Probability

In this section, we present some results in probability and statistics within the framework of this course. For more information, you can refer to the STT-1000 course, Wasserman (2010) (in English) and Delmas (2013) (in French).

1 Modeling randomness

Many real-world phenomena are unpredictable, and their outcomes generally contain a certain amount of variability. This variability is taken into account using a measure of uncertainty called **probability measure**.

> Definition
>
> The **sample space** S is the set of all possible outcomes of a phenomenon. An **event** is a subset of the sample space S.

> Examples
>
> 1. If the experiment consists of tossing a coin, $S = \{0, 1\}$. The result of this experiment cannot be known in advance. For example, $E = \{1\}$ is an event of S.
>
> 2. If we are interested in the lifespan of a phone, $S = \mathbb{R}_+$. We can also choose $S = [0, M]$, because this lifespan is probably not infinite! The event $E = [10, \infty)$ represents the event "the lifespan of more than 10 time units."
>
> 3. For the number of days without snow in Quebec City in a year, we can choose $S = \mathbb{N}$. The event $E = (0, 5]$ represents the event "fewer than 5 days without snow in Quebec City in a year."

> **Definition**
>
> A **probability measure** $\mathbb{P}$ on S is an application (function) defined on the sample space and satisfying the following properties:
>
> 1. For each event E, $\mathbb{P}(E) \in [0, 1]$.
>
> 2. $\mathbb{P}(S) = 1$.
>
> 3. Let $E_1, E_2, ...$ be a sequence of mutually exclusive events (finite or infinite), i.e. $\forall i \neq j, E_i \cap E_j = \emptyset$. We have
> $$\mathbb{P}(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \mathbb{P}(E_n).$$
>
> We call $\mathbb{P}(E)$ the probability of event E.

The definition of probability measures can be subjective and linked to the statistician's experience. Let's take example 3 on the number of days without snow in Quebec City during the year. Someone who has just arrived in Canada may want to give the same probability to each day, while a Quebecer will have more information and will be able to vary the probabilities based on this knowledge.

> **Definition**
>
> Two events E and F are said to be **independent** if $\mathbb{P}(E \cap F) = \mathbb{P}(E) \times \mathbb{P}(F)$.

> **Definition**
>
> Let E and F be two events. The **conditional probability** that E occurs given that F has occurred is defined by:
> $$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}.$$

Intuitively, two events are independent if knowledge of one provides no information about the occurrence of the other. We also have $\mathbb{P}(E \mid F) = \mathbb{P}(E)$.

2 Random variables

In probability theory, it is customary to express the outcome of experiments as the value of a function called a **random variable**. This characterization is always possible.

> **Definition**
>
> Let X be a random variable. The **distribution** of this random variable is defined by the application $A \mapsto \mathbb{P}(X \in A)$.

> **Definition**
>
> Let X be a random variable. This random variable is **discrete** if it takes, at most, a countable number of values. In this case, the distribution of X is given by the probabilities $\mathbb{P}(X = x)$ for all outcomes x.

> **Definition**
>
> Let X be a random variable. This random variable is **continuous** if the probabilities $\mathbb{P}(X \in A)$ are given by integrals of the form $\int_A f(x)dx$ where $f : \mathbb{R}^d \to \mathbb{R}_+$ is an integrable function such that $\int_{\mathbb{R}^d} f(x)dx = 1$. Note that, for a fixed outcome x, $\mathbb{P}(X = x) = 0$.

> **Definition**
>
> Let X be a random variable. The **mathematical expectation** $\mathbb{E}(X)$ of X is the average value of the outcome of X with respect to its probability distribution. The expectation is usually denoted by μ.

Let F be a countable set. A discrete random variable X has an expectation $\mathbb{E}(X) = \sum_{x \in F} x \mathbb{P}(X = x)$. Let X be a continuous random variable with density f. Its expectation is given by $\mathbb{E}(X) = \int_{\mathbb{R}^d} x f(x)dx$.

> **Transfer theorem**
>
> Let X be a random variable. Let $g : \mathbb{R}^d \mapsto \mathbb{R}$ be a function such that $\mathbb{E}[g(X)]$ exists. We have:
>
> 1. If X is a discrete random variable, $\mathbb{E}[g(X)] = \sum_{x \in F} g(x)\mathbb{P}(X = x)$;
>
> 2. If X is a continuous random variable with density f, $\mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(x)f(x)dx$.

> **Properties: Linearity of expectation**
>
> Let X and Y be two random variables whose expectations exist, and let $\lambda \in R$. We have:
>
> 1. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$;
>
> 2. $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$.

> **Proof**
>
> The proof is derived from the transfer theorem and the linearity of addition and integration.

