Lecture 2:
Probability and Random Variables

ELE 525: Random Processes in Information Systems

Hisashi Kobayashi

Department of Electrical Engineering
Princeton University
September 18, 2013

Textbook: Hisashi Kobayashi, Brian L. Mark and William Turin, ***Probability, Random Processes and Statistical Analysis*** (Cambridge University Press, 2012)

2 Probability

2.1 Randomness in the Real World

2.1.1 Repeated experiments and statistical regularity

- **Randomly** varying quantities
- In repeated *experiments*, the averages may exhibit **statistical regularity,** and may converge as more trials are made.
- A mathematical model to study random phenomena. This is the domain of **probability and statistics**.

2.1.2 Random experiments and relative frequencies

- A random experiment, a trial and an outcome
- The relative frequency of an outcome *A*

$$f_N(A) = \frac{N(A)}{N} \tag{2.1}$$

- How to interpret

$$f_N(A) \to P[A] \text{ as } N \to \infty. \tag{2.2}$$

2.1.2 Random experiments and relative frequencies-cont'd

- Construct an abstract model of probabilities such that the probabilities behave like the limits of the relative frequencies.

- **The law of large numbers**.

2.2 Axioms of Probability

Note: Axiomatic method in mathematics:

- The oldest and most famous example of axioms is Euclid's axioms in geometry. In his book entitled ***Elements*** , **Euclid** (aka Euclid of Alexandria 330BC?-375BC?) deduced all propositions (or theorems) of what is now called Euclidean geometry from the following five **axioms** (or sometimes called postulates).

Axiom 1: We can draw a straight line segment joining any two points.

Axiom 2: We can extend any straight line segment indefinitely in a straight line.

Axiom 3: We can draw a circle with any point as its center and any distance as its radius.

Axiom 4: Are all right angles are congruent (i.e., equal to each other)
Axiom 5: If two straight lines intersect a third straight line in such a way that the sum of the inner angles on one side is less than two right angles, then the two lines inevitably must intersect each other on that side if extended indefinitely (known as the ***parallel postulate***).

Many mathematicians attempted to deduce Axiom 5 from Axioms 1-4, but failed. Around 1823-1831, the Hungarian mathematician **János Bolyai** (1802-1860) and the Russian mathematician **Nicolai Labachevsky** (1792-1856) independently discovered a **non-Euclidean geometry ,** now known as **hyperbolic geometry**, based entirely on Axioms 1-4.

Carl Friedrich Gauss (1777-1855) had also discovered the existence of non-Euclidean geometry by around 1824, but did not publish his theory for fear of getting involved in a philosophical dispute concerning the notion of space and time held by the German philosopher **Immanuel Kant** (1724=1804).

Kant states that: "Space is not an empirical concept which has been derived from outer experiences." On the contrary: "…it is the subjective condition of sensibility, under which alone outer intuition is possible for us."

2.1.2 Random experiments and relative frequencies-cont'd

- Construct an abstract model of probabilities such that the probabilities behave like the limits of the relative frequencies.

- **The law of large numbers**.

2.2 Axioms of Probability

2.2.1 Sample space

- The **sample space** Ω : a mathematical abstraction of the collection of all possible outcomes.
- A **sample point** ω in Ω: a possible outcome of the experiment.
- **Example 2.1**: Tossing of two coins:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}, \tag{2.3}$$

where

$$\omega_1 = (\mathrm{hh}), \qquad \omega_2 = (\mathrm{ht}), \qquad \omega_3 = (\mathrm{th}), \qquad \omega_4 = (\mathrm{tt}). \tag{2.4}$$

2.2.2 Event

- An **event** is a set of sample points.
- **Example 2.3**: Tossing of two coins

$$A = \{(\mathrm{hh}), (\mathrm{ht}), (\mathrm{th})\}, \tag{2.7}$$

2.2.3 Probability measure

- A **probability measure** is an assignment of real numbers to the events defined on Ω.
- The set of properties that the assignment must satisfy are called the **axioms of probabilities**.
- **A. N. Kolmogorov's paper in 1933**.

2.2.4 Properties of probability measures

- Review of Set Theory:
  - The **complement**, **union**, **intersection**,
  - The **null event,** denoted Ø, and the **sure event,** denoted Ω .
  - **Disjoint events** (or mutually exclusive events)

❖ Review of Set Theory-cont'd

- Commutative law, Associative law, Distributive law and **de Morgan's laws** See (2.14) through (2.18)

- **Venn diagrams** (Figure 2.1) are useful in verifying the above laws.

