Probabilites

Anders Hansson and Martin Andersen

Linköping University and Technical Unversity of Denmark

March 26, 2024

Probability Space

A probability space is defined by a triplet $(\Omega, \mathcal{F}, P)$.

Ω is called the *sample space* and it is a set that contains all possible outcomes of an experiment.

When throwing a dice we can take $\Omega = \{1, 2, 3, 4, 5, 6\}$.

If outcome of experiment measurement error $\Omega = \mathbf{R}$.

If we throw two dices $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$.

Sometimes infinite dimensional vectors or functions from **Z** to **R**: $\Omega = \mathbf{R}^{\mathbf{Z}}$.

Set of all functions from **R** to **R**: $\Omega = \mathbf{R}^{\mathbf{R}}$.

σ-algebra

The second element $\mathcal{F}$ of a probability space should contain all events that we are interested in assigning probabilities to.

This is a set of subsets of Ω, and it should be a so-called *σ-algebra* i.e.

1. $\Omega \in \mathcal{F}$
2. $\Omega \setminus A \in \Omega, \ \forall A \in \mathcal{F}$
3. $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}, \ \forall A_1, A_2, \ldots \in \mathcal{F}$

For finite Ω one may take $\mathcal{F}$ to contain all subsets of Ω. The smallest possible σ-algebra is $\mathcal{F} = \{\Omega, \emptyset\}$.

Probability measure

The *probability measure* P is a function $P : \mathcal{F} \rightarrow [0, 1]$ such that

1. $P(\Omega) = 1$
2. If $A_1, A_2, \ldots \in \mathcal{F}$ is a collection of pairwise disjoint sets, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Well-known properties such as $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ all follow from the above axioms.

Probability function

When Ω contains countable number of elements may define a *probability function* (pf) $p : \Omega \to [0,1]$:

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

Then the probability measure is

$$P(A) = \sum_{\omega \in A} p(\omega)$$

for any $A \in \mathcal{F}$. Notice that the summations above may be infinite.

When p is a function of an integer we often write p_k instead of $p(k)$.

We may also use the vector $p = (p_1, \ldots, p_n) \in [0,1]^n$ instead of the function p to describe the probability measure.

Categorial distribution

Let $\Omega = \mathbf{N}_n$ and let $p : \mathbf{N}_n \to [0,1]$ be defined as

$$p(k) = \begin{cases} \frac{e^{z_k}}{1+\sum_{l=1}^{n-1} e^{z_l}}, & k \in \mathbf{N}_{n-1} \\ \frac{1}{1+\sum_{l=1}^{n-1} e^{z_l}}, & k = n \end{cases}$$

where $z_k \in \mathbf{R}$, $k \in \mathbf{N}_{n-1}$, which is called the *categorial* probability function. It is a *finite probability function*.

Distribution function and Probability density function

When $\Omega = \mathbf{R}$ we define the *distribution function* (df) $F : \mathbf{R} \rightarrow [0, 1]$ as

$$F(\omega) = P([-\infty, \omega])$$

For any set $A \subset \Omega$ we have that

$$P(A) = \int_{\omega \in A} dF(\omega)$$

When F is differentiable the derivative $f : \mathbf{R} \rightarrow \mathbf{R}_+$ of F exists, and we may write

$$P(A) = \int_{\omega \in A} f(\omega) d\omega$$

The function f is called the *probability density function* (pdf). It is straight forward to generalize the results to $\Omega = \mathbf{R}^n$.

Conditional probabilities

Interested in probabilites of an event A if we know that another event B has happened.

Example: what is the probability that we get a one when we know that we have observed an odd number when throwing a dice?

We are interested in probability of $A = \{1\}$ given $B = \{1, 3, 5\}$.

Whole sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. Can look at a smaller sample space defined by B and then we investigate how frequent A is in this sample space and we obtain the probability $1/3$.

Alternatively use Ω and compute

$$\frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1/6}{1/2} = \frac{1}{3}$$

i.e. we normalize the probability of both events occurring with the probability of the event that we know have occurred.

