Combinatorial Optimization in Computer Vision (IN2245)

Frank R. Schmidt
Csaba Domokos

Winter Semester 2015/2016

Reasoning under uncertainty

We often want to understand a system when we have *imperfect* or *incomplete* information due to, for example, noisy measurement. There are two main reasons why we might reason under uncertainty:

- *Laziness*: modeling every detail of a complex system is costly
- *Ignorance*: we may not completely understand

Probability $P(A)$ refers to a degree of confidence that an event A with uncertain nature will occur.

It is common to assume that $0 \leqslant P(A) \leqslant 1$. If $P(A) = 1$, we are certain that A occurs, while $P(A) = 0$ asserts that A will not occur.

Interpretations of probability

Objective probability: It quantifies uncertainty regarding the occurrence of events. After repeating an experiment under identical conditions one can calculate the **relative frequency** of an event A as

$$h_A = \frac{m_A}{m} ,$$

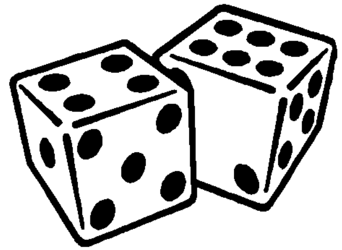
where m_A is the number of times when A occurs and m is the total number of experiments performed.
Example: flipping a coin, the relative frequencies of heads and tails are around one half.

Subjective probability: measures a personal belief.
Example: The probability of rain tomorrow in Munich is 50%.

Experiment, event space, event

An **experiment** is a (random) process that can be infinitely many times repeated and has a well-defined set of possible **outcomes**. In case of repeated experiments the individual repetitions are also called **trials**.

Example: throwing two "fair dice" (i.e. we assume equally likely chance of landing on any face) with six faces.



The **event space**, denoted by Ω, is the set of possible outcomes.

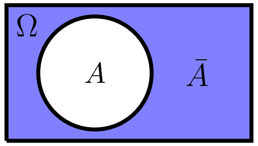
Example: $\Omega = \{(i, j) : 1 \leqslant i, j \leqslant 6\}$.

A set of outcomes $A \subseteq \Omega$ is called an **event**. An **atomic event** is an event that contains a single outcome $\omega \in \Omega$.
Example: $A = \{(i, j) : i + j = 11\}$, i.e. the sum of the numbers showing on the top is equal to eleven.

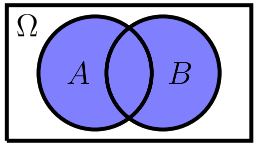
Basic notations

Let A and B be two events from an event space Ω. We will use the following notations:

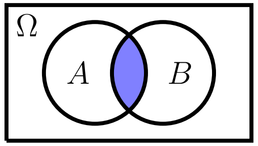
A does not occur: $\bar{A} = \Omega \backslash A$



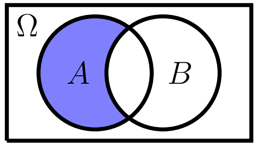
either A or B occur: $A \cup B$



both A and B occur: $A \cap B$



A occurs and B does not: $A \backslash B$



- The $\emptyset$ is called the **impossible event**; and
- Ω is the **sure event**.

Discrete probability space

A probability space represents our uncertainty regarding an experiment.

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

- Ω is not empty and **countable** (i.e. $\exists \mathcal{S} \subseteq \mathbb{N}$ such that $|\Omega| = |\mathcal{S}|$),
- $\mathcal{A}$ is the power set $\mathcal{P}(\Omega)$, and
- $P : \mathcal{A} \to \mathbb{R}$ is a function, called a *probability measure*, with the following properties:
  1. $P(A) \geqslant 0$ for all $A \in \mathcal{A}$
  2. $P(\Omega) = 1$
  3. **σ-additivity** holds: if $A_n \in \mathcal{A}, n = 1, 2, \ldots$ and $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) \, .$$

The conditions 1-3. are called **Kolmogorov's axioms**.

Example: throwing two "fair dice"

For this case a discrete probability space $(\Omega, \mathcal{A}, P)$ is given by

- Event space: $\Omega = \{(i, j) : 1 \leqslant i, j \leqslant 6\}$.
- $\mathcal{A} = \mathcal{P}(\Omega) = \{\{(1,1)\}, \ldots, \{(1,1),(1,2)\}, \ldots, \{(1,1),(1,2),(1,3)\}, \ldots\}$.
- The probability measure

$$P(A) = \frac{|A|}{36} = \frac{k}{36} \, ,$$

  where k is the number of atomic events in A.

