Parameter point estimators

Lecturer:

Francesco Dolce

Department of Applied Mathematics

Faculty of Information Technology

Czech Technical University in Prague

Probability and Statistics

BIE-PST, WS 2023/24, Lecture 9

Content

- **Probability theory:**
  - Events, probability, conditional probability, Bayes' Theorem, independence of events.
  - Random variables, distribution function, functions of random variables, characteristics of random variables: expected value, variance, moments, generating function, quantiles, critical values, important discrete and continuous distributions.
  - Random vectors, joint and marginal distributions, functions of random vectors, independence of random variables, conditional distribution, conditional expected value, covariance and correlation.
  - Markov's and Chebyshev's inequality, weak law of large numbers, strong law of large numbers, Central limit theorem.
- **Mathematical statistics:**
  - Point estimators, sample mean, sample variance, properties of point estimators, Maximum likelihood method.
  - Interval estimators, hypothesis testing, one-sided vs. two-sided alternatives, linear regression, estimators of regression parameters, testing of linear model.

Recap

- A **random variable** X is a measurable function, which assigns real values to the outcomes of a random experiment.
- The **distribution** of X gives the information of the probabilities of its values and is uniquely given by the **distribution function**:

$$F_X(x) = \mathrm{P}(X \leq x).$$

- Often we observe a **sequence of independent and identically distributed** (`i.i.d.`) random variables $X_1, X_2, \dots$. Let each of them have expectation μ and variance σ^2.
- If we denote the **sum** and the arithmetic **mean** of n such variables as

$$S_n = \sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

we get that

$$\mathrm{E}\, S_n = n \cdot \mu, \qquad \mathrm{E}\, \bar{X}_n = \mu,$$
$$\operatorname{var} S_n = n \cdot \sigma^2, \qquad \operatorname{var} \bar{X}_n = \sigma^2 / n.$$

- According to the **law of large numbers**, the arithmetic mean converges to the expectation, provided that it exists:

$$\bar{X}_n \xrightarrow{n \to \infty} \mu.$$

Introduction to statistics

So far we have dealt with **probabilistic** problems with known parameters. For example if we have a box with r red and b blue balls, we can:

- find the probability of drawing a blue ball,
- find the probability of drawing a certain number of blue balls in three draws with or without replacement,
- find the expected number of blue balls in 10 draws with replacement,
- make statements about a sequence of 1000 draws,
- etc.

Introduction to statistics

So far we have dealt with **probabilistic** problems with known parameters. For example if we have a box with r red and b blue balls, we can:

- find the probability of drawing a blue ball,
- find the probability of drawing a certain number of blue balls in three draws with or without replacement,
- find the expected number of blue balls in 10 draws with replacement,
- make statements about a sequence of 1000 draws,
- etc.

Now we will deal with **statistical** problems. For example if we have a box with an unknown number of red and blue balls, we can take a sample and:

- estimate the proportion of red and blue balls,
- test whether there are 50% of blue balls or more,
- test whether the red/blue proportion is the same among two separate boxes,
- etc.

Introduction to statistics

Probability theory deals with mathematical models of processes (experiments, tests, etc.) with random results. These models are then utilized for prediction of possible outcomes, i.e., we determine probabilities of events, distributions and expected values of random variables, etc.

Introduction to statistics

Probability theory deals with mathematical models of processes (experiments, tests, etc.) with random results. These models are then utilized for prediction of possible outcomes, i.e., we determine probabilities of events, distributions and expected values of random variables, etc.

Mathematical statistics proceeds, to some extent, reversely. On the grounds of real outcomes we choose an appropriate model and estimate its parameters. Then we can test hypotheses about these parameters and verify how well does the model fit the data.

Random sample

Statistics uses specific terminology.

Definition

An n-tuple of independent and identically distributed random variables (i.i.d.) $X_1, \ldots, X_n$ with distribution function F is called a **random sample** from the distribution F.

Random sample

Statistics uses specific terminology.

Definition

An n-tuple of independent and identically distributed random variables (i.i.d.) $X_1, \ldots, X_n$ with distribution function F is called a **random sample** from the distribution F.

Examples

- Measurement of a given variable in n independent repetitions of some experiment.
- Time to execute an algorithm in n repeated runs.
- Measurement of body height of n different people.

Random sample

Statistics uses specific terminology.

Definition

An n-tuple of independent and identically distributed random variables (i.i.d.) $X_1, \ldots, X_n$ with distribution function F is called a **random sample** from the distribution F.

Examples

- Measurement of a given variable in n independent repetitions of some experiment.
- Time to execute an algorithm in n repeated runs.
- Measurement of body height of n different people.

Definition

The **random sample realization** (random vector of observations or simply **data**) is an n-tuple of particular observed values $x_1, \ldots, x_n$.

Steps of statistical inference

Consider a random sample from an unknown distribution. On the grounds of measured data (random sample realizations) we want to learn as much as possible about the underlying distribution.

Steps of statistical inference

Consider a random sample from an unknown distribution. On the grounds of measured data (random sample realizations) we want to learn as much as possible about the underlying distribution.

Typical steps of statistical inference:

- **Estimate the shape of the distribution** – restrict the inference to a family of distributions F_θ with a parameter θ. This can follow from prior knowledge, intuition or experience.

Steps of statistical inference

Consider a random sample from an unknown distribution. On the grounds of measured data (random sample realizations) we want to learn as much as possible about the underlying distribution.

Typical steps of statistical inference:

- **Estimate the shape of the distribution** – restrict the inference to a family of distributions F_θ with a parameter θ. This can follow from prior knowledge, intuition or experience.
- **Estimate the parameters of the distribution**

Steps of statistical inference

Consider a random sample from an unknown distribution. On the grounds of measured data (random sample realizations) we want to learn as much as possible about the underlying distribution.

Typical steps of statistical inference:

- **Estimate the shape of the distribution** – restrict the inference to a family of distributions F_θ with a parameter θ. This can follow from prior knowledge, intuition or experience.
- **Estimate the parameters of the distribution**
  - **Point estimation** – determine the “most probable” value of θ.

Steps of statistical inference

Consider a random sample from an unknown distribution. On the grounds of measured data (random sample realizations) we want to learn as much as possible about the underlying distribution.

Typical steps of statistical inference:

- **Estimate the shape of the distribution** – restrict the inference to a family of distributions F_θ with a parameter θ. This can follow from prior knowledge, intuition or experience.
- **Estimate the parameters of the distribution**
  - **Point estimation** – determine the “most probable” value of θ.
  - **Interval estimation** – determine an interval (region) in which θ lies with a given large probability.

Steps of statistical inference

Consider a random sample from an unknown distribution. On the grounds of measured data (random sample realizations) we want to learn as much as possible about the underlying distribution.

Typical steps of statistical inference:

- **Estimate the shape of the distribution** – restrict the inference to a family of distributions F_θ with a parameter θ. This can follow from prior knowledge, intuition or experience.
- **Estimate the parameters of the distribution**
  - **Point estimation** – determine the "most probable" value of θ.
  - **Interval estimation** – determine an interval (region) in which θ lies with a given large probability.
- **Verify the model – hypothesis testing**

Steps of statistical inference

Consider a random sample from an unknown distribution. On the grounds of measured data (random sample realizations) we want to learn as much as possible about the underlying distribution.

Typical steps of statistical inference:

- **Estimate the shape of the distribution** – restrict the inference to a family of distributions F_θ with a parameter θ. This can follow from prior knowledge, intuition or experience.
- **Estimate the parameters of the distribution**
  - **Point estimation** – determine the “most probable” value of θ.
  - **Interval estimation** – determine an interval (region) in which θ lies with a given large probability.
- **Verify the model – hypothesis testing**
  - **Goodness-of-fit tests** – we verify hypothesis about the shape of the probability distribution –e.g., whether the investigated variable has the normal distribution.

Steps of statistical inference

Consider a random sample from an unknown distribution. On the grounds of measured data (random sample realizations) we want to learn as much as possible about the underlying distribution.

Typical steps of statistical inference:

- **Estimate the shape of the distribution** – restrict the inference to a family of distributions F_θ with a parameter θ. This can follow from prior knowledge, intuition or experience.
- **Estimate the parameters of the distribution**
  - **Point estimation** – determine the “most probable” value of θ.
  - **Interval estimation** – determine an interval (region) in which θ lies with a given large probability.
- **Verify the model – hypothesis testing**
  - **Goodness-of-fit tests** – we verify hypothesis about the shape of the probability distribution –e.g., whether the investigated variable has the normal distribution.
  - **Parametric tests** – we state a hypothesis about the parameter θ (e.g., $\theta = 0$) and on the grounds of measured data we try to decide whether this hypothesis can be true or not.

Estimation of the shape of the distribution – model selection

The distribution of an investigated random variable usually may not be absolutely arbitrary.

Based on previous experience, intuition or the type of underlying data we can often

- determine whether the variable is discrete or continuous;

Estimation of the shape of the distribution – model selection

The distribution of an investigated random variable usually may not be absolutely arbitrary.

Based on previous experience, intuition or the type of underlying data we can often

- determine whether the variable is discrete or continuous;
- approximate the shape of the distribution (e.g., exponential, normal, etc.);

Estimation of the shape of the distribution – model selection

The distribution of an investigated random variable usually may not be absolutely arbitrary.

Based on previous experience, intuition or the type of underlying data we can often

- determine whether the variable is discrete or continuous;
- approximate the shape of the distribution (e.g., exponential, normal, etc.);
- establish other possible determining properties (e.g., range of values, zero expectation, etc.).

Estimation of the shape of the distribution – model selection

The distribution of an investigated random variable usually may not be absolutely arbitrary.

Based on previous experience, intuition or the type of underlying data we can often

- determine whether the variable is discrete or continuous;
- approximate the shape of the distribution (e.g., exponential, normal, etc.);
- establish other possible determining properties (e.g., range of values, zero expectation, etc.).

This information leads us to a choice of a particular **model**, thus to the

Estimation of the shape of the distribution – model selection

The distribution of an investigated random variable usually may not be absolutely arbitrary.

Based on previous experience, intuition or the type of underlying data we can often

- determine whether the variable is discrete or continuous;
- approximate the shape of the distribution (e.g., exponential, normal, etc.);
- establish other possible determining properties (e.g., range of values, zero expectation, etc.).

This information leads us to a choice of a particular **model**, thus to the

- choice of parametric distribution family $\{F_\theta(x) | \theta \in \Theta\}$, where Θ is a set of all possible values of parameter θ;

Estimation of the shape of the distribution – model selection

The distribution of an investigated random variable usually may not be absolutely arbitrary.

Based on previous experience, intuition or the type of underlying data we can often

- determine whether the variable is discrete or continuous;
- approximate the shape of the distribution (e.g., exponential, normal, etc.);
- establish other possible determining properties (e.g., range of values, zero expectation, etc.).

