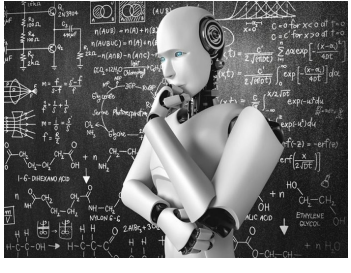
Supervised Learning

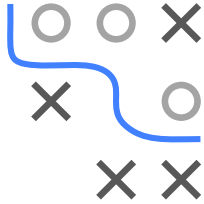
Refreshing Mathematical Tools



Learning goals

- Refresher on the basics of probability theory

PROBABILITY SPACE

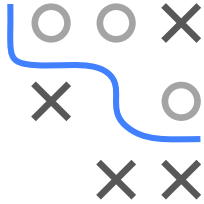


Probability space. A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ which is modeling a specific random experiment/process. The components are

- a sample space Ω, which is the set of all possible outcomes of the random process modeled by the probability space.
- a σ-algebra $\mathcal{F}$, which is a family of sets representing the allowable events of the random process modeled by the probability space. In particular, each set in $\mathcal{F}$ is a subset of the sample space Ω.
- a probability measure $\mathbb{P}$, which assigns each allowable event a probability value, i.e., $\mathbb{P} : \mathcal{F} \to [0, 1]$. It satisfies the following axioms of probability:
  - *Completeness* — $\mathbb{P}(\Omega) = 1$,
  - *σ-additivity* — For any finite or countably infinite sequence of mutually disjoint events $E_1, E_2, \ldots,$ it holds that

$$\mathbb{P}(\cup_{i\geq 1} E_i) = \sum\nolimits_{i\geq 1} \mathbb{P}(E_i).$$

PROBABILITY SPACE / 2



Examples:

- *Coin Tossing* — Possible outcomes are $\Omega = \{\mathtt{H}, \mathtt{T}\}$ with $\mathtt{H}$ resp. $\mathtt{T}$ representing "heads" resp. "tails". The allowable events are contained in $\mathcal{F} = \{\varnothing, \{\mathtt{H}\}, \{\mathtt{T}\}, \{\mathtt{H}, \mathtt{T}\}\}$. If the coin is fair, then $\mathbb{P}(\mathtt{H}) = \mathbb{P}(\mathtt{T}) = 1/2$.
- *Dice Rolling*— Possible outcomes are $\Omega = \{1, 2, 3, 4, 5, 6\}$. The allowable events are contained in 2^{Ω}, i.e., the power set of Ω. If the dice is fair, then $\mathbb{P}(i) = 1/6$ for $i = 1, \ldots, 6$.

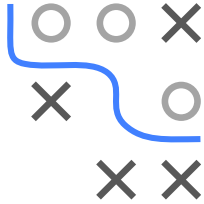
Further properties. For any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ the following properties hold

- *Monotonicity* — $A, B \in \mathcal{F}$, and $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- *Union bound* — For any finite or countably infinite sequence events $E_1, E_2, \ldots,$

$$\mathbb{P}(\cup_{i \geq 1} E_i) \leq \sum\nolimits_{i \geq 1} \mathbb{P}(E_i).$$

RUNNING EXERCISE

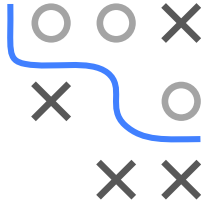
An urn contains five blue, three green, and one red ball. Two balls are randomly selected (without replacement).



What is the sample space of this experiment?

What is the probability of each point in the sample space?

INDEPENDENCE



Independence of events:

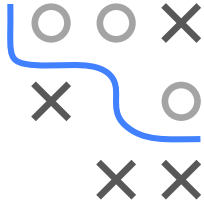
- Two events $A, B \in \mathcal{F}$ are *independent* iff $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
- Events $E_1, \ldots, E_n \in \mathcal{F}$ are called *pairwise independent* iff any pair E_i, E_j with $i \neq j$ is independent.
- Events $E_1, \ldots, E_n \in \mathcal{F}$ are called *mutually independent* iff for any subset $I \subset \{1, \ldots, n\}$ it holds that $\mathbb{P}(\bigcap_{i \in I} E_i) = \prod_{i \in I} \mathbb{P}(E_i)$.

Note that pairwise independence does not imply mutual independence!

