Summer school: Machine Learning and Applied Statistics
Week 2

Almut E. D. Veraart
Room 551, Huxley Building
Department of Mathematics
Imperial College London
180 Queen's Gate, London, SW7 2AZ
E-Mail: a.veraart@imperial.ac.uk

Room 551, Huxley Building
Department of Mathematics
Imperial College London
180 Queen's Gate, London, SW7 2AZ
E-Mail: a.veraart@imperial.ac.uk

Overview

In the second week of the summer school, we will be focussing on **time series analysis** with a particular focus on the following questions:

- What is a time series?

Overview

In the second week of the summer school, we will be focussing on **time series analysis** with a particular focus on the following questions:

- What is a time series?
- Which statistical models can be used for describing time series data?

Overview

In the second week of the summer school, we will be focussing on **time series analysis** with a particular focus on the following questions:

- What is a time series?
- Which statistical models can be used for describing time series data?
- How can such models be estimated and the goodness of the fit be assessed?

Overview

In the second week of the summer school, we will be focussing on **time series analysis** with a particular focus on the following questions:

- What is a time series?
- Which statistical models can be used for describing time series data?
- How can such models be estimated and the goodness of the fit be assessed?
- How can we forecast time series data?

Overview

In the second week of the summer school, we will be focussing on **time series analysis** with a particular focus on the following questions:

- What is a time series?
- Which statistical models can be used for describing time series data?
- How can such models be estimated and the goodness of the fit be assessed?
- How can we forecast time series data?
- How can we estimate risk measures such as *Value at Risk (VaR)* and *Expected shortfall (ES)* for (financial) time series?

Overview

In the second week of the summer school, we will be focussing on **time series analysis** with a particular focus on the following questions:

- What is a time series?
- Which statistical models can be used for describing time series data?
- How can such models be estimated and the goodness of the fit be assessed?
- How can we forecast time series data?
- How can we estimate risk measures such as *Value at Risk (VaR)* and *Expected shortfall (ES)* for (financial) time series?

Throughout the week, we will be looking at various case studies with a particular focus on financial time series.

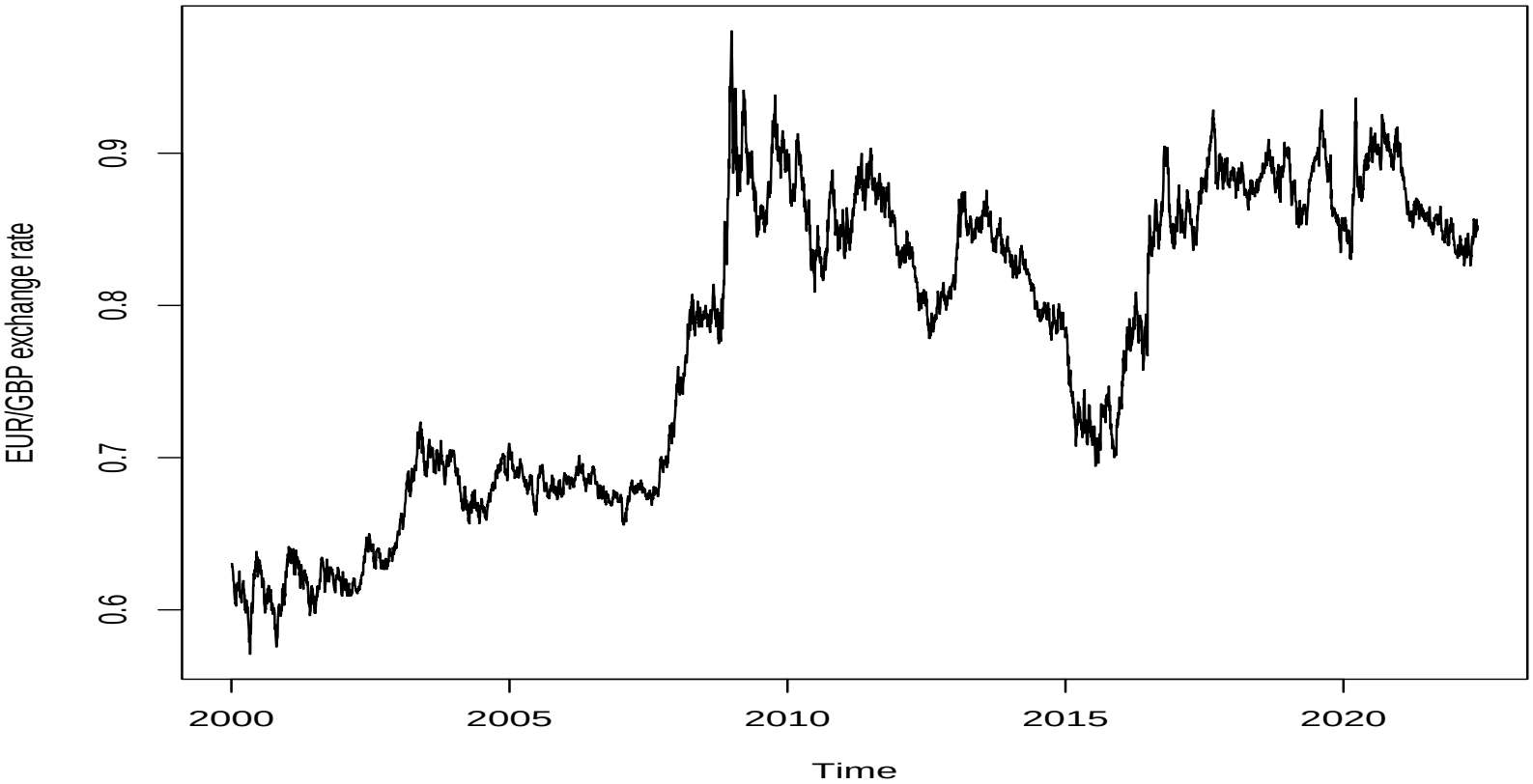
What is a time series?

What is a time series?

Definition

A *time series* is a collection of observations x_t, where t denotes the time point at which the observation is recorded.

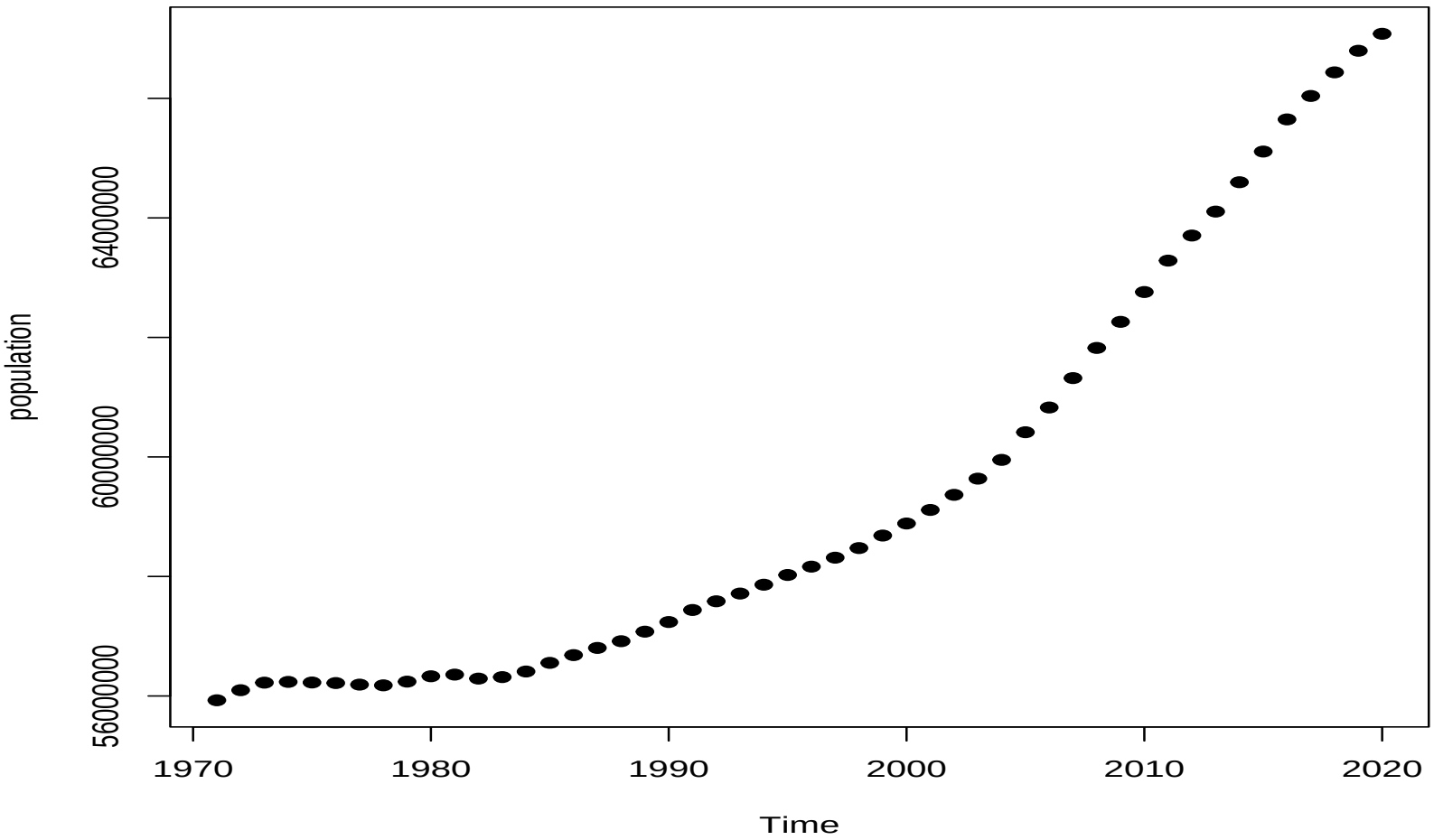
```
#Read in the data:
MyData <- read.csv("EUR-GBP-Rates.csv", header =
    TRUE, sep = ",", dec = ".",fileEncoding="UTF-8-
    BOM")

#View the data
View(MyData)

#Print the first part of the data
head(MyData)

#Print the last part of the data
tail(MyData)

#Convert the dates into the correct date format
MyDates <- as.Date(MyData$Date, "%d/%m/%Y")

#Plot the data
```



Exercise

Describe the time series. Can you think of any events which might have influenced the exchange rate?

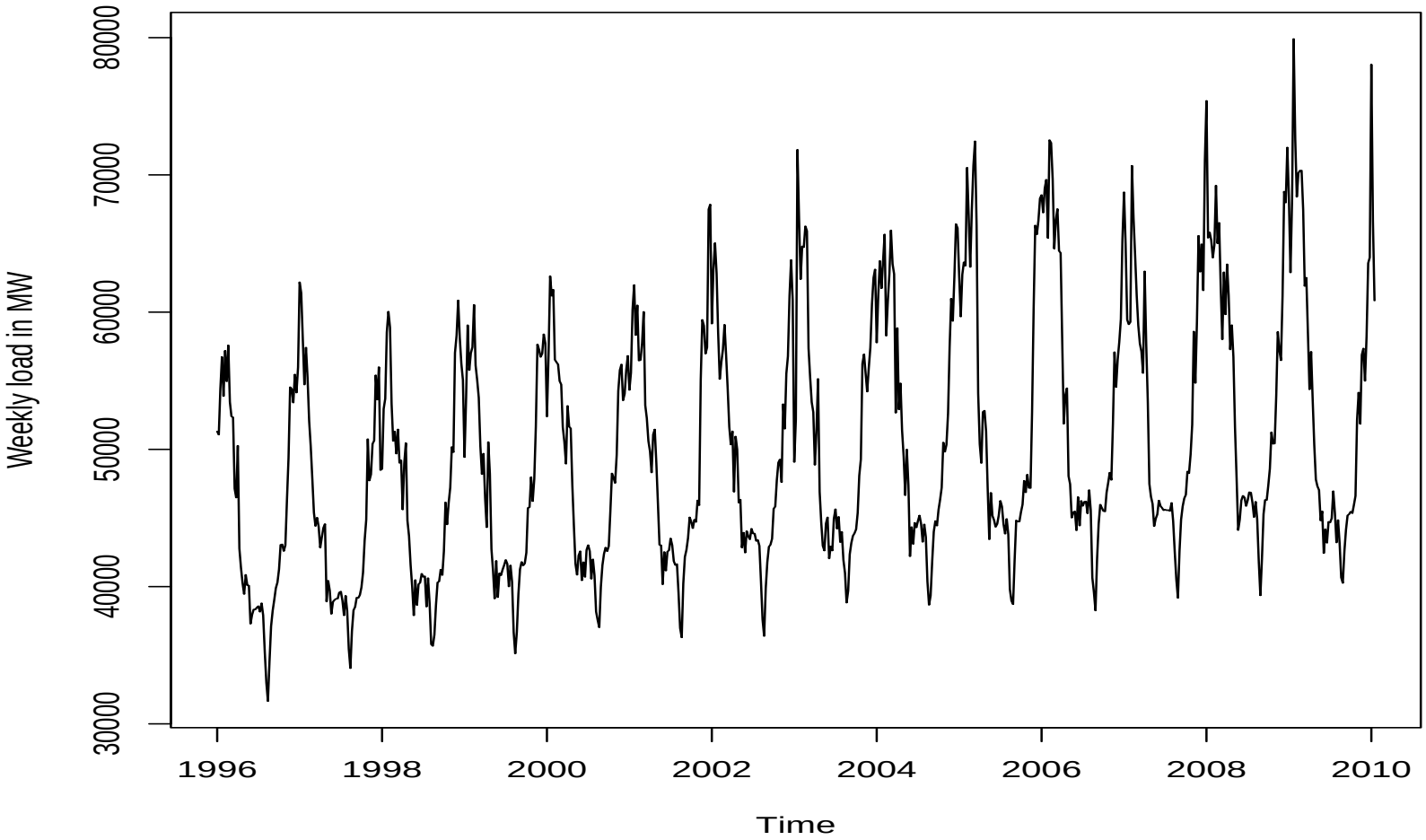
```
1  #Read in the data (ignoring the first seven rows
       which include the data description)
2  MyData <- read.csv("UK-Population.csv", header =
       TRUE, sep = ",", dec=".", skip=7)
3
4  #View the data
5  View(MyData)
6
7  #Print the first part of the data
8  head(MyData)
9
10 #Print the last part of the data
11 tail(MyData)
12
13 #Briefly check the structure of the data
14 str(MyData) #outcome: a data.frame containing
       integers
15
```



Exercise

Describe the time series. Can you spot any trend in the data?

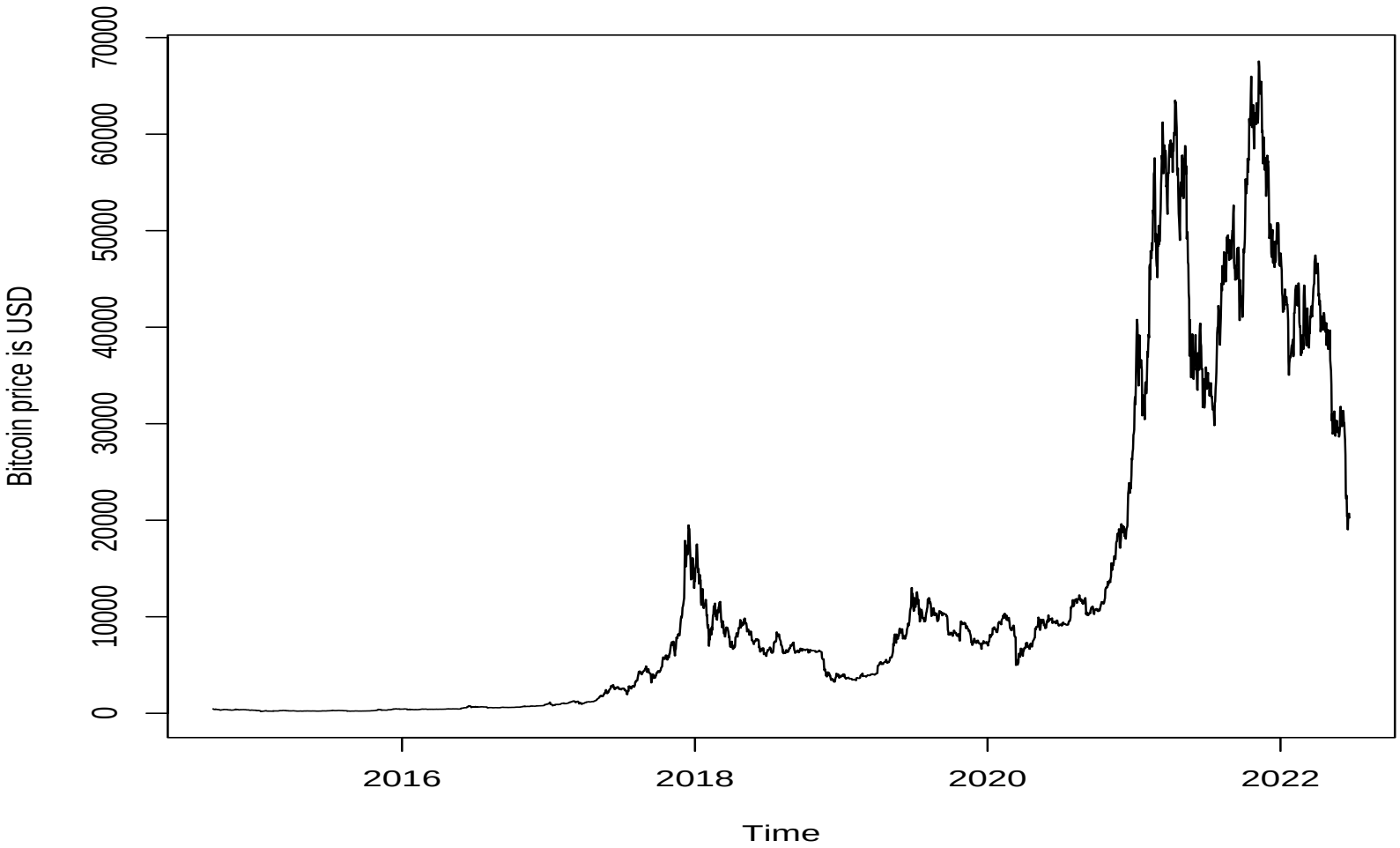
```
1  #Install the package opera (this only needs to be
       done once!)
2  #install.packages("opera")
3  #Load the package opera
4  library(opera)
5
6  #Load the data set
7  attach(electric_load)
8
9  #Convert the data into a time series using the
       function ts
0  LoadTS <-ts(Load,start=1996, frequency =52)
1
2  #Check the length of the time series
3  length(LoadTS)
4
5  #Plot the data
6  plot(LoadTS,type ="l", ylab="Weekly load in MW")
```



Exercise

Describe the time series. Can you spot any seasonality in the data?

```
#Read in the data:
MyData <- read.csv("BTC-USD.csv", header = TRUE, sep
    = ",", dec = ".")

#View the data
View(MyData)

#Print the first part of the data
head(MyData)

#Print the last part of the data
tail(MyData)

#Convert the dates into the correct date format
MyDates <- as.Date(MyData$Date, "%d/%m/%Y")

#Plot the data
plot(MyDates,MyData$Close,type="l", xlab="Time",ylab
```



Exercise

Describe the time series. Do you think it is likely that such a time series can be described by a stationary stochastic process?