Example: Let A denote the event when the sum of the numbers showing on the top is equal to eleven that is $A = \{(i,j) : i + j = 11\} = \{(5,6),(6,5)\}$. Hence

$$P(A) = P(\{(5,6),(6,5)\}) = \frac{2}{36} \, .$$

Example: throwing two "fair dice"

Events	Set of corresponding atomic events	Probability
2	{(1,1)}	$1/36 \approx 3\%$
3	{(1,2), (2,1)}	$2/36 \approx 6\%$
4	{(1,3), (2,2), (3,1)}	$3/36 \approx 8\%$
5	{(1,4), (2,3), (3,2), (4,1)}	$4/36 \approx 11\%$
6	{(1,5), (2,4), (3,3), (4,2), (5,1)}	$5/36 \approx 14\%$
7	{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}	$6/36 \approx 17\%$
8	{(2,6), (3,5), (4,4), (5,3), (6,2)}	$5/36 \approx 14\%$
9	{(3,6), (4,5), (5,4), (6,3)}	$4/36 \approx 11\%$
10	{(4,6), (5,5), (6,4)}	$3/36 \approx 8\%$
11	{(5,6), (6,5)}	$2/36 \approx 6\%$
12	{(6,6)}	$1/36 \approx 3\%$

σ-algebra, measure, measure space

Assume an arbitrary set Ω and $\mathcal{A} \subseteq \mathcal{P}(\Omega)$. The set $\mathcal{A}$ is a **σ-algebra** over Ω if the following conditions are satisfied:

1. $\emptyset \in \mathcal{A}$,
2. $A \in \mathcal{A} \Rightarrow \bar{A} \in \mathcal{A}$ (i.e. it is closed under complementation),
3. $A_i \in \mathcal{A}$ $(i \in \mathbb{N}) \Rightarrow \bigcup_{i=0}^{\infty} A_i \in \mathcal{A}$ (i.e. it is closed under countable union).

It is a consequence of this definition that $\Omega \in \mathcal{A}$ is also satisfied.

Assume an arbitrary set Ω and a σ-algebra $\mathcal{A}$ over Ω. A function $P : \mathcal{A} \to [0, \infty]$ is called a **measure** if the following conditions are satisfied:

1. $P(\emptyset) = 0$,
2. P is σ-additive.

Let $\mathcal{A}$ be a σ-algebra over Ω and $P : \mathcal{A} \to [0, \infty]$ is a measure. $(\Omega, \mathcal{A})$ is said to be a **measurable space** and the triple $(\Omega, \mathcal{A}, P)$ is called a **measure space**.

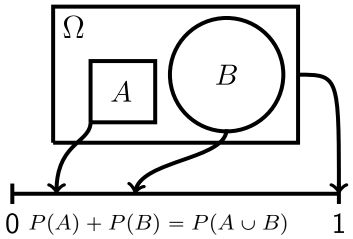
Probability space

A **probability space** is a triple $(\Omega, \mathcal{A}, P)$, where $(\Omega, \mathcal{A})$ is a measurable space, and P is a measure such that $P(\Omega) = 1$, called a **probability measure**.

To summarize:
A triple $(\Omega, \mathcal{A}, P)$ is called probability space, if

- the event space Ω is not empty,
- $\mathcal{A}$ is a σ-algebra over Ω, and
- $P : \mathcal{A} \to \mathbb{R}$ is a function with the following properties:
  1. $P(A) \geqslant 0$ for all $A \in \mathcal{A}$
  2. $P(\Omega) = 1$
  3. σ-additive: if $A_n \in \mathcal{A}$, $n = 1, 2, , \ldots$ and $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) \; .$$



Example: throwing a dart

Suppose a dart is thrown at a round board modeled as a unit circle. The sample space contains the location of the dart if it lands in the board only. The sample space is given by $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\}$.

We denote the **area** of an the event $A \subseteq \Omega$ by $\mu(A)$, which is defined as the *Riemann-integral* of the characteristic function of A

$$\mu(A) := \int_\Omega \chi_A(x) \mathrm{d}x \ , \quad \text{where} \quad \chi_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \ . \end{cases}$$

The σ-algebra $\mathcal{A}$ over Ω is defined as follows

$$\mathcal{A} = \{A \subseteq \Omega : \mu(A) \text{ exists}\} \ .$$

The probability measure $P : \Omega \to [0, 1]$ is given as $P(A) = \frac{\mu(A)}{\pi}$.