Example: One urn with four balls with the labels $110, 101, 011, 000$. We select one ball randomly.

- For $i = 1, 2, 3$ let
  $E_i := \{\text{Selected ball has zero at the } i\text{-th position}\}$.
- $\mathbb{P}(E_i) = 1/2$, $i = 1, 2, 3$ and
  $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1 \cap E_3) = \mathbb{P}(E_2 \cap E_3) = 1/4$.
- But $\mathbb{P}(E_1 \cap E_2 \cap E_3) = 1/4 \neq 1/8 = \mathbb{P}(E_1)\mathbb{P}(E_2)\mathbb{P}(E_3)$.

INDEPENDENCE / 2



Conditional probability. The *conditional probability* that event $A \in \mathcal{F}$ occurs given that event $B \in \mathcal{F}$ occurs is $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$. The conditional probability is well-defined only if $\mathbb{P}(B) > 0$.

Bayes rule. For two events $A, B \in \mathcal{F}$ it holds that

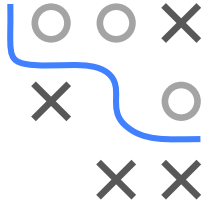
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

The law of total probability. Let $E_1, \ldots, E_n \in \mathcal{F}$ be mutually disjoint events, such that $\cup_{i=1}^{n} E_i = \Omega$, then $\forall A \in \mathcal{F}$,

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

RUNNING EXERCISE

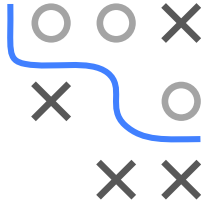
An urn contains five blue, three green, and one red ball. Two balls are randomly selected (without replacement).



Consider the event $A = \{\text{First ball is red}\}$ and the event $B = \{\text{Second ball is red}\}$. Are these events independent?

Are the events independent if we put a ball back into the urn after each selection?

RANDOM VARIABLES



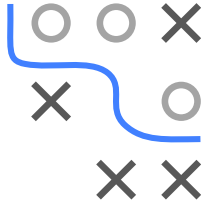
Random variables. A random variable X on a sample space Ω is a real-valued function on Ω, that is $X : \Omega \to \mathbb{R}$. The following observations can be made:

- A random variable defines a probability space $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ with $\Omega_X = Im(X)$ and $\mathbb{P}_X(A) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\})$ for any $A \subset \Omega_X$. Usually, one writes just $\mathbb{P}(X \in A)$ to denote the latter term, which is the *probability distribution of* X.

  (Technical remark: $\mathcal{F}_X$ is usually the Borel-σ-algebra on $\mathbb{R}$.)

- In practical applications oftentimes the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is not the interesting object, but rather the induced probability space by X. One is rather interested in the probability of the outcome of the random variable:
  - *Coin tossing* — Let X be the random variable counting the number of tails after 100 flips.
  - *Financial market* — Let X_t be the price of some asset in a future time t.

RANDOM VARIABLES / 2

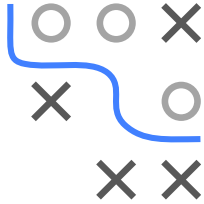


- Functions of random variables are again random variables, i.e., if $f : \mathbb{R} \to \mathbb{R}$ is some (measurable) function, then $Z = f(X)$ is also a random variable.
- Identically distributed — Two random variables X and Y are identically distributed if their probability distributions coincide, i.e., $\mathbb{P}_X = \mathbb{P}_Y$.

One distinguishes between two types of random variables:

- A *discrete random variable* is a random variable that can take only a finite or countably infinite number of values. Its probability distribution is determined by the *probability mass function* which assigns a probability to each value in the image of X.
- A *continuous random variable* is a random variable which can take uncountably infinite number of values. Usually its probability distribution is determined by a *density function,* which assigns probabilities to intervals of the image of X.

DISCRETE RANDOM VARIABLES



If the image Ω_X of X is discrete (e.g., finite or countably infinite), then X is called a *discrete RV*.

For a discrete RV X, the function

$$p: \Omega_X \to [0,1],\ x \mapsto \mathbb{P}(X \in \{x\}) = \mathbb{P}(X = x)$$

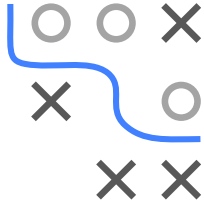
is called a probability function or probability mass function of X.