Break :-)

Now we want to formalise the concept of a time series and explain that a time series can be viewed as a realisation of a stochastic process. For this we need to recall the concept of a *random variable* and define what we mean by a *stochastic process*.

The sample space

We call Ω the *sample space*, i.e. the set of all possible outcomes of an experiment. The elements of Ω are typically denoted by ω and called *sample points*.

The sample space

We call Ω the *sample space*, i.e. the set of all possible outcomes of an experiment. The elements of Ω are typically denoted by ω and called *sample points*. Let's recap some concepts from set theory:

- Subsets of Ω are collections of elements of Ω and called *events*. Notation: A is a subset of Ω can be written as $A \subseteq \Omega$ meaning that every element of A is also an element of Ω.

The sample space

We call Ω the *sample space*, i.e. the set of all possible outcomes of an experiment. The elements of Ω are typically denoted by ω and called *sample points*. Let's recap some concepts from set theory:

- Subsets of Ω are collections of elements of Ω and called *events*. Notation: A is a subset of Ω can be written as $A \subseteq \Omega$ meaning that every element of A is also an element of Ω.
- We write $\omega \in A$ if the element ω is a member of A and $\omega \notin A$ if the element ω is not a member of A.

The sample space

We call Ω the *sample space*, i.e. the set of all possible outcomes of an experiment. The elements of Ω are typically denoted by ω and called *sample points*. Let's recap some concepts from set theory:

- Subsets of Ω are collections of elements of Ω and called *events*. Notation: A is a subset of Ω can be written as $A \subseteq \Omega$ meaning that every element of A is also an element of Ω.
- We write $\omega \in A$ if the element ω is a member of A and $\omega \notin A$ if the element ω is not a member of A.
- We denote the empty set by $\emptyset$. Note that the empty set contains no points, i.e. $\omega \notin \emptyset$ for all $\omega \in \Omega$.

The sample space

We call Ω the *sample space*, i.e. the set of all possible outcomes of an experiment. The elements of Ω are typically denoted by ω and called *sample points*. Let's recap some concepts from set theory:

- Subsets of Ω are collections of elements of Ω and called *events*. Notation: A is a subset of Ω can be written as $A \subseteq \Omega$ meaning that every element of A is also an element of Ω.
- We write $\omega \in A$ if the element ω is a member of A and $\omega \notin A$ if the element ω is not a member of A.
- We denote the empty set by $\emptyset$. Note that the empty set contains no points, i.e. $\omega \notin \emptyset$ for all $\omega \in \Omega$.
- Every subset A of the sample space Ω satisfies $\emptyset \subseteq A \subseteq \Omega$.

Notation from basic set theory

Definition

$\mathcal{F}$ is a *set of events* which we are allowed to consider (technically: a σ-algebra),

Notation from basic set theory

Definition

$\mathcal{F}$ is a *set of events* which we are allowed to consider (technically: a σ-algebra),

Suppose that $A, B \subset \Omega$ are events (i.e. $A, B \in \mathcal{F}$), then

- the union $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$ is the event that at least one of A and B occurs (this is the *inclusive* "or"),

Notation from basic set theory

Definition

$\mathcal{F}$ is a *set of events* which we are allowed to consider (technically: a σ-algebra),

Suppose that $A, B \subset \Omega$ are events (i.e. $A, B \in \mathcal{F}$), then

- the union $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$ is the event that at least one of A and B occurs (this is the *inclusive* "or"),
- the intersection $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$ is the event that both A and B occur,

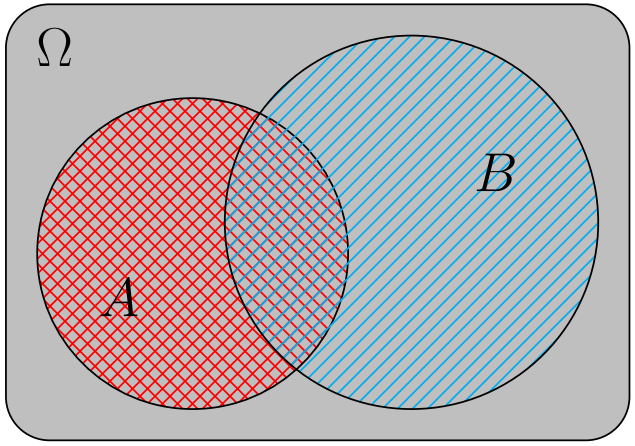
Notation from basic set theory

Definition

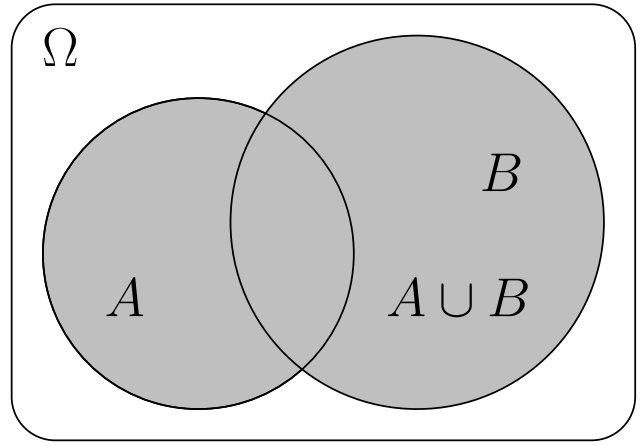
$\mathcal{F}$ is a *set of events* which we are allowed to consider (technically: a σ-algebra),

Suppose that $A, B \subset \Omega$ are events (i.e. $A, B \in \mathcal{F}$), then

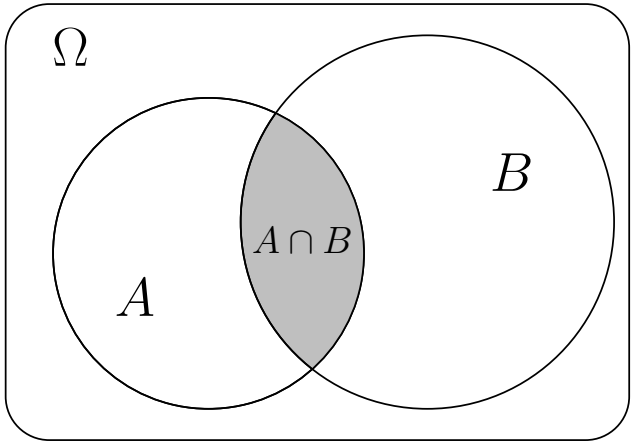
- the union $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$ is the event that at least one of A and B occurs (this is the *inclusive* "or"),
- the intersection $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$ is the event that both A and B occur,
- the complement $A^c = \{\omega \in \Omega : \omega \notin A\}$ is the event that occurs if and only if A does not occur.



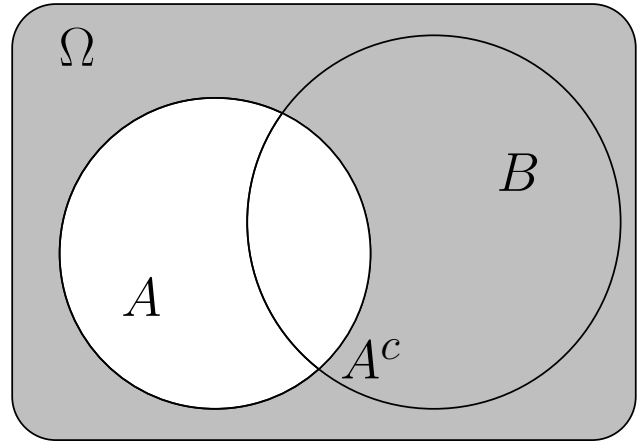
(a) $A, B \subseteq \Omega$



(b) $A \cup B$



(c) $A \cap B$



(d) A^c

Definition of probability measure

Definition

The probability measure P is a function from $\mathcal{F}$ into the real numbers $\mathbb{R}$ which satisfies three conditions:

Definition of probability measure

Definition

The probability measure P is a function from $\mathcal{F}$ into the real numbers $\mathbb{R}$ which satisfies three conditions:

(i) $0 \leq \mathrm{P}(A) \leq 1$ for all events $A \in \mathcal{F}$,

(ii) $\mathrm{P}(\Omega) = 1$ and $\mathrm{P}(\emptyset) = 0$ (where $\emptyset$ denotes the empty set),

(iii) For any sequence of disjoint events $A_1, A_2, A_3, \cdots \in \mathcal{F}$ we have

$$\mathrm{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$

[Note that by "disjoint events" we mean that $A_i \cap A_j = \emptyset$ for all $i \neq j$.]

Example: Flipping a fair coin

We start with the classical example of flipping a fair coin. We write H for heads and T for tail.

Example: Flipping a fair coin

We start with the classical example of flipping a fair coin. We write H for heads and T for tail.

- The sample space is given by $\Omega = \{H, T\}$.

Example: Flipping a fair coin

We start with the classical example of flipping a fair coin. We write H for heads and T for tail.

- The sample space is given by $\Omega = \{H, T\}$.
- The set of events can be taken as $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$ which is the collection of all subsets of Ω.

Example: Flipping a fair coin

We start with the classical example of flipping a fair coin. We write H for heads and T for tail.

- The sample space is given by $\Omega = \{H, T\}$.
- The set of events can be taken as $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$ which is the collection of all subsets of Ω.
- Since we are considering a "fair" coin, we have that $\mathrm{P}(\{H\}) = \mathrm{P}(\{T\}) = \frac{1}{2}$, where we typically shorten the notation to:

$$\mathrm{P}(H) = \mathrm{P}(T) = \frac{1}{2}.$$

Moreover, we have $\mathrm{P}(\emptyset) = 0$ and $\mathrm{P}(\Omega) = 1$.

Random variable

Definition

Let Ω be a sample space. A *random variable* (r.v.) is a function from Ω into the real numbers $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$.

Random variable

Definition

Let Ω be a sample space. A *random variable* (r.v.) is a function from Ω into the real numbers $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$.

Note that

- Despite the name, a random variable is a *function* and not a variable.

Random variable

Definition

Let Ω be a sample space. A *random variable* (r.v.) is a function from Ω into the real numbers $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$.

Note that

- Despite the name, a random variable is a *function* and not a variable.
- We typically use capital letters such as X, Y, Z to denote random variables.

Random variable

Definition

Let Ω be a sample space. A *random variable* (r.v.) is a function from Ω into the real numbers $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$.

Note that

- Despite the name, a random variable is a *function* and not a variable.
- We typically use capital letters such as X, Y, Z to denote random variables.
- The value of the random variable X at the sample point ω is given by $X(\omega)$ and is called a *realisation* of X.

Random variable

Definition

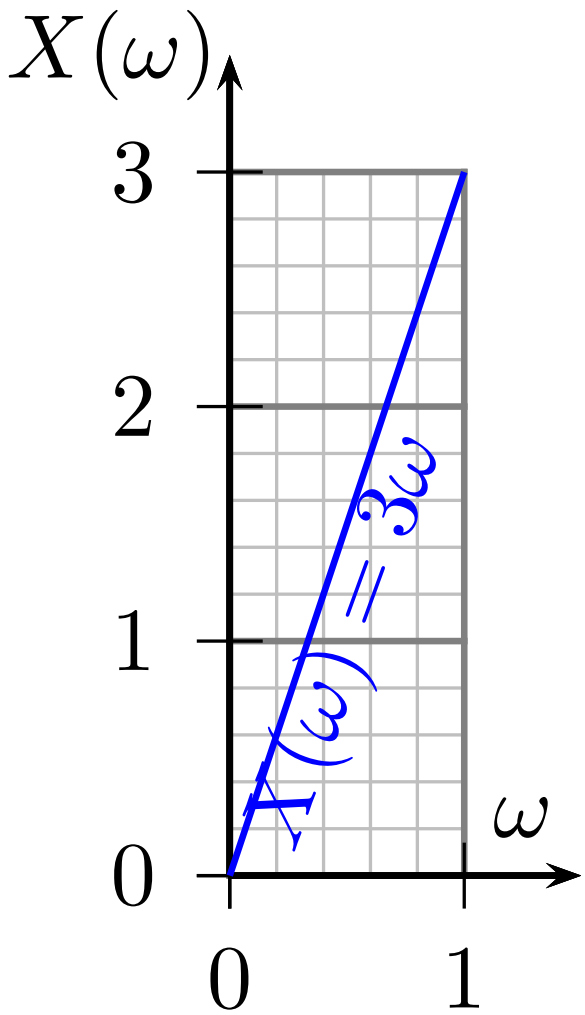
Let Ω be a sample space. A *random variable* (r.v.) is a function from Ω into the real numbers $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$.

Note that

- Despite the name, a random variable is a *function* and not a variable.
- We typically use capital letters such as X, Y, Z to denote random variables.
- The value of the random variable X at the sample point ω is given by $X(\omega)$ and is called a *realisation* of X.
- The randomness stems from $\omega \in \Omega$ (we don't know which outcome ω appears in the random experiment), the mapping itself given by X is deterministic.

Example

Let $\Omega = [0, 1]$ and define the random variable $X : \Omega \to \mathbb{R}$ by $X(\omega) = 3\omega$.



Example cont'd

Suppose that the probability measure P has the following property: For $0 \leq a \leq b \leq 1$, we have that $\mathrm{P}([a, b]) = b - a$, i.e. the probability of an event given by an interval is given by the *length of that interval*.

Example cont'd

Suppose that the probability measure P has the following property: For $0 \leq a \leq b \leq 1$, we have that $\mathrm{P}([a, b]) = b - a$, i.e. the probability of an event given by an interval is given by the *length of that interval*. We want to find $\mathrm{P}(X \leq x)$ for $x \in \mathbb{R}$.

Example cont'd

Suppose that the probability measure P has the following property: For $0 \leq a \leq b \leq 1$, we have that $\mathrm{P}([a,b]) = b - a$, i.e. the probability of an event given by an interval is given by the *length of that interval*. We want to find $\mathrm{P}(X \leq x)$ for $x \in \mathbb{R}$. Note that, for $x \in [0,3]$,

$$\begin{aligned}\{X \leq x\} &= \{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : 3\omega \leq x\} \\ &= \{\omega \in \Omega : \omega \leq x/3\} = [0, x/3].\end{aligned}$$

Example cont'd

Suppose that the probability measure P has the following property: For $0 \leq a \leq b \leq 1$, we have that $\mathrm{P}([a, b]) = b - a$, i.e. the probability of an event given by an interval is given by the *length of that interval*. We want to find $\mathrm{P}(X \leq x)$ for $x \in \mathbb{R}$. Note that, for $x \in [0, 3]$,

$$\begin{aligned} \{X \leq x\} &= \{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : 3\omega \leq x\} \\ &= \{\omega \in \Omega : \omega \leq x/3\} = [0, x/3]. \end{aligned}$$

Then $\mathrm{P}(X \leq x) = x/3$ for $x \in [0, 3]$,

Example cont'd

Suppose that the probability measure P has the following property: For $0 \leq a \leq b \leq 1$, we have that $\mathrm{P}([a,b]) = b - a$, i.e. the probability of an event given by an interval is given by the *length of that interval*. We want to find $\mathrm{P}(X \leq x)$ for $x \in \mathbb{R}$. Note that, for $x \in [0,3]$,

$$\begin{aligned}\{X \leq x\} &= \{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : 3\omega \leq x\} \\ &= \{\omega \in \Omega : \omega \leq x/3\} = [0, x/3].\end{aligned}$$

Then $\mathrm{P}(X \leq x) = x/3$ for $x \in [0,3]$, and $\mathrm{P}(X \leq x) = \mathrm{P}(\emptyset) = 0$ for $x < 0$,

Example cont'd

Suppose that the probability measure P has the following property: For $0 \leq a \leq b \leq 1$, we have that $\mathrm{P}([a,b]) = b - a$, i.e. the probability of an event given by an interval is given by the *length of that interval*. We want to find $\mathrm{P}(X \leq x)$ for $x \in \mathbb{R}$. Note that, for $x \in [0,3]$,

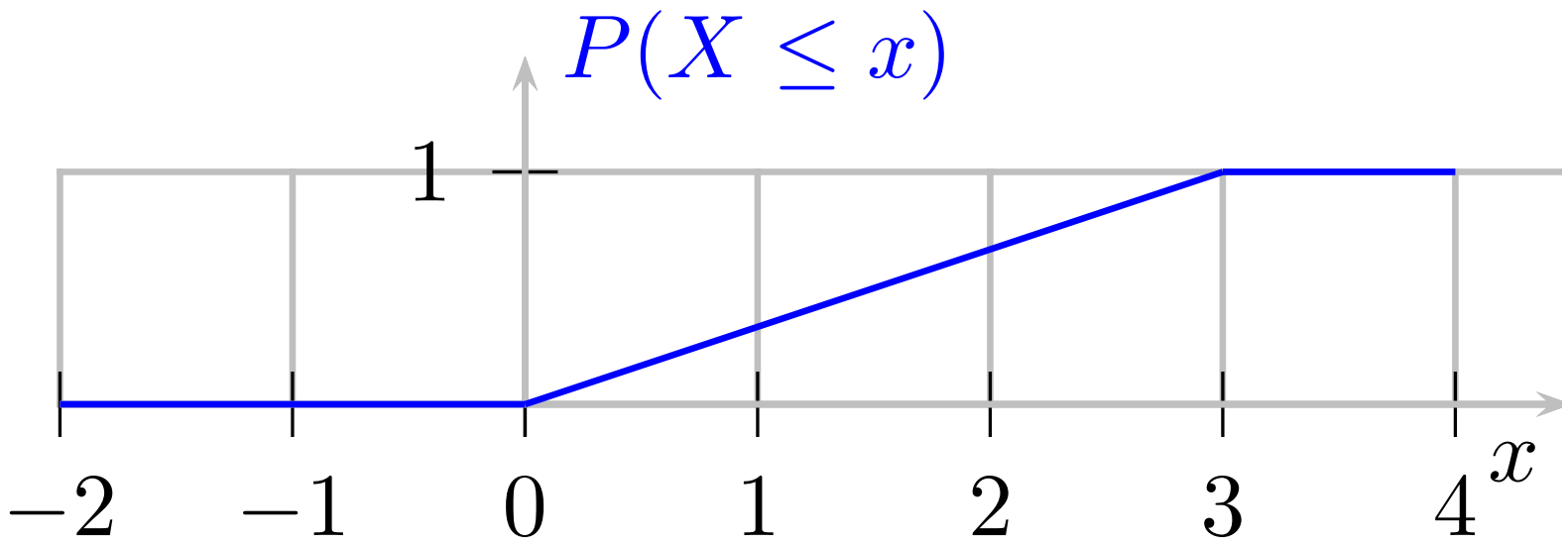
$$\begin{aligned}\{X \leq x\} &= \{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : 3\omega \leq x\} \\ &= \{\omega \in \Omega : \omega \leq x/3\} = [0, x/3].\end{aligned}$$

Then $\mathrm{P}(X \leq x) = x/3$ for $x \in [0,3]$, and $\mathrm{P}(X \leq x) = \mathrm{P}(\emptyset) = 0$ for $x < 0$, and $\mathrm{P}(X \leq x) = \mathrm{P}(\Omega) = 1$ for $x > 3$.

