Chapter 2: Random Variables and Probability Distributions

By Joan Llull*

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

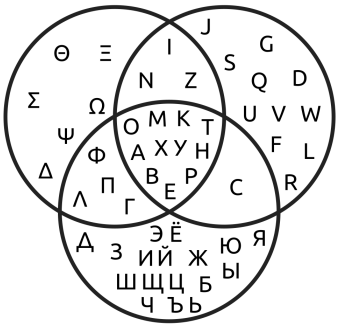
Main references:

— Mood: I.3.2-5; II; III.2.1, III.2.4, III.3.1-2; V.5.1; Appendix A.2.2, A.2.4

— Lindgren: 1.2; 3.1 to 3.3, 3.5; 4.1 to 4.6, 4.9, 4.10; 6.1, 6.5, 6.8; (3.2)

I. Preliminaries: An Introduction to Set Theory

We start this chapter with the introduction of some tools that we are going to use throughout this course (and you will use in subsequent courses). First, we introduce some definitions, and then describe some operators and properties of these operators. Consider a collection of objects, including all objects under consideration in a given discussion. Each object in our collection is an ***element*** or a point. The totality of all these elements is called the ***space***, also known as the universe, or the universal set, and is denoted by Ω. We denote an element of the set Ω by ω. For example, a set can be all the citizens of a country, or all the points in a plane (i.e. $\Omega = \mathbb{R}^2$, and $\omega = (x, y)$ for any pair of real numbers x and y). A partition of the space Ω is called a ***set***, which we denote by calligraphic capital Latin letters, with or without subscripts. When we opt for the second, we define the catalog of all possible incides as the ***index set***, which we denote by Λ (for example, if we consider the sets $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, then $\Lambda = \{1, 2, 3\}$). A ***Venn diagram*** is a diagram that shows all possible logical relations between a finite collection of sets. For example, we would represent the sets of capital letters in the Latin, Greek and Cyrillic scripts as:



* Departament d'Economia i Història Econòmica. Universitat Autònoma de Barcelona. Facultat d'Economia, Edifici B, Campus de Bellaterra, 08193, Cerdanyola del Vallès, Barcelona (Spain). E-mail: joan.llull[at]movebarcelona[dot]eu. URL: http://pareto.uab.cat/jllull.

To express that an element ω is part of a set $\mathcal{A}$, we write $w \in \mathcal{A}$, and to state the opposite, we write $w \notin \mathcal{A}$. We can define sets by explicitly specifying all its elements (e.g. $\mathcal{A} = \{1, 2, 3, 4, 5, 6\}$), or implicitly, by specifying properties that describe its elements (e.g. $\mathcal{A} = \{(x, y) : x \in \mathbb{R}, y \in \mathbb{R}^+\}$). The set that includes no elements is called the ***empty set***, and is denoted by $\varnothing$.

Now we define a list of operators for sets:

- ***Subset***: when all elements of a set $\mathcal{A}$ are also elements of a set $\mathcal{B}$ we say that $\mathcal{A}$ is a subset of $\mathcal{B}$, denoted by $\mathcal{A} \subset \mathcal{B}$ ("$\mathcal{A}$ is contained in $\mathcal{B}$") or $\mathcal{B} \supset \mathcal{A}$ ("$\mathcal{B}$ contains $\mathcal{A}$").

- ***Equivalent set:*** two sets $\mathcal{A}$ and $\mathcal{B}$ are equivalent or equal, denoted $\mathcal{A} = \mathcal{B}$ if $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{A}$.

- ***Union***: the set that consists of all points that are either in $\mathcal{A}$, in $\mathcal{B}$, or in both $\mathcal{A}$ and $\mathcal{B}$ is defined to be the union between $\mathcal{A}$ and $\mathcal{B}$, and is denoted by $\mathcal{A} \cup \mathcal{B}$. More generally, let Λ be an index set, and $\{\mathcal{A}_\lambda\} \equiv \{A_\lambda : \lambda \in \Lambda\}$, a collection of subsets of Ω indexed by Λ. The set that consists of all points that belong to $\mathcal{A}_\lambda$ for at least one $\lambda \in \Lambda$ is called the union of the sets $\{\mathcal{A}_\lambda\}$, denoted by $\bigcup\limits_{\lambda \in \Lambda} \mathcal{A}_\lambda$. If $\Lambda = \varnothing$, we define $\bigcup\limits_{\lambda \in \varnothing} \mathcal{A}_\lambda \equiv \varnothing$.