Obviously, $p(x) \geq 0$ and $\sum_{x \in \Omega_X} p(x) = 1$.
Examples:

- *Bernoulli distribution*: For a binary RV with $\Omega_X = \{0, 1\}$, $X \sim \mathrm{Ber}(\theta)$ if $p(1) = \theta$ and $p(0) = 1 - \theta$.
- *Binomial distribution*: $X \sim \mathrm{Bin}(n, \theta)$ if

$$p(k) = \begin{cases} \binom{n}{k}\theta^k(1-\theta)^{n-k} & \text{if } k \in \{0, \ldots, n\} \\ 0 & \text{otherwise} \end{cases}.$$

CONTINUOUS RANDOM VARIABLES



X is a *continuous RV* if Ω_X is non-discrete and there exists a function $p : \mathbb{R} \to \mathbb{R}$ such that

- $p(x) \geq 0$ for all $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} p(x)\, dx = 1$,
- for all $a \leq b$ it holds that $\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\, dx$.

The function p is called the probability density function (PDF) of X.

Examples.

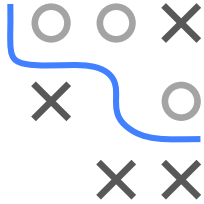
- *Uniform distribution*: $X \sim U(a, b)$ if

$$p(x) = \begin{cases} 1/(b-a) & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} .$$

- *Normal/Gaussian distribution*: $X \sim \mathcal{N}(\mu, \sigma)$ if

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) .$$

CUMULATIVE DISTRIBUTION FUNCTION



The cumulative distribution function (CDF) of a random variable X is the function

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

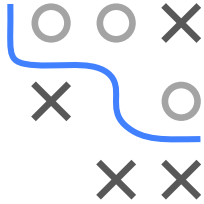
A CDF fully characterizes a RV: If $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$, then X and Y are identically distributed.

If X is

- discrete with probability mass function p, then for all $x \in \mathbb{R}$, $F_X(x) = \sum_{y \in \Omega_X \cap (-\infty, x]} p(y)\,.$
- continuous with probability density function p, then $F_X(x) = \int_{-\infty}^{x} p(t)\,dt$ for all $x \in \mathbb{R}$, and $p(x) = F_X'(x)$ whenever F_X is differentiable at x.

RUNNING EXERCISE

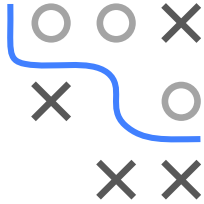
An urn contains five blue, three green, and one red ball. Two balls are randomly selected (without replacement).



Let X be the number of green balls selected. What are the possible values of X?

What is the cumulative distribution function of X?

EXPECTED VALUE/EXPECTATION



Expectation is the most basic characteristic of a random variable. Let X be a random variable, then the expectation of X, denoted by $\mathbb{E}(X)$, is

$$\mathbb{E}(X) = \int x \, dF(x) = \begin{cases} \sum_{x \in \Omega_X} x \, p(x) & \text{if } X \text{ is discrete} \\ \int_{\Omega_X} x \, p(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

provided the sum resp. the integral is well-defined and exists.
Some important properties of the expected value are:

- Linearity — For any constants $c_1, c_2 \in \mathbb{R}$ and any pair of random variables X and Y it holds that $\mathbb{E}(c_1 X + c_2 Y) = c_1 \mathbb{E}(X) + c_2 \mathbb{E}(Y)$.
- Transformations — If $f : \mathbb{R} \to \mathbb{R}$ is a (measurable) function, then the expectation of $f(X)$ is

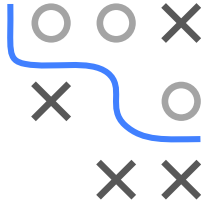
$$\mathbb{E}(f(X)) = \int f(x) \, dF(x) = \begin{cases} \sum_{x \in \Omega_X} f(x) \, p(x) & \text{if } X \text{ is discrete} \\ \int_{\Omega_X} f(x) \, p(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

(provided the sum resp. integral exists.)

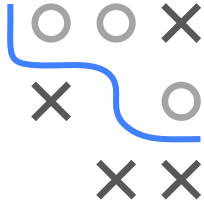
RUNNING EXERCISE

An urn contains five blue, three green, and one red ball. Two balls are randomly selected (without replacement).

Let X be the number of green balls selected. What is the expected value of X?



