Summary of basic probability theory (weeks 1–3)

November 6, 2023

1 Probability space

Let Ω be a set, $\mathcal{F} \subset \mathcal{P}(\Omega) = \{A \mid A \subset \Omega\}$ a set of subsets of Ω , and let $\mathbb{P}: \mathcal{F} \to [0, 1]$ be a function.

- The set $\mathcal{F}$ is called a *σ-algebra* if it satisfies

  (i) $\varnothing \in \mathcal{F}$;

  (ii) $A \in \mathcal{F} \Rightarrow A^{\complement} \in \mathcal{F}$;

  (iii) $\{A_n\}_{n\geq 1}$ is a countable set with $A_n \in \mathcal{F}$, $n \geq 1 \Rightarrow \bigcup_{n\geq 1} A_n \in \mathcal{F}$.

  The elements of $\mathcal{F}$ are called *events*. The σ-algebra structure ensures that if A is an event, then $A^{\complement}$ ("not A") is also an event by condition (ii) and that the set-theoretic union and intersection operations can be used to "build" new events by condition (iii). For example, if A and B are events, then $A \cup B$ ("A or B") is an event and that $A \cap B$ ("A and B") is an event.

- The function $\mathbb{P}$ is called a *probability measure* if it satisfies

  (iv) $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{F}$;

  (v) $\mathbb{P}(\Omega) = 1$;

  (vi) $\{A_n\}_{n\geq 1}$ is a countable set of *disjoint* events $A_n \in \mathcal{F}$, $n \geq 1$, i.e., $A_i \cap A_j = \varnothing$ whenever $i \neq j$, then
  $$\mathbb{P}\Big(\bigcup_{n\geq 1} A_n\Big) = \sum_{n\geq 1} \mathbb{P}(A_n).$$

We call the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space*. The set Ω is called the *sample space*. In our treatment of probability, the set $\mathcal{F}$ is *implicitly defined* depending on the context, and we simply write $(\Omega, \mathbb{P}) = (\Omega, \mathcal{F}, \mathbb{P})$.

These definitions ensure that the usual "rules" for computing using probabilities hold:

- $\mathbb{P}(\varnothing) = 0$.
- If A and B are two events satisfying $A \subset B$, then $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$.
- If A and B are two events satisfying $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- For any event A, there holds $\mathbb{P}(A^{\complement}) = 1 - \mathbb{P}(A)$.
- For any two events A and B (not necessarily disjoint), there holds
  $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$
  Note that if A and B are mutually disjoint events, i.e., $A \cap B = \varnothing$, then the above states that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ ("addition rule of disjoint events").

- For any countable sequence of events $\{A_n\}_{n\geq 1}$, not necessarily pairwise disjoint, there holds
$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} \mathbb{P}(A_n).$$

Definition 1. Let A and B be two events such that $\mathbb{P}(B) > 0$. The *conditional probability* of A, given that B has already happened, is

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

Definition 2. Two events A and B are said to be *independent* if

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

This notion can be expressed in terms of conditional probability.

Lemma 1. *Assume* $\mathbb{P}(B) > 0$. *Then*

- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$;
- *the events* A *and* B *are independent if and only if* $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Theorem 1 (Law of total probability)**.** *Let* $A_1, \ldots, A_k$ *be events that form a partition of* Ω, *i.e.,* $A_i \cap A_j = \varnothing$ *whenever* $i \neq j$ *and* $\Omega = \bigcup_{i=1}^{k} A_i$. *Then, for any event* B, *there holds*

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

This means that we can form the unconditional probability $\mathbb{P}(B)$ given knowledge of $\mathbb{P}(B|A_i)$ and $\mathbb{P}(A_i)$.

Theorem 2 (Bayes' theorem)**.** *Let* A *and* B *be events and assume that* $\mathbb{P}(B) > 0$. *Then*

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

The conditional probability for $A|B$ (the "cause" A given the "effect" B) can be written in terms of the conditional probability for the $B|A$ (the "effect" B given the "cause" A).

2 Random variables

A *random variable (RV)* X with values in a set E is a function $X\colon \Omega \to E$. The set E is called the *outcome* or *target space*.

- When $E \subset \mathbb{R}$, we say that X is a *real-valued random variable*.
- When $E \subset \mathbb{R}^d$, $d \geq 2$, we say that X is a *vector-valued random variable*.
- When $E \subset \mathbb{R}$ is countable, we say that X is a *discrete random variable*.