Multiplication theorem and Baye's theorem

If $P(B) > 0$ we define the *conditional probability* that A occurs given that B occurs as

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

From this it immediately follows that

$$P(A \cap B) = P(B)P(A|B) \tag{1}$$

and by induction that for $A_i \in \mathcal{F}$ it holds that

$$P\left(\cap_{i=1}^{n} A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \times \cdots \\ \times P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

which is called the *multiplication theorem*. From (1) and its symmetric counterpart $P(B \cap A) = P(A)P(B|A)$ it follows that

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{2}$$

for any $A, B \in \mathcal{F}$ such that $P(A) > 0$ and $P(B) > 0$. This is called *Bayes' theorem*.

Formula of total probablity

Let $A_i \in \mathcal{F}$, $i \in \mathbf{N}_n$ be pairwise disjoint sets such that $\Omega = \cup_{i=1}^{n} A_i$. Then for any $X \in \mathcal{F}$ it holds that

$$P(X) = \sum_{i=1}^{n} P(A_i \cap X)$$

which is called the *formula of total probability*. Moreover, by (1) we have $P(A_i \cap X) = P(A_i)P(X|A_i)$. Hence from (2) it follows that

$$P(A_i|X) = \frac{P(A_i)P(X|A_i)}{\sum_{j=1}^{n} P(A_j)P(X|A_j)}, \quad i \in \mathbf{N}_n$$

Here we have tacitly assumed that all involved events have non-zero probability.

Independence

We have that the occurrence of the event $B \in \mathcal{F}$ changes the probability of the event $A \in \mathcal{F}$ to occur from $P(A)$ to $P(A|B)$.

If $P(A|B) = P(A)$ or equivalently

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

we say that A and B are *independent*.

This also implies that $P(B|A) = P(B)$.

Random Variables

A *random variable* or *stochastic variable* is a function from the sample space Ω to some set $\mathcal{D}$.

If this set is countable, like $\mathcal{D} = \mathbf{Z}$, we say that the random variable is *discrete*, and if it is uncountable, like $\mathcal{D} = \mathbf{R}$, we say that it is *continuous*.

We assume that the set $\mathcal{D}$ is partially ordered. This implies that the set $A = \{\omega \in \Omega : \ X(\omega) \leq x\}$ is well-defined.

We also assume that the random variable $X : \Omega \to \mathcal{D}$ is $\mathcal{F}$-*measurable*, i.e. the set $A \in \mathcal{F}$ for all $x \in \mathcal{D}$.

This means that the distribution function of the random variable $F : \mathcal{D} \to [0, 1]$ given by

$$F(x) = P(\{\omega : \ X(\omega) \leq x\})$$

is well-defined.

Random Variables ctd.

We will often use the shorter notation $F(x) = P(X \leq x)$.

When $X(\omega) = \omega$ the distribution function of the random variable is the same as the distribution function associated with probability space itself.

For n-dimensional random variable X we have

$$F(x) = P(\{\omega : \ X_1(\omega) \leq x_1, \dots, X_n(\omega) \leq x_n\})$$

Often we write $F(x) = P(X \leq x)$ where the inequality now is to be interpreted as component-wise inequalities.

The dimensions of Ω and of the random variable do not have to be the same, but sometimes they are.

Gaussian Distribution

We say that a random variable $X : \mathbf{R}^n \to \mathbf{R}^n$ defined as $X(\omega) = \omega$ that has pdf $f : \mathbf{R}^n \to \mathbf{R}_+$ given by

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{3}$$

where $m \in \mathbf{R}^n$ and $\Sigma \in \mathbf{S}^n_{++}$ has a *Gaussian* or *normal* distribution.[1]

To emphasize the dependence on parameters μ and Σ we use $\mathcal{N} : \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{S}^n_{++} \to \mathbf{R}_+$ defined as $\mathcal{N}(x, \mu, \Sigma) = f(x)$.

---

[1] We remark that it is the random variable that has a Gaussian distribution and that f is not a Gaussian distribution but a Gaussian pdf.

Marginal Distribution Function

For a two-dimensional random variable $X : \Omega \to \mathbf{R}^2$ we define the *marginal distribution functions* as

$$F_1(x_1) = P(X_1 \leq x_1) = F((x_1, \infty));\; F_2(x_2) = P(X_2 \leq x_2) = F((\infty, x_2))$$

where F is the distribution function for $X = (X_1, X_2)$, sometimes called the *joint distribution function*.