Example cont'd

We plot the function

$$\mathrm{P}(X \leq x) = \begin{cases} 0, & \text{for } x < 0, \\ \frac{x}{3}, & \text{for } x \in [0, 3], \\ 1, & \text{for } x > 3 \end{cases}$$



Break :-)

Probability distribution

Definition

Let X be a random variable on a sample space Ω. The *probability distribution* of X is the collection of probabilities $\mathrm{P}(X \in B)$ for sets B of real numbers.

Discrete random variable

Definition

A random variable is said to be *discrete* if there exists a finite or countably infinite set $\{k_1, k_2, \dots\}$ of real numbers such that

$$\sum_i \mathrm{P}(X = k_i) = 1,$$

where the sum ranges over all points in $\{k_1, k_2, \dots\}$.

Discrete random variable

Definition

A random variable is said to be *discrete* if there exists a finite or countably infinite set $\{k_1, k_2, \dots\}$ of real numbers such that

$$\sum_i \mathrm{P}(X = k_i) = 1,$$

where the sum ranges over all points in $\{k_1, k_2, \dots\}$.

Definition

The probability mass function (p.m.f.) of a discrete random variable X is the function p_X defined by $p_X(k) = \mathrm{P}(X = k)$ for all possible values k of X (for which $\mathrm{P}(X = k) > 0$).

Often we drop the subscript X and write p for the p.m.f. of X.

Continuous random variable

Definition

Let X be a random variable. If there is a function f_X such that

$$\mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)dt,$$

for all $x \in \mathbb{R}$, then f_X is called the *probability density function* (p.d.f.) of X and X is called a *continuous random variable*.

Continuous random variable

Definition

Let X be a random variable. If there is a function f_X such that

$$\mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)dt,$$

for all $x \in \mathbb{R}$, then f_X is called the *probability density function* (p.d.f.) of X and X is called a *continuous random variable*.

- Often we drop the subscript X and write f for the p.d.f. of X.

Continuous random variable

Definition

Let X be a random variable. If there is a function f_X such that

$$\mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)dt,$$

for all $x \in \mathbb{R}$, then f_X is called the *probability density function* (p.d.f.) of X and X is called a *continuous random variable*.

- Often we drop the subscript X and write f for the p.d.f. of X.
- Note that a probability density function must satisfy $f(t) \geq 0$ for all $t \in \mathbb{R}$ and $\int_{-\infty}^{\infty} f(t)dt = 1$.

Continuous random variable

Definition

Let X be a random variable. If there is a function f_X such that

$$\mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)dt,$$

for all $x \in \mathbb{R}$, then f_X is called the *probability density function* (p.d.f.) of X and X is called a *continuous random variable*.

- Often we drop the subscript X and write f for the p.d.f. of X.
- Note that a probability density function must satisfy $f(t) \geq 0$ for all $t \in \mathbb{R}$ and $\int_{-\infty}^{\infty} f(t)dt = 1$.
- For a continuous random variable X,we always have that $P(X = x) = 0$ for all $x \in \mathbb{R}$ and $\mathrm{P}(a < X \leq b) = \mathrm{P}(a < X < b) = \int_a^b f(t)dt$.

Cumulative distribution function

We note that p.m.f.s are only defined for discrete random variables and p.d.f.s only for continuous random variables. However, the so-called cumulative distribution function can be used to describe the distribution of *any* random variable.

Cumulative distribution function

We note that p.m.f.s are only defined for discrete random variables and p.d.f.s only for continuous random variables. However, the so-called cumulative distribution function can be used to describe the distribution of *any* random variable.

Definition

Let X be a random variable on the sample space Ω. Its *cumulative distribution function* (c.d.f.) is defined as

$$F_X(x) = \mathrm{P}(X \leq x), \qquad \text{for all } x \in \mathbb{R}.$$

Often we drop the subscript X and write F for the c.d.f. of X.

Cumulative distribution function

We note that p.m.f.s are only defined for discrete random variables and p.d.f.s only for continuous random variables. However, the so-called cumulative distribution function can be used to describe the distribution of *any* random variable.

Definition

Let X be a random variable on the sample space Ω. Its *cumulative distribution function* (c.d.f.) is defined as

$$F_X(x) = \mathrm{P}(X \leq x), \qquad \text{for all } x \in \mathbb{R}.$$

Often we drop the subscript X and write F for the c.d.f. of X. Note that

$$\mathrm{P}(a < X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) = F_X(b) - F_X(a).$$

C.d.f. for disrete and continuous r.v.s

If X is a discrete random variable, then

$$F(x) = \mathrm{P}(X \leq x) = \sum_{k:k \leq x} \mathrm{P}(X = k).$$

C.d.f. for disrete and continuous r.v.s

If X is a discrete random variable, then

$$F(x) = \mathrm{P}(X \leq x) = \sum_{k:k \leq x} \mathrm{P}(X = k).$$

If X is a continuous random variable, then

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Expectation

Definition

For a discrete random variable X, the *expectation* or *mean* is given by

$$\mathrm{E}(X) = \sum_k k\mathrm{P}(X = k),$$

where we sum over all possible values k of X.

Expectation

Definition

For a discrete random variable X, the *expectation* or *mean* is given by

$$\mathrm{E}(X) = \sum_k k\mathrm{P}(X = k),$$

where we sum over all possible values k of X. For a continuous random variable, the *expectation/mean* is given by

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Expectation

Suppose that X is a r.v.s and let g be a real-valued function defined on the range of X.

Expectation

Suppose that X is a r.v.s and let g be a real-valued function defined on the range of X. If X is discrete, then

$$\mathrm{E}(g(X)) = \sum_k g(k)\mathrm{P}(X = k),$$

whereas if X is continuous, then

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

nth moment

Example

Choose $g(x) = x^n$ for $n \in \mathbb{N}$. Then, we obtain for discrete X:

$$\mathrm{E}(X^n) = \sum_k k^n \mathrm{P}(X = k),$$

and for continuous X:

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

which we call the nth moment of X (which might not always exist!).

Variance

Definition

The variance of a random variable X with mean $\mu = \mathrm{E}(X)$ is defined as

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu)^2].$$

Often we write $\sigma^2 = \mathrm{Var}(X)$.

Variance

Definition

The variance of a random variable X with mean $\mu = \mathrm{E}(X)$ is defined as

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu)^2].$$

Often we write $\sigma^2 = \mathrm{Var}(X)$.

We note that

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

Properties of expectation and variance

We have the following important properties: For a random variable X with finite mean and variance and real numbers a, b, we have

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b, \qquad \mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

Bernoulli random variable

A so-called *Bernoulli* random variable records the outcome of a single experiment with two possible outcomes which are denoted by $\{0, 1\}$. We call $p \in [0, 1]$ the *success probability*.

Bernoulli random variable

A so-called *Bernoulli* random variable records the outcome of a single experiment with two possible outcomes which are denoted by $\{0, 1\}$. We call $p \in [0, 1]$ the *success probability*. Then

$$\mathrm{P}(X = 1) = p, \qquad \mathrm{P}(X = 0) = 1 - p.$$

Bernoulli random variable

A so-called *Bernoulli* random variable records the outcome of a single experiment with two possible outcomes which are denoted by $\{0, 1\}$. We call $p \in [0, 1]$ the *success probability*. Then

$$\mathrm{P}(X = 1) = p, \qquad \mathrm{P}(X = 0) = 1 - p.$$

We typically write $X \sim \mathrm{Ber}(p)$ for a Bernoulli random variable with parameter p. We can now derive the c.d.f. of a Bernoulli random variable and obtain

Bernoulli random variable

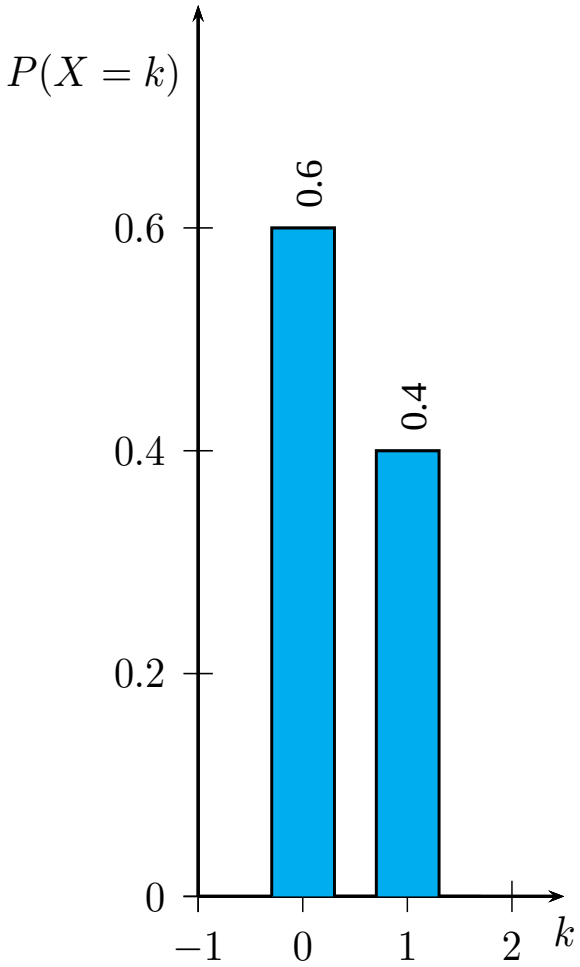
A so-called *Bernoulli* random variable records the outcome of a single experiment with two possible outcomes which are denoted by $\{0, 1\}$. We call $p \in [0, 1]$ the *success probability*. Then

$$\mathrm{P}(X = 1) = p, \qquad \mathrm{P}(X = 0) = 1 - p.$$

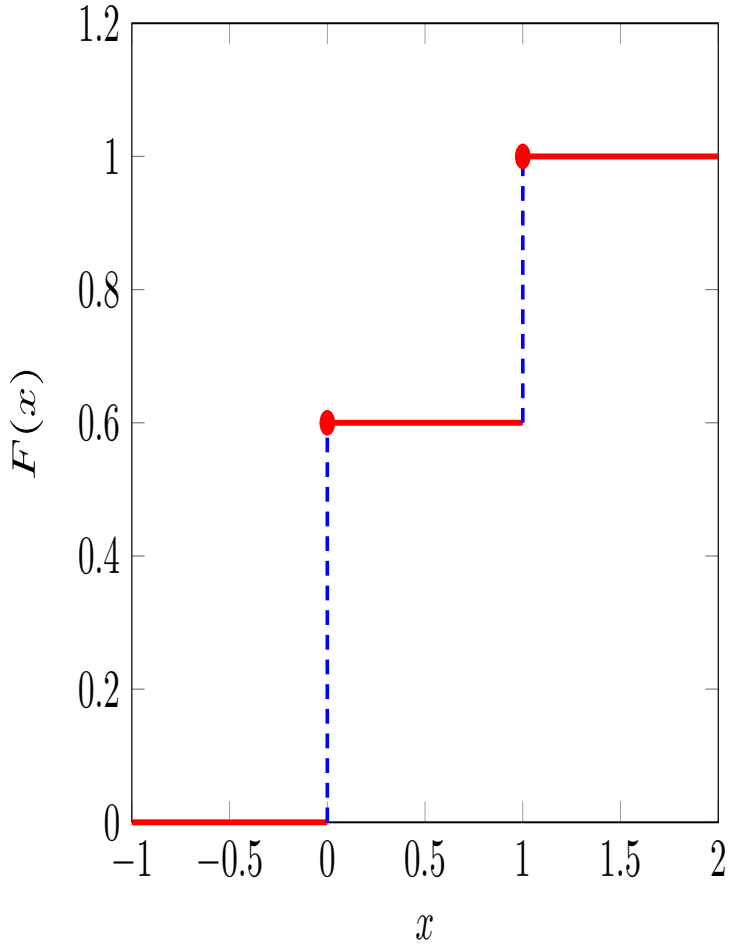
We typically write $X \sim \mathrm{Ber}(p)$ for a Bernoulli random variable with parameter p. We can now derive the c.d.f. of a Bernoulli random variable and obtain

$$F(x) = \mathrm{P}(X \leq x) = \sum_{k \leq x} \mathrm{P}(X = k) = \begin{cases} 0, & \text{for } x < 0, \\ 1 - p, & \text{for } x \in [0, 1), \\ 1, & \text{for } x \geq 1. \end{cases}$$

Bernoulli random variable



(e) P.m.f. of $X \sim \text{Ber}(0.4)$



(f) C.d.f. of $X \sim \text{Ber}(0.4)$

Bernoulli random variable

We can now compute the mean and the variance as follows:

$$\mathrm{E}(X) = \sum_k k\mathrm{P}(X = k) = 0\mathrm{P}(X = 0) + 1\mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

Bernoulli random variable

We can now compute the mean and the variance as follows:

$$\mathrm{E}(X) = \sum_k k\mathrm{P}(X = k) = 0\mathrm{P}(X = 0) + 1\mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

The second moment is given by

$$\mathrm{E}(X^2) = \sum_k k^2\mathrm{P}(X = k) = 0^2\mathrm{P}(X = 0) + 1^2\mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

Hence

Bernoulli random variable

We can now compute the mean and the variance as follows:

$$\mathrm{E}(X) = \sum_k k\mathrm{P}(X = k) = 0\mathrm{P}(X = 0) + 1\mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

The second moment is given by

$$\mathrm{E}(X^2) = \sum_k k^2\mathrm{P}(X = k) = 0^2\mathrm{P}(X = 0) + 1^2\mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

Hence

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = p - p^2 = p(1 - p).$$

Break :-)

Searching for a time series model...

- Recall: A *time series* is a collection of observations x_t for $t = 1, \ldots, n$.

Searching for a time series model...

- Recall: A *time series* is a collection of observations x_t for $t = 1, \ldots, n$.
- Definition: A stochastic process $X = (X_t)_{t \in \{1,\ldots,n\}}$ is a collection of random variables X_t.

Searching for a time series model...

- Recall: A *time series* is a collection of observations x_t for $t = 1, \ldots, n$.
- Definition: A stochastic process $X = (X_t)_{t \in \{1,\ldots,n\}}$ is a collection of random variables X_t.
- For fixed $\omega \in \Omega$, we call $(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$ the *realisation/sample path* of the stochastic process X or the *time series*.

Searching for a time series model...

- Recall: A *time series* is a collection of observations x_t for $t = 1, \ldots, n$.
- Definition: A stochastic process $X = (X_t)_{t \in \{1,\ldots,n\}}$ is a collection of random variables X_t.
- For fixed $\omega \in \Omega$, we call $(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$ the *realisation/sample path* of the stochastic process X or the *time series*.
- **Goal:** Find a time series model, i.e. a probabilistic description of the random vector $(X_1, \ldots, X_n)$ such that it is reasonable to assume that our (observed) time series $(x_1, \ldots, x_n)$ is a realisation of the random vector $(X_1, \ldots, X_n)$.

Random vector

Definition

Let $n \in \mathbb{N}$. An n-dimensional random vector is a column vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = (X_1, X_2, \ldots, X_n)^\top,$$

where X_i is a random variable for $i = 1, \ldots, n$.

Note that the sign $^\top$ stands for the *transpose* of a vector [or a matrix] and it means that a row (column) vector is changed into a column (row) vector where the elements remain in the same order.

Joint distribution

Definition

The *joint/multivariate cumulative distribution function* of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ is defined as

$$F(x_1, \ldots, x_n) = \mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

for all real vectors $\mathbf{x} = (x_1, \ldots, x_n)^\top$.

Joint distribution

Definition

The *joint/multivariate cumulative distribution function* of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ is defined as

$$F(x_1, \ldots, x_n) = \mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

for all real vectors $\mathbf{x} = (x_1, \ldots, x_n)^\top$.

Note that you can find the joint distribution of a subcollection of random variables by setting the corresponding x value to ∞. More precisely, we have

$$F_{X_1}(x_1) = \mathrm{P}(X_1 \leq x_1) = F(x_1, \infty, \ldots, \infty),$$
$$F_{(X_1, X_3)}(x_1, x_3) = \mathrm{P}(X_1 \leq x_1, X_3 \leq x_3) = F(x_1, \infty, x_3, \infty, \ldots, \infty).$$

Multivariate probability density

Exactly as in the univariate setting, we say that a random vector with c.d.f. F is *continuous* if F has a density function f which satisfies

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(t_1, \ldots, t_n) dt_1 \cdots dt_n.$$

Multivariate probability density

Exactly as in the univariate setting, we say that a random vector with c.d.f. F is *continuous* if F has a density function f which satisfies

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(t_1, \ldots, t_n) dt_1 \cdots dt_n.$$

In that case, the density can be derived by differentiating F with respect to all variables (i.e. by taking n *partial* derivatives)

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}.$$

Multivariate probability density

Exactly as in the univariate setting, we say that a random vector with c.d.f. F is *continuous* if F has a density function f which satisfies

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(t_1, \ldots, t_n) dt_1 \cdots dt_n.$$

In that case, the density can be derived by differentiating F with respect to all variables (i.e. by taking n *partial* derivatives)

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}.$$

Also, we say that a random vector is *discrete* if there exist real-valued vectors $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ and a probability mass function $p(\mathbf{x}^{(i)}) = \mathrm{P}(\mathbf{X} = \mathbf{x}^{(i)}) = \mathrm{P}(X_1 = x_1^{(i)}, \ldots, X_n = x_n^{(i)})$ such that

$$\sum_{i=0}^{\infty} p(\mathbf{x}^{(i)}) = 1.$$

Independence of random variables

Definition

The random variables $X_1, \ldots, X_n$ are said to be *independent* if their joint cumulative distribution function factorises as follows:

$$\mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathrm{P}(X_1 \leq x_1) \cdots \mathrm{P}(X_n \leq x_n),$$

i.e.,

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n),$$

for all real numbers $x_1, \ldots, x_n$.

Independence of random variables

Note that in the discrete case, independence is equivalent to the factorisation of the joint probability mass function,

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n),$$

and in the continuous case to the factorisation of the joint probability density function

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Expectation of functions of random vectors

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a deterministic function, then we define the expectation of function g of a random vector $\mathbf{X}$ as

$$\begin{aligned}\mathrm{E}(g(\mathbf{X})) &= \int g(x_1, \ldots, x_n) dF(x_1, \ldots, x_n) \\ &= \begin{cases} \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n, & \text{if } \mathbf{X} \text{ is continuo} \\ \sum_{i_1} \cdots \sum_{i_n} g(x_{i_1}, \ldots, x_{i_n}) p(x_{i_1}, \ldots, x_{i_n}), & \text{if } \mathbf{X} \text{ is discrete,} \end{cases}\end{aligned}$$

provided these integrals/sums are finite.