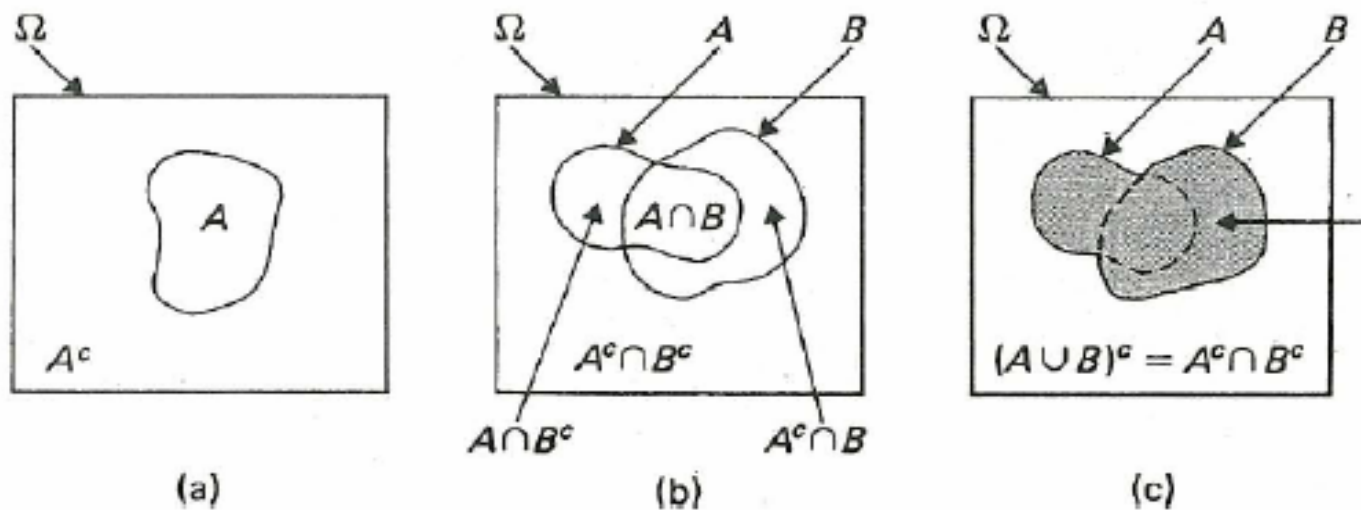

Figure 2.1 Venn diagrams.

❖ Axioms of Probability (A. N. Kolmogorov)

Axiom 1. $P[A] \geq 0$ for all events A. (2.19)

Axiom 2. $P[\Omega] = 1$, that is, the probability of the *sure* event is 1. (2.20)

Axiom 3. If A and B are *mutually exclusive* events — i.e., if $A \cap B = \emptyset$— then $P[A \cup B] = P[A] + P[B]$. (2.21)

A consequence of Axiom 3 is that if $A_1, A_2, \ldots, A_M$ are M mutually exclusive events, then their union $A_1 \cup A_2 \cup \cdots \cup A_M$, which we denote $\bigcup_{m=1}^{M} A_m$, has the probability

$$P\left[\bigcup_{m=1}^{M} A_m\right] = \sum_{m=1}^{M} P[A_m]. \tag{2.22}$$

We can derive

$$\boxed{P[A \cup B] = P[A] + P[B] - P[A \cap B].} \tag{2.27}$$

❖ We must extend Axiom 3 to deal with infinitely many events:

Axiom 4. If $A_1, A_2, \ldots$ are mutually exclusive events, then their union $A_1 \cup A_2 \cup \cdots$, denoted $\bigcup_{m=1}^{\infty} A_m$, has the probability

$$P\left[\bigcup_{m=1}^{\infty} A_m\right] = \sum_{m=1}^{\infty} P[A_m]. \tag{2.29}$$

Because of Axiom 4, we require that the collection of events be ***closed under countable unions:***

Definition 2.1 (σ-field[7]). *A collection $\mathcal{F}$ of subsets of Ω is called a σ-field, if it satisfies the following properties:*

$$\boxed{\begin{aligned} &\text{(a)}\ \ \emptyset \in \mathcal{F};\\ &\text{(b)}\ \ \text{if } A \in \mathcal{F},\ \text{then} A^c \in \mathcal{F};\\ &\text{(c)}\ \ \text{if } A_1, A_2, \ldots \in \mathcal{F},\ \text{then} \bigcup_{m=1}^{\infty} A_m \in \mathcal{F}. \end{aligned}} \tag{2.30}$$

Definition 2.2 (Probability Measure and Probability Space). *A **probability measure** P defined on $(\Omega, \mathcal{F})$ is a function that maps any element of $\mathcal{F}$ into $[0,1]$ such that*

(a) $P[\emptyset] = 0, \quad P[\Omega] = 1$

(b) If $A_1, A_2, \ldots \in \mathcal{F}$ *and* $A_m \cap A_n = \emptyset \ \ (m \neq n)$, *then*

$$P\left[\bigcup_{m=1}^{\infty} A_m\right] = \sum_{m=1}^{\infty} P[A_m]. \tag{2.31}$$

The triple $(\Omega, \mathcal{F}, P)$ *is called a **probability space**.* □

Example 2.9: Tossing of two coins and a product space. (pp. 25-26)

the first coin (coin "1") is denoted $\Omega_1 = \{\mathrm{h}, \mathrm{t}\}$. Its σ-field is $\mathcal{F}_1 = \{\emptyset, \{\mathrm{h}\}, \{\mathrm{t}\}, \Omega_1\}$. A possible probability measure P_1 is given by

$$P_1[\emptyset] = 0, \quad P_1[\{\mathrm{h}\}] = p_1, \quad P_1[\{\mathrm{t}\}] = 1 - p_1, \text{and } P_1[\Omega_1] = 1,$$

