Stochastic Signals and Systems

Contents

1 Probability Theory

1.1 Terminology

$\mathfrak{E}$: random experiment (activity where the outcome is randomly influenced)

Ξ: sample space (set of all possible outcomes of $\mathfrak{E}$ which may consist of a finite, infinite countable or uncountable number of elements)

ξ: elementary event (possible outcome of $\mathfrak{E}$, i.e. $\xi \in \Xi$)

$\emptyset$: impossible event (empty set $\emptyset = \{\ \}$)

E: event (collection of some of the possible outcomes of $\mathfrak{E}$, i.e. $E \subset \Xi$)

$\mathbb{S}$: σ-field, i.e. a system of subsets of Ξ satisfying

1) $\Xi \in \mathbb{S}$
2) if $E \in \mathbb{S}$ then $\bar{E} = \Xi \backslash E \in \mathbb{S}$
3) if $E_i \in \mathbb{S}$ for $i = 1, 2, \ldots$ then $\bigcup_{i=1}^{\infty} E_i \in \mathbb{S}$

Corollary:

1) $\emptyset \in \mathbb{S}$
2) if $E_1, E_2 \in \mathbb{S}$ then $E_1 \cap E_2 \in \mathbb{S}$ and $E_1 \backslash E_2 \in \mathbb{S}$
3) if $E_i \in \mathbb{S}$ for $i = 1, 2, \ldots$ then $\bigcap_{i=1}^{\infty} E_i \in \mathbb{S}$

$(\Xi, \mathbb{S})$: measurable space

1.2 Definition of Probability

1.2.1 Relative Frequency and Probability

If a random experiment is performed n times and where the event of interest E is observed with frequency $h_n(E)$, then the relative frequency of the occurrence of E is defined by

$$H_n(E) = \frac{h_n(E)}{n} \quad \text{with} \quad 0 \le H_n(E) \le 1.$$

Empirical law of large numbers

For sufficiently large n we can write with a high degree of certainty that

$$P(E) \cong H_n(E).$$

1.2.2 Axiomatic Approach to Probability

Consider an experiment $\mathfrak{E}$ with measurable space $(\Xi,\mathbb{S})$. A probability measure P is then defined as a mapping

$$P:\ \mathbb{S} \to \mathbb{R}$$

which satisfies the following axioms

1) if $E \in \mathbb{S}$ then $P(E) \geq 0$,

2) $P(\Xi) = 1$,

3) if $E_i \in \mathbb{S}$ for $i = 1,2,\ldots$ and $E_i \cap E_j = \varnothing$ for $i \neq j$ then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

The triple $(\Xi, \mathbb{S}, P)$ is called probability space.

Implications:

1) $P(\varnothing) = 0$,
2) if $E_1, E_2 \in \mathbb{S}$ with $E_1 \subset E_2$ then
$P(E_1) \le P(E_2)$,
3) if $E_1, E_2 \in \mathbb{S}$ then
$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.

1.2.3 Classical Definition of Probability

Suppose that an experiment has a finite number *n* of possible outcomes, $\xi_1, \xi_2, \ldots, \xi_n$ and we are interested in an event $E = \{\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}\}$ with $\{i_1, i_2, \ldots, i_m\} \subset \{1, 2, \ldots, n\}$. If we assume that all outcomes $\xi_1, \xi_2, \ldots, \xi_n$ are equally likely, then

$$P(E) = \frac{\text{number of outcomes favorable to } E}{\text{total number of outcomes}} = \frac{m}{n}.$$

This is a basic result which assigns probabilities to events purely on the basis of combinatorial arguments.

However, its application is strictly limited to experiments of a finite number of equally likely outcomes.

1.3 Conditional Probability

Let $(\Xi, \mathbb{S}, P)$ be a probability space with $E_1, E_2 \in \mathbb{S}$ and $P(E_2) > 0$. The conditional probability of E_1 given that E_2 occurred is defined by

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}.$$

One can easily show that the conditional probability satisfies the axioms of a probability measure.

Implications:

1) Bayes' Formula

$$P(E_2 \mid E_1) = P(E_1 \mid E_2)\, P(E_2) \,/\, P(E_1).$$

Furthermore, assuming

$$E_i \cap E_j = \emptyset,\ i \neq j \quad \text{and} \quad E \subset \bigcup_i E_i$$

we can derive the Total Probability

$$P(E) = \sum_i P(E \mid E_i) P(E_i)$$

and the generalised Bayes' Formula

$$P(E_k \mid E) = \frac{P(E_k)P(E \mid E_k)}{\sum_i P(E_i)P(E \mid E_i)}, \quad P(E) > 0.$$

2) Two events E_1 and E_2 are called independent if

$$P(E_1 \mid E_2) = P(E_1)$$

holds. Consequently, we can stipulate

$$P(E_1 \cap E_2) = P(E_1)\, P(E_2).$$

1.4 Random Variables

A mapping

$$X:\ \Xi \to \mathbb{R},$$

such that to each $\xi \in \Xi$ there corresponds a unique real number $X(\xi) \in \mathbb{R}$, is called random variable or measurable function with respect to $\mathbb{S}$, if for each set $B \subset \mathbb{R}$ the inverse image

$$X^{-1}(B) = \{\xi :\ X(\xi) \in B\}$$

is element of $\mathbb{S}$.

To assign probabilities to random variables one has to translate statements about the values of random variables as follows.

$$P_X(B) = P(X^{-1}(B)) = P(\{\xi : X(\xi) \in B\})$$

Furthermore, a σ-field has to be defined over $\mathbb{R}$. One can show that such a σ-field should include all intervals of the kind $(-\infty, x]$.

The power set $\mathbb{P}(\mathbb{R})$ includes the desired intervals but its cardinality is to high to be able to implement the measurability properties.

However, one can show that a particular σ-field exists, called Borel-field $\mathbb{B}$, that is the smallest possible including all the intervals $(-\infty, x]$ and that guarantees the measurability of all sets element of $\mathbb{B}$

Thus, we can define by

$(\mathbb{R},\mathbb{B})$ the measurable space

and

$(\mathbb{R},\mathbb{B},P_X)$ the probability space

of a random variable X.

1.5 Distribution Functions

Given a random variable X, the distribution function of X, $F_X(x)$, is defined by

$$F_X(x) = P_X((-\infty, x]) = P(\{\xi : X(\xi) \le x\}) = P(X \le x).$$

One can show that $F_X(x)$ uniquely determines all the probabilistic properties of the random variable X.

In particular, for any $a, b \in \mathbb{R}$ with $a \le b$ we have

$$P(X \le b) = P(X \le a) + P(a < X \le b),$$

cf. Axiom 3. Hence

$$P(a < X \le b) = P(X \le b) - P(X \le a) = F_X(b) - F_X(a).$$

Distribution functions have the properties:

(1) $0 \leq F_X(x) \leq 1$ for all $x \in \mathbb{R}$,
(since $F_X(x)$ is a probability)

(2) $\lim_{x\to-\infty} F_X(x) = 0, \lim_{x\to\infty} F_X(x) = 1,$
(since $\lim_{x\to-\infty} X^{-1}((-\infty, x]) = \emptyset \wedge \lim_{x\to\infty} X^{-1}((-\infty, x]) = \Xi$)

(3) $F_X(x)$ is a non-decreasing function, i.e. for any $h \geq 0$ and all x, $F_X(x+h) \geq F_X(x)$,
(since $F_X(x+h) - F_X(x) = P(x < X \leq x+h) \geq 0$)

(4) $F_X(x)$ is right-continuous, i.e. for all x

$$\lim_{h\to 0+} F_X(x+h) = F_X(x).$$

(the limit $h \to 0$ is taken through positive values only)

Any distribution function $F_X(x)$, can be expressed by

$$F_X(x) = a_1\, F_{X,1}(x) + a_2\, F_{X,2}(x) + a_3\, F_{X,3}(x),$$

where

$$a_i \geq 0 \text{ for } i = 1, 2, 3, \quad a_1 + a_2 + a_3 = 1$$

and

$F_{X,1}(x)$ is continuous everywhere and differentiable for almost all x, i.e. absolute continuous,

$F_{X,2}(x)$ is a step-function with a finite or countable infinite number of jumps,

$F_{X,3}(x)$ is a singular function, that is continuous with zero derivative almost everywhere.

$F_{X,1}(x)$ and $F_{X,2}(x)$ correspond to the two basic types of probability distributions one usually encounters in practise, i.e. the

continuous and *discrete* distribution,

respectively.

Since $F_{X,3}(x)$ is highly pathological, it can be safely assumed that it does not arise in real applications.

In practice we therefore ignore $F_{X,3}(x)$ and assume that all distribution functions can be simply represented by

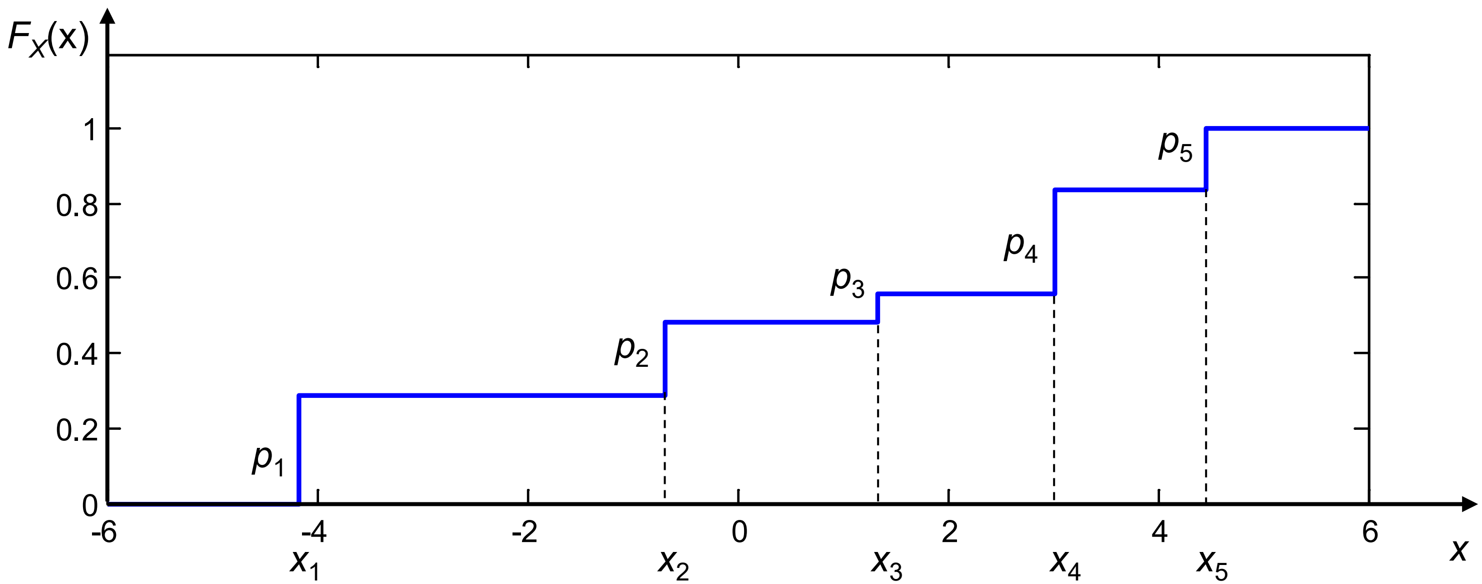
$$F_X(x) = \lambda F_{X,1}(x) + (1 - \lambda)\, F_{X,2}(x)$$

with $0 \le \lambda \le 1$.

1.5.1 Purely discrete case ($\lambda = 0$)

The distribution function $F_X(x) = F_{X,2}(x)$ is a simple step-function with jumps p_i at the points x_i for $i = 1, 2, 3, \ldots$

$F_X(x)$ would typically have the form



If an interval $(a, b]$ does not contain any of the jump points x_i, then clearly

$$P(a < X \leq b) = F_X(b) - F_X(a) = 0.$$

Hence, X cannot take any value lying between to successive jump points.

For each i and any small $h > 0$ we can write

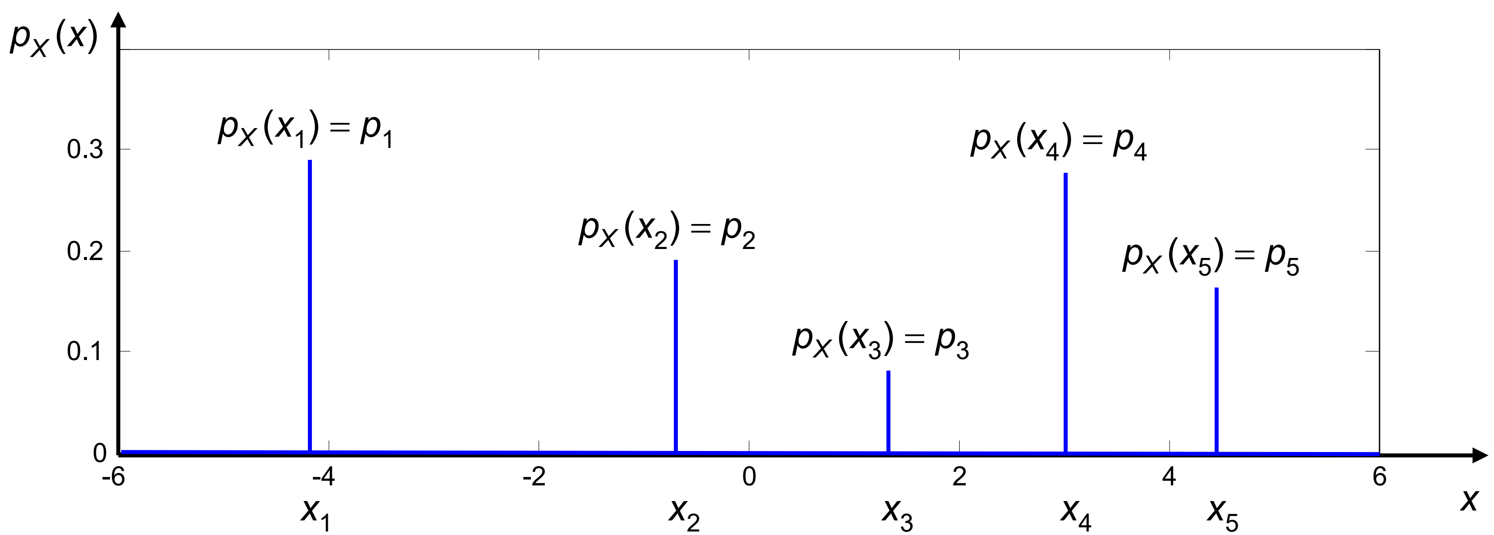
$$P(x_i - h < X \leq x_i + h) = F_X(x_i + h) - F_X(x_i - h) = p_i\,.$$

Letting $h \to 0$, we obtain

$$P(X = x_i) = p_i \quad i = 1, 2, 3, \ldots$$

Thus the only values X can take are those corresponding to the jump points. Therefore, X is called discrete random variable.

The jump p_i at point x_i represents the probability that X takes the value x_i. Furthermore, $(x_1,p_1),(x_2,p_2),\ldots$ are used to define the so-called probability mass function $p_X(x)$.



Discrete distribution functions possess the properties:

(1) $F_X(x) = \sum_{i, x_i \le x} p_X(x_i)$, where the summation extents over all values of i for which $x_i \le x$,

(2) $0 \le p_X(x) \le 1$,

(since $p_X(x)$ is a probability mass function)

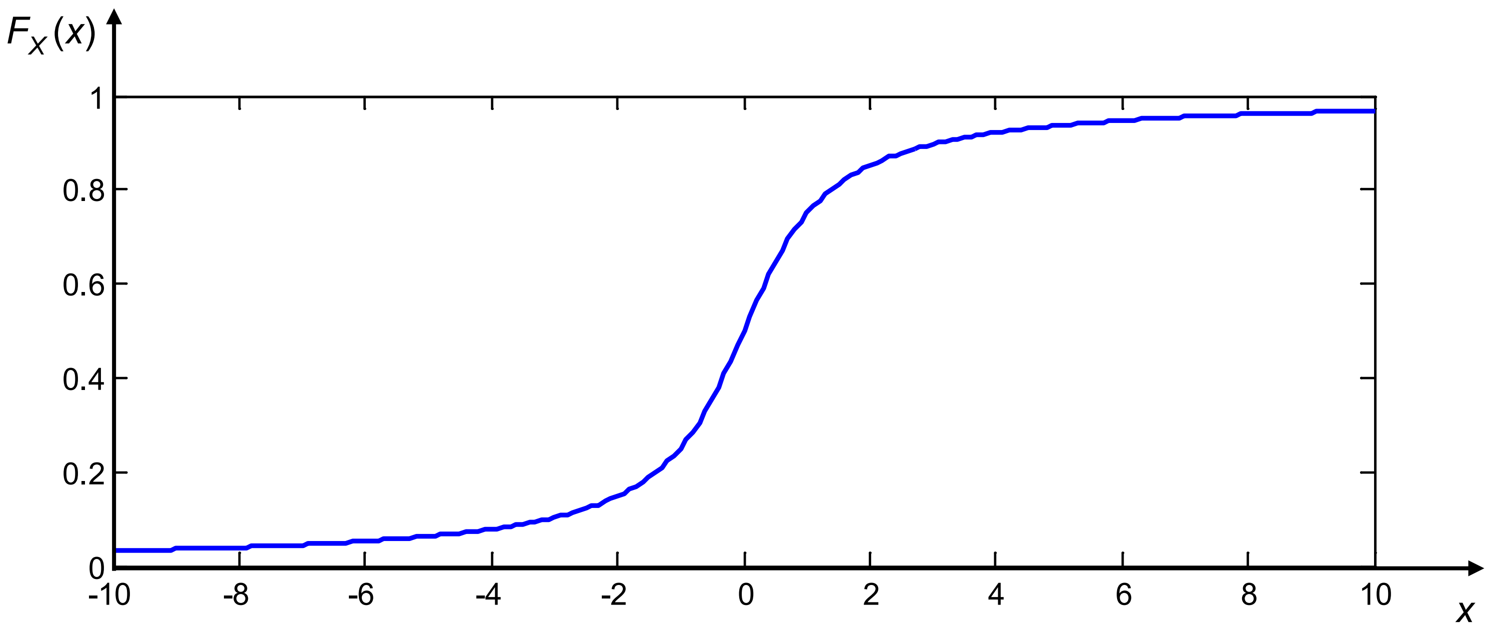
(3) $\sum_i p_X(x_i) = 1$.

(since $\lim_{x\to\infty} F_X(x) = \Sigma_i\, p_X(x_i) = 1$)

1.5.2 Purely continuous case ($\lambda = 1$)

The distribution function $F_X(x) = F_{X,1}(x)$ is absolutely continuous, i.e. differentiable for almost all x.

$F_X(x)$ would typically possess a graph as shown below.