Expectation of functions of random vectors

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a deterministic function, then we define the expectation of function g of a random vector $\mathbf{X}$ as

$$\begin{aligned}\mathrm{E}(g(\mathbf{X})) &= \int g(x_1, \ldots, x_n) dF(x_1, \ldots, x_n) \\ &= \begin{cases} \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n, & \text{if } \mathbf{X} \text{ is continuo} \\ \sum_{i_1} \cdots \sum_{i_n} g(x_{i_1}, \ldots, x_{i_n}) p(x_{i_1}, \ldots, x_{i_n}), & \text{if } \mathbf{X} \text{ is discrete,} \end{cases}\end{aligned}$$

provided these integrals/sums are finite.
In particular, for two random variables X, Y, we have (when choosing $g(x, y) = xy$):

$$\mathrm{E}(XY) = \begin{cases} \int \int xyf(x, y) dxdy, & \text{if } (X, Y)^\top \text{ is continuous,} \\ \sum_i \sum_j x_i y_j p(x_i, y_j), & \text{if } (X, Y)^\top \text{ is discrete,} \end{cases}$$

Expectation of functions of random vectors

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a deterministic function, then we define the expectation of function g of a random vector $\mathbf{X}$ as

$$\begin{aligned}\mathrm{E}(g(\mathbf{X})) &= \int g(x_1, \ldots, x_n) dF(x_1, \ldots, x_n) \\ &= \begin{cases} \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n, & \text{if } \mathbf{X} \text{ is continuo} \\ \sum_{i_1} \cdots \sum_{i_n} g(x_{i_1}, \ldots, x_{i_n}) p(x_{i_1}, \ldots, x_{i_n}), & \text{if } \mathbf{X} \text{ is discrete,} \end{cases}\end{aligned}$$

provided these integrals/sums are finite.
In particular, for two random variables X, Y, we have (when choosing $g(x, y) = xy$):

$$\mathrm{E}(XY) = \begin{cases} \int \int xyf(x, y) dxdy, & \text{if } (X, Y)^\top \text{ is continuous,} \\ \sum_i \sum_j x_i y_j p(x_i, y_j), & \text{if } (X, Y)^\top \text{ is discrete,} \end{cases}$$

Exercise: Show that for independent continuous random variables X and Y, we have $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

Break :-)

Covariance between random variables

Definition

Consider two (one-dimensional) random variables X and Y on the same sample space with expectations $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$. The *covariance* of X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$$

if the expectation on the right hand side takes a finite value.

Covariance between random variables

Definition

Consider two (one-dimensional) random variables X and Y on the same sample space with expectations $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$. The *covariance* of X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$$

if the expectation on the right hand side takes a finite value.

When we set $X = Y$, then the covariance simplifies to the variance:

$$\mathrm{Cov}(X, X) = \mathrm{E}[(X - \mu_X)^2] = \mathrm{Var}(X).$$

Covariance between random variables

Definition

Consider two (one-dimensional) random variables X and Y on the same sample space with expectations $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$. The *covariance* of X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$$

if the expectation on the right hand side takes a finite value.

When we set $X = Y$, then the covariance simplifies to the variance:

$$\mathrm{Cov}(X, X) = \mathrm{E}[(X - \mu_X)^2] = \mathrm{Var}(X).$$

Exercise: Use the linearity of the expectation to derive the alternative formula for the covariance:

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

Independent random variables have zero covariance

It is important to note that independent random variables always have zero covariance, but the converse does not hold in general!

Independent random variables have zero covariance

It is important to note that independent random variables always have zero covariance, but the converse does not hold in general!
Consider the case of independent continuous random variables X and Y. Then from Exercise **??** we know that $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y) = \mu_X \mu_Y$.
Hence

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mu_X \mu_Y - \mu_X \mu_Y = 0.$$

The proof of the discrete case works similarly.

Correlation between random variables

Definition

Consider two (one-dimensional) random variables X and Y on the same sample space with expectations $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$. The *correlation* of X and Y is defined as

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}},$$

if $\mathrm{Cov}(X, Y), \mathrm{Var}(X), \mathrm{Var}(Y)$ take finite values.

Correlation between random variables

Definition

Consider two (one-dimensional) random variables X and Y on the same sample space with expectations $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$. The *correlation* of X and Y is defined as

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}},$$

if $\mathrm{Cov}(X, Y), \mathrm{Var}(X), \mathrm{Var}(Y)$ take finite values.

When we set $X = Y$, then we have

$$\mathrm{Cor}(X, X) = \frac{\mathrm{Cov}(X, X)}{\sqrt{(\mathrm{Var}(X))^2}} = \frac{\mathrm{Var}(X)}{\sqrt{(\mathrm{Var}(X))^2}} = 1.$$

Correlation between random variables

One can show that the correlation between two random variables X and Y can only take values in the interval $[-1, 1]$. We call X and Y

- *uncorrelated* if $\mathrm{Cor}(X, Y) = 0$,
- *positive correlated* if $\mathrm{Cor}(X, Y) > 0$,
- *negative correlated* if $\mathrm{Cor}(X, Y) < 0$.

A second order perspective to time series modelling

We think of a stochastic process $X = (X_t)_{t=0,\pm1,\ldots}$ as being stationary if its statistical properties are the same as the ones of the "time-shifted" process $(X_{t+h})_{t=0,\pm1,\ldots}$ for any $h \in \mathbb{Z}$. Rather than considering the joint distribution of the entire stochastic process, we rather use a second order perspective in time series modelling, which means that we focus on the statistical properties related to the first and second moment of the stochastic process X.

A second order perspective to time series modelling

Definition

Consider a stochastic process $X = (X_t)$ with $\mathrm{E}(X_t^2) < \infty$ for all t.

- The *mean function* of X is defined as

$$\mu_X(t) = \mathrm{E}(X_t).$$

- The *covariance function* of X is defined as

$$\gamma_X(s,t) = \mathrm{Cov}(X_s, X_t) = \mathrm{E}[(X_s - \mu_X(s))(X_t - \mu_X(t))], \text{ for all } s, t.$$

A second order perspective to time series modelling

Definition

Consider a stochastic process $X = (X_t)$ with $\mathrm{E}(X_t^2) < \infty$ for all t.

- The *mean function* of X is defined as

$$\mu_X(t) = \mathrm{E}(X_t).$$

- The *covariance function* of X is defined as

$$\gamma_X(s,t) = \mathrm{Cov}(X_s, X_t) = \mathrm{E}[(X_s - \mu_X(s))(X_t - \mu_X(t))], \text{ for all } s, t.$$

Moreover, X is called *(weakly) stationary*[a] if (i) $\mu_X(t)$ is independent of t, and (ii) $\gamma_X(t+h, t)$ is independent of t for each h.

[a] There is a stronger concept of stationarity, sometimes referred to as *strict* stationarity, which we will not study in this course in more detail. However, it is important to note that the type of stationarity we have defined is sometimes called *weak/covariance/second order stationarity* to distinguish it more clearly from strict

Autocovariance and autocorrelation function

Definition

Consider a stationary stochastic process $X = (X_t)$.

- We define the *autocovariance function* (ACVF) of X at lag h as

$$\gamma_X(h) = \text{Cov}(X_{t+h}, X_t),$$

- and the *autocorrelation function* (ACF) of X at lag h as

$$\rho_X(h) = \text{Cor}(X_{t+h}, X_t) = \frac{\gamma_X(h)}{\gamma_X(0)}.$$

Sample mean, autocovariance and autocorrelation

Suppose we have a time series with observations denoted by $x_1, \ldots, x_n$.

- The sample mean is given by $\overline{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$.

Sample mean, autocovariance and autocorrelation

Suppose we have a time series with observations denoted by $x_1, \ldots, x_n$.

- The sample mean is given by $\overline{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$.
- We define the *sample variance* of our time series as

$$\hat{\gamma}(0) := \frac{1}{n} \sum_{t=1}^{n} (x_t - \overline{x})^2,$$

Sample mean, autocovariance and autocorrelation

Suppose we have a time series with observations denoted by $x_1, \ldots, x_n$.

- The sample mean is given by $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$.
- We define the *sample variance* of our time series as

$$\hat{\gamma}(0) := \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^2,$$

and, more generally, the *sample autocovariance function* as

$$\hat{\gamma}(h) := \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \overline{x})(x_t - \overline{x}), \text{ for } -n < h < n.$$

Sample mean, autocovariance and autocorrelation

Suppose we have a time series with observations denoted by $x_1, \ldots, x_n$.

- The sample mean is given by $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$.
- We define the *sample variance* of our time series as

$$\hat{\gamma}(0) := \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^2,$$

and, more generally, the *sample autocovariance function* as

$$\hat{\gamma}(h) := \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \overline{x})(x_t - \overline{x}), \text{ for } -n < h < n.$$

- The *sample autocorrelation function* is given by

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \text{ for } -n < h < n.$$

Break :-)

I.i.d. noise

Suppose we have random variables $X_1, X_2, \ldots$ which are all independent of each other and have identical distribution with zero mean. We call such a collection of random variables *i.i.d. noise*, where i.i.d. stands for independent and identically distributed.

I.i.d. noise

Suppose we have random variables $X_1, X_2, \ldots$ which are all independent of each other and have identical distribution with zero mean. We call such a collection of random variables *i.i.d. noise*, where i.i.d. stands for independent and identically distributed. Let us briefly verify that i.i.d. noise constitutes a stationary process (provided its second moment exist). So let us assume that $\mathrm{E}(X_t^2) = \sigma^2 < \infty$ for all t.

I.i.d. noise

Suppose we have random variables $X_1, X_2, \ldots$ which are all independent of each other and have identical distribution with zero mean. We call such a collection of random variables *i.i.d. noise*, where i.i.d. stands for independent and identically distributed. Let us briefly verify that i.i.d. noise constitutes a stationary process (provided its second moment exist). So let us assume that $\mathrm{E}(X_t^2) = \sigma^2 < \infty$ for all t. Then $\mathrm{E}(X_t) = 0$ does not depend on t. Also, $\gamma_X(t,t) = \mathrm{Var}(X_t) = \sigma^2$ and, for $h \neq 0$ we have $\gamma_X(t+h,t) = \mathrm{Cov}(X_{t+h}, X_t) = 0$. Hence $\gamma_X(t+h,t)$ does not depend on t for each h.

I.i.d. noise

Suppose we have random variables $X_1, X_2, \ldots$ which are all independent of each other and have identical distribution with zero mean. We call such a collection of random variables *i.i.d. noise*, where i.i.d. stands for independent and identically distributed. Let us briefly verify that i.i.d. noise constitutes a stationary process (provided its second moment exist). So let us assume that $\mathrm{E}(X_t^2) = \sigma^2 < \infty$ for all t. Then $\mathrm{E}(X_t) = 0$ does not depend on t. Also, $\gamma_X(t,t) = \mathrm{Var}(X_t) = \sigma^2$ and, for $h \neq 0$ we have $\gamma_X(t+h,t) = \mathrm{Cov}(X_{t+h}, X_t) = 0$. Hence $\gamma_X(t+h,t)$ does not depend on t for each h. We often write $(X_t) \sim IID(0, \sigma^2)$.

Standard normal distribution

Let us consider the case of $n = 100$ i.i.d. random variables $X_1, \ldots, X_n$, where each random variable follows the standard normal distribution.

Standard normal distribution

Let us consider the case of $n = 100$ i.i.d. random variables $X_1, \ldots, X_n$, where each random variable follows the standard normal distribution.

Definition

A random variable X has the *standard normal/standard Gaussian* distribution if it has density function $f(x) = \phi(x)$ with

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{for } x \in \mathbb{R}.$$

Standard normal distribution

Let us consider the case of $n = 100$ i.i.d. random variables $X_1, \ldots, X_n$, where each random variable follows the standard normal distribution.

Definition

A random variable X has the *standard normal/standard Gaussian* distribution if it has density function $f(x) = \phi(x)$ with

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{for } x \in \mathbb{R}.$$

Note that we typically write $X \sim N(0, 1)$ since a standard normal random variable has mean zero and variance one.

Standard normal distribution

Let us consider the case of $n = 100$ i.i.d. random variables $X_1, \ldots, X_n$, where each random variable follows the standard normal distribution.

Definition

A random variable X has the *standard normal/standard Gaussian* distribution if it has density function $f(x) = \phi(x)$ with

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{for } x \in \mathbb{R}.$$

Note that we typically write $X \sim N(0, 1)$ since a standard normal random variable has mean zero and variance one. The c.d.f. is then denoted by $F(x) = \Phi(x)$ with

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt, \quad \text{for } x \in \mathbb{R}.$$

Unfortunately there is no explicit formula for the integral appearing in the c.d.f.!

Standard normal distribution

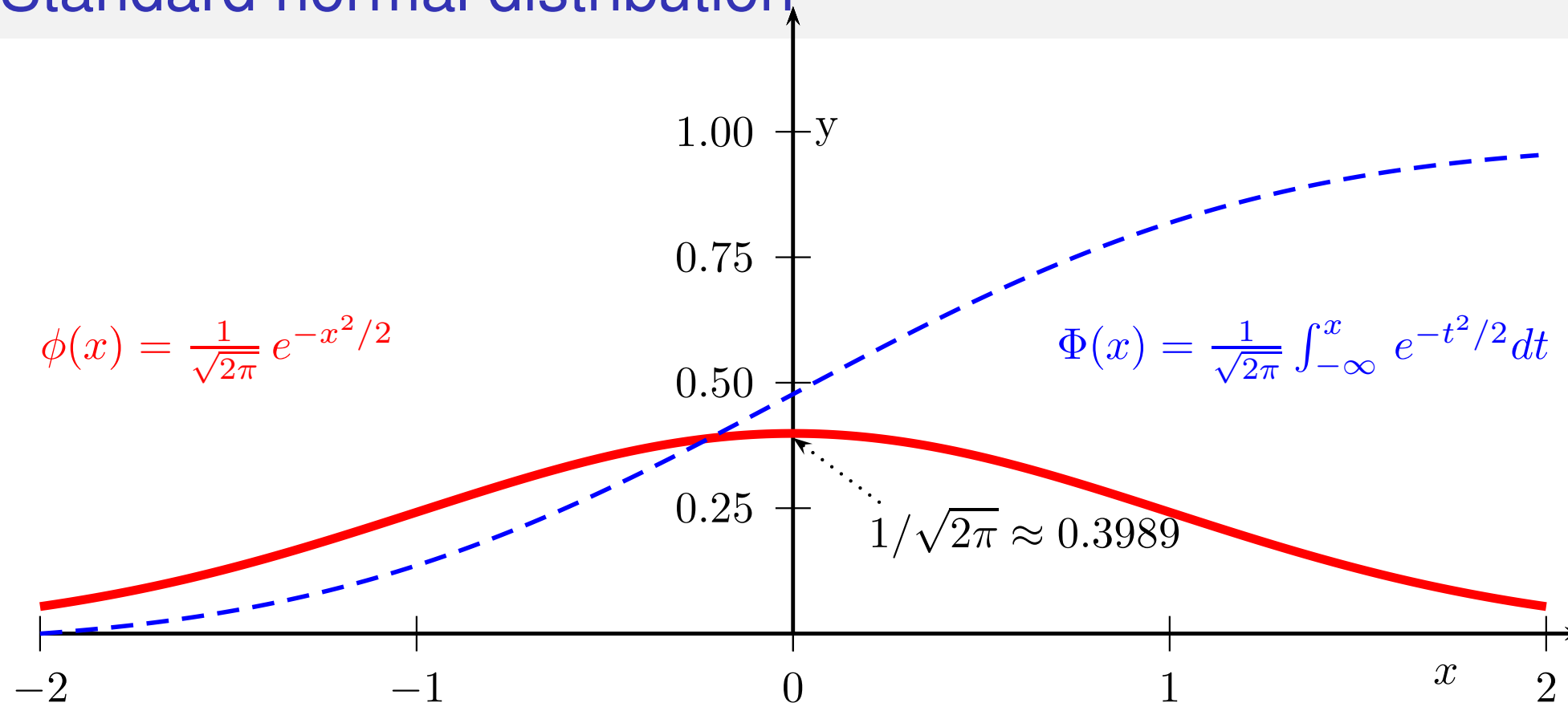


Figure: The red solid line depicts the standard Gaussian probability density function and the blue dashed line the corresponding cumulative distribution function.

Symmetry of the standard normal distribution

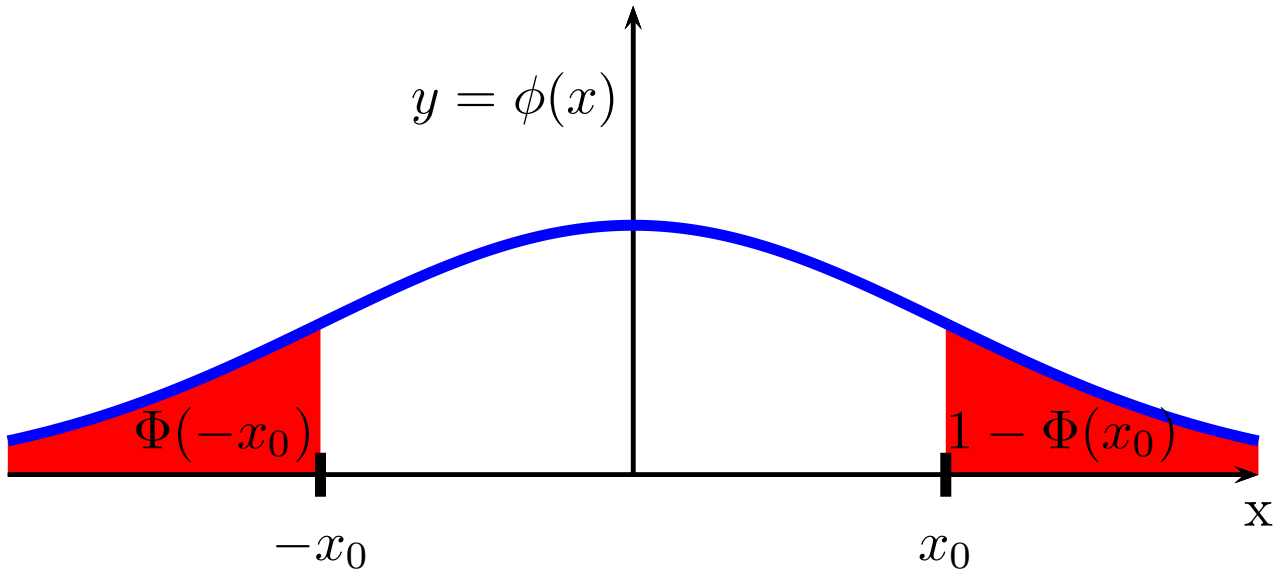


Figure: Note that the standard normal density is symmetric around 0, i.e. $\phi(x) = \phi(-x)$ for all x. This also implies that $\Phi(-x) = 1 - \Phi(x)$.

Normal distribution

Definition

Let μ denote a real number and let $\sigma > 0$. A random variable X has the *normal/ Gaussian* distribution with mean μ and variance σ^2 if it has density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{for } x \in \mathbb{R}.$$

Normal distribution

Definition

Let μ denote a real number and let $\sigma > 0$. A random variable X has the *normal/ Gaussian* distribution with mean μ and variance σ^2 if it has density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{for } x \in \mathbb{R}.$$

Note that we typically write $X \sim N(\mu, \sigma^2)$.