When F is differentiable, the *marginal probability density functions* satisfy

$$f_1(x_1) = \int_{x_2 \in \mathbf{R}} f(x_1, x_2) dx_2; \qquad f_2(x_2) = \int_{x_1 \in \mathbf{R}} f(x_1, x_2) dx_1$$

For discrete-valued random variables similar formulas hold using summation.

For n-dimensional random variables marginal pdf:s of dimension $n_1 < n$ obtained by integrating or summing over the remaining $n_2 = n - n_1$ variables.

Independent Random Variables

Two random variables $X : \Omega \to \mathbf{R}$ and $Y : \Omega \to \mathbf{R}$ are independent if the events $A = \{\omega : X(\omega) \leq x\}$ and $B = \{\omega : Y(\omega) \leq y\}$ are independent for all $x, y \in \mathbf{R}$.

The independence of the random variables X and Y is equivalent to

$$F(x, y) = F_X(x)F_Y(y)$$

and to

$$f(x, y) = f_X(x)f_Y(y)$$

For independence of $n > 2$ random variables we realize that we define this as independence of n events, and that the criteria in terms of distribution functions and probability density functions are that we can factorize them in n factors, where these factors are the marginals.

Functions of Random Variables

Consider a random variable $X : \Omega \to \mathcal{D}$. If $g : \mathcal{D} \to \mathcal{E}$ is such that $g(X(\omega))$ is $\mathcal{F}$-measurable, then $g(X)$ is a well-defined random variable that is a function of the random variable X.

In case the distribution function for X is $F : \mathcal{D} \to [0,1]$, then the distribution function $F_Y : \mathcal{E} \to [0,1]$ for $Y = g(X)$ can be obtained as

$$F_Y(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F(g^{-1}(y))$$

when g is invertible. For the general case of continuous random variables, e.g. when $\mathcal{D} = \mathcal{E} = \mathbf{R}$, it holds that

$$F_Y(y) = P(g(X) \leq y) = \int_{x : g(x) \leq y} f(x) dx$$

where $f : \mathbf{R} \to \mathbf{R}_+$ is the pdf of X.

Conditional Distribution and Probability Functions

Given two random variables $X : \Omega \to \mathcal{D}$ and $Y : \Omega \to \mathcal{E}$, where $\mathcal{D}$ and $\mathcal{E}$ are countable, we define the *conditional distribution function* $F_{Y|X} : \mathcal{E} \to [0,1]$ as

$$F_{Y|X}(y) = P\left(Y \le y | X = x\right)$$

for any x such that $P(X = x) > 0$.

We define the *conditional probability function* $p_{Y|X}(y) : \mathcal{E} \to [0,1]$ as

$$p_{Y|X}(y) = P\left(Y = y | X = x\right)$$

for any x such that $P(X = x) > 0$.

Sometimes we write $F_{Y|X}(y|x)$ and $p_{Y|X}(y|x)$ to emphasize the dependence on x. However, strictly speaking we have just defined functions of the variable y for all values of x, i.e. a family of functions. We may of course consider them to be functions of x as well if we so desire.

Some Formulas

It should be clear that

$$P\left(Y = y | X = x\right) = \frac{P\left(Y = y, X = x\right)}{P\left(X = x\right)}$$

which can be computed from the joint and marginal probability functions for (X, Y) and X, respectively, i.e.

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

where

$$p_{X,Y}(x, y) = P\left(Y = y, X = x\right)$$

and

$$p_X(x) = \sum_{y \in \mathcal{D}} p_{X,Y}(x, y)$$

It is straight forward to verify that

$$F_{Y|X}(y|x) = \sum_{z \leq y} p_{Y|X}(z|x)$$

Conditional Probability Density Functions

Consider random variables $X : \Omega \to \mathcal{D}$ and $Y : \Omega \to \mathcal{E}$, where $\mathcal{D}$ and $\mathcal{E}$ are uncountable, e.g. $\mathbf{R}^m$ and $\mathbf{R}^n$, respectively. Then

$$\begin{aligned} P(Y \leq y | x \leq X \leq x + dx) &= \frac{P(Y \leq y, x \leq X \leq x + dx)}{P(x \leq X \leq x + dx)} \\ &\approx \frac{\int_{v \leq y} f_{X,Y}(x,y) dx dv}{f_X(x) dx} \\ & \int_{v \leq y} \frac{f_{X,Y}(x,y)}{f_X(x)} dv \end{aligned}$$

for a small $dx > 0$, where $f_{X,Y} : \mathcal{D} \times \mathcal{E} \to \mathbf{R}_+$ is the joint pdf for (X, Y) and where $f_X : \mathcal{D} \to \mathbf{R}_+$ is the marginal pdf for X.