- ***Intersection***: the set that consists of all points that are both in $\mathcal{A}$ and in $\mathcal{B}$ is defined to be the intersection between $\mathcal{A}$ and $\mathcal{B}$, and is written $\mathcal{A} \cap \mathcal{B}$ or $\mathcal{A}\,\mathcal{B}$. More generally, with the notation from the previous point, the set that consists of all points that belong to $\mathcal{A}_\lambda$ for every $\lambda \in \Lambda$ is called the intersection of the sets $\{\mathcal{A}_\lambda\}$, and is denoted by $\bigcap\limits_{\lambda \in \Lambda} \mathcal{A}_\lambda$. If $\Lambda = \varnothing$, we define $\bigcap\limits_{\lambda \in \varnothing} \mathcal{A}_\lambda \equiv \Omega$.

- ***Set difference***: the set that consists of all points in $\mathcal{A}$ that are not in $\mathcal{B}$ is defined to be the set difference between $\mathcal{A}$ and $\mathcal{B}$, and is denoted by $\mathcal{A} \setminus \mathcal{B}$ (or $\mathcal{A} - \mathcal{B}$ when the context is clear).

- ***Complement***: the complement of a set $\mathcal{A}$ with respect to the space Ω, denoted by $\mathcal{A}^c$ (or $\overline{\mathcal{A}}$) is the set that consists of all points that are in the space Ω and are not in $\mathcal{A}$, that is $\Omega \setminus \mathcal{A}$.

- ***Disjoint/mutually exclusive sets***: $\mathcal{A} \subset \Omega$ and $\mathcal{B} \subset \Omega$ are defined to be mutually exclusive or disjoint if $\mathcal{A} \cap \mathcal{B} = \varnothing$. Subsets $\{\mathcal{A}_\lambda\}$ are defined to be mutually exclusive is $\mathcal{A}_\lambda \cap \mathcal{A}_{\lambda'} = \varnothing$ for every λ and λ' such that $\lambda \neq \lambda'$.

- ***Cartesian product***: the set of all possible ordered pairs (a, b) where $a \in \mathcal{A}$

and $b \in \mathcal{B}$ is defined to be the Cartesian product of $\mathcal{A}$ and $\mathcal{B}$, and is denoted by $\mathcal{A} \times \mathcal{B}$.

- ***Power set***: the power set of a set $\mathcal{A}$, denoted by $2^{\mathcal{A}}$ (or $\mathcal{P}(\mathcal{A})$), is the set of all possible subsets of $\mathcal{A}$, including the empty set $\varnothing$, and $\mathcal{A}$ itself. For example, if $\mathcal{A} = \{x, y, z\}$, $2^{\mathcal{A}} = \{\varnothing, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$. If $\mathcal{A}$ includes n elements, $2^{\mathcal{A}}$ includes 2^n elements (hence its notation).
- ***Finite and countable sets***: a finite set is a set that has a finite number of elements (and an infinite set is a set with an infinite number of elements). A countable set is a set with the same number of elements as some subset of the set of natural numbers (can be finite or infinite).
- ***Sigma-algebra***: a sigma algebra (or σ-algebra), Σ, on a set $\mathcal{A}$ is a subset of the power set of $\mathcal{A}$, $\Sigma \subset 2^{\mathcal{A}}$, that satisfies three properties: i) it includes $\mathcal{A}$; ii) if the subset $\mathcal{B} \subset \mathcal{A}$ is included in Σ, $\mathcal{B}^c$ is also included; and iii) if a countable collection of subsets $\{\mathcal{A}_\lambda\}$ is included, its union $\underset{\lambda \in \Lambda}{\cup} \mathcal{A}_\lambda$ is also included.

