Fundamentals of Probability Theory
- Notes -

Lim Zi Xiang

August 11, 2025

Contents

Chapter 1

Random variables and random vectors

1.1 Random vectors

Definition 1.1.1: Random vector

Let Ω be a sample space. A **random vector $\boldsymbol{X}$** is a function from the sample space Ω to the set of K-dimensional real vectors $\mathbb{R}^K$:

$$\boldsymbol{X} : \Omega \to \mathbb{R}^K.$$

To put it simply, a random vector is a vector whose value depends on the outcome of the probabilistic experiment. The real vector $\boldsymbol{X}(\omega)$ associated to a sample point $\omega \in \Omega$ is a **realisation** of the random vector. The set of all possible realisations is the **support**, denoted R_X.

Note:-

We denote the probability of an event $E \subseteq \Omega$ by $P(E)$. We use the following conventions when dealing with random vectors:

- For $A \subseteq \mathbb{R}^K$, $P(\boldsymbol{X} \in A) = P(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in A\})$.
- For $A \subseteq \mathbb{R}^K$, $P_{\boldsymbol{X}}(A) = P(\boldsymbol{X} \in A)$.
  It is very common in applied work to build statistical models where a random vector X is defined by directly specifying P_X and omitting the specification of the sample space Ω.
- We often write $\boldsymbol{X}$ to mean $\boldsymbol{X}(\omega)$.

Example 1.1.1 (Defining a random vector on a sample space)

Two coins are toseed. The possible outcomes of each toss can be either tail (T) or head (H). The sample space is

$$\Omega = \{TT, TH, HT, HH\}.$$

The four possible outcomes are assigned equal probabilities:

$$P(\{TT\}) = P(\{TH\}) = P(\{HT\}) = P(\{HH\}) = \frac{1}{4}.$$

If the outcome is tails, we win a dollar, otherwise we lose one dollar. A 2D random vector $\boldsymbol{X}$ indicates the amount we win on each toss:

$$\boldsymbol{X}(\omega) = \begin{cases} \begin{pmatrix} 1 & 1 \end{pmatrix} & \text{if } \omega = TT \\ \begin{pmatrix} 1 & -1 \end{pmatrix} & \text{if } \omega = TH \\ \begin{pmatrix} -1 & 1 \end{pmatrix} & \text{if } \omega = HT \\ \begin{pmatrix} -1 & -1 \end{pmatrix} & \text{if } \omega = HH \end{cases}$$

The probability of winning one dollar on both tosses is

$$P\left(\boldsymbol{X} = \begin{pmatrix} 1 & 1 \end{pmatrix}\right) = P\left(\{\omega \in \Omega : \boldsymbol{X}(\omega) = \begin{pmatrix} 1 & 1 \end{pmatrix}\}\right) = P(\{TT\}) = \frac{1}{4}.$$

The probability of losing one dollar on the second toss is

$$P(X_2 = -1) = P(\{\omega \in \Omega : X_2(\omega) = -1\}) = P(\{TH, HH\}) = \frac{1}{2}.$$

1.1.1 Discrete random vectors

Definition 1.1.2: Discrete random vector

A random vector $\boldsymbol{X}$ is **discrete** iif

1. its support $R_{\boldsymbol{X}}$ is a countable set;
2. there is a function $p_{\boldsymbol{X}} : \mathbb{R}^K \to [0, 1]$, called the **joint probability mass function** of $\boldsymbol{X}$, such that for any $\boldsymbol{x} \in \mathbb{R}^K$:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} P(\boldsymbol{X} = \boldsymbol{x}) & \text{if } \boldsymbol{x} \in R_{\boldsymbol{X}}; \\ 0 & \text{if } \boldsymbol{x} \notin R_{\boldsymbol{X}}. \end{cases}$$

Note:-

The following are equivalent notations used interchangeably to indicate the joint pmf:

$$p_{\boldsymbol{X}}(x) = p_{\boldsymbol{X}}(x_1, \ldots, x_K) = p_{X_1,\ldots,X_k}(x_1, \ldots, x_K).$$