Normal distribution

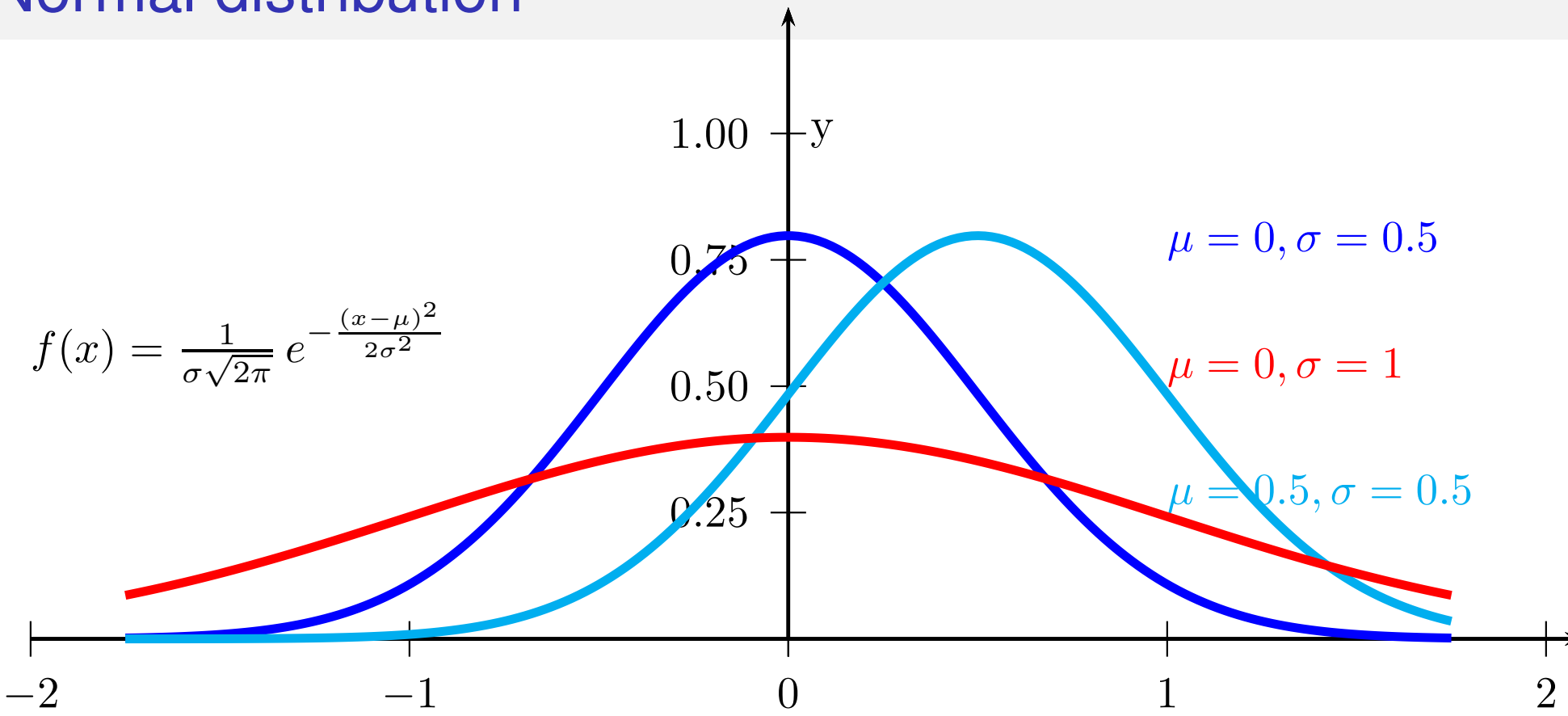


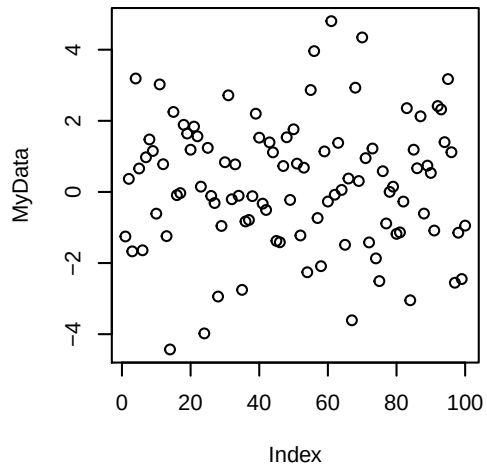
Figure: The red line depicts the standard Gaussian probability density function and the two blue lines show non-standard Gaussian probability density functions.

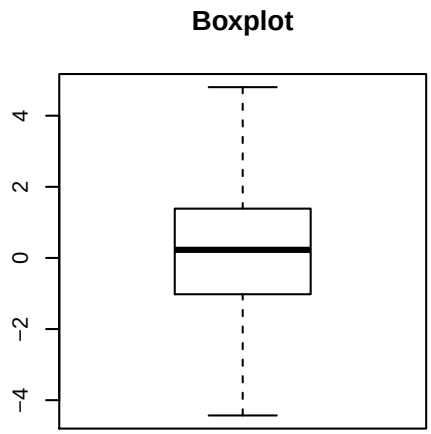
Let us consider the case of $n = 100$ i.i.d. random variables $X_1, \ldots, X_n$, where each random variable follows the standard normal distribution. We note that random variables with various distributions can be simulated in R. We will just use these tools without studying which algorithms are implemented to generated random variables.

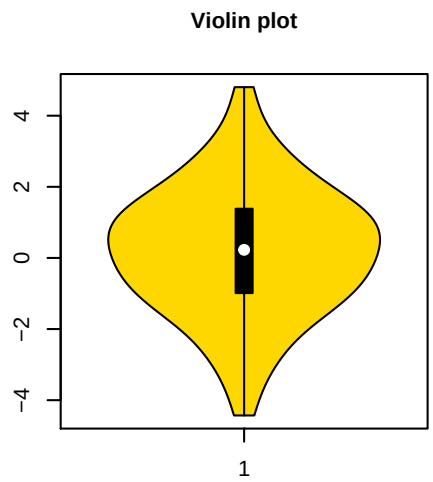
Let us consider the case of $n = 100$ i.i.d. random variables $X_1, \ldots, X_n$, where each random variable follows the standard normal distribution. We note that random variables with various distributions can be simulated in R. We will just use these tools without studying which algorithms are implemented to generated random variables.

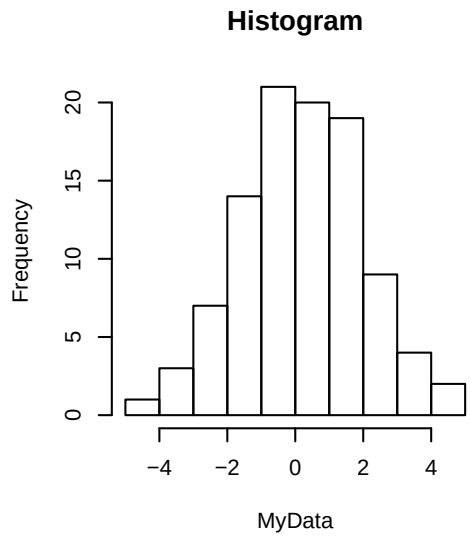
```
1 #First we fix the seed of R's random number
     generator
2 #so that we can generate a simulation which can be
     reproduced
3 set.seed(1)
4 #Simulate 100 i.i.d N(0,4) variables
5 MyData <- rnorm(100,mean=0,sd=2)
```

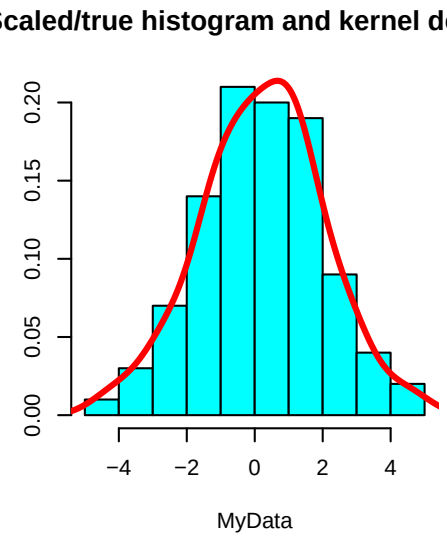
Visualising the IID N(0,4) realisations

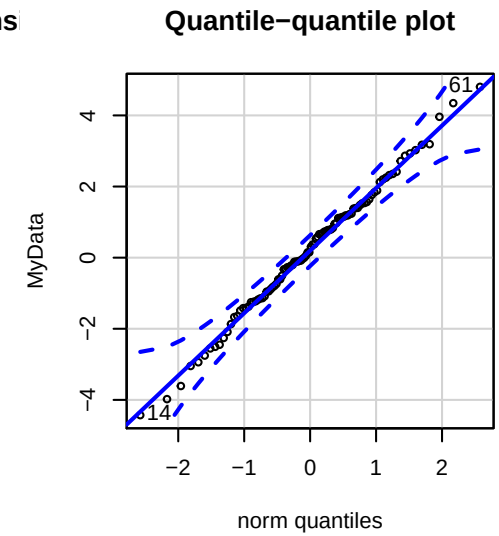












Visualisation tools

As long as we are not dealing with too many observations, it is typically very useful to plot the data. We have already studied the basic `plot` function in R.

Visualisation tools

As long as we are not dealing with too many observations, it is typically very useful to plot the data. We have already studied the basic `plot` function in R. In addition, there are various functions which can be used to give us some insights on the *distribution* of the empirical data. For instance, we often study

- box plots,
- violin plots,
- histograms and scaled histograms,
- quantile-quantile plots (short: QQ-plots).

Visualisation tools

As long as we are not dealing with too many observations, it is typically very useful to plot the data. We have already studied the basic `plot` function in R. In addition, there are various functions which can be used to give us some insights on the *distribution* of the empirical data. For instance, we often study

- box plots,
- violin plots,
- histograms and scaled histograms,
- quantile-quantile plots (short: QQ-plots).

In order to be able to interpret these plots, we need to know what *quantiles* of a distribution are.

Quantiles provide you with cut points such that a certain percentage of your distribution/your observations fall below a certain value.

Quantiles provide you with cut points such that a certain percentage of your distribution/your observations fall below a certain value.

Definition

Let $0 < p < 1$, then the *pth quantile* of a random variable X is defined as any real value x satisfying both

$$\mathrm{P}(X \geq x) \geq 1 - p \text{ and } \mathrm{P}(X \leq x) \geq p.$$

Quantiles provide you with cut points such that a certain percentage of your distribution/your observations fall below a certain value.

Definition

Let $0 < p < 1$, then the *pth quantile* of a random variable X is defined as any real value x satisfying both

$$\mathrm{P}(X \geq x) \geq 1 - p \text{ and } \mathrm{P}(X \leq x) \geq p.$$

Note that we call the 0.5 quantile the *median*, m say. Since at least half the probability is on $\{X \geq m\}$ and $\{X \leq m\}$, you can view the median as the *midpoint* of the distribution.

Quantiles provide you with cut points such that a certain percentage of your distribution/your observations fall below a certain value.

Definition

Let $0 < p < 1$, then the *pth quantile* of a random variable X is defined as any real value x satisfying both

$$\mathrm{P}(X \geq x) \geq 1 - p \text{ and } \mathrm{P}(X \leq x) \geq p.$$

Note that we call the 0.5 quantile the *median*, m say. Since at least half the probability is on $\{X \geq m\}$ and $\{X \leq m\}$, you can view the median as the *midpoint* of the distribution. Note that the theoretical quantiles are rarely available to us, so in practice we typically work with their *empirical counterparts*.

Quantiles provide you with cut points such that a certain percentage of your distribution/your observations fall below a certain value.

Definition

Let $0 < p < 1$, then the *pth quantile* of a random variable X is defined as any real value x satisfying both

$$\mathrm{P}(X \geq x) \geq 1 - p \text{ and } \mathrm{P}(X \leq x) \geq p.$$

Note that we call the 0.5 quantile the *median*, m say. Since at least half the probability is on $\{X \geq m\}$ and $\{X \leq m\}$, you can view the median as the *midpoint* of the distribution. Note that the theoretical quantiles are rarely available to us, so in practice we typically work with their *empirical counterparts*. Note that different computer packages typically use different methods for computing sample/empirical quantiles, so these estimates might vary depending on the software you use.

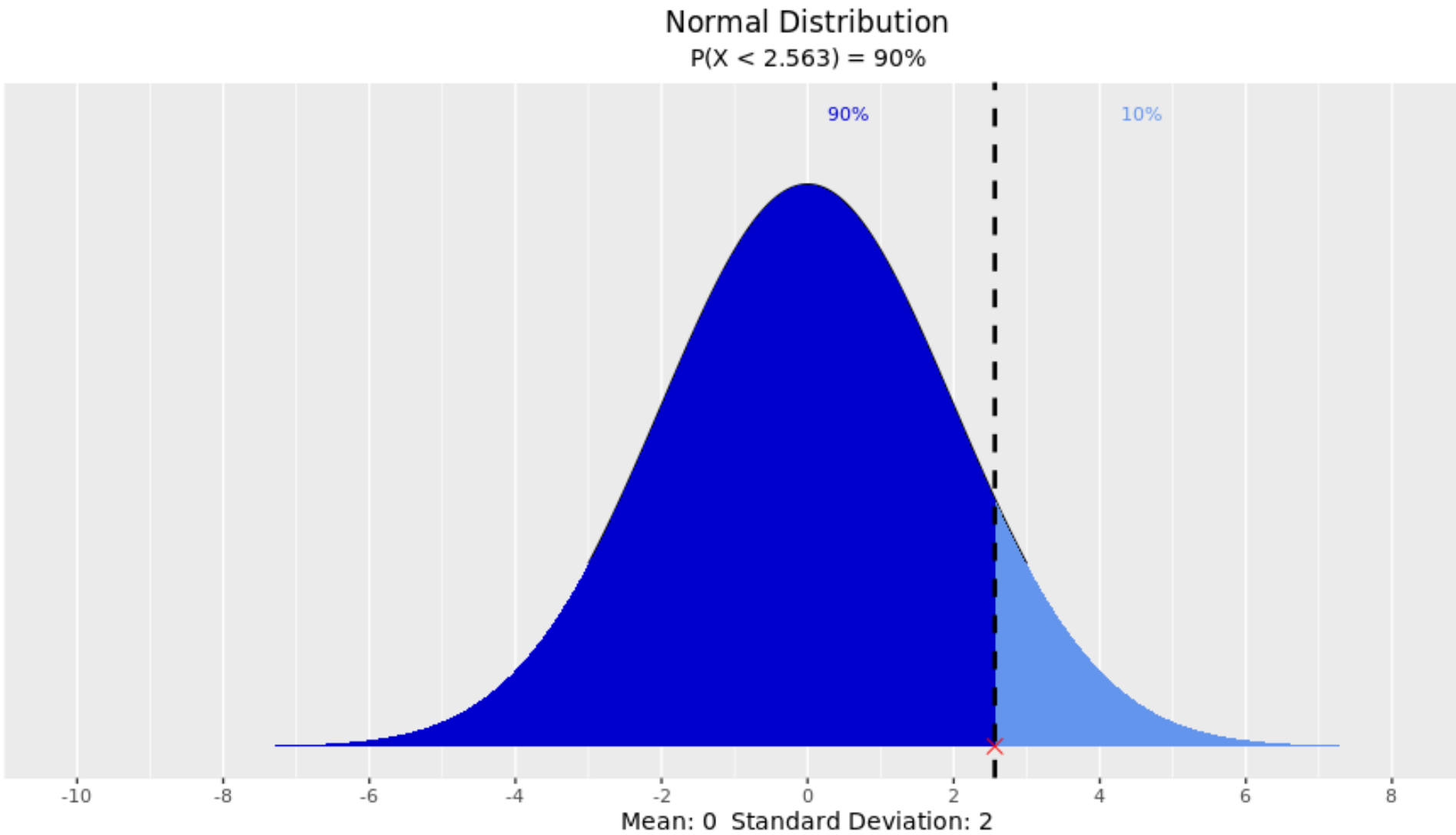


Figure: The pictures shows a plot of the probability density of the $N(0, 4)$ distribution and the red cross indicates the corresponding 90% quantile of the

A **box plot** graphically displays the following information of a vector of observations:

- The bounds of the ”box” represent the first sample quartile (i.e. the 25% quantile, let’s call it Q1) and the third sample quartile (i.e. the 75% quantile, let’s call it Q3).
- The line inside the box represents the sample median (i.e. the 50% quantile).

A **box plot** graphically displays the following information of a vector of observations:

- The bounds of the ”box” represent the first sample quartile (i.e. the 25% quantile, let’s call it Q1) and the third sample quartile (i.e. the 75% quantile, let’s call it Q3).
- The line inside the box represents the sample median (i.e. the 50% quantile).

A **violin plot** combines a boxplot with a so-called kernel density plot which estimates the probability density of the observations.

A **box plot** graphically displays the following information of a vector of observations:

- The bounds of the ”box” represent the first sample quartile (i.e. the 25% quantile, let’s call it Q1) and the third sample quartile (i.e. the 75% quantile, let’s call it Q3).
- The line inside the box represents the sample median (i.e. the 50% quantile).

A **violin plot** combines a boxplot with a so-called kernel density plot which estimates the probability density of the observations.
A **histogram** depicts the number of times an observation falls into each of the bins. Sometimes we scale the histogram such that it approximates a probability density function.

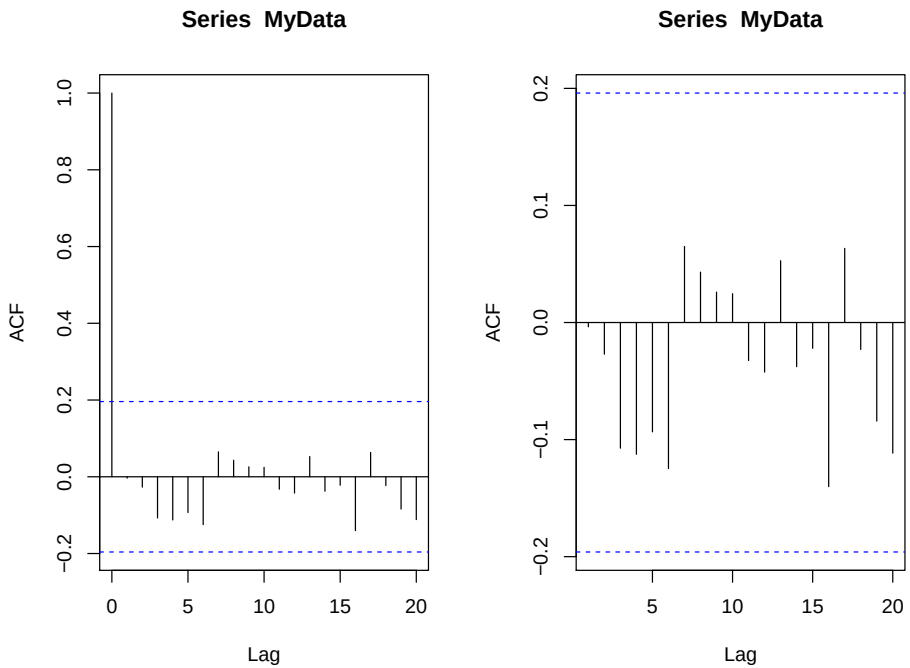
A **box plot** graphically displays the following information of a vector of observations:

- The bounds of the ”box” represent the first sample quartile (i.e. the 25% quantile, let's call it Q1) and the third sample quartile (i.e. the 75% quantile, let's call it Q3).
- The line inside the box represents the sample median (i.e. the 50% quantile).

A **violin plot** combines a boxplot with a so-called kernel density plot which estimates the probability density of the observations.
A **histogram** depicts the number of times an observation falls into each of the bins. Sometimes we scale the histogram such that it approximates a probability density function.
A **quantile-quantile** (qq) plot plots the quantiles of two distributions against each other. Typically we plot the sample quantiles against an estimated distribution (such as the estimated normal distribution). If the two distributions are similar, we expect to see a straight line through $y = x$ in the qq plot.

