are red and two are white. Four ball are drawn without replacement. Find the the probability that the first ball is black, the second is red, the third is white and the fourth is black.*

Let A_1 be the event that the first ball is black, A_2 be the event that the second ball is red, A_3 be the event that the third ball is white and A_4 be the event that the fourth ball is black. Then

$$P(A_1 \cap A_2 \cap A_3 \cap A_4) = P(A_4|A_1 \cap A_2 \cap A_3)P(A_3|A_1 \cap A_2)P(A_2|A_1)P(A_1),$$

and using the fact that each of the balls is equally likely to be drawn, we have

$$P(A_1) = \tfrac{5}{10}, \quad P(A_2|A_1) = \tfrac{3}{9}, \quad P(A_3|A_1 \cap A_2) = \tfrac{2}{8}, \quad P(A_4|A_1 \cap \cdots \cap A_1) = \tfrac{4}{7}.$$

Thus the required probability is equal to $\frac{1}{42}$.

Example 0.5 *A multiple choice test question lists five alternative answers, of which only one is correct. If a student has done the homework, then he/she is certain to identify the correct answer; otherwise he/she chooses an answer at random. If A is the event "the student does the homework" and B is the event "the student answers the question correctly". Find the expression of the conditional probability $P(A|B)$ in terms of $p = P(A)$.*

A and A^c form a partition of the sample space, i.e. the student either does or doesn't do the homework, thus the probability that the student does the homework given a correct answer is

$$P(A|B) = \tfrac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)} = \tfrac{1 \cdot p}{1 \cdot p + \frac{1}{5}(1-p)} = \tfrac{5p}{4p+1}.$$

For instance if $p = 1/2$, i.e. the chance that the students does the homework is one half, given that he guessed correctly the answer to a multiple choice question with five alternatives, the chances that he has done the homework is now 0.83. As we will see in Lecture 7, in Bayesian statistics p is also called the prior probability and $P(A|B)$ is called the posterior probability.

0.3.2 Independence

For any event A and B with $P(B) > 0$ we defined $P(A|B) = P(A \cap B)/P(B)$. Now since $P(A|B)$ is a real number in the interval $[0, 1]$, by the trichotomy law of real numbers only one $P(A|B) < P(A), P(A|B) > P(A)$ and $P(A|B) = P(A)$ holds. As an illustration consider an urn containing 10 balls, seven of which are red and 3 are black. Except for the colour, the balls are identical. Suppose that two balls are drawn successively and without replacement. Then the conditional probability that the second ball is red, given that the first ball was red is $\frac{6}{9}$, whereas the conditional probability that the second ball is red given that the first ball was black is $\frac{7}{9}$. Without any knowledge on the first ball, the probability that the second ball is red equals (by the Low of Total Probability) $\frac{7}{10}\frac{6}{9} + \frac{3}{10}\frac{7}{9} = \frac{7}{10}$. On the other hand, if the balls are drawn with replacement, the probability that the second ball is red, given that the first ball was red is $\frac{7}{10}$. This probability is the same even if the the first ball was black. In other words, knowledge of the event which occurred in the first drawing provides no additional information in the calculation of the event that the second ball is red. These latter events are called *independent*.

Definition 0.3 *Events A and B are independent if $P(A \cap B) = P(A)P(B)$.*

Another alternative definition of independence is $P(A|B) = P(A)$, provided $P(B) > 0$. For more than two events we need stronger conditions, stated in the next definition.

Definition 0.4 *Events $A_1, \ldots, A_n$ are independent or completely or independent if*

$$P(A_{j_1} \cap \cdots \cap A_{j_k}) = P(A_{j_1}) \cdots P(A_{j_k}),$$

for $j_1, \ldots, j_k = 1, 2, \ldots, n$ and $k = 2, \ldots, n$ such that $1 \le j_1 < j_2 < \cdots < j_k \le n$. The events are said to be pairwise independent if $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j = 1, 2, \ldots, n$.

The following example that illustrates the concept.

Example 0.6 *Let $\mathcal{S} = \{1, 2, 3, 4\}$ and assume that $P(\{1\}) = \cdots = P(\{4\}) = \frac{1}{4}$. Let $A_1 = \{1, 2\}$, $A_2 = \{1, 3\}$, $A_3 = \{1, 4\}$. Then*

$$A_1 \cap A_2 = A_1 \cap A_3 = A_2 \cap A_3 = \{1\}, \quad \textit{and} \quad A_1 \cap A_2 \cap A_3 = \{1\}.$$

Thus $P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = P(A_1 \cap A_2 \cap A_3) = \frac{1}{4}$. Now,

$$\begin{aligned} P(A_1 \cap A_2) &= \tfrac{1}{4} = \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_2), \\ P(A_1 \cap A_3) &= \tfrac{1}{4} = \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_3), \\ P(A_2 \cap A_3) &= \tfrac{1}{4} = \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_2)P(A_3), \end{aligned}$$

but

$$P(A_1 \cap A_2 \cap A_3) = \tfrac{1}{4} \neq \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_2)P(A_3).$$

Thus, A_1, A_2, A_3 are pairwise independent but not completely independent. As another example, let $\mathcal{S} = \{1, 2, 3, 4, 5\}$ and define P as follows: $P(\{1\}) = \frac{1}{8}$, $P(\{2\}) = P(\{3\}) = P(\{4\}) = \frac{3}{16}$, $P(\{5\}) = \frac{5}{16}$. Consider

$$A_1 = \{1, 2, 3\}, \quad A_2 = \{1, 2, 4\}, \quad A_3 = \{1, 3, 4\}.$$

Then

$$A_1 \cap A_2 = \{1, 2\}, \quad A_1 \cap A_2 \cap A_3 = \{1\}.$$

Thus

$$P(A_1 \cap A_2 \cap A_3) = \tfrac{1}{8} = \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_2)P(A_3),$$

but

$$P(A_1 \cap A_2) = \tfrac{5}{16} \neq \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_2).$$