Some simple consequences of the axioms

The following rules are frequently used in applications:

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

  Proof. Note that A and $\Omega \backslash A$ are disjoint.
  $1 = P(\Omega) = P(A \cup (\Omega \backslash A)) = P(A) + P(\Omega \backslash A)$ □

- $P(\varnothing) = 0$

  Proof. $P(\varnothing) = 1 - P(\Omega \backslash \varnothing) = 1 - P(\Omega) = 1 - 1 = 0$ □

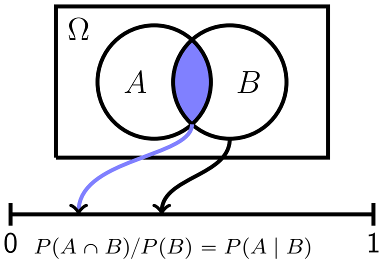
- If $A \subseteq B$, then $P(A) \leqslant P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leqslant P(A) + P(B)$
- $P(A \backslash B) = P(A) - P(A \cap B)$
- ...

Conditional probability

Conditional probability allows us to reason with partial information.
If $P(B) > 0$, the **conditional probability of A given B** is defined as

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

This is the probability that A occurs, given we have observed B, i.e. we know the experiment's actual outcome will be in B.



Note that the axioms and rules of probability theory are fulfilled for the conditional probability. (e.g. $P(A \mid B) = 1 - P(\bar{A} \mid B)$).

Example

Consider two producing machines creating identical product in a factory. Assume we are given the following table with probabilities

	Machine I	Machine II	
The product is good	0.56	0.41	0.97
The product is waste	0.01	0.02	0.03
	0.57	0.43	1

What is the probability of that a product was created by Machine I, when it is good?
Let A denote the event that the product was created by Machine I and let B denote the event that the product is good.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.56}{0.97} \approx 0.58$$

The chain rule

The product rule: starting with the definition of conditional probability $P(B \mid A)$ and multiplying by P(A) we get that

$$P(A \cap B) = P(A)P(B \mid A) \ .$$

The chain rule:

$$P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid \cap_{i=1}^{n-1} A_i) \ . \tag{1}$$

Proof. By induction. For $n = 2$ we get the product rule. Let $n \in \mathbb{N}$ be given and suppose Eq. (1) is true for $k \leqslant n$. Then

$$P(\cap_{i=1}^{n+1} A_i) = P(A_{n+1} \cap (\cap_{i=1}^{n} A_i)) = P(A_{n+1} \mid \cap_{i=1}^{n} A_i)P(\cap_{i=1}^{n} A_i) \ .$$

□

The chain rule will become important later when we discuss conditional independence.

Bayes' rule

By making use of the product rule we can get

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

$P(A \mid B)$ is often called the **posteriori probability**, and $P(B \mid A)$ is called the **likelihood**, and $P(A)$ is called the **prior probability**.

A more general version of Bayes' rule, when we have a background event C:

$$P(A \mid B \cap C) = \frac{P(B \mid A \cap C)P(A \mid C)}{P(B \mid C)} \ .$$

Example: What is the probability of that a product is good, if it was created by Machine I? We are given $P(A \mid B) = 0.58$, $P(A) = 0.57$ and $P(B) = 0.97$.

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} = \frac{0.58 \cdot 0.97}{0.57} \approx 0.98 \ .$$

Independence

Two events A and B are **independent**, denoted by $A \perp B$ if

$$P(A \mid B) = P(A)$$

or, equivalently, iff

$$P(A \cap B) = P(A)P(B) \ .$$

If A and B are independent, learning that B happened does not make A more or less likely to occur.

Example: Suppose we roll a die. Let us consider the event A denoting "the die outcome is even" and B denoting "the die outcome is 1 or 2". If the die is fair, then $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{3}$. Moreover $A \cap B$ means the event that the outcome is two, so $P(A \cap B) = \frac{1}{6}$.

$$P(A \cap B) = \frac{1}{6} = \frac{1}{2} \cdot \frac{1}{3} = P(A)P(B) \quad \Rightarrow \quad A \text{ and } B \text{ are independent.}$$

Conditional independence

Let A, B and C be events. A and B is **conditionally independent** given C, iff

$$P(A \mid C) = P(A \mid B \cap C) ,$$

or, equivalently, iff

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C) .$$

A and B are independent given C means that once we learned C, learning B gives us no additional information about A.