We can compute and plot the sample autocorrelation function using the following code:

```
#Plot the sample autocorrelation of the data
par(mfrow=c(1,2))
acf(MyData)
#Remove the lag 0
plot(acf(MyData,plot=F)[1:20])
```



Break :-)

Traditional time series analysis is typically performed using the following procedure:

Traditional time series analysis is typically performed using the following procedure:

Step 1: Plot the time series and analyse the main characteristics of the graph, in particular, you should look out for

- a trend,
- a seasonal component,
- any change points, where the behaviour of the time series changes abruptly,
- any extreme observations/outliers.

Traditional time series analysis is typically performed using the following procedure:

Step 1: Plot the time series and analyse the main characteristics of the graph, in particular, you should look out for

- a trend,
- a seasonal component,
- any change points, where the behaviour of the time series changes abruptly,
- any extreme observations/outliers.

Step 2: Remove the trend and seasonal component and possibly apply a data transformation to produce a time series which appears to be *stationary* .

Traditional time series analysis is typically performed using the following procedure:

Step 1: Plot the time series and analyse the main characteristics of the graph, in particular, you should look out for

- a trend,
- a seasonal component,
- any change points, where the behaviour of the time series changes abruptly,
- any extreme observations/outliers.

Step 2: Remove the trend and seasonal component and possibly apply a data transformation to produce a time series which appears to be *stationary* .

Step 3: Fit a (stationary) time series model to the (transformed) data.

Traditional time series analysis is typically performed using the following procedure:

Step 1: Plot the time series and analyse the main characteristics of the graph, in particular, you should look out for

- a trend,
- a seasonal component,
- any change points, where the behaviour of the time series changes abruptly,
- any extreme observations/outliers.

Step 2: Remove the trend and seasonal component and possibly apply a data transformation to produce a time series which appears to be *stationary* .

Step 3: Fit a (stationary) time series model to the (transformed) data.

Step 4: Assess the model fit and possibly compare models.

Traditional time series analysis is typically performed using the following procedure:

Step 1: Plot the time series and analyse the main characteristics of the graph, in particular, you should look out for
- a trend,
- a seasonal component,
- any change points, where the behaviour of the time series changes abruptly,
- any extreme observations/outliers.

Step 2: Remove the trend and seasonal component and possibly apply a data transformation to produce a time series which appears to be *stationary* .

Step 3: Fit a (stationary) time series model to the (transformed) data.

Step 4: Assess the model fit and possibly compare models.

Step 5: If you want to do any forecasting, then you will need to forecast observations from your stationary time series model and then invert your procedures from Step 2 to obtain forecasts of your original time series.

When you plot the data and observe any apparent discontinuities or outliers, you might want to consider breaking the series up into homogeneous segments and analysing those segments separately. Recall out bitcoin example:

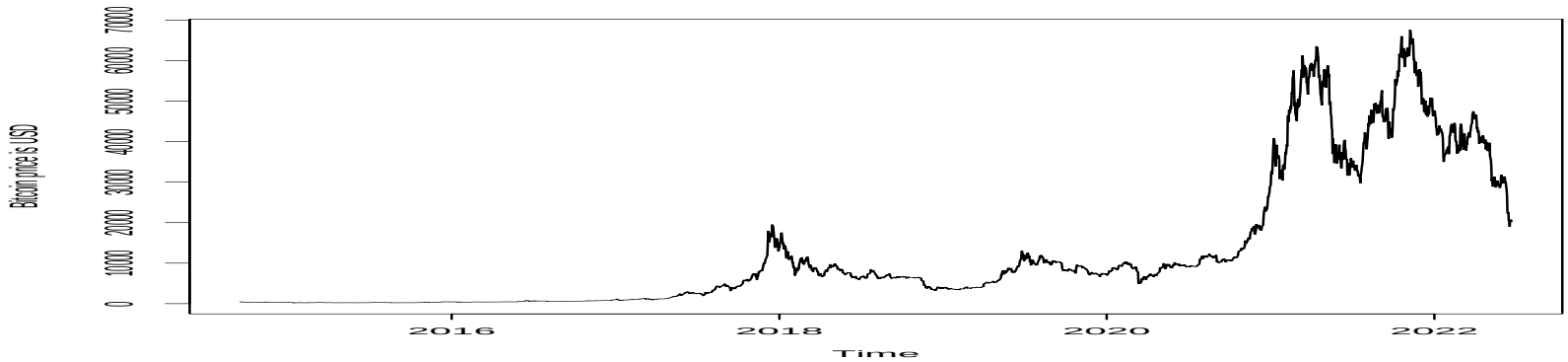


Figure: Daily bitcoin closing prices from 17/09/2014 to 22/06/2022.

You clearly observe that the behaviour of the time series changes substantially in mid 2017.

Classical decomposition model

When you study the graph of your time series, you might think that your data could be realisations of a stochastic process obtained from the *classical decomposition model*, which is defined as

$$X_t = m_t + s_t + Y_t,$$

where

- m_t is called the *trend component*, which is typically a slowly changing function,
- s_t is called the *seasonal component*, which is a periodic function with known period d,
- Y_t is a *stochastic process* which is stationary (see Definition 22).

Remark

You can use the R function `stl` *to decompose a time series into these three components, see Figure* **??**.

Step 2: Removing trend and seasonality

Two approaches are widely used for dealing with trend and seasonality: Either one can *difference* the data to generate a stationary time series, or one can fit a trend and/or seasonal model and remove the fitted model from the data to obtain a stationary time series.

Step 2: Removing trend and seasonality

Two approaches are widely used for dealing with trend and seasonality: Either one can *difference* the data to generate a stationary time series, or one can fit a trend and/or seasonal model and remove the fitted model from the data to obtain a stationary time series. We will only study the former approach in this lecture, for the latter, see for instance (**?**, Chapter 1.5) and the next problem class.

Consider the stochastic process (X_t) and define the *lag-1 difference operator* ∇ by

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t,$$

where B denotes the *back shift operator*

$$BX_t = X_{t-1}.$$

Consider the stochastic process (X_t) and define the *lag-1 difference operator* ∇ by

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t,$$

where B denotes the *back shift operator*

$$BX_t = X_{t-1}.$$

We can generalise the definitions of the difference and back shift operator to higher powers, i.e. $B^j(X_t) = X_{t-j}$ and $\nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t))$, for $j \in \mathbb{N}$ with $\nabla^0(X_t) = X_t$.

Consider the stochastic process (X_t) and define the *lag-1 difference operator* ∇ by

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t,$$

where B denotes the *back shift operator*

$$BX_t = X_{t-1}.$$

We can generalise the definitions of the difference and back shift operator to higher powers, i.e. $B^j(X_t) = X_{t-j}$ and $\nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t))$, for $j \in \mathbb{N}$ with $\nabla^0(X_t) = X_t$. Consider a linear trend function: $m_t = c_0 + c_1 t$. Then

$$\nabla m_t = m_t - m_{t-1} = c_0 + c_1 t - (c_0 + c_1(t-1)) = c_1,$$

which is a constant independent of time t.

Consider the stochastic process (X_t) and define the *lag-1 difference operator* ∇ by

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t,$$

where B denotes the *back shift operator*

$$BX_t = X_{t-1}.$$

We can generalise the definitions of the difference and back shift operator to higher powers, i.e. $B^j(X_t) = X_{t-j}$ and $\nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t))$, for $j \in \mathbb{N}$ with $\nabla^0(X_t) = X_t$. Consider a linear trend function: $m_t = c_0 + c_1 t$. Then

$$\nabla m_t = m_t - m_{t-1} = c_0 + c_1 t - (c_0 + c_1(t-1)) = c_1,$$

which is a constant independent of time t. Similarly, one can remove any polynomial trend by applying the suitable (higher order) difference operator.

Using the `diff` function in R, we can display the original data and the first and second order differences:

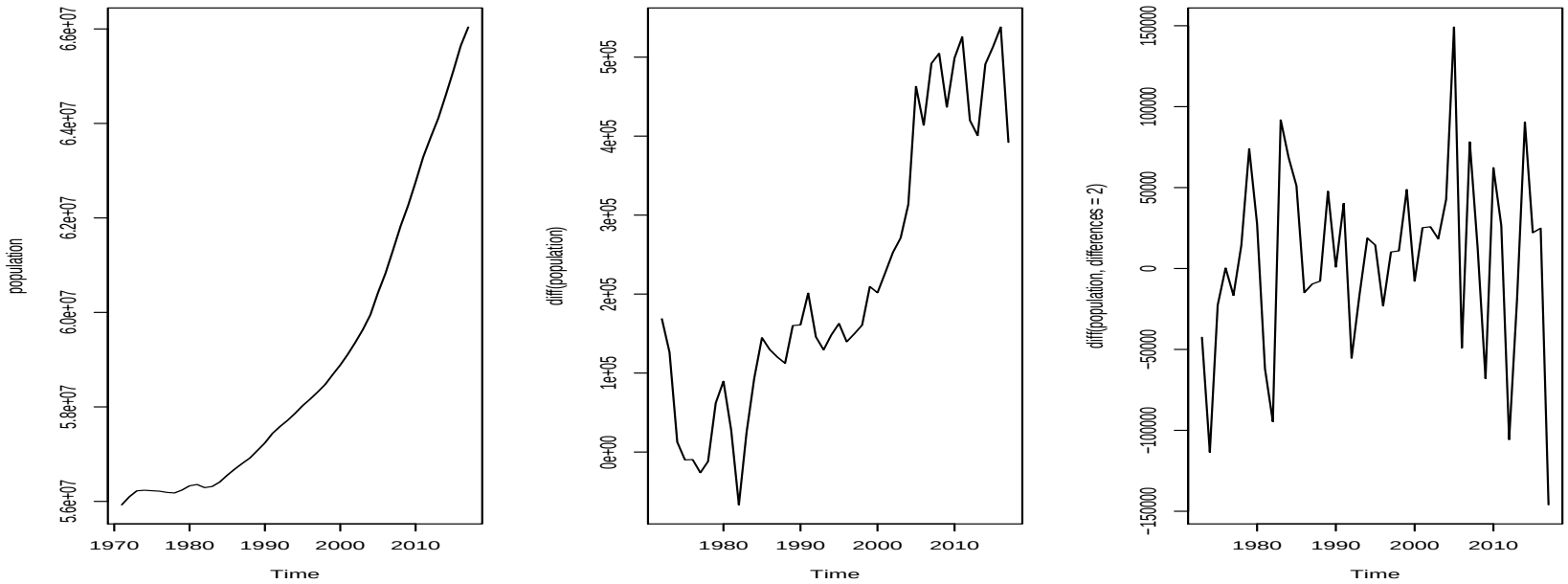


Figure: Yearly UK population from 1971 to 2017: Original data, first order differences and second order differences.

Using the `diff` function in R, we can display the original data and the first and second order differences:

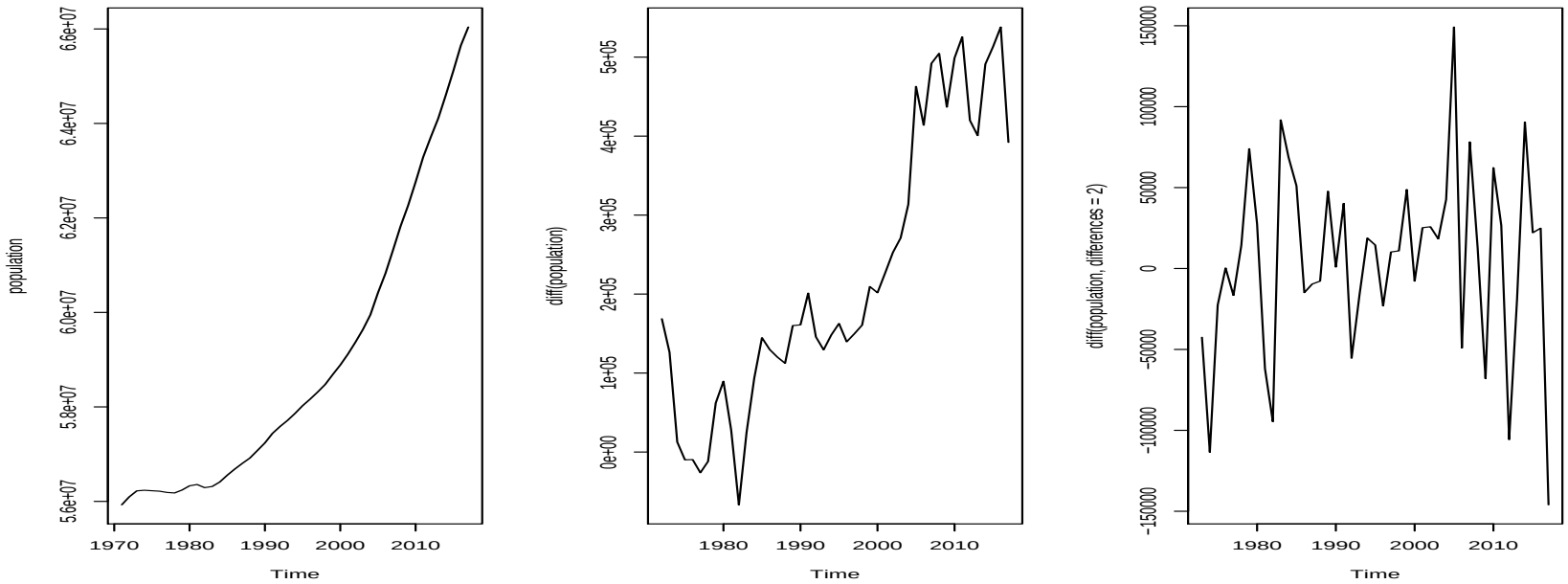


Figure: Yearly UK population from 1971 to 2017: Original data, first order differences and second order differences.

We observe that taking second order differences seems to remove the trend in the data and the new time series "looks" stationary. This conclusion still needs to be justified more precisely!

Seasonal differencing

Differencing can not only be used for removing trends, but also for removing seasonal features in a time series. For $d \in \mathbb{N}$, we define the *lag-d difference operator* ∇_d by

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d) X_t$$

Seasonal differencing

Differencing can not only be used for removing trends, but also for removing seasonal features in a time series. For $d \in \mathbb{N}$, we define the *lag-d difference operator* ∇_d by

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d) X_t$$

Consider the classical decomposition model, where s is assumed to have period d. Then

$$\nabla_d X_t = m_t - m_{t-d} + s_t - s_{t-d} + Y_t - Y_{t-d} = m_t - m_{t-d} + Y_t - Y_{t-d}.$$

Seasonal differencing

Differencing can not only be used for removing trends, but also for removing seasonal features in a time series. For $d \in \mathbb{N}$, we define the *lag-d difference operator* ∇_d by

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d) X_t$$

Consider the classical decomposition model, where s is assumed to have period d. Then

$$\nabla_d X_t = m_t - m_{t-d} + s_t - s_{t-d} + Y_t - Y_{t-d} = m_t - m_{t-d} + Y_t - Y_{t-d}.$$

Note that the trend component $m_t - m_{t-d}$ can be removed by applying a suitable power of the lag-1 difference operator.

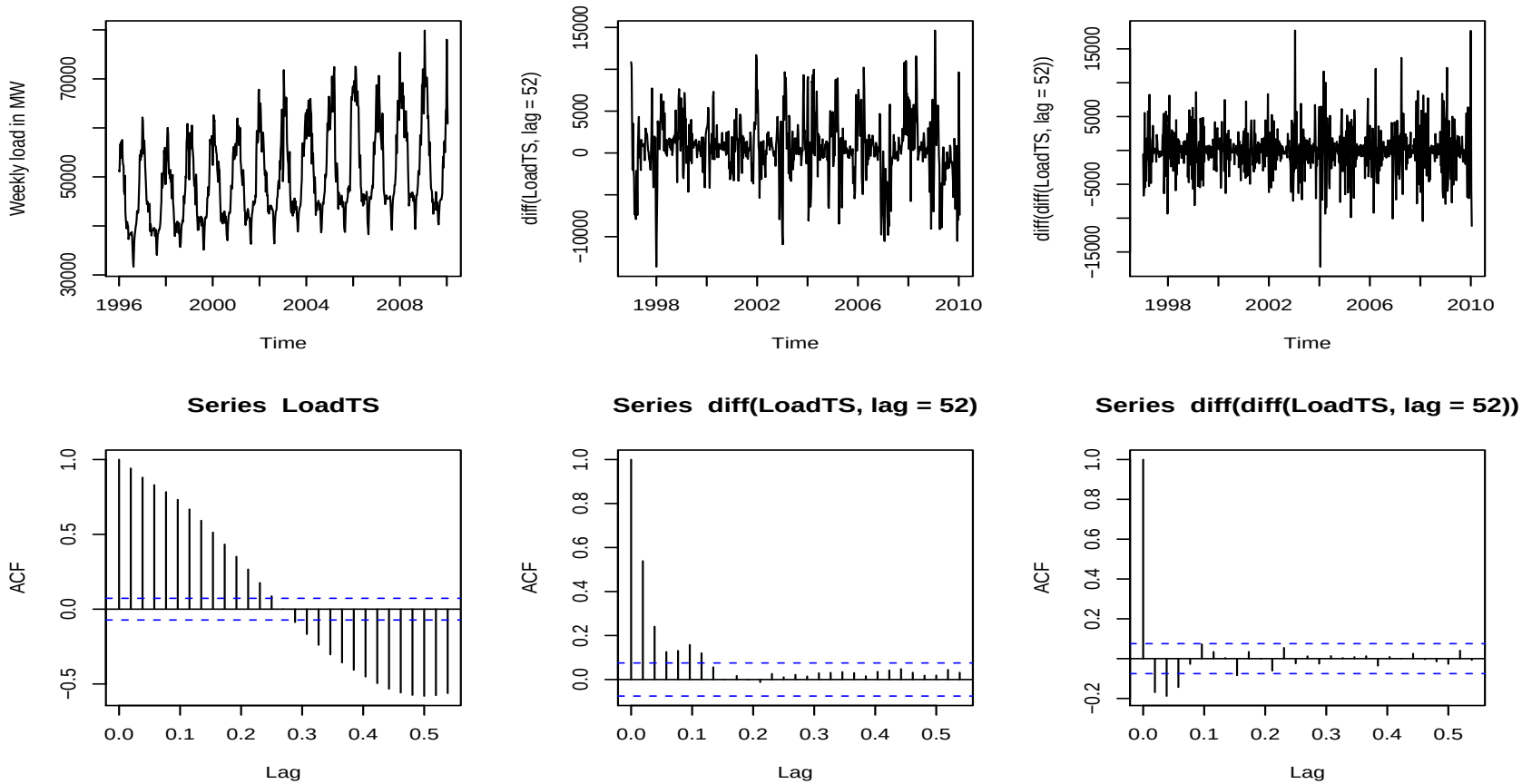


Figure: Weekly French electricity consumption data (in MW) from 1996 to 2009: The first row depicts the original time series (X_t), the seasonal differences $(\nabla_{52} X_t)$, and the differences $(\nabla\nabla_{52} X_t)$. The second row shows the corresponding empirical autocorrelation functions.

Break :-)

White noise

Definition

The stochastic process $(X_t)_{t\in\mathbb{Z}}$ is a *white noise process* if it is stationary with autocorrelation function given by

$$\rho(h) = \begin{cases} 1, & \text{for } h = 0, \\ 0, & \text{for } h \neq 0. \end{cases}$$

If the white noise has zero mean and variance denoted by $\sigma^2 = \mathrm{Var}(X_t)$, then we typically write $WN(0, \sigma^2)$.