Example 1.1.2

Suppose $\boldsymbol{X}$ is a 2D random vector whose components (X_1 and X_2) can take only two values: 1 or 0, and the four possible combinations of 0 and 1 are equally likely. The support of the discrete vector $\boldsymbol{X}$ is

$$R_{\boldsymbol{X}} = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}.$$

Its pmf is

$$p_{\boldsymbol{X}} = \begin{cases} 0.25 & \text{if } x = \begin{pmatrix} 1 & 1 \end{pmatrix}^T; \\ 0.25 & \text{if } x = \begin{pmatrix} 1 & 0 \end{pmatrix}^T; \\ 0.25 & \text{if } x = \begin{pmatrix} 0 & 1 \end{pmatrix}^T; \\ 0.25 & \text{if } x = \begin{pmatrix} 0 & 0 \end{pmatrix}^T; \\ 0 & \text{otherwise.} \end{cases}$$

1.1.2 Continuous random vectors

Definition 1.1.3: Continuous random vector

A random vector $\boldsymbol{X}$ is **continuous** (or **absolutely continuous**) iif

1. its support $R_{\boldsymbol{X}}$ is uncountable;
2. there exists a function $f_{\boldsymbol{X}} : \mathbb{R}^K \to [0, \infty]$, called the **joint probability density function** of $\boldsymbol{X}$, such that for any set $A \subseteq \mathbb{R}^K$ where

$$A = [a_1, b_1] \times \ldots \times [a_K, b_K].$$

The probability that $\boldsymbol{X} \in A$ is calculated by

$$P(\boldsymbol{X} \in A) = \int_{a_1}^{b_1} \ldots \int_{a_K}^{b_K} f_{\boldsymbol{X}}(x_1, \ldots, x_K) dx_K \ldots dx_1$$

provided the multiple integral is well defined.

Example 1.1.3

Suppose $\boldsymbol{X}$ is a 2D random vector whose components X_1 and X_2 are independent uniform random variables on the interval $[0, 1]$. Then, $\boldsymbol{X}$ is an example of a continuous vector with support

$$R_{\boldsymbol{X}} = [0, 1] \times [0, 1].$$

Its joint pmf is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in [0, 1] \times [0, 1]; \\ 0 & \text{otherwise.} \end{cases}$$

The probability that the realisation of $\boldsymbol{X}$ falls in the rectangle $[0, 0.5] \times [0, 0.5]$ is

$$\begin{aligned} P(\boldsymbol{X} \in [0, 0.5] \times [0, 0.5]) &= \int_0^{0.5} \int_0^{0.5} f_{\boldsymbol{X}}(x_1, x_2) dx_2 dx_1 \\ &= \int_0^{0.5} \int_0^{0.5} (1) dx_2 dx_1 \\ &= \int_0^{0.5} [x_2]_0^{0.5} dx_1 \\ &= \int_0^{0.5} 0.5 dx_1 \\ &= [0.5x_1]_0^{0.5} \\ &= 0.25 \end{aligned}$$

1.1.3 Random vectors in general

Definition 1.1.4: Joint distribution function

Let $\boldsymbol{X}$ be a random vector. The **joint (cumulative) distribution function** of $\boldsymbol{X}$ is a function $F_{\boldsymbol{X}} : \mathbb{R}^K \to [0,1]$ such that

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_K \leq x_K), \forall \boldsymbol{x} \in \mathbb{R}^K,$$

where the components of $\boldsymbol{X}$ and $\boldsymbol{x}$ are denoted by X_k and x_k respectively, for $k = 1, \ldots, K$.

Similarly for the case of joint pmf/pdf, the following notations are used interchangeably to indicate the joint cdf:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{\boldsymbol{X}}(x_1, \ldots, x_K) = F_{X_1,\ldots,X_K}(x_1, \ldots, x_K).$$

1.1.4 Joint distribution

Sometimes we talk about the **joint distribution** of a random vector without specifying whether we mean the joint cdf, pmf, or pdf. And this is justified, since the joint pmf/pdf completely determines and is complete determined by the joint cdf of a discrete/continuous vector.

1.1.5 Random matrices

Definition 1.1.5: Random matrix

A random matrix is a matrix whose entries are random variables.

A random matrix can always be written as a random vector by vectorising it: given a $K \times L$ random matrix $\boldsymbol{A}$, its vectorisation, denoted vec$(\boldsymbol{A})$ is the $KL \times 1$ random vector obtained by stacking the columns of $\boldsymbol{A}$ on top of each other.

