Probability Theory

Motonobu Kanagawa

Introduction to Statistics, EURECOM

March 17, 2023

Outline

What is the Meaning of "Probability"?

Subjective Probability (used in Bayesian statistics):

- A probability represents one's degree of belief or knowledge about a certain statement, expressed as a number between 0 and 1.

e.g.,

- This drug cures this disease with probability 0.6.

- It will rain today with probability 0.6.

What is the Meaning of "Probability"?

Objective Probability (used in Frequentist statistics):

- A probability represents the degree of randomness.

- Given as the frequency of a statement to be true over infinitely many repeated experiments.

- e.g., think about a biased coin, with the "probability of head 0.6".

- Assume that you toss the coin n times: then the probability statement can be understood as

$$\lim_{n\to\infty} \frac{\text{the number of heads out of } n \text{ trials}}{n} = 0.6.$$

What is the Meaning of "Probability"?

- This lecture introduces mathematical definition of probabilities, which may be used for both (subjective and objective) interpretations.

- Note that this lecture may be the most mathematical among the other lectures in this STATS course.

    - Being mathematically rigorous is similar to being rigorous about grammar in a language course

- But please don't be scared: the other lectures are less mathematical.

    - Only very basic questions may appear in the exam.
    - Don't quite the course because of the lecture today!

Outline

Probability Space: Definition

A triplet $(\Omega, \mathcal{F}, P)$ is called a **probability space**, if

i) Ω is a set (e.g., $\Omega = \mathbb{R}$):

— this is called a sample space, and each $\omega \in \Omega$ is called a sample or an elementary event.

ii) $\mathcal{F}$ is a σ-algebra, i.e., a set of subsets of Ω satisfying

1. $\phi \in \mathcal{F}$ and $\Omega \in \mathcal{F}$; (ϕ is an empty set)
2. If $A \in \mathcal{F}$, then $\Omega \backslash A \in \mathcal{F}$; ($\Omega \backslash A := \{\omega \in \Omega \mid \omega \notin A\}$.)
3. If $A_1, A_2, \ldots, \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ and $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

- Each $A \in \mathcal{F}$ is called a measurable set
- Each $A \in \mathcal{F}$ can be understood as a certain logical statement (we'll see later).
- Thus, the σ-algebra is a set of statements for which probabilities are defined.

Probability Space: Definition

iii) P: a probability measure (distribution), i.e.,

1. P is a function from $\mathcal{F}$ to $[0, 1]$:
    - The probability $P(A)$ of any statement $A \in \mathcal{F}$ being true between 0 and 1.
2. $P(\phi) = 0$ and $P(\Omega) = 1$.
3. For $A_1, A_2, \cdots \in \mathcal{F}$ such that $A_i \cap A_j = \phi$ with $i \neq j$, we have

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

Probability Space: Interpretation

For any two statements $A, B \in \mathcal{F}$,

- The intersection

$$A \cap B := \{\omega \mid \omega \in A \text{ and } \omega \in B\}$$

may be understood as the statement that "both A and B are true."

- The union

$$A \cup B := \{\omega \mid \omega \in A \text{ or } \omega \in B\}$$

may be understood as the statement that "either A or B is true."

- $A \cap B = \phi$ means that A and B cannot be true simultaneously.

Probability Space: Interpretation

Thus, for statements $A, B \in \mathcal{F}$ with $A \cap B = \phi$,

- $P(A \cap B) = P(\phi) = 0$:

  - The probability of both A and B being true is 0.

- $P(A \cup B) = P(A) + P(B)$:

  - The probability that either A or B is true is the sum of the probability of A being true and the probability of B being true.

Probability Space: Interpretation

Note that for any $A \in \mathcal{F}$, the complement

$$\Omega \backslash A := \{\omega \in \Omega \mid \omega \notin A\}$$

maybe understood as the negation of A.

Thus,

$$1 = P(\Omega) = P(A \cup (\Omega \backslash A)) = P(A) + P(\Omega \backslash A).$$

and

$$P(\Omega \backslash A) = 1 - P(A).$$

- The probability of A being not true is 1 minus the probability of A being true.

Examples of Probability Spaces: Finite Discrete Sample Space

Consider fair coin tossing

i.e., the probabilities of "head" and "tail" are the same: $1/2$.

i) Sample space: $\Omega := \{H, T\}$.

i.e., the sample space consists of two elements:

"H" (Head) and "T" (Tail).

Examples of Probability Spaces: Finite Discrete Sample Space

ii) σ-algebra: $\mathcal{F} := \{\phi, \{H\}, \{T\}, \{H, T\}\}$

– i.e., $\mathcal{F} = 2^{\Omega}$ is the power set (= the set of all subsets of Ω).

- $\{H\}$: the statement that "the head appears"
- $\{T\}$: the statement that "the tail appears"
- $\{H, T\} = \{H\} \cup \{T\} = \Omega$: "the head or the tail appears"
- ϕ: the statement that "nothing appears".

iii) Probability measure:

$P(\phi) = 0$, $P(\{H\}) = 1/2$, $P(\{T\}) = 1/2$, $P(\{H, T\}) = 1$.

Examples of Probability Spaces: Infinite Discrete Sample Space

i) Sample space: $\Omega := \mathbb{N} := \{1, 2, \dots\}$ (i.e., all natural numbers).

ii) σ-algebra: $\mathcal{F} := 2^{\mathbb{N}}$ (i.e., all the subsets of $\mathbb{N}$).

iii) Probability measure: $P(\{n\}) := 2^{-n}$ for all $n \in \mathbb{N}$.

Exercise:

- Verify that this example satisfies the definition of a probability space.

Examples of Probability Spaces: Uncountable Sample Space

i) Sample space: $\Omega := [0, 1] \subset \mathbb{R}$.

ii) σ-algebra: $\mathcal{F} =$ the Borel σ-algebra (i.e., the smallest σ-algebra that contains all the open subsets of $[0, 1]$).

iii) Probability measure: $P((a, b)) := b - a$ for all $0 \leq a < b \leq 1$. (i.e., the uniform measure on $[0, 1]$).

- You can think about throwing a needle onto the interval [0,1] uniformly at random.

- Then $(a, b) \in \mathcal{F}$ is the statement that the needle lies between a and b.

Exercises:

- Verify that this example satisfies the definition of a probability space.
- Verify that $\mathcal{F}$ contains all the closed subsets of $[0, 1]$.

Outline

Random Variables

A random variable = a variable that is random.

- e.g., consider rolling a dice:

  - Then the number (1, 2, ..., or 6) that appears in the top is a random variable.

How can we define a random variable mathematically?

Random Variables: Definition

Let $(P, \Omega, \mathcal{F})$ be a probability space.

Let $(\Omega_X, \mathcal{F}_X)$ be a measurable space, i.e.,

- Ω_X is a nonempty set;
- $\mathcal{F}_X$ is a σ-algebra of subsets of Ω_X.

Definition. A function $X : \Omega \to \Omega_X$ is called a random variable, if it is a measurable function, i.e.,

$$\text{For all } S \in \mathcal{F}_X, \text{ we have } X^{-1}(S) \in \mathcal{F};$$

where $X^{-1}(S) := \{\omega \in \Omega \mid X(\omega) \in S\}$ is the inverse image of S.