Next we list some properties of the operators defined above. Some of the proofs will be done in class, others will be listed as exercises, and others are recommended to be done at own initiative. The properties are:

- ***Commutative laws***: $\mathcal{A} \cup \mathcal{B} = \mathcal{B} \cup \mathcal{A}$, and $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$.
- ***Associative laws***: $\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cup \mathcal{C}$, and $\mathcal{A} \cap (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cap \mathcal{C}$
- ***Distributive laws***: $\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C})$, and $\mathcal{A} \cup (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C})$.
- $\mathcal{A} \cap \Omega = \mathcal{A}$; $\mathcal{A} \cup \Omega = \Omega$; $\mathcal{A} \cap \varnothing = \varnothing$; $\mathcal{A} \cup \varnothing = \mathcal{A}$.
- $\mathcal{A} \cap \mathcal{A}^c = \varnothing$; $\mathcal{A} \cup \mathcal{A}^c = \Omega$; $\mathcal{A} \cap \mathcal{A} = \mathcal{A} \cup \mathcal{A} = \mathcal{A}$.
- $(\mathcal{A}^c)^c = \mathcal{A}$.
- ***DeMorgan's laws***: $(\mathcal{A} \cup \mathcal{B})^c = \mathcal{A}^c \cap \mathcal{B}^c$ and $(\mathcal{A} \cap \mathcal{B})^c = \mathcal{A}^c \cup \mathcal{B}^c$. Likewise, $\left(\underset{\lambda \in \Lambda}{\cup} \mathcal{A}_\lambda \right)^c = \underset{\lambda \in \Lambda}{\cap} \mathcal{A}_\lambda^c$, and $\left(\underset{\lambda \in \Lambda}{\cap} \mathcal{A}_\lambda \right)^c = \underset{\lambda \in \Lambda}{\cup} \mathcal{A}_\lambda^c$.
- $\mathcal{A} \setminus \mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$ and $(\mathcal{A} \setminus \mathcal{B})^c = \mathcal{A}^c \cup \mathcal{B}$.
- $(\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{B}^c) = \mathcal{A}$ and $(\mathcal{A} \cap \mathcal{B}) \cap (\mathcal{A} \cap \mathcal{B}^c) = \varnothing$.
- $\mathcal{A} \subset \mathcal{B} \Rightarrow \mathcal{A} \cap \mathcal{B} = \mathcal{A}$ and $\mathcal{A} \cup \mathcal{B} = \mathcal{B}$.
- $\mathcal{A} \times \mathcal{B} \neq \mathcal{B} \times \mathcal{A}$.

II. Statistical Inference, Random Experiments, and Probabilities

In Chapter 1 we started from a data sample, and we learned how to describe these data. Now we want to make general claims about the population the sample is meant to represent. These general claims are known as ***statistical inference***.

In order to do inference, we need a statistical model, this is, a data generating process. A ***random experiment*** or trial is a conceptual description of the process that generated the data that we observe. We call it random because, even though the process can be replicated under similar conditions, results are not known with certainty (there is more than one possible outcome). This is unlike a deterministic experiment, which has only one possible outcome. For example, a random experiment could be to roll a dice or toss a coin; the outcome of this experiment is random because if we repeat it (roll the dice again, or toss the coin again) we are not necessarily obtain the same result.

The ***probability space*** is a mathematical construct that formalizes a random experiment. The probability space consists of three parts:

- A ***sample space*** Ω, which is the set of all possible outcomes.
- A σ***-algebra***, $\mathcal{F} \subset 2^{\Omega}$, which is a set of events, $\mathcal{F} = \{\mathcal{A}_1, \mathcal{A}_2, ...\}$, where each event $\mathcal{A}_\lambda \subset \Omega$ is a subset of Ω that contains zero or more outcomes. An event $\mathcal{A}_\lambda$ is said to occur if the experiment at hand results in an outcome ω that belongs to $\mathcal{A}_\lambda$.
- A ***probability measure*** $P : \mathcal{F} \to [0, 1]$, which is a function on $\mathcal{F}$ that satisfies three axioms:

  1) $P(\mathcal{A}) \geq 0$ for every $\mathcal{A} \in \mathcal{F}$.
  2) $P(\Omega) = 1$ (Ω is sometimes called the sure event)
  3) If $\mathcal{A}_1, \mathcal{A}_2, ...$ is a sequence of mutually exclusive events in $\mathcal{F}$, then $P(\cup_{\lambda=1}^{\infty} \mathcal{A}_\lambda) = \sum_{\lambda=1}^{\infty} P(\mathcal{A}_\lambda)$.

  The third axiom has two implications. The first is that $P(\varnothing) = 0$. The second is that if we partition the space Ω in mutually exclusive events $\mathcal{A}_1, \mathcal{A}_2, ...$ the sum of the probabilities of all mutually exclusive events that form Ω equals 1.