Events that are not independent are called *dependent.* Although here we dealt with events defined on discrete sample spaces, events can be defined on continuous sample spaces as well, e.g. spaces $\mathcal{S} = \mathbb{R}$. In this case, the σ-field of subspaces of $\mathcal{S}$ is the Borell σ-field, which is treated in advanced probability theory courses and is too much technical for our purposes. For us it is enough to known that whatever is the sample space $\mathcal{S}$, we can always attach to it a suitable σ-filed and thus build a suitable probability space $(\mathcal{S}, \mathcal{A}, P)$.

0.4 Random variables and their distributions

In a probability space $(\mathcal{S}, \mathcal{A}, P)$, the sample space $\mathcal{S}$ may be quite an abstract set for use in real-life applications. Indeed, in practical applications we typically measure quantities such as temperature, wind speed, precipitation, electricity consumption, life-time of an electronic circuit, etc. and we'd like to be able to characterise these measurements by means of probability theory. For instance, we might want to calculate the probability that the life-time of a certain electronic circuit is at least t, with $t > 0$. This can be done by a suitable transformation or mapping of events of $\mathcal{S}$ to real values, called *random variable.*

Definition 0.5 *A random variable (r.v.) X is a function*[2] $X(s) : \mathcal{S} \to \mathbb{R}$*. For a subset $B \subseteq \mathbb{R}$, we write*

$$\{X \in B\} = \{s \in \mathcal{S} : X(s) \in B\}.$$

Mapping $s \in \mathcal{S}$ to say $t \in \mathbb{R}$ by $t = X(s)$ induces a probability for t and to distinguish it from P, we call the probability distribution of all possible values of X, P_X. This probability

[2] Technically, we require this function to be a *measurable* mapping between the two sets. However, such a technicality is not relevant to us since all r.v. that we will meet in this course are measurable.

distribution is defined as follows. Let $B \subseteq \mathbb{R}$, then

$$P_X(B) = P(\{X \in B\}) = P(\{s \in \mathcal{S} : X(s) \in B\}).$$

We also use the shorthand notation $P(X \in B)$ for $P(\{X \in B\})$. To fix the idea consider the following simple example.

Example 0.7 *Consider throwing a regular die with* $\mathcal{S} = \{\text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅}\}$ *once. You win* $-1\$$ *(i.e. you loose 1\$) if the it realises* ⚀, ⚁ *or* ⚂, *you win* $0\$$ *if it realises* ⚃ *and you win* $1\$$ *if it realises* ⚄ *or* ⚅. *Let* X *be the r.v. "amount of win" in dollars. Thus* X *takes values in* $\{-1, 0, 1\}$ *and is defined by*

$$X(s) = \begin{cases} -1 & \textit{if } s \in \{\text{⚀}, \text{⚁}, \text{⚂}\} \\ 0 & \textit{if } s \in \{\text{⚂}\} \\ 1 & \textit{otherwise.} \end{cases}$$

and its probability distribution is

$$P_X(x)^3 = \begin{cases} 1/2 & \textit{if } x = -1 \\ 1/6 & \textit{if } x = 0 \\ 1/3 & \textit{if } x = 1. \end{cases}$$

(check!).

A r.v. is called *discrete* if there are countable points $x_1, x_2, \ldots,$ such that $P_X(x_j) > 0, j \geq 1$ and $\sum_j P_X(x_j) = \sum_j P(X = x_j) = 1$. Then the function f_X defined on $\mathbb{R}$ by the relationship

$$f_X(x) = \begin{cases} P_X(x) = P(X = x) & \text{if} \quad x = x_j \\ 0 & \text{otherwise} \end{cases},$$

has the properties $f_X(x) \geq 0$ for all x, and $\sum_j f_X(x_j) = 1$. Furthermore, for any event B, $P(X \in B) = \sum_{x_j \in B} f_X(x_j)$. The function $f_X(x)$ is called the *probability density function* (p.d.f.) of X.

[3] The precise but more pedantic notation is $P_X(\{x\})$.

Suppose now that X is a r.v. which takes values in $\mathbb{R}$ or a subset of it and $P(X = x) = 0$ for every $x \in \mathbb{R}$. Then X is called a *continuous* r.v.. All continuous r.v.s that we will meet in this course possess a function f_X such that $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and $P(X \in B) = \int_B f_X(x)\,dx$, for any interval B of $\mathbb{R}$. The function f_X is called again probability density function of X; though in this case, $f_X(x)$ is different from $P(X = x)$, which is always zero, but still $\int_{-\infty}^{\infty} f_X(x) = 1$.