This information leads us to a choice of a particular **model**, thus to the

- choice of parametric distribution family $\{F_\theta(x) | \theta \in \Theta\}$, where Θ is a set of all possible values of parameter θ;
- and the assumption that our random sample is governed by distribution from this family.

Examples of possible models

- **Bernoulli distribution** – tossing with an unknown coin

$$\{\mathsf{Be}(p) \mid p \in [0,1]\}$$

Parameter $\theta = p$ and $\Theta = [0,1]$.

Examples of possible models

- **Bernoulli distribution** – tossing with an unknown coin

$$\{\mathsf{Be}(p) \mid p \in [0,1]\}$$

  Parameter $\theta = p$ and $\Theta = [0,1]$.

- **Exponential distribution** – times between two incoming request on a database server

$$\{\mathsf{Exp}(\lambda) \mid \lambda \in (0,+\infty)\}$$

  Parameter $\theta = \lambda$ and $\Theta = (0,+\infty)$.

Examples of possible models

- **Bernoulli distribution** – tossing with an unknown coin

$$\{\mathsf{Be}(p) \mid p \in [0,1]\}$$

Parameter $\theta = p$ and $\Theta = [0,1]$.

- **Exponential distribution** – times between two incoming request on a database server

$$\{\mathsf{Exp}(\lambda) \mid \lambda \in (0,+\infty)\}$$

Parameter $\theta = \lambda$ and $\Theta = (0,+\infty)$.

- **Normal distribution** – results of an IQ test in a given population

$$\{\mathsf{N}(\mu,\sigma^2) \mid \mu \in (-\infty,+\infty), \sigma^2 \in (0,+\infty)\}$$

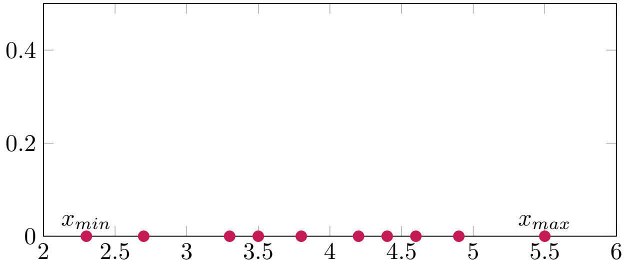
Two dimensional parameter $\theta = (\mu,\sigma^2)$ and $\Theta = (-\infty,+\infty) \times (0,+\infty)$.

Estimation of the shape of the distribution – histogram

The shape of the **density** can be estimated by the **histogram**:

Estimation of the shape of the distribution – histogram

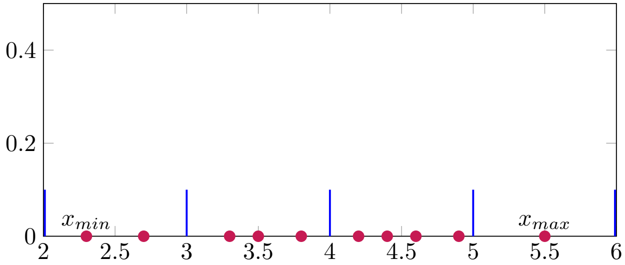
The shape of the **density** can be estimated by the **histogram**:



- Determine the data range.

Estimation of the shape of the distribution – histogram

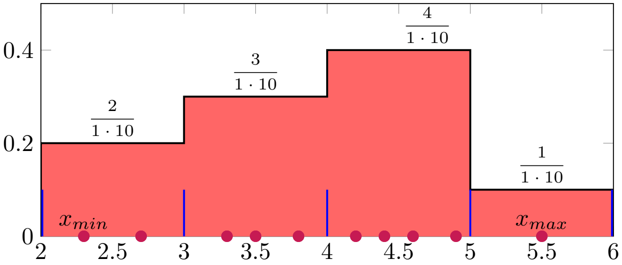
The shape of the **density** can be estimated by the **histogram**:



- Determine the data range.
- Choose a number of bins k and their size h (here $k = 4$ and $h = 1$).

Estimation of the shape of the distribution – histogram

The shape of the **density** can be estimated by the **histogram**:



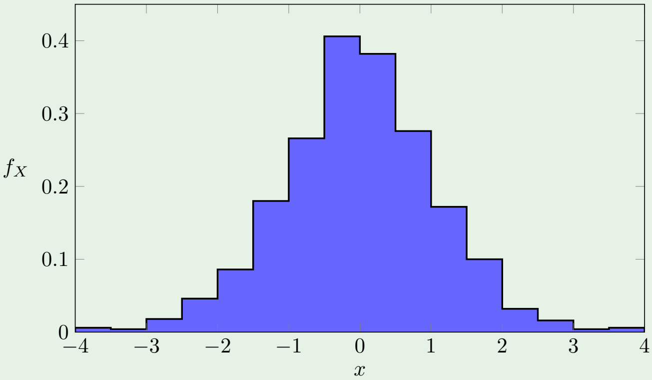
- Determine the data range.
- Choose a number of bins k and their size h (here $k = 4$ and $h = 1$).
- Over each bin, plot a column of the size

$$\frac{\text{number of observation in bin}}{h \cdot \text{total number of observations}} \overset{\text{denote}}{=} \frac{m_i}{h \cdot n}.$$

Estimation of the shape of the distribution

Example

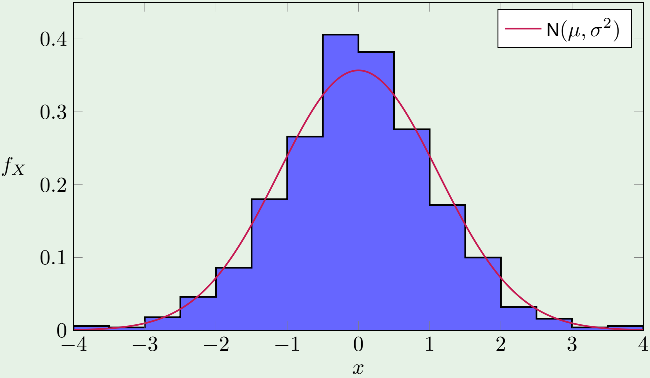
We measured 1000 values from an unknown distribution. The histogram of these values is:



Estimation of the shape of the distribution

Example

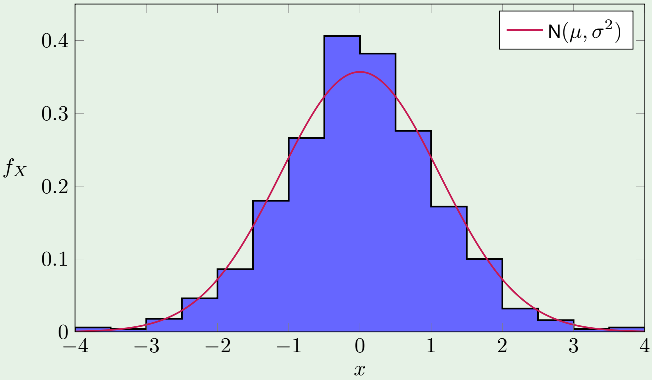
We measured 1000 values from an unknown distribution. The histogram of these values is:



Estimation of the shape of the distribution

Example

We measured 1000 values from an unknown distribution. The histogram of these values is:



We can assume that we deal with values from the normal distribution with unknown parameters μ and σ^2.

Empirical distribution function

The shape of the **distribution function** can be estimated by the **empirical distribution function**:

$$F_n(x, X_1, \dots, X_n) = F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

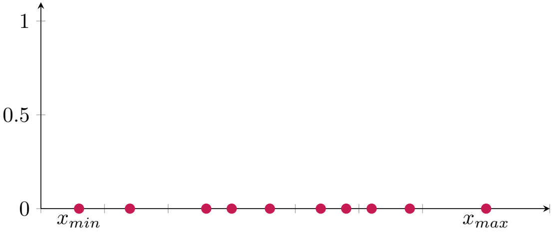
In other words, the probability that the random variable in question is less than or equal x can be estimated by the proportion of data points which are less than or equal to x.

Empirical distribution function

The shape of the **distribution function** can be estimated by the **empirical distribution function**:

$$F_n(x, X_1, \dots, X_n) = F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}.$$

In other words, the probability that the random variable in question is less than or equal x can be estimated by the proportion of data points which are less than or equal to x.

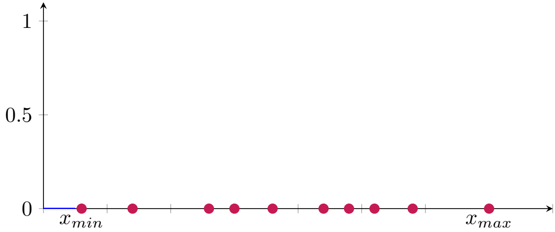


Empirical distribution function

The shape of the **distribution function** can be estimated by the **empirical distribution function**:

$$F_n(x, X_1, \ldots, X_n) = F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

In other words, the probability that the random variable in question is less than or equal x can be estimated by the proportion of data points which are less than or equal to x.

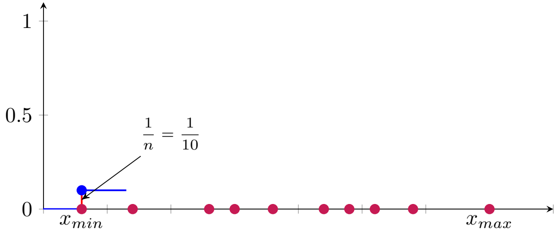


Empirical distribution function

The shape of the **distribution function** can be estimated by the **empirical distribution function**:

$$F_n(x, X_1, \dots, X_n) = F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

In other words, the probability that the random variable in question is less than or equal x can be estimated by the proportion of data points which are less than or equal to x.

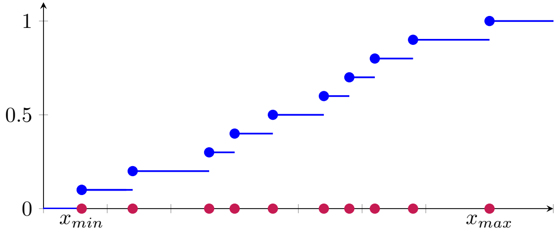


Empirical distribution function

The shape of the **distribution function** can be estimated by the **empirical distribution function**:

$$F_n(x, X_1, \ldots, X_n) = F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

In other words, the probability that the random variable in question is less than or equal x can be estimated by the proportion of data points which are less than or equal to x.

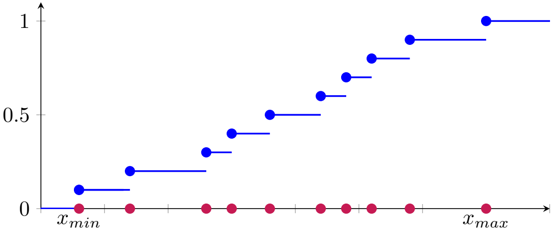


Empirical distribution function

The shape of the **distribution function** can be estimated by the **empirical distribution function**:

$$F_n(x, X_1, \dots, X_n) = F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

In other words, the probability that the random variable in question is less than or equal x can be estimated by the proportion of data points which are less than or equal to x.