For example, tossing a coin is a random experiment. The probability space of this experiment is as follows. The sample space is a set that includes heads and tails: $\Omega = \{head, tail\}$. The σ-algebra is the set of all possible combinations

of outcomes, which is $\{\{\varnothing\}, \{head\}, \{tail\}, \{\Omega\}\}$. An event is an element of the σ-algebra, i.e. one of the four listed subsets. And the probability measure is the function that transforms this event into a probability, expressed between 0 and 1. The probability measure satisfies the three axioms: the probability of each of the four events is greater or equal than zero, it is sure that the outcome of the experiment will be either heads or tails, and if we consider the three mutually exclusive sets of the σ-algebra $\{\varnothing\}$, $\{heads\}$, and $\{tails\}$, the probability of the union of the three (which, since the union of the three is Ω, is equal to 1) is equal to the sum of the probabilities of the three events.

III. Finite Sample Spaces and Combinatorial Analysis

In the previous section, we generalized the concept of probability for any sample space. Now we focus on a particular type of sample spaces: those with a finite number of points, $\Omega = \{\omega_1, ..., \omega_N\}$. Let the operator $N(\mathcal{A})$ denote the number of elements of a finite set $\mathcal{A}$. We define $N \equiv N(\Omega)$ as the total number of possible outcomes of a random experiment with a finite number of outcomes.

Initially, consider the case of a finite sample space with ***equally likely*** points. In this case, the probability of each outcome is $1/N$. However, we can also implement the axiomatic definition of probability introduced in the previous section. Define a probability function $P(\cdot)$ over a finite sample space that satisfies two properties:

- $P(\{\omega_1\}) = P(\{\omega_2\}) = ... = P(\{\omega_N\})$.
- If $\mathcal{A} \subset \Omega$ includes $N(\mathcal{A})$ elements, then $P(\mathcal{A}) = N(\mathcal{A})/N$.

We shall call such function an equally likely probability function. It is trivial to check that an equally likely probability function satisfies the three axioms and hence is a probability function.

In this environment, the only problem left in determining the probability of a given event is a problem of counting: count the number of points in $\mathcal{A}$, $N(\mathcal{A})$ and the number of points in Ω, N. For example consider the experiment of tossing a coin twice. Let $\Omega = \{head, tail\} \times \{head, tail\} = \{(z_1, z_2) : z_1 \in \{head, tail\}, z_2 \in \{heads, tail\}\}$. There are $N = 2 \cdot 2 = 4$ sample points. It seems reasonable to attach a probability of $\frac{1}{4}$ to each point. Let $\mathcal{A} = \{$at least one head$\}$. Then, $\mathcal{A} = \{(head, tail), (tail, head), (head, head)\}$, and $P(\mathcal{A}) = \frac{3}{4}$.

In this example, counting was quite simple. However, in higher dimensional cases, we may need to count in a systematic way. For that purpose, it is useful to introduce an important tool: ***combinatorial analysis***. Specifically, let us introduce the following definitions:

- n **factorial**: a product of a positive integer n by all the positive integers smaller than it $[n! \equiv n(n-1)(n-2)...1 = \prod_{j=0}^{n-1}(n-j)]$. We define $0! \equiv 1$.
- $(n)_k$: a product of a positive integer n by the next $k-1$ smaller positive integers $[(n)_k \equiv n(n-1)...(n-k+1) = \prod_{j=0}^{k-1}(n-j) = \frac{n!}{(n-k)!}]$.
- **Combinatorial symbol (or n pick k)**: it is defined as:

$$\binom{n}{k} \equiv \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!}, \quad \text{with} \quad \binom{n}{k} \equiv 0 \text{ if } k < 0 \text{ or } k > n. \tag{1}$$

- **Binomial theorem**: the binomial theorem states that:

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}. \tag{2}$$

Consider an experiment such that each outcome can be represented as an n-tuple (as in our example before, where we expressed each outcome as a 2-tuple. Another example (that will be central in Chapter 4) is drawing a *sample* of size n from an *urn* with M *balls*, numbered from 1 to M. There are two basic ways of drawing a sample: *with replacement* and *without replacement*. In the case of sampling with replacement, the sample space is $\Omega = \{(z_1, ..., z_n) : z_1 \in \{1, ..., M\}, ..., z_n \in \{1, ..., M\}\}$, and in the case of sampling without replacement, it is $\Omega = \{(z_1, ..., z_n) : z_1 \in \{1, ..., M\}, z_2 \in \{1, ..., M\}\backslash\{z_1\}..., z_n \in \{1, ..., M\}\backslash\{z_1, ..., z_{n-1}\}\}$.