White noise

Definition

The stochastic process $(X_t)_{t\in\mathbb{Z}}$ is a *white noise process* if it is stationary with autocorrelation function given by

$$\rho(h) = \begin{cases} 1, & \text{for } h = 0, \\ 0, & \text{for } h \neq 0. \end{cases}$$

If the white noise has zero mean and variance denoted by $\sigma^2 = \mathrm{Var}(X_t)$, then we typically write $WN(0, \sigma^2)$.

Remark

Note that the definition of the white noise does not include any distributional assumption other than that its first and second moments exist. For example, white noise does not need to follow the Gaussian distribution.

Strict white noise

We have already introduced i.i.d. random variables yesterday. We will now introduce another name for them:

Definition

A stochastic process $(X_t)_{t\in\mathbb{Z}}$ is called a *strict white noise* (SWN), if it is a series of i.i.d. random variables with finite mean and finite variance. If the strict white noise has zero mean and variance denoted by $\sigma^2 = \mathrm{Var}(X_t)$, then we typically write $SWN(0, \sigma^2)$.

ARMA(p,q) processes

Definition

Let $(\epsilon_t)_{t\in\mathbb{Z}}$ be WN(0,σ_ϵ^2). The stochastic process $(X_t)_{t\in\mathbb{Z}}$ is a *zero mean ARMA(p,q) process* if it is stationary and satisfies, for all $t \in \mathbb{Z}$,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \qquad (20.1)$$

where $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q \in \mathbb{R}$ are the model parameters.[a]

[a] We also require that the (complex) polynomials $p_{AR}(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $p_{MA}(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ have no common factors.

Remark

Note that equation (20.1) *is equivalent to*

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \quad \forall t \in \mathbb{Z}.$$

ARMA(p,q) processes with mean μ

Note that we say that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA(p,q) process with mean μ, if the centred process $(X_t - \mu)_{t\in\mathbb{Z}}$ is a zero-mean ARMA(p,q) process, i.e. it satisfies

$$(X_t - \mu) - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}. \tag{20.2}$$

ARIMA processes

Definition

Let $p, q, d \in \{0, 1, 2, \ldots\}$. A stochastic process (X_t) is called an *ARIMA(p,d,q)* process if the differenced series (Y_t), where $Y_t = \nabla^d X_t$, is an ARMA(p,q) process.

First order moving average process (MA(1))

Let $(\epsilon_t)_{t\in\mathbb{Z}}$ be WN(0, σ_ϵ^2). The stochastic process $(X_t)_{t\in\mathbb{Z}}$ is a *zero mean MA(1) process* if it is stationary and satisfies

$$X_t = \epsilon_t + \theta\epsilon_{t-1}, \quad \text{for all } t \in \mathbb{Z}.$$

First order moving average process (MA(1))

Let $(\epsilon_t)_{t\in\mathbb{Z}}$ be WN(0, σ_ϵ^2). The stochastic process $(X_t)_{t\in\mathbb{Z}}$ is a *zero mean MA(1) process* if it is stationary and satisfies

$$X_t = \epsilon_t + \theta\epsilon_{t-1}, \quad \text{for all } t \in \mathbb{Z}.$$

Exercise

Compute the mean, variance and autocorrelation function of the MA(1) process.[a]

[a] You may use without proof that for random variables X, Y, X_1, X_2, Y_1, Y_2 we have $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$ and for deterministic constants $a, b, a_1, a_2, b_1, b_2 \in \mathbb{R}$ we have $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X,Y)$ and $\mathrm{Cov}(a_1X_1 + a_2X_2, b_1Y_1 + b_2Y_2) = a_1b_1\mathrm{Cov}(X_1,Y_1) + a_1b_2\mathrm{Cov}(X_1,Y_2) + a_2b_1\mathrm{Cov}(X_2,Y_1) + a_2b_2\mathrm{Cov}(X_2,Y_2)$.

First order autoregressive process (AR(1))

Let $(\epsilon_t)_{t\in\mathbb{Z}}$ be WN(0,σ_ϵ^2). The stochastic process $(X_t)_{t\in\mathbb{Z}}$ is a *zero mean AR(1) process* if it is stationary and satisfies

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \text{for all } t \in \mathbb{Z}.$$

First order autoregressive process (AR(1))

Let $(\epsilon_t)_{t\in\mathbb{Z}}$ be WN(0,σ_ϵ^2). The stochastic process $(X_t)_{t\in\mathbb{Z}}$ is a *zero mean AR(1) process* if it is stationary and satisfies

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \text{for all } t \in \mathbb{Z}.$$

Suppose that $|\phi| < 1$ (such AR(1) processes are called *causal*). Then

$$\begin{aligned} X_t &= \phi X_{t-1} + \epsilon_t \\ &= \phi(\phi X_{t-2} + \epsilon_{t-1}) + \epsilon_t = \phi^2 X_{t-2} + \phi\epsilon_{t-1} + \epsilon_t = \cdots \\ &= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j \epsilon_{t-j} \longrightarrow \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}, \text{ as } k \to \infty. \end{aligned}$$

The above asymptotic results requires a proper proof and a statement in which sense we have convergence. Such a discussion is beyond the scope of this course.

Exercise

Compute the mean, variance and autocorrelation function of the AR(1)

Break :-)

Model estimation by maximum likelihood

Consider a general ARMA(p,q) process with mean μ and suppose you know the order of the process, i.e. p and q. We need to estimate the $p + q + 1$ parameters $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$.

Model estimation by maximum likelihood

Consider a general ARMA(p,q) process with mean μ and suppose you know the order of the process, i.e. p and q. We need to estimate the $p+q+1$ parameters $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$. Let us shorten the notation and write $\zeta = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)^\top$ for the $p+q+1$-dimensional parameter vector.

Model estimation by maximum likelihood

Consider a general ARMA(p,q) process with mean μ and suppose you know the order of the process, i.e. p and q. We need to estimate the $p+q+1$ parameters $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$. Let us shorten the notation and write $\zeta = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)^\top$ for the $p+q+1$-dimensional parameter vector. In addition, you might want to estimate the variance σ_ϵ^2 of the white noise process which appears in the definition of the ARMA process.

Model estimation by maximum likelihood

Consider a general ARMA(p,q) process with mean μ and suppose you know the order of the process, i.e. p and q. We need to estimate the $p+q+1$ parameters $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$. Let us shorten the notation and write $\zeta = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)^\top$ for the $p+q+1$-dimensional parameter vector. In addition, you might want to estimate the variance σ_ϵ^2 of the white noise process which appears in the definition of the ARMA process. The maximum likelihood estimation technique finds the parameter values which maximise the probability of obtaining the data which we have observed.

Model estimation by maximum likelihood

More precisely, consider our time series $x_1, \ldots, x_n$ which we regard as realisations of the random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$. Suppose that the random vector has a joint density function denoted by f. Rather than writing $f(x_1, \ldots, x_n)$ we will write $f(x_1, \ldots, x_n|\boldsymbol{\zeta})$ to indicate that the joint density function depends on the parameter vector $\boldsymbol{\zeta}$. Then the likelihood function is given by $L(\boldsymbol{\zeta}) = f(x_1, \ldots, x_n|\boldsymbol{\zeta})$ and the maximum likelihood estimator $\widehat{\boldsymbol{\zeta}}$ is a vector, out of all possible choices of $\boldsymbol{\zeta}$, which maximises $L(\boldsymbol{\zeta})$.

Model estimation by maximum likelihood

Example

In the case when $\mathbf{x} = (x_1, \ldots, x_n)^\top$ is assumed to be a realisation of a Gaussian time series model $\mathbf{X} = (X_1, \ldots, X_n)^\top$ with zero mean, then the likelihood of $\mathbf{x} = (x_1, \ldots, x_n)^\top$ is given by

$$L(\zeta) = (2\pi)^{-n/2}(\det(\Gamma_n))^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^\top \Gamma_n^{-1}\mathbf{x}\right),$$

where $\Gamma_n = \mathrm{E}(\mathbf{X}^\top \mathbf{X})$ is the covariance matrix which is assumed to be nonsingular. Note that Γ_n depends on the model parameters ζ.

Order selection

We typically use an information criterion for estimating the order (i.e. p and q) in the ARMA(p,q) process. Information criteria give you a method for finding the right balance between having a model with high likelihood, but not too many parameters. So they typically include the likelihood, denoted by L, and a penalty term involving the number of parameters. Suppose our time series has n observations.

Order selection

Three information criteria are widely used:

AIC Akaike's Information Criterion for ARMA(p,q) processes:

$$AIC = -2\log(L) + 2(p+q+b+1),$$

where $b = 1$ for non-zero mean $\mu \neq 0$ and $b = 0$ for zero mean $\mu = 0$.

AICC the Corrected AIC:

$$AICC = AIC + \frac{2(p+q+b+1)(p+q+b+2)}{n-p-q-b-2},$$

BIC the Bayesian Information Criterion

$$BIC = AIC + (\log(n) - 2)(p+q+b+1).$$

We select the best model by minimising AIC, AICC or BIC. Note that the penalty terms in the AIC and AICC are asymptotically equivalent when $n \to \infty$, but AICC puts a higher penalty on large-order models, whereas the AIC is often prone to overfitting.

Computing the residuals of an ARMA process

Consider (for simplicity) a zero-mean ARMA(p,q) process as defined in equation (20.1):

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}.$$

We can rearrange this equation to obtain:

$$\epsilon_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} - (\theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}).$$

Let $\hat{\phi}_1, \ldots \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q$ denote the parameter estimates obtained through (maximum likelihood) estimation. Given our time series $(x_1, \ldots, x_n)$ we can compute the residuals as the difference between the original data and the so-called fitted values:

$$\hat{\epsilon}_t = x_t - \hat{\phi}_1 x_{t-1} - \cdots - \hat{\phi}_p x_{t-p} - (\hat{\theta}_1 \hat{\epsilon}_{t-1} + \cdots + \hat{\theta}_q \hat{\epsilon}_{t-q}),$$

for $t = 1, \ldots, n$. Note that we need to choose initial values $x_0, \ldots, x_{1-p}, \hat{\epsilon}_0, \ldots, \hat{\epsilon}_{1-q}$.

Are the residuals (strict) white noise?

For our ARMA(p,q) model to describe the data well, we expect that the residuals are (strict) white noise.
For a time series $(x_1, \ldots, x_n)^\top$ (being a realisation of $(X_1, \ldots, X_n)^\top$), we denote the sample mean by $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$ and the *sample autocovariance function* by

$$\hat{\gamma}_x(h) := \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \overline{x})(x_t - \overline{x}), \text{ for } -n < h < n.$$

Then the *sample autocorrelation function* is given by

$$\hat{\rho}_x(h) := \frac{\hat{\gamma}_x(h)}{\hat{\gamma}_x(0)}, \text{ for } -n < h < n.$$

We have already used the R function `acf` to display the *correlogram*, which is the plot

$$\{(h, \hat{\rho}_x(h)) : h = 0, 1, 2, \ldots\}.$$

Did you notice the blue dotted lines in the correlogram produced in R??

95% bounds in the correlogram

Suppose briefly that $(X_1, \ldots, X_n)$ is a strict white noise. In that case, one can show that, for large n, the sample autocorrelations $(\hat{\rho}_X(1), \ldots, \hat{\rho}_X(h))^\top$ behave like observations from i.i.d. random variables with $N(0, 1/n)$ distribution. That means that we expect that 95% of our observations fall between the 0.025 and 0.975 quantiles of the $N(0, 1/n)$ distribution, which are given by $-1.96/\sqrt{n}$ and $1.96/\sqrt{n}$, respectively, and that 5% of our observations fall outside these bands[1]. Hence, we typically plot these bounds with the sample autocorrelations. If more than 5% of the estimated correlations are outside the bounds, we would conclude that our data are not realisations of a white noise process[2].

95% bounds in the correlogram

Suppose briefly that $(X_1, \ldots, X_n)$ is a strict white noise. In that case, one can show that, for large n, the sample autocorrelations $(\hat{\rho}_X(1), \ldots, \hat{\rho}_X(h))^\top$ behave like observations from i.i.d. random variables with $N(0, 1/n)$ distribution. That means that we expect that 95% of our observations fall between the 0.025 and 0.975 quantiles of the $N(0, 1/n)$ distribution, which are given by $-1.96/\sqrt{n}$ and $1.96/\sqrt{n}$, respectively, and that 5% of our observations fall outside these bands[1]. Hence, we typically plot these bounds with the sample autocorrelations. If more than 5% of the estimated correlations are outside the bounds, we would conclude that our data are not realisations of a white noise process[2].

Remark

When checking the goodness of fit of your estimated ARMA process, you would compute the sample acf of the residuals, i.e. $(\rho_{\hat{\epsilon}}(1), \ldots, \rho_{\hat{\epsilon}}(h))^\top$*, to check whether they can be regarded as realisations of a white noise process*

Portmanteau tests

Box-Pierce test The Box-Pierce test statistics is given by

$$Q_{BP} = n \sum_{j=1}^{H} \hat{\rho}(j)^2.$$

Ljung-Box test The Ljung-Box test statistics is given by

$$Q_{LB} = n(n+2) \sum_{j=1}^{H} \frac{\hat{\rho}(j)^2}{n-j}.$$

In both cases, a large value of Q_{BP} or Q_{LB} suggests that the data do not come from a strict white noise series. In order to decide what we mean by a "large value", we use the fact, that in the case of a strict white noise, both statistics follow asymptotically a χ^2-distribution with $H - K$ degrees of freedom, where K denotes the number of model parameters (so in the ARMA(p,q)-case with zero mean, we have $K = p + q$ and if the tests are applied to raw data, then we set $K = 0$)[3] Note that the Ljung-Box test is generally preferred to the Box-Pierce test.

Break :-)

Forecasting an ARMA(1,1) process

Consider an ARMA(1,1) process with zero mean given by

$$X_t = \phi X_{t-1} + \epsilon_t + \theta \epsilon_{t-1}.$$

Forecasting an ARMA(1,1) process

Consider an ARMA(1,1) process with zero mean given by

$$X_t = \phi X_{t-1} + \epsilon_t + \theta \epsilon_{t-1}.$$

Suppose we have realisations $x_1, \ldots, x_n$ from the ARMA(1,1) process and we would like to make a one (or even h) step ahead prediction, i.e. we would like to find the value $\hat{x}_{n+1}$ (or $\hat{x}_{n+h}$, for $h \in \mathbb{N}$). How do we proceed?

Forecasting an ARMA(1,1) process

Consider an ARMA(1,1) process with zero mean given by

$$X_t = \phi X_{t-1} + \epsilon_t + \theta \epsilon_{t-1}.$$

Suppose we have realisations $x_1, \ldots, x_n$ from the ARMA(1,1) process and we would like to make a one (or even h) step ahead prediction, i.e. we would like to find the value $\hat{x}_{n+1}$ (or $\hat{x}_{n+h}$, for $h \in \mathbb{N}$). How do we proceed?

Step 1: Write down the ARMA(1,1) equation for time $n+1$

$$X_{n+1} = \phi X_n + \epsilon_{n+1} + \theta \epsilon_n.$$

Step 2: Replace all unknown quantities by suitable forecasts based on the information available to you up to time n. We will use the notation $\hat{x}_{n+1}$ for the predicted value of the time series at time $n+1$ given the information up to time n.

Forecasting an ARMA(1,1) process

- Note that we assume that we have estimated the parameters ϕ, θ and hence replace them with their estimates $\hat{\phi}, \hat{\theta}$.

Forecasting an ARMA(1,1) process

- Note that we assume that we have estimated the parameters ϕ, θ and hence replace them with their estimates $\hat{\phi}, \hat{\theta}$.
- We know the realisation x_n of X_n. Hence we use $\hat{x}_n = x_n$.

Forecasting an ARMA(1,1) process

- Note that we assume that we have estimated the parameters ϕ, θ and hence replace them with their estimates $\hat{\phi}, \hat{\theta}$.
- We know the realisation x_n of X_n. Hence we use $\hat{x}_n = x_n$.
- What about $\hat{\epsilon}_{n+1}$? Given that we work under the white noise assumption, we can set $\hat{\epsilon}_{n+1} = 0$ (and even $\hat{\epsilon}_{n+h} = 0$).

Forecasting an ARMA(1,1) process

- Note that we assume that we have estimated the parameters ϕ, θ and hence replace them with their estimates $\hat{\phi}, \hat{\theta}$.
- We know the realisation x_n of X_n. Hence we use $\hat{x}_n = x_n$.
- What about $\hat{\epsilon}_{n+1}$? Given that we work under the white noise assumption, we can set $\hat{\epsilon}_{n+1} = 0$ (and even $\hat{\epsilon}_{n+h} = 0$).
- Finally we need to find $\hat{\epsilon}_n$. However, this residual is known to us after we have estimated the ARMA(1,1) process. We can write:

$$\epsilon_t = X_t - \phi X_{t-1} - \theta \epsilon_{t-1}.$$

Forecasting an ARMA(1,1) process

- Hence (given starting values X_0, ϵ_0)

$$\begin{aligned}
\epsilon_1 &= X_1 - \phi X_0 - \theta \epsilon_0, \\
\epsilon_2 &= X_2 - \phi X_1 - \theta \epsilon_1, \\
&\vdots \\
\epsilon_n &= X_n - \phi X_{n-1} - \theta \epsilon_{n-1}.
\end{aligned}$$

Forecasting an ARMA(1,1) process

- Hence (given starting values X_0, ϵ_0)

$$\begin{aligned}
\epsilon_1 &= X_1 - \phi X_0 - \theta \epsilon_0, \\
\epsilon_2 &= X_2 - \phi X_1 - \theta \epsilon_1, \\
&\vdots \\
\epsilon_n &= X_n - \phi X_{n-1} - \theta \epsilon_{n-1}.
\end{aligned}$$

Hence (given starting values $x_0, \hat{\epsilon}_0$), we can recursively compute

$$\begin{aligned}
\hat{\epsilon}_1 &= x_1 - \hat{\phi} x_0 - \hat{\theta} \hat{\epsilon}_0, \\
\hat{\epsilon}_2 &= x_2 - \hat{\phi} x_1 - \hat{\theta} \hat{\epsilon}_1, \\
&\vdots \\
\hat{\epsilon}_n &= x_n - \hat{\phi} x_{n-1} - \hat{\theta} \hat{\epsilon}_{n-1}.
\end{aligned}$$

Forecasting an ARMA(1,1) process

- Hence (given starting values X_0, ϵ_0)

$$\begin{aligned}
\epsilon_1 &= X_1 - \phi X_0 - \theta \epsilon_0, \\
\epsilon_2 &= X_2 - \phi X_1 - \theta \epsilon_1, \\
&\vdots \\
\epsilon_n &= X_n - \phi X_{n-1} - \theta \epsilon_{n-1}.
\end{aligned}$$

Hence (given starting values $x_0, \hat{\epsilon}_0$), we can recursively compute

$$\begin{aligned}
\hat{\epsilon}_1 &= x_1 - \hat{\phi} x_0 - \hat{\theta} \hat{\epsilon}_0, \\
\hat{\epsilon}_2 &= x_2 - \hat{\phi} x_1 - \hat{\theta} \hat{\epsilon}_1, \\
&\vdots \\
\hat{\epsilon}_n &= x_n - \hat{\phi} x_{n-1} - \hat{\theta} \hat{\epsilon}_{n-1}.
\end{aligned}$$

We then get the following formula for the one-step ahead prediction:

$$\hat{x}_{n+1} = \hat{\phi} x_n + \hat{\theta} \hat{\epsilon}_n.$$

Forecasting an ARMA(1,1) process two/h steps ahead

Step 1: We write

$$X_{n+2} = \phi X_{n+1} + \epsilon_{n+2} + \theta \epsilon_{n+1}.$$

Forecasting an ARMA(1,1) process two/h steps ahead

Step 1: We write

$$X_{n+2} = \phi X_{n+1} + \epsilon_{n+2} + \theta \epsilon_{n+1}.$$

Step 2: We note that both white noise terms will vanish as mentioned above: $\hat{\epsilon}_{n+h} = 0$ for all $h \in \mathbb{N}$.