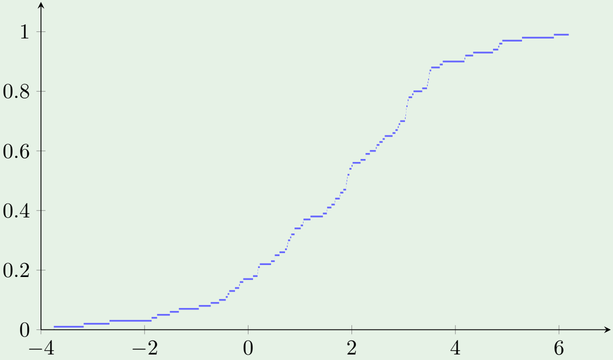


✓ The empirical distribution function is a piecewise constant function with jumps of size $\frac{1}{n}$ in the observed data points.

Empirical distribution function

Example

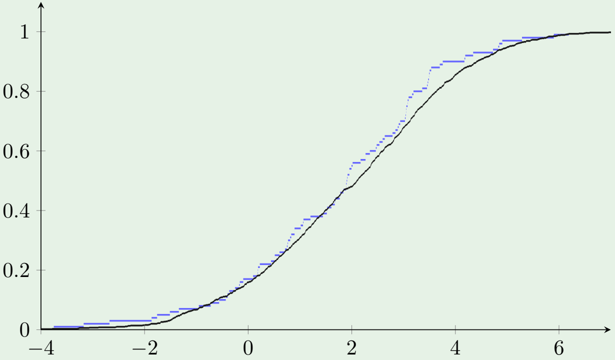
We measured 100 and 1000 values from an unknown distribution. The empirical distribution functions are:



Empirical distribution function

Example

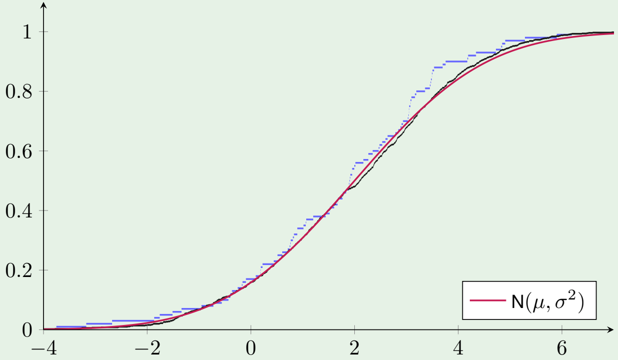
We measured 100 and 1000 values from an unknown distribution. The empirical distribution functions are:



Empirical distribution function

Example

We measured 100 and 1000 values from an unknown distribution. The empirical distribution functions are:



We can assume that we deal with values from the normal distribution with unknown parameters μ and σ^2.

Estimating the shape of the distribution – example

Example – waiting for a bus

Every morning we measure the time which we spend waiting for a bus on our way to school. After 15 days, we have observed the following data (in minutes, sorted):

0.1	0.3	0.5	0.7	1.0
1.9	2.8	3.4	3.5	3.8
5.3	7.7	8.6	8.7	11.1

Suppose that the waiting times form a random sample $(X_1, \ldots, X_{15})$ from an unknown distribution *Find the histogram and the empirical distribution function of this distribution.*

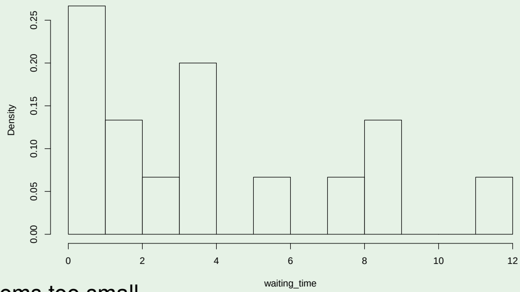
Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. If we take the bandwidth h too small or too large, the histogram may be inaccurate:

0.1	0.3	0.5	0.7	1.0	1.9	2.8	3.4	3.5	3.8	5.3	7.7	8.6	8.7	11.1

```
>hist(waiting_time,prob=T,breaks=12)
```



The bandwidth seems too small.

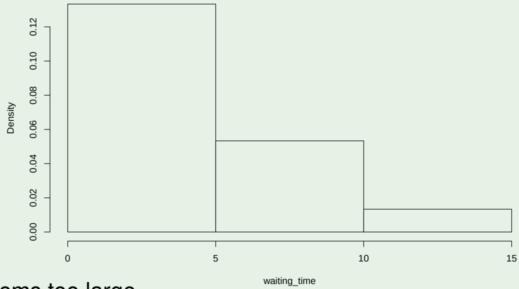
Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. If we take the bandwidth h too small or too large, the histogram may be inaccurate:

0.1	0.3	0.5	0.7	1.0	1.9	2.8	3.4	3.5	3.8	5.3	7.7	8.6	8.7	11.1

```
>hist(waiting_time,prob=T,breaks=2)
```



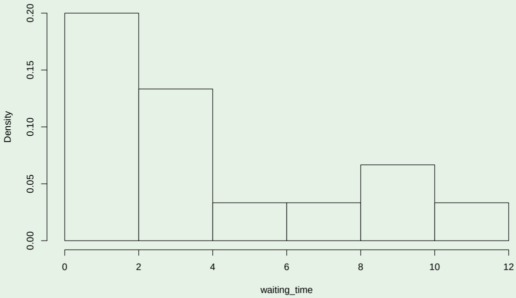
The bandwidth seems too large.

Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. It seems reasonable to divide them into six parts, each covering two minutes. Each data point constitutes $\frac{1}{h \cdot n} = \frac{1}{2 \cdot 15} = 0.0\bar{3}3$:

```
>hist(waiting_time,prob=T)
```



The histogram might seem similar to the exponential distribution.

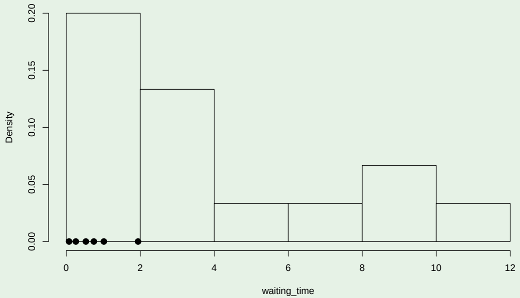
Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. It seems reasonable to divide them into six parts, each covering two minutes. Each data point constitutes $\frac{1}{h \cdot n} = \frac{1}{2 \cdot 15} = 0.0\bar{3}\bar{3}$:

0.1 0.3 0.5 0.7 1.0 1.9

```
>hist(waiting_time,prob=T)
```



The histogram might seem similar to the exponential distribution.

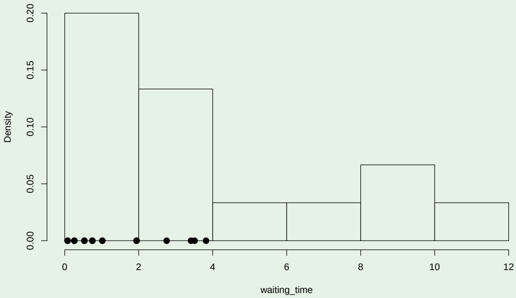
Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. It seems reasonable to divide them into six parts, each covering two minutes. Each data point constitutes $\frac{1}{h \cdot n} = \frac{1}{2 \cdot 15} = 0.0\bar{3}\bar{3}$:

0.1 0.3 0.5 0.7 1.0 1.9 2.8 3.4 3.5 3.8

```
>hist(waiting_time,prob=T)
```



The histogram might seem similar to the exponential distribution.

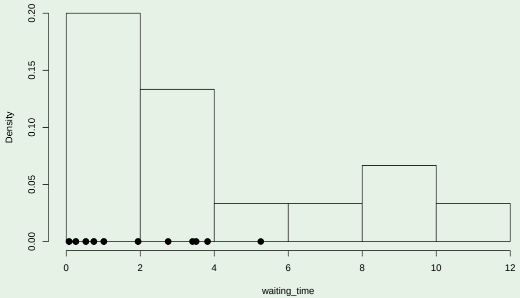
Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. It seems reasonable to divide them into six parts, each covering two minutes. Each data point constitutes $\frac{1}{h \cdot n} = \frac{1}{2 \cdot 15} = 0.0\bar{3}\bar{3}$:

0.1 0.3 0.5 0.7 1.0 1.9 2.8 3.4 3.5 3.8 5.3

```
>hist(waiting_time,prob=T)
```



The histogram might seem similar to the exponential distribution.

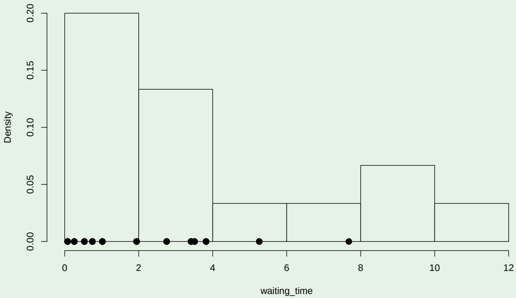
Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. It seems reasonable to divide them into six parts, each covering two minutes. Each data point constitutes $\frac{1}{h \cdot n} = \frac{1}{2 \cdot 15} = 0.0\bar{3}\bar{3}$:

0.1	0.3	0.5	0.7	1.0	1.9	2.8	3.4	3.5	3.8	5.3	7.7

```
>hist(waiting_time,prob=T)
```



The histogram might seem similar to the exponential distribution.

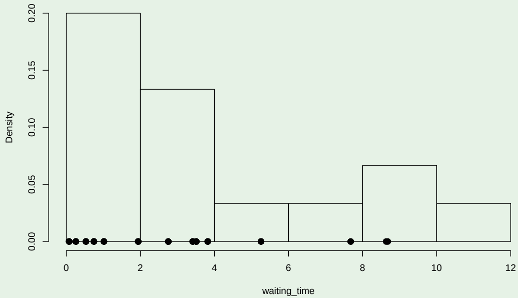
Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. It seems reasonable to divide them into six parts, each covering two minutes. Each data point constitutes $\frac{1}{h \cdot n} = \frac{1}{2 \cdot 15} = 0.0\bar{3}\bar{3}$:

0.1	0.3	0.5	0.7	1.0	1.9	2.8	3.4	3.5	3.8	5.3	7.7	8.6	8.7

```
>hist(waiting_time,prob=T)
```



The histogram might seem similar to the exponential distribution.