- In words,

$$\text{If } S \subset \Omega_X \text{ is measurable, then } X^{-1}(S) \subset \Omega \text{ is also measurable.}$$

Random Variables: The Distribution of a Random Variable

Random variable $X : \Omega \to \Omega_X$ induces a probability measure $P_X : \mathcal{F}_X \to [0,1]$ by

$$P_X(S) := P(X^{-1}(S)), \quad S \in \mathcal{F}_X.$$

- $(P_X, \Omega_X, \mathcal{F}_X)$ is a probability space.

- P_X is called the distribution (or the law) of X.

— We write $X \sim P_X$.

- It is said that X takes values in Ω_X.

Exercise:

- Verify that $(\Omega_X, \mathcal{F}_X, P_X)$ is a probability space.

Random Variables: Discrete and Continuous

A random variable $X : \Omega \to \Omega_X$ is said to be

- discrete if X takes countably many (finite or countably infinite) values.

- continuous if X takes uncountably infinitely many values.

Random Variables: A Discrete Example

Consider a biased dice:

– Sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

– σ-algebra: $\mathcal{F} := 2^{\Omega}$ (= the set of all subsets of Ω)

– Probabilities: $P(\{1\}) = 7/12$, $P(\{2\}) = \cdots P(\{6\}) = 1/12$.

Define a random variable X as follows:

– Sample space: $\Omega_X := \{0, 1\}$.

– Random variable: $X : \Omega \to \Omega_X$ defined by

$$X(\omega) := \begin{cases} 0 & \text{if } \omega = 1, 3, \text{ or } 5 \\ 1 & \text{if } \omega = 2, 4, \text{ or } 6. \end{cases}$$

i.e., $X(\omega)$ takes the value 0 if ω is odd, and takes 1 if ω is even.

Random Variables: A Discrete Example

– σ-algebra: $\mathcal{F}_X := 2^{\Omega_X} := \{\phi, \{0\}, \{1\}, \{0,1\}\}$.

The measurability of $X : \Omega \to \Omega_X$ can be checked easily:

- $X^{-1}(\phi) = \phi \in \mathcal{F}$.
- $X^{-1}(\{0\}) = \{1, 3, 5\} \in \mathcal{F}$.
- $X^{-1}(\{1\}) = \{2, 4, 6\} \in \mathcal{F}$.
- $X^{-1}(\{0, 1\}) = \{1, 2, 3, 4, 5, 6\} \in \mathcal{F}$.

The distribution of X is thus given by

- $P_X(\phi) = P(\phi) = 0$.
- $P_X(\{0\}) = P(\{1, 3, 5\}) = 7/12 + 1/12 + 1/12 = 9/12$.
- $P_X(\{1\}) = P(\{2, 4, 6\}) = 1/12 + 1/12 + 1/12 = 3/12$.
- $P_X(\{0, 1\}) = P(\{1, 2, 3, 4, 5, 6\}) = 1$.

Random Variables: A Continuous Example

Consider a random variable following the uniform measure on the unit interval $[0, 1]$.

- The sample space: $\Omega = \Omega_X = [0, 1]$.

- The σ-algebra: $\mathcal{F} = \mathcal{F}_X =$ the Borel-σ algebra (i.e., the smallest σ-algebra containing all open subsets of $[0, 1]$).

- Random variable $X : \Omega \to \Omega_X$ by the identity, i.e., $X(\omega) = \omega$ for all $\omega \in [0, 1]$.

- Probability measure: $P = P_X$ is defined for any interval $(a, b) \subset [0, 1]$ as

$$P((a, b)) = P_X((a, b)) = b - a.$$

For instance, you can think of $X(\omega)$ as the needle that you throw on the interval $[0, 1]$ uniformly at random.

Outline

Expectation of a Random Variable

The expectation (or the mean, the average) of a random variable is an important concept in probability and statistics.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable, with probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

- Let $f : \Omega_X \to \mathbb{R}$ be a measurable function (with respect to the Borel σ-algebra in $\mathbb{R}$).

  - This implies, e.g., $f^{-1}(B) \in \mathcal{F}_X$ for any open (or closed) subset $B \subset \mathbb{R}$.

- Then $f(X)$ is a real-valued random variable.

  - You can interpret $f(X)$ as a mapping $f(X) : \Omega \to \mathbb{R}$:

$$f(X)(\omega) := f(X(\omega)) \in \mathbb{R}.$$

Expectation of a Random Variable

- For a discrete random variable X, the expectation of $f(X)$ is defined by the sum of values $f(x)$ weighted by their probabilities $P_X(\{x\})$:

$$\mathbb{E}[f(X)] := \sum_{x \in \Omega_X} f(x) P_X(\{x\})$$

- For a continuous random variable X, the expectation is defined by the Lebesgue integral:

$$\mathbb{E}[f(X)] := \int_{\Omega_X} f(x) dP_X(x).$$

We will quickly look at the definition of the Lebesgue integral for completeness.

Lebesgue Integration: Simple Functions

First, consider a function $f : \Omega_X \to \mathbb{R}$ of the form (such a f is called a simple function)

$$f(x) = \sum_{i=1}^{n} a_i 1_{S_i}(x),$$

where

- $a_1, \ldots, a_n \in \mathbb{R}$ are constants;
- $S_i \in \mathcal{F}_X$ are disjoint to each other: $S_i \cap S_j = \phi$ for $i \neq j$;
- $1_{S_i} : \Omega_X \to \mathbb{R}$ are indicator functions:

$$1_{S_i}(x) = \begin{cases} 1 & \text{if } x \in S_i, \\ 0 & \text{if } x \notin S_i. \end{cases}$$

Lebesgue Integration: Simple Functions

The integral of the simple function $f = \sum_{i=1}^{n} a_i 1_{S_i}$ is defined by

$$\int_{\Omega_X} f(x) dP_X(x) := \sum_{i=1}^{n} a_i P_X(S_i)$$

- Note that $f(x) = a_i$ for all $x \in S_i$.
- Thus, this takes the same form as the discrete case: the sum of values $f(x)$ weighted by the probabilities $P_X(S_i)$.

In particular, for an indicator function $f(x) := 1_S(x)$ with $S \in \mathcal{F}_X$, the integral is the probability of S:

$$\int_{\Omega_X} 1_S(x) dP_X(x) = P_X(S), \quad S \in \mathcal{F}_X.$$

Lebesgue Integration: Non-negative Functions

Assume that f is measurable and non-negative, i.e., $f(x) \geq 0$ for any $x \in \Omega_X$. (See e.g. [Dudley, 2002, 4.1.5] for the details below.)

- For any $n \in \mathbb{N}$, consider the division of $[0, \infty]$ into disjoint $n \times 2^n + 1$ intervals:

$$
\begin{aligned}
&[0, \infty] \\
&= \left[0, \frac{1}{2^n}\right] \cup \left(\frac{1}{2^n}, \frac{2}{2^n}\right] \cup \left(\frac{2}{2^n}, \frac{3}{2^n}\right] \cup \cdots \cup \left(\frac{n \times 2^n - 1}{2^n}, n\right] \cup (n, \infty] \\
&= \left[0, \frac{1}{2^n}\right] \cup \bigcup_{j=1}^{n \times 2^n - 1} \left(\frac{j}{2^n}, \frac{j+1}{2^n}\right] \cup (n, \infty].
\end{aligned}
$$

- Define the corresponding subsets in Ω_X given by the inverse mapping f^{-1}:

$$
S_{nj} := f^{-1}\left(\left(\frac{j}{2^n}, \frac{j+1}{2^n}\right]\right), \quad U_n := f^{-1}((n, \infty]).
$$

Lebesgue Integration: Non-negative Functions

- These subsets S_{nj} (and U_n) are disjoint to each other.
- $S_{nj} \in \mathcal{F}_X$ and $U_n \in \mathcal{F}_X$ because f is Borel-measurable.
- Note that if $n > \max_{x \in \Omega_X} f(x)$, then $U_n = \phi$.