Example 1.1.4

Let $\boldsymbol{A}$ be the following 2×2 random matrix:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

The vectorisation of $\boldsymbol{A}$ is the following 4×1 vector:

$$\text{vec}(\boldsymbol{A}) = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{pmatrix}.$$

If vec$(\boldsymbol{A})$ is a discrete/continuous vector, then $\boldsymbol{A}$ is a **discrete/continuous random matrix**, the joint pmf of $\boldsymbol{A}$ is just the joint pmf/pdf of vec$(\boldsymbol{A})$.

1.1.6 The marginal distribution of a random vector

Let X_i be the i-th component of a K-dimensional random vector $\boldsymbol{X}$. The cdf $F_{X_i}(\boldsymbol{x})$ of X_i is the marginal distribution function of X_i.

If $\boldsymbol{X}$ is discrete/continuous, then X_i is a discrete/continuous random variable and its pmf $p_{X_i}(\boldsymbol{x})$/pdf $f_{X_i}(\boldsymbol{x})$ is the **marginal pmf/pdf** of X_i.

1.1.7 Marginalisation of a joint distribution

Marginalisation is the process of deriving the distribution of a component X_i of a random vector $\boldsymbol{X}$ from the joint distribution of $\boldsymbol{X}$.

It can also have a broader meaning of deriving the joint distribution of a subset of the set of components of $\boldsymbol{X}$ from the joint distribution of $\boldsymbol{X}$, e.g. if $\boldsymbol{X}$ has three components X_1, X_2, X_3, we can marginalise their joint distribution to find the joint distribution of X_1 and X_2. In this case, X_3 is said to be marginalised out of the joint distribution of X_1, X_2, and X_3.

1.1.8 The marginal distribution of a discrete vector

Let X_i be the i-th component of a K-dimensional discrete random vector $\boldsymbol{X}$. The marginal pmf of X_i is derived from the joint pmf by:

$$p_{X_i}(x) = \sum_{(x_1,\ldots,x_K)\in R_{\boldsymbol{X}}:x_i=x} p_{\boldsymbol{X}}(x_1, \ldots, x_K),$$

where the sum is over the set

$$\{(x_1, \ldots, x_K) \in R_{\boldsymbol{X}} : x_i = x\},$$

i.e. the probability that $X_i = x$ is obtained as the sum of the probabilites of all the vectors in $R_{\boldsymbol{X}}$ such that their i-th component is equal to x.

1.1.9 Marginalisation of a discrete distribution

Let X_i be the i-th component of a K-dimensional discrete random vector $\boldsymbol{X}$. By marginalising $\boldsymbol{X}_i$ out of the joint distribution of $\boldsymbol{X}$, we obtain the joint distribution of the remaining components of $\boldsymbol{X}$, $\boldsymbol{X}_{-i}$:

$$\boldsymbol{X}_{-i} = \begin{pmatrix} X_1 & \ldots & X_{i-1} & X_{i+1} & \ldots & X_K \end{pmatrix}.$$

The joint pmf of $\boldsymbol{X}_{-i}$ is computed as follows:

$$p_{\boldsymbol{X}_{-i}}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_K) = \sum_{x_i\in R_{X_i}} p_{\boldsymbol{X}}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_K),$$

i.e. the joint pmf of $\boldsymbol{X}_{-i}$ is computed by summing the joint pmf od $\boldsymbol{X}$ over all values of x_i that belong to the support of X_i.

1.1.10 The marginal distribution of a continuous vector

Let X_i be the i-th component of a K-dimensional continuous random vector $\boldsymbol{X}$. The **marginal pdf** of X_i is derived from the joint pdf of $\boldsymbol{X}$ by

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_K) dx_K \ldots dx_{i+1} dx_{i-1} dx_1,$$

i.e. the joint pdf evaluated at $x_i = x$ is integrated with respect to all variables except x_i.

1.1.11 Marginalisation of a continuous distribution

Let X_i be the i-th component of a continuous random vector $\boldsymbol{X}$. By marginalising X_i out of the joint distribution of $\boldsymbol{X}$, we obtain the joint distribution of the remaining components of $\boldsymbol{X}$, $\boldsymbol{X}_{-i}$.