As a general rule, counting the number of elements of a set $\mathcal{A}$ composed of points that are n-tuples satisfying certain conditions consists of determining the number of points that may be used as each of the n components, say $N_1, ..., N_n$, and this way obtain $N(\mathcal{A}) = N_1 \cdot N_2 \cdot ... \cdot N_n$. Thus, in the case of sampling with replacement, M^n different samples could possibly be drawn. In the case without replacement, $N(\mathcal{A}) = M \cdot (M-1) \cdot ... \cdot (M-n+1) = (M)_n$.

To determine the size of the power set of a finite sample space with M elements, we can also use combinatorial analysis. For every subset of Ω that contains n elements, we can create $n!$ different (note that, for instance, $\{1, 2, 3\}$ is not different than $\{2, 3, 1\}$, since both contain the same three objects) combinations of n elements, drawing from the set without replacement. Denote the number of different sets of size n can be formed off Ω by x_n. Now, because we know from the previous paragraph that we can draw $(M)_n$ different size n samples from Ω, then $n!x_n = (M)_n$, which implies that $x_n = \frac{(M)_n}{n!} = \binom{M}{n}$. Therefore, the total number of sets that can be formed off Ω is $\sum_{n=0}^{M}\binom{M}{n}$. Thus, using the binomial theorem for $a = b = 1$, we see that $N(2^\Omega) = 2^M$.

We can also consider finite sample spaces ***without equally likely points***. In this case, we have to define our probability function in a different way. We can completely define values for $P(\mathcal{F})$ for each of the $2^{N(\Omega)}$ events by specifying a value of $P(\cdot)$ for each of the $N = N(\Omega)$ elementary elements. Let $\Omega = \{\omega_1, ..., \omega_N\}$, and define $p_j \equiv P(\{\omega_j\})$ for $j = 1, ..., N$. To satisfy the second and third axioms of the probability function, p_j for $j = 1, ..., N$ need to be such that $\sum_{\mathrm{j=1}}^{\mathrm{N}} p_j = 1$, since:

$$\sum_{\mathrm{j=1}}^{\mathrm{N}} p_j = \sum_{\mathrm{j=1}}^{\mathrm{N}} P(\{\omega_j\}) = P\left(\bigcup_{j=1}^{N} \{\omega_j\}\right) = P(\Omega) = 1. \tag{3}$$

For any event $\mathcal{A}$, define $P(\mathcal{A}) \equiv \sum_{\{j:\omega_j \in \mathcal{A}\}} p_j$. It is easy to prove that this function also satisfies the three axioms, and hence is a probability function.

IV. Definition of Random Variable and Cumulative Density Function

A ***random variable***, denoted by $X : \Omega \to \mathbb{R}$, is a function from Ω to the real line such that the set $\mathcal{A}_r$, defined by $\mathcal{A}_r \equiv \{\omega : X(\omega) \leq r\}$, belongs to $\mathcal{F}$ for every real number r. What is important from this definition is that a random variable is a transformation of an event into a numeric value.

In our example of tossing the coin, the number of heads obtained is a random variable because i) it is a transformation of elements of Ω into real numbers (e.g. $X(head) = 1$ and $X(tail) = 0$), and ii) the indicated condition is satisfied for any $r \in \mathbb{R}$: if $r < 0$, $\{\omega : X(\omega) \leq r\} = \varnothing$, if $r \geq 1$, $\{\omega : X(\omega) \leq r\} = \Omega$, and if $0 \leq r < 1$ $\{\omega : X(\omega) \leq r\} = \{tail\}$, and all three belong to $\mathcal{F}$.

A random variable is represented by its ***cumulative distribution function*** (cdf), denoted by F_X, which transforms real numbers into probabilities as follows:

$$F_X : \mathbb{R} \to [0, 1], \quad F_X(x) \equiv P(X \leq x). \tag{4}$$

This concept is analogous to the cumulative frequency that we defined in Chapter 1. In the coin tossing example, the cdf is as follows:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{2}, & \text{if } x \in [0, 1) \\ 1, & \text{if } x \geq 1 \end{cases}. \tag{5}$$

Notice that the main reason why we impose the mathematical condition $\mathcal{A}_r \subset \mathcal{F}$ for all $r \in \mathbb{R}$ is to ensure that the cdf of the random variable is well defined over the entire real line.