Then define a simple function $f_n : \Omega_X \to \mathbb{R}$ by

$$f_n(x) := \sum_{j=1}^{n \times 2^n - 1} \frac{j}{2^n} 1_{S_{nj}}(x) + n 1_{U_n}(x),$$

By construction,

- For $x \in S_{nj} = f^{-1}((\frac{j}{2^n}, \frac{j+1}{2^n}])$, we have $f_n(x) = \frac{j}{2^n} < f(x)$.
- For $x \in U_n = f^{-1}((n, \infty])$, we have $f_n(x) = n < f(x)$.
- For $x \in \Omega \backslash \left(\bigcup_j S_{nj} \cup U_n \right)$, we have $f_n(x) = 0 \leq f(x)$.

Therefore $f_n(x) \leq f(x)$ for all $x \in \Omega_X$.

Lebesgue Integration: Non-negative Functions

- We can also show that $f_n(x) \to f(x)$ for $n \to \infty$ for all $x \in \Omega_X$.

- Since f_n is a simple function, we can define the integral

$$\int_{\Omega_X} f_n(x) dP_X(x) := \sum_{j=1}^{n \times 2^n - 1} \frac{j}{2^n} P_X(S_{nj}) + nP_X(U_n)$$
$$= \sum_{j=1}^{n \times 2^n - 1} \frac{j}{2^n} P_X\left(f^{-1}(\left(\frac{j}{2^n}, \frac{j+1}{2^n}\right])\right) + nP_X\left(f^{-1}((n, \infty])\right)$$

- Then the integral of f can be defined as the limit of the integral of f_n as $n \to \infty$:

$$\int_{\Omega_X} f(x) dP_X(x) := \lim_{n \to \infty} \int_{\Omega_X} f_n(x) dP_X(x).$$

If $\int_{\Omega_X} f(x) dP_X(x)$ defined above is finite, f is called integrable.

Lebesgue Integration: General Functions

For a general measurable function $f : \Omega_X \to \mathbb{R}$, we can consider the following decomposition:

$$f(x) = f^+(x) - f^-(x),$$

where

$$f^+(x) := \max(f(x), 0), \quad f^-(x) := \max(-f(x), 0).$$

- These are both non-negative measurable functions:

$$f^+(x) \geq 0, \quad f^-(x) \geq 0.$$

- Thus we can define their integrals as in the previous slides:

$$\int_{\Omega_X} f^+(x) dP_X(x), \quad \int_{\Omega_X} f^-(x) dP_X(x).$$

- If both integrals are finite (i.e., f^+ and f^- are integrable), then f is called integrable, and the integral is given by

$$\int_{\Omega_X} f(x) dP_X(x) := \int_{\Omega_X} f^+(x) dP_X(x) - \int_{\Omega_X} f^-(x) dP(x).$$

Lebesgue Integration: Vector-valued Functions

- Consider a vector-valued function $\mathbf{f} : \Omega_X \to \mathbb{R}^d$ such that

$$\mathbf{f}(x) := (f_1(x), \ldots, f_d(x))^\top \in \mathbb{R}^d$$

where each $f_i : \Omega_X \to \mathbb{R}$ is measurable.

- Then the integral can be defined as

$$\int_{\Omega_X} \mathbf{f}(x) dP_X(x) := \left(\int_{\Omega_X} f_1(x) dP_X(x), \ldots, \int_{\Omega_X} f_d(x) dP_X(x) \right)^\top \in \mathbb{R}^d.$$

Important Examples: The Mean of a Random Variable

Consider the case $\Omega_X = \mathbb{R}^d$: i.e., X takes values in $\mathbb{R}^d$.

- Define $\mathbf{f} : \Omega_X \to \mathbb{R}^d$ as the identity: $\mathbf{f}(x) = x$.

- Then we can define the expected value (or the mean) of X as

$$\mu_X := \mathbb{E}[X] := \int \mathbf{f}(x) dP_X(x) = \int x \; dP_X(x).$$

- This is the average value that X takes.

Important Examples: The Variance of a Random Variable

- Let $g : \Omega_X \to \mathbb{R}$ be a measurable function.

  - Then $g(X)$ is a random variable.

- Let μ_g be its mean: $\mu_g := \mathbb{E}[g(X)] := \int g(x)\, dP_X(x)$.

- The variance of $g(X)$ can be defined as

$$\begin{aligned}\mathrm{Var}[g(X)] &:= \mathbb{E}[(g(X) - \mu_g)^2] \\ &= \int_{\Omega_X} (g(x) - \mu_g)^2 dP_X(x) = \int_{\Omega_X} f(x) dP_X(x).\end{aligned}$$

where we defined $f : \Omega_X \to \mathbb{R}$ by

$$f(x) := (g(x) - \mu_g)^2.$$

- The variance quantifies how much $g(X)$ may vary around the mean $\mu_g = \mathbb{E}[g(X)]$.

Important Examples: The Variance of a Random Variable

- In particular, for $\Omega_X = \mathbb{R}$, the variance of X is

$$\mathrm{Var}[X] := \mathbb{E}[(X - \mu_X)^2] = \int (x - \mu_X)^2 dP_X(x),$$

where

$$\mu_X := \mathbb{E}[X] := \int_{\Omega_X} x \, dP_X(x).$$

Notation

- I will often write the integral without writing the sample space Ω_X (which is obvious from the context):

$$\int_{\Omega_X} f(x) dP_X(x) =: \int f(x) dP_X(x).$$

- Some people also use the following notation

$$P_X f = \int f \ dP_X = \int_{\Omega_X} f(x) P_X(dx) = \int_{\Omega_X} f(x) dP_X(x)$$

- There are also variations in the notation of the expectation:

$$\mathbb{E}[f(X)] = \mathbb{E}_X[f(X)] = \mathbb{E}_{X \sim P_X}[f(X)] = \mathbb{E}_{P_X}[f(X)] = \int_{\Omega_X} f(x) dP_X(x)$$

- Anyway always pay attention to the definition!

Outline

Probability Density Functions

- Probability density functions are an important concept in probability and statistics.

- People often confuse probability density functions and probability distributions (measures)

- Distributions always exist, but density functions may not.

- An important example is Dirac distributions, another key concept in statistics

— This gives a representation of data.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable with probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

Base Measure

Density functions are defined with respect to another measure ν, which is usually called a base measure (or a reference measure).

A set function $\nu : \mathcal{F}_X \to \mathbb{R}$ is called a measure on the measurable space $(\Omega_X, \mathcal{F}_X)$, if it satisfies

- $\nu(A) \geq 0$ for all $A \in \mathcal{F}_X$.
- $\nu(\phi) = 0$.
- For any $A_1, A_2, \cdots \in \mathcal{F}_X$ with $A_i \cap A_j = \phi$ with $i \neq j$,

$$\nu\left(\bigcup_i A_i\right) = \sum_i \nu(A_i).$$

Note that ν is a probability measure if it in addition satisfies

$$\nu(\Omega_X) = 1.$$

But in general, we may have $\nu(\Omega_X) > 1$ or even $\nu(\Omega) = \infty$.