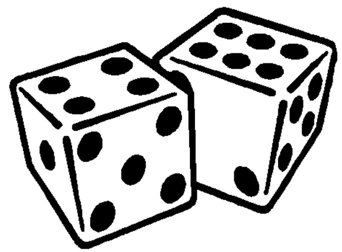
Examples:

- The operation of a car's starter motor is conditionally independent its radio given the status of the battery.
- Symptoms are conditionally independent given the disease.

Example: throwing two "fair" dice

In many cases it would be more natural to consider *attributes* of the outcomes. A random variable is a way of reporting an attribute of the outcome.

We have the state space $\Omega = \{(i, j) : 1 \leqslant i, j \leqslant 6\}$ and the (uniform) probability measure $P(\{(i, j)\}) = \frac{1}{36}$, where $(\Omega, \mathcal{P}(\Omega), P)$ forms a probability space.



We are interested in the sum of the numbers showing on the dice, defined by define the mapping $X : \Omega \to \Omega'$, $X(i, j) = i + j$, where $\Omega' = \{2, 3, \ldots, 12\}$.

It can be seen that this mapping leads a probability space $(\Omega', \mathcal{P}(\Omega'), P')$ such that the probability measure is defined as $P' : \mathcal{P}(\Omega') \to [0, 1]$,

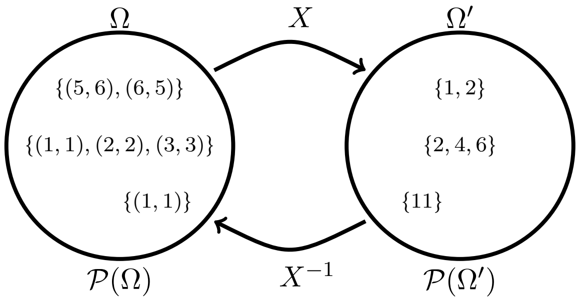
$$P'(A') = P(\{(i, j) : X(i, j) \in A'\}) \ .$$

For example: $P'(\{11\}) = P(\{(5, 6), (6, 5)\}) = \frac{2}{36}$.

Preimage mapping

Let $X : \Omega \to \Omega'$ be an arbitrary mapping. The **preimage mapping** $X^{-1} : \mathcal{P}(\Omega') \to \mathcal{P}(\Omega)$ is defined as

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



Note that the preimage of a σ-algebra is a σ-algebra.

Random variable

Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ measurable spaces. A mapping $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ is called **measurable**, if

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

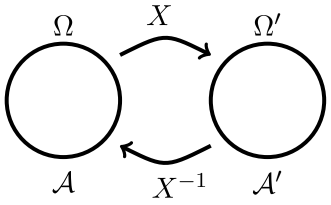
A measurable mapping $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{A}')$ is called **random variable**.

Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ be a random variable and P a measure over $\mathcal{A}$. Then

$$P'(A') := P_X(A') := P(X^{-1}(A'))$$

defines a measure over $\mathcal{A}'$.

P_X is called is called the **image measure** of P by X. Specially, if P is a probability measure then P_X is a probability measure over $\mathcal{A}'$.



Example: throwing two "fair" dice

We are given the event spaces $\Omega = \{(i, j) : 1 \leqslant i, j \leqslant 6\}$ and $\Omega' = \{2, 3, \ldots, 12\}$. We assume the uniform probability measure P over $(\Omega, \mathcal{P}(\Omega))$.

Define a mapping $X : (\Omega, \mathcal{P}(\Omega)) \to (\Omega', \mathcal{P}(\Omega'))$, $X(i, j) = i + j$. Is X a random variable?

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\} \in \mathcal{P}(\Omega)$$

is satisfied, since for any $\omega' \in \Omega'$ one can find an $\omega \in \Omega$ such that $X(\omega) = \omega'$. Therefore X is measurable, thus it is a random variable.

Moreover, P is a probability measure, hence the image measure $P_X(A') = P(X^{-1}(A'))$ is a probability measure on $(\Omega', \mathcal{P}(\Omega'))$.

Example: $P_X(\{2, 4, 5\}) = P(X^{-1}(\{2, 4, 5\})) = P(\{(1, 1), (1, 3), (2, 2), (3, 1), (1, 4), (2, 3), (3, 2), (4, 1)\}) = \frac{8}{36} = \frac{2}{9}$.

Probability distributions

Probability distribution

A random variable is a measurable mapping from a probability space to a measure space. It is *neither a variable nor random*.