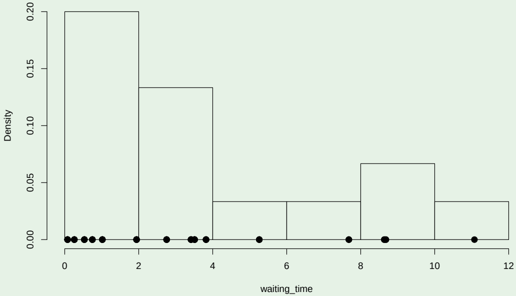
Estimating the shape of the distribution – example

Example – waiting for a bus – histogram

The data are in the interval $[0, 12]$. It seems reasonable to divide them into six parts, each covering two minutes. Each data point constitutes $\frac{1}{h \cdot n} = \frac{1}{2 \cdot 15} = 0.0\bar{3}\bar{3}$:

0.1	0.3	0.5	0.7	1.0	1.9	2.8	3.4	3.5	3.8	5.3	7.7	8.6	8.7	11.1

```
>hist(waiting_time,prob=T)
```



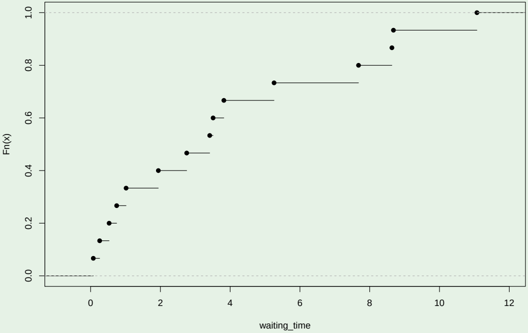
The histogram might seem similar to the exponential distribution.

Estimating the shape of the distribution – example

Example – waiting for a bus – empirical distribution function

We proceed from the left and add a jump of $1/15$ at each data point encountered:

```
>plot(ecdf(waiting_time))
```

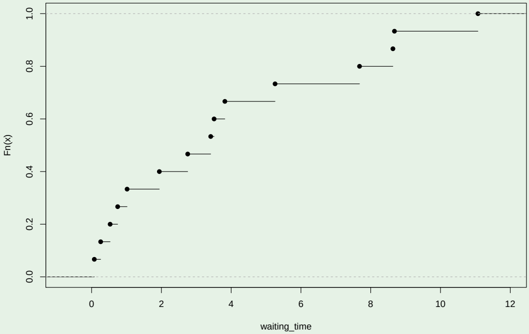


Estimating the shape of the distribution – example

Example – waiting for a bus – empirical distribution function

We proceed from the left and add a jump of $1/15$ at each data point encountered:

```
>plot(ecdf(waiting_time))
```



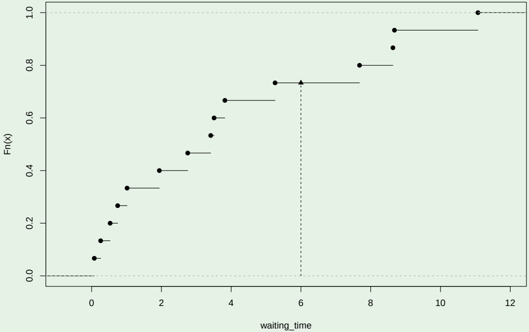
Now we can estimate probabilities of the type $\mathrm{P}(X \leq x)$ using $F_n(x)$.

Estimating the shape of the distribution – example

Example – waiting for a bus – empirical distribution function

We proceed from the left and add a jump of $1/15$ at each data point encountered:

```
>plot(ecdf(waiting_time))
```



Now we can estimate probabilities of the type $\mathrm{P}(X \leq x)$ using $F_n(x)$.

The probability that we do not need to wait for more than six minutes is estimated as $F_n(6) = 11/15 \doteq 0.733$, which is the proportion of data points less than or equal to 6.

Estimating the shape of the distribution – quantiles

The **quantiles** q_α divide the population so that there are $\alpha\%$ of values under the α-quantile and $(1-\alpha)\%$ above. The 50%-quantile is called the **median** and divides the population into two equally large parts with respect to probability.

Estimating the shape of the distribution – quantiles

The **quantiles** q_α divide the population so that there are $\alpha\%$ of values under the α-quantile and $(1-\alpha)\%$ above. The 50%-quantile is called the **median** and divides the population into two equally large parts with respect to probability.

If we denote the ordered data as

$$\left(x_{(1)}, x_{(2)}, \dots, x_{(n)}\right),$$

the $\alpha\%$-quantile can be estimated as $x_{(\lceil n\alpha \rceil)}$. This is then the inverse of the empirical distribution function.

Estimating the shape of the distribution – quantiles

The **quantiles** q_α divide the population so that there are $\alpha\%$ of values under the α-quantile and $(1-\alpha)\%$ above. The 50%-quantile is called the **median** and divides the population into two equally large parts with respect to probability.

If we denote the ordered data as

$$\left(x_{(1)}, x_{(2)}, \dots, x_{(n)}\right),$$

the $\alpha\%$-quantile can be estimated as $x_{(\lceil n\alpha \rceil)}$. This is then the inverse of the empirical distribution function.

The median $q_{0.5}$ can then be estimated as the **middle value** of the ordered data, $x_{(\lceil \frac{n}{2} \rceil)}$. If there is an even number of data points, some software estimates the median as the average of $x_{(\frac{n}{2})}$ and $x_{(\frac{n}{2}+1)}$.

Estimating the shape of the distribution – quantiles

The **quantiles** q_α divide the population so that there are $\alpha\%$ of values under the α-quantile and $(1-\alpha)\%$ above. The 50%-quantile is called the **median** and divides the population into two equally large parts with respect to probability.

If we denote the ordered data as

$$\left(x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right),$$

the $\alpha\%$-quantile can be estimated as $x_{(\lceil n\alpha \rceil)}$. This is then the inverse of the empirical distribution function.

The median $q_{0.5}$ can then be estimated as the **middle value** of the ordered data, $x_{(\lceil \frac{n}{2} \rceil)}$. If there is an even number of data points, some software estimates the median as the average of $x_{(\frac{n}{2})}$ and $x_{(\frac{n}{2}+1)}$.

Example – waiting for a bus – median

Estimate the median of the time spent waiting for the bus using the observed data:

0.1 0.3 0.5 0.7 1.0 1.9 2.8 **3.4** 3.5 3.8 5.3 7.7 8.6 8.7 11.1

Estimating the shape of the distribution – quantiles

The **quantiles** q_α divide the population so that there are $\alpha\%$ of values under the α-quantile and $(1-\alpha)\%$ above. The 50%-quantile is called the **median** and divides the population into two equally large parts with respect to probability.

If we denote the ordered data as

$$\left(x_{(1)}, x_{(2)}, \dots, x_{(n)}\right),$$

the $\alpha\%$-quantile can be estimated as $x_{(\lceil n\alpha \rceil)}$. This is then the inverse of the empirical distribution function.

The median $q_{0.5}$ can then be estimated as the **middle value** of the ordered data, $x_{(\lceil \frac{n}{2} \rceil)}$. If there is an even number of data points, some software estimates the median as the average of $x_{(\frac{n}{2})}$ and $x_{(\frac{n}{2}+1)}$.

Example – waiting for a bus – median

Estimate the median of the time spent waiting for the bus using the observed data:

0.1	0.3	0.5	0.7	1.0	1.9	2.8	**3.4**	3.5	3.8	5.3	7.7	8.6	8.7	11.1

The median is estimated as the middle observed value. Therefore with a probability of about 50% we will be waiting for the bus for less than 3.4 minutes and also for more than 3.4 minutes.

Point estimators

From the measured data we can estimate the real value of the parameter θ using a **point estimator**:

Point estimators

From the measured data we can estimate the real value of the parameter θ using a **point estimator**:

Definition

A **point estimator** of a parameter θ is a function $\hat{\theta}_n(X_1, \ldots, X_n)$ of the random sample which does not depend on θ.

Point estimators

From the measured data we can estimate the real value of the parameter θ using a **point estimator**:

Definition

A **point estimator** of a parameter θ is a function $\hat{\theta}_n(X_1, \ldots, X_n)$ of the random sample which does not depend on θ.

Notes:

- A point estimator is an example of a statistic. A **statistic** is an arbitrary function of the random sample which does not depend on the parameter θ.

Point estimators

From the measured data we can estimate the real value of the parameter θ using a **point estimator**:

Definition

A **point estimator** of a parameter θ is a function $\hat{\theta}_n(X_1, \ldots, X_n)$ of the random sample which does not depend on θ.

Notes:

- A point estimator is an example of a statistic. A **statistic** is an arbitrary function of the random sample which does not depend on the parameter θ.
- Generally, we can also construct a point estimator of a function of a parameter $g(\theta)$.

Point estimators

From the measured data we can estimate the real value of the parameter θ using a **point estimator**:

Definition

A **point estimator** of a parameter θ is a function $\hat{\theta}_n(X_1, \dots, X_n)$ of the random sample which does not depend on θ.

Notes:

- A point estimator is an example of a statistic. A **statistic** is an arbitrary function of the random sample which does not depend on the parameter θ.
- Generally, we can also construct a point estimator of a function of a parameter $g(\theta)$.
- A typical example is $g(\lambda) = \dfrac{1}{\lambda} = \operatorname{E} X$ for the exponential distribution.

Most common point estimators

- **Sample mean** – point estimator of the expectation $\mathrm{E}\,X$:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Most common point estimators

- **Sample mean** – point estimator of the expectation $\mathrm{E}\,X$:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- **Sample variance** – point estimator the of variance $\operatorname{var} X$:

$$s_n^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Most common point estimators

- **Sample mean** – point estimator of the expectation $\mathrm{E}\,X$:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- **Sample variance** – point estimator the of variance $\operatorname{var} X$:

$$s_n^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

- **Sample standard deviation** – point estimator of the standard deviation $\sqrt{\operatorname{var} X}$:

$$s_n = \sqrt{s_n^2}.$$

Most common point estimators

- **Sample mean** – point estimator of the expectation $\mathrm{E}\,X$:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- **Sample variance** – point estimator the of variance $\operatorname{var} X$:

$$s_n^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

- **Sample standard deviation** – point estimator of the standard deviation $\sqrt{\operatorname{var} X}$:

$$s_n = \sqrt{s_n^2}.$$

- **k-th sample moment** – point estimator of k-th moment $\mu_k = \mathrm{E}\,X^k$:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Most common point estimators

- **Sample covariance** – point estimator of the covariance $\operatorname{cov}(X, Y)$:

$$s_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

Most common point estimators

- **Sample covariance** – point estimator of the covariance $\mathrm{cov}(X, Y)$:

$$s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

- **Sample correlation coefficient** – point estimator of the correlation coefficient $\rho(X, Y)$:

$$r_{X,Y} = r = \frac{s_{X,Y}}{s_X s_Y},$$

where s_X and s_Y are square roots of the sample variances of X and Y.

Properties of point estimators

A point estimator as a function of the random sample is itself also a random variable with some distribution which obviously depends on the parameter θ.