Conditional Probability Density Functions ctd.

As dx goes to zero we obtain $P(Y \leq y|X = x)$ and define the *conditional distribution function* $F_{Y|X} : \mathcal{E} \to [0,1]$ as

$$F_{Y|X}(y) = \int_{v \leq y} \frac{f_{X,Y}(x,y)}{f_X(x)} dv$$

The *conditional probability density function* $f_{Y|X} : \mathcal{E} \to \mathbf{R}_+$ is given by

$$f_{Y|X}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{4}$$

Expected Values

Let us assume that we are interested in estimation how frequent a certain event $A \subset \Omega$ is when the experiment is repeated. Hence it is enough to define $\mathcal{F} = \{\Omega, \emptyset, A, A^c\}$, where $A^c = \Omega \setminus A$. Let us assume that $P(A) = p$ and that $P(A^c) = 1 - p$. We then define the random variable $X : \Omega \to \{0, 1\}$ as $X(\omega) = 1$ when $\omega \in A$ and $X(\omega) = 0$ when $\omega \notin A$. If we repeat the experiment N times it is reasonable to estimate the relative frequency of A with the *sample average*

$$\frac{1}{N}\sum_{i=1}^{N} X(\omega_i)$$

where ω_i is the outcome of the $i$the experiment. We realize that this quantity is very close to

$$\begin{aligned} 0 \times P(\{\omega : X(\omega) = 0\}) + 1 \times P(\{\omega : X(\omega) = 1\}) \\ = 0 \times P(A^c) + 1 \times P(A) = p \end{aligned}$$

Expected Values ctd.

We define the *expected value* of any discrete random variable $X : \Omega \to \mathcal{D}$ as

$$E(X) = \sum_{x \in \mathcal{D}} x p(x)$$

where $p(x) = P(\{\omega : X(\omega) = x\})$.

The expected value is close to the sample average of the random variable for large values of N and is sometimes called the *mean* of the random variable.

For continuous random variables we define the expected value of a random variable $X : \Omega \to \mathbf{R}$ as

$$E(X) = \int_{x \in \mathbf{R}} x f(x) dx$$

where f is the pdf of the random variable.

Moments

For any scalar-valued random variable X we define the kth *moment* as

$$m_k = E\left(X^k\right)$$

and the kth *central moment* as

$$\mu_k = E\left((X - m_1)^k\right)$$

The moment m_1 is the mean of X and μ_2 is called the *variance* of X also denoted by σ^2 or $\mathrm{Var}(X)$, where σ is called the *standard deviation*

It holds that $\mu_2 = m_2 - m_1^2$.

Expected Value of Functions of Random Variables

For a function g of a random variable X it holds that

$$E(g(X)) = \int_{x \in \mathbf{R}} g(x) f(x) dx$$

for the continuous case, where f is the pdf for X.

For the special case $g(x) = Ax$, where $A \in \mathbf{R}^{m \times n}$ it holds that $E(AX) = AE(X)$.

Expectation is a *linear functional* on the space of random variables.

Correlation

For two scalar-valued random variables X and Y the product XY is a special case of a function g of the two-dimensional random variable (X, Y). Expected value if given by

$$E(XY) = \int_{(x,y)\in\mathbf{R}^2} xyf_{X,Y}(x, y)dxdy$$

where $f_{X,Y}$ is the pdf of (X, Y).

If X and Y are independent, then $E(XY) = E(X)E(Y)$.

The converse is in general not true.

In case $E(XY) = E(X)E(Y)$ we say that X and Y are *uncorrelated*.. Otherwise they are said to be *correlated*.

Vector-Valude Case

Generalization to vector-valued random variables obtained by considering outer product.

The *covariance* between two random variables is

$$\operatorname{Cov}(X, Y) = E(X - m_X)(Y - m_Y)^T = E\left(XY^T\right) - m_X m_Y^T$$

where m_X and m_Y are the means of X and Y respectively.

The variance of X is $\operatorname{Var}(X) = \operatorname{Cov}(X, X)$.