As a rule of thumb, $f : \mathbb{R} \to \mathbb{R}$ is a probability density function iff

(i) $f(x) \geq 0$, for all $x \in \mathbb{R}$

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

Remark 0.1 *At this point it might seem that the triple $(\mathcal{S}, \mathcal{A}, P)$ is superfluous once we know f or the distribution function of F (see the next section). At some degree this is true. Indeed, once we know f or F we know everything about the r.v. X and we can always attach to a given X a probability space $(\mathcal{S}, \mathcal{A}, P)$ if needed. The triple $(\mathcal{S}, \mathcal{A}, P)$ comes into play when we wish to deal with certain theoretical aspects of X.*

0.4.1 The distribution function

If we consider subsets B of $\mathbb{R}$ that are intervals closed on the right, i.e. $B = \{y \in \mathbb{R} : y \leq x\}$ for a given x, then $P_X(B)$ is denoted by $F_X(x)$ and is called the *cumulative distribution function* of X or just *distribution function* (d.f.) of X. To ease notation, we will omit the subscript from F_X or f_X when no confusion arises. Thus F, is an ordinary point function with values in [0,1], i.e. $F : \mathbb{R} \to [0, 1]$.

Remark 0.2 *(i) $F(x)$ can be used to find probabilities, such as $P(a < X \leq b)$, for $a < b$,*

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

Indeed, $\{-\infty < X \leq b\} = \{-\infty < X \leq a\} \cup \{a < X \leq b\}$, thus

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

(ii) If X is discrete, its d.f. is a step function and is defined by

$$F(x) = \sum_{x_j \leq x} f(x_j), \quad \text{and} \quad f(x_j) = F(x_j) - F(x_{j-1}),$$

assuming $x_1 \leq x_2 \leq \cdots$.

(iii) With F in our toolbox, the concept of a continuous random variable can be rephrased as follows:
X is a continuous r.v. if its distribution function $F(x)$ is continuous at each $x \in \mathbb{R}$. If, in addition, $F(x) = \int_{-\infty}^{x} f(t)dt$, then X is absolutely continuous and by the Fundamental Theorem of Calculus

$$\tfrac{dF(x)}{dx} = f(x), \quad \text{at all continuity points of } f.$$

Here is an Example of a discrete r.v..

Example 0.8 *Consider the r.v. in Example 0.7. Then the d.f. of X is*

$$F(x) = \begin{cases} 0 & \text{if} \quad x < -1 \\ \frac{3}{6} & \text{if} \quad -1 \leq x < 0 \\ \frac{4}{6} & \text{if} \quad 0 \leq x < 1 \\ \frac{6}{6} & \text{if} \quad x \geq 1. \end{cases}$$

Similarly, we find that $P(X = 0) = \frac{1}{6}$ and $P(X = 1) = \frac{2}{6}$. The d.f. and the density function are depicted in Figure 0.1.

Here is another example, this time involving a continuous r.v..

Example 0.9 *Let X be a real number chosen at random from the interval $(-1, 1)$. For example, we could happen to see $X = 1/2$ or $X = -1/5$ or any other value, say, $x \in (-1, 1)$. Values of $x < -1$ are impossible so $P(X \leq x) = 0$; values $x \leq 1$ will appear with certainty, thus $P(X \leq x) = 1$.*

Now let $x \in [-1, 1)$. Since we are drawing at random, it is reasonable to expect that the "likelihood" of drawing a value from $[-1, x)$ must be relative or inversely proportional to the

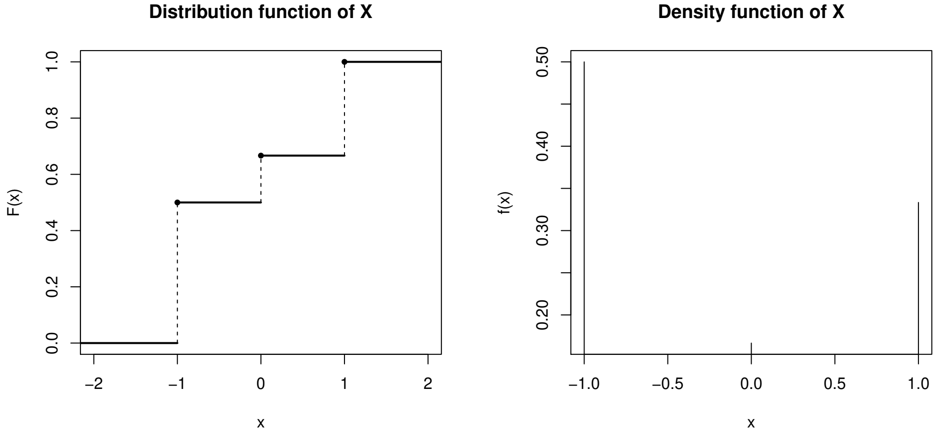

Figure 0.1: Example 0.8. Left: the distribution function of X, the filled dots suggest that the value is included; right: the probability density function.

likelihood of drawing from the whole interval $[-1, 1)$*. If we let this likelihood be represented by the length of the interval we arrive at the result*

$$P(X \le x) = \frac{\mathit{length}((-1,x))}{\mathit{length}((-1,1))} = \frac{x+1}{2}, \quad \mathit{for}\ x \in [-1, 1).$$

We have thus built the d.f. of X*, which is (see Fig. 0.2)*

$$F(x) = \begin{cases} 0 & \mathit{if} \quad x < -1 \\ \frac{x+1}{2} & \mathit{if} \quad -1 \le x < 1 \\ 1 & \mathit{if} \quad x \ge 1. \end{cases}$$

Following Remark 0.2(iii), we see that $F(x) = \int_{-\infty}^{x} f(t)dt$ *holds with*

$$f(x) = \begin{cases} 0 & \mathit{if}\, x \le -1 \\ 1/2 & \mathit{if}\, -1 < x < 1 \\ 0 & \mathit{otherwise}. \end{cases}$$

Furthermore, $f(x) = \frac{dF(x)}{dx}$*, depicted in the left plot of Figure 0.2.*

Any r.v. whose p.d.f is constant on its domain is called <u>uniform</u>. For instance, the distribution