A random variable $X\colon \Omega \to E$ induces a probability measure P_X on E, defined by

$$P_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) \quad \text{for all subsets } B \subset E,$$

which is called the *probability distribution* (or *law*) of X.

It is common to simply write

$$\{X \in B\} = \{\omega \in \Omega \mid X(\omega) \in B\}$$

and

$$P_X(B) = \mathbb{P}(X \in B).$$

Two random variables X and Y with the same target space are said to be *equal in law* if they have the same probability distribution:

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in B) \quad \text{for all subsets } B \subset E.$$

Usually, we are ultimately interested in the laws of random variables rather than the random variables *per se*. (This is also why the probability space $(\Omega, \mathbb{P})$ is typically suppressed when working with random variables.)

2.1 Discrete random variables

If X is a discrete random variable, i.e., the target space E is countable, then the *probability mass function (PMF)* $p_X : E \to [0,1]$ is simply the probability associated with each value that the random variable can take:

$$p_X(x) = \mathbb{P}(X = x), \quad x \in E.$$

This means that the probability distribution can be written as

$$\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x), \quad B \subset E,$$

which implies that the PMF p_X *determines* the law of X completely.

The *cumulative distribution function (CDF)* $F_X : \mathbb{R} \to [0,1]$ of a real-valued, discrete random variable is

$$F_X(x) = \sum_{\substack{a \le x \\ a \in E}} p_X(a), \quad x \in \mathbb{R}.$$

The CDF satisfies

- $a \le b \;\Rightarrow F_X(a) \le F_X(b)$;
- F_X is right-continuous: $F_X(a) = \lim_{x \to a+} F_X(x)$ for all $a \in \mathbb{R}$;
- $F_X(-\infty) = \lim_{x \to -\infty} F_X(x) = 0$ and $F_X(\infty) = \lim_{x \to \infty} F_X(x) = 1$;
- $a < b \Rightarrow \mathbb{P}(a < X \le b) = F_X(b) - F_X(a)$;
- $\mathbb{P}(X > a) = 1 - F_X(a)$ for $a \in \mathbb{R}$;
- $p_X(x) = \mathbb{P}(X = x) = F_X(x) - \lim_{y \to x-} F_X(y)$ for $x \in \mathbb{R}$,

The generalized inverse of the CDF is called the *quantile function* $F_X^{-1} : (0,1) \to \mathbb{R}$, defined by

$$F_X^{-1}(q) = \inf\{x \in \mathbb{R} \mid F_X(x) \ge q\}, \quad q \in (0,1).$$

The quantile function of a discrete random variable satisfies $F_X(F_X^{-1}(q)) \ge q$ for all $q \in (0,1)$.

2.1.1 Joint distribution (discrete random variables)

If $X : \Omega \to E$ and $Y : \Omega \to F$ are discrete random variables, then the joint PMF $p_{X,Y} : E \times F \to [0,1]$ is defined as

$$p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y), \quad (x,y) \in E \times F.$$

In consequence, the joint probability distribution is

$$P_{X,Y}(C) = \mathbb{P}((X,Y) \in C) = \sum_{(x,y)\in C} p_{X,Y}(x,y) \quad \text{for all } C \subset E \times F.$$

One can obtain the marginal PMFs of X and Y, respectively, by summation over the "nuisance" RVs:

$$p_X(x) = \sum_{y\in F} p_{X,Y}(x,y), \quad x \in E,$$
$$p_Y(y) = \sum_{x\in E} p_{X,Y}(x,y), \quad y \in F,$$

and likewise for the marginal distributions of X and Y, respectively:

$$P_X(A) = \mathbb{P}(X \in A, Y \in F) = \sum_{x\in A,\ y\in F} p_{X,Y}(x,y), \quad A \subset E,$$
$$P_Y(B) = \mathbb{P}(X \in E, Y \in B) = \sum_{x\in E,\ y\in B} p_{X,Y}(x,y), \quad B \subset F.$$

Definition 3. The random variables X and Y are said to independent if, for any subsets $A \subset E$ and $B \subset F$, there holds

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

Equivalently, the random variables X and Y are independent if and only if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \quad \text{for all } (x,y) \in E \times F.$$

The concepts of joint probability distribution, joint PMF, marginal distributions, and independence of random variables can be generalized in a natural way to arbitrarily many random variables.