The joint pdf of $\boldsymbol{X}_{-i}$ is then computed by

$$f_{\boldsymbol{X}_{-i}}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_K) = \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_K) dx_i,$$

i.e. the joint pdf of $\boldsymbol{X}_{-i}$ is computed by integrating the joint pdf of $\boldsymbol{X}$ with respect to x_i.

1.1.12 Partial derivatives of the distribution function of a continuous vector

We know that if $\boldsymbol{X}$ is continuous, then

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_K} f_{\boldsymbol{X}}(t_1, \ldots, t_K) dt_K \ldots dt_1.$$

So by taking the K-th order cross-partial derivative with respect to $x_1, \ldots, x_K$ of both sides of the above equation, we get

$$\frac{\partial^K F_{\boldsymbol{X}}(\boldsymbol{x})}{\partial x_1 \ldots, \partial x_K} = f_{\boldsymbol{X}}(\boldsymbol{x}).$$

1.1.13 A more rigorous definition of random vectors

The following is a more rigorous definition of random vector using the formalism of measure theory. (I'll ignore this part for the time being.)

Definition 1.1.6

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\mathcal{B}(\mathbb{R}^K)$ be the Borel sigma-algebra of $\mathbb{R}^K$ (i.e. the smallest sigma-algebra containing all open hyper-rectangles in $\mathbb{R}^K$). A function $\boldsymbol{X} : \Omega \to \mathbb{R}^K$ such that

$$\{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\} \in \mathcal{F}$$

for any $B \in \mathcal{B}(\mathbb{R}^K)$ is said to be a random vector on Ω.

This definition ensures that the probability that the realisation of $\boldsymbol{X}$ belongs to a set $B \in \mathcal{B}(\mathbb{R}^K)$ that can be defined as

$$P(\boldsymbol{X} \in B) := P(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\})$$

because the set $\{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\}$ belongs to the sigma-algebra $\mathcal{F}$ and hence its probability is well-defined.

1.1.14 Exercises

Question 1

Let $\boldsymbol{X}$ be a 2×1 discrete random vector and denote its components by X_1 and X_2.
Let the support of X be the set of all 2×1 vectors such that their entries belong to the set of the first three natural numbers, that is,

$$R_X = \{\boldsymbol{x} = \begin{pmatrix} x_1 & x_2 \end{pmatrix}^T : x_1 \in N_3 \text{ and } x_2 \in N_3\}.$$

1.2 Expected value

In this section we give an informal definition of expected value. A formal definition involves the Lebesgue integral which I will ignore for the time being.

Definition 1.2.1: Expected value

The **expected value** of a random variable $\boldsymbol{X}$, $E[X]$ is the weighted average of the values that $\boldsymbol{X}$ can take on, where each possible value is weighted by its respective probability.

1.2.1 Expected value of a discrete random variable

Definition 1.2.2: Expected value of a discrete random variable

Let X be a discrete random variable with support $R_{\boldsymbol{X}}$ and pmf $p_X(x)$. The expected value of $\boldsymbol{X}$ is

$$E[X] = \sum_{x \in R_X} x p_X(x),$$

provided that we have **absolute summability**

$$\sum_{x \in R_X} |x| p_X(x) < \infty,$$

ensuring that the summation is well-defined when R_X contained infinitely many elements.

Note:-

When summing infinitely many terms, the order in which you sum them can change the results and the expected value of $\boldsymbol{X}$ is not well-defined or does not exist. However this is not true if the terms are absolutely summable.

Example 1.2.1 (Expected value)

Let $\boldsymbol{X}$ be a random variable with support $R_X = \{0, 1\}$ and pmf

$$p_X(x) = \begin{cases} 0.5 & \text{if } x = 1; \\ 0.5 & \text{if } x = 0; \\ 0 & \text{otherwise.} \end{cases}$$

Its expected value is

$$\begin{aligned} E[X] &= \sum_{x \in R_X} x p_X(x) \\ &= (1)(0.5) + (0)(0.5) \\ &= 0.5 \end{aligned}$$