Properties of point estimators

A point estimator as a function of the random sample is itself also a random variable with some distribution which obviously depends on the parameter θ.

A "good estimator" $\hat{\theta}_n$ should be in some way close to the true value of θ for all values θ and for all realizations of the random sample from F_θ.

Properties of point estimators

A point estimator as a function of the random sample is itself also a random variable with some distribution which obviously depends on the parameter θ.

A "good estimator" $\hat{\theta}_n$ should be in some way close to the true value of θ for all values θ and for all realizations of the random sample from F_θ.

Usually we want an estimator to be **unbiased**:

Definition

An estimator $\hat{\theta}_n$ of the parameter θ is called **unbiased** if

$$\mathrm{E}\,\hat{\theta}_n(X_1,\dots,X_n) = \theta \qquad \text{for all } \theta \in \Theta.$$

Properties of point estimators

A point estimator as a function of the random sample is itself also a random variable with some distribution which obviously depends on the parameter θ.

A "good estimator" $\hat{\theta}_n$ should be in some way close to the true value of θ for all values θ and for all realizations of the random sample from F_θ.

Usually we want an estimator to be **unbiased**:

Definition

An estimator $\hat{\theta}_n$ of the parameter θ is called **unbiased** if

$$\mathrm{E}\,\hat{\theta}_n(X_1, \ldots, X_n) = \theta \qquad \text{for all } \theta \in \Theta.$$

Unbiasedness means that an estimator does not have a systematic error, e.g., that it does not produce systematically larger or smaller values.

Properties of point estimators

The next desirable property of estimators is **consistency**:

Definition

An estimator $\hat{\theta}_n$ of the parameter θ is called **consistent** if for all $\theta \in \Theta$:

$$\hat{\theta}_n \xrightarrow{P} \theta \qquad \text{for} \quad n \to \infty.$$

Properties of point estimators

The next desirable property of estimators is **consistency**:

Definition

An estimator $\hat{\theta}_n$ of the parameter θ is called **consistent** if for all $\theta \in \Theta$:

$$\hat{\theta}_n \xrightarrow{P} \theta \qquad \text{for} \quad n \to \infty.$$

In other words if for all $\varepsilon > 0$: $\mathrm{P}(|\hat{\theta}_n(X_1, \ldots, X_n) - \theta| \geq \varepsilon) \to 0$. Consistency means that by choosing a large n, the error of the estimate will be sufficiently small.

Properties of point estimators

The next desirable property of estimators is **consistency**:

Definition

An estimator $\hat{\theta}_n$ of the parameter θ is called **consistent** if for all $\theta \in \Theta$:

$$\hat{\theta}_n \xrightarrow{P} \theta \qquad \text{for} \quad n \to \infty.$$

In other words if for all $\varepsilon > 0$: $\mathrm{P}(|\hat{\theta}_n(X_1, \ldots, X_n) - \theta| \geq \varepsilon) \to 0$. Consistency means that by choosing a large n, the error of the estimate will be sufficiently small.

Theorem

Let $\mathrm{E}\,\hat{\theta}_n^2 < +\infty$ for all n. If for $n \to +\infty$ it holds that

$$\mathrm{E}\,\hat{\theta}_n \to \theta \quad \textit{and} \quad \operatorname{var} \hat{\theta}_n \to 0,$$

*then $\hat{\theta}_n$ is a **consistent** estimator.*

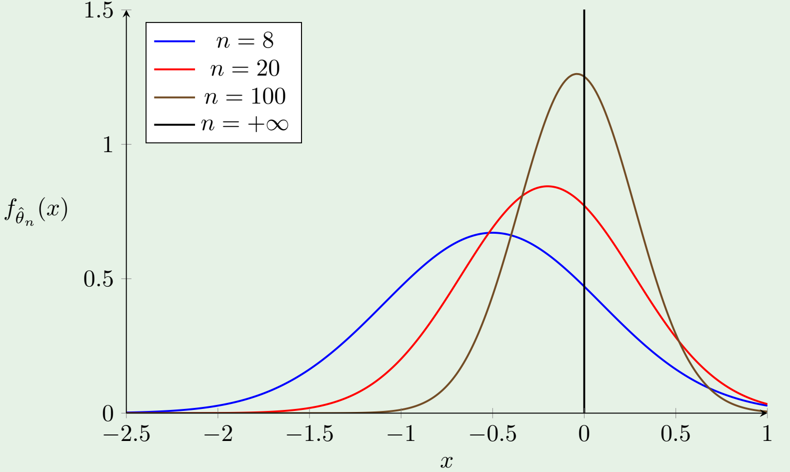
Proof

Proof can be found in bibliography. □

Estimator consistency

Example

Convergence of the densities of a consistent estimator $\hat{\theta}_n$ with the true value of $\theta = 0$.



Sample mean

Consider a random sample $X_1, \ldots, X_n$ from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

Sample mean

Consider a random sample $X_1, \ldots, X_n$ from a distribution $F_{(\mu,\sigma^2)}$ where $\operatorname{E} X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

- The sample mean $\bar{X}_n$ is **unbiased**:

$$\operatorname{E} \bar{X}_n = \operatorname{E} \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \operatorname{E} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} \operatorname{E} X_i = \frac{n\mu}{n} = \mu.$$

Sample mean

Consider a random sample $X_1, \dots, X_n$ from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

- The sample mean $\bar{X}_n$ is **unbiased**:

$$\mathrm{E}\, \bar{X}_n = \mathrm{E}\, \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n}\, \mathrm{E} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\, X_i = \frac{n\mu}{n} = \mu.$$

- It is also **consistent**: from the weak law of large numbers we get that

$$\bar{X}_n \xrightarrow{\mathrm{P}} \mu \quad \text{for } n \to \infty.$$

- The same follows from previous theorem and the fact that $\operatorname{var} \bar{X}_n = \dfrac{\sigma^2}{n} \to 0$.

Sample mean

Consider a random sample $X_1, \ldots, X_n$ from a distribution $F_{(\mu,\sigma^2)}$ where $\operatorname{E} X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

- The sample mean $\bar{X}_n$ is **unbiased**:

$$\operatorname{E} \bar{X}_n = \operatorname{E} \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \operatorname{E} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} \operatorname{E} X_i = \frac{n\mu}{n} = \mu.$$

- It is also **consistent**: from the weak law of large numbers we get that

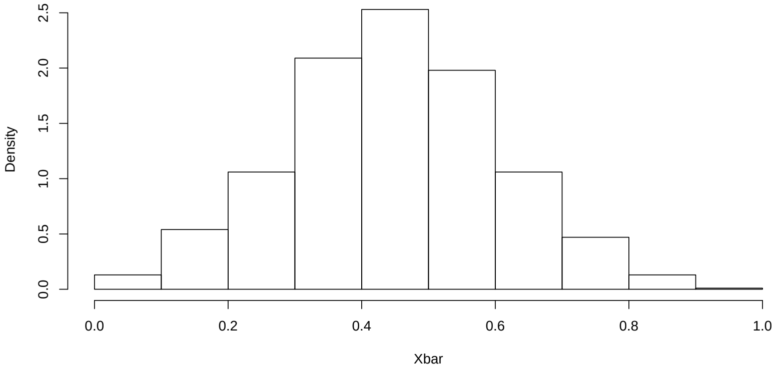
$$\bar{X}_n \xrightarrow{\mathrm{P}} \mu \quad \text{for } n \to \infty.$$

- The same follows from previous theorem and the fact that $\operatorname{var} \bar{X}_n = \dfrac{\sigma^2}{n} \to 0$.

The sample mean $\bar{X}_n$ is thus an unbiased and consistent estimator of the expectation.

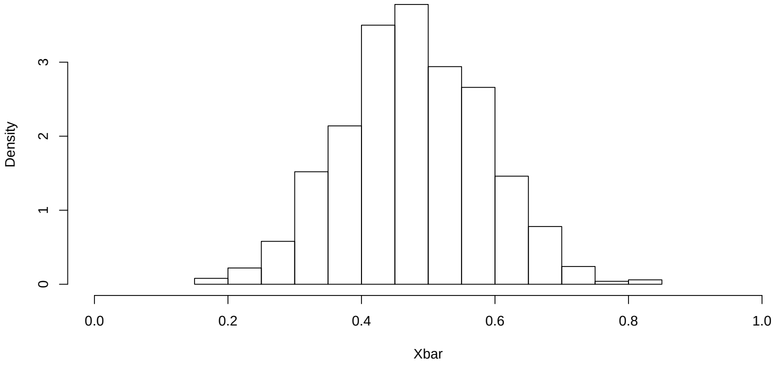
Distribution of the sample mean

Histogram of the proportion of heads among 10 coin tosses (1000 simulations).



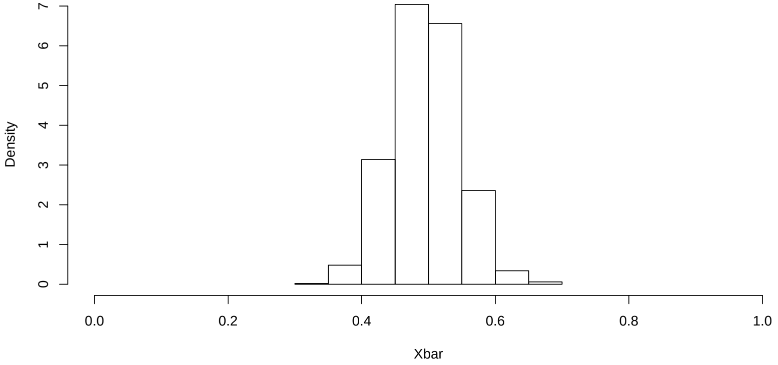
Distribution of the sample mean

Histogram of the proportion of heads among 20 coin tosses (1000 simulations).



Distribution of the sample mean

Histogram of the proportion of heads among 100 coin tosses (1000 simulations).



Sample variance

Consider a random sample $X_1, \dots, X_n$ from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

We want to estimate the variance σ^2 using the sample variance s_n^2.

Sample variance

Consider a random sample $X_1, \ldots, X_n$ from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\mathrm{var}\, X_i = \sigma^2$.