The sample space Ω of the experiment of tossing the two coins is the Cartesian product of the two sample spaces defined above:

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \ \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}. \tag{2.32}$$

2.3 Bernoulli trials and Bernoulli's theorem

- Bernoulli trials: repeated independent trials in which there are only two possible outcomes ("success" and "failure"), and their probabilities remain unchanged.

$$\Omega = \{\mathrm{s}, \mathrm{f}\}$$

$$P[\{\mathrm{s}\}] = p, \quad 0 \le p \le 1, \tag{2.34}$$

$$P[\{f\}] = 1 - p \triangleq q, \tag{2.35}$$

(2.34) and (2.35) constitute the **Bernoulli distribution** for a single trial.

- Consider an event *E* = { *k* successes and *n-k* failures in n Bernoulli trials}. The number of sample points in *E*:

$$\binom{n}{k} \triangleq \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k\cdot(k-1)\cdots 2\cdot 1}. \tag{2.37}$$

This number is called the **binomial coefficient**.

Then

$$P[E] = B(k; n, p) \triangleq \binom{n}{k} p^k q^{n-k}, \quad k = 0,\ 1,\ 2,\ \ldots,\ n. \tag{2.38}$$

The set of probabilities $B(k; n, p)$ is called the **binomial distribution**.

$$\frac{B(k; n, p)}{B(k-1; n, p)} = \frac{n! p^k q^{n-k}}{(n-k)!k!} \frac{(n-k+1)!(k-1)!}{n! p^{k-1} q^{n-k+1}} = \frac{n-k+1}{k} \frac{p}{q}. \tag{2.40}$$

Thus, B(k; n, p), as a function of k, increases until

$$k_{\max} = \lfloor (n+1)p \rfloor, \tag{2.41}$$

$$p - \frac{q}{n} < \frac{k_{\max}}{n} \leq p + \frac{p}{n}. \tag{2.43}$$

THEOREM 2.1 (Bernoulli's Theorem (weak law of large numbers)).
For any ε >0,

$$P\left[\left|\frac{k}{n} - p\right| > \epsilon\right] \leq \frac{p(1-p)}{n\epsilon^2}. \tag{2.45}$$

Then

$$\boxed{P\left[\left|\frac{k}{n} - p\right| > \epsilon\right] \to 0, \ \ as\ n \to \infty.} \tag{2.46}$$

THEOREM 2.1 –cont'd

We say that *k/n* ***converges*** to *p* ***in probability***. The above result is called the ***weak law of large numbers***.

Proof. See p. 29 , Eqs. (2.47) through (2.50) .

- The term "law of large numbers (la loi des grands nombres)" was coined by Poisson (1781-1840) to describe the above theorem by Jacob Bernoulli (1654-1705). The theorem can be paraphrased as

$$\boxed{\lim_{n\to\infty} P\left[\left|\frac{k}{n} - p\right| < \epsilon\right] = 1,} \tag{2.51}$$

- Émile **Borel** (1871-1956) and Francesco Paolo **Cantelli** (1875-1966) showed stronger versions of the law of large numbers, i.e., *k/n* converges to *p* not only in probability, but also **with probability 1 (or almost surely)**, i.e.,

$$\boxed{P\left[\lim_{n\to\infty}\left|\frac{k}{n} - p\right| < \epsilon\right] = 1,} \tag{2.52}$$

- A simpler proof of the weak law of large numbers (WLLN) see p. 250 .
- The difference between the WLLN and the strong law of large numbers (SLLN) is equivalent to that between *convergence in probability* and *almost sure convergence* . (see pp. 280-288).

Figure 11.2 illustrates the difference between these two types of convergence.