Let $X : (\Omega, \mathcal{A}, P) \to (\Omega', \mathcal{A}')$ be a random variable. Then the image measure P_X of P by X is called **probability distribution**.

Assume an event $A \in \Omega$ and let $x = X(A)$. We use the notation $P(x)$ for $P(\{X = x\})$, where $\{X = x\}$, which means that the mapping X has the value x, is also considered as an event for an $x \in \Omega'$.

Similarly, $\{X < x\}$, which corresponds to the set of atomic events $\{\omega \in \Omega : X(\omega) < x\}$, also defines an event in Ω' .

Let $X : (\Omega, \mathcal{A}, P) \to (\Omega', \mathcal{A}')$ be a random variable. Then $F_P : \mathbb{R} \to \mathbb{R}$ is called **cumulative distribution function** of P.

$$F_P(x) = P(X < x) \,, \quad x \in \mathbb{R} \,.$$

Each probability measure is defined uniquely by its distribution function.

Properties of the cumulative distributive function

The cumulative distributive function $F_P : \mathbb{R} \to \mathbb{R}$, $F_P(x) = P(X < x)$ for a probability measure P has the following properties:

1. F_P is monotonously increasing
2. F_P is left continuous
3. $\lim_{x \to -\infty} F_P(x) = 0$
4. $\lim_{x \to \infty} F_P(x) = 1$

$$P(a \leqslant X < b) = P(X < b) - P(X < a) = F_P(b) - F_P(a) \ .$$

Density function

A random variable $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ is said to be **discrete random variable** if Ω' is countable.

Let $F_P : \mathbb{R} \to \mathbb{R}$ be the cumulative distribution function of a probability measure P. A measurable function $f(x)$ is called a **density function**, if

$$F_P(x) = \int_{-\infty}^{x} f(t) \mathsf{d}t \ , \quad x \in \mathbb{R} \ .$$

A **measurable function** we mean to be a function with improper Riemann-integral.

Continuous random variable

A random variable $X : (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{A}')$ is called **continuous random variable**, if it has a density function $f(x)$. Then the following are held:

1. $f(x)$ is non-negative,
2. $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$,
3. $F_P(a \leqslant X < b) = \int_a^b f(x)\mathrm{d}x$.

Proof.

1. F_P is non-negative and monotonously increasing which implies $f(x) \geqslant 0$.
2. $$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = F_P(\infty) - F_P(-\infty) = 1 - 0 = 1 \ .$$
3. $$F_P(a \leqslant X < b) = F_P(b) - F_P(a) = \int_{-\infty}^{b} f(x)\mathrm{d}x - \int_{-\infty}^{a} f(x)\mathrm{d}x = \int_{a}^{b} f(x)\mathrm{d}x \ .$$

□

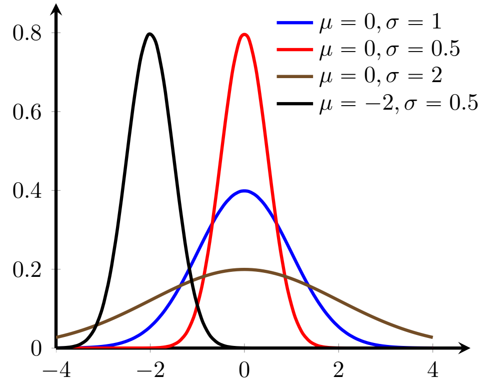
The Normal (Gaussian) distribution

A continuous random variable $X : \mathbb{R} \to \mathbb{R}$ with density function

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

is said the have **Normal distribution** (or **Gaussian distribution** with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$.

Standard Normal distribution:
$\mu = 0$ and $\sigma = 1$.



Three-sigma rule of thumb:

68.27% of the area under curve lie within the interval $\mu \pm \sigma$.
95.45% of the area under curve lie within the interval $\mu \pm 2\sigma$.
99.73% of the area under curve lie within the interval $\mu \pm 3\sigma$.

Joint distribution

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be discrete random variables, where $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote the values of Y.

We introduce the notation

$$p_{ij} = P(X = x_i, Y = y_j) \quad i, j = 1, 2, \ldots$$

for the probability of the events $\{X = x_i, Y = y_j\} := \{X = x_i\} \cap \{Y = y_j\}$.

These probabilities p_{ij} form a distribution, called the **joint distribution** of X and Y. Therefore,

$$\sum_i \sum_j p_{ij} = 1 \; .$$

Marginal distributions

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be discrete random variables, where $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote the values of Y.