VARIANCE AND COVARIANCE



The variance of a RV X is defined as follows:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \int_{\Omega_X} (x - \mathbb{E}(X))^2 \, dF(x)\,,$$

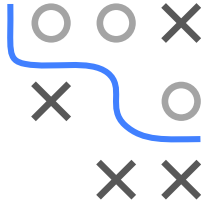
provided the integral on the right-hand side exists.

The *standard deviation* is defined by $\sqrt{\mathrm{Var}(X)}$.

The *covariance* between RVs X and Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

MULTIVARIATE RANDOM VARIABLES



RVs $X_1, \ldots, X_n$ over the same probability space can be combined into a *random vector* $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Their joint distribution is specified by the joint mass/density function $p_{\boldsymbol{X}}$, such that for any measurable set A it holds that

$$\mathbb{P}(\boldsymbol{X} \in A) = \begin{cases} \sum_{(x_1,\ldots,x_n)\in A} p_{\boldsymbol{X}}(x_1, \ldots, x_n) & \text{if } X_1, \ldots, X_n \text{ are discrete} \\ \int_A p_{\boldsymbol{X}}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n & \text{if } X_1, \ldots, X_n \text{ are continuous} \end{cases}$$

The marginal distribution p_1 of X_1 is given by

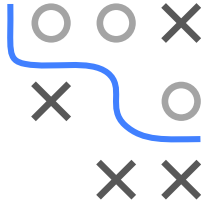
$$p_1(x_1) = \begin{cases} \sum_{(x_2,\ldots,x_n)\in \Omega_{X_2}\times\ldots\times\Omega_{X_n}} p_{\boldsymbol{X}}(x_1, \ldots, x_n) & \text{if discrete} \\ \int_{\Omega_{X_2}\times\ldots\times\Omega_{X_n}} p_{\boldsymbol{X}}(x_1, \ldots, x_n)\, dx_2 \ldots dx_n & \text{if continuous} \end{cases}$$

In the same way, the marginal distributions of $X_2, \ldots, X_n$ are defined. The same type of projection (summation/integration over all remaining variables) is used to define marginal distributions on subsets of variables $(X_{i_1}, \ldots, X_{i_k})$ with $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$.

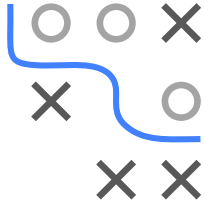
RUNNING EXERCISE

An urn contains five blue, three green, and one red ball. Two balls are randomly selected (without replacement).

Let X_1 be the number of green balls selected and X_2 the number of blue balls selected. What is the joint mass function of (X_1, X_2)?



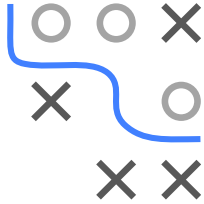
INDEPENDENCE OF RANDOM VARIABLES



- Two discrete random variables X and Y are independent iff for any $x \in \Omega_X, y \in \Omega_Y$ it holds that $p_{XY}(x, y) = p_X(x)p_Y(y)$.
- Random variables $X_1, \ldots, X_n$ are pairwise independent iff for any pair i, j and any $x_i \in \Omega_{X_i}, x_j \in \Omega_{X_j}$ it holds that $p_{X_iX_j}(x_i, x_j) = p_{X_i}(x_i)p_{X_j}(x_j)$.
- Random variables $X_1, \ldots, X_n$ are mutually independent iff for any subset $I \subset \{1, \ldots, n\}$ it holds that the joint probability mass/density function of $(X_i)_{i \in I}$ is given by $\prod_{i \in I} p_{X_i}(x_i)$ for any $x_i \in \Omega_{X_i}$, $i \in I$.

Similar to independence of events, pairwise independence of random variables does not imply their mutual independence. If we say that random variables are independent without further specifications we are referring to mutual independence.

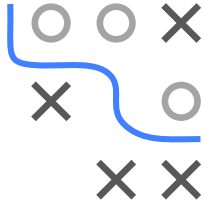
INDEPENDENCE OF RANDOM VARIABLES / 2



Some important properties and concepts with respect to independence are:

- The iid assumption — Random variables $X_1, \ldots, X_n$ are called *independent and identically distributed* (iid) iff they are mutually independent and each random variable has the same probability distribution as the others.
- Independence under transformations — Let X and Y be independent random variables and $f, g : \mathbb{R} \to \mathbb{R}$ are (measurable) functions. Then, $f(X)$ and $g(Y)$ are independent as well.