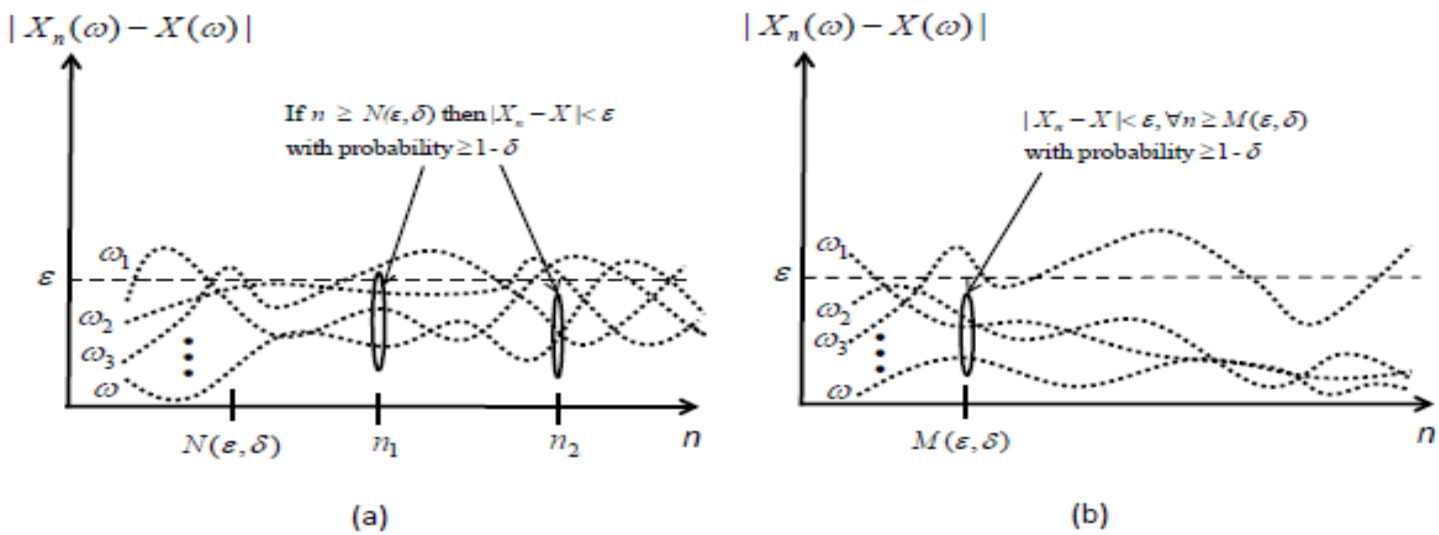

Figure 11.2 Comparison of two types of convergence: (a) convergence in probability; (b) almost sure convergence.

- Further discussions of the WLLN and SLLN, see Sections 11.3.2 and 11.3.3.

2.4 Conditional Probability, Bayes' Theorem and Statistical Independence

- Joint Probability of events *A* and *B*: $P[A, B] = P[A \cap B]$.
- From the axioms of probability

$$0 \leq P[\mathrm{A}, \mathrm{B}] \leq 1. \tag{2.53}$$

- If M possible events A_m's are mutually exclusive, and N possible events B_n's are also mutually exclusive,

$$\sum_{m=1}^{M} \sum_{n=1}^{N} P[A_m, B_n] = 1. \tag{2.54}$$

Definition 2.3 (Conditional Probability). *The conditional probability that event B occurs given that event A occurs is defined as*

$$\boxed{P[B \mid A] \triangleq \frac{P[A, B]}{P[A]},} \tag{2.58}$$

provided that $P[A] > 0$. The conditional probability $P[B \mid A]$ is undefined if $P[A] = 0$. □

2.4.2 Bayes' theorem

- Let $A_1, A_2, \ldots, A_n$ be a partition of Ω, i.e., a set of *mutually exclusive* and *exhaustive* events in Ω. Then for any event *B*,

$$\bigcup_{j=1}^{n} \{B \cap A_j\} = B, \tag{2.60}$$

and then

$$\sum_{j=1}^{n} P[B, A_j] = \sum_{j=1}^{n} P[A_j] P[B \mid A_j] = P[B]. \tag{2.61}$$

The above formula is called the **total probability theorem**.

Theorem 2.2 (Bayes' theorem). *Let B be an event in a sample space Ω and $A_1, A_2, \ldots A_n$ be a partition of Ω. Then it can be shown that*

$$\boxed{P[A_j \mid B] = \frac{P[A_j]\, P[B \mid A_j]}{P[B]} = \frac{P[A_j]\, P[B \mid A_j]}{\sum_{i=1}^{n} P[B \mid A_i] P[A_i]}.} \tag{2.63}$$

Example 2.10: Medical test. Consider some disease and its medical diagnosis test. The following statistics are known about this disease and its medical test.

- For a person with this disease, the test yields a positive result 99% of the time, and a negative result 1%.
- For a person without this disease, the test yields a negative result 99%, and a positive result 1%.
- Suppose that 1% of the population is infected by this disease, and 99% of the population are not.

Suppose that you have taken this test, and unfortunately, the test result is positive. What is the chance that you are indeed infected by this disease?

Answer:

Let

$$A = \text{“Not infected by the disease”}, \quad A^c = \text{“Infected by the disease”}$$
$$B = \text{“Negative test result”}, \quad B^c = \text{“Positive test result”}$$

Then

$$P[A^c|B^c] = \frac{P[A^c, B^c]}{P[A, B^c] + P[A^c, B^c]} = \frac{0.01 \times 0.99}{0.99 \times 0.01 + 0.01 \times 0.99} = \frac{0.0099}{0.0198} = 0.5.$$

Frequentist probabilities and Baysian probabilities

- Frequentists' view: probabilities can be assigned only to outcomes of an experiment that can be repeated many times.
- Bayesians' view: the notion of probability is applicable to any situation or event to which we attach some uncertainty or belief.
- The Bayesian view is appealing to those who investigate probabilistic learning theory.
- The Bayes' theorem provides fundamental principles of learning based on data or evidence.
- You may find the following article of interest and helpful.

  "Are you a Bayeisan or a Frequentist? (Or Bayesian Statistics 101)" in Panos Ipeirotis's blog

  http://www.behind-the-enemy-lines.com/2008/01/are-you-bayesian-or-frequentist-or.html
- See also "Section 4.5 Bayesian inference and conjugate priors" (pp. 97-102") of the textbook on this subject.