We want to estimate the variance σ^2 using the sample variance s_n^2. First we rewrite s_n^2 as

$$
\begin{aligned}
s_n^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i\bar{X}_n + n\bar{X}_n^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2 \right).
\end{aligned}
$$

Sample variance

- **Unbiasedness**: since $\mathrm{E}\, X_i^2 = \sigma^2 + \mu^2$ and $\mathrm{E}\, \bar{X}_n^2 = \sigma^2/n + \mu^2$, we get

$$\mathrm{E}\, s_n^2 = \frac{1}{n-1}\, \mathrm{E}\left(\sum_i X_i^2 - n\bar{X}_n^2\right) = \frac{1}{n-1}\left(n\, \mathrm{E}\, X_i^2 - n\, \mathrm{E}\, \bar{X}_n^2\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - n\sigma^2/n - n\mu^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.$$

Sample variance

- **Unbiasedness**: since $\mathrm{E}\,X_i^2 = \sigma^2 + \mu^2$ and $\mathrm{E}\,\bar{X}_n^2 = \sigma^2/n + \mu^2$, we get

$$\mathrm{E}\,s_n^2 = \frac{1}{n-1}\,\mathrm{E}\left(\sum_i X_i^2 - n\bar{X}_n^2\right) = \frac{1}{n-1}\left(n\,\mathrm{E}\,X_i^2 - n\,\mathrm{E}\,\bar{X}_n^2\right)$$
$$= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - n\sigma^2/n - n\mu^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.$$

- **Consistency**: from the law of large numbers we get $\bar{X}_n \xrightarrow{n\to\infty} \mu = \mathrm{E}\,X_i$ and also $\frac{1}{n}\sum_i X_i^2 \xrightarrow{n\to\infty} \mathrm{E}\,X_i^2$. Thus we get

$$s_n^2 = \frac{1}{n-1}\left(\sum_i X_i^2 - n\bar{X}_n^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_i X_i^2 - \bar{X}_n^2\right)$$
$$\xrightarrow{n\to\infty} 1\cdot\left(\mathrm{E}\,X_i^2 - \mu^2\right) = \mathrm{E}\,X_i^2 - (\mathrm{E}\,X_i)^2 = \operatorname{var} X_i = \sigma^2.$$

Sample variance

- **Unbiasedness**: since $\mathrm{E}\, X_i^2 = \sigma^2 + \mu^2$ and $\mathrm{E}\, \bar{X}_n^2 = \sigma^2/n + \mu^2$, we get

$$\mathrm{E}\, s_n^2 = \frac{1}{n-1}\, \mathrm{E}\left(\sum_i X_i^2 - n\bar{X}_n^2\right) = \frac{1}{n-1}\left(n\, \mathrm{E}\, X_i^2 - n\, \mathrm{E}\, \bar{X}_n^2\right)$$
$$= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - n\sigma^2/n - n\mu^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.$$

- **Consistency**: from the law of large numbers we get $\bar{X}_n \xrightarrow{n\to\infty} \mu = \mathrm{E}\, X_i$ and also $\frac{1}{n}\sum_i X_i^2 \xrightarrow{n\to\infty} \mathrm{E}\, X_i^2$. Thus we get

$$s_n^2 = \frac{1}{n-1}\left(\sum_i X_i^2 - n\bar{X}_n^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_i X_i^2 - \bar{X}_n^2\right)$$
$$\xrightarrow{n\to\infty} 1 \cdot \left(\mathrm{E}\, X_i^2 - \mu^2\right) = \mathrm{E}\, X_i^2 - (\mathrm{E}\, X_i)^2 = \operatorname{var} X_i = \sigma^2.$$

The sample variance s_n^2 is thus an unbiased and consistent estimator of the variance σ^2.

Quality of an unbiased estimator

Often we can construct several unbiased estimators of a given parameter. In this case we try to find the best of them, meaning the one with the smallest variance.

Quality of an unbiased estimator

Often we can construct several unbiased estimators of a given parameter. In this case we try to find the best of them, meaning the one with the smallest variance.

Definition

An estimator $\widehat{\theta}_n^{\text{best}}(X_1, \ldots, X_n)$ is called the best unbiased estimator of the parameter θ if it is unbiased and for all other unbiased estimators $\widehat{\theta}_n$ of parameter θ it holds that

$$\operatorname{var}(\widehat{\theta}_n) \geq \operatorname{var}(\widehat{\theta}_n^{\text{best}}) \quad \text{for all } \theta \in \Theta$$

Quality of an unbiased estimator

Often we can construct several unbiased estimators of a given parameter. In this case we try to find the best of them, meaning the one with the smallest variance.

Definition

An estimator $\widehat{\theta}_n^{\text{best}}(X_1, \ldots, X_n)$ is called the best unbiased estimator of the parameter θ if it is unbiased and for all other unbiased estimators $\widehat{\theta}_n$ of parameter θ it holds that

$$\operatorname{var}(\widehat{\theta}_n) \geq \operatorname{var}(\widehat{\theta}_n^{\text{best}}) \quad \text{for all } \theta \in \Theta$$

There exists a lower bound for the variance of an unbiased estimator (Rao - Cramer lower bound). If we find an unbiased estimator with the variance equal to this lower bound, we have the best unbiased estimator.

Quality of an unbiased estimator

Often we can construct several unbiased estimators of a given parameter. In this case we try to find the best of them, meaning the one with the smallest variance.

Definition

An estimator $\widehat{\theta}_n^{\text{best}}(X_1, \ldots, X_n)$ is called the best unbiased estimator of the parameter θ if it is unbiased and for all other unbiased estimators $\widehat{\theta}_n$ of parameter θ it holds that

$$\operatorname{var}(\widehat{\theta}_n) \geq \operatorname{var}(\widehat{\theta}_n^{\text{best}}) \quad \text{for all } \theta \in \Theta$$

There exists a lower bound for the variance of an unbiased estimator (Rao - Cramer lower bound). If we find an unbiased estimator with the variance equal to this lower bound, we have the best unbiased estimator.

Theorem

For binomial, Poisson, exponential, and normal distribution the sample mean is the best unbiased estimator of the expected value. For the normal distribution the sample variance is the best unbiased estimator of the variance.

Method of moments

For a simple and quick (but sometimes not optimal) estimate of the parameters, the **method of moments** can be used. Let $X_1, \ldots, X_n$ be a sample from a distribution with a d-dimensional parameter $\theta = (\theta_1, \ldots, \theta_d)$.

Method of moments

For a simple and quick (but sometimes not optimal) estimate of the parameters, the **method of moments** can be used. Let $X_1, \dots, X_n$ be a sample from a distribution with a d-dimensional parameter $\theta = (\theta_1, \dots, \theta_d)$.

Steps of the method of moments:

- Compute the theoretical moments $\mathrm{E}\, X_i^k$, for $k = 1, \dots, d$.

Method of moments

For a simple and quick (but sometimes not optimal) estimate of the parameters, the **method of moments** can be used. Let $X_1, \ldots, X_n$ be a sample from a distribution with a d-dimensional parameter $\theta = (\theta_1, \ldots, \theta_d)$.

Steps of the method of moments:

- Compute the theoretical moments $\mathrm{E}\, X_i^k$, for $k = 1, \ldots, d$.
- Express the parameters as functions of the moments.

Method of moments

For a simple and quick (but sometimes not optimal) estimate of the parameters, the **method of moments** can be used. Let $X_1, \ldots, X_n$ be a sample from a distribution with a d-dimensional parameter $\theta = (\theta_1, \ldots, \theta_d)$.

Steps of the method of moments:

- Compute the theoretical moments $\mathrm{E}\, X_i^k$, for $k = 1, \ldots, d$.
- Express the parameters as functions of the moments.
- Estimate the theoretical moments by their empirical versions:

$$\widehat{\mathrm{E}\, X_i^k} = m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Method of moments

For a simple and quick (but sometimes not optimal) estimate of the parameters, the **method of moments** can be used. Let $X_1, \ldots, X_n$ be a sample from a distribution with a d-dimensional parameter $\theta = (\theta_1, \ldots, \theta_d)$.

Steps of the method of moments:

- Compute the theoretical moments $\mathrm{E}\, X_i^k$, for $k = 1, \ldots, d$.
- Express the parameters as functions of the moments.
- Estimate the theoretical moments by their empirical versions:

$$\widehat{\mathrm{E}\, X_i^k} = m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

- Insert the estimated moments and find the parameter estimates by solving the corresponding equations.

Method of moments

For a simple and quick (but sometimes not optimal) estimate of the parameters, the **method of moments** can be used. Let $X_1, \dots, X_n$ be a sample from a distribution with a d-dimensional parameter $\theta = (\theta_1, \dots, \theta_d)$.

Steps of the method of moments:

- Compute the theoretical moments $\mathrm{E}\, X_i^k$, for $k = 1, \dots, d$.
- Express the parameters as functions of the moments.
- Estimate the theoretical moments by their empirical versions:

$$\widehat{\mathrm{E}\, X_i^k} = m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

- Insert the estimated moments and find the parameter estimates by solving the corresponding equations.

The method is useful because the law of large numbers implies that $m_k \to \mathrm{E}\, X_i^k$ for $n \to +\infty$. The estimates are thus always consistent.

Methods of moments – estimator of the variance

Suppose $X_1, \dots, X_n$ form a random sample from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

Methods of moments – estimator of the variance

Suppose $X_1, \dots, X_n$ form a random sample from a distribution $F_{(\mu,\sigma^2)}$ where $\operatorname{E} X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

- The first two theoretical moments are

$$\operatorname{E} X_i = \mu, \qquad \operatorname{E} X_i^2 = \operatorname{var} X_i + (\operatorname{E} X_i)^2 = \sigma^2 + \mu^2.$$

Methods of moments – estimator of the variance

Suppose $X_1, \ldots, X_n$ form a random sample from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\mathrm{var}\, X_i = \sigma^2$.

- The first two theoretical moments are

$$\mathrm{E}\, X_i = \mu, \qquad \mathrm{E}\, X_i^2 = \mathrm{var}\, X_i + (\mathrm{E}\, X_i)^2 = \sigma^2 + \mu^2.$$

- The parameters can be expressed as functions of the moments:

$$\mu = \mathrm{E}\, X_i, \qquad \sigma^2 = \mathrm{E}\, X_i^2 - (\mathrm{E}\, X_i)^2.$$

Methods of moments – estimator of the variance

Suppose $X_1, \ldots, X_n$ form a random sample from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

- The first two theoretical moments are

$$\mathrm{E}\, X_i = \mu, \qquad \mathrm{E}\, X_i^2 = \operatorname{var} X_i + (\mathrm{E}\, X_i)^2 = \sigma^2 + \mu^2.$$

- The parameters can be expressed as functions of the moments:

$$\mu = \mathrm{E}\, X_i, \qquad \sigma^2 = \mathrm{E}\, X_i^2 - (\mathrm{E}\, X_i)^2.$$

- We estimate $\mathrm{E}\, X_i$ using m_1 and $\mathrm{E}\, X_i^2$ using m_1 and m_2.

Methods of moments – estimator of the variance

Suppose $X_1, \dots, X_n$ form a random sample from a distribution $F_{(\mu,\sigma^2)}$ where $\operatorname{E} X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

- The first two theoretical moments are

$$\operatorname{E} X_i = \mu, \qquad \operatorname{E} X_i^2 = \operatorname{var} X_i + (\operatorname{E} X_i)^2 = \sigma^2 + \mu^2.$$

- The parameters can be expressed as functions of the moments:

$$\mu = \operatorname{E} X_i, \qquad \sigma^2 = \operatorname{E} X_i^2 - (\operatorname{E} X_i)^2.$$

- We estimate $\operatorname{E} X_i$ using m_1 and $\operatorname{E} X_i^2$ using m_1 and m_2.
- The estimators of the expectation and variance are then

$$\hat{\mu}_n = \bar{X}_n \qquad \text{and} \qquad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2.$$

Methods of moments – estimator of the variance

Suppose $X_1, \dots, X_n$ form a random sample from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\mathrm{var}\, X_i = \sigma^2$.