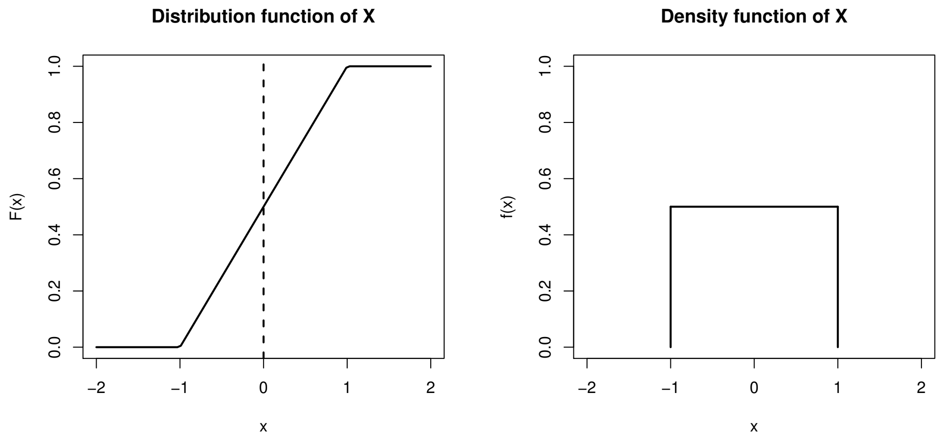

Figure 0.2: Example 0.9. Left: the distribution function of X; right: the probability density function.

in Example 0.9 is uniform in (-1,1). The following theorem states four fundamental properties of the distribution function.

Theorem 0.5 *For the distribution function F of a r.v. X, the following holds.*

(i) *F is nondecreasing;*

(ii) *F is continuous from the right;*

(iii) $\lim_{x\to-\infty} F(x) = 0$*;*

(iv) $\lim_{x\to\infty} F(x) = 1$.

Proof: We prove (i) and (iii).
(i): The non decreasing property means that for all reals a and b, if $a < b$, then $F(a) \leq F(b)$. Consider the sets $A = \{s : X(s) \leq a\}$ and $B = \{s : X(s) \leq b\}$ and note that $P(A) = P(X \leq a) = F(a)$ and similarly for B (see Remark 0.1). We see that $A \subseteq B$ since $s \in A$ implies that $X(s) \leq a < b$, thus s is also in B. By Theorem 0.1(c), $P(A) \leq P(B)$ and the result follows.
(iii): Let $x_1, x_2, \ldots$ be a sequence of real numbers decreasing to $-\infty$ and let $A_1, A_2, \ldots$ be a sequence of sets with $A_k = \{s : X(s) \leq x_n\}$. Then $A_k\downarrow$ (check!) and $\cap_{k=1}^{\infty} A_k = \emptyset$. Indeed, let $s \in \cap_{k=1}^{\infty} A_k$ then this would mean that there is a real number $X(s) \leq x_n$ for every x_n.

Stated differently, there is a lower bound to the unbounded from below sequence $x_1, x_2, \ldots$, which is a contradiction. Such an s cannot exist, thus $\cap_{k=1}^{\infty} A_k$ is empty. By Theorem 0.1(g),

$$0 = P(\emptyset) = P(\cap_{k=1}^{\infty} A_k) = \lim_{k\to\infty} P(A_k) = \lim_{x_k\to-\infty} P(X \le x_k) = \lim_{x_n\to-\infty} F(x_n).$$

■

0.4.2 The quantile function

Let X be a r.v. with d.f. F and consider p such that $0 < p < 1$. A pth *quantile* of X is a number ξ_p with the property:

$$P(X < \xi_p) \le p \quad \text{and} \quad F(\xi_p) \ge p.$$

Some quantiles or functions thereof have special names: $\xi_{0.25}$ is called the *first quartile*, $\xi_{0.5}$ is the *median* (or second quartile), $\xi_{0.75}$ is the *third quartile* and $iqr = \xi_{0.75} - \xi_{0.25}$ is called interquartile range. Median and iqr are often used as a measures of center and dispression, respectively, of the distribution of X.

Quantiles are unique provided X is absolutely continuous (Remark 0.2(iii)). In this case, the pth quantile is

$$\xi_p = F^{-1}(p).$$

The function $F^{-1}(p) : (0, 1) \to \mathbb{R}$ is the inverse of F and is called the *quantile function*. The notation $Q(p)$ is also used for denoting the quantile function.

If X is absolutely continuous, then F is strictly increasing, and $F(F^{-1}(p)) = p$ and $F^{-1}(F(x)) = x$ for all $p \in (0, 1)$ and $x \in \mathbb{R}$.

When the r.v. is not absolutely continuous, then F has jumps and it may be piecewise constant, thus the equation $F(x) = p$, for a given p, may have no solutions, exactly one solution or infinitely many solutions; thus F may not be invertible. The following general definition for the quantile function is thus preferred

$$Q(p) = \inf\{x : F(x) \ge p\}, \quad p \in (0, 1),$$

where the infimum is taken over all x in the domain of X. The quantile function can be defined also on the boundary points of the domain, and the definition depends on the range of X. In particular, if the lower limit of the range of X is $-\infty$, then by definition $Q(0) = -\infty$; if the upper limit of the range of X is ∞, then by definition $Q(1) = \infty$; see the example below for an X with finite range.