2.4.3 Statistical Independence of Events

Definition 2.4 (Statistical independence). *Event B is said to be* statistically independent[12] *of event A, if*

$$P[B \mid A] = P[B], \tag{2.64}$$

or equivalently, if

$$\boxed{P[A, B] = P[A]P[B].} \tag{2.65}$$

Then the events A and B are said to be statistically independent. □

Equations (2.64) and (2.65) are also equivalent to

$$P[A \mid B] = P[A], \tag{2.66}$$

that is, A is also statistically independent of B.

❖ When more than two events are involved:

A set of M events A_m (m=1, 2, …, M) is said to be *mutually independent* if and only if the probability of every intersection of M or fewer events equals the product of the probabilities of the constituents.

❖ For M=3. Three events A, B and C are mutually independent if and only if

$$\begin{aligned} P[A,B] &= P[A]P[B], \\ P[B,C] &= P[B]P[C], \\ P[A,C] &= P[A]P[C], \end{aligned} \tag{2.67}$$

and

$$P[A,B,C] = P[A]P[B]P[C]. \tag{2.68}$$

No three of these relations necessarily implies the fourth. If only Eqs. (2.67) are satisfied, we say that the events are *pairwise independent*. Pairwise independence does not imply mutual independence.

3 Discrete Random Variables

3.1 Random Variables

- **Random variable :** A real-valued function $X(\omega)$, $\omega \in \Omega$, is called a random variable (RV). Thus $X(\omega)$ maps Ω into the real line. $X(\omega)$ is often simply written as X.

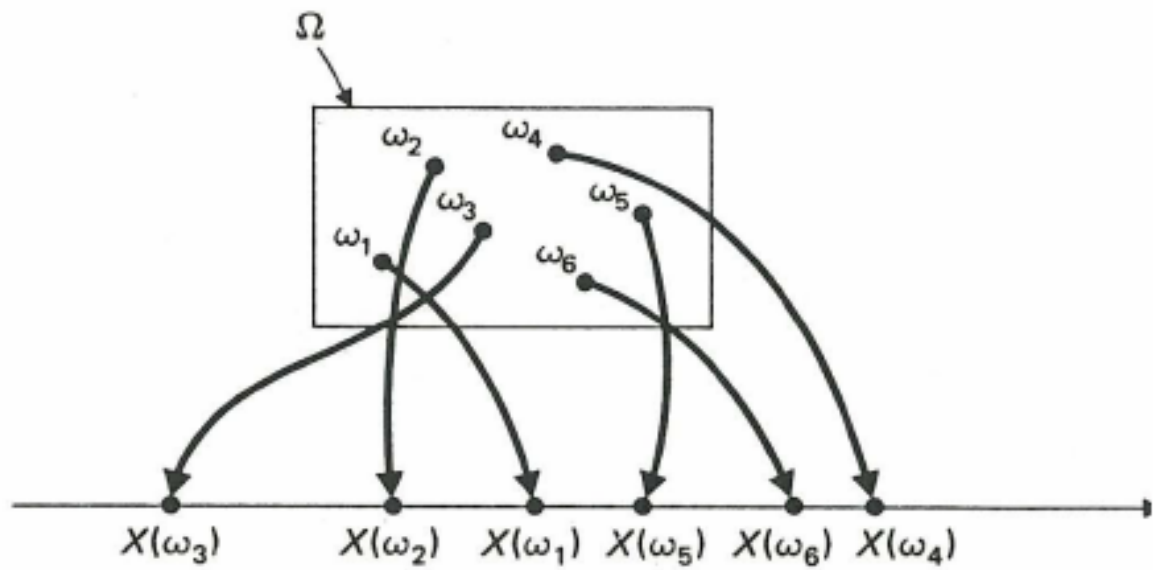

Figure 3.1 A random variable $X(\omega)$ as a mapping from Ω to the real line.

3.1.1 Distribution function

A random variable X is characterized by its *distribution function* $F_X(x)$:

$$\boxed{F_X(x) \triangleq P\left[\{\omega : X(\omega) \leq x\}\right]} \tag{3.1}$$

or simply

$$F_X(x) = P\left[X \leq x\right]. \tag{3.2}$$

The properties of distribution functions listed below follow directly from the definition (3.1) or (3.2).

$$\boxed{\begin{array}{ll} \textit{Property 1.} & F_X(x) \geq 0, \quad \text{for } -\infty < x < \infty. \\ \textit{Property 2.} & F_X(-\infty) = 0. \\ \textit{Property 3.} & F_X(\infty) = 1. \\ \textit{Property 4.} & \text{If } b > a,\ F_X(b) - F_X(a) = P\left[a < X \leq b\right] \geq 0. \end{array}} \tag{3.3}$$

3.1.2 Two random variables and joint distribution function

We define the joint distribution function of the RVs *X* and *Y* by

$$F_{XY}(x,y) \triangleq P[\{\omega : X(\omega) \le x, Y(\omega) \le y\}] = P[X \le x, Y \le y]. \tag{3.4}$$