X can, in general, take any value either on a finite or an infinite interval and is therefore called continuous random variable.

Thus continuous random variables are suitable models for measuring physical quantities such as pressures, voltages, temperatures, etc.

Furthermore, $F_X(x)$ can be represented by

$$F_X(x) = \int_{-\infty}^{x} f_X(x')\,dx',$$

where $f_X(x)$ is said to be the probability density function (PDF) of X.

If $f_X(x)$ is continuous at x, then

$$F_X'(x) = \frac{dF_X(x)}{dx} = f_X(x)$$

exists. For a small interval $(x, x+\Delta x]$ we can now write

$$P(x < X \leq x+\Delta x) = F_X(x+\Delta x) - F_X(x) = \int_x^{x+\Delta x} f_X(x')dx'$$

or

$$P(x < X \leq x+\Delta x) = f_X(x)\Delta x + \mathrm{o}(\Delta x),$$

where $\mathrm{o}(\Delta x)$ represents a term of smaller order of magnitude than Δx.

The latter equation forms the basis for interpreting $f_X(x)$ as a density function, namely, $f_X(x)$ defines the density of probability in the neighbourhood of the point x.

Remarks:

- $f_X(x)$ itself does not represent a probability,
- $f_X(x) \cdot \Delta x$ has a probabilistic interpretation,
- $f_X(x)$ completely determines $F_X(x)$ and therefore completely specifies the properties of a continuous random variable.

Probability density functions satisfy the properties:

(1) $f_X(x) \geq 0$ for all $x \in \mathbb{R}$,

(since $F_X(x)$ is a non-decreasing function)

(2) $\int_{-\infty}^{\infty} f_X(x)\,dx = 1,$

(since $\lim_{x\to\infty} F_X(x) = \int_{-\infty}^{\infty} f_X(x)\,dx = 1$)

(3) For any $a, b \in \mathbb{R}$ with $a \leq b$

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx.$$

1.6 Some Special Distributions

1.6.1 Discrete Distributions

Binomial distribution

Consider an experiment which has only two possible outcomes, “success” and “failure”, with probability p and $(1-p)$, respectively.

The number of “successes” occurring in n independent repetitions of the experiment is a random variable X that can take the values $k = 0,1,\ldots,n$.

Within a sequence of n independent trials, k successes can occur in

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

different arrangements. The probability of a specific arrangement is obviously

$$p^k(1-p)^{n-k}.$$

Thus, the probability for observing k successes in n independent trials is given by

$$P(\{\xi : X(\xi) = k\}) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = b_{n,p}(k).$$

The $b_{n,p}(k)$ $(k = 0, 1, \ldots, n)$ are called binomial probabilities.

Exploiting the binomial theorem one can verify that

$$\sum_{k=0}^{n} b_{n,p}(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \big(p + (1-p)\big)^n = 1.$$

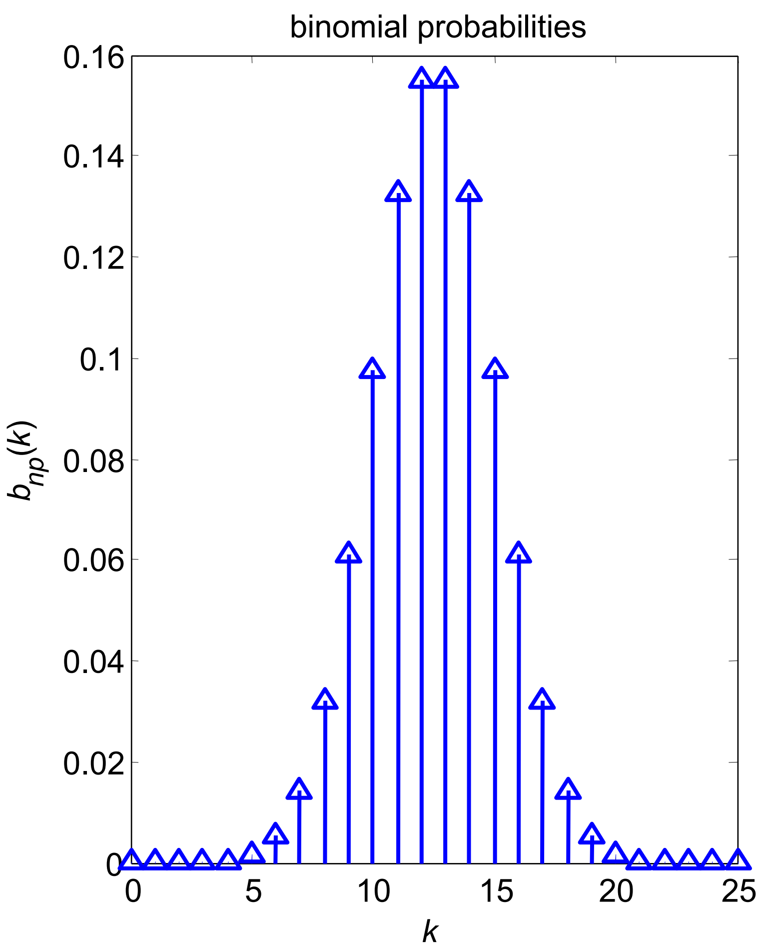
Thus, the distribution function of the so-called binomial distribution $\mathcal{B}(n,p)$ can be defined by

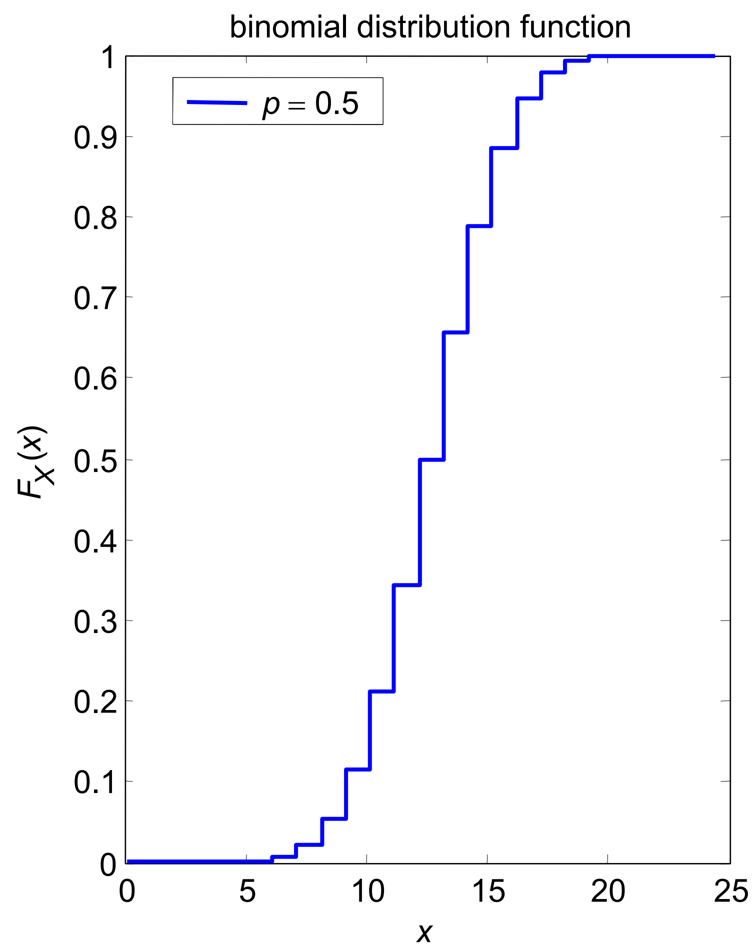
$$F_X(x) = P(X \le x) = \sum_{k=0}^{m} b_{n,p}(k) = \sum_{k=0}^{m} \binom{n}{k} p^k (1-p)^{n-k},$$

where $m = \lfloor x \rfloor \in \mathbb{Z}$, i.e. $m \le x < m+1$ (Gauss bracket).

Example:

1,0,0,0,0 | 1,1,0,0,0; 1,0,1,0,0; 1,0,0,1,0; 1,0,0,0,1 |
0,1,0,0,0 | 1,1,0,0,0; 0,1,1,0,0; 0,1,0,1,0; 0,1,0,0,1 |
0,0,1,0,0 | 1,0,1,0,0; 0,1,1,0,0; 0,0,1,1,0; 0,0,1,0,1 | …
0,0,0,1,0 | 1,0,0,1,0; 0,1,0,1,0; 0,0,1,1,0; 0,0,0,1,1 |
0,0,0,0,1 | 1,0,0,0,1; 0,1,0,0,1; 0,0,1,0,1; 0,0,0,1,1 |


binomial probabilities
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
k
$b_{np}(k)$


binomial distribution function
$p = 0.5$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
x
$F_X(x)$

Poisson distribution

Consider the limiting form of the binomial probabilities when $n \to \infty$ and $p \to 0$ in such a way that $np = \alpha_n \to \alpha$, a positive constant.

Substituting p by α_n / n, we obtain

$$b_{n,\frac{\alpha_n}{n}}(k) = \frac{n!}{k!(n-k)!}\left(\frac{\alpha_n}{n}\right)^k \left(1-\left(\frac{\alpha_n}{n}\right)\right)^{n-k}$$

$$= \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n}\left(1-\left(\frac{\alpha_n}{n}\right)\right)^{-k} \left\{\left(1-\left(\frac{\alpha_n}{n}\right)\right)^n \frac{\alpha_n^k}{k!}\right\}.$$

As n tends to infinity, we have

$$b_{n,\frac{\alpha_n}{n}}(k) \xrightarrow[n\to\infty]{} e^{-\alpha}\frac{\alpha^k}{k!} = p_\alpha(k).$$

The $p_\alpha(k)$ $(k = 0,1,\ldots)$ are called Poisson probabilities.

Note that the sum of the infinitely many but countable Poisson probabilities satisfies

$$\sum_{k=0}^{\infty} p_\alpha(k) = e^{-\alpha}\sum_{k=0}^{\infty}\frac{\alpha^k}{k!} = e^{-\alpha}e^{\alpha} = 1$$

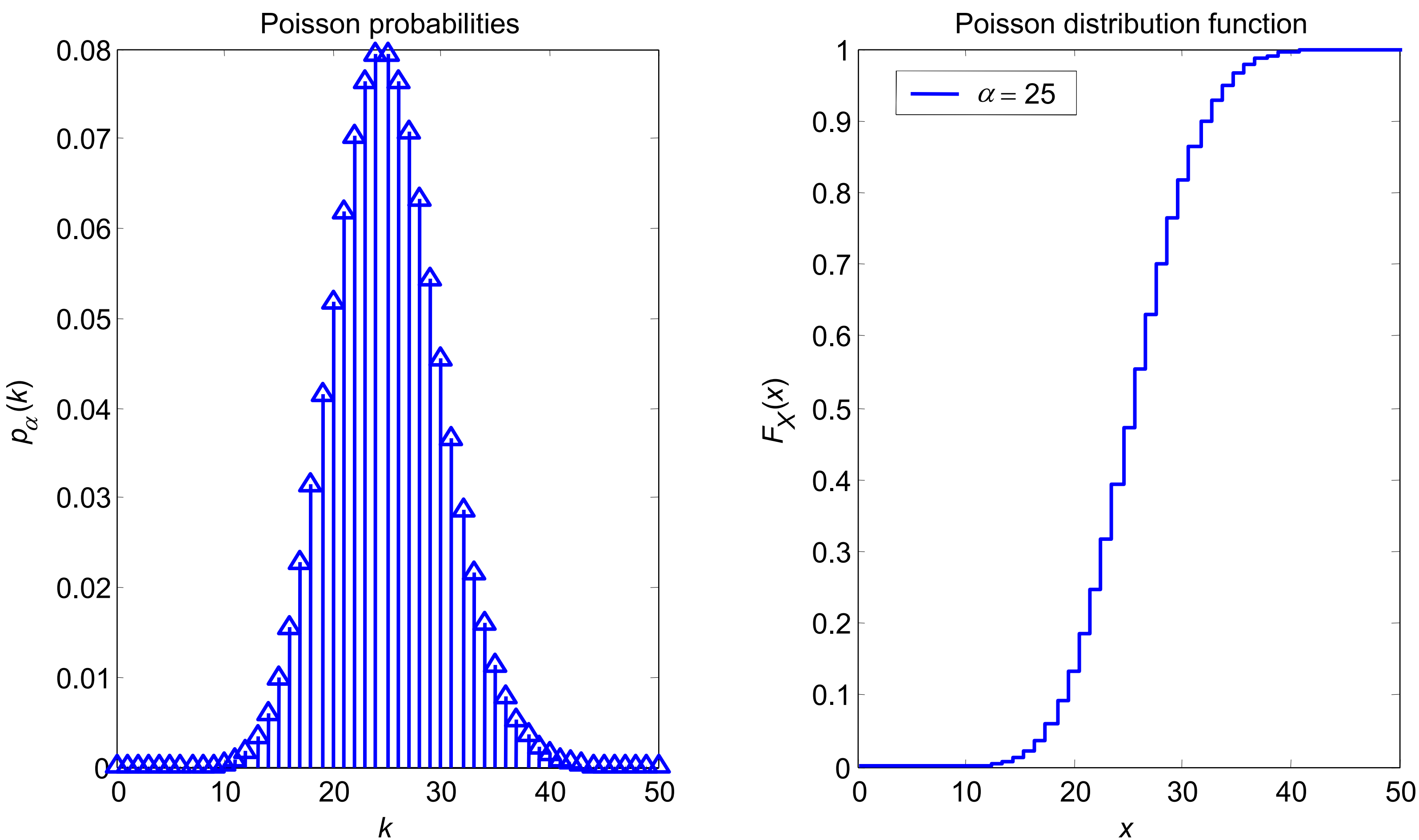
for all α.

Hence, the distribution function of the so-called Poisson distribution $\mathcal{P}(\alpha)$ is given by

$$F_X(x) = P(X \leq x) = \sum_{k=0}^{m} p_\alpha(k) = e^{-\alpha} \sum_{k=0}^{m} \frac{\alpha^k}{k!},$$

where $m = \lfloor x \rfloor \in \mathbb{Z}$, i.e. $m \leq x < m+1$ (Gauss bracket).

In practice, the Poisson distribution is used for approximating the binomial distribution in cases, where in a large number of independent trials (large n) the number of occurrences of a rare event (small p) is of interest.


Poisson probabilities
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
$p_\alpha(k)$
0
10
20
30
40
50
k
Poisson distribution function
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
$\alpha = 25$
$F_X(x)$
0
10
20
30
40
50
x

1.6.2 Continuous Distributions

Uniform (rectangular) distribution

A continuous random variable X is uniformly distributed on the interval $[a,b]$ (in abbreviated form $X \sim \mathcal{R}(a,b)$), if the probability density function is defined by

$$f_X(x) = \frac{1}{b-a} 1_{[a,b]}(x) \qquad x \in \mathbb{R},$$

where $1_M(x)$ denotes the indicator function of the set $M \subset \mathbb{R}$, i.e.

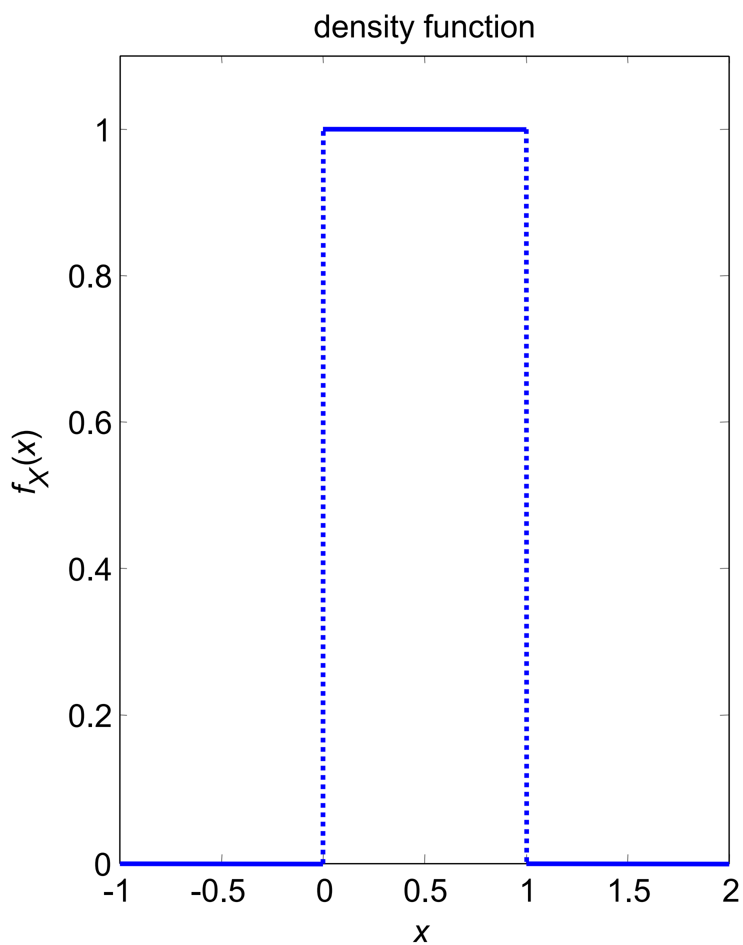
$$1_M(x) = \begin{cases} 1 & x \in M \\ 0 & \text{otherwise} \end{cases} .$$

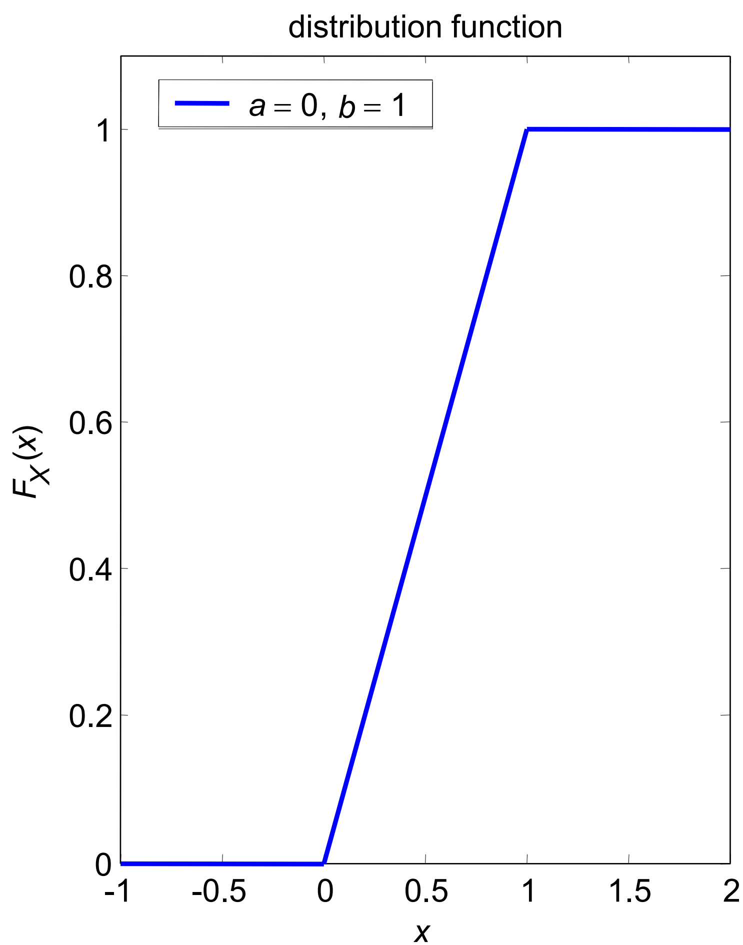
The distribution function can be expressed as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(x')dx' = \frac{1}{b-a}\int_{-\infty}^{x} 1_{[a,b]}(x')dx'$$

$$= \begin{cases} 0 & x \le a \\ (x-a)/(b-a) & x \in [a,b] \\ 1 & x \ge b \end{cases}.$$

A uniformly distributed random variable $X \sim \mathcal{R}(-\pi,\pi)$ is often used for modeling a random initial phase of a sinu-soidal signal.


density function
$f_X(x)$
1
0.8
0.6
0.4
0.2
0
-1
-0.5
0
0.5
1
1.5
2
x


distribution function
$a = 0,\ b = 1$
$F_X(x)$
1
0.8
0.6
0.4
0.2
0
-1
-0.5
0
0.5
1
1.5
2
x

Normal (Gaussian) distribution

A continuous random variable X is said to be normally distributed with parameters $\mu \in \mathbb{R}$ and σ^2 ($X \sim \mathcal{N}(\mu,\sigma^2)$), if the probability density function is defined by

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

The normal distribution function

$$F_X(x) = \int_{-\infty}^{x} f_X(x')\,dx' = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left\{-\frac{(x'-\mu)^2}{2\sigma^2}\right\} dx'$$

can not be expressed in explicit form.