Base Measure

For $\Omega_X \subset \mathbb{R}^d$, a standard choice is ν being the Lebesgue measure:

- For any rectangle

$$A := [a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d, \quad -\infty < a_i < b_i < \infty,$$

the Lebesgue measure outputs its “volume”:

$$\nu(A) = \prod_{i=1}^{n} (b_i - a_i).$$

For simplicity, we often write an integral in the following way, when ν is the Lebesgue measure:

$$\int f(x) d\nu(x) = \int f(x) dx$$

Probability Density Function

-We say that probability measure P_X (or random variable X) has a probability density function $p_X : \Omega_X \to [0, \infty)$ with respect to the base measure ν, if

$$P_X(A) = \int_A p_X(x) d\nu(x) := \int 1_A(x) p_X(x) d\nu(x), \quad \forall A \in \mathcal{F}_X,$$

where 1_A is the indicator function of A:

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

By taking $A := \Omega_X$, it follows that

$$\int_{\Omega_X} p_X(x) d\nu(x) = P_X(\Omega_X) = 1.$$

Example: Gaussian distributions and Gaussian densities

We consider a Gaussian random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$.

- The sample space: $\Omega = \Omega_X = \mathbb{R}$ (the real line).

- The σ-algebra: $\mathcal{F} = \mathcal{F}_X$ = the Borel-σ algebra (i.e., the smallest σ-algebra containing all open subsets of $\mathbb{R}$).

- Random variable $X : \Omega \to \Omega_X$ is the identity, i.e., $X(\omega) = \omega$.

- The Gaussian distribution $P = P_X$ is given by, for all $S \in \mathcal{F}_X$,

$$P(S) = P_X(S) = \int_S p_{\mu,\sigma^2}(x)dx, \quad S \in \mathcal{F}_X$$

where $p_{\mu,\sigma^2} : \mathbb{R} \to [0, \infty)$ is the Gaussian density

$$p_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Probability Distributions and Density Functions

Note that the probability distribution (measure) P_X and the probability density function p_X are **different**!

- Probability distribution (measure) P_X: a function that maps a measurable set to a value in $[0, 1]$:

$$P_X : S \in \mathcal{F}_X \to P_X(S) \in [0, 1].$$

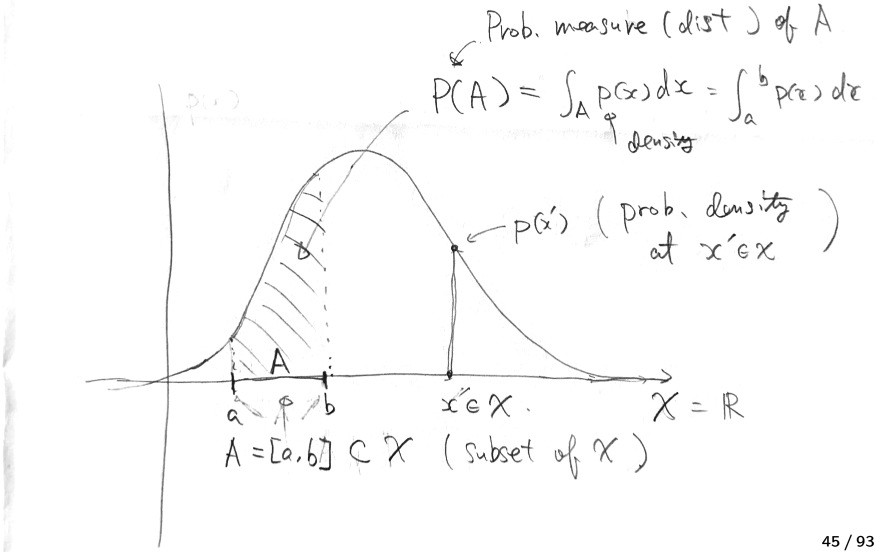
- Probability density function p_X: a function that maps a sample point $x \in \Omega_X$ to a value in $[0, \infty)$.

$$p_X : x \in \Omega_X \to p_X(x) \in \mathbb{R}.$$

Not all probability distributions have density functions.

- Representative examples include Dirac distributions.

Probability Distributions and Density Functions



Dirac Distributions

- For any $z \in \Omega_X$, the Dirac distribution (or Dirac measure) at z, denoted by $\delta_z : \mathcal{F}_X \to \mathbb{R}$ is defined as

$$\delta_z(A) = \begin{cases} 1 & \text{if } z \in A \\ 0 & \text{if } z \notin A. \end{cases}, \quad A \in \mathcal{F}_X.$$

- This is the distribution of a random variable $X : \Omega \to \Omega_X$ such that

$$X(\omega) := z, \quad \forall \omega \in \Omega.$$

- i.e., X takes only the value z, whatever the sample ω is.

This can be seen as follows.

Dirac Distributions

Since $X(\omega) = z$ for all $\omega \in \Omega$, for any $A \in \mathcal{F}_X$ we have

$$X^{-1}(A) := \{\omega \in \Omega \mid X(\omega)(:= z) \in A\}$$
$$= \begin{cases} \Omega & \text{if } z \in A \\ \phi & \text{if } z \notin A. \end{cases}$$

Therefore,

$$\delta_z(A) := P_X(A) := P(X^{-1}(A)) = \begin{cases} P(\Omega) = 1 & \text{if } z \in A \\ P(\phi) = 0 & \text{if } z \notin A. \end{cases}$$

Dirac Distributions

- For any measurable function $f : \Omega_X \to \mathbb{R}$, the expected value of $f(X)$ with $X \sim \delta_z$ is given by

$$\mathbb{E}_{X \sim \delta_x}[f(X)] = \int_{\Omega_X} f(x) d\delta_z(x) = f(z).$$

- This is intuitively obvious, because $X(\omega) = z$ for all $\omega \in \Omega$.

For instance, assume that f is a simple function $f(x) = \sum_i a_i 1_{S_i}(x)$ with disjoint subsets $S_i \in \mathcal{F}_X$. Then:

$$\int f(x) d\delta_z(x) = \sum_i a_i \delta_z(S_i) = \begin{cases} a_j = f(z) & \text{if } z \in S_j \text{ for some } j \\ 0 & \text{otherwise .} \end{cases}$$

Exercise: Prove $\int f(x) d\delta_z(x) = f(z)$ for a general measurable function f. (Recall the definition of the Lebesgue integral)

Dirac Distributions do not Have Density Functions with respect to the Lebesgue Measure.

For instance, assume $\Omega_X = \mathbb{R}$ and let $z = 0$.

Assume that the Dirac distribution δ_z has a probability density function $p_z(x)$. (We'll show a contradiction)

Then for any $a > 0$, we have $z := 0 \in [-a, a]$, and thus

$$1 = \delta_z([-a, a]) = \int_{\Omega_X} 1_{[-a,a]}(x) p_z(x) d\nu(z) \leq 2a \max_{-a<x<a} p_z(x)$$

Therefore,

$$\frac{1}{2a} \leq \max_{-a<x<a} p_z(x).$$

This holds for all $a > 0$. Thus,

$$\infty = \lim_{a \to +0} \frac{1}{2a} \leq \lim_{a \to +0} \max_{-a<x<a} p_z(x).$$

Therefore, p_z is diverging at 0, which is a contradiction.