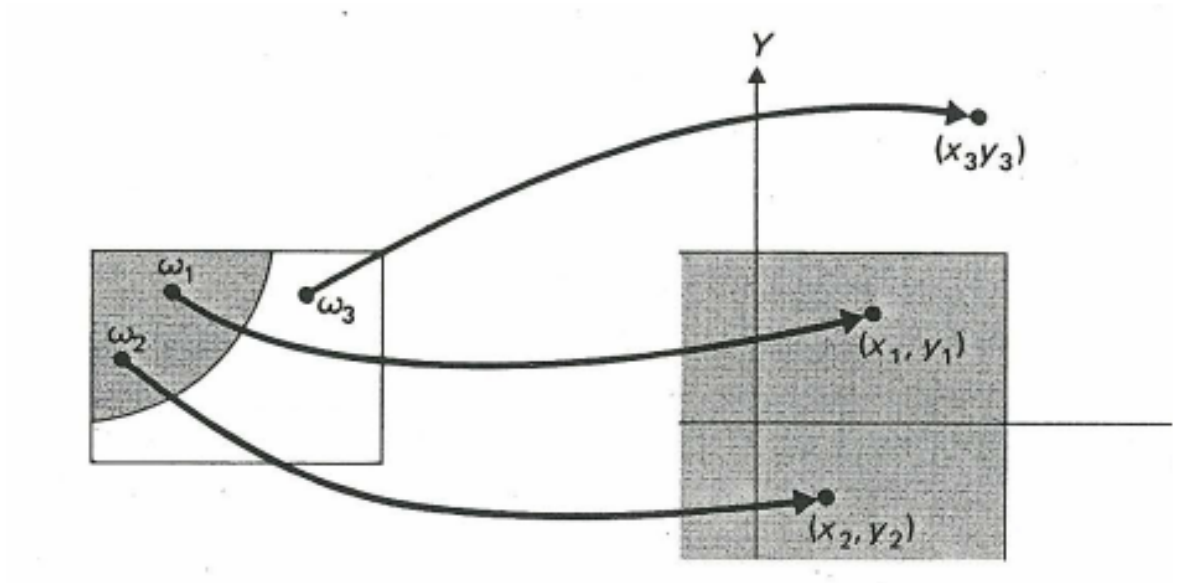

Figure 3.2 Random variables $X(\omega)$ and $Y(\omega)$ as a mapping from Ω to the two-dimensional Euclidean space.

The properties of joint distribution functions listed below follow directly from the definition (3.4).

$$
\begin{aligned}
&\textit{Property 1.} && F_{XY}(x,y) \geq 0; \text{ for } -\infty < x < \infty, \ -\infty < y < \infty. \\
&\textit{Property 2.} && F_{XY}(x,-\infty) = 0; \text{ for } -\infty < x < \infty, \\
& && F_{XY}(-\infty,y) = 0; \text{ for } -\infty < y < \infty. \\
&\textit{Property 3.} && F_{XY}(\infty,\infty) = 1. \\
&\textit{Property 4.} && \text{If } b > a \text{ and } d > c, \\
& && F_{XY}(b,d) \geq F_{XY}(b,c) \geq F_{XY}(a,c). \\
&\textit{Property 5.} && F_{XY}(x,\infty) = F_X(x), \\
& && F_{XY}(\infty,y) = F_Y(y).
\end{aligned}
\tag{3.5}
$$

Property 5 is a consequence of

$$
\begin{aligned}
\{\omega : X(\omega) \leq x\} \cap \{\omega : Y(\omega) < \infty\} &= \{\omega : X(\omega) \leq x\} \cap \Omega \\
&= \{\omega : X(\omega) \leq x\}.
\end{aligned}
$$

3.2 Discrete random variables and probability distributions

- Random variable X is called a *discrete random variable*, if it takes on only a finite or countably infinite number of values $\{x_1, x_2, x_3, \ldots\}$.
- We denote by $p_X(x_i)$, the probability that X takes on x_i :

$$\boxed{p_X(x_i) \triangleq P[X = x_i], \quad i = 1, 2, \ldots,} \tag{3.7}$$

- $\{p_X(x_i)\}$ is called the **probability distribution** of X.
- The distribution function defined by (3.1) is given by

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i). \tag{3.8}$$

- The above distribution function is often called the **cumulative distribution function** (CDF).
- $p_X(x)$: $\mathbb{R} \to [0, 1]$, defined by

$$\boxed{p_X(x) = P[X = x], \quad x \in \mathbb{R},} \tag{3.9}$$

Is called the **probability mass function** (PMF).