CONDITIONAL DISTRIBUTIONS



If (X, Y) have a joint distribution with mass function $p_{X,Y}$, then the *conditional probability mass function* for X given Y is defined by

$$p_{X|Y}(x \mid y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

provided $p(Y = y) > 0$.

Likewise, in the continuous case, the *conditional probability density function* is given by

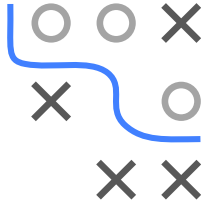
$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

provided $p_Y(y) > 0$. Then,

$$p(X \in A \mid Y = y) = \int_A p_{X|Y}(x \mid y)\, dx\,.$$

The soundness of this definition is less obvious than in the discrete case, due to conditioning on an event of probability 0 here.

CONDITIONAL EXPECTED VALUE/EXPECTATION



Let X and Y be random variables, then the conditional expectation of X given $Y = y$, denoted by $\mathbb{E}(X|Y = y)$, is given by

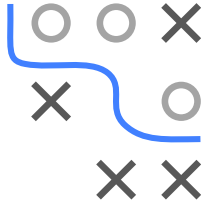
$$\mathbb{E}(X|Y = y) = \begin{cases} \sum_{x \in \Omega_X} x \, p_{X|Y}(x \mid y) & \text{discrete case} \\ \int_{\Omega_X} x \, p_{X|Y}(x \mid y) \, dx & \text{continuous case} \end{cases}$$

Interpretation: The expected value of X under the condition that $Y = y$ holds.

Note that $\mathbb{E}(X|Y = y)$ induces a mapping from Ω_Y to $\mathbb{R}$ according to $y \mapsto \mathbb{E}(X|Y = y)$. This function is called the *conditional expectation of X given Y* and simply denoted by $\mathbb{E}(X|Y)$. Note that $\mathbb{E}(X|Y)$ is a random variable!

$\mathbb{E}(X|Y)$ can be also interpreted as a prediction of X under the information encoded by the random variable Y.

CONDITIONAL EXPECTED VALUE/EXPECTATION / 2



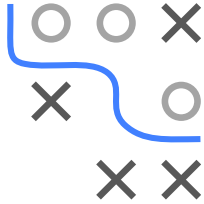
Some important properties of the conditional expected value are the following. For any random variables X, Y, Z it holds that

- Linearity — For any constants $c_1, c_2 \in \mathbb{R}$ it holds that $\mathbb{E}(c_1 X + c_2 Y|Z) = c_1 \mathbb{E}(X|Z) + c_2 \mathbb{E}(Y|Z)$.
- Independence — If X and Y are independent random variables, then $\mathbb{E}(X|Y) = \mathbb{E}(X)$.
- Transformations — If $f : \mathbb{R} \to \mathbb{R}$ is a (measurable) function, then the conditional expectation of $f(X)$ given $Y = y$ is

$$\mathbb{E}(f(X)|Y = y) = \begin{cases} \sum_{x \in \Omega_X} f(x)\, p_{X|Y}(x \mid y) & \text{discrete case} \\ \int_{\Omega_X} f(x)\, p_{X|Y}(x \mid y)\, dx & \text{continuous case} \end{cases}$$

- Law of total expectation — $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$.
- Tower property — $\mathbb{E}(\mathbb{E}(X|Y, Z)|Y) = \mathbb{E}(X|Y)$.

CONDITIONAL VARIANCE



Let X and Y be random variables, then the conditional variance of X given $Y = y$, denoted by $\text{Var}(X|Y = y)$, is given by

$$\text{Var}(X|Y = y) = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y = y])^2 \mid Y = y\right]$$

Interpretation: Variance of the prediction $\mathbb{E}[X \mid Y]$ for X.

Note that $\text{Var}(X|Y = y)$ induces a mapping from Ω_Y to $\mathbb{R}$ according to $y \mapsto \text{Var}(X|Y = y)$. This function is called the *conditional variance of X given Y* and simply denoted by $\text{Var}(X|Y)$. Note that $\text{Var}(X|Y)$ is a random variable!

An important property is the law of total variance:

$$\text{Var}(X) = \mathbb{E}(\text{Var}(X \mid Y)) + \text{Var}(\mathbb{E}(X \mid Y)).$$