Example 0.10 *Consider the r.v. in Example 0.8. We find that (see Figure 0.1)* $\xi_{0.25} = \xi_{0.5} = -1$ *and* $\xi_{0.8} = 1$*. Regarding* $Q(p)$ *we have,* $Q(0) = -\inf Q(p) = -1$ *for* $p \in (0, 1/2]$, $Q(p) = 0$ *for* $p \in (1/2, 2/3]$ *and* $Q(p) = 1$ *for* $p \in (2/3, 1]$*. It turns out that* $Q(0) = 0$ *and* $Q(1) = 2$.

0.4.3 Transformations

We might want to transform the values of X to $Y = g(X)$, by some suitable function $g : \mathbb{R} \to \mathbb{R}$. For instance, suppose we are measuring the load of a certain virus in a patient's blood. Instead of reporting x, the number of viruses counted in a certain amount of blood, we may wish to report its base 10 logarithm, $y = \log_{10}(x)$. The next theorem shows how to get f_Y and F_Y from F_X under some conditions on g.

Theorem 0.6 *Let* X *be a continuous r.v. with d.f.* F_X *and consider the bijective differentiable function* $g(x) : \mathbb{R} \to \mathbb{R}$*, with inverse is* $g^{-1}(y)$*. Then* $Y = g(X)$ *is a r.v. with d.f.* F_Y *and density function* f_Y *given respectively by*

$$F_Y(y) = F_X(g^{-1}(y)), \quad \textit{and} \quad f_Y(y) = \tfrac{dF_X(g^{-1}(y))}{dy} = f_X(g^{-1}(y)) \left| \tfrac{dg^{-1}(y)}{dy} \right|.$$

If g is only surjective, i.e. there may be more than one $x \in \mathbb{R}$ such that $g(x) = y$, then the $F_Y(y) = F_X(g^{-1}(y))$ is still valid, whereas the computation of f_Y is more complicated.

0.4.4 Moments

Let X be a continuous r.v. with density function f; if X is discrete replace all the integrals below by sums. For a suitable function $g : \mathbb{R} \to \mathbb{R}$ we define the expectation of $g(X)$ by

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

provided $\int_{-\infty}^{\infty} |g(x)| f(x) dx$ exists and is finite.

Noteworthy examples of $E(g(X))$ are:

- $g(x) = x$, the identity function, leads to $E(X) = \mu_X$[4], called the *expectation* of X;
- $g(x) = (x - c)^n$ for some constant c and a positive integer n, leads to the nth *moment of X about c*
$$E[(X - c)^n] = \int_{-\infty}^{\infty} (x - c)^n f(x) dx,$$
provided the integral exists.
- $g(x) = (x - E(X))^n$, leads to the so-called $n$the *central moments*; the 2nd central moment of X, is also called the *variance* of X, denoted by $\mathrm{var}(X) = \sigma_X^2$.

Example 0.11 *Consider the r.v. in Example 0.8. We find that* $E(X) = -1 \cdot \frac{3}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{2}{6} = -\frac{1}{6}$.

Here are some remarkable properties about moments.

Theorem 0.7 *Let $X, X_1, X_2, \ldots, X_n$ be r.v.s with finite expectations and finite variances. Then*

(i) $E(a) = a$ *and* $\mathrm{var}(a) = 0$*, for all* $a \in \mathbb{R}$.

(ii) $E(a + bX) = a + bE(X)$ *and* $\mathrm{var}(a + bX) = b^2 \mathrm{var}(X)$*, for all* $a, b \in \mathbb{R}$.

(iii) $E(\sum_{j=1}^{n} c_j X_j) = \sum_{j=1}^{n} c_j E(X_j)$*, for all* $c_1, \ldots, c_n \in \mathbb{R}$.

(iv) $\mathrm{var}(X) = E(X^2) - [E(X)]^2$.

(iv) If $X \geq 0$ *then* $E(X) \geq 0$*; more generally, if* $X \geq X_1$ *then* $E(X) \geq E(X_1)$.

(v) $|E(X)| \leq E(|X|)$.

(vi) For any $k > 0$*, then* $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$*, where* $\mu = E(X), \sigma^2 = \mathrm{var}(X)$*; this is known as the Chebyshev inequality.*

[4] We will drop the subscript when no confusion arises.

(vii) If t is a convex function[5] *on an open interval of the real numbers, X has support on this interval and has finite expectation, then $t(E(X)) \leq E(t(X))$*

Proof: We assume X is continuous and prove (i), (iv).
(i): $E(a) = \int_{-\infty}^{\infty} af(x)dx = a\int_{-\infty}^{\infty} f(x)dx = a$, since the p.d.f. over its domain integrates to one by definition.
(iv): $f(x) \geq 0$ for all x, thus $xf(x) \geq 0$ and by the properties of the integral $E(X) = \int_{-\infty}^{\infty} xf(x)dx \geq 0$.
■

0.5 Examples of discrete random variables

0.5.1 The Binomial distribution

A *Bernoullian experiment* is random experiment with only two possible outcomes: either the event A realises or it does not, i.e. A^c realises. Examples are: the toss of a coin where it could either realise heads (A) or tails (A^c), the roll of a die where numbers are partitioned in even (A) and odd (A^c), etc. We classify the event of interest to us as the "success" (say A) and its complement is called "failure" (A^c). The *Bernoulli r.v.* is defined by

$$X(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \in A^c \end{cases}.$$

Since A is an event pertaining to some probability space $(\mathcal{S}, \mathcal{A}, \mathcal{P})$ it must have a probability. Let thus $\theta = P(A)$, where $\theta \in [0, 1]$. Then the p.d.f. of the Bernoulli r.v. is defined by

$$P(X = x) = \theta^x(1 - \theta)^{1-x}.$$

We see that $E(X) = \sum_x xP(X = x) = 0 \cdot (1 - \theta) + 1 \cdot \theta = \theta = P(A)$. Thus the expectation of a Bernoulli r.v. equals its success probability. Similarly we find $\text{var}(X) = \theta(1 - \theta)$. We use the notation $X \sim \text{Ber}(\theta)$, to mean that the r.v. X has a Bernoulli distribution with a single parameter θ, called *success probability*.