However, to check that $f_X(x)$, where $f_X(x) > 0 \ \forall x \in \mathbb{R}$, represents a valid form of a probability density function we have to show that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

With the substitution $x = (x' - \mu)/\sigma$ we can derive

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x'-\mu)^2}{2\sigma^2}\right\} dx' = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2}{2}\right\} dx.$$

Since $f_X(x) > 0 \ \ \forall x \in \mathbb{R}$, it is equivalent to proof that

$$\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2}{2}\right\} dx\right)^2 = 1.$$

Thus, after introducing a double integral, employing polar coordinates and finally substituting $u = r^2/2$, the validity of the equation can be shown:

$$\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-x^2/2)\,dx \right)^2 =$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-(x^2 + y^2)/2\right\} dx\,dy$$

$$= \frac{1}{2\pi} \int_{0}^{2\pi} \int_{0}^{\infty} \exp(-r^2/2)\, r\,dr\,d\varphi$$

$$= \int_{0}^{\infty} \exp(-r^2/2)\, r\,dr = \int_{0}^{\infty} \exp(-u)\,du = 1.$$

The special form $\mathcal{N}(0,1)$, i.e. $\mu = 0$ and $\sigma^2 = 1$, is called standardized normal distribution.

Its distribution function is usually denoted by $\Phi(x)$, i.e.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{x'^2}{2}\right\} dx'.$$

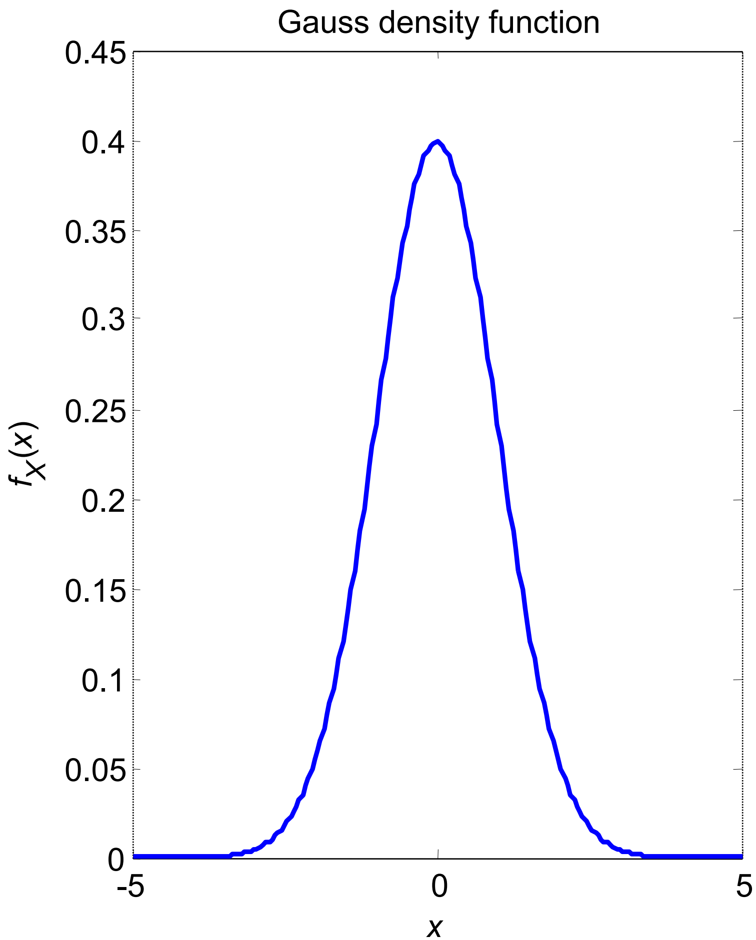
There are extensive tables of the function $\Phi(x)$ available in the literature. These tables enable us to evaluate the distribution function of $X \sim \mathcal{N}(\mu,\sigma^2)$ as follows.

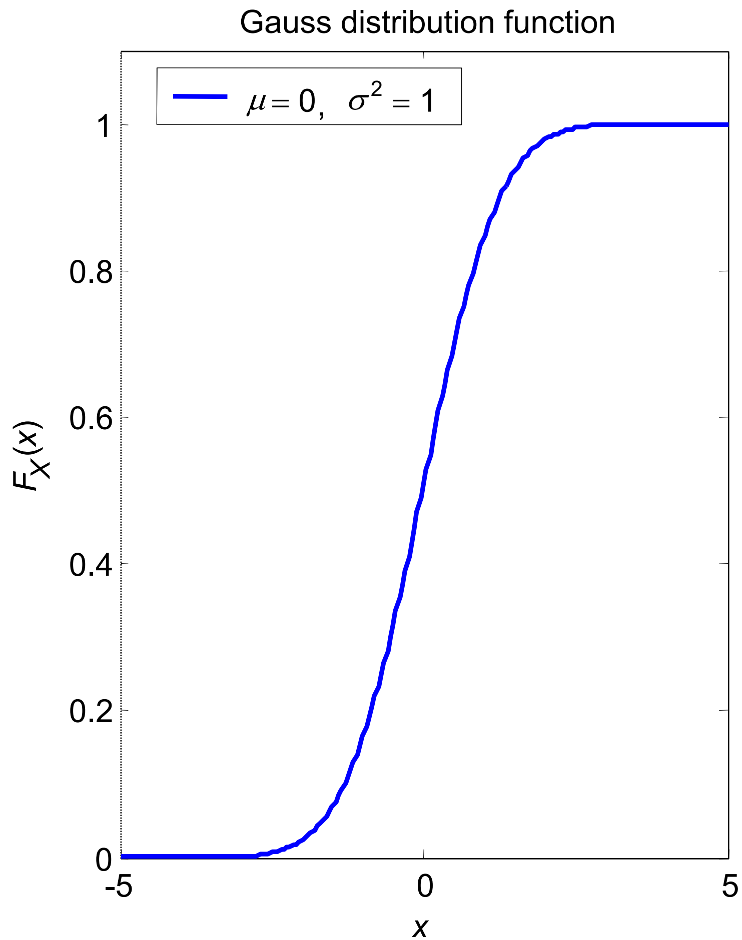
Let $X \sim \mathcal{N}(\mu,\sigma^2)$ and therefore $(X-\mu)/\sigma \sim \mathcal{N}(0,1)$ then

$$F_X(x) = P\left(\{\xi : -\infty < X(\xi) \le x\}\right)$$
$$= P\left(\left\{\xi : -\infty < \frac{X(\xi)-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right\}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

The normal distribution is by far the most important distribution in probability theory and statistical inference.

Its prominence is attributed to central limit theorems, which roughly state, that the sum of a large number of (independent) random variables obeys an approximate normal distribution.


Gauss density function
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-5
0
5
x
$f_X(x)$


Gauss distribution function
$\mu = 0, \quad \sigma^2 = 1$
1
0.8
0.6
0.4
0.2
0
-5
0
5
x
$F_X(x)$

Exponential distribution

A continuous random variable X, taking positive values only, is said to satisfy an exponential distribution with parameter $\lambda > 0$ ($X \sim \mathcal{E}(\lambda)$), if its probability density function is of the form

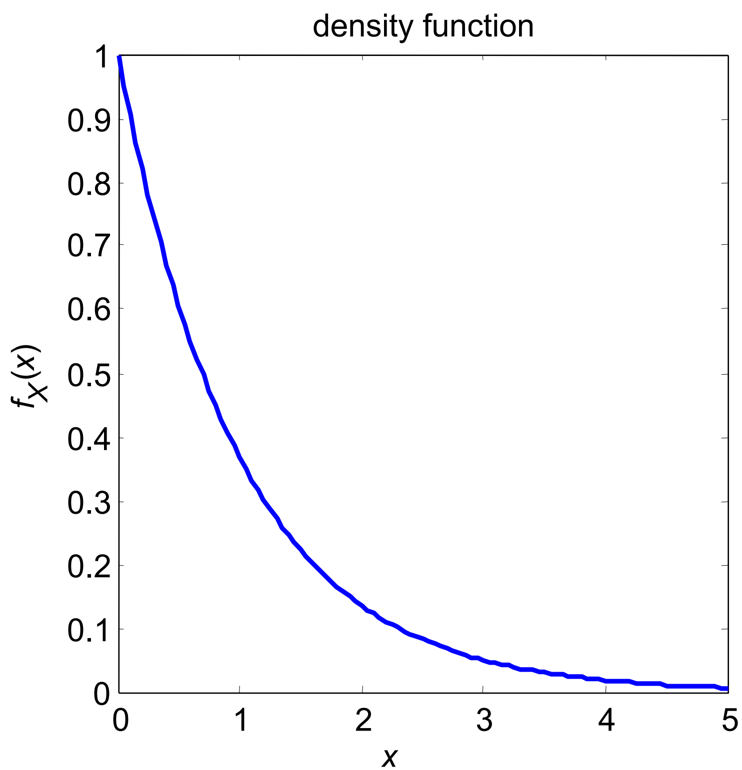
$$f_X(x) = \lambda \exp(-\lambda x) 1_{[0,\infty)}(x) \qquad x \in \mathbb{R}.$$

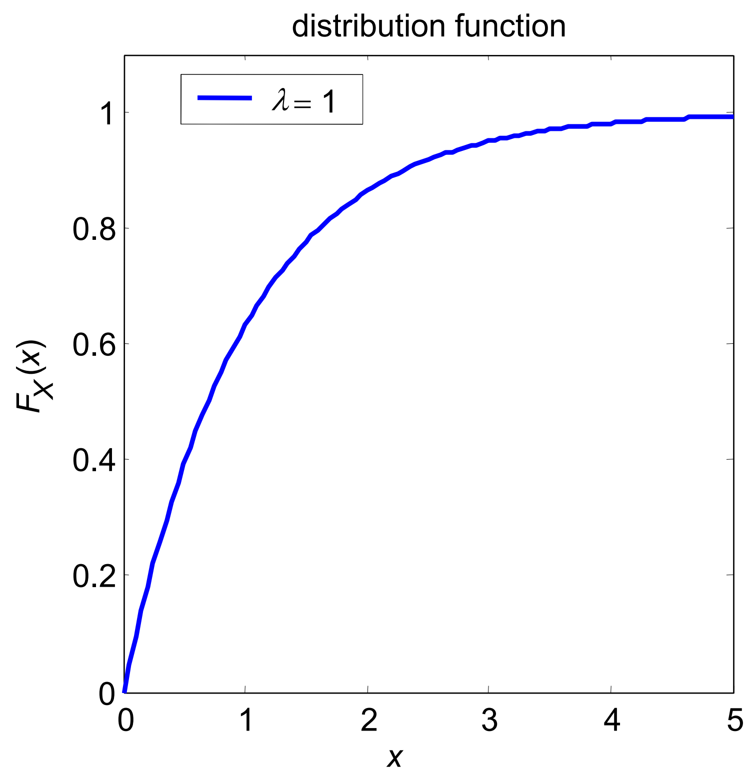
Hence, its distribution function is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(x')dx' = \int_{-\infty}^{x} \lambda \exp(-\lambda x') 1_{[0,\infty)}(x')dx'$$

$$= [1 - \exp(-\lambda x)] 1_{[0,\infty)}(x).$$

The exponential distribution is used as a model for the life times of items when ageing processes can be neglected.





Weibull distribution

When either ageing or initial failures processes have to be taken into account, a suitable model for the life time of an item can be provided by a continuous random variable X obeying a Weibull distribution with parameters $\lambda > 0, \eta > 0$ ($X \sim \mathcal{W}(\lambda,\eta)$).

The probability density function of the Weibull distribution is defined by

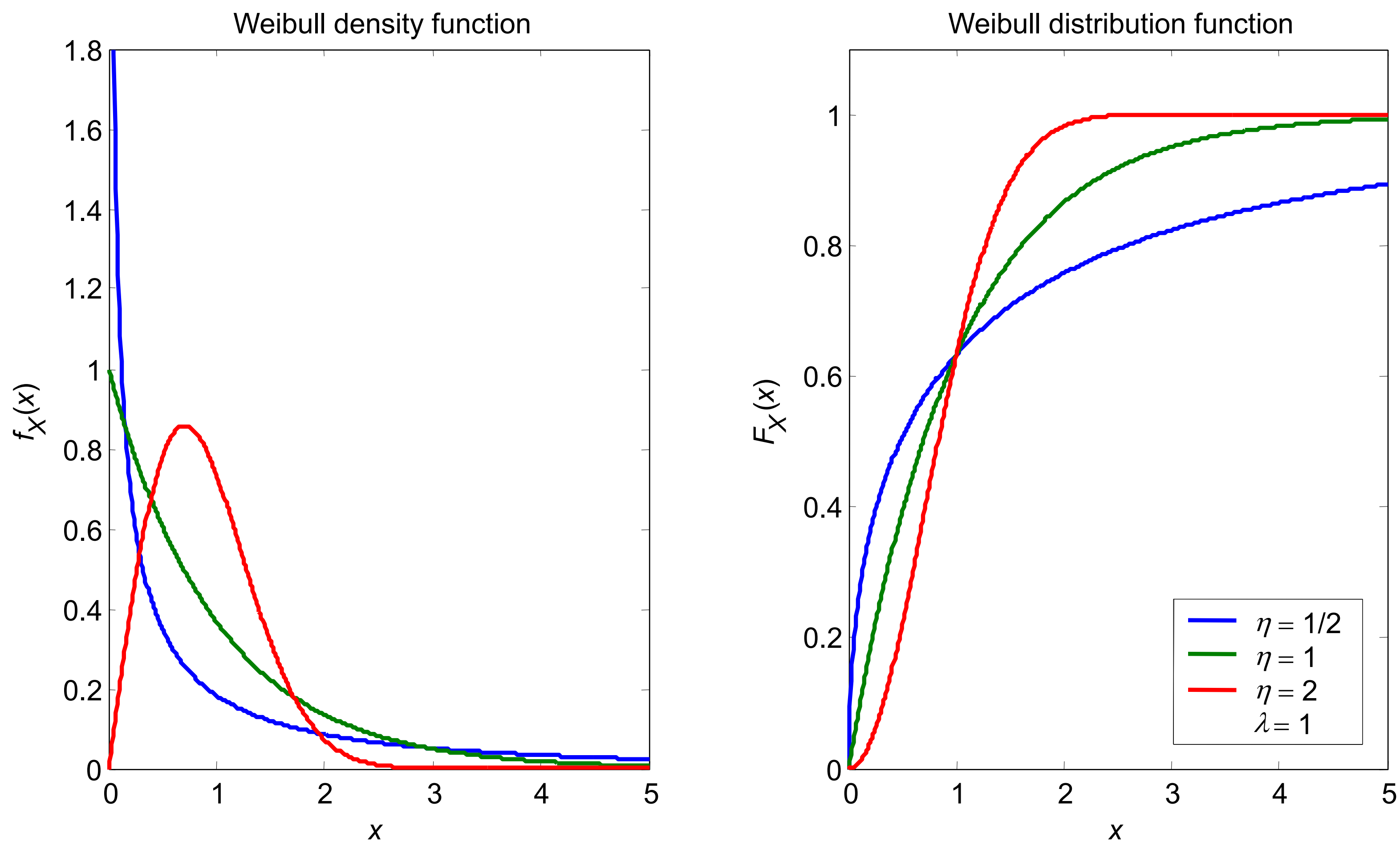
$$f_X(x) = \lambda\eta x^{\eta-1}\exp(-\lambda x^{\eta})1_{[0,\infty)}(x) \qquad x \in \mathbb{R}.$$

For the distribution function we obtain

$$F_X(x) = \int_{-\infty}^{x} f_X(x')\,dx' = \int_{-\infty}^{x} \lambda\eta x'^{\eta-1} \exp(-\lambda x'^{\eta}) 1_{[0,\infty)}(x')\,dx'$$
$$= \left(1 - \exp(-\lambda x^{\eta})\right) 1_{[0,\infty)}(x).$$

By selecting a value for η, the following three cases can be qualitatively distinguished:

$\eta = 1$	obviously $\mathcal{W}(\lambda, 1) = \mathcal{E}(\lambda)$,
$\eta > 1$	ageing process is incorporated,
$0 < \eta < 1$	initial failure process is considered.