A cdf satisfies the following properties:

- In the limit, $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
- F_X is nondecreasing (because if $x_1 < x_2$ then $\{\omega : X(\omega) \leq x_1\} \subseteq \{\omega : X(\omega) \leq x_2\}$).
- F_x is continuous from the right (not necessarily from the left, as in the coin tossing example).

V. Continuous and Discrete Random Variables

We say that a random variable is ***discrete*** if its support includes a finite (or countably infinite) number of points of support. The cdf of a discrete random variable is a step function, with the discrete jumps occurring at the points of support. The cdf is fully characterized by the ***probability mass function*** (pmf), which is defined as $P(X = x_a)$, since:

$$F_X(x) \equiv \sum_{\{a:x_a \leq x\}} P(X = x_a). \tag{6}$$

Note that the concept of pmf is closely tied to the relative frequency defined in Chapter 1.

Analogously, we define a random variable as ***continuous*** if there exists a non-negative function $f_X(\cdot)$ such that:

$$F_X(x) = \int_{-\infty}^{x} f_X(z)dz, \quad \forall x \in \mathbb{R}\,. \tag{7}$$

The function $f_X(\cdot)$ is known as ***probability density function*** (pdf). The pdf indicates the rate at which the probability is accumulated in the neighborhood of a point, which is also connected to the relative frequency explained in Chapter 1. This is easy to see using the definition of derivative (as, from Equation (7), if F_X is differentiable, the pdf is the derivative of the cdf at a given point):

$$f_X(x) = \lim_{h \to 0} \frac{F_X(x+h) - F_X(x)}{h}. \tag{8}$$

Continuous random variables (and their pdfs and cdfs) satisfy the following:

- $f_X(x) \geq 0$ in all the support where $F_X(x)$ is differentiable.
- $\int_{-\infty}^{\infty} f_X(z)dz = 1$, even though nothing prevents $f_X(x) > 1$ at some point x.
- F_X is continuous (from both sides).
- $P(X = x) = 0$ for all x in (and out of) the support of X.

- $P(x_1 < X < x_2) = \int_{x_1}^{x_2} f_X(z)dz$.
- $f_X(x) = \frac{d}{dx}F_X$ at all points where F_X is differentiable.

A random variable can also be ***mixed***: it is continuous in a part of its domain, but also has some points at which there is positive probability mass. More formally, a random variable is mixed if its cdf is of the form:

$$F_X(x) = pF_X^{(d)}(x) + (1-p)F_X^{(c)}(x), \quad 0 < p < 1, \tag{9}$$

where $F_X^{(d)}(\cdot)$ is the cdf of the discrete part, and $F^{(c)X(\cdot)}$ is the cdf of the continuous part. This type of cdf, formed as a convex combination of cdfs of continuous and discrete random variables is called a ***mixture***.

VI. Commonly Used Univariate Distributions

In this section we introduce a set of widely used discrete and continuous parametric families of distributions. In the problem sets throughout the course you may see additional distributions that are also commonly used. In any of the listed manuals, you can find a more extensive list of distributions.

The ***Bernoulli distribution*** is a discrete distribution with pmf given by:

$$f_X(x) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x \in \{0,1\} \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

where the parameter p satisfies $0 \leq p \leq 1$.

The ***binomial distribution*** is a discrete distribution function with pmf given by:

$$f_X(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & \text{for } x = 0, 1, ..., n \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

where $0 \leq p \leq 1$, and n ranges over the positive integers.

The ***Poisson distribution*** is a discrete distribution with pmf given by:

$$f_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{if } x = 0, 1, 2, ... \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where the parameter λ satisfies $\lambda > 0$.

The ***uniform distribution*** is a continuous distribution (there is a discrete version of it) with pdf given by:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

where a and b are the inferior and superior limits of the support, and with cdf given by:

$$F_X(x) = \begin{cases} 0 & \text{if } x \in (-\infty, a) \\ \dfrac{x-a}{b-a} & \text{if } x \in [a, b] \\ 1 & \text{if } x \in (b, \infty) \end{cases} . \tag{14}$$

If X is uniformly distributed, we denote $X \sim \mathcal{U}(a, b)$.

The ***standard normal distribution*** is a continuous distribution with pdf given by:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \tag{15}$$

and cdf given by:

$$\Phi(x) = \int_{-\infty}^{x} \phi(z) dz. \tag{16}$$

Its pdf is symmetric around 0, its only maximum is at $x = 0$, and it has two inflection points at ± 1. The indication that a random variable X is distributed as a standard normal is denoted as $X \sim \mathcal{N}(0, 1)$. The cdf of the normal distribution does not have a closed form solution, but its values are tabulated, and incorporated in most statistical softwares (even in spreadsheets!).