Another Example where Densities do not Exist

Define $\Omega := \Omega_X := [0,1]^2 \subset \mathbb{R}^2$.

Assume P is the uniform distribution on $[0,1]^2$.

Define $X : \Omega \to \Omega_X$ by

$$X(\omega) = (\omega_1, 1/2), \quad \omega := (\omega_1, \omega_2) \in \Omega.$$

Then the distribution P_X of X does not have a density function with respect to the Lebesgue measure.

$$P_X([a_1, b_1] \times [a_2, b_2]) = \begin{cases} b_1 - a_1 & \text{if } 1/2 \in [a_2, b_2] \\ 0 & \text{otherwise }. \end{cases}$$

Outline

Dealing with Several Random Variables

- So far we have considered only one random variable $X : \Omega \to \Omega_X$.

- There might be another random variable, say $Y : \Omega \to \Omega_Y$ with sample space Ω_Y.

- By sharing the common probability space $(\Omega, \mathcal{F}, P)$, these random variables may be related to each other.

For instance:

- X may be whether or not it will rain tomorrow.

- Y may be whether or not your flight will be delayed tomorrow.

Dealing with Several Random Variables

Here we look at

- how to model several random variables and their joint probability distribution.

- how to quantify the degree of relatedness (independence, covariance etc.)

Joint Random Variable: The Sample Space

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable, with the associated probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

- Let $Y : \Omega \to \Omega_Y$ be a random variable, with the associated probability space $(\Omega_Y, \mathcal{F}_Y, P_Y)$.

Define a joint sample space as the product set

$$\Omega_X \times \Omega_Y := \{(\omega_1, \omega_2) \mid \omega_1 \in \Omega_X,\ \omega_2 \in \Omega_Y\}.$$

Joint Random Variable: The σ-algebra

Define $\mathcal{F}_X \otimes \mathcal{F}_Y \subset 2^{\Omega_X \times \Omega_Y}$ as the product σ-algebra, i.e., the smallest σ-algebra containing all subsets ("rectangles") of the form

$$\begin{aligned} S &= A \times B \\ &= \{(\omega_1, \omega_2) \in \Omega_X \times \Omega_Y \mid \omega_1 \in A,\ \omega_2 \in B\}, \quad A \in \mathcal{F}_X,\ B \in \mathcal{F}_Y. \end{aligned}$$

- Note that each $A \in \mathcal{F}_X$ is a certain statement for which a probability can be defined;

- Similarly, each $B \in \mathcal{F}_Y$ is a certain statement for which a probability can be defined.

- The measurable set $A \times B \in \mathcal{F}_X \otimes \mathcal{F}_Y$ is thus a combined statement that both A and B are true, for which we define a probability.

Joint Random Variable: The Definition

We define the joint random variable of X and Y as a mapping $(X, Y) : \Omega \to \Omega_X \times \Omega_Y$ by

$$(X, Y)(\omega) := (X(\omega), Y(\omega)) \in \Omega_X \times \Omega_Y, \quad \omega \in \Omega.$$

The inverse map $(X, Y)^{-1} : \mathcal{F}_{(X,Y)} \to \Omega$ is defined by

$$(X, Y)^{-1}(S) := \{\omega \in \Omega \mid (X(\omega), Y(\omega)) \in S\}, \quad S \in \mathcal{F}_{(X,Y)}.$$

In particular, for a set of the form $S = A \times B$,

$$\begin{aligned}(X, Y)^{-1}(A \times B) &:= \{\omega \in \Omega \mid X(\omega) \in A \text{ and } Y(\omega) \in B\} \\ &= X^{-1}(A) \cap Y^{-1}(B), \quad A \in \mathcal{F}_X,\ B \in \mathcal{F}_Y\end{aligned}$$

- i.e., the inverse map $(X, Y)^{-1}(A \times B)$ is a subset in Ω for which both $X(\omega) \in A$ and $Y(\omega) \in B$ hold.

- Intuitively, $(X, Y)^{-1}(A \times B) \in \mathcal{F}$ is the statement that A and B are both true.

Joint Random Variable: The Joint Distribution

- $P_{(X,Y)}$: Joint distribution defined as

$$P_{(X,Y)}(S) := P((X,Y)^{-1}(S)), \quad S \in \mathcal{F}_X \otimes \mathcal{F}_Y$$

In particular, for $A \in \mathcal{F}_X$ and $B \in \mathcal{F}_Y$, we have

$$\begin{aligned} P_{(X,Y)}(A \times B) &:= P((X,Y)^{-1}(A \times B)) \\ &= P(\{\omega \in \Omega \mid X(\omega) \in A \text{ and } Y(\omega) \in B\}). \end{aligned}$$

Then the triplet

$$\left(\Omega_X \times \Omega_Y,\ \mathcal{F}_X \otimes \mathcal{F}_Y,\ P_{(X,Y)}\right)$$

is a probability space.

Joint Random Variable: Some Properties

If we take $A = \Omega_X$, then for any $B \in \mathcal{F}_Y$ we have

$$\begin{aligned}(X, Y)^{-1}(\Omega_X \times B) &= X^{-1}(\Omega_X) \cap Y^{-1}(B)\\ &= \Omega \cap Y^{-1}(B)\\ &= Y^{-1}(B).\end{aligned}$$

Therefore,

$$\begin{aligned}P_{(X,Y)}(\Omega_X \times B) &= P((X, Y)^{-1}(\Omega_X \times B))\\ &= P(Y^{-1}(B))\\ &= P_Y(B).\end{aligned}$$

Similarly, we have

$$P_{(X,Y)}(A \times \Omega_Y) = P_X(A), \quad \forall A \in \mathcal{F}_X$$

In this context, P_X and P_Y are called marginal distributions of $P_{(X,Y)}$.

Example: A Fair Dice

Let's consider a fair dice.

- **Sample space:** $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- σ-**algebra:** $\mathcal{F} = 2^\Omega$ (the power set, i.e., the set of all subsets of Ω).
- **Probability:** $P(\{1\}) = P(\{2\}) \cdots = P(\{6\}) = 1/6$.

Define random variables

$X : \Omega \to \Omega_X := \{a, b\}$:

$$X(\omega) := \begin{cases} a & \text{if } \omega \text{ is odd } (i.e., 1, 3, 5) \\ b & \text{if } \omega \text{ is even } (i.e., 2, 4, 6) \end{cases}$$

$Y : \Omega \to \Omega_Y := \{c, d\}$:

$$Y(\omega) := \begin{cases} c & \text{if } \omega = 1 \\ d & \text{if } \omega = 2, 3, 4, 5, 6 \end{cases}$$

Example: A Fair Dice

The probability distribution P_X of X:

$$P_X(\{a\}) = P(X^{-1}(a)) = P(\{1, 3, 5\}) = \frac{1}{2},$$
$$P_X(\{b\}) = P(X^{-1}(b)) = P(\{2, 4, 6\}) = \frac{1}{2}.$$

The probability distribution P_Y of Y:

$$P_Y(\{c\}) = P(Y^{-1}(\{c\})) = P(\{1\}) = \frac{1}{6},$$
$$P_Y(\{d\}) = P(Y^{-1}(\{d\})) = P(\{2, 3, 4, 5, 6\}) = \frac{5}{6}.$$

Example: A Fair Dice

The product σ-algebra is given by:

$$\mathcal{F}_X \otimes \mathcal{F}_Y = \{\phi, \{a\}, \{b\}, \{a, b\}\} \ \times \ \{\phi, \{c\}, \{d\}, \{c, d\}\}$$

For instance, consider $\{a\} \times \{c\} \in \mathcal{F}_X \otimes \mathcal{F}_Y$. Since

$$X^{-1}(\{a\}) = \{1, 3, 5\}, \quad Y^{-1}(\{c\}) = \{1\}$$

we have

$$(X, Y)^{-1}(\{a\} \times \{c\}) = X^{-1}(\{a\}) \cap Y^{-1}(\{c\}) = \{1\}.$$

Therefore the joint probability of $\{a\} \times \{c\}$ is

$$P_{(X,Y)}(\{a\} \times \{c\}) = P((X, Y)^{-1}(\{a\} \times \{c\})) = P(\{1\}) = 1/6.$$

Joint Probability Density Function

A related key concept is joint probability density functions.