The distributions defined by the probabilities

$$p_i = P(X = x_i) \quad \text{and} \quad q_j = P(Y = y_j)$$

are called the **marginal distributions** of X and of Y, respectively.

Let us consider the marginal distribution of X. Then

$$p_i = P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij} \; .$$

Similarly,

$$q_j = P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij} \; .$$

Example: marginal distribution

Consider two producing machines creating identical product in a factory. Assume we are given the following table with probabilities

	Machine I	Machine II	
The product is good	0.56	0.41	0.97
The product is waste	0.01	0.02	0.03
	0.57	0.43	1

The marginal distributions of discrete random variables corresponding to the values of $\{\text{good}, \text{waste}\}$ and $\{\text{I}, \text{II}\}$ are shown in the last column and last row, respectively.

The following also holds

$$\sum_i p_i = \sum_i P(X = x_i) = \sum_i \sum_j P(X = x_i, Y = y_i) = \sum_i \sum_j p_{ij} = 1 \ .$$

Joint density

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be random variables. The **joint cumulative distribution function** of X and Y, denoted by $F_P : \mathbb{R}^2 \to \mathbb{R}$, is defined as

$$F_P(x, y) = P(X < x, Y < y) \quad x, y \in \mathbb{R} \ .$$

If both X and Y are continuous random variables, then the **joint density function** $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ is defined as

$$F_P(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) \mathrm{d}u \mathrm{d}v \ .$$

The joint density function $f_{XY}(x, y)$ also satisfies the following property:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(u, v) \mathrm{d}u \mathrm{d}v = 1 \ .$$

Marginal densities

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be random variables with joint cumulative distribution function $F_P : \mathbb{R}^2 \to \mathbb{R}$. The **marginal cumulative distribution functions** of X and Y are given by

$$F_{P_X} := F_P(x, \infty) = \lim_{y \to \infty} F_P(x, y) \;, \quad \text{and}$$
$$F_{P_Y} := F_P(\infty, y) = \lim_{x \to \infty} F_P(x, y) \;.$$

If both X and Y are continuous random variables with the joint density function $F_{XY}(x, y)$, then the **marginal density functions** $f_X, f_Y : \mathbb{R} \to \mathbb{R}$ are defined as

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) \mathrm{d}y \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \mathrm{d}x \;.$$

Conditional distribution

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let X and Y be discrete random variables, where $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote the values of Y.

The **conditional distribution** of X given Y is defined by

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{\sum_k p_{kj}} = \frac{p_{ij}}{q_j} \;.$$

Therefore, $\sum_i P(X = x_i \mid Y = y_j) = \sum_i \frac{p_{ij}}{\sum_k p_{kj}} = 1$ is also held.

The **conditional cumulative distribution function** is defined as

$$F_P(x \mid y) = \lim_{h \to 0} F_P(x \mid y \leqslant Y < y + h) \;,$$

where

$$F_P(x \mid y \leqslant Y < y + h) = P(X < x \mid y \leqslant Y < y + h) = \frac{P(X < x, y \leqslant Y < y + h)}{P(y \leqslant Y < y + h)} \;.$$

Conditional density

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let X and Y be random variables with joint density function $f_{XY}(x, y)$. If the marginal density function $f_Y(y) \neq 0$, then the **conditional density function** of X given Y is defined as

$$f(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} \, .$$

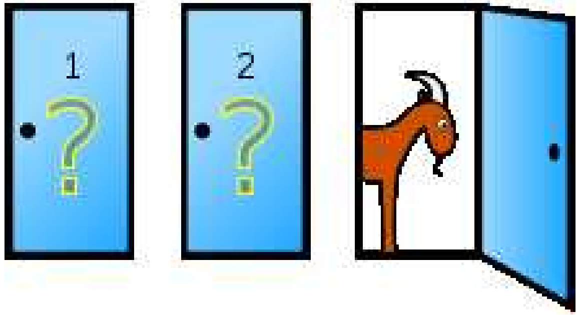
Literature

1. Marek Capiński and Ekkerhard Kopp. **Measure, Integral and Probability**. Springer, 1998.
2. Daphne Koller and Nir Friedman. **Probabilistic Graphical Models: Principles and Techniques**. The MIT Press, 2009. Note: Chapter 2.1

A brain teaser

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats.

You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.



He then says to you, "Do you want to pick door No. 2?"

Is it to your advantage to switch your choice?