1.2.2 Expected value of a continuous random variable

Definition 1.2.3: Expected value of a continuous random variable

Let X be a continuous random variable with pdf $f_X(x)$. The expected value of X is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx,$$

provided that we have absolute integrability

$$\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty.$$

Example 1.2.2 (Expected value of continuous random variable)

Let X be a continuous random variable with support $R_X = [0, \infty)$ and pdf

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \in [0, \infty); \\ 0 & \text{otherwise.} \end{cases}$$

where $\lambda > 0$. Its expected value is

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{0}^{\infty} x \lambda(-\lambda x) dx \\ &= \frac{1}{\lambda} \int_{t=0}^{t=\infty} t \exp(-t) dt \\ &= \frac{1}{\lambda} \left\{ [-t \exp(-t)]_{t=0}^{t=\infty} + \int_{0}^{\infty} \exp(-t) dt \right\} \end{aligned}$$

1.2.3 Expected value of a random variable in general: the Riemann-Stieltjes integral

Definition 1.2.4: Expected value (general)

Let X be a random variable with cdf $F_X(x)$. The expected value of X is

$$E[X] = \int_{-\infty}^{\infty} x dF_X(x),$$

where the integral is a Riemann-Stieltjes integral and the expected value exists and is well-defined iif the integral is well-defined.

This definition gives a formal notation which allows for a unified treatment of discrete and continuous random variables and can be treated as a sum in one case and as ordinary Riemann integral in the other.

Example 1.2.3

Let X be a random variable with support $R_X = [0, 1]$ and distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.5x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 0 \end{cases}.$$

Its expected value is

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x dF_X(x) \\ &= \int_0^1 x dF_X(x) + 1 \cdot \left[F_X(1) - \lim_{x \to 1^+} F_X(x) \right] \\ &= \int_0^1 x \frac{d}{dx}(\frac{1}{2}x) dx + 1 \cdot [1 - \frac{1}{2}] \\ &= \left[\frac{1}{4} x^2 \right]_0^1 + \frac{1}{2} \\ &= \frac{3}{4} \end{aligned}$$

1.2.4 Expected value of a random variable in general: the Lebesgue integral

Definition 1.2.5: Expected value (rigorous)

Let Ω be a sample space, P a probability measure defined on the events of Ω and X a random variable defined on Ω. The expected value of X is

$$E[X] = \int X dP,$$

provided the Lebesgue integral of X with respect to P exists and is well-defined.

1.2.5 The transformation theorem

Let X be a random variable, $g : \mathbb{R} \to \mathbb{R}$ be a real function. Define a new random variable Y as $Y = g(X)$. Then,

$$E[Y] = \int_{-\infty}^{\infty} g(x) dF_X(x)$$

provided that the integral exists. For discrete random variables, the formula becomes

$$E[Y] = \sum_{x \in R_X} g(x) p_X(x)$$

while for continuous random variables,

$$E[Y] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

When $\boldsymbol{X}$ is a discrete random vector and $p_{\boldsymbol{X}}(x)$ is its joint pmf, then

$$E[Y] = \sum_{x \in R_X} g(\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}).$$

When $\boldsymbol{X}$ is an continuous random vector and $f_{\boldsymbol{X}}(\boldsymbol{x})$ is its joint pdf, then

$$E[Y] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) dx_1 \dots dx_K.$$

1.2.6 Linearity of the expected value

1.2.7 Expected value of random vectors

1.2.8 Expected value of random matrices

1.2.9 Integrability and Lp spaces

1.2.10 Exercises

1.3 Properties of the expected value

1.3.1 Scalar multiplication of a random variable

1.3.2 Sums of random variables

1.3.3 Linear combination of random variables

1.3.4 Expected value of a constant

1.3.5 Expectation of a product of random variables

1.3.6 Non-linear transformations

1.3.7 Addition of a constant matrix and a matrix with random entries

1.3.8 Multiplication of a constant matrix and a matrix with random entries

1.3.9 Expectation of a positive random variable

1.3.10 Preservation of almost sure inequalities

1.3.11 Exercises

1.4 Variance

1.5 Covariance

1.6 Linear correlation

1.7 Covariance matrix

1.8 Indicator functions

1.9 Quantile

Chapter 2

Conditional distributions and independence

2.0.1 Rigorous conditional probebility