- Let ν_X be a base measure on $(\Omega_X, \mathcal{F}_X)$.

- Let ν_Y be a base measure on $(\Omega_Y, \mathcal{F}_Y)$.

- Define $\nu_X \otimes \nu_Y$ as the product measure of $\nu_X \otimes \nu_Y$: i.e., a measure on $(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y)$ such that

$$\nu_X \otimes \nu_Y(A \times B) = \nu_X(A)\nu_Y(B), \quad A \in \mathcal{F}_X,\ B \in \mathcal{F}_Y.$$

For instance, assume that

- $\Omega_X = \mathbb{R}^p$ and ν_X is the Lebesgue measure on $\mathbb{R}^p$.

- $\Omega_Y = \mathbb{R}^q$ and ν_Y is the Lebesgue measure on $\mathbb{R}^q$.

Then, $\Omega_X \times \Omega_Y = \mathbb{R}^{p+q}$ and $\nu_X \otimes \nu_Y$ is the Lebesgue measure on $\mathbb{R}^{p+q}$.

Joint Probability Density Function

If the joint distribution $P_{(X,Y)}$ has a probability density function

$$p_{(X,Y)} : \Omega_X \times \Omega_Y \to [0, \infty)$$

with respect to the base measure $\nu_X \otimes \nu_Y$ such that

$$P_{(X,Y)}(S) = \int_S p_{(X,Y)}(x, y) d\nu_X \otimes \nu_Y(x, y), \quad \forall S \in \mathcal{F}_X \otimes \mathcal{F}_Y,$$

then we call $p_{(X,Y)}$ the joint probability density function of X and Y.

In particular, for $S = A \times B$ with $A \in \mathcal{F}_X$ and $B \in \mathcal{F}_Y$, the joint density function satisfies

$$P_{(X,Y)}(A \times B) = \int_B \int_A p_{(X,Y)}(x, y) d\nu_X(x) d\nu_Y(y),$$

Joint Probability Density Function

We look at some important properties of the joint density function.

- Assume that P_X has a density function $p_X : \Omega_X \to \mathbb{R}$ with respect to the base measure ν_X.

- Then we have

$$p_X(x) = \int_{\Omega_Y} p_{(X,Y)}(x, y) dP_Y(y), \quad x \in \Omega_X.$$

This operation is called the marginalization of Y, or the sum rule.

- This can be shown as follows.

Joint Probability Density Function

- Define $f(x) := \int_{\Omega_Y} p_{(X,Y)}(x,y)d\nu_Y(y)$.
- Then for any $A \in \mathcal{F}_X$, this function satisfies

$$\int_A f(x)d\nu_X(x) = \int_A \int_{\Omega_Y} p_{(X,Y)}(x,y)d\nu_Y(y)d\nu_X(x)$$
$$= P_{(X,Y)}(A \times \Omega_Y) = P_X(A).$$

- Thus, f defined here satisfies the definition of a density function of P_X.

Similarly, we have

$$p_Y(y) = \int_{\Omega_X} p_{(X,Y)}(x,y)dP_X(x), \quad y \in \Omega_Y.$$

Outline

Conditional Probabilities and Conditional Distributions

Another important concept is conditional probabilities and conditional distributions.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable with probability space $(\Omega_X, \mathcal{F}_X, P_X)$

- Let $Y : \Omega \to \Omega_Y$ be a random variable with probability space $(\Omega_Y, \mathcal{F}_Y, P_Y)$

Conditional Probabilities

- Take a measurable set $A \in \mathcal{F}_X$, which is a certain statement for which a probability $P_X(A)$ is defined.

- Take a measurable set $B \in \mathcal{F}_Y$, which is another statement regarding the random variable Y.

We are interested in the probability of B being true, given that the statement A is true.

This is the conditional probability of B given A, which we write

$$P_{Y|X}(B|A), \quad A \in \mathcal{F}_X,\ B \in \mathcal{F}_Y.$$

Conditional Probabilities

Statement A can be expressed in the probability space $(\Omega, \mathcal{F}, P)$ as

$$X^{-1}(A) := \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{F}.$$

- Thus, "A is true" can be interpreted as "$X^{-1}(A)$ is true" in the original probability space $(\Omega, \mathcal{F}, P)$.

- Therefore, the condition that "A is true" can be formulated as the restriction of the probability space $(\Omega, \mathcal{F}, P)$ onto $X^{-1}(A) \in \mathcal{F}$.

Conditional Probabilities

- i.e., we consider the restricted probability space

$$(\Omega_{|X^{-1}(A)},\ \mathcal{F}_{|X^{-1}(A)},\ P_{|X^{-1}(A)}),$$

defined with

- Restricted sample space: $\Omega_{|X^{-1}(A)} := X^{-1}(A)$;

- Restricted σ-algebra

$$\mathcal{F}_{|X^{-1}(A)} := \{S \cap X^{-1}(A) \mid S \in \mathcal{F}\} \subset \mathcal{F}.$$

- Restricted probability measure:

$$P_{|X^{-1}(A)}(C) := \frac{P(C)}{P(X^{-1}(A))}, \quad C \in \mathcal{F}_{|X^{-1}(A)}.$$

Conditional Probabilities

- Here, we assumed that $P_X(A) = P(X^{-1}(A)) > 0$, i.e., the statement A has a non-zero probability.

- The division by $P(X^{-1}(A))$ is needed to ensure

$$P_{|X^{-1}(A)}(X^{-1}(A)) = \frac{P(X^{-1}(A))}{P(X^{-1}(A))} = 1.$$

- We can check that $(\Omega_{|X^{-1}(A)}, \mathcal{F}_{|X^{-1}(A)}, P_{|X^{-1}(A)})$ satisfies the definition of a probability space (exercise).

Conditional Probabilities

- Take a statement $B \in \mathcal{F}_Y$, which is expressed as $Y^{-1}(B) \in \mathcal{F}$ in the original probability space $(\Omega, \mathcal{F}, P)$.

- In the restricted probability space $(\Omega_{|X^{-1}(A)}, \mathcal{F}_{|X^{-1}(A)}, P_{|X^{-1}(A)})$, $Y^{-1}(B)$ is expressed as

$$Y^{-1}(B) \cap X^{-1}(A) \in \mathcal{F}_{|X^{-1}(A)}.$$

- Thus, the conditional probability of B given A is defined by

$$\begin{aligned}
P_{Y|X}(B|A) &:= P_{|X^{-1}(A)}(Y^{-1}(B) \cap X^{-1}(A)) \\
&= \frac{P(Y^{-1}(B) \cap X^{-1}(A))}{P(X^{-1}(A))} \\
&= \frac{P_{(X,Y)}(A \times B)}{P_X(A)}.
\end{aligned}$$

Example: A Fair Dice

Let's consider a fair dice.