Alternatively, we can write

$$\boxed{F_X(x) = \sum_i p_X(x_i) u(x - x_i), \quad -\infty < x < \infty,} \tag{3.10}$$

where $u(t)$ is the *unit step function* defined by

$$u(x) = \begin{cases} 1, & \text{for } x \geq 0, \\ 0, & \text{for } x < 0. \end{cases} \tag{3.11}$$

The formal derivative of this last equation is

$$\boxed{f_X(x) = \frac{dF_X(x)}{dx} = \sum_i p_X(x_i) \delta(x - x_i), \quad -\infty < x < \infty,} \tag{3.12}$$

where $\delta(t)$ is the *Dirac*[3] *delta function* [80] or the *impulse function* defined by

$$\boxed{\delta(x) \triangleq \frac{du(x)}{dx} = 0, \quad \text{for } x \neq 0,} \tag{3.13}$$

Alternatively, we can write

$$\boxed{F_X(x) = \sum_i p_X(x_i) u(x - x_i), \quad -\infty < x < \infty,} \tag{3.10}$$

where $u(t)$ is the *unit step function* defined by

$$u(x) = \begin{cases} 1, & \text{for } x \geq 0, \\ 0, & \text{for } x < 0. \end{cases} \tag{3.11}$$

The formal derivative of this last equation is

$$\boxed{f_X(x) = \frac{dF_X(x)}{dx} = \sum_i p_X(x_i) \delta(x - x_i), \quad -\infty < x < \infty,} \tag{3.12}$$

where $\delta(t)$ is the *Dirac*[3] *delta function* [80] or the *impulse function* defined by

$$\boxed{\delta(x) \triangleq \frac{du(x)}{dx} = 0, \quad \text{for } x \neq 0,} \tag{3.13}$$

3.2.1 Joint and conditional probability distributions

The joint probability distribution of two discrete RVs X and Y is defined by

$$p_{XY}(x_i, y_j) \triangleq P[X = x_i, Y = y_j], \tag{3.17}$$

The corresponding joint distribution function defined by (3.4) is given by

$$F_{XY}(x, y) = \sum_{x_i \le x} \sum_{y_j \le y} p_{XY}(x_i, y_j). \tag{3.18}$$

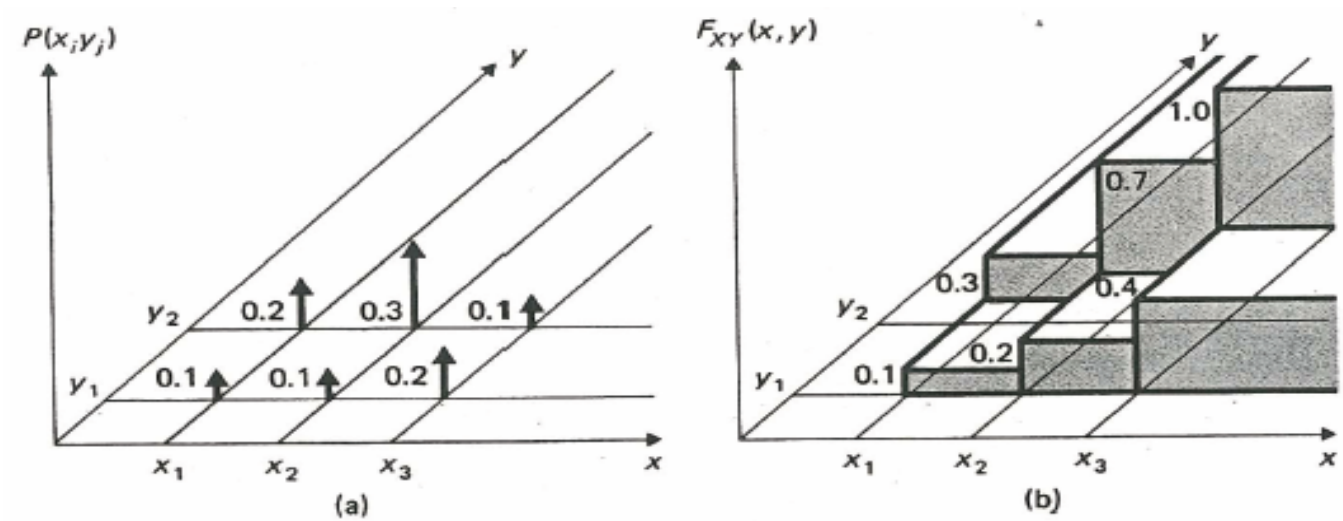

Figure 3.4 (a) The joint probability distribution and (b) the joint distribution function

- We define the *conditional probability distributions*:

$$p_{Y|X}(y_j|x_i) \triangleq P[Y = y_j|X = x_i] \text{ and } p_{X|Y}(x_i|y_j) \triangleq P[X = x_i|Y = y_j]. \tag{3.20}$$

- The following relations hold:

$$p_{XY}(x_i, y_j) = p_X(x_i)p_{Y|X}(y_j|x_i) = p_Y(y_j)p_{X|Y}(x_i|y_j), \tag{3.21}$$

$$\sum_{\text{all } j} p_{Y|X}(y_j|x_i) = \sum_{\text{all } i} p_{X|Y}(x_i|y_j) = 1, \tag{3.22}$$

$$p_X(x_i) = \sum_{\text{all } j} p_{XY}(x_i, y_j) \text{ and } p_Y(y_j) = \sum_{\text{all } i} p_{XY}(x_i, y_j), etc. \tag{3.23}$$

- We define the conditional distribution function of *X* given by *Y* as

$$F_{X|Y}(x|y) \triangleq P[X \leq x|Y = y] = \sum_{x_i < x} p_{X|Y}(x_i|y), \tag{3.24}$$

when $P[Y = y] > 0$, i.e., $y = y_j$ for some j.