The random variables X and Y are uncorrelated if $E(XY^T) = E(X)\,(E(Y))^T = m_X m_Y^T$.

This is equivalent to $\operatorname{Cov}(X, Y) = 0$. As before independence implies being uncorrelated. Notice that X and Y can have different dimensions.

Gaussian Random Variable

For a variable X with Gaussian pdf

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

it holds that $E(X) = \mu$ and that $\mathrm{Var}(X) = \Sigma$.

Conditional Expectations

Given two random variables $X : \Omega \to \mathcal{D}$ and $Y : \Omega \to \mathcal{E}$ the *conditional expectation* of Y given $X = x$ is

$$E(Y|X = x) = \sum_{y \in \mathcal{E}} y p_{Y|X}(y|x)$$

for discrete random variables and

$$E(Y|X = x) = \int_{y \in \mathcal{E}} y f_{Y|X}(y|x) dy$$

for continuous random variables, where $p_{Y|X} : \mathcal{E} \to [0, 1]$ is the conditional probability function and where $f_{Y|X} : \mathcal{E} \to \mathbf{R}_+$ is the conditional pdf, respectively.

For a given value of x the conditional expectation is just a number, but it is possible to consider all values of $x \in \mathcal{D}$

Define $\Psi : \mathcal{D} \to \mathbf{R}$ as $\Psi(x) = E(Y|X = x)$.

Conditional Expectations

Let us now investigate the random variable $\Psi(X)$. Its expectation is for discrete random variables given by

$$E(\Psi(X)) = \sum_{x \in \mathcal{D}} \left(\sum_{y \in \mathcal{E}} y p_{Y|X}(y|x) \right) p_X(x) = \sum_{y \in \mathcal{E}} y \sum_{x \in \mathcal{D}} p_{X,Y}(x,y)$$

$$= \sum_{y \in \mathcal{E}} y p_Y(y) = E(Y)$$

where $p_{X,Y} : \mathcal{D} \times \mathcal{E} \to [0,1]$ is the joint probability function for (X,Y), and where $p_X : \mathcal{D} \to [0,1]$ is the marginal probability function for X.

The same result holds for continuous random variables.

We often write $\Psi(X) = E(Y|X)$ and hence the result can be summarized as

$$E(E(Y|X)) = E(Y)$$

Random Processes

Consider random variables $X : \Omega \to \mathcal{D}^{\mathbf{Z}_+}$. Hence the random variable is a function of $k \in \mathbf{Z}_+$, i.e. $(\omega, k) \mapsto X(\omega, k)$.

We often instead interpret it as an infinite-dimensional vector $X = (X(\omega, 0), X(\omega, 1), \ldots)$ and we often instead write $X_k(\omega) = X(k, \omega)$ to ease the notation.

We may interpret the random variable X as an infinite sequence of random variables $X_k : \Omega \to \mathcal{D}$, $k \in \mathbf{Z}_+$. We may interpret $k \in \mathbf{Z}_+$ as a discrete time-index.

Such a random variable X is called a *discrete-time random process*. or *stochastic process*

The set $\mathcal{D}$ could be a finite set like $\mathbf{N}_n$, a countable set like $\mathbf{Z}$ or an uncountable set like $\mathbf{R}$. It may also be vector-valued.

Two Views

We may observe the evolution of a random process in two different ways.

For each fixed outcome $\omega \in \Omega$, we obtain a *realization* or *sample path* $X(\omega)$ of X at ω. We can study properties of this sample path.

Alternatively we investigate a finite subset of components of the infinite dimensional vector X, i.e. let $K = \{k_1, k_2, \ldots, k_n\} \subset \mathbf{Z}_+$. Then consider the joint distribution function $F_K : \mathcal{D}^n \to [0,1]$ defined as

$$F_K(x) = P(X_{k_1} \leq x_1, \ldots, X_{k_n} \leq x_n)$$

The collection $\{F_K\}$ where K ranges over all finite-dimensional $K \subset \mathbf{Z}_+$ is called the collection of *finite-dimensional distribustions* (fdds) of X or the *name* of X. This contains all the information that is available about X from finitely many components X_k.

Continuous Time Random Processes and Random Fields

If we replace $\mathbf{Z}_+$ with $\mathbf{R}_+$ we obtain a *continuous-time random process*.