- **Sample space:** $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- σ-**algebra:** $\mathcal{F} = 2^{\Omega}$ (the power set, i.e., the set of all subsets of Ω).
- **Probability:** $P(\{1\}) = P(\{2\}) \cdots = P(\{6\}) = 1/6$.

Define random variables

$X : \Omega \to \Omega_X := \{a, b\}$: $\quad (\mathcal{F}_X := 2^{\{a,b\}})$

$$X(\omega) := \begin{cases} a & \text{if } \omega \text{ is odd } (i.e., 1, 3, 5) \\ b & \text{if } \omega \text{ is even } (i.e., 2, 4, 6) \end{cases}$$

$Y : \Omega \to \Omega_Y := \{c, d\}$: $\quad (\mathcal{F}_Y := 2^{\{c,d\}})$

$$Y(\omega) := \begin{cases} c & \text{if } \omega = 1 \\ d & \text{if } \omega = 2, 3, 4, 5, 6 \end{cases}$$

Example: A Fair Dice

Let's consider conditioning with $A := \{a\} \in \mathcal{F}_X$.

Then

$$X^{-1}(\{a\}) = \{1, 3, 5\}, \quad P(X^{-1}(\{a\})) = 1/2.$$

The restricted sample space is

$$\Omega_{X^{-1}(\{a\})} = X^{-1}(\{a\}) = \{1, 3, 5\}.$$

The restricted σ-algebra is

$$\mathcal{F}_{X^{-1}(\{a\})} = \{S \cap \{1, 3, 5\} \mid S \in 2^{\{1,2,3,4,5,6\}}\} = 2^{\{1,3,5\}}.$$

The restricted probability measure is

$$P_{X^{-1}(\{a\})} := \frac{P(C)}{P(\{1, 3, 5\})}, \quad C \in \mathcal{F}_{X^{-1}(\{a\})}.$$

Example: A Fair Dice

Therefore, the conditional probability of $B \in \mathcal{F}_Y$ given $\{a\}$ is

$$P_{Y|X}(B \mid \{a\}) = \frac{P(Y^{-1}(B) \cap \{1,3,5\})}{P(\{1,3,5\})}.$$

For instance, since $Y^{-1}(\{c\}) = \{1\}$,

$$P_{Y|X}(\{c\} \mid \{a\}) = \frac{P(\{1\} \cap \{1,3,5\})}{P(\{1,3,5\})} = \frac{P(\{1\})}{P(\{1,3,5\})} = 1/3.$$

Similarly, since $Y^{-1}(\{d\}) = \{2,3,4,5,6\}$,

$$P_{Y|X}(\{d\} \mid \{a\}) = \frac{P(\{2,3,4,5,6\} \cap \{1,3,5\})}{P(\{1,3,5\})} = \frac{P(\{3,5\})}{P(\{1,3,5\})} = 2/3.$$

Conditional Distributions

By construction,

$$P_{Y|X}(\cdot \mid A)$$

defines the probability distribution (measure) on $(\Omega_Y, \mathcal{F}_Y)$.

- This is the conditional distribution of Y given the statement A about X.

- In the case $A = \Omega_X$, $P_{Y|X}(\cdot \mid \Omega_X)$ is the marginal distribution of Y, i.e., P_Y:

In fact, since $X^{-1}(\Omega_X) = \Omega$

$$P_{Y|X}(B|\Omega_X) = \frac{P(Y^{-1}(B) \cap X^{-1}(\Omega_X))}{P(X^{-1}(\Omega_X))}$$

$$= \frac{P(Y^{-1}(B) \cap \Omega)}{P(\Omega)} = P(Y^{-1}(B)) = P_Y(B).$$

Conditional Distributions

- In the case where A is a singleton set, i.e.,

$$A = \{x\} \text{ for some } x \in \Omega_X,$$

then $P_{Y|X}(\cdot|\{x\})$ is usually called "the conditional distribution of Y given $X = x$."

- In this case we implicitly assume $P_X(\{x\}) > 0$, since otherwise the division

$$P_{Y|X}(B \mid \{x\}) = \frac{P_{(X,Y)}(\{x\} \times B)}{P_X(\{x\})}$$

is not well-defined.

- However, $P_X(\{x\}) = 0$ occurs in many important settings, in particular when X is a continuous random variable.

Conditional Distributions: Abstract Definition

- Thus, the conditional distribution of Y given $X = x$ is defined in a rather abstract way, as follows.

- For each $x \in \Omega_X$, let $P_{Y|X=x}$ be a probability distribution on $(\Omega_Y, \mathcal{F}_Y)$.

- Assume that for any $B \in \mathcal{F}_Y$, a function $f_B : \Omega_X \to \mathbb{R}$ defined by

$$f_B(x) := P_{Y|X=x}(B), \quad x \in \Omega_X$$

is a measurable function.

- Then $P_{Y|X=x}$ is called the conditional distribution of Y given $X = x$, if it satisfies

$$P_Y(B) = \int_{\Omega_X} P_{Y|X=x}(B) dP_X(x), \quad \forall B \in \mathcal{F}_Y.$$

Conditional Distributions: Abstract Definition

- For instance, assume that X is a discrete random variable and that $P_X(\{x\}) > 0$ for all $x \in \Omega_X$.

- In this case, the conditional probability of $B \in \mathcal{F}_Y$ given $\{x\} \in \mathcal{F}_X$ is given by

$$P_{Y|X}(B \mid \{x\}) = \frac{P_{(X,Y)}(\{x\} \times B)}{P_X(\{x\})}.$$

- This is consistent with the above abstract definition, since

$$\begin{aligned}
&\int_{\Omega_X} \frac{P_{(X,Y)}(\{x\} \times B)}{P_X(\{x\})} dP_X(x) \\
&= \sum_{x \in \Omega_X} \frac{P_{(X,Y)}(\{x\} \times B)}{P_X(\{x\})} P_X(\{x\}) \\
&= \sum_{x \in \Omega_X} P_{(X,Y)}(\{x\} \times B) = P_{(X,Y)}(\Omega_X, B) = P_Y(B).
\end{aligned}$$

Conditional Probability Density Functions

Assume that

- the joint distribution $P_{(X,Y)}$ has a density function

$$p_{(X,Y)} : \Omega_X \times \Omega_Y \to [0, \infty)$$

with respect to the base measure $\nu_X \otimes \nu_Y$:

$$\begin{aligned} P_{(X,Y)}(A \times B) &= \int_{A \times B} p_{(X,Y)}(x, y) d\nu_X \otimes \nu_Y(x, y) \\ &= \int_B \int_A p_{(X,Y)}(x, y) d\nu_X(x) d\nu_Y(y) \end{aligned}$$

- the marginal distribution P_X has a density function

$$p_X : \Omega_X \to [0, \infty)$$

such that

$$P_X(A) = \int_A p_X(x) d\nu_X(x)$$

Conditional Probability Density Functions

Then, the conditional probability density function

$$p_{Y|X} : \Omega_X \times \Omega_Y \to [0, \infty)$$

is defined by

$$p_{Y|X}(y \mid x) := \frac{p_{(X,Y)}(x, y)}{p_X(x)}, \quad x \in \Omega_X, \; y \in \Omega_Y,$$

assuming $p_X(x) > 0$.