Weibull density function
$f_X(x)$
x
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
0
1
2
3
4
5
Weibull distribution function
$F_X(x)$
x
0
0.2
0.4
0.6
0.8
1
$\eta = 1/2$
$\eta = 1$
$\eta = 2$
$\lambda = 1$

Cauchy distribution

A continuous random variable X is said to follow a Cauchy distribution with parameters $\mu \in \mathbb{R}$ and $\nu > 0$ ($X \sim \mathcal{C}(\mu, \nu)$), if the probability density function is given by

$$f_X(x) = \frac{1}{\pi} \frac{\nu}{\nu^2 + (x-\mu)^2} = \frac{1}{\pi \nu} \frac{1}{1 + \left[(x-\mu)/\nu\right]^2} \qquad x \in \mathbb{R}.$$

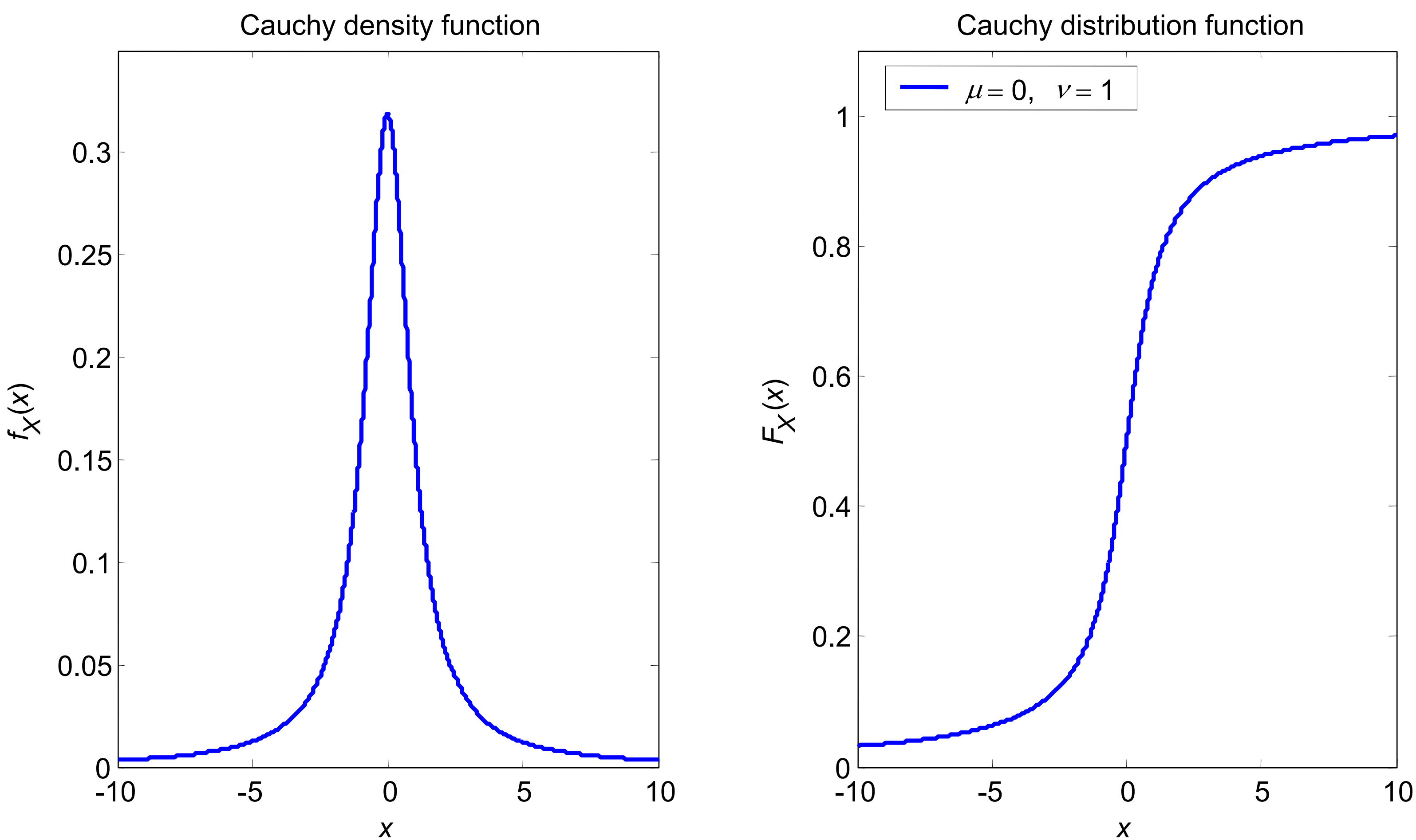
The distribution function can be derived as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(x')dx' = \frac{1}{\pi \nu} \int_{-\infty}^{x} \frac{1}{1 + \left[(x'-\mu)/\nu\right]^2} dx' = \cdots$$

$$\cdots = \frac{1}{\pi}\int_{-\infty}^{(x-\mu)/\nu} \frac{1}{1+x''^2}dx'' = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\nu}\right).$$

The Cauchy distribution possess so-called long tails, i.e. its density function decays slowly as x tends either to plus or minus infinity.

Consequently, Cauchy distributed random variables are suitable models for experiments where large measurement values can be observed with certain likelihood.


Cauchy density function
$f_X(x)$
0.3
0.25
0.2
0.15
0.1
0.05
0
-10
-5
0
5
10
x
Cauchy distribution function
$\mu = 0, \quad \nu = 1$
$F_X(x)$
1
0.8
0.6
0.4
0.2
0
-10
-5
0
5
10
x

1.7 Bivariate Distribution

The theory of random variables discussed so far dealt only with univariate distributions, i.e. the probability distributions describe the properties of single random variables.

However, the modeling of experimental results often requires several random variables, e.g. the results of measuring the simultaneous values of pressure and temperature in a gas of constant volume have to be described by two random variables.

1.7.1 Bivariate Distribution Function

A bivariate distribution function $F_{XY}(x,y)$ is defined by

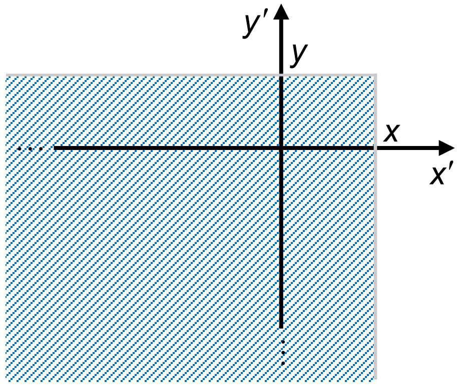
$$F_{XY}(x,y) = P\left(\left\{\xi : X(\xi) \le x, Y(\xi) \le y\right\}\right) = P(X \le x, Y \le y).$$

It can be computed in the discrete and continuous case by

$$F_{XY}(x,y) = \sum_{i, x_i \le x} \sum_{j, y_j \le y} p_{XY}(x_i, y_j)$$

and

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x',y')\,dx'dy',$$



respectively, where $p_{XY}(x,y)$ denotes the bivariate probability mass function and $f_{XY}(x,y)$ the bivariate density function.

Bivariate distribution functions possess the properties:

(1) $\lim_{x\to\infty,\, y\to\infty} F_{XY}(x,y) = F_{XY}(\infty,\infty) = 1.$

(2) $\lim_{x\to-\infty} F_{XY}(x,y) = F_{XY}(-\infty, y) = 0,$

$\lim_{y\to-\infty} F_{XY}(x,y) = F_{XY}(x,-\infty) = 0.$

(3) $F_{XY}(x,y)$ is right-continuous in x and y, respectively,

i.e. $\lim_{h\to 0+} F_{XY}(x+h,y) = F_{XY}(x,y),$

$\lim_{h\to 0+} F_{XY}(x,y+h) = F_{XY}(x,y).$

(4) F_{XY}(x,y) is a non-decreasing function, i.e. for any

$h \geq 0$ is $F_{XY}(x+h,y) \geq F_{XY}(x,y)$ for all x and any y,

$F_{XY}(x,y+h) \geq F_{XY}(x,y)$ for all y and any x.

(5) second difference (n-th difference for $n = 2$)

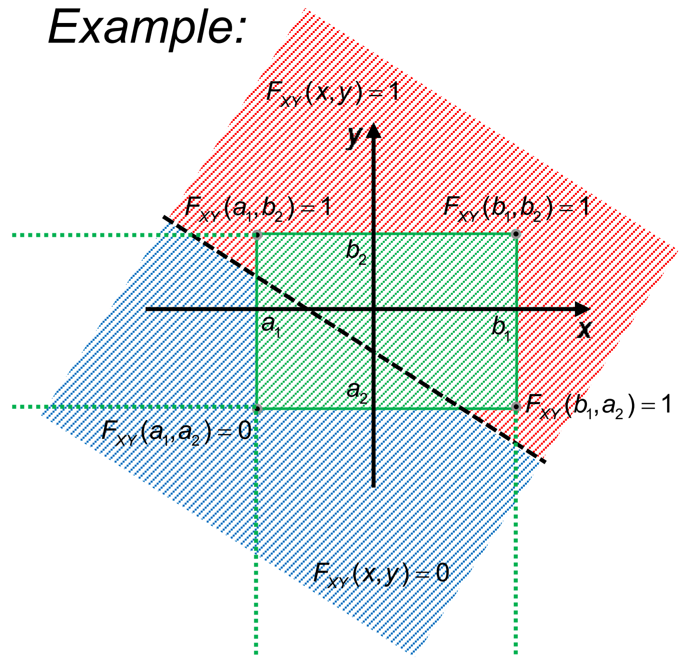
$$\Delta F_{XY}((\mathbf{a},\mathbf{b}]) =$$

$$= F_{XY}(b_1,b_2) - F_{XY}(b_1,a_2) - F_{XY}(a_1,b_2) + F_{XY}(a_1,a_2)$$

$$= P(\{\xi : (X(\xi),Y(\xi))^T \in (\mathbf{a},\mathbf{b}]\})$$

$$= P((X,Y)^T \in (\mathbf{a},\mathbf{b}]) \geq 0,$$

Example:



where

$$(\mathbf{a},\mathbf{b}] = ((a_1,a_2)^T,(b_1,b_2)^T]$$

$$= (a_1,b_1] \times (a_2,b_2].$$

1.7.2 Bivariate Density Function

If $f_{XY}(x,y)$ is continuous at (x,y), then

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x\,\partial y} = \frac{\partial^2 F_{XY}(x,y)}{\partial y\,\partial x}$$

exists. Bivariate density functions satisfy the properties:

(1) $f_{XY}(x,y) \geq 0$ for all $(x,y)^T \in \mathbb{R}^2$.

(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = 1.$

(3) For any $(\mathbf{a},\mathbf{b}] = (a_1,b_1] \times (a_2,b_2] \subset \mathbb{R}^2$

$$P\big((X,Y)^T \in (\mathbf{a},\mathbf{b}]\big) = \int_{a_2}^{b_2}\int_{a_1}^{b_1} f_{XY}(x,y)\,dx\,dy.$$

1.7.3 Marginal Distribution and Density Function

For a given bivariate distribution of (X,Y), the univariate distributions of the individual random variables X and Y can be deduced from the expressions

$$\lim_{y\to\infty} F_{XY}(x,y) = F_{XY}(x,\infty) = P(X \le x, Y \le \infty)$$
$$= P(X \le x) = F_X(x)$$

and

$$\lim_{x\to\infty} F_{XY}(x,y) = F_{XY}(\infty,y) = P(X \le \infty, Y \le y)$$
$$= P(Y \le y) = F_Y(y).$$

For the marginal distribution and density functions of X and Y we may write

$$F_X(x) = \int_{-\infty}^{x} f_X(x')dx' = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{XY}(x',y)dydx'$$

with

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

and

$$F_Y(y) = \int_{-\infty}^{y} f_Y(y')dy' = \int_{-\infty}^{y}\int_{-\infty}^{\infty} f_{XY}(x,y')dxdy'$$

with

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx,$$

respectively.

1.7.4 Conditional Distribution and Density Function

The conditional probability of $\{\xi : X(\xi) \le x\}$ knowing that $\{\xi : Y(\xi) \le y\}$ occurred is given by, cf. Chapter 1.3,

$$P(X \le x \,|\, Y \le y) = \frac{P\big(\{X \le x\} \cap \{Y \le y\}\big)}{P(Y \le y)}, \text{ if } P(Y \le y) > 0.$$

Hence, the conditional distribution function of X under the condition $\{Y \le y\}$ can be defined by

$$\begin{aligned} F_X(x \,|\, Y \le y) &= P(X \le x \,|\, Y \le y) \\ &= P(X \le x, Y \le y) / P(Y \le y) \\ &= F_{XY}(x,y)/F_Y(y), \quad \text{if } F_Y(y) > 0. \end{aligned}$$

The conditional density function of X under the condition $\{Y \leq y\}$ can be deduced from

$$F_X(x \mid Y \leq y) = \int_{-\infty}^{x} f_X(x' \mid Y \leq y)\,dx' = F_{XY}(x,y)/F_Y(y)$$

$$= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(x',y')\,dy'dx' \Big/ F_Y(y)$$

such that

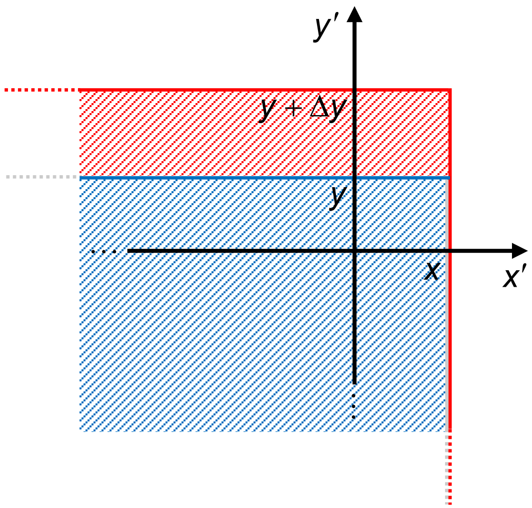
$$f_X(x \mid Y \leq y) = \frac{\int_{-\infty}^{y} f_{XY}(x,y')\,dy'}{F_Y(y)}.$$

If $f_{XY}(x,y)$ is continuous at (x,y), we can also write

$$f_X(x \mid Y \leq y) = \frac{dF_X(x \mid Y \leq y)}{dx} = \frac{\partial F_{XY}(x,y)/\partial x}{F_Y(y)}.$$

Let now $\{\xi : y < Y(\xi) \leq y + \Delta y\}$ denote an event satisfying $P(y < Y \leq y + \Delta y) > 0$. Then, we can derive

$$F_X(x \mid y < Y \leq y + \Delta y) =$$



$$= \frac{P(X \leq x, y < Y \leq y + \Delta y)}{P(y < Y \leq y + \Delta y)}$$

$$= \frac{P(X \leq x, Y \leq y + \Delta y) - P(X \leq x, Y \leq y)}{P(Y \leq y + \Delta y) - P(Y \leq y)}$$

$$= \frac{F_{XY}(x, y + \Delta y) - F_{XY}(x, y)}{F_Y(y + \Delta y) - F_Y(y)} \cdot \frac{1/\Delta y}{1/\Delta y}.$$

Assuming $f_{XY}(x,y)$ to be continuous at (x,y) and $f_Y(y) > 0$, the limit

$$\lim_{\Delta y \to 0} F_X(x \mid y < Y \le y + \Delta y) = \frac{\partial F_{XY}(x,y)/\partial y}{dF_Y(y)/dy}$$

exists and the conditional distribution function of X under the condition $\{Y = y\}$ can be expressed by

$$F_X(x \mid Y = y) = \frac{\partial F_{XY}(x,y)/\partial y}{dF_Y(y)/dy} = \frac{\partial F_{XY}(x,y)/\partial y}{f_Y(y)}$$

$$= \frac{\int_{-\infty}^{x} f_{XY}(x',y)\,dx'}{f_Y(y)} = \frac{\int_{-\infty}^{x} f_{XY}(x',y)\,dx'}{\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx}.$$

Furthermore, exploiting

$$F_X(x \mid Y=y)=\int_{-\infty}^{x} f_X(x' \mid Y=y)\,dx'=\int_{-\infty}^{x} \frac{f_{XY}(x',y)}{f_Y(y)}\,dx'$$

the conditional density function of X under the condition $\{Y=y\}$ can be represented by

$$f_X(x \mid Y=y)=\frac{f_{XY}(x,y)}{f_Y(y)}=\frac{f_{XY}(x,y)}{\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx}$$

$$=\frac{\partial^2 F_{XY}(x,y)/\partial x \partial y}{dF_Y(y)/dy}=\frac{dF_X(x \mid Y=y)}{dx}.$$

Conditional density functions exhibit the properties:

(1) $$f_X(x \mid Y = y) \cdot \Delta x = \frac{f_{XY}(x,y) \cdot \Delta x \cdot \Delta y}{f_Y(y) \cdot \Delta y}$$

$$\approx \frac{P(x < X \le x + \Delta x, y < Y \le y + \Delta y)}{P(y < Y \le y + \Delta y)}.$$

(2) $$f_{XY}(x,y) = f_X(x \mid Y = y) f_Y(y) = f_Y(y \mid X = x) f_X(x).$$

Bayes' Formula:

$$f_X(x \mid Y = y) = f_Y(y \mid X = x) f_X(x) / f_Y(y), \text{ if } f_Y(y) > 0.$$

(3) $$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy = \int_{-\infty}^{\infty} f_X(x \mid Y = y) f_Y(y)\,dy.$$

1.7.5 Independent Random Variables

We say that two continuous random variables, X and Y are independent if the events $\{X \le x\}$ and $\{Y \le y\}$ are independent for all $x,y \in \mathbb{R}$, i.e. cf. Chapter 1.3 that

$$F_{XY}(x,y) = F_X(x)F_Y(y), \qquad f_{XY}(x,y) = f_X(x)f_Y(y)$$

and consequently

$$F_X(x|Y=y) = F_X(x), \qquad F_Y(y|X=x) = F_Y(y),$$

$$f_X(x|Y=y) = f_X(x) \quad \text{and} \quad f_Y(y|X=x) = f_Y(y).$$

For notational convenience we define

$$f_X(x|y) := f_X(x|Y=y), \qquad f_Y(y|x) := f_Y(y|X=x),$$

$$F_X(x|y) := F_X(x|Y=y) \quad \text{and} \quad F_Y(y|x) := F_Y(y|X=x).$$

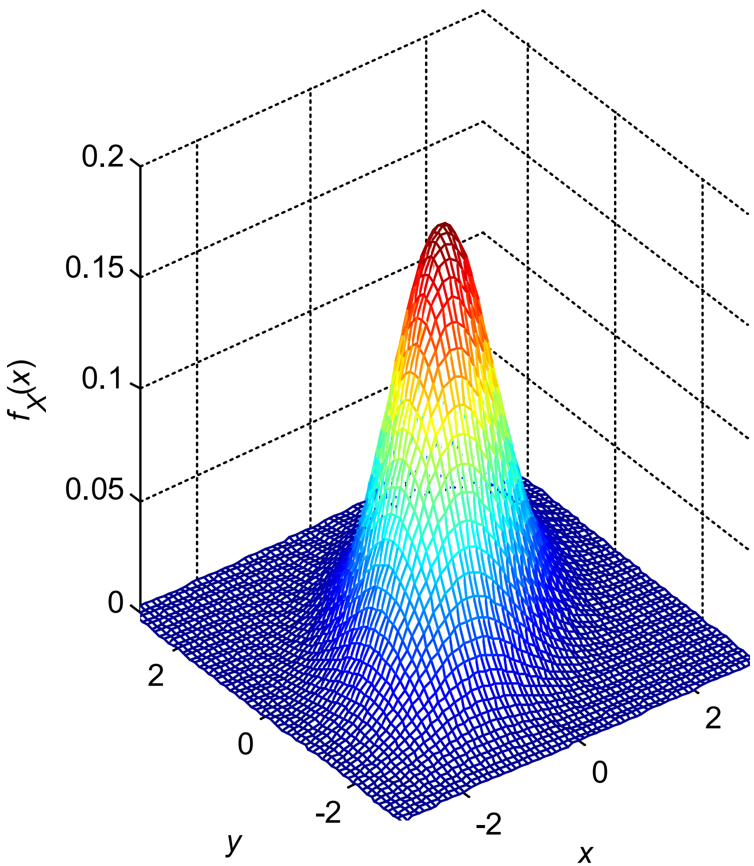
1.7.6 Bivariate Normal Distribution

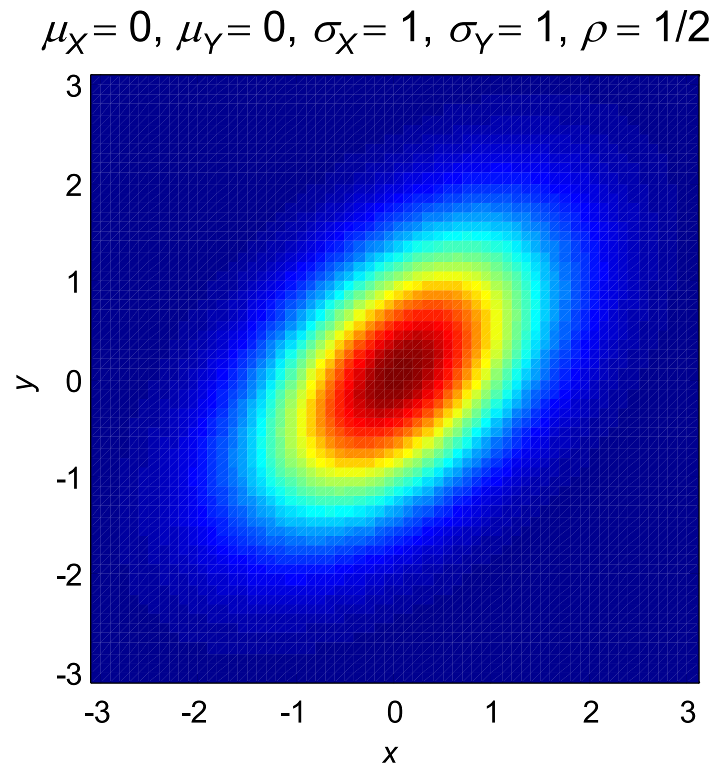
Two continuous random variables X and Y are said to obey a bivariate normal distribution with parameters μ_X, μ_Y, $|\rho| < 1$, $\sigma_X > 0$, $\sigma_Y > 0$ if their probability density function is given by

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \times \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

for all $x, y \in \mathbb{R}$.