We then often write $X(t)$ instead of X_k, where $t \in \mathbf{R}_+$. The fdds are the often denoted F_T where T is now a finite subset of $\mathbf{R}_+$.

An even more general concept is a *random field*, which is obtained by replacing $\mathbf{Z}_+$ with $\mathbf{R}^n$.

Stationarity

A discrete-time random process is *strongly stationary* if $\{X_{k_1}, \ldots, X_{k_n}\}$ and $\{X_{k_1+l}, \ldots, X_{k_n+l}\}$ have the same joint distribution for all $k_1, \ldots, k_n$ and $l > 0$.

A discrete-time random process is *weakly stationary* if $E(X_{k_1}) = E(X_{k_2})$ and $\mathrm{Cov}(X_{k_1}, X_{k_2}) = \mathrm{Cov}(X_{k_1+l}, X_{k_2+l})$ for all k_1, k_2 and $l > 0$.

A random process is weakly stationary if and only if it has a constant mean and the *autocovariance function* $c : \mathbf{Z}_+^2 \to \mathbf{R}$ given by

$$c(k, k+l) = \mathrm{Cov}(X_k, X_{k+l})$$

satisfies

$$c(k, k+l) = c(0, l)$$

for all k and $l \geq 0$. For weakly stationary processes we define the autocovariance function as a function of only l and write $c : \mathbf{Z}_+ \to \mathbf{R}$.

Stationarity ctd.

Strong stationarity implies weak stationarity.

One example when strong stationarity is equivalent to weak stationarity is when the fdds are all Gaussian.

The definitions of weak and strong stationarity for a continuous-time random process are similar as for a discrete -time random process, and this also goes for a random field.

Markov Processes

Consider a discrete-time random processes $X : \Omega \to \mathcal{D}^{\mathbf{Z}_+}$ which satisfies the so-called *Markov property*

$$P(X_{k_n} \leq x_n \mid X_{k_{n-1}} \leq x_{n-1}) = P(X_{k_n} \leq x_n \mid X_{k_{n-1}} \leq x_{n-1}, \ldots, X_{k_1} \leq x_1) \tag{5}$$

for all $k_1 \leq k_2 \leq \cdots \leq k_n$. Such random processes are called *Markov processes*.

Example for $\mathcal{D} = \mathbf{R}^n$:

$$X_{k+1} = AX_k + E_k, \quad k \in \mathbf{Z}_+$$

where $A \in \mathbf{R}^{n \times n}$, E_k are i.i.d. n-dimensional random vectors, and where X_0 is a random vector with known distribution.

Markov Chain

When $\mathcal{D}$ is a finite set like $\mathbf{N}_N$ the process is often called a *Markov chain*. We say that a Markov process is *homogeneous* if

$$P(X_n \leq x \mid X_{n-1} \leq y) = P(X_1 \leq x \mid X_0 \leq y)$$

for all x, y and n. For a homogeneous Markov chain we define the *transition matrix* P with elements p_{ij} called the *transition probabilities* defined as

$$p_{ij} = P(X_n = j \mid X_{n-1} = i)$$

Chapman-Komogorov Equation

The *n-step transition matrix* P_n has elements $p_{ij}(n)$ called the *n-step transition probabilities* defined as

$$p_{ij}(n) = P(X_{m+n} = j \mid X_m = i)$$

It holds that $P_{m+n} = P_m P_n$, which is called the *Chapman-Kolmogorov equation*, and hence $P_n = P^n$.

If π^k is the row vector of

$$\pi_i^k = P(X_k = i)$$

it follows that $\pi^k = \pi^0 P^k$ for $k \geq 0$. Hence the transition matrix fully characterizes a homogeneous Markov chain.

Stationary Solution

If $\pi \in \mathbf{R}^{1\times N}$ is such that $\pi_i \geq 0$, $\sum_i \pi_i = 1$ and $\pi = \pi P$, then π is called a *stationary distribution* for the Markov chain.

If $\pi^0 = \pi$, then $\pi^k = \pi$ for all $k \geq 0$, i.e. the distribution remains the same for all times, and hence the Markov chain is for this initial distribution a stationary random process, both weakly and strongly.

For other initial values the Markov chain may or may not converge to the stationary distribution as time goes to infinity.