- For each $x \in \Omega_X$, $p_{Y|X}(\cdot \mid x)$ is a probability density function on Ω_Y: In fact,

$$\int_{\Omega_Y} p_{Y|X}(y \mid x) d\nu_Y(y) = \int_{\Omega_Y} \frac{p_{(X,Y)}(x, y)}{p_X(x)} d\nu_Y(y)$$
$$= \frac{1}{p_X(x)} \int_{\Omega_Y} p_{(X,Y)}(x, y) d\nu_Y(y) = \frac{p_X(x)}{p_X(x)} = 1,$$

where we used the sum rule.

Conditional Probability Density Functions

Conditional density function $p_{Y|X}(\cdot \mid x)$ is consistent with the abstract definition of conditional distributions.

To see this, let

$$P_{Y|X=x}(B) := \int_B p_{Y|X}(y \mid x) d\nu_Y(y), \quad B \in \mathcal{F}_Y.$$

Then

$$\int_{\Omega_X} P_{Y|X=x}(B) dP_X(x) = \int_{\Omega_X} \int_B p_{Y|X}(y \mid x) d\nu_Y(y) dP_X(x)$$

$$= \int_{\Omega_X} \int_B \frac{p_{(X,Y)}(x,y)}{p_X(x)} d\nu_Y(y) dP_X(x)$$

$$= \int_B \int_{\Omega_X} \frac{p_{(X,Y)}(x,y)}{p_X(x)} p_X(x) d\nu_X(x) d\nu_Y(y)$$

$$\int_B \int_{\Omega_X} p_{(X,Y)}(x,y) d\nu_X(x) d\nu_Y(y) = \int_B p_Y(y) d\nu_Y(y) = P_Y(B),$$

where we used the sum rule.

Outline

Independence of Random Variables

The independence of random variables is another key concept in probability and statistics.

This characterizes "unrelatedness" of two random variables.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable, with the associated probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

- Let $Y : \Omega \to \Omega_Y$ be a random variable, with the associated probability space $(\Omega_Y, \mathcal{F}_Y, P_Y)$.

- Let

$$\left(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y, P_{(X,Y)}\right)$$

be the joint probability space.

Independence of Random Variables

Definition. Random variables X and Y are called independent if

$$P_{(X,Y)}(A \times B) = P_X(A)P_Y(B), \quad \forall A \in \mathcal{F}_X, \ \forall B \in \mathcal{F}_Y.$$

In other words, X and Y are independent, if the joint distribution $P_{(X,Y)}$ is equal to the product measure, i.e.,

$$P_{(X,Y)} = P_X \otimes P_Y,$$

where $P_X \otimes P_Y$ is the product measure such that

$$P_X \otimes P_Y(A \times B) = P_X(A)P_Y(B).$$

Independence of Random Variables

- If there exists a joint density function

$$p_{X,Y} : \Omega_X \times \Omega_Y \to [0, \infty),$$

then X and Y are independent if

$$p_{(X,Y)}(x, y) = p_X(x) p_Y(y), \quad \forall x \in \Omega_X, \ \forall y \in \Omega_Y,$$

where

- $p_X : \Omega_X \to [0, \infty)$ is the density function of X

- $p_Y : \Omega_Y \to [0, \infty)$ is the density function of Y

Exercise:

- Show that the above characterization leads to the definition of independence.

Independence of Random Variables: An Interpretation

The independence of X and Y implies that X does not have any information about Y (and vice versa).

- In fact, the independence implies that the conditional probability $P_{Y|X}(A|B)$ is equal to the marginal $P_Y(B)$:

$$P_{Y|X}(B|A) := \frac{P_{(X,Y)}(A \times B)}{P_X(A)} = \frac{P_X(A)P_Y(B)}{P_X(A)} = P_Y(B).$$

- Similarly, if there exist density functions, the conditional density function $p_{Y|X}(y|x)$ equals the marginal density function p_Y(y):

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

Intuitively, this means that the conditioning $X = x$ does not affect the distribution of Y.

Consequences of Independence

Let $f : \Omega_X \to \mathbb{R}$ and $g : \Omega_Y \to \mathbb{R}$ be any measurable functions.

Then, if X and Y are independent, we have

$$\mathbb{E}_{X,Y}[f(X)g(Y)] = \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)]$$

This is because, since $P_{(X,Y)} = P_X \otimes P_Y$ by the independence,

$$\begin{aligned}
\mathbb{E}_{(X,Y)}[f(X)g(Y)] &:= \int_{\Omega_X \times \Omega_Y} f(x)g(y)dP_{(X,Y)}(x,y) \\
&= \int_{\Omega_X} \int_{\Omega_Y} f(x)g(y)dP_Y(y)dP_X(x) \\
&= \int_{\Omega_X} f(x)dP_X(x) \int_{\Omega_Y} g(y)dP_Y(x) \\
&= \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)].
\end{aligned}$$

Independently and Identically Distributed (i.i.d.)

The i.i.d. is a very important concept, ubiquitous in statistics and machine learning.

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Let $X : \Omega_X \to \Omega$ be a random variable with probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

Consider n random variables $X_1, X_2, \ldots, X_n$ such that

- $X_i : \Omega \to \Omega_{X_i}$ is a random variable with probability space $(\Omega_{X_i}, \mathcal{F}_{X_i}, P_{X_i})$ $(i = 1, \ldots, n)$.

Independently and Identically Distributed (i.i.d.)

Definition. Random variables $X_1, X_2, \ldots, X_n$ independently and identically distributed (i.i.d.) with X (or with P_X), if they satisfy the following:

– $(\Omega_{X_i}, \mathcal{F}_{X_i}, P_{X_i}) = (\Omega_X, \mathcal{F}_X, P_X)$ $(i = 1, \ldots, n)$.

- i.e., X_i is identically distributed with X.

– X_i and X are independent $(i = 1, \ldots, n)$.

- i.e., X_i is independently distributed with X.

– X_i and X_j are independent for $i \neq j$.

- i.e., X_i is independently (and identically) distributed with X_j.

We often write as $X_1, \ldots, X_n \sim P$ (i.i.d.).

Outline

Some Important Points to Remember

When talking about random variables, people often omit mentioning the underlying probability space $(\Omega, \mathcal{F}, P)$.

- Often, we say something like "Let X be a random variable taking values in Ω_X".

- But remember that there is always such a probability space.

In this lecture, I will use a different notation like $\mathcal{X}$ in place of Ω_X.

- So, "Let X be a random variable taking values in $\mathcal{X}$"

should be understood as

"Let $(P, \mathcal{F}, \Omega)$ be a probability space, and let $X : \Omega \to \Omega_X$ be a random variable with probability space $(P_X, \mathcal{F}_X, \Omega_X)$ with $\mathcal{X} := \Omega_X$."

Further Reading

If you are interested in more detail of probability theory, you may look at the books in the references.

Specifically:

- [Dudley, 2002, Chapters 3, 4, 8, 10].

- [Rao, 1973, Chapter 2].

Dudley, R. M. (2002).
Real Analysis and Probability.
Cambridge University Press.

Rao, C. R. (1973).
Linear Statistical Inference and Its Applications.
Wiley New York.