Bivariate Normal Density Function





Employing vector/matrix notation the bivariate normal density function can be expressed by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det(\mathbf{\Sigma}_{\mathbf{X}})}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T \mathbf{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})\right\},$$

where $\boldsymbol{\mu}_{\mathbf{X}}$ denotes the vector of the expected/mean values and $\mathbf{\Sigma}_{\mathbf{X}}$ represents the so-called covariance matrix, i.e.

$$\boldsymbol{\mu}_{\mathbf{X}} = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \end{pmatrix}, \quad \mathbf{\Sigma}_{\mathbf{X}} = \begin{pmatrix} \sigma_{X_1}^2 & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{pmatrix},$$

and one writes

$$\mathbf{X} = (X_1, X_2)^T \sim \mathcal{N}_2(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{\Sigma}_{\mathbf{X}}).$$

Marginal density functions

The marginal density functions of *X* and *Y* are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y')\,dy' = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\}$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x', y)\,dx' = \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_Y}\right)^2\right\}.$$

Hence, the marginal probability distributions of *X* and *Y* are $\mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathcal{N}(\mu_Y, \sigma_Y^2)$, respectively.

Conditional density functions

The conditional density function of X under the condition $\{Y = y\}$ is given by

$$f_X(x \mid Y = y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{1}{\sqrt{2\pi\sigma_X^2(1-\rho^2)}} \times$$

$$\exp\left\{ -\frac{1}{2\sigma_X^2(1-\rho^2)} \left[x - \left(\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y) \right) \right]^2 \right\}$$

and we can write in abbreviated form

$$X \mid Y = y \sim \mathcal{N}\left(\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y), \sigma_X^2(1-\rho^2) \right).$$

Analogous, the conditional density function of Y under the condition $\{X = x\}$ is given by

$$f_Y(y \mid X = x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}} \times$$

$$\exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y - \left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right)\right]^2\right\}.$$

Thus, we can write

$$Y \mid X = x \sim \mathcal{N}\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2)\right).$$

1.8 Transformations of Random Variables

1.8.1 Function of One Random Variable

Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function, i.e. $\forall y \in \mathbb{R}$ is

$$g^{-1}\big((-\infty, y]\big) = \{x : g(x) \le y\} \in \mathbb{B}.$$

Then, we can define a random variable $Y : \Xi \to \mathbb{R}$ by

$$\xi \mapsto Y(\xi) = g\big(X(\xi)\big)$$

possessing a distribution function determined by

$$F_Y(y) = P(Y \le y) = P\big(g(X) \le y\big)$$

$$= P\big(\{\xi : g\big(X(\xi)\big) \le y\}\big) = P\Big(X^{-1}\big(g^{-1}((-\infty, y])\big)\Big).$$

Strictly Monotonic Function

a) Suppose that $g(x)$ is a strictly monotonic increasing function. The distribution function can be written as

$$F_Y(y) = P(Y \le y) = P\left(g(X) \le y\right)$$
$$= P\left(X \le g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right)$$
$$= \int_{-\infty}^{g^{-1}(y)} f_X(x)\,dx.$$

Moreover, if $f_X(x)$ is continuous and $g(x)$ continuously differentiable at $x = g^{-1}(y)$ we can derive the density function $f_Y(y)$ by applying the chain role of calculus.

Hence, we obtain

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\dfrac{dg^{-1}(y)}{dy} & \text{for } a < y < b \\ 0 & \text{elsewhere} \end{cases}$$

$$= \begin{cases} \dfrac{f_X\left(g^{-1}(y)\right)}{dg(x)/dx\big|_{x=g^{-1}(y)}} & \text{for } a < y < b \\ 0 & \text{elsewhere} \end{cases}$$

with

$$a = g(-\infty) \quad \text{and} \quad b = g(\infty).$$

b) Let $g(x)$ be a strictly monotonic decreasing function. Consequently, we have

$$F_Y(y) = P(Y \le y) = P\left(g(X) \le y\right) = P\left(X \ge g^{-1}(y)\right)$$

$$= 1 - P\left(X < g^{-1}(y)\right) = 1 - F_X\left(g^{-1}(y) - 0\right)$$

$$= 1 - \int_{-\infty}^{g^{-1}(y)} f_X(x)\,dx = \int_{g^{-1}(y)}^{\infty} f_X(x)\,dx.$$

Again, if $f_X(x)$ is continuous and $g(x)$ continuously differentiable at $x = g^{-1}(y)$ the density function $f_Y(y)$ can be determined by employing the chain role of calculus.

Thus, we obtain

$$f_Y(y) = \begin{cases} -f_X\left(g^{-1}(y)\right)\dfrac{dg^{-1}(y)}{dy} & \text{for } a < y < b \\ 0 & \text{elsewhere} \end{cases}$$

$$= \begin{cases} -\dfrac{f_X\left(g^{-1}(y)\right)}{dg(x)/dx\big|_{x=g^{-1}(y)}} & \text{for } a < y < b \\ 0 & \text{elsewhere} \end{cases}$$

with

$$a = g(\infty) \quad \text{and} \quad b = g(-\infty).$$

Exploiting the property

$$dg(x)/dx > 0 \quad \text{for } g(x) \text{ strictly monotonic increasing}$$
$$< 0 \quad \text{for } g(x) \text{ strictly monotonic decreasing}$$

the results of case a) and b) can be summarised.

Let $f_X(x)$ be continuous at $x = g^{-1}(y)$ and $g(x)$ any strictly monotonic function that is continuously differentiable at $x = g^{-1}(y)$. Then $f_Y(y)$ can be calculated by

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right| = \frac{f_X\left(g^{-1}(y)\right)}{\left|dg(x)/dx\big|_{x=g^{-1}(y)}\right|} \quad \text{for } a < y < b$$

with

$$a = \min\{g(-\infty), g(\infty)\} \quad \text{and} \quad b = \max\{g(-\infty), g(\infty)\}.$$

Exercise 1.8-1:
Linear function

Non-Monotonic Function

Theorem:

Let $f_X(x)$ denote the continuous density function of the random variable X and let $g(\mathrm{x})$ be a continuously differentiable function. Furthermore, suppose that equation $y = g(\mathrm{x})$ may possess n solutions for a particular y, i.e.

$$y = g(x_1) = \ldots = g(x_n).$$

Then, $f_Y(y)$, the continuous density function of the random variable $Y = g(X)$ can be determined by

$$f_Y(y) = \sum_{i=1}^{n} \frac{f_X(x_i)}{|dg(x_i)/dx|}\Bigg|_{x_i = g_i^{-1}(y)}.$$

Exercise 1.8-2:
Quadratic function

1.8.2 One Function of Two Random Variables

Suppose (X,Y) are random variables with bivariate density function $f_{XY}(x,y)$. Let $g(x,y)$ be a function such that

$$Z = g(X,Y)$$

represents a random variable, i.e. for all $z \in \mathbb{R}$ is

$$D_z = \{(x,y) : g(x,y) \le z\} \in \mathbb{B}.$$

Then the distribution function of Z is given by

$$F_Z(z) = P(Z \le z) = P\big(g(X,Y) \le z\big)$$
$$= P\big((X,Y) \in D_z\big) = \iint_{D_z} f_{XY}(x,y)\,dx\,dy.$$

Exercise 1.8-3:
Sum of two random variables

Exercise 1.8-4:
Magnitude of the difference of two independent random variables

1.8.3 Two Functions of Two Random Variables

Let (X,Y) be random variables with bivariate density function $f_{XY}(x,y)$. Suppose $g_1(x,y)$ and $g_2(x,y)$ are functions such that

$$U = g_1(X,Y) \quad \text{and} \quad V = g_2(X,Y)$$

are random variables, i.e. for all $u,v \in \mathbb{R}$ are

$$D_u = \{(x,y) : g_1(x,y) \le u\} \in \mathbb{B},$$

$$D_v = \{(x,y) : g_2(x,y) \le v\} \in \mathbb{B}$$

and consequently

$$D_{uv} = D_u \cap D_v = \{(x,y) : g_1(x,y) \le u,\, g_2(x,y) \le v\} \in \mathbb{B}.$$

The bivariate distribution function of (U,V) is given by

$$\begin{aligned} F_{UV}(u,v) &= P(U \le u, V \le v) \\ &= P\left(g_1(X,Y) \le u, g_2(X,Y) \le v\right) \\ &= P\left((X,Y) \in D_{uv}\right) = \iint_{D_{uv}} f_{XY}(x,y)\,dx\,dy. \end{aligned}$$

Assume that the equation system

$$(u,v) = \left(g_1(x,y), g_2(x,y)\right)$$

has the unique solution

$$(x,y) = \left(g_1^{-1}(u,v), g_2^{-1}(u,v)\right).$$

Furthermore, suppose $g_1(x,y)$, $g_2(x,y)$ have continuous partial derivatives and the determinant of the Jacobian

$$\mathbf{J}(x,y) = \frac{\partial(g_1,g_2)}{\partial(x,y)} = \begin{pmatrix} \partial g_1/\partial x & \partial g_1/\partial y \\ \partial g_2/\partial x & \partial g_2/\partial y \end{pmatrix}$$

does not vanish, i.e.

$$\det\left(\mathbf{J}(x,y)\right) \neq 0.$$

Then the bivariate density function of (U,V) is given by

$$f_{UV}(u,v) = \frac{f_{XY}\left(g_1^{-1}(u,v), g_2^{-1}(u,v)\right)}{\left|\det\left(\mathbf{J}\left(g_1^{-1}(u,v), g_2^{-1}(u,v)\right)\right)\right|}.$$

Alternatively, using the Jacobian

$$\tilde{\mathbf{J}}(u,v) = \frac{\partial(g_1^{-1}, g_2^{-1})}{\partial(u,v)} = \begin{pmatrix} \partial g_1^{-1}/\partial u & \partial g_1^{-1}/\partial v \\ \partial g_2^{-1}/\partial u & \partial g_2^{-1}/\partial v \end{pmatrix}$$

and exploiting the well known results

$$\tilde{\mathbf{J}}(u,v) = \mathbf{J}\left(g_1^{-1}(u,v), g_2^{-1}(u,v)\right)^{-1}$$

and

$$\det\left(\tilde{\mathbf{J}}(u,v)\right) = 1 \Big/ \det\left(\mathbf{J}\left(g_1^{-1}(u,v), g_2^{-1}(u,v)\right)\right)$$

we can write

$$f_{UV}(u,v) = f_{XY}\left(g_1^{-1}(u,v), g_2^{-1}(u,v)\right) \left|\det\left(\tilde{\mathbf{J}}(u,v)\right)\right|.$$

<u>Exercise 1.8-5:</u>
Linear transformation

<u>Exercise 1.8-6:</u>
Product of two random variables

<u>Exercise 1.8-7:</u>
Quotient of two random variables

<u>Exercise 1.8-8:</u>
Rayleigh distribution

1.9 Expectation Operator

1.9.1 Expectation for Univariate Distributions

Expected Value of a Random Variable

The expected value of a random variable *X*, also called mean value, is defined by

$$\mu_X = \mathrm{E}(X) = \begin{cases} \sum_i x_i \, p_X(x_i) & \text{when } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x \, f_X(x) \, dx & \text{when } X \text{ is continuous} \end{cases}$$

provided that the sum respectively integral converges absolutely.

The two cases can be summarized by introducing the so-called Stieltjes-Integral

$$\mu_X = \mathrm{E}(X) = \int_{-\infty}^{\infty} x \, dF_X(x).$$

Remark:

Let $a = x_0 < x_1 < \ldots < x_n = b$ be points that provide a partition of $[a, b]$ into n subintervals $(x_k, x_{k+1}]$ $(k = 0, \ldots, n-1)$ and $\tilde{x}_k \in (x_k, x_{k+1}]$.

Then the Stieltjes-Integral is defined by

$$\int_a^b g(x) dF(x) = \lim_{\substack{n \to \infty \\ \max_k (x_{k+1} - x_k) \to 0}} \sum_{k=0}^{n-1} g(\tilde{x}_k) \big(F(x_{k+1}) - F(x_k) \big).$$

Exercise 1.9-1:
Expected value for Poisson distribution

Exercise 1.9-2:
Expected value for exponential distribution

Exercise 1.9-3:
Expected value for normal distribution

Exercise 1.9-4:
Expected value for Cauchy distribution

Expected Value of a Function of a Random Variable

For a measurable function, $g(X)$, of the random variable X, we define the expected value of $g(X)$ by

$$\mathrm{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x)\,dF_X(x)$$

$$= \begin{cases} \sum_i g(x_i)\,p_X(x_i) & \text{when } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)\,f_X(x)\,dx & \text{when } X \text{ is continuous} \end{cases}$$

provided that the sum respectively integral converges absolutely.

Let $F_Y(y)$ denote the distribution function of the random variable $Y = g(X)$, then by using the definition of integration it is easy to establish that

$$\mathrm{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x)\, dF_X(x) = \int_{-\infty}^{\infty} y\, dF_Y(y) = \mathrm{E}(Y).$$

Consequently, the expected value of a function of a random variable $g(X)$ can be computed directly without determining first the distribution function of the random variable $Y = g(X)$.

Moments

If $g(X) = X^k$ with $k > 0$, the expected value of $g(X)$, i.e.

$$m_k = \mathrm{E}(X^k) = \int_{-\infty}^{\infty} x^k \, dF_X(x)$$

$$= \begin{cases} \sum_i x_i^k p_X(x_i) & \text{when } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f_X(x)\,dx & \text{when } X \text{ is continuous} \end{cases},$$

is called k-th moment of X provided that the integral converges absolutely. For $k = 1$, we obtain the mean value $\mu_X = m_1 = \mathrm{E}(X)$.

Centralised Moments

Suppose $g(X) = (X - \mu_X)^k$ with $k > 0$, then the expected value of $g(X)$, i.e.

$$c_k = \mathrm{E}\left((X - \mu_X)^k\right) = \mathrm{E}\left(\sum_{m=0}^{k} \binom{k}{m} (-\mu_X)^m X^{k-m}\right)$$

$$= \sum_{m=0}^{k} \binom{k}{m} (-\mu_X)^m \mathrm{E}(X^{k-m})$$

is called k-th centralised moment of X.

For $k = 2$, the centralised moment given by

$$\sigma_X^2 = \mathrm{Var}(X) = c_2 = \mathrm{E}\left((X - \mu_X)^2\right) = \mathrm{E}(X^2) - \mu_X^2$$

is called variance.

The positive root of the variance is denoted by σ_X and is called standard deviation.

Absolute Moments

The k-th absolute moment of X is defined by

$$\mathrm{E}\left(|X|^k\right) = \int_{-\infty}^{\infty} |X|^k \, dF_X(x).$$

Because of the inequality

$$|X|^{k-1} \leq 1 + |X|^k \quad \text{for} \quad k = 1, 2, \ldots,$$

we can state that the existence of the *k*-th absolute moment insures the existence of the (*k*–1)-th moment.

Chebyschev Inequality

Let X be a random variable. For $k \geq 1$ and any $\varepsilon > 0$ the inequality

$$P\left(|X| \geq \varepsilon\right) \leq \frac{\mathrm{E}\left(|X|^k\right)}{\varepsilon^k}$$

holds.

Exercise 1.9-5:
Second order moments for Poisson distribution

Exercise 1.9-6:
Second order moments for exponential distribution

Exercise 1.9-7:
Second order moments for normal distribution

Exercise 1.9-8:
Application and proof of the Chebyschev inequality

Characteristic Function

The characteristic function of the random variable X is defined by taking the expected value of $g(X) = e^{jsX}$, i.e.

$$\Phi_X(s) = \mathrm{E}(e^{jsX}) = \int_{-\infty}^{\infty} e^{jsx} dF_X(x)$$

$$= \begin{cases} \sum_i e^{jsx_i} p_X(x_i) & \text{when } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{jsx} f_X(x) dx & \text{when } X \text{ is continuous} \end{cases},$$

where $s \in \mathbb{R}$.