Definition 4. Let (X,Y) be a discrete random variable in $E \times F$ with joint PMF $p_{X,Y}$ and marginal PMFs p_X and p_Y. The *conditional PMF* $p_{X|Y}$ of X, given a realization of Y, is defined as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \quad \text{for all } x \in E,$$

provided that $y \in F$ satisfies $p_Y(y) > 0$.

2.1.2 Change of variables (discrete random variables).

Proposition 1. *Let $X : \Omega \to E$ and $Y : \Omega \to F$ be discrete random variables such that $Y = g(X)$, where $g \colon E \to F$. Then the PMF of Y is given by*

$$p_Y(y) = \sum_{x\in g^{-1}(\{y\})} p_X(x) = \sum_{\substack{x\in E \\ g(x)=y}} p_X(x).$$

In other words, the PMF of Y at point y is obtained by summing up the PMF of X over the preimage $g^{-1}(\{y\})$.

2.2 Continuous random variables.

A function $f\colon \mathbb{R} \to \mathbb{R}$ is called a *probability density function (PDF)* if

- $f(x) \geq 0$ for all $x \in \mathbb{R}$;
- $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

A real-valued random variable X is said to be a *continuous random variable* if there exists a PDF $f_X\colon \mathbb{R} \to \mathbb{R}$ such that, for all $a \leq b$, there holds

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x. \tag{1}$$

Then f_X is called the *probability density function (PDF)* of X.

Note that (1) implies for any (measurable) subset $A \subset \mathbb{R}$ that

$$P_X(A) = \mathbb{P}(X \in A) = \int_A f_X(x)\,\mathrm{d}x,$$

meaning that the PDF f_X *determines* the law of X completely.

Note also an important difference to discrete random variables: for continuous random variables, there holds

$$\mathbb{P}(X = x) = \int_x^x f_X(t)\,\mathrm{d}t = 0 \quad \text{for all } x \in \mathbb{R}.$$

In consequence, $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b)$ for all $a < b$.

The *cumulative distribution function (CDF)* $F_X\colon \mathbb{R} \to [0,1]$ of a real-valued, continuous random variable is

$$F_X(x) = \int_{-\infty}^x f_X(t)\,\mathrm{d}t.$$

In addition, if F_X is differentiable at $x \in \mathbb{R}$, then

$$f_X(x) = F_X'(x). \qquad \text{(“}F_X \text{ is the antiderivative of } f_X\text{”)}$$

The CDF satisfies

- $a \leq b \Rightarrow F_X(a) \leq F_X(b)$;
- F_X is continuous;
- $F_X(-\infty) = \lim_{x\to-\infty} F_X(x) = 0$ and $F_X(\infty) = \lim_{x\to\infty} F_X(x) = 1$;
- $a \leq b \Rightarrow \mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$;
- $\mathbb{P}(X \geq a) = 1 - F_X(a)$ for $a \in \mathbb{R}$.

The generalized inverse of the CDF is called the *quantile function* $F_X^{-1}\colon (0,1) \to \mathbb{R}$, defined by

$$F_X^{-1}(q) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq q\}, \quad q \in (0,1).$$

The quantile function of a continuous random variable satisfies $F_X(F_X^{-1}(q)) = q$ for all $q \in (0,1)$.

Remark. if the CDF has a function inverse G in the sense that $F_X(G(q)) = q$, then the inverse CDF coincides with the function inverse $F_X^{-1}(q) = G(q)$.

Definition 5 (Continuous joint probability distribution / density)**.** A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is called a *probability density function (PDF)* if the following conditions hold:

- $f(x_1, \dots, x_n) \geq 0$ for all $(x_1, \dots, x_n) \in \mathbb{R}^n$;
- $\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, \dots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1$.

The real-valued random variables $X_1, \dots, X_n$ admit a *continuous joint distribution* (resp. admit a *joint density*) if there exists a PDF $f_{X_1,\dots,X_n} : \mathbb{R}^n \to \mathbb{R}$ such that, for all subsets $A \subset \mathbb{R}^n$, there holds

$$\mathbb{P}((X_1, \dots, X_n) \in A) = \int_A f_{X_1,\dots,X_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Then we call $f_{X_1,\dots,X_n}$ the *probability density function (PDF)* of X.