[5] Check your calculus textbook for the definition of a convex function.

Now suppose we toss n identical coins, each having probability θ of heads (H) and we ask about the probability of observing 2 heads out of n. If we let E be the event "2 events out of n are head" and denote tails by T, then E realises if

$$E = \{\underbrace{HHT\cdots T}_{n}, \underbrace{HTHT\cdots T}_{n}, \underbrace{THHT\cdots T}_{n}, \ldots\},$$

Each sample point $s \in E$ involves 2 H's and $n-2$ T's and because the tosses are independent, the probability of say $s = HHT\cdots T$ is

$$P(\{s\}) = P(\{HHT\cdots T\}) = P(H)P(H)P(T)\cdots P(T) = \theta^2(1-\theta)^{n-2},$$

It is clear that for any $s \in E$, $P(\{s\}) = \theta^2(1-\theta)^{n-2}$. Note also that, elements of E are sample points, thus only one can be observed. This means that

$$\begin{aligned} P(E) &= P(\{s_1\}) + P(\{s_2\}) + \cdots + P(\{s_{C_{2,n}}\}) \\ &= C_{2,n}P(\{s_1\}) \\ &= C_{2,n}\theta^2(1-\theta)^{n-2}, \end{aligned}$$

.

where $C_{2,n} = \binom{n}{2} = \frac{n!}{2!(n-2)!}$ is the number of possible orderings of n elements in which there two identical F's and $n-2$ identical T's.

More generally, the probability of getting $y \leq n$ heads in n tosses is given by

$$P(Y = y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y \in \{0, 1, 2, \ldots, n\}.$$

The r.v. Y with the above probability distribution is called *binomial r.v.* and is denoted by $Y \sim \text{Bin}(n, \theta)$. The Binomial distribution thus also has a single parameter, θ, which has the same interpretation as in the Bernoulli distribution and n is called *index.* The binomial r.v. can also be obtained as the sum of n i.i.d Bernoulli r.v.. Indeed, if $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, for $i = 1, \ldots, n$, then for the r.v. $Y = \sum_{i=1}^n X_i$, it follows that $Y \sim \text{Bin}(n, \theta)$. Using this last fact we have that

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n\theta\,, \quad \mathrm{var}(Y) = n\theta(1-\theta).$$

In calculating the mean and average of the binomial r.v. we used the linearity property of the expectation (L0, Theorem 0.9) and the fact that this r.v. arises as the sum of n i.i.d Bernoulli r.v. with parameter θ.

0.5.2 The Negative binomial distribution

Consider a sequence of Bernoullian experiments each with probability of success equal to θ. Instead of counting the number of successes in a fixed number of trials, we let Y to be the number of failures observed until a fixed number of success $r \in \mathbb{N}$ are obtained. Then Y is an r.v. and is called *negative binomial* r.v., denoted $Y \sim \mathrm{NegBin}(r, \theta)$. It has p.d.f. equal to

$$P(Y = y) = \binom{y + r - 1}{y} \theta^r (1-\theta)^y, \quad y \in \mathbb{Z}_{\geq 0}.$$

where $\mathbb{Z}_{\geq 0} = \{0, 1, 2, \ldots, \}$. If $Y \sim \mathrm{NegBin}(1, \theta)$, then Y is also called *geometric r.v.*. Here the number of parameters are r, called *index* and θ the success probability.

For this r.v. it holds that, if $Y_1 \sim \mathrm{NegBin}(n_1, \theta), Y_2 \sim \mathrm{NegBin}(n_2, \theta)$ and $Y = Y_1 + Y_2$, then $Y \sim \mathrm{NegBin}(r, \theta), r = n_1 + n_1$. It holds that $E(Y) = r(1-\theta)/\theta$ and $\mathrm{var}(Y) = r(1-\theta)/\theta^2$.

Remark 0.3 *In the definition of the* $\mathrm{NegBin}(r, \theta)$ *r.v., instead of counting the number of failures, one could take the number of trials performed in order to obtain r success. The latter counting approach leads to an alternative definition of the* $\mathrm{NegBin}(r, \theta)$*, in which the range of the r.v. is $\{r, r+1, r+2, \ldots\}$. However, we prefer the latter definition since it is more useful from a statistical modelling perspective.*

0.5.3 The Possion distribution

A r.v. Y is defined to be Poisson with parameter $\lambda \in \mathbb{R}_{>0}$, written $Y \sim \text{Poi}(\lambda)$, if its p.d.f. is

$$P(Y = y) = \tfrac{e^{-\lambda}\lambda^y}{y!}, \quad y \in \mathbb{Z}_{\geq 0}.$$

The Poisson distribution has parameter λ, which is called *rate* or *mean* (see below why). This distribution is appropriate for modelling the number of phone calls arriving at a given telephone exchange within a certain period of time, the number of particles emitted by a radioactive source within a certain period of time, etc.

For a $Y \sim \text{Poi}(\lambda)$ r.v. it holds that $E(Y) = \text{var}(Y) = \lambda$.