The standard normal distribution can be generalized by means of an affine transformation. This transformation is simply called the ***normal distribution***, and is denoted by $\mathcal{N}(\mu, \sigma^2)$. More specifically, let $Z \sim \mathcal{N}(0, 1)$, and let $X \equiv \mu + \sigma Z$, with $\sigma > 0$; then $X \sim \mathcal{N}(\mu, \sigma^2)$. The cdf of the normal distribution is given by:

$$F_X(x) \equiv P(X \leq x) = P(\mu + \sigma Z \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right), \tag{17}$$

and its pdf is equal to:

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right). \tag{18}$$

In this case, $f_X(\cdot)$ is symmetric with respect to μ, its only maximum is at $x = \mu$, and it has two inflection points at $\pm\sigma$.

VII. Transformations of Random Variables

In this section we want to learn what is the distribution of $Y \equiv g(X)$, given that we know that $X \sim F_X(\cdot)$. For example, suppose that we roll a dice once, and

our random variable X is the number of points we obtain. Let $Y = X^2 - 7X + 10$. In this example:

X	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
Y	4	0	-2	-2	0	4

and thus:

Y	-2	0	4
$P(Y = y)$	1/3	1/3	1/3

More formally, what we have done is the following:

$$P(Y = y) = \sum_{\{i:g(x_i)=y\}} P(X = x_i), \tag{19}$$

this is, we summed the probability mass of all values of the support of X that generate the same value for $g(X)$.

When X is continuous, assuming that $g(\cdot)$ is invertible and differentiable, and that $g'(\cdot) \neq 0$, the cdf of Y is given by:

$$F_Y(y) \equiv P(Y \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)), \tag{20}$$

and the pdf is obtained differentiating:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = f_X\left(g^{-1}(y)\right) \left| \frac{1}{g'\left[g^{-1}(y)\right]} \right|. \tag{21}$$

In words, the probability in Y is accumulated the same way as it accumulates for X (first term) times the rate at which X is transformed —regardless of sign— into Y (second term). If $g(\cdot)$ is not invertible, it is still possible to follow a similar procedure if the function can be divided in invertible pieces.

VIII. Expectation and Moments

The mathematical ***expectation*** of a random variable X, denoted by $\mathbb{E}[X]$, is defined as follows:

- if X is discrete: $\mathbb{E}[X] \equiv \sum_a x_a P(X = x_a)$,
- and if X is continuous: $\mathbb{E}[X] \equiv \int_{-\infty}^{\infty} x f_X(x) dx$.

Note the analogy with the sample mean described in Chapter 1. This is not coincidental, as the expectation is the population equivalent to the sample mean. The

two expressions above can be unified using the ***Riemann-Stieltjes integral***:

$$\mathbb{E}[X] \equiv \int_{-\infty}^{\infty} x dF_X(x). \tag{22}$$

The ***variance*** of a random variable X, denoted by $\text{Var}(X)$, is the expected quadratic deviation with respect to the mean $\mu_X \equiv \mathbb{E}[X]$:

- if X is discrete: $\text{Var}(X) \equiv \sum_a [(x_a - \mu_X)^2 P(X = x_a)]$,
- and if X is continuous: $\text{Var}(X) \equiv \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$.

In general, using the Riemann-Stieltjes integral:

$$\text{Var}(X) \equiv \int_{-\infty}^{\infty} (x - \mu_X)^2 dF_X(x). \tag{23}$$

The variance is a measure of dispersion of the probability mass (or density) of X around its mean, and it is the population counterpart of the sample variance. It is always nonnegative. And we can define the ***standard deviation*** as:

$$\sigma_X \equiv +\sqrt{\text{Var}(X)}, \tag{24}$$

where the positive sign indicates that it is given by the positive root only. More generally, the $\boldsymbol{k^{th}}$ ***central moment*** of the distribution of X is defined as:

$$\mathbb{E}[(x - \mu_X)^k] \equiv \int_{-\infty}^{\infty} (x - \mu_X)^k dF_X(x). \tag{25}$$

Its interpretation is analogous to the sample moments described in Chapter 1, and we can normalize the third and fourth moments in the way described there to obtain the coefficients of ***skewness*** and ***kurtosis***. We similarly define the $\boldsymbol{k^{th}}$ ***uncentered moment*** as $\mathbb{E}[X^k]$.