Definition 3.1 (Independent random variables). *We say that random variables X and Y are* independent *or* statistically independent *if and only if*

$$\boxed{p_{XY}(x_i, y_j) = p_X(x_i)p_Y(y_j), \quad \text{for all values } (x_i, y_j),} \tag{3.25}$$

or, equivalently, if and only if

$$F_{XY}(x_i, y_j) = F_X(x_i)F_Y(y_j), \quad \text{for all values of } x_i \text{ and } y_j. \tag{3.26}$$

□

Similarly, the discrete RVs $X, Y, \ldots, Z$ are said to be independent RVs if and only if

$$p_{XY\cdots Z}(x_k, y_l, \ldots, z_m) = p_X(x_k)p_Y(y_l)\cdots p_Z(z_m) \tag{3.27}$$

is satisfied for all values $(x_k, y_l, \ldots, z_m)$ or, equivalently,

$$F_{XY\cdots Z}(x_k, y_l, \ldots, z_m) = F_X(x_k)F_Y(y_l)\cdots F_Z(z_m) \tag{3.28}$$

for all values of $x_k, y_l, \cdots, z_m$.

Definition 3.2 (Expectation). *The* expectation, *the* expected value, *or the* mean *of a discrete RV X with probability distribution $\{p_X(x_i\}$ is defined as*

$$\boxed{\mu_X = E[X] \triangleq \sum_{\text{all } i} x_i p_X(x_i),} \tag{3.32}$$

provided the sum converges absolutely[5]. □

Given a conditional probability distribution $\{p_{X|Y}(x_i|y_j)\}$, the **conditional expectation** of X conditioned on the event $\{Y = y_j\}$ is defined by

$$\boxed{E[X|Y = y_j] \triangleq \sum_{\text{all } i} x_i p_{X|Y}(x_i|y_j).} \tag{3.36}$$

Definition 3.3 (Conditional expectation). *Let $\{p_{X|Y}(x_i|y_j)\}$ be the conditional probability distribution of a discrete RV X conditioned on another discrete RV Y to be equal to y_j and define a function $\psi : \{y_j\} \to \mathbb{R}$, by $\psi(y_j) = E[X|Y = y_j]$. Then the conditional expectation of X given Y is given by*

$$\boxed{E[X|Y] \triangleq \psi(Y).} \tag{3.37}$$

□

- Note that $E[X|Y]$ is a function of the RV Y only.
- Since $E[X|Y]$ is itself a discrete RV, we can take its expectation and show

$$\boxed{E[E[X|Y]] = E[X].} \tag{3.38}$$

- The property (3.38) of the conditional expectation is called the **law of iterated expectations**, the **law of total expectation**, or the **tower property**.

3.2.2 Moments, central moments, and variance

Definition 3.4 (Moments and central moments). *For a positive integer k,*

$$E[X^k] = \sum_{\text{all } i} x_i^k p_X(x_i), \tag{3.39}$$

is called the kth moment *of X, provided the series converges absolutely. Similarly,*

$$E\left[(X - \mu_X)^k\right] = \sum_{\text{all } i} (x_i - \mu_X)^k p_X(x_i). \tag{3.40}$$

is called the kth central moment *of X.* □

Definition 3.5 (Variance and standard deviation). *Let X be a RV with finite second moment $E[X^2]$ and mean μ_X. We define the* variance *of X as*

$$\boxed{\sigma_X^2 = \mathrm{Var}[X] \triangleq E[(X-\mu_X)^2] = E[X^2] - \mu_X^2.} \tag{3.41}$$

The square root of the variance, σ_X, is called the standard deviation. □

Definition 3.6 (Conditional variance). *Let X and Y be discrete RVs. The conditional variance of X given Y is defined as*

$$\boxed{\mathrm{Var}[X|Y] \triangleq E[(X - E[X|Y])^2|Y].} \tag{3.42}$$

□

3.2.3 Covariance and correlation coefficient

$$\boxed{\sigma_{X,Y} \triangleq \mathrm{Cov}[X,Y] \triangleq E[(X-\mu_X)(Y-\mu_Y)]} \tag{3.47}$$

is called the *covariance* between X and Y. Expanding the above expression gives

$$\sigma_{X,Y} = \mathrm{Cov}[X,Y] = E[XY] - \mu_X\mu_Y. \tag{3.48}$$

Let us *normalize* X and Y by their standard deviations:

$$X^* = \frac{X}{\sigma_x}, \text{ and } Y^* = \frac{Y}{\sigma_Y}. \tag{3.49}$$

The covariance between X^* and Y^* is called the *correlation coefficient* of X and Y, and is denoted by $\rho(X, Y)$:

$$\boxed{\rho(X, Y) \triangleq \mathrm{Cov}[X^*, Y^*] = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.} \tag{3.50}$$

- From the Cauchy-Schwarz inequality (see pp. 241-245)

$$|\rho(X, Y)| \leq 1.$$

Definition 3.7 (Uncorrelated random variables). *We say X and Y are uncorrelated if*

$$\boxed{\mathrm{Cov}[X, Y] = \rho(X, Y) = 0.} \tag{3.51}$$

□