0.6 Examples of continuous random variables

0.6.1 The normal (Gaussian) distribution

A r.v. Y is said to have *normal distribution* if its p.d.f. is

$$f(y) = \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad y \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0,$$

We denote this by $Y \sim \text{N}(\mu, \sigma^2)$, where the mean $\mu = E(Y)$ and the variance $\sigma^2 = \text{var}(Y)$ are the parameters of the distribution. For $\mu = 0$ and $\sigma^2 = 1$, the distribution is known as the *standard normal distribution*, denoted by $\text{N}(0, 1)$.

The normal distribution has bell-like shape, and is symmetric around μ. It is a good approximation to the distribution of grades, heights or weights of a large group of individuals, the diameters of hail hitting the ground during a storm, errors in numerous measurements, etc. However, its main significance drives from the Central Limit Theorem (to be seen in L1). If Y is a standard normal r.v., then its d.f. is

$$F(y) = \int_{-\infty}^{y} \tfrac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \Phi(y).$$

Many other distributions are based on the normal r.v. as we will see below. Here are some useful properties about this distribution.

Theorem 0.8 *(i) If $Y \sim \mathrm{N}(\mu, \sigma^2)$ and $T = a + bY$, then $T \sim \mathrm{N}(a + b\mu, b^2\sigma^2)$.*

(ii) If $Y \sim \mathrm{N}(\mu, \sigma^2)$ and $Z = \frac{Y-\mu}{\sigma}$, then $Z \sim \mathrm{N}(0, 1)$; that is, a standardised normal r.v. has standard normal distribution.

0.6.2 The exponential distribution

A r.v. Y with p.d.f. given by

$$f(y) = \lambda e^{-\lambda y}, \quad \text{for all } y > 0, \quad \lambda > 0,$$

is called *exponential distribution.* We denote this by $Y \sim \mathrm{Exp}(\lambda)$, where λ is the parameter of the distribution called *scale.* The exponential r.v. occurs frequently in waiting time problems.

The exponential distribution is "memoryless". In an application in which the event is failure of an electronic component of an equipment, this means that how long the electronic component will be working does not depend on how long it has been working already.

$$P(Y > r + y | Y > r) = \frac{P(Y>r+y)}{P(Y>r)} = \frac{e^{-\lambda(r+y)}}{e^{-\lambda r}} = e^{-\lambda y},$$

which is the same as $P(Y > y)$. The future looks the same no matter when you start watching the process, and no matter hwo much you may have learned about the previous failures.

For $Y \sim \mathrm{Exp}(\lambda)$ it holds that $E(Y) = \frac{1}{\lambda}$ and $\mathrm{var}(Y) = \frac{1}{\lambda^2}$.

0.6.3 The gamma distribution

A r.v. Y with p.d.f.

$$f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}, \quad y > 0, \quad \alpha > 0, \quad \lambda > 0,$$

is called *gamma distribution*, where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-\lambda y}dy$ is the *gamma function*. We denote this by $Y \sim \mathrm{Ga}(\alpha, \lambda)$, where α, λ are the parameters of the distribution, α is called the *shape parameter* and λ is called the *scale*. For this r.v., $E(Y) = \frac{\alpha}{\lambda}, \mathrm{var}(Y) = \frac{\alpha}{\lambda^2}$.

Remark 0.4 *When working with the gamma distribution the following properties of the gamma function are worth remembering.*

(i) $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, $\alpha > 1$

(ii) $\Gamma(1) = \int_0^\infty e^{-y}dy = 1$

(iii) *For* $n \in \mathbb{N}$, $\Gamma(n) = (n-1)(n-1)\cdots 1 = (n-1)!$.

(iv) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ *and* $\Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}$.

Here are some useful properties of the gamma distribution

Theorem 0.9 *(i) If* $Y \sim \mathrm{Ga}(\alpha, \lambda)$ *then* $kY \sim \mathrm{Ga}(\alpha, \frac{\lambda}{k})$, $k > 0$.

(ii) If $Y_1 \sim \mathrm{Ga}(\alpha_1, \lambda)$ *and* $Y_2 \sim \mathrm{Ga}(\alpha_2, \lambda)$, *with* Y_1 *independent from* Y_2 *and* $Y = Y_1 + Y_2$, *then*

$$Y \sim \mathrm{Ga}(\alpha_1 + \alpha_2, \lambda).$$

0.6.4 The Weibull distribution

A Weibull r.v. results from a power of an exponential r.v.. Suppose $X \sim \mathrm{Exp}(1/\beta)$, so that $E(X) = \beta$ and let $Y = X^{1/\alpha}$ for $\alpha > 0$. Then the r.v. Y is said to have Weibull distribution and has d.f.

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X^{1/\alpha} \le y) \\ &= P(X \le y^\alpha) \\ &= F_X(y^\alpha) \\ &= 1 - e^{-\frac{y^\alpha}{\beta}}. \end{aligned}$$

Differentiating the d.f. we get the p.d.f

$$f(y) = \frac{\alpha}{\beta} y^{\alpha-1} e^{-\frac{y^\alpha}{\beta}}, \quad y > 0.$$

We denote it by $Y \sim \mathrm{Wei}(\alpha, \beta)$ where α (the shape) and β (the rate) are the parameters of the distribution. The Weibull distribution is extensively used for modelling life time data as an alternative to exponential and gamma distributions.