Forecasting an ARMA(1,1) process two/h steps ahead

Step 1: We write

$$X_{n+2} = \phi X_{n+1} + \epsilon_{n+2} + \theta \epsilon_{n+1}.$$

Step 2: We note that both white noise terms will vanish as mentioned above: $\hat{\epsilon}_{n+h} = 0$ for all $h \in \mathbb{N}$. Hence we have

$$\hat{x}_{n+2} = \hat{\phi}\hat{x}_{n+1} = \hat{\phi}(\hat{\phi}x_n + \hat{\theta}\hat{\epsilon}_n) = \hat{\phi}^2 x_n + \hat{\phi}\hat{\theta}\hat{\epsilon}_n.$$

Forecasting an ARMA(1,1) process two/h steps ahead

Step 1: We write

$$X_{n+2} = \phi X_{n+1} + \epsilon_{n+2} + \theta \epsilon_{n+1}.$$

Step 2: We note that both white noise terms will vanish as mentioned above: $\hat{\epsilon}_{n+h} = 0$ for all $h \in \mathbb{N}$. Hence we have

$$\hat{x}_{n+2} = \hat{\phi}\hat{x}_{n+1} = \hat{\phi}(\hat{\phi}x_n + \hat{\theta}\hat{\epsilon}_n) = \hat{\phi}^2 x_n + \hat{\phi}\hat{\theta}\hat{\epsilon}_n.$$

This procedure can be iterated: Since, for $h \in \mathbb{N}, h \geq 2$, we have

$$X_{n+h} = \phi X_{n+h-1} + \epsilon_{n+h} + \theta \epsilon_{n+h-1},$$

Forecasting an ARMA(1,1) process two/h steps ahead

Step 1: We write

$$X_{n+2} = \phi X_{n+1} + \epsilon_{n+2} + \theta \epsilon_{n+1}.$$

Step 2: We note that both white noise terms will vanish as mentioned above: $\hat{\epsilon}_{n+h} = 0$ for all $h \in \mathbb{N}$. Hence we have

$$\hat{x}_{n+2} = \hat{\phi}\hat{x}_{n+1} = \hat{\phi}(\hat{\phi}x_n + \hat{\theta}\hat{\epsilon}_n) = \hat{\phi}^2 x_n + \hat{\phi}\hat{\theta}\hat{\epsilon}_n.$$

This procedure can be iterated: Since, for $h \in \mathbb{N}, h \geq 2$, we have

$$X_{n+h} = \phi X_{n+h-1} + \epsilon_{n+h} + \theta \epsilon_{n+h-1},$$

the h-step ahead forecast is given by

$$\hat{x}_{n+h} = \hat{\phi}\hat{x}_{n+h-1} = \hat{\phi}^h x_n + \hat{\phi}^{h-1}\hat{\theta}\hat{\epsilon}_n.$$

Case study in R

Break :-)

Value at Risk

Value at risk (VaR) (despite various shortcomings) is still a very widely used risk measure in financial risk management. We will discuss how VaR can be estimated and we also introduce a related risk measure called expected shortfall (ES).

Value at Risk

Value at risk (VaR) (despite various shortcomings) is still a very widely used risk measure in financial risk management. We will discuss how VaR can be estimated and we also introduce a related risk measure called expected shortfall (ES). Suppose we have a financial portfolio consisting of one or more risky financial assets. In addition, let us fix a time horizon $\Delta > 0$ say. We associate a loss L with the portfolio whose cumulative distribution function is given by $F_L(l) = \mathrm{P}(L \leq l)$. The intuition behind the definition of the value at risk is to choose a maximum loss which is very unlikely to be exceeded. Let's formalise this idea.

Value at Risk

Value at risk (VaR) (despite various shortcomings) is still a very widely used risk measure in financial risk management. We will discuss how VaR can be estimated and we also introduce a related risk measure called expected shortfall (ES). Suppose we have a financial portfolio consisting of one or more risky financial assets. In addition, let us fix a time horizon $\Delta > 0$ say. We associate a loss L with the portfolio whose cumulative distribution function is given by $F_L(l) = \mathrm{P}(L \leq l)$. The intuition behind the definition of the value at risk is to choose a maximum loss which is very unlikely to be exceeded. Let's formalise this idea.

Definition

Let $\alpha \in (0, 1)$ denote a confidence level. The value at risk of the portfolio at confidence level α is the α-quantile of the loss distribution, denoted by VaR_α.

Typically, we choose $\alpha = 0.95$ or $\alpha = 0.99$.

Value at Risk

Adapting the definition of a quantile introduced earlier (Definition 26): Let $0 < \alpha < 1$, then the VaR_α of a random variable L is defined as any real value satisfying both

$$\mathrm{P}(L \geq VaR_\alpha) \geq 1 - \alpha \text{ and } \mathrm{P}(L \leq VaR_\alpha) \geq \alpha.$$

Note that in the case when F_L is continuous and strictly increasing, then the VaR can be expressed in terms of the inverse of the cumulative distribution function: $VaR_\alpha = F_L^{-1}(\alpha)$.

Example 1: Value at Risk

Suppose that $L \sim N(\mu, \sigma^2)$. For a confidence level $\alpha \in (0, 1)$, we have

$$VaR_\alpha = \mu + \sigma \Phi^{-1}(\alpha).$$

Example 1: Value at Risk

Suppose that $L \sim N(\mu, \sigma^2)$. For a confidence level $\alpha \in (0, 1)$, we have

$$VaR_\alpha = \mu + \sigma\Phi^{-1}(\alpha).$$

Here $\Phi^{-1}(\alpha)$ denotes the αth quantile of the standard normal distribution. We can now verify that the VaR_α satisfies our definition: For this note that $L \sim N(\mu, \sigma^2)$ implies that $\frac{L-\mu}{\sigma} \sim N(0, 1)$.

Example 1: Value at Risk

Suppose that $L \sim N(\mu, \sigma^2)$. For a confidence level $\alpha \in (0,1)$, we have

$$VaR_\alpha = \mu + \sigma\Phi^{-1}(\alpha).$$

Here $\Phi^{-1}(\alpha)$ denotes the αth quantile of the standard normal distribution. We can now verify that the VaR_α satisfies our definition: For this note that $L \sim N(\mu, \sigma^2)$ implies that $\frac{L-\mu}{\sigma} \sim N(0,1)$. Then

$$\begin{aligned} \mathrm{P}(L \geq VaR_\alpha) &= \mathrm{P}\left(L \geq \mu + \sigma\Phi^{-1}(\alpha)\right) = \mathrm{P}\left(\frac{L-\mu}{\sigma} \geq \Phi^{-1}(\alpha)\right) \\ &= 1 - \mathrm{P}\left(\frac{L-\mu}{\sigma} \leq \Phi^{-1}(\alpha)\right) = 1 - \Phi(\Phi^{-1}(\alpha)) = 1 - \alpha, \end{aligned}$$

and

$$\mathrm{P}(L \leq VaR_\alpha) = \mathrm{P}(L \leq \mu + \sigma\Phi^{-1}(\alpha)) = \mathrm{P}\left(\frac{L-\mu}{\sigma} \leq \Phi^{-1}(\alpha)\right) = \Phi(\Phi^{-1}(\alpha)$$

Example 2: Value at Risk

Suppose that $(L-\mu)/\sigma$ follows a standard Student t distribution with $\nu > 0$ degrees of freedom, i.e. the corresponding probability density[4] is given by

$$f_{(L-\mu)/\sigma}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \text{ for } x \in \mathbb{R}.$$

For a confidence level $\alpha \in (0,1)$, we have

$$VaR_\alpha = \mu + \sigma t_\nu^{-1}(\alpha),$$

where $t_\nu^{-1}(\alpha)$ denotes the αth quantile of the standard Student t distribution.

Shortcomings of Value at Risk

- Since the VaR_α is just the quantile of the loss distribution, it gives you no information on the *severity* of the loss which happens with a probability smaller than $1 - \alpha$.

Shortcomings of Value at Risk

- Since the VaR_α is just the quantile of the loss distribution, it gives you no information on the *severity* of the loss which happens with a probability smaller than $1 - \alpha$.
- The VaR is not subadditive, which means that if you have two portfolios with losses L_1 and L_2 the VaR of the aggregated portfolio loss $L = L_1 + L_2$ is not necessarily bounded by the individual VaRs, i.e. we do not necessarily have that

$$VaR_\alpha(L) \leq VaR_\alpha(L_1) + VaR_\alpha(L_2).$$

This is counter-intuitive to the fact that we expect the risk of a portfolio to decrease when we diversify the assets.

Shortcomings of Value at Risk

- Since the VaR_α is just the quantile of the loss distribution, it gives you no information on the *severity* of the loss which happens with a probability smaller than $1 - \alpha$.
- The VaR is not subadditive, which means that if you have two portfolios with losses L_1 and L_2 the VaR of the aggregated portfolio loss $L = L_1 + L_2$ is not necessarily bounded by the individual VaRs, i.e. we do not necessarily have that

$$VaR_\alpha(L) \leq VaR_\alpha(L_1) + VaR_\alpha(L_2).$$

  This is counter-intuitive to the fact that we expect the risk of a portfolio to decrease when we diversify the assets.
- When computing the VaR (and other risk measures), there is significant model risk. What if the normal/Student t assumption is wrong?

Shortcomings of Value at Risk

- Since the VaR_α is just the quantile of the loss distribution, it gives you no information on the *severity* of the loss which happens with a probability smaller than $1-\alpha$.
- The VaR is not subadditive, which means that if you have two portfolios with losses L_1 and L_2 the VaR of the aggregated portfolio loss $L = L_1 + L_2$ is not necessarily bounded by the individual VaRs, i.e. we do not necessarily have that

$$VaR_\alpha(L) \leq VaR_\alpha(L_1) + VaR_\alpha(L_2).$$

  This is counter-intuitive to the fact that we expect the risk of a portfolio to decrease when we diversify the assets.
- When computing the VaR (and other risk measures), there is significant model risk. What if the normal/Student t assumption is wrong?
- You also need to be careful about which confidence level to choose in the VaR computations. Do you have enough observations to be able to reliably estimate extreme quantiles? You might need to apply tools from extreme value theory in some applications.

Expected shortfall(ES)

Definition

Consider a loss L with $\mathrm{E}(|L|) < \infty$. Then the expected shortfall (ES) at confidence level $\alpha \in (0, 1)$ is defined as

$$ES_\alpha = \frac{1}{1-\alpha} \int_\alpha^1 VaR_u(L) du.$$

Expected shortfall(ES)

Definition

Consider a loss L with $\mathrm{E}(|L|) < \infty$. Then the expected shortfall (ES) at confidence level $\alpha \in (0,1)$ is defined as

$$ES_\alpha = \frac{1}{1-\alpha} \int_\alpha^1 VaR_u(L) du.$$

We observe that the ES computes the average over all VaR with confidence levels $u \geq \alpha$. Hence it does contain information on the severity of potential losses. Also, one can show that $ES_\alpha \geq VaR_\alpha$.

Expected shortfall(ES)

Example

Suppose that $L \sim N(\mu, \sigma^2)$. Then, for $\alpha \in (0, 1)$, we get

$$ES_\alpha = \frac{1}{1-\alpha} \int_\alpha^1 VaR_u(L) du = \frac{1}{1-\alpha} \int_\alpha^1 (\mu + \sigma \Phi^{-1}(u)) du$$

$$= \cdots = \mu + \sigma \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}.$$

A similar result can be derived in the case of a Student t distributed loss function.

Case study in R

For an empirical example, we revisit the time series of the Google stock price we studied in yesterday's problem class: Recall that the data consists of daily closing prices (in USD) for the Google stock price recorded over a two year time period from 01 January 2020 to 22 June 2022.

Case study in R

For an empirical example, we revisit the time series of the Google stock price we studied in yesterday's problem class: Recall that the data consists of daily closing prices (in USD) for the Google stock price recorded over a two year time period from 01 January 2020 to 22 June 2022.
Suppose that the daily closing price is denoted by S_t. We compute the daily log returns as follows:

$$r_t = r(t,1) = \log(S_{t+1}) - \log(S_t).$$

We then define the process of the losses as $L_t = -r_t$.

Case study in R

For an empirical example, we revisit the time series of the Google stock price we studied in yesterday's problem class: Recall that the data consists of daily closing prices (in USD) for the Google stock price recorded over a two year time period from 01 January 2020 to 22 June 2022.
Suppose that the daily closing price is denoted by S_t. We compute the daily log returns as follows:

$$r_t = r(t, 1) = \log(S_{t+1}) - \log(S_t).$$

We then define the process of the losses as $L_t = -r_t$. Recall that in the problem class we found that the NIG distribution describes the log returns best. Hence we compute the VaR and ES based on the estimated NIG distribution and for comparison also for the estimated Gaussian distribution:

Advice: Use more advanced methodology in practice!

In the above computations, we treated the process of losses $(L_t)_{t\in\mathbb{Z}}$ as being i.i.d. random variables.

Advice: Use more advanced methodology in practice!

In the above computations, we treated the process of losses $(L_t)_{t\in\mathbb{Z}}$ as being i.i.d. random variables. A more flexible approach is to consider a (stationary) model of the form

$$L_t = \mu_t + \sigma_t Z_t,$$

where (Z_t) is a strict white noise process with zero mean and variance equal to 1. Moreover, we typically assume that μ_t and σ_t are known given the information up to time $t-1$.

Advice: Use more advanced methodology in practice!

In the above computations, we treated the process of losses $(L_t)_{t\in\mathbb{Z}}$ as being i.i.d. random variables. A more flexible approach is to consider a (stationary) model of the form

$$L_t = \mu_t + \sigma_t Z_t,$$

where (Z_t) is a strict white noise process with zero mean and variance equal to 1. Moreover, we typically assume that μ_t and σ_t are known given the information up to time $t-1$. One can then show that the VaR and ES given the information until time t are given by

$$VaR_\alpha^t = \mu_{t+1} + \sigma_{t+1} VaR_\alpha(Z),$$
$$ES_\alpha^t = \mu_{t+1} + \sigma_{t+1} ES_\alpha(Z).$$

Break :-)

Quiz: Rolling a die

Consider the experiment where we roll a standard six-sided fair die.

The sample space associated with this experiment is given by

a) $\Omega = \{1, 2, 3, 4, 5, 6\}$.

b) $\Omega = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$.

c) $\Omega = \{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$.

d) $\Omega = \{1, 6\}$.

e) None of the above.

Quiz: Rolling a die

Consider the experiment where we roll a standard six-sided fair die.

The sample space associated with this experiment is given by

a) $\Omega = \{1, 2, 3, 4, 5, 6\}$.

b) $\Omega = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$.

c) $\Omega = \{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$.

d) $\Omega = \{1, 6\}$.

e) None of the above.

Answer: a)

Quiz: Rolling a die

We set $A = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$,$B = \{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\}\}$, $C = \{\{2,3\},\{2,4\},\{2,5\},\{2,6\}\}$, $D = \{\{3,4\},\{3,5\},\{3,6\},\{4,5\},\{4,6\},\{5,6\}\}$.
The set of all subsets of the sample space Ω is given by

a) $\mathcal{F} = \{1,2,3,4,5,6\}$

b) $\mathcal{F} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$

c) $\mathcal{F} = A \cup B \cup C \cup D$.

d) $\mathcal{F} = \{\emptyset, \Omega\}$.

e) None of the above.

Quiz: Rolling a die

We set $A = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$,$B = \{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\}\}$, $C = \{\{2,3\},\{2,4\},\{2,5\},\{2,6\}\}$, $D = \{\{3,4\},\{3,5\},\{3,6\},\{4,5\},\{4,6\},\{5,6\}\}$.

The set of all subsets of the sample space Ω is given by

a) $\mathcal{F} = \{1,2,3,4,5,6\}$

b) $\mathcal{F} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}$

c) $\mathcal{F} = A \cup B \cup C \cup D$.

d) $\mathcal{F} = \{\emptyset, \Omega\}$.

e) None of the above.

Answer: e)

Quiz: Rolling a die

The probability that the outcome of the experiment is odd is given by

a) $\mathrm{P}(1,3,5) = 1/6$.
b) $\mathrm{P}(1,3,5) = 1/3$.
c) $\mathrm{P}(1,3,5) = 1/2$.
d) $\mathrm{P}(1,3,5) = 1/4$.
e) None of the above.

Quiz: Rolling a die

The probability that the outcome of the experiment is odd is given by

a) $\mathrm{P}(1,3,5) = 1/6$.
b) $\mathrm{P}(1,3,5) = 1/3$.
c) $\mathrm{P}(1,3,5) = 1/2$.
d) $\mathrm{P}(1,3,5) = 1/4$.
e) None of the above.

Answer: c)

Quiz: Time series

Suppose you want to plot the correlogram of a time series. Which R function is best suited for the task?

a) The `auto.arima` function from the `forecast` package.

b) The `stl` function from the `stats` package.

c) The `acf` function from the `stats` package.

d) The `median` function from the `stats` package.

e) The `window` function from the `stats` package.

Quiz: Time series

Suppose you want to plot the correlogram of a time series. Which R function is best suited for the task?

a) The `auto.arima` function from the `forecast` package.

b) The `stl` function from the `stats` package.

c) The `acf` function from the `stats` package.

d) The `median` function from the `stats` package.

e) The `window` function from the `stats` package.

Answer: c)

Quiz: Time series

Let $(\epsilon_t)_{t\in\mathbb{Z}}$ denote a white noise process with zero mean. Which of the following stochastic processes $(X_t)_{t\in\mathbb{Z}}$ describes an MA(1) process?

a) $X_t = \epsilon_t + 0.4\epsilon_{t-1}$.

b) $X_t = 0.3X_{t-1} + \epsilon_t + 0.4\epsilon_{t-1}$.

c) $X_t = \epsilon_t + 0.4\epsilon_{t-1} + 0.2\epsilon_{t-2}$.

d) $X_t = 0.3X_{t-1} + \epsilon_t$.

e) None of the above.

Quiz: Time series

Let $(\epsilon_t)_{t\in\mathbb{Z}}$ denote a white noise process with zero mean. Which of the following stochastic processes $(X_t)_{t\in\mathbb{Z}}$ describes an MA(1) process?

a) $X_t = \epsilon_t + 0.4\epsilon_{t-1}$.

b) $X_t = 0.3X_{t-1} + \epsilon_t + 0.4\epsilon_{t-1}$.

c) $X_t = \epsilon_t + 0.4\epsilon_{t-1} + 0.2\epsilon_{t-2}$.

d) $X_t = 0.3X_{t-1} + \epsilon_t$.

e) None of the above.

Answer: a)