- The first two theoretical moments are

$$\mathrm{E}\, X_i = \mu, \qquad \mathrm{E}\, X_i^2 = \mathrm{var}\, X_i + (\mathrm{E}\, X_i)^2 = \sigma^2 + \mu^2.$$

- The parameters can be expressed as functions of the moments:

$$\mu = \mathrm{E}\, X_i, \qquad \sigma^2 = \mathrm{E}\, X_i^2 - (\mathrm{E}\, X_i)^2.$$

- We estimate $\mathrm{E}\, X_i$ using m_1 and $\mathrm{E}\, X_i^2$ using m_1 and m_2.
- The estimators of the expectation and variance are then

$$\hat{\mu}_n = \bar{X}_n \qquad \text{and} \qquad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2.$$

- After some algebra

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X}_n + (\bar{X}_n)^2\right) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n} s_n^2.$$

Methods of moments – estimator of the variance

Suppose $X_1, \dots, X_n$ form a random sample from a distribution $F_{(\mu,\sigma^2)}$ where $\mathrm{E}\, X_i = \mu$ and $\mathrm{var}\, X_i = \sigma^2$.

- The first two theoretical moments are

$$\mathrm{E}\, X_i = \mu, \qquad \mathrm{E}\, X_i^2 = \mathrm{var}\, X_i + (\mathrm{E}\, X_i)^2 = \sigma^2 + \mu^2.$$

- The parameters can be expressed as functions of the moments:

$$\mu = \mathrm{E}\, X_i, \qquad \sigma^2 = \mathrm{E}\, X_i^2 - (\mathrm{E}\, X_i)^2.$$

- We estimate $\mathrm{E}\, X_i$ using m_1 and $\mathrm{E}\, X_i^2$ using m_1 and m_2.
- The estimators of the expectation and variance are then

$$\hat{\mu}_n = \bar{X}_n \qquad \text{and} \qquad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2.$$

- After some algebra

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X}_n + (\bar{X}_n)^2\right) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n} s_n^2.$$

This estimator of the variance is consistent, but not unbiased. However, the extent of the bias will decrease, as $\frac{n-1}{n} \to 1$ for $n \to \infty$.

Maximum likelihood method – motivation

Example

Suppose that among four coin tosses we obtained the sequence H,T,H,H. How can we estimate the expected proportion of Heads?

Maximum likelihood method – motivation

Example

Suppose that among four coin tosses we obtained the sequence H,T,H,H. How can we estimate the expected proportion of Heads?

X_1, X_2, X_3, X_4 form a random sample from the Bernoulli distribution with the parameter p with realizations $1, 0, 1, 1$. The probability of such realization is:

Maximum likelihood method – motivation

Example

Suppose that among four coin tosses we obtained the sequence H,T,H,H. How can we estimate the expected proportion of Heads?

X_1, X_2, X_3, X_4 form a random sample from the Bernoulli distribution with the parameter p with realizations $1, 0, 1, 1$. The probability of such realization is:

$$L(p) = \mathrm{P}(\text{H,T,H,H}) = \mathrm{P}(X_1 = 1 \cap X_2 = 0 \cap X_3 = 1 \cap X_4 = 1) = p^3(1-p).$$

Maximum likelihood method – motivation

Example

Suppose that among four coin tosses we obtained the sequence H,T,H,H. How can we estimate the expected proportion of Heads?

X_1, X_2, X_3, X_4 form a random sample from the Bernoulli distribution with the parameter p with realizations $1, 0, 1, 1$. The probability of such realization is:

$$L(p) = \mathrm{P}(\text{H,T,H,H}) = \mathrm{P}(X_1 = 1 \cap X_2 = 0 \cap X_3 = 1 \cap X_4 = 1) = p^3(1-p).$$

As an estimate of the parameter p we take the value for which the obtained realization has the largest probability. Thus we find the maximum of the function $L(p)$.

$$\frac{\mathrm{d}L}{\mathrm{d}p}(p) = \frac{\mathrm{d}}{\mathrm{d}p}(p^3 - p^4) = 3p^2 - 4p^3 = p^2(3 - 4p) = 0.$$

Maximum likelihood method – motivation

Example

Suppose that among four coin tosses we obtained the sequence H,T,H,H. How can we estimate the expected proportion of Heads?

X_1, X_2, X_3, X_4 form a random sample from the Bernoulli distribution with the parameter p with realizations $1, 0, 1, 1$. The probability of such realization is:

$$L(p) = \mathrm{P}(\text{H,T,H,H}) = \mathrm{P}(X_1 = 1 \cap X_2 = 0 \cap X_3 = 1 \cap X_4 = 1) = p^3(1-p).$$

As an estimate of the parameter p we take the value for which the obtained realization has the largest probability. Thus we find the maximum of the function $L(p)$.

$$\frac{\mathrm{d}L}{\mathrm{d}p}(p) = \frac{\mathrm{d}}{\mathrm{d}p}(p^3 - p^4) = 3p^2 - 4p^3 = p^2(3-4p) = 0.$$

Stationary points are 0 and $\frac{3}{4}$ and the maximum is achieved at point $\frac{3}{4}$. Hence we obtain the estimate $\widehat{p}_n = \frac{3}{4}$, which can be guessed from the set up.

Maximum likelihood method

Consistent estimators with desirable properties can be obtained using the **maximum likelihood** method. The aim is to maximize the **likelihood function** for given observations.

Maximum likelihood method

Consistent estimators with desirable properties can be obtained using the **maximum likelihood** method. The aim is to maximize the **likelihood function** for given observations.

Definition

Let the random sample $X_1, \dots, X_n$ have a distribution given by the joint density

$$f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) \quad \text{for a continuous distribution or}$$

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{n} \mathrm{P}_\theta(X_i = x_i) \quad \text{for a discrete distribution.}$$

With values of $\mathbf{x} = (x_1, \dots, x_n)$ fixed, the function $f_\theta(\mathbf{x})$, or $p_\theta(\mathbf{x})$, as a function of θ is called the **likelihood function** and is denoted as $L(\theta; \mathbf{x})$ or simply $L(\theta)$.

Maximum likelihood method

Consistent estimators with desirable properties can be obtained using the **maximum likelihood** method. The aim is to maximize the **likelihood function** for given observations.

Definition

Let the random sample $X_1, \ldots, X_n$ have a distribution given by the joint density

$$f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) \quad \text{for a continuous distribution or}$$

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{n} \mathrm{P}_\theta(X_i = x_i) \quad \text{for a discrete distribution.}$$

With values of $\mathbf{x} = (x_1, \ldots, x_n)$ fixed, the function $f_\theta(\mathbf{x})$, or $p_\theta(\mathbf{x})$, as a function of θ is called the **likelihood function** and is denoted as $L(\theta; \mathbf{x})$ or simply $L(\theta)$.

The likelihood function depends only on the parameter θ. The values $x_1, \ldots, x_n$ are treated as known and fixed.

Maximum likelihood method

Definition

The value $\hat{\theta}_n$ of the parameter θ maximizing the likelihood function $L(\theta; \mathbf{x})$ for a given random sample realization $\mathbf{X} = \mathbf{x}$ is called the **maximum likelihood estimator** (MLE) of the parameter θ. It means that

$$L(\hat{\theta}_n; \mathbf{x}) \geq L(\theta; \mathbf{x}) \quad \text{for all } \theta \in \Theta.$$

Maximum likelihood method

Definition

The value $\hat{\theta}_n$ of the parameter θ maximizing the likelihood function $L(\theta; \mathbf{x})$ for a given random sample realization $\mathbf{X} = \mathbf{x}$ is called the **maximum likelihood estimator** (MLE) of the parameter θ. It means that

$$L(\hat{\theta}_n; \mathbf{x}) \geq L(\theta; \mathbf{x}) \quad \text{for all } \theta \in \Theta.$$

Notes:

- We can take $g(\hat{\theta}_n)$ as the maximum likelihood estimator of a function $g(\theta)$.

Maximum likelihood method

Definition

The value $\hat{\theta}_n$ of the parameter θ maximizing the likelihood function $L(\theta; \mathbf{x})$ for a given random sample realization $\mathbf{X} = \mathbf{x}$ is called the **maximum likelihood estimator** (MLE) of the parameter θ. It means that

$$L(\hat{\theta}_n; \mathbf{x}) \geq L(\theta; \mathbf{x}) \quad \text{for all } \theta \in \Theta.$$

Notes:

- We can take $g(\hat{\theta}_n)$ as the maximum likelihood estimator of a function $g(\theta)$.
- Often it is advantageous to maximize the function $\ln L(\theta; \mathbf{x})$, because the logarithm turns a product into a sum.

Maximum likelihood method

Definition

The value $\hat{\theta}_n$ of the parameter θ maximizing the likelihood function $L(\theta; \mathbf{x})$ for a given random sample realization $\mathbf{X} = \mathbf{x}$ is called the **maximum likelihood estimator** (MLE) of the parameter θ. It means that

$$L(\hat{\theta}_n; \mathbf{x}) \geq L(\theta; \mathbf{x}) \quad \text{for all } \theta \in \Theta.$$

Notes:

- We can take $g(\hat{\theta}_n)$ as the maximum likelihood estimator of a function $g(\theta)$.
- Often it is advantageous to maximize the function $\ln L(\theta; \mathbf{x})$, because the logarithm turns a product into a sum.
- In the case of a k-dimensional parameter $\theta = (\theta_1, \dots, \theta_k)$ we usually solve a system of equations

$$\frac{\partial \ln L(\theta_1, \dots, \theta_k; \mathbf{x})}{\partial \theta_j} = 0 \quad \text{for } j = 1, \dots, k.$$

Maximum likelihood method

Definition

The value $\hat{\theta}_n$ of the parameter θ maximizing the likelihood function $L(\theta; \mathbf{x})$ for a given random sample realization $\mathbf{X} = \mathbf{x}$ is called the **maximum likelihood estimator** (MLE) of the parameter θ. It means that

$$L(\hat{\theta}_n; \mathbf{x}) \geq L(\theta; \mathbf{x}) \quad \text{for all } \theta \in \Theta.$$

Notes:

- We can take $g(\hat{\theta}_n)$ as the maximum likelihood estimator of a function $g(\theta)$.
- Often it is advantageous to maximize the function $\ln L(\theta; \mathbf{x})$, because the logarithm turns a product into a sum.
- In the case of a k-dimensional parameter $\theta = (\theta_1, \dots, \theta_k)$ we usually solve a system of equations

$$\frac{\partial \ln L(\theta_1, \dots, \theta_k; \mathbf{x})}{\partial \theta_j} = 0 \quad \text{for } j = 1, \dots, k.$$

- If certain *regularity conditions* are met (see literature), the maximum likelihood estimates are consistent, asymptotically unbiased and asymptotically normal.