0.6.5 The chi-square distribution

The r.v. $Y \sim \mathrm{Ga}(\nu/2, 1/2)$ with $\nu \in \mathbb{N}$ is known as the χ^2 (chi-square) distribution, denoted χ^2_ν (or $Y \sim \chi^2_\nu$ if we wish to express that the r.v. Y follows a chi-square distribution) where the parameter ν is called *degrees of freedom*. For this r.v., $E(Y) = \nu, \mathrm{var}(Y) = 2\nu$. The χ^2_ν distribution is nothing more than a special case of the gamma and like the latter it has many interesting properties. Here are some of them

Theorem 0.10 *(i) If $X \sim \mathrm{N}(0,1)$ and $Y = X^2$ then $Y \sim \chi^2_1$.*

(ii) If $X_1, \ldots, X_n$ are indepenent with $X_i \sim \mathrm{N}(0,1)$ and $Y = \sum_{i=1}^n X_i^2$, then $Y \sim \chi^2_n$.

As we will see in the incoming lectures, the chi-square distribution arises as sampling distribution to many functions of r.v. used in inferential statistics.

0.6.6 The t-Student distribution

Let $Z \sim \mathrm{N}(0,1)$ and $U \sim \chi^2_\nu$, where Z and U are independent, then the r.v. $T = \frac{Z}{\sqrt{\frac{U}{\nu}}}$ is called t-Student r.v. with degrees of freedom $\nu \in \mathbb{N}$. This distribution is denoted by t_ν, has domain equal to $\mathbb{R}$ and it holds that $E(T) = 0$ whenever $\nu > 1$ and $\mathrm{var}(T) = \frac{\nu}{\nu-2}$, whenever $\nu > 2$.

The t-Student distribution is symmetric around 0 and has a bell-like shape, just like the normal distribution. With respect to the latter, the t-Student has heavier tails, at least for low values of ν. It can be shown that $t \xrightarrow{d}$ approaches $\mathrm{N}(0,1)$ as $\nu \to \infty$.

Also the t-Student distribution arises as sampling distribution in inferential statistics, but it is also used as a robust statistical model since it can accommodate extremely low and extremely large observations which under the normal model are unlikely.

0.6.7 The F distribution

Let $U_1 \sim \chi^2_{\nu_1}$ and $U_2 \sim \chi^2_{\nu_2}$ with U_1 independent from U_2. Then the r.v.

$$Y = \tfrac{U_1/\nu_1}{U_2/\nu_2},$$

is said to have F distribution with $\nu_1, \nu_2 \in \mathbb{N}$ *degrees of freedom*, denoted $Y \sim F_{\nu_1,\nu_2}$

Here are some useful properties.

Theorem 0.11 *(i) If $Y \sim F_{n_1,n_2}$, then $\frac{1}{Y} \sim F_{n_2,n_1}$*

(ii) If $T \sim t_\nu$, then $T^2 = \frac{Z^2}{\frac{U}{\nu}} \sim F_{1,\nu}$

Also the F distribution arises as sampling distribution of functions of r.v. used in inferential statistics.

0.7 Generating random variates

So far we have been concerned with mathematical properties of r.v. and their probability distributions. However, the study of mathematical properties per se may not be fully satisfactory for at least two reasons. Firstly, there are many r.v.'s which have p.d.f.'s or d.f.'s with complicated expressions, and sometimes are not even available. In this case, mathematical properties such as expectations are typically impossible to study analytically. Secondly, we may want to check how good is a given probability model as an approximation of a given set of observed experimental data. In both cases, we can still achieve our goal if we could generate random draws from the given distribution.

There are many methods for generating random draws from a given distribution function, here we limit ourselves to the *inverse transform sampling* method. Before explaining the method, we need to introduce another continuous r.v., the *uniform distribution.* A r.v. X is said to have the uniform distribution in the interval $[0, 1]$ if it has density

$$f(x) = 1_{[0,1]}(x), \quad x \in [0, 1]$$

where $1_{[a,b]}(x)$ is the indicator function which takes 1 if $x \in [a, b]$ and 0 otherwise. We denote this by $X \sim \text{Unif}(0, 1)$, and we say that the r.v. X is uniform between 0 and 1. Although it is not of direct interest as an approximating model for observed data, the uniform distribution is of vital importance in statistics and probability since it is the heart of many algorithms for simulating random variables. In fact, the uniform distribution is at the heart of the inverse transform sampling method.

Theorem 0.12 *Let X be a r.v. with d.f. F, assumed to be continuous with quantile function F^{-1}, and let $U \sim \text{Unif}(0, 1)$. Define the r.v. $Y = F^{-1}(U)$, Then for any $x \in \mathbb{R}$, we have*

$$\begin{aligned} F_Y(x) &= P(Y \leq x) = P(F^{-1}(U) \leq x) \\ &= P(U \leq F(x)) \\ &= \int_0^{F(x)} 1_{[0,1]}(u) du \\ &= F(x). \end{aligned}$$

Where the last equality holds because $F(x) \in [0, 1]$.

This theorem is valid for any cumulative distribution function, and tells us that if we want to generate random variates X, with d.f. F, then it is enough to generate random variates U and then transform them by applying $F^{-1}(U)$.

Example 0.12 *Let $X \sim \text{Exp}(1)$. Then*

$$F(x) = \int_0^x e^{-t}\, dt = 1 - e^{-x}.$$

It follows that $F^{-1}(t) = -\log(1 - t) = \log(1/(1 - t))$. Therefore by Theorem 0.12, if $U \sim \text{Unif}(0, 1)$, then $Y = F^{-1}(U) = \log(1/(1 - U))$ and $Y \sim \text{Exp}(1)$.

References

[HMC20] Hogg, R. V., McKeen, J. W. and Craig, A. T. (2018) *Introduction to Mathematical Statistics* (8th edition, global ed.), Pearson Education, Chapp. 1, 3.