In the following, we focus on the case $n = 2$ (the generalization to $n > 2$ is natural). The joint probability distribution is

$$P_{X,Y}(C) = \mathbb{P}((X, Y) \in C) = \int_C f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \quad \text{for all } C \subset \mathbb{R} \times \mathbb{R}.$$

One can obtain the marginal PDFs of X and Y, respectively, by integrating out the "nuisance" RVs:

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}} f_{X,Y}(x, y)\, \mathrm{d}y, \quad x \in \mathbb{R}, \\ f_Y(y) &= \int_{\mathbb{R}} f_{X,Y}(x, y)\, \mathrm{d}x, \quad y \in \mathbb{R}, \end{aligned}$$

and likewise for the marginal distributions of X and Y, respectively:

$$\begin{aligned} P_X(A) &= \mathbb{P}(X \in A, Y \in \mathbb{R}) = \int_A \int_{\mathbb{R}} f_{X,Y}(x, y)\, \mathrm{d}y\, \mathrm{d}x, \quad A \subset \mathbb{R}, \\ P_Y(B) &= \mathbb{P}(X \in \mathbb{R}, Y \in B) = \int_{\mathbb{R}} \int_B f_{X,Y}(x, y)\, \mathrm{d}y\, \mathrm{d}x, \quad B \subset \mathbb{R}. \end{aligned}$$

Definition 6. The random variables X and Y are said to independent if, for any subsets $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, there holds

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

Equivalently, the random variables X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad \text{for all } (x, y) \in \mathbb{R} \times \mathbb{R}.$$

The concepts of joint probability distribution, joint PMF, marginal distributions, and independence of random variables can be generalized in a natural way to arbitrarily many random variables.

Definition 7. Let (X, Y) be a continuous random variable in $\mathbb{R}^d \times \mathbb{R}^k$ with joint PDF $f_{X,Y}$ and marginal PMFs f_X and f_Y. The *conditional PDF* $f_{X|Y}$ of X, given a realization of Y, is defined as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad \text{for all } x \in \mathbb{R}^d,$$

provided that $y \in \mathbb{R}^k$ satisfies $f_Y(y) > 0$.

2.2.1 Inverse transform sampling

Theorem 3. *Let X be a continuous, real-valued random variable with CDF F_X and quantile function F_X^{-1}.*

1. *The random variable $U = F_X(X) \sim \mathcal{U}(0,1)$.*
2. *If $U \sim \mathcal{U}(0,1)$, then $F_X^{-1}(U)$ has the same distribution as X (they are equal in law).*

The previous theorem implies the following algorithm.

Algorithm 1 (Inverse transform sampling)**.**

1. *Draw $U \sim \mathcal{U}(0,1)$.*
2. *Calculate $X = F_X^{-1}(U)$.*

If a closed form expression for the inverse CDF is not available, then a computationally attractive formula for approximating the value $F_X^{-1}(U)$ is given by the generalized inverse:

$$F_X^{-1}(q) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq q\}.$$

2.2.2 Change of variables (continuous random variables)

Let $X_1, \dots, X_k$ be real-valued random variables and let $g\colon \mathbb{R}^k \to \mathbb{R}$. In order to derive the PDF of $Z = g(X_1, \dots, X_k)$, one can proceed as follows:

1. Compute the CDF F_Z of Z by
$$F_Z(z) = \mathbb{P}(g(X_1, \dots, X_k) \leq z).$$
2. If F_Z is differentiable, then its PDF is given by $f_Z = F_Z'$.

Theorem 4. *Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuously differentiable and strictly monotonic function. Let X and Y be continuous, real-valued random variables satisfying $Y = g(X)$. Then*

$$\begin{aligned} f_X(x) &= f_Y(f(x))|g'(x)|, \quad x \in \mathbb{R}; \\ f_Y(y) &= f_X(g^{-1}(y))|(g^{-1})'(y)| = f_X(g^{-1}(y))\left|\frac{1}{g'(g^{-1}(y))}\right|, \quad y \in \mathbb{R}. \end{aligned}$$

Theorem 5. *Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be a C^1-diffeomorphism (i.e., g is a bijection and both g and its inverse g^{-1} are continuously differentiable). Let X and Y be continuous random variables with values in $\mathbb{R}^n$ satisfying $Y = g(X)$. Then*

$$\begin{aligned} f_X(x) &= f_Y(g(x))|\det Dg(x)|, \quad x \in \mathbb{R}^n, \\ f_Y(y) &= f_X(g^{-1}(y))|\det Dg^{-1}(y)|, \quad y \in \mathbb{R}^n, \end{aligned}$$

where Dg denotes the Jacobian matrix of g and Dg^{-1} the Jacobian matrix of g^{-1}, respectively.