Characteristic functions have the properties:

- $\Phi_X(s)$ is continuous in s.
  (absolute and uniform convergence of the sum resp. integral)
- $|\Phi_X(s)| \leq 1$ for all $s \in \mathbb{R}$.
- $\Phi_X(0) = \mathrm{E}(e^0) = \mathrm{E}(1) = 1.$
  (characteristic function takes its maximum at $s = 0$)

Note:

$$\left|\Phi_X(s)\right| = \left|\sum_i e^{jsx_i} p_X(x_i)\right| \leq \sum_i \left|e^{jsx_i}\right| p_X(x_i) = \sum_i p_X(x_i) = 1$$

$$\left|\Phi_X(s)\right| = \left|\int_{-\infty}^{\infty} e^{jsx} f_X(x)\,dx\right| \leq \int_{-\infty}^{\infty} \left|e^{jsx}\right| f_X(x)\,dx = \int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

Moreover, one can easily observe that $\Phi_X(-\omega)$ equals the Fourier transform of $f_X(x)$.

Hence, the properties of a characteristic function are essentially equivalent to the properties of a Fourier transform (one-to-one mapping).

Consequently, the probability distribution of a random variable is uniquely defined by the inverse Fourier transform of the characteristic function $\Phi_X(s)$, i.e.

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-jsx}\Phi_X(s)ds.$$

Let X be a random variable with characteristic function $\Phi_X(s)$ and $Y = aX + b$. Thus, the characteristic function of Y can be easily determined by

$$\Phi_Y(s) = \mathrm{E}(e^{jsY}) = \mathrm{E}(e^{js(aX+b)}) = e^{jbs}\,\mathrm{E}(e^{jasX}) = e^{jbs}\,\Phi_X(as).$$

Its probability density function can be derived by applying the inverse Fourier Transform as follows

$$f_Y(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-jsy}\,\Phi_Y(s)\,ds = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-js(y-b)}\,\Phi_X(as)\,ds$$

$$= \frac{1}{|a|}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-js'(y-b)/a}\,\Phi_X(s')\,ds' = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Moment Theorem

Suppose that $E(X^k)$ exists for any $k \geq 1$, i.e. $E(|X|^k) < \infty$, and therefore

$$\frac{d^k \Phi_X(s)}{ds^k} = \frac{d^k E(e^{jsX})}{ds^k} = E\left(\frac{\partial^k e^{jsX}}{\partial s^k}\right) = j^k E(X^k e^{jsX})$$

holds, i.e. the order of differentiation and integration can be interchanged, we can deduce the so-called moment theorem

$$m_k = E(X^k) = \frac{1}{j^k} \left.\frac{d^k \Phi_X(s)}{ds^k}\right|_{s=0} .$$

Exercise 1.9-9:
Characteristic function of univariate normal distributions

Exercise 1.9-10:
Higher order moments of univariate normal distributions

Exercise 1.9-11:
Non-negative definiteness of the characteristic function

1.9.2 Expectation for Bivariate Distributions

Expected Value of a Function of two Random Variables

For a measurable function $g(X,Y)$ of the random variables X and Y, the expected value of $g(X,Y)$ is defined by

$$\mathrm{E}\big(g(X,Y)\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\,d^2F_{XY}(x,y) =$$

$$= \begin{cases} \sum\limits_i \sum\limits_j g(x_i,y_j)\,p_{XY}(x_i,y_j) & \text{in the discrete case} \\ \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} g(x,y)\,f_{XY}(x,y)\,dx\,dy & \text{in the continuous case} \end{cases}$$

provided that the sum respectively integral converges ab-solutely.

$F_Z(z)$ may denote the distribution function of the random variable $Z = g(X,Y)$. Then analogue to the univariate case, we can show that

$$\mathrm{E}\big(g(X,Y)\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\,d^2F_{XY}(x,y)$$
$$= \int_{-\infty}^{\infty} z\,dF_Z(z) = \mathrm{E}(Z).$$

That is, the expected value of a function of two random variables $g(X,Y)$ can be computed directly without determining first the distribution function of the random variable $Z = g(X,Y)$.

Expected Value of a Linear Combination

The expected value of a linear combination leads to

$$\mathrm{E}\left(\sum_i a_i\, g_i(X,Y)\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_i a_i\, g_i(x,y)\, d^2F_{XY}(x,y) =$$

$$= \sum_i \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a_i\, g_i(x,y)\, d^2F_{XY}(x,y) = \sum_i a_i\, \mathrm{E}\big(g_i(X,Y)\big).$$

Thus, the expected value of a linear combination equals the linear combination of the expected values.

We note in particular that

$$\mathrm{E}(a X^k + b Y^k) = a\mathrm{E}(X^k) + b\mathrm{E}(Y^k) \quad \text{for } k \geq 1.$$

Bivariate Moments

The (k,l)-th moment and centralised moment of discrete distributed random variables are defined by

$$m_{kl} = \mathrm{E}(X^k Y^l) = \sum_i \sum_j x_i^k y_j^l \, p_{XY}(x_i, y_j) \quad k = 1,2,\ldots; \, l = 1,2,\ldots$$

and

$$c_{kl} = \mathrm{E}\Big((X - \mathrm{E}(X))^k \, (Y - \mathrm{E}(Y))^l \Big) = \mathrm{E}\Big((X - \mu_X)^k (Y - \mu_Y)^l \Big)$$

$$= \sum_i \sum_j (x_i - \mu_X)^k (y_j - \mu_Y)^l \, p_{XY}(x_i, y_j)$$

$$k = 1,2,\ldots; \, l = 1,2,\ldots,$$

respectively.

In case of continuously distributed random variables, the (k,l)-th moment and centralised moment are defined by

$$m_{kl} = \mathrm{E}(X^k Y^l) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^l f_{XY}(x,y)\,dx\,dy$$

$$k = 1,2,\ldots;\ l = 1,2,\ldots$$

and

$$c_{kl} = \mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)^k \big(Y - \mathrm{E}(Y)\big)^l\Big) = \mathrm{E}\big((X-\mu_X)^k (Y-\mu_Y)^l\big)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)^k (y-\mu_Y)^l f_{XY}(x,y)\,dx\,dy$$

$$k = 1,2,\ldots;\ l = 1,2,\ldots,$$

respectively.

On setting $k = 0$ or $l = 0$, the moments reduce to the corresponding moments of the marginal distributions of X and Y.

However, if $k \geq 1$ and $l \geq 1$, the moments become functions of the complete bivariate distribution.

In particular, setting $k = l = 1$, the centralised moment, called the covariance between X and Y, is given by

$$c_{11} = \mathrm{Cov}(X,Y) = \mathrm{E}\big((X - \mu_X)(Y - \mu_Y)\big)$$
$$= \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y)\, p_{XY}(x_i, y_j)$$

and

$$c_{11} = \mathrm{Cov}(X,Y) = \mathrm{E}\big((X-\mu_X)(Y-\mu_Y)\big)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f_{XY}(x,y)\,dx\,dy$$

for the discrete and continuous case, respectively.

The covariance can be expressed by first and second order moments as follows

$$c_{11} = \mathrm{Cov}(X,Y) = \mathrm{E}\big((X-\mu_X)(Y-\mu_Y)\big) = \mathrm{E}(XY) - \mu_X\mu_Y .$$

Furthermore, we have

$$c_{20} = \mathrm{Cov}(X,X) = \mathrm{E}\big((X-\mu_X)^2\big)$$

$$= \mathrm{Var}(X) = \mathrm{E}(X^2) - \mu_X^2 = \sigma_X^2 .$$

The covariance measures the degree of linear association between X, Y; i.e. the larger resp. smaller the magnitude of the covariance the larger resp. smaller is the linear association.

To achieve an unified understanding about what is large and small, we introduce the normalized quantity

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{c_{11}}{\sqrt{c_{20}\, c_{02}}},$$

which is called the correlation coefficient between X, Y.

Exercise 1.9-12:
Covariance and correlation coefficient of bivariate normal distributions

Theorem:

For all bivariate distributions with finite second order moments the Cauchy Schwarz inequality

$$\left(\mathrm{E}(XY)\right)^2 \le \mathrm{E}(X^2)\,\mathrm{E}(Y^2)$$

holds and the correlation coefficient satisfies the inequality

$$\left|\rho(X,Y)\right| \le 1,$$

where the equality sign is taken, if, with probability 1, a linear relationship between X and Y exists.

Exercise 1.9-13:
Proof of the inequalities

Uncorrelatedness, Orthogonality and Independence

Let X, Y be random variables. Then X, Y are called

(1) uncorrelated, if

$$\rho(X,Y) = 0 \Rightarrow \mathrm{Cov}(X,Y) = 0$$
$$\Rightarrow \mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y),$$

(2) orthogonal, if

$$\mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f_{XY}(x,y)\,dx\,dy = 0,$$

(3) independent, if for all $x, y \in \mathbb{R}$

$$f_{XY}(x,y) = f_X(x) \cdot f_y(y).$$

Implications:

1) If X and Y are independent random variables and $g_1(x)$ and $g_2(y)$ are measurable functions then $U = g_1(X)$ and $V = g_2(Y)$ are independent and uncorrelated random variables.

2) If X and Y are orthogonal random variables then $\mathrm{E}(X+Y)^2 = \mathrm{E}(X^2) + \mathrm{E}(Y^2)$ holds.

3) If X and Y are orthogonal random variables and $\mathrm{E}(X) = 0$ or/and $\mathrm{E}(Y) = 0$ then X and Y are uncorrelated random variables.

Remarks:

- If X and Y are uncorrelated random variables then $U = g_1(X)$ and $V = g_2(Y)$ are not necessarily uncorrelated random variables.
- If X and Y are uncorrelated random variables then X and Y are not necessarily independent random variables.
- If X and Y are uncorrelated and normally distributed random variables then X and Y are also independent random variables.

Exercise 1.9-14:
Verification of the remarks

Conditional Expected Value

Suppose that X and Y are bivariate distributed continuous random variables. The conditional expected value of X, given $Y = y$, written as $\mathrm{E}_X(X|Y = y)$ is defined by

$$\mathrm{E}_X(X \,|\, Y = y) = \int_{-\infty}^{\infty} x f_X(x \,|\, Y = y)\,dx$$

$$= \int_{-\infty}^{\infty} x \frac{f_{XY}(x,y)}{f_Y(y)}\,dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} x f_{XY}(x,y)\,dx$$

provided that the integral converges absolutely.

The corresponding expression for the discrete case is obtained with obvious modifications.

In general, the value of $E_X(X|Y=y)$ will vary as we vary the value of y.

Thus, $E_X(X|Y=y)$ will be a function of y and we can write

$$E_X(X|Y=y)=\psi_X(y),$$

where $\psi_X(y)$ is called the regression function of X on Y.

Analogously, the conditional expected value of Y, given $X=x$, is defined by

$$E_Y(Y|X=x)=\psi_Y(x),$$

where $\psi_Y(x)$ denotes the regression function of Y on X.

More generally, if we consider a measurable function of X, $g(X)$, whose expected value exists, then the conditional expected value of $g(X)$, given $Y = y$, is given by

$$\mathrm{E}_X\big(g(X)\,|\,Y=y\big)=\int_{-\infty}^{\infty} g(x) f_X(x\,|\,Y=y)\,dx$$
$$=\frac{1}{f_Y(y)}\int_{-\infty}^{\infty} g(x) f_{XY}(x,y)\,dx=\psi_{g(X)}(y).$$

Similarly, the conditional expected value of $g(Y)$, given $X = x$, is defined by

$$\mathrm{E}_Y\big(g(Y)\,|\,X=x\big)=\psi_{g(Y)}(x).$$

Conditional Expectation

Let us now consider the random variable $\psi_{g(X)}(Y)$, which we obtain by replacing y by Y in the function $\psi_{g(X)}(\cdot)$.

The random variable $\psi_{g(X)}(Y)$ is called the conditional expectation of $g(X)$, given Y, and we write

$$\psi_{g(X)}(Y) = \mathrm{E}_X\left(g(X)\,|\,Y\right).$$

For the random variable $\psi_{g(Y)}(X)$, we analogous write

$$\psi_{g(Y)}(X) = \mathrm{E}_Y\left(g(Y)\,|\,X\right).$$

Properties of conditional expectations:

(1) $\mathrm{E}_Y\big(\mathrm{E}_X(X|Y)\big)=\mathrm{E}(X),\ \mathrm{E}_X\big(\mathrm{E}_Y(Y|X)\big)=\mathrm{E}(Y)$

or more generally $\mathrm{E}_Y\big(\mathrm{E}_X(g(X)|Y)\big)=\mathrm{E}\big(g(X)\big)$,

$$\mathrm{E}_X\big(\mathrm{E}_Y(g(Y)|X)\big)=\mathrm{E}\big(g(Y)\big).$$

(2) Moreover, if $h(Y)$ is a function such that $\mathrm{E}(g(X)h(Y))$ exists, then

$$\mathrm{E}_X\big(g(X)h(Y)|Y=y\big)=h(y)\mathrm{E}_X\big(g(X)|Y=y\big)$$

(conditional on $Y=y$, $h(Y)$ can be treated as constant)

and hence we have

$$\mathrm{E}_Y\big(\mathrm{E}_X(g(X)h(Y)|Y)\big)=\mathrm{E}_Y\big(h(Y)\mathrm{E}_X(g(X)|Y)\big)$$
$$=\mathrm{E}\big(g(X)h(Y)\big).$$

Exercise 1.9-15:
Verification of the properties

Exercise 1.9-16:
Determination of $\mathrm{E}(XY)$ *for bivariate normal distributed random variables*

Bivariate Characteristic Function

The bivariate characteristic function of the random variables X and Y is defined by

$$\Phi_{XY}(s_1,s_2) = \mathrm{E}\Big(\exp\big(j\,(s_1X + s_2Y)\big)\Big) \quad \text{with} \quad (s_1,s_2)^T \in \mathbb{R}^2,$$

and we can write

$$\Phi_{XY}(s_1,s_2) = \sum_n \sum_m \exp\big(j\,(s_1x_n + s_2y_m)\big)\,p_{XY}(x_n,y_m)$$

and

$$\Phi_{XY}(s_1,s_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\big(j\,(s_1x + s_2y)\big)\,f_{XY}(x,y)\,dx\,dy$$

for the discrete and continuous case respectively.

Properties of bivariate characteristic functions:

- $\Phi_{XY}(s_1,s_2)$ is continuous in s_1 and s_2.
- $|\Phi_{XY}(s_1,s_2)| \leq 1$ for all $(s_1,s_2)^T \in \mathbb{R}^2$.
- $\Phi_{XY}(s_1,0) = \Phi_X(s_1)$ and $\Phi_{XY}(0,s_2) = \Phi_Y(s_2)$.
- $\Phi_{XY}(0,0) = \mathrm{E}(e^0) = \mathrm{E}(1) = 1$.
- $\Phi_{XY}(-\omega_1,-\omega_2)$ is the 2d-Fourier transform of $f_{XY}(x,y)$

  $$\Rightarrow \; f_{XY}(x,y) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j(s_1 x + s_2 y)}\, \Phi_{XY}(s_1,s_2)\, ds_1 ds_2 .$$

- The random variables X and Y are independent, iff $\Phi_{XY}(s_1,s_2) = \Phi_X(s_1) \cdot \Phi_Y(s_2)$.

Exercise 1.9-17:

Determine the density of $Z = X + Y$, if X and Y are independent random variables

Moment Theorem

Supposing the moments $m_{kl} = \mathrm{E}(X^k Y^l)$ exist and therefore

$$\frac{\partial^{k+l}\Phi_{XY}(s_1,s_2)}{\partial s_1^k \partial s_2^l} = \frac{\partial^{k+l}\mathrm{E}\left(e^{j(s_1X+s_2Y)}\right)}{\partial s_1^k \partial s_2^l}$$
$$= j^{k+l}\,\mathrm{E}\left(X^k Y^l e^{j(s_1X+s_2Y)}\right),$$

holds, i.e. the order of differentiation and integration can be interchanged, we can deduce the moment theorem

$$m_{kl} = \mathrm{E}\left(X^k Y^l\right) = \frac{1}{j^{k+l}} \left.\frac{\partial^{k+l}\Phi_{XY}(s_1,s_2)}{\partial s_1^k \partial s_2^l}\right|_{s_1=0,s_2=0}.$$

1.9.3 Mean Square Error Estimation

Non-linear Mean Square Error Estimation

We wish to estimate the random variable Y by a function of the random variable X. Our aim is to find any function $g(X)$ such that the mean square error

$$q(g) = \mathrm{E}\left(\left(Y - g(X)\right)^2\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(y - g(x)\right)^2 f_{XY}(x,y)\,dx\,dy$$

is minimum.

Over the class of all functions g for which the expected value exists, $q(g)$ is minimized by choosing

$$\hat{g}(x) = \psi_Y(x) = \mathrm{E}_Y(Y \mid X = x).$$

Exercise 1.9-18: *Proof of the non-linear mean square error estimation result*

Linear Mean Square Error Estimation

Now we wish to estimate the random variable Y by a linear function of X, i.e. $g(X) = a_1 X + a_0$. The objective is to minimise the mean square error

$$q(a_1, a_0) = \mathrm{E}\left(\left(Y - (a_1 X + a_0)\right)^2\right)$$

by varying the parameters a_1 and a_0.

That is, we want to solve the minimisation problem

$$\hat{\mathbf{a}} = \left(\hat{a}_1, \hat{a}_0\right)^T = \underset{a_1, a_0}{\arg\min}\left(q(a_1, a_0)\right).$$

The mean square error is minimum if the equation system

$$\left.\begin{pmatrix} \partial q(a_1,a_0)/\partial a_1 \\ \partial q(a_1,a_0)/\partial a_0 \end{pmatrix}\right|_{\substack{a_1=\hat{a}_1 \\ a_0=\hat{a}_0}} = -2\begin{pmatrix} \mathrm{E}\left(\left(Y-\hat{a}_1X-\hat{a}_0\right)X\right) \\ \mathrm{E}\left(Y-\hat{a}_1X-\hat{a}_0\right) \end{pmatrix} = \mathbf{0}$$

is satisfied. Solving the equation system provides

$$\hat{a}_1 = \frac{\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)}{\mathrm{E}(X^2)-\left(\mathrm{E}(X)\right)^2} = \frac{m_{11}-m_{10}m_{01}}{m_{20}-m_{10}^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

$$= \rho(X,Y)\sigma_Y/\sigma_X$$

and

$$\hat{a}_0 = \mathrm{E}(Y)-\hat{a}_1\mathrm{E}(X) = m_{01}-\hat{a}_1 m_{10} = \frac{m_{20}m_{01}-m_{10}m_{11}}{m_{20}-m_{10}^2}.$$

Substitution of a_1 and a_0 by $\hat{a}_1$ and $\hat{a}_0$ in $q(a_1,a_0)$ gives the minimum mean square error

$$\begin{aligned} q(\hat{a}_1,\hat{a}_0) &= \mathrm{E}\left(\left(Y-(\hat{a}_1X+\hat{a}_0)\right)^2\right) \\ &= \mathrm{E}\left(\left(Y-\mathrm{E}(Y)-\hat{a}_1\left(X-\mathrm{E}(X)\right)\right)^2\right) \\ &= \mathrm{Var}(Y)-2\hat{a}_1\mathrm{Cov}(X,Y)+\hat{a}_1^2\mathrm{Var}(X) \\ &= \sigma_Y^2\left(1-\rho^2(X,Y)\right). \end{aligned}$$

Moreover, the minimum mean square errors of the linear and non-linear approach obviously satisfy the relationship

$$q(\hat{a}_1,\hat{a}_0) \geq q(\hat{g}).$$

Exercise 1.9-19:

Mean square error estimation for bivariate normal distributed random variables

1.10 Vector-valued Random Variables

1.10.1 Multivariate Distributions

Multivariate Distributions and Density Functions

The basic ideas of bivariate distributions are easily extended to the general case, where instead of two, n random variables $X_1, X_2, \ldots, X_n$ are considered.