The expectation (and analogously any moment) of a transformation of X, $Y \equiv g(X)$, can be calculated directly with a transformation of Equation (22), without a need of obtaining the cdf of Y first:

$$\mathbb{E}[Y] \equiv \int_{-\infty}^{\infty} y dF_Y(y) = \int_{-\infty}^{\infty} g(x) dF_X(x). \tag{26}$$

This implication is very useful to establish a set of ***general properties*** of the expectation (and the variance). Let c be a constant, and let $g(X)$ and $h(X)$ denote two arbitrary functions of the random variable X. Then:

- $\mathbb{E}[c] = c$,
- $\mathbb{E}[cX] = c\,\mathbb{E}[X]$,

- $\mathbb{E}[g(X) + h(X)] = \mathbb{E}[g(X)] + \mathbb{E}[h(X)]$,
- $\mathbb{E}[g(X)] \geq \mathbb{E}[h(X)]$ if $g(X) \geq h(X)$ for every possible value of X,

and:

- $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
- $\text{Var}(c) = 0$,
- $\text{Var}(cX) = c^2 \,\text{Var}(X)$,
- $\text{Var}(c + X) = \text{Var}(X)$.

An additional interesting property of the expectations is known as the ***Jensen's inequality***, which is as follows. Let X denote a random variable, and let $g(\cdot)$ denote a continuous and convex function. Then:

$$\mathbb{E}[g(X)] \geq g\left(\mathbb{E}[X]\right). \tag{27}$$

The opposite is true if $g(\cdot)$ is concave. If the function is strictly convex (or concave), the inequality holds strictly. And if the function is linear, then Equation (27) is satisfied with equality. The proof is simple. If a function is (globally) convex, for every point x of its domain crosses a line $h(X) = a_x + b_x X$ that satisfies $h(X) \leq g(X)$ for all X (the tangent). Given that, we know (from the last property of the expectations listed above) that $\mathbb{E}[g(X)] \geq \mathbb{E}[h(X)]$. However, we know that $h(x) = g(x)$ by construction, and thus, at the point $x = \mathbb{E}[X]$, $h(\mathbb{E}[X]) = g(\mathbb{E}[X])$. Because $h(\mathbb{E}[X]) = \mathbb{E}[h(\mathbb{E}[X])]$, the result follows.

Another property is known as the ***Chebyshev's inequality***, which is satisfied by any distribution, and is given by:

$$P(|X - \mu_X| \geq c) \leq \frac{\sigma_X^2}{c^2} \quad \Leftrightarrow \quad P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}, \tag{28}$$

where X is a random variable, μ_X is its mean, σ_X^2 is its variance, c is an arbitrary positive constant, and $k \equiv \frac{c}{\sigma_X}$. Equation (28) states that not more than $\frac{1}{k^2}$ of the distribution's values can be more than k standard deviations away from the mean. For example, in the case of the normal distribution, $P(|X - \mu_X| \geq \sigma_X) \approx 1 - 0.6827 \leq 1$, $P(|X - \mu_X| \geq 2\sigma_X) \approx 1 - 0.9545 \leq 0.25$, and $P(|X - \mu_X| \geq 3\sigma_X) \approx 1 - 0.9973 \leq \frac{1}{9}$.

More generally, the ***Markov's inequality*** establishes that, for any positive constant c and nonnegative function $g(\cdot)$:

$$P(g(X) \geq c) \leq \frac{\mathbb{E}[g(X)]}{c}, \tag{29}$$

provided that the expectation $\mathbb{E}[g(X)]$ exists.

IX. Quantiles, the Median, and the Mode

The τ***th quantile*** of a distribution indicates the minimum value of X below which there is a fraction τ of the density of the distribution:

$$q_\tau \equiv \min\{x : F_X(x) \geq \tau\}, \tag{30}$$

for $\tau \in [0, 1]$. When $F_X(\cdot)$ is invertible, $q_\tau = F_X^{-1}(\tau)$. Thus, the quantiles also characterize the distribution of X, as so does the cdf. The ***median*** (which is the population equivalent to the sample median seen in Chapter 1) is the 0.5th quantile, $q_{0.5}$. The ***mode*** is the value of X that has the maximum density (or mass if X is discrete).