> **Definition**
>
> Let X be a random variable such that the expectation of its square exists. The **variance** of X is defined by
> $$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2.$$

The variance measures the dispersion of a random variable around its mean. We can also look at the **standard deviation**, defined as the square root of the variance: $\sigma(X) = \sqrt{\text{Var}(X)}$.

> **Definition**
>
> Let X and Y be two random variables and A and B be two events. If the events $\{X \in A\}$ and $\{Y \in B\}$ are independent, then we say that the random variables X and Y are **independent**.

From this definition, we can deduce that:

1. for functions f and g, the random variables $f(X)$ and $g(Y)$ are independent;

2. if the random variables X and Y are real-valued and their expectation exists, then the expectation of the product XY exists and $\mathbb{E}(XY) = \mathbb{E}(X) \times \mathbb{E}(Y)$.

> **Definition**
>
> Let X be a random variable. The **distribution function** $F : \mathbb{R} \mapsto [0, 1]$ of X is defined by
> $$F(t) = \mathbb{P}(X \leq t), \quad t \in \mathbb{R}.$$

3 Random vectors

Suppose that $X = (X_1, X_2)$ is a random variable of dimension 2 with density f_X. Random variables of dimension greater than 1 are generally referred to as **random vectors**. The densities of X_1 and X_2 are called **marginal densities**. When X_1 and X_2 are independent, we have:
$$f_X(x, y) = f_{X_1}(x) \cdot f_{X_2}(y), \quad (x, y) \in \mathbb{R}^2.$$

> **Example of the multivariate normal distribution**
>
> A random vector X of dimension p is said to follow a multivariate normal distribution with mean μ and variance Σ if its density is given by
>
> $$f_X(x) = \frac{1}{(2\pi)^{p/2}} \cdot \frac{1}{(\det\Sigma)^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p.$$
>
> We denote $X \sim \mathcal{N}_p(\mu, \Sigma)$.

In statistics, an important quantity to measure is the linear dependence between X_1 and X_2. For this, we can use the covariance or the correlation.

> **Definition**
>
> Let $X = (X_1, X_2)$ be a random vector such that the expectation of the square of X_1 and X_2 exists. The **covariance** between X_1 and X_2 is given by
>
> $$\text{Cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right].$$
>
> The **correlation** between X_1 and X_2 is a version of the covariance normalized by the standard deviation of the random variables. It is given by
>
> $$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sigma(X_1)\sigma(X_2)}.$$

The sign of the covariance and correlation can be interpreted. If they are strictly positive, X_1 and X_2 tend to move in the same direction. If X_1 increases, then X_2 also increases, and vice versa. If they are strictly negative, X_1 and X_2 tend to move in opposite directions. If X_1 increases, then X_2 decreases, and vice versa. If the covariance is equal to 0, there are no rules and X_1 and X_2 are said to be orthogonal.

> **Properties**
>
> Let $X = (X_1, X_2)$ be a random vector. We have
>
> 1. $\text{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$;
> 2. $\text{Cov}(X_1, X_2) = \text{Cov}(X_2, X_1)$;
> 3. $\text{Cov}(X_1 + \lambda Y_1, X_2) = \text{Cov}(X_1, X_2) + \lambda\text{Cov}(Y_1, X_2)$.

> **Proof**
>
> 1. We find the result by expanding the product in the definition of the covariance.
>
> 2. Using point 1. and the commutativity of multiplication.
>
> 3. Using point 1. and the linearity of expectation.

4 Estimation

In practice, we do not have perfect knowledge of our random vectors, but only some of their realizations (called samples). Let $x_1, \dots, x_n$ be n independent realizations of a random vector X with mean μ and variance Σ.

The estimator of the mean μ is given by

$$\hat{\mu} = \overline{X} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The estimator of the variance Σ is given by

$$\widehat{\Sigma} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top.$$

Why do we divide this sum by $n-1$ and not by n to estimate the variance? If we divide by n, $\widehat{\Sigma}$ is a biased estimator of the variance. Indeed, we must take into account that we are using a biased estimator of the mean in the variance estimator and therefore correct for this estimate.

Let $D = \{\mathrm{diag}(\widehat{\Sigma})\}^{1/2}$ be the matrix of standard deviations calculated on the sample. We can estimate the correlation matrix on the sample by

$$\widehat{R} = D^{-1} \widehat{\Sigma} D^{-1}.$$

Delmas, Jean-François. 2013. *Introduction au calcul des probabilités et à la statistique : exercices, problèmes et corrections (2e édition)*. Les Presses de l'ENSTA.

Wasserman, Larry. 2010. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.