Thus, the distribution function of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ is defined by

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n)$$

and the density function is given by

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{X_1 \ldots X_n}(x_1, \ldots, x_n).$$

Marginal Distributions and Density Functions

For a given multivariate distribution function of $X_1, X_2, \ldots,$ X_n, the marginal distribution and density function of $X_1, X_2,$ $\ldots, X_k$ can be expressed by

$$F_{X_1 \ldots X_k}(x_1, \ldots, x_k) = F_{X_1 \ldots X_k \ldots X_n}(x_1, \ldots, x_k, \infty, \ldots, \infty)$$

and

$$f_{X_1 \ldots X_k}(x_1, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 \ldots X_k \ldots X_n}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) dx_{k+1} \cdots dx_n$$

respectively.

Conditional Distributions and Density Functions

The conditional distribution and density function of $X_1,\ldots,X_k$ under the condition $\{X_{k+1}=x_{k+1},\ldots,X_n=x_n\}$ is given by

$$F_{X_1\ldots X_k}(x_1,\ldots,x_k \mid X_{k+1}=x_{k+1},\ldots,X_n=x_n)=$$
$$=P(X_1\le x_1,\ldots,X_k\le x_k \mid X_{k+1}=x_{k+1},\ldots,X_n=x_n)$$

and

$$f_{X_1\ldots X_k}(x_1,\ldots,x_k \mid X_{k+1}=x_{k+1},\ldots,X_n=x_n)=\frac{f_{X_1\ldots X_n}(x_1,\ldots,x_n)}{f_{X_{k+1}\ldots X_n}(x_{k+1},\ldots,x_n)}.$$

Notation:

$$F_{X_1 \ldots X_k}(x_1, \ldots, x_k \,|\, x_{k+1}, \ldots, x_n) = $$
$$= F_{X_1 \ldots X_k}(x_1, \ldots, x_k \,|\, X_{k+1} = x_{k+1}, \ldots, X_n = x_n)$$

and

$$f_{X_1 \ldots X_k}(x_1, \ldots, x_k \,|\, x_{k+1}, \ldots, x_n) = $$
$$= f_{X_1 \ldots X_k}(x_1, \ldots, x_k \,|\, X_{k+1} = x_{k+1}, \ldots, X_n = x_n).$$

Exercise 1.10-1:

Calculations with conditional densities

Independent Random Variables

The random variables $X_1, X_2, \ldots, X_n$ are said to be independent, if the multivariate distribution or density function breaks down into the product of n marginal distributions or density functions.

Thus $X_1, X_2, \ldots, X_n$ are independent if the multivariate distribution can be written in the form

$$F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

(valid for the discrete and continuous case) or if the density function can be written in the form

$$f_{X_1 \dots X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

(applies for the continuous case).

The random variables $X_1, \dots, X_k$ are independent from the random variables $X_{k+1}, \dots, X_n$ if the multivariate distribution or the density function can be expressed by

$$F_{X_1 \dots X_n}(x_1, \dots, x_n) = F_{X_1 \dots X_k}(x_1, \dots, x_k) \cdot F_{X_{k+1} \dots X_n}(x_{k+1}, \dots, x_n)$$

or

$$f_{X_1 \dots X_n}(x_1, \dots, x_n) = f_{X_1 \dots X_k}(x_1, \dots, x_k) \cdot f_{X_{k+1} \dots X_n}(x_{k+1}, \dots, x_n).$$

1.10.2 Transformation of Vector-valued Random Variables

Suppose $X_1, X_2, \ldots, X_n$ are random variables and $g_1, \ldots, g_m$ are functions with $m \leq n$ such that

$$Y_1 = g_1(X_1, \ldots, X_n), \ldots, Y_m = g_m(X_1, \ldots, X_n)$$

are random variables. Then the multivariate distribution of $Y_1, Y_2, \ldots, Y_m$ is given by

$$F_{Y_1 \ldots Y_m}(y_1, \ldots, y_m) =$$

$$= P\left(g_1(X_1, \ldots, X_n) \leq y_1, \ldots, g_m(X_1, \ldots, X_n) \leq y_m\right).$$

The multivariate density function of $Y_1, Y_2, \ldots, Y_m$ can be determined as follows. In case of $m < n$ we define $n - m$ auxiliary variables (functions)

$$y_i = g_i(x_1, \ldots, x_n) \qquad \text{for} \quad i = 1, \ldots, m$$

$$y_i = x_i = g_i(x_1, \ldots, x_n) \qquad \text{for} \quad i = m+1, \ldots, n.$$

Assume that the equation system

$$(y_1, \ldots, y_n)^T = \big(g_1(x_1, \ldots, x_n), \ldots, g_n(x_1, \ldots, x_n)\big)^T$$

has the unique solution

$$(x_1, \ldots, x_n)^T = \big(g_1^{-1}(y_1, \ldots, y_n), \ldots, g_n^{-1}(y_1, \ldots, y_n)\big)^T.$$

Furthermore, suppose $g_1(x_1,x_2,\ldots,x_n),\ldots,g_n(x_1,x_2,\ldots,x_n)$ have continuous partial derivatives and the determinant of the Jacobian

$$\mathbf{J}(x_1,\ldots,x_n)=\begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix}$$

does not vanish, i.e.

$$\det\left(\mathbf{J}(x_1,\ldots,x_n)\right)\neq 0,$$

then the multivariate density function of $Y_1, Y_2, \dots, Y_n$ is given by

$$f_{Y_1 \dots Y_n}(y_1, \dots, y_n) = \left. \frac{f_{X_1 \dots X_n}(x_1, \dots, x_n)}{\left| \det \mathbf{J}(x_1, \dots, x_n) \right|} \right|_{\substack{x_1 = g_1^{-1}(y_1, \dots, y_n) \\ \vdots \\ x_n = g_n^{-1}(y_1, \dots, y_n)}} .$$

Finally, integration over $y_{m+1}, \dots, y_n$ provides the multivariate density function of $Y_1, Y_2, \dots, Y_m$.

$$f_{Y_1 \dots Y_m}(y_1, \dots, y_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{Y_1 \dots Y_n}(y_1, \dots, y_n) dy_{m+1} \cdots dy_n$$

Exercise 1.10-2:
Density of a linear combination of random variables

1.10.3 Expectations for Vector-valued Random Variables

Expected Value

For a measurable function $g(X_1, X_2, \ldots, X_n)$ the expected value of $g(X_1, X_2, \ldots, X_n)$ is defined by

$$\mathrm{E}\big(g(X_1,\ldots,X_n)\big) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\ldots,x_n)\,d^nF_{X_1\ldots X_n}(x_1,\ldots,x_n)$$

$$= \begin{cases} \displaystyle\sum_{i_1}\cdots\sum_{i_n} g(x_{1,i_1},\ldots,x_{n,i_n})\,p_{i_1\ldots i_n} \\ \displaystyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\ldots,x_n)\,f_{X_1\ldots X_n}(x_1,\ldots,x_n)\,dx_1\cdots dx_n \end{cases}$$

provided that the sum resp. integral converges absolutely.

Conditional Expected Value

Let $X_1, X_2, \ldots, X_n$ are multivariate distributed continuous random variables.

The conditional expected value of X_1, given $X_2 = x_2, \ldots, X_n = x_n$ is defined by

$$
\begin{aligned}
\mathrm{E}\left(X_1 \mid X_2 = x_2, \ldots, X_n = x_n\right) &= \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1 \mid x_2, \ldots, x_n)\, dx_1 \\
&= \int_{-\infty}^{\infty} x_1 \frac{f_{X_1 \ldots X_n}(x_1, \ldots, x_n)}{f_{X_2 \ldots X_n}(x_2, \ldots, x_n)}\, dx_1 \\
&= \psi_{X_1}(x_2, \ldots, x_n).
\end{aligned}
$$

Conditional Expectation

Now, replacing $x_2,\ldots,x_n$ by $X_2,\ldots,X_n$ in $\psi_{X_1}(x_2,\ldots,x_n)$ we obtain the random variable

$$\psi_{X_1}(X_2,\ldots,X_n) = \mathrm{E}\left(X_1 \mid X_2,\ldots,X_n\right)$$

which is called the conditional expectation of X_1, given $X_2,\ldots,X_n$.

Properties of conditional expectations:

$$\mathrm{E}\left(\mathrm{E}\left(X_1 \mid X_2,\ldots,X_n\right)\right) = \mathrm{E}(X_1),$$

$$\begin{aligned}\mathrm{E}\left(X_1 X_2 \mid X_3\right) &= \mathrm{E}\left(\mathrm{E}\left(X_1 X_2 \mid X_2, X_3\right) \mid X_3\right) \\ &= \mathrm{E}\left(X_2 \mathrm{E}\left(X_1 \mid X_2, X_3\right) \mid X_3\right).\end{aligned}$$

Uncorrelated and Orthogonal Random Variables

The random variables $X_1,\ldots,X_n$ are called uncorrelated resp. called orthogonal if for all $i \neq j$

$$\mathrm{E}(X_i X_j) = \mathrm{E}(X_i)\mathrm{E}(X_j) \quad \text{resp.} \quad \mathrm{E}(X_i X_j) = 0$$

holds.

Consequently, if $X_1,\ldots,X_n$ are uncorrelated resp. are orthogonal and

$$U = \sum_{i=1}^{n} X_i,$$

we can write

$$\sigma_U^2 = \sum_{i=1}^{n} \sigma_{X_i}^2 \quad \text{resp.} \quad \mathrm{E}(U^2) = \sum_{i=1}^{n} \mathrm{E}\left(X_i^2\right).$$

Moments

Let $X_1,\ldots,X_n$ be multivariate distributed continuous random variables. Then common moments of $X_1,\ldots,X_n$ can be determined by

$$m_{k_1\ldots k_n} = \mathrm{E}\left(X_1^{k_1},\ldots,X_n^{k_n}\right)$$

$$= \int_{-\infty}^{\infty} x_1^{k_1}\cdots x_n^{k_n}\, f_{X_1\ldots X_n}(x_1,\ldots,x_n)\,dx_1\cdots dx_n$$

and where the order of the moments is defined by

$$r = k_1 + \cdots + k_n.$$

Characteristic Functions

The characteristic function of the multivariate distributed random variables $X_1,\ldots,X_n$ is given by

$$\Phi_{\mathbf{X}}(\mathbf{s}) = \Phi_{X_1\ldots X_n}(s_1,\ldots,s_n)$$

$$= \mathrm{E}\left(\exp\left(j\sum_{i=1}^{n} s_i X_i\right)\right) = \mathrm{E}\left(\exp\left(j\,\mathbf{s}^T\mathbf{X}\right)\right).$$

Hence, if the random variables $X_1,\ldots,X_n$ are independent, we obtain

$$\Phi_{\mathbf{X}}(\mathbf{s}) = \mathrm{E}\left(\prod_{i=1}^{n}\exp(js_iX_i)\right) = \prod_{i=1}^{n}\mathrm{E}\left(\exp(js_iX_i)\right) = \prod_{i=1}^{n}\Phi_{X_i}(s_i).$$

Let $X_1,\ldots,X_n$ be now independent random variables and

$$U=\sum_{i=1}^{n} X_i\,,$$

then the characteristic function of U is given by

$$\Phi_U(s)=\mathrm{E}\big(\exp(j\,sU)\big)=\mathrm{E}\left(\exp\left(j\sum_{i=1}^{n} sX_i\right)\right)=\prod_{i=1}^{n}\Phi_{X_i}(s).$$

If in addition, the $X_1,\ldots,X_n$ possess the density functions $f_{X_1},\ldots,f_{X_n}$, the density function of U can be determined by

$$f_U(u)=\left(f_{X_1}*f_{X_2}*\cdots*f_{X_n}\right)(u).$$

Exercise 1.10-3:
Sample mean and sample variance of independent normally distributed random variables

Moment Theorem

Suppose that the moments $m_{k_1\ldots k_n} = \mathrm{E}\left(X_1^{k_1}\cdots X_n^{k_n}\right)$ exist and therefore

$$\frac{\partial^{k_1+\cdots+k_n}\Phi_{X_1,\ldots,X_n}(s_1,\ldots,s_n)}{\partial s_1^{k_1}\cdots\partial s_n^{k_n}} = \frac{\partial^{k_1+\cdots+k_n}\mathrm{E}\left(e^{j(s_1X_1+\cdots+s_nX_n)}\right)}{\partial s_1^{k_1}\cdots\partial s_n^{k_n}}$$

$$= j^{k_1+\cdots+k_n}\mathrm{E}\left(X_1^{k_1}\cdots X_n^{k_n}\, e^{j(s_1X_1+\cdots+s_nX_n)}\right),$$

holds, we can deduce the moment theorem

$$m_{k_1\ldots k_n} = \mathrm{E}\left(X_1^{k_1}\cdots X_n^{k_n}\right) = \frac{1}{j^{k_1+\cdots+k_n}}\left.\frac{\partial^{k_1+\cdots+k_n}\Phi_{X_1\cdots X_n}(s_1,\ldots,s_n)}{\partial s^{k_1}\cdots\partial s^{k_n}}\right|_{\substack{s_1=0\\ \vdots\\ s_n=0}}.$$

Exercise 1.10-4:
Application of the Moment Theorem

1.10.4 Mean Square Error Estimation

Non-linear Mean Square Error Estimation

To estimate the random variable X_0 by means of the random variables $X_1,\dots,X_n$, we want to find any function $g(X_1,\dots,X_n)$ that minimizes the mean square error

$$q(g) = \mathrm{E}\left(\left(X_0 - g(X_1,\dots,X_n)\right)^2\right)$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(x_0 - g(x_1,\dots,x_n)\right)^2 f_{X_0\dots X_n}(x_0,\dots,x_n)\,dx_0\cdots dx_n.$$

Over the class of all functions g for which the expected value exists, $q(g)$ is minimised by choosing

$$\hat{g}(x_1,\dots,x_n) = \psi_{X_0}(x_1,\dots,x_n) = \mathrm{E}_{X_0}\left(X_0 \mid X_1 = x_1,\dots,X_n = x_n\right).$$

Linear Mean Square Error Estimation

Now we wish to estimate the random variable X_0 by a linear function of $X_1,\ldots,X_n$, i.e.

$$g(X_1,\ldots,X_n) = a_1X_1 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i = \mathbf{a}^T\mathbf{X},$$

such that the mean square error

$$q(\mathbf{a}) = q(a_1,\ldots,a_n) = \mathrm{E}\left(\left(X_0 - \sum_{i=1}^{n} a_iX_i\right)^2\right) = \mathrm{E}\left(\left(X_0 - \mathbf{a}^T\mathbf{X}\right)^2\right)$$

is minimized by varying the parameter vector **a**.

Thus, we have to solve the minimisation problem

$$\hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_n)^T = \underset{a_1, \ldots, a_n}{\arg\min} \big(q(a_1, \ldots, a_n) \big).$$

The mean square error is minimum if

$$\left. \frac{\partial q(\mathbf{a})}{\partial a_i} \right|_{\mathbf{a}=\hat{\mathbf{a}}} = -2\mathrm{E}\Big(\big(X_0 - \hat{\mathbf{a}}^T \mathbf{X} \big) X_i \Big) = 0$$

is satisfied for $i = 1, \ldots, n$. After some manipulations we obtain the equation system

$$\mathbf{R}\hat{\mathbf{a}} = \mathbf{r},$$

where

$$\mathbf{R} = \mathrm{E}(\mathbf{X}\mathbf{X}^T) \text{ and } \mathbf{r} = \mathrm{E}(X_0 \mathbf{X}).$$

With the estimate $\hat{\mathbf{a}} = \mathbf{R}^{-1}\mathbf{r}$ the minimum mean square error is given by

$$
\begin{aligned}
q(\hat{\mathbf{a}}) &= \mathrm{E}\left(\left(X_0 - \hat{\mathbf{a}}^T\mathbf{X}\right)^2\right) \\
&= \mathrm{E}\left(\left(X_0 - (\mathbf{R}^{-1}\mathbf{r})^T\mathbf{X}\right)^2\right) = \mathrm{E}\left(\left(X_0 - \mathbf{r}^T\mathbf{R}^{-1}\mathbf{X}\right)^2\right) \\
&= \mathrm{E}\left(X_0 X_0 - 2\mathbf{r}^T\mathbf{R}^{-1}\mathbf{X}X_0 + \mathbf{r}^T\mathbf{R}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{r}\right) \\
&= \mathrm{E}\left(X_0^2\right) - 2\mathbf{r}^T\mathbf{R}^{-1}\mathbf{r} + \mathbf{r}^T\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{r} \\
&= \mathrm{E}\left(X_0^2\right) - \mathbf{r}^T\mathbf{R}^{-1}\mathbf{r}.
\end{aligned}
$$

Orthogonality Principle

The coefficient vector $\hat{\mathbf{a}}$ that minimizes the mean square error satisfies the equation

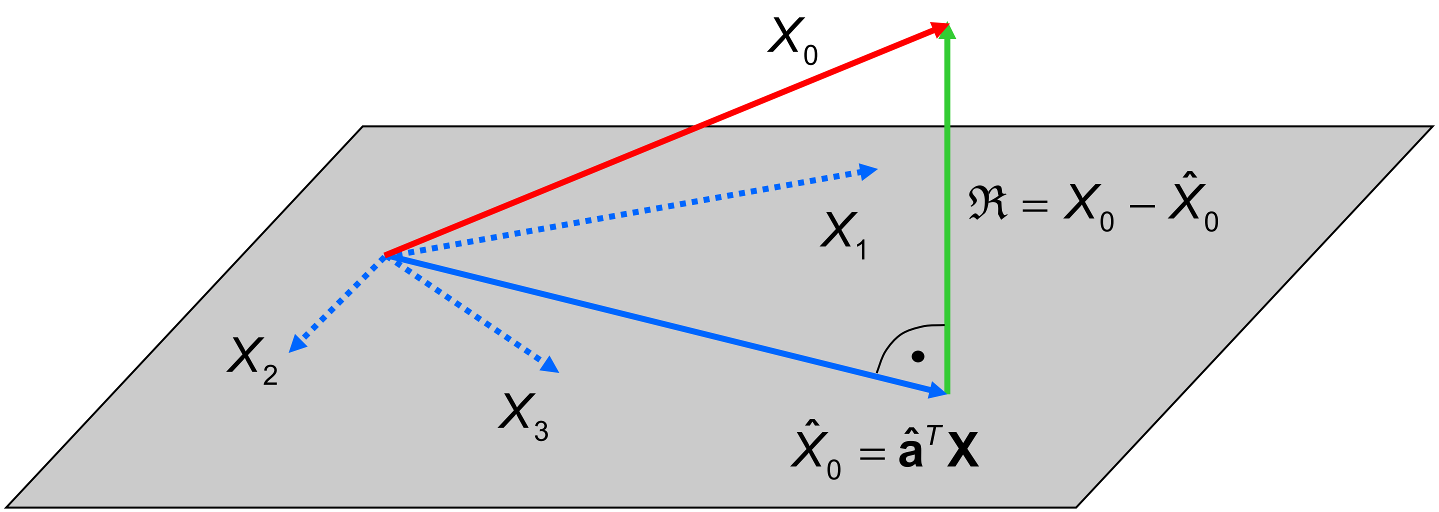
$$\mathrm{E}\left(\left(X_0 - \hat{\mathbf{a}}^T \mathbf{X}\right) X_i\right) = 0, \quad i = 1, \ldots, n.$$

Thus, the residual

$$\mathfrak{R} = X_0 - \hat{\mathbf{a}}^T \mathbf{X} = X_0 - \hat{X}_0$$

is orthogonal to the observations $X_1, \ldots, X_n$. This is known as the orthogonality principle.