Maximum likelihood method – continuous example

Example – parameter of the exponential distribution

Construct the MLE estimate of the parameter $\lambda > 0$ of the exponential distribution Exp(λ).

Maximum likelihood method – continuous example

Example – parameter of the exponential distribution

Construct the MLE estimate of the parameter $\lambda > 0$ of the exponential distribution Exp(λ). The likelihood function for n observed values $x_1, \ldots, x_n$ (random sample realization) is clearly:

$$L(\lambda; \mathbf{x}) = f_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}.$$

Maximum likelihood method – continuous example

Example – parameter of the exponential distribution

Construct the MLE estimate of the parameter $\lambda > 0$ of the exponential distribution Exp(λ). The likelihood function for n observed values $x_1, \dots, x_n$ (random sample realization) is clearly:

$$L(\lambda; \mathbf{x}) = f_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}.$$

In this case it is advantageous to maximize $\ln L(\lambda; \mathbf{x}) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$.

Maximum likelihood method – continuous example

Example – parameter of the exponential distribution

Construct the MLE estimate of the parameter $\lambda > 0$ of the exponential distribution Exp(λ). The likelihood function for n observed values $x_1, \ldots, x_n$ (random sample realization) is clearly:

$$L(\lambda; \mathbf{x}) = f_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}.$$

In this case it is advantageous to maximize $\ln L(\lambda; \mathbf{x}) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$.

After differentiating we obtain:

$$\frac{\mathrm{d} \ln L(\lambda; \mathbf{x})}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$$

Maximum likelihood method – continuous example

Example – parameter of the exponential distribution

Construct the MLE estimate of the parameter $\lambda > 0$ of the exponential distribution Exp(λ). The likelihood function for n observed values $x_1, \dots, x_n$ (random sample realization) is clearly:

$$L(\lambda; \mathbf{x}) = f_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}.$$

In this case it is advantageous to maximize $\ln L(\lambda; \mathbf{x}) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$.

After differentiating we obtain:

$$\frac{\mathrm{d} \ln L(\lambda; \mathbf{x})}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$$

A solution of this equation is the **maximal likelihood estimator** $\hat{\lambda}_n = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}_n}$.

Maximum likelihood method – continuous example

Example – parameter of the exponential distribution

Construct the MLE estimate of the parameter $\lambda > 0$ of the exponential distribution $\text{Exp}(\lambda)$. The likelihood function for n observed values $x_1, \ldots, x_n$ (random sample realization) is clearly:

$$L(\lambda; \mathbf{x}) = f_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}.$$

In this case it is advantageous to maximize $\ln L(\lambda; \mathbf{x}) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$.

After differentiating we obtain:

$$\frac{\mathrm{d} \ln L(\lambda; \mathbf{x})}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$$

A solution of this equation is the **maximal likelihood estimator** $\hat{\lambda}_n = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}_n}$.

Using the second derivative we can check that the obtained point is indeed the maximum.

Estimating the distribution – example

Example – waiting for a bus – comparison of distributions

Try fitting known continuous distributions on the observed waiting times from before. Estimate their parameters and compare the densities with the histogram.

Estimating the distribution – example

Example – waiting for a bus – comparison of distributions

Try fitting known continuous distributions on the observed waiting times from before. Estimate their parameters and compare the densities with the histogram.

0.1	0.3	0.5	0.7	1.0	1.9	2.8	3.4	3.5	3.8	5.3	7.7	8.6	8.7	11.1

We try fitting the uniform Unif$(0, b)$, exponential Exp(λ) and normal N(μ, σ^2) distributions with estimated parameters:

Estimating the distribution – example

Example – waiting for a bus – comparison of distributions

Try fitting known continuous distributions on the observed waiting times from before. Estimate their parameters and compare the densities with the histogram.

0.1 0.3 0.5 0.7 1.0 1.9 2.8 3.4 3.5 3.8 5.3 7.7 8.6 8.7 11.1

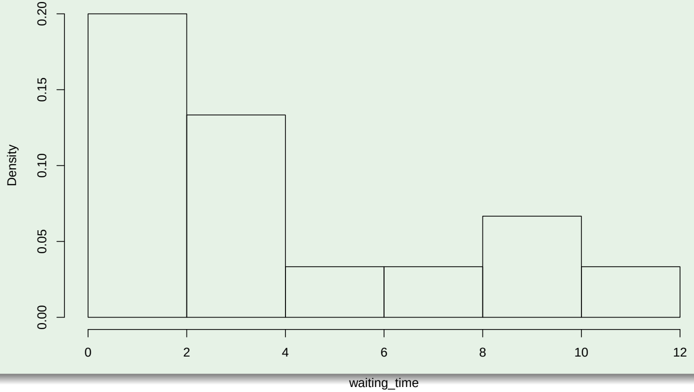
We try fitting the uniform Unif$(0, b)$, exponential Exp(λ) and normal N(μ, σ^2) distributions with estimated parameters:

Distribution	Estimated parameters	
Uniform	$a = 0$	$\hat{b}_n = \max(x_1, \ldots, x_{15}) \doteq 11.1$
Exponential	$\hat{\lambda}_n = \frac{1}{\bar{x}_n} \doteq 0.25$	–
Normal	$\hat{\mu}_n = \bar{x}_n \doteq 3.96$	$s_n^2 \doteq 12.56.$

Estimating the distribution – example

Example – waiting for a bus – comparison of distributions, continued

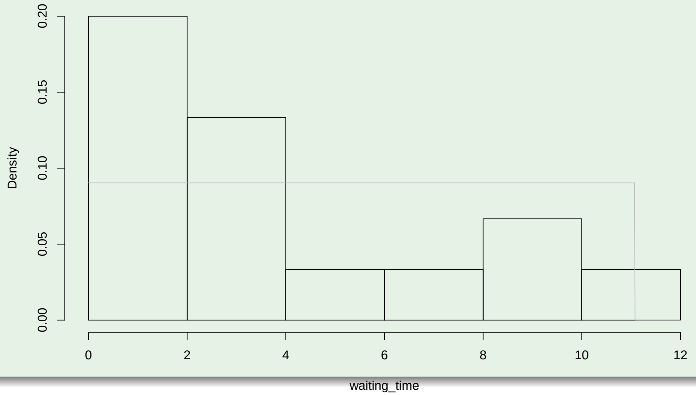
Compare the histogram with the fitted densities.



Estimating the distribution – example

Example – waiting for a bus – comparison of distributions, continued

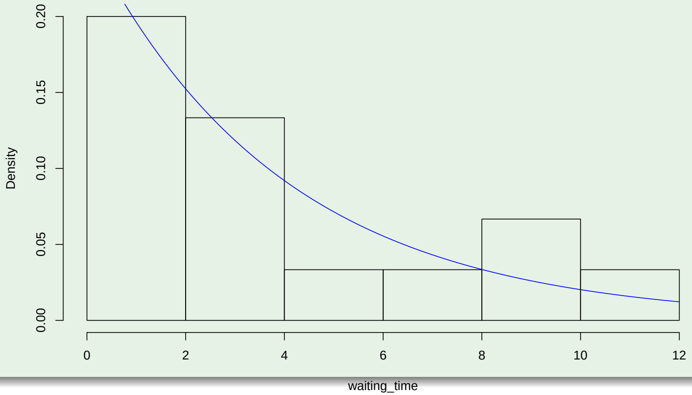
Compare the histogram with the fitted densities.



Estimating the distribution – example

Example – waiting for a bus – comparison of distributions, continued

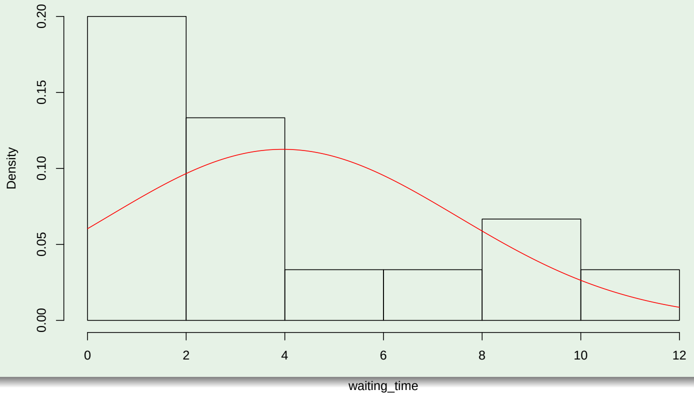
Compare the histogram with the fitted densities.



Estimating the distribution – example

Example – waiting for a bus – comparison of distributions, continued

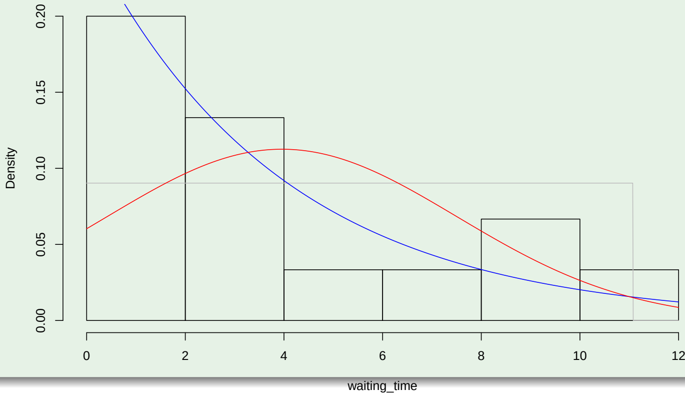
Compare the histogram with the fitted densities.



Estimating the distribution – example

Example – waiting for a bus – comparison of distributions, continued

Compare the histogram with the fitted densities.



The exponential distribution seems to provide the best fit.

Recap

Suppose we observe a **random sample** $X_1, ..., X_n$ (independent and identically distributed random variables) from an **unknown distribution**. We aim to estimate:

- the **shape** of the distribution – its type and parametric family;
- the **parameters** of the distribution.

To get a graphical overview of the shape of the distribution, we can find:

- The **histogram**, which is an approximation of the **density**.
- The **empirical distribution function**, which estimates the real **distribution function**.

To estimate the parameters θ we use **point estimators** $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$. We want them to be:

- **unbiased**, meaning that $\mathrm{E}\,\hat{\theta}_n = \theta$;
- **consistent**, meaning that $\hat{\theta}_n \overset{n\to\infty}{\longrightarrow} \theta$.

Estimates with reasonable properties may be found using:

- the **method of moments**;
- the **maximum likelihood method** (MLE).