A geometric interpretation of this result is shown in the figure below, where X_0, $\mathfrak{R}$ and $X_1,\ldots,X_n$ are thought of as vectors.



Consequently,

1) $\hat{X}_0$ is the orthogonal projection of X_0 onto the subspace U spanned by the observations $X_1,\ldots,X_n$.

2) $\mathfrak{R}$ is orthogonal to the subspace U and therefore orthogonal to the observations $X_1,\ldots,X_n$.

3) $q(\hat{\mathbf{a}}) = \mathrm{E}\left((X_0 - \hat{X}_0)^2\right) = \mathrm{E}\left((X_0 - \hat{X}_0)X_0\right) = 0$ if and only if X_0 lies within the subspace U, i.e. X_0 is linearly dependent on $X_1,\ldots,X_n$, thus $X_0 = \hat{\mathbf{a}}^T\mathbf{X}$.

4) $q(\hat{\mathbf{a}}) = \mathrm{E}\left((X_0 - \hat{X}_0)X_0\right) = \mathrm{E}\left(X_0^2\right)$ if and only if X_0 is orthogonal to the subspace U.

1.10.5 Multivariate Normal Distribution

Characteristic Function of a Normally Distributed Vector-valued Random Variable

Let $U_1,\ldots,U_m$ be m independent standardized normally distributed random variables, i.e. $U_k \sim \mathcal{N}(0,1)$.

The characteristic function of U_k is given by

$$\Phi_{U_k}(r_k) = \exp\left(-\frac{r_k^2}{2}\right),$$

cf. Exercise 1.9-9.

Consequently, the characteristic function of the random vector $\mathbf{U} = (U_1,\ldots,U_m)^T$ is

$$\Phi_{\mathbf{U}}(\mathbf{r}) = \mathrm{E}\left(\exp(j\,\mathbf{r}^T\mathbf{U})\right) = \mathrm{E}\left(\prod_{k=1}^{m}\exp(j\,r_k U_k)\right)$$

$$= \prod_{k=1}^{m}\mathrm{E}\left(\exp(j\,r_k U_k)\right) = \prod_{k=1}^{m}\Phi_{U_k}(r_k)$$

$$= \exp\left(-\frac{1}{2}\sum_{k=1}^{m} r_k^2\right) = \exp\left(-\frac{1}{2}\mathbf{r}^T\mathbf{r}\right).$$

Transformation of the random vector **U** by a $(n \times m)$ matrix **A** provides the random vector

$$\mathbf{V} = (V_1,\ldots,V_n)^T = \mathbf{A}\mathbf{U}$$

that possesses the characteristic function

$$\Phi_{\mathbf{V}}(\mathbf{s}) = \mathrm{E}\left(\exp(j\,\mathbf{s}^T\mathbf{V})\right) = \mathrm{E}\left(\exp(j\,\mathbf{s}^T\mathbf{A}\mathbf{U})\right)$$

$$= \Phi_{\mathbf{U}}\left((\mathbf{s}^T\mathbf{A})^T\right) = \exp\left(-\frac{1}{2}\mathbf{s}^T\mathbf{A}(\mathbf{s}^T\mathbf{A})^T\right)$$

$$= \exp\left(-\frac{1}{2}\mathbf{s}^T\mathbf{A}\mathbf{A}^T\mathbf{s}\right) = \exp\left(-\frac{1}{2}\mathbf{s}^T\boldsymbol{\Sigma}\mathbf{s}\right),$$

where

$$\boldsymbol{\Sigma} = \mathrm{E}(\mathbf{V}\mathbf{V}^T) = \mathrm{E}\left(\mathbf{A}\mathbf{U}(\mathbf{A}\mathbf{U})^T\right) = \mathrm{E}\left(\mathbf{A}\mathbf{U}\mathbf{U}^T\mathbf{A}^T\right)$$

$$= \mathbf{A}\mathrm{E}(\mathbf{U}\mathbf{U}^T)\,\mathbf{A}^T = \mathbf{A}\mathbf{I}\mathbf{A}^T = \mathbf{A}\mathbf{A}^T$$

represents the covariance matrix of **V**.

Translation of the random vector **V** by a constant vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ leads to the random vector

$$\mathbf{W} = (W_1, \ldots, W_n)^T = \mathbf{V} + \boldsymbol{\mu} = \mathbf{A}\mathbf{U} + \boldsymbol{\mu}.$$

The characteristic function of **W** can be expressed by

$$\Phi_{\mathbf{W}}(\mathbf{s}) = \mathrm{E}\left(\exp(j\mathbf{s}^T\mathbf{W})\right) = \mathrm{E}\left(\exp(j\mathbf{s}^T(\mathbf{V} + \boldsymbol{\mu}))\right)$$

$$= \exp(j\mathbf{s}^T\boldsymbol{\mu})\mathrm{E}\left(\exp(j\mathbf{s}^T\mathbf{V})\right) = \exp\left(j\mathbf{s}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{s}^T\boldsymbol{\Sigma}\mathbf{s}\right).$$

$\Phi_{\mathbf{W}}(\mathbf{s})$ represents the characteristic function of the normally distributed random vector, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the vector-valued expected value and the symmetric and non-negative definite covariance matrix, respectively, i.e.

$$E(\mathbf{W}) = E(\mathbf{V}+\boldsymbol{\mu}) = E(\mathbf{V})+\boldsymbol{\mu} = \boldsymbol{\mu}$$

and $$E\left((\mathbf{W}-\boldsymbol{\mu})(\mathbf{W}-\boldsymbol{\mu})^T\right) = E\left(\mathbf{V}\mathbf{V}^T\right) = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T,$$

$$\mathbf{s}^T\boldsymbol{\Sigma}\mathbf{s} \ge 0 \quad \forall \mathbf{s} \in \mathbb{R}^n.$$

Theorem:

Let $\mathbf{X} \sim \mathcal{N}_m(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$ and $\mathbf{Y} = \mathbf{A}\mathbf{X}+\mathbf{b}$, where the entries of the $(n \times m)$ matrix $\mathbf{A}$ and the $(n \times 1)$ vector $\mathbf{b}$ are constants. Then $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}_\mathbf{Y}, \boldsymbol{\Sigma}_\mathbf{Y})$ with

$$\boldsymbol{\mu}_\mathbf{Y} = E(\mathbf{Y}) = \mathbf{A}\boldsymbol{\mu}_\mathbf{X} + \mathbf{b}$$

and

$$\boldsymbol{\Sigma}_\mathbf{Y} = E\left((\mathbf{Y}-\boldsymbol{\mu}_\mathbf{Y})(\mathbf{Y}-\boldsymbol{\mu}_\mathbf{Y})^T\right) = \mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^T.$$

Exercise 1.10-5:
Proof of the Theorem

Density Function of a Normally Distributed Vector-valued Random Variable

Let $U_1,\ldots,U_m$ be m independent standardized normally distributed random variables, i.e. $U_k \sim \mathcal{N}(0,1)$.

Thus, the density function of the random vector $\mathbf{U} = (U_1, \ldots, U_m)^T$ can be written as

$$f_{\mathbf{U}}(\mathbf{u}) = \prod_{k=1}^{m} f_{U_k}(u_k) = \prod_{k=1}^{m}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u_k^2\right)\right)$$

$$= (2\pi)^{-\frac{m}{2}}\exp\left(-\frac{1}{2}\sum_{k=1}^{m}u_k^2\right) = (2\pi)^{-\frac{m}{2}}\exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{u}\right).$$

Now, we transform the random vector $\mathbf{U}$ by

$$\mathbf{V} = (V_1,\ldots,V_m)^T = \mathbf{A}\mathbf{U} + \boldsymbol{\mu},$$

where $\mathbf{A}$ denotes a regular $(m \times m)$ matrix, i.e. $\det(\mathbf{A}) \neq 0$.

With the inverse transform

$$\mathbf{U} = \mathbf{A}^{-1}(\mathbf{V} - \boldsymbol{\mu}),$$

the determinant of the Jacobian

$$\det(\mathbf{J}(\mathbf{u})) = \det\left(\frac{\partial(v_1,\ldots,v_m)}{\partial(u_1,\ldots,u_m)}\right) = \det(\mathbf{A}),$$

the well known identities

$$\det(\mathbf{A}) = \det(\mathbf{A}^T)$$

$$\det(\mathbf{\Sigma}) = \det(\mathbf{A}\mathbf{A}^T)$$

$$= \det(\mathbf{A})\det(\mathbf{A}^T) = \left(\det(\mathbf{A})\right)^2 > 0$$

$$\mathbf{\Sigma}^{-1} = (\mathbf{A}\mathbf{A}^T)^{-1} = (\mathbf{A}^T)^{-1}\mathbf{A}^{-1} = (\mathbf{A}^{-1})^T\mathbf{A}^{-1}$$

and the result derived for determining the density of a transformed random vector

$$f_{\mathbf{V}}(\mathbf{v}) = \frac{f_{\mathbf{U}}\left(\mathbf{A}^{-1}(\mathbf{v}-\boldsymbol{\mu})\right)}{\left|\det(\mathbf{A})\right|},$$

we obtain

$$
\begin{aligned}
f_{\mathbf{v}}(\mathbf{v}) &= (2\pi)^{-\frac{m}{2}} \left|\det(\mathbf{A})\right|^{-1} \exp\left(-\frac{1}{2}\left(\mathbf{A}^{-1}(\mathbf{v}-\boldsymbol{\mu})\right)^T \mathbf{A}^{-1}(\mathbf{v}-\boldsymbol{\mu}) \right) \\
&= (2\pi)^{-\frac{m}{2}} \left|\det(\mathbf{A})\right|^{-1} \exp\left(-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^T (\mathbf{A}^{-1})^T \mathbf{A}^{-1}(\mathbf{v}-\boldsymbol{\mu}) \right) \\
&= (2\pi)^{-\frac{m}{2}} \left|\det(\mathbf{A})\right|^{-1} \exp\left(-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^T (\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{v}-\boldsymbol{\mu}) \right) \\
&= (2\pi)^{-\frac{m}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{v}-\boldsymbol{\mu}) \right).
\end{aligned}
$$

Exercise 1.10-6:
Transformation to standardized normal distribution

Composed Vector-valued Random Variables

Theorem:

Let $\mathbf{X}$ be a random vector composed of the random vectors $\mathbf{X}_1$, $\mathbf{X}_2$ and obeying the distribution

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim \mathcal{N}_k\left(\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right).$$

Then $\mathbf{X}_1$ and $\mathbf{X}_2$ possessing the marginal distributions

$$\mathbf{X}_1 \sim \mathcal{N}_{k_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}),\ \ \mathbf{X}_2 \sim \mathcal{N}_{k_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

and the conditional distribution

$$\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2 \sim \mathcal{N}_{k_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right).$$

Exercise 1.10-7:
Proof of the Theorem

Exercise 1.10-8:
Mean square error estimation

1.11 Sequences of Random Variables

1.11.1 Convergence Concepts

Let $(X_n) = X_1, X_2, \ldots$ be a sequence of random variables. For any specific ξ, $(X_n(\xi))$ is a sequence that might or might not converge and where the notion of convergence can be given several interpretations.

Recall, that a deterministic sequence (x_n) tends to a limit x, if for any given $\varepsilon > 0$, we can find a number N_ε such that

$$\left| x_n - x \right| < \varepsilon \quad \forall n > N_\varepsilon .$$

Convergence (everywhere)

If the sequence $(X_n(\xi))$ tends to a number $X(\xi)$ for every $\xi \in \Xi$, then we say that the random sequence (X_n) converges everywhere to the random variable X and we write this as

$$\lim_{n\to\infty} X_n = X \quad \text{or} \quad X_n \xrightarrow[n\to\infty]{} X.$$

Convergence almost surely (a.s.)

If the probability of the set of all events ξ that satisfy $\lim_{n\to\infty} X_n(\xi) = X(\xi)$, equals 1, i.e.

$$P\left(\left\{\xi : \lim_{n\to\infty} X_n(\xi) = X(\xi)\right\}\right) = P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

or equivalently

$$\lim_{m\to\infty} P\left(\sup_{n\geq m}\left|X_n - X\right| \geq \varepsilon\right) = 0 \quad \forall \varepsilon > 0$$

then we say that the random sequence (X_n) converges almost surely (with probability 1) to the random variable X and we write this as

$$\lim_{n\to\infty} X_n = X \quad \text{or} \quad X_n \xrightarrow[n\to\infty]{a.s.} X.$$

Convergence in Mean Square (m.s.)

Let (X_n) be a sequence of random variables. We say that the sequence converges in mean square to a random variable X if

$$\lim_{n\to\infty} \mathrm{E}\left(\left(X_n - X\right)^2\right) = 0$$

holds and we write this as

$$\underset{n\to\infty}{\mathrm{l.i.m.}}\, X_n = X \quad \text{or} \quad X_n \xrightarrow[n\to\infty]{m.s.} X,$$

where l.i.m. denotes the limit in mean.

Convergence in Probability (p)

A sequence of random variables (X_n) is said to converge in probability to a random variable X if for every $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P\left(\left|X_n - X\right| \geq \varepsilon\right) = 0$$

and we write this as

$$p\lim_{n\to\infty} X_n = X \quad \text{or} \quad X_n \xrightarrow[n\to\infty]{p} X,$$

where $p\lim$ denotes the limit in probability.

Convergence in Distribution (d)

Let (F_{X_n}) be the sequence of distribution functions of the sequence of random variables (X_n). Then (X_n) is said to converge in distribution (or in Law) to a random variable X with the distribution function F_X if

$$F_{X_n} \xrightarrow[n\to\infty]{} F_X$$

at all continuity points of F_X. Such a convergence is expressed by

$$X_n \xrightarrow[n\to\infty]{d} X \quad \text{or} \quad X_n \xrightarrow[n\to\infty]{L} X.$$

Relationships among the various types of convergence

1) Convergence with probability 1 implies convergence in probability.

2) Convergence with probability 1 implies convergence in mean square, provided second order moments exist.

3) Convergence in mean square implies convergence in probability.

4) Convergence in probability implies convergence in distribution.

Exercise 1.11-1:
Proof of statement 1) and 3)

1.11.2 Laws of Large Numbers

Chebyschev's Theorem (Weak Law of Large Numbers)

Let (X_k) be a sequence of pairwise uncorrelated random variables with

$$\mathrm{E}(X_k) = \mu_k, \;\; \mathrm{Var}(X_k) = \sigma_k^2 \;\; \text{and} \;\; \lim_{n\to\infty} \frac{1}{n^2} \sum_{k=1}^{n} \sigma_k^2 = 0.$$

Then we have

$$\bar{X}_n \xrightarrow[n\to\infty]{m.s.} \mu,$$

where

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \;\; \text{and} \;\; \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mu_k = \mu.$$

Exercise 1.11-2:
Proof of Chebyschev's Theorem

Kolmogorov's Theorem (Strong Law of Large Numbers)

Let (X_k) be a sequence of independent and identically distributed (i.i.d.) random variables. Furthermore, the moments $\mathrm{E}(X_k)$ exist and are equal to μ. Then we have

$$\bar{X}_n \xrightarrow[n\to\infty]{a.s.} \mu,$$

where

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

1.11.3 Central Limit Theorems

Lindeberg-Levy's Theorem

Let (X_k) be a sequence of independent and identically distributed random variables, such that $\mathrm{E}(X_k) = \mu$ and $\mathrm{Var}(X_k) = \sigma^2 \neq 0$. Then the distribution function of the random variable

$$Y_n = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$$

tends to that of a standardized normal distribution as n approaches infinity, i.e.

$$F_{Y_n}(y) \xrightarrow[n\to\infty]{} \Phi(y).$$

Exercise 1.11-3:
Proof of Lindeberg-Levy's Theorem

Liaponov's Theorem

Let (X_k) be a sequence of independent distributed random variables, such that $E(X_k)=\mu_k$, $Var(X_k)=\sigma_k^2 \neq 0$ and $E|X_k-\mu_k|^3=\beta_k$. Furthermore, let

$$B_n=\left(\sum_{k=1}^{n}\beta_k\right)^{1/3},\ C_n=\left(\sum_{k=1}^{n}\sigma_k^2\right)^{1/2} \text{ and } \lim_{n\to\infty}\frac{B_n}{C_n}=0.$$

Then the distribution function of the random variable

$$Y_n=\frac{1}{C_n}\sum_{k=1}^{n}(X_k-\mu_k)$$

tends to that of a standardized normal distribution as *n* approaches infinity, i.e.

$$F_{Y_n}(y)\xrightarrow[n\to\infty]{}\Phi(y).$$

References to Chapter 1

[1] J.F. Böhme, *Stochastische Signale*, Teubner, 1998

[2] M. Fisz, *Probability Theory and Mathematical Statistics*, Krieger Publishing Company, 1980

[3] G. Hänsler, *Statistische Signale*, Springer, 2001

[4] S. Kay, Intuitive Probability and Random Processes using MATLAB, Springer, 2006

[5] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 1991

[6] C.R. Rao, *Linear Statistical Inference and Its Application*, John Wiley, 1973

[7] G. Wunsch, H. Schreiber, *Stochastische Systeme*, VEB-Verlag Technik, 1982