

# Modèles stochastiques II

## *INFO 154*

Gianluca Bontempi

Département d'Informatique  
Boulevard de Triomphe - CP 212  
<http://www.ulb.ac.be/di>

# Approaches to estimation

There are two main approaches to estimation

**Classical or frequentist:** it is based on the idea that sample data are the sole quantifiable form of relevant information and that the parameters are **fixed but unknown**. It is related to the frequency view of probability.

**Bayesian approach:** the parameters are supposed to be random variables, having a distribution prior to data observation and a distribution posterior to data observation. This approach assumes that there exists something beyond data, (i.e. a subjective degree of belief), and that this belief can be described in probabilistic form.

# Classical approach: some history

It dates back to the period 1920-35

- J. Neyman and E.S. Pearson, stimulated by problems in biology and industry, concentrated on the principles for testing hypothesis
- R.A. Fisher who was interested in agricultural issues gave attention to the estimation problem

# Estimation

Consider a r.v.  $\mathbf{z}$ . Suppose that

1. we do not know completely the density  $p_{\mathbf{z}}(z)$  (the distribution if discrete) but that we can write it in a parametric form

$$p_{\mathbf{z}}(z) = p(\theta, z)$$

where  $\theta \in \Theta$  is a parameter,

2. we have access to a set  $D_N$  of  $N$  measurements of  $\mathbf{z}$ , called *sample data*.

- Goal of the **estimation** procedure: to find a value  $\hat{\theta}$  of the parameter  $\theta$  so that the parametrized distribution  $p(\hat{\theta}, z)$  closely matches the distribution of data.
- We assume that the  $N$  observations are the observed values of  $N$  i.i.d. random variables  $\mathbf{z}_i$ , each having a density identical to  $p_{\mathbf{z}}(z)$ .

# Some estimation problems

1. Consider the model of the traffic in the boulevard.  
Suppose that the measures of the inter-arrival times are  
 $D_N = \{10, 11, 1, 21, 2, \dots\}$  seconds.  
What does this imply about the average inter-arrival time?
2. Consider the students of the last year of Computer Science. What is the variance of their grades?

# Estimation (II)

Estimation is a mapping from the space of the sample data to the space of parameters  $\Theta$ . Two are the possible outcomes

1. some specific value of  $\theta$ . In this case we have the so-called **point estimation**.
2. some particular region of  $\Theta$ . In this case we obtain the **interval of confidence**.

# Point estimation

- Consider a random variable  $z$  with a parametric distribution  $F(z, \theta)$ ,  $\theta \in \Theta$ .
- The parameter can be written as a function of  $F$

$$\theta = t(F)$$

This corresponds to the fact that  $\theta$  is a characteristic of the population described by  $F$ .

- Suppose we have a set of  $N$  observations  $D_N = \{z_1, z_2, \dots, z_N\}$ .
- Any function of the sample data  $D_N$  is called a **statistic**. A *point estimate* is an example of statistic.
- A *point estimate* is a function

$$\hat{\theta} = g(D_N)$$

of the sample dataset  $D_N$ .

# Sample average

- Consider a normal r.v.  $\mathbf{z} \sim \mathcal{N}(\theta, 1)$ , where  $\theta$  is unknown.
- The distribution function is parametric and the parameter  $\theta$  (in this case the mean) can be written in the form

$$\theta = E[\mathbf{z}] = t(F) = \int z dF(z)$$

- Suppose we have available the sample  $F_{\mathbf{z}} \rightarrow D_n$ .
- A typical point estimate of  $\theta$  is given by the **sample average**

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^N z_i = \hat{\mu}$$

which is indeed a statistic, i.e. a function of the data set.

# Sample variance

- Consider a normal r.v.  $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ , where  $\mu$  and  $\sigma^2$  are unknown.
- Suppose we have available the sample  $F_{\mathbf{z}} \rightarrow D_n$ .
- Once we have the sample average  $\hat{\mu}$ , a typical estimate of  $\sigma^2$  is given by the **sample variance**

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (z_i - \hat{\mu})^2$$

- Note the presence of  $N - 1$  instead of  $N$  at the denominator (it will be explained later).
- Note that the following relation holds

$$\frac{1}{N} (z_i - \hat{\mu})^2 = \frac{1}{N} z_i^2 - \hat{\mu}^2$$

# Empirical distribution

Suppose we have observed a random sample of size  $N$  from a probability distribution  $F_{\mathbf{z}}$

$$F_{\mathbf{z}} \rightarrow \{z_1, z_2, \dots, z_N\}$$

The **empirical distribution**  $\hat{F}$  is defined to be the **discrete distribution** that puts probability  $1/N$  on each value  $z_i, i = 1, \dots, N$ .

In other words,  $\hat{F}$  assigns to a set  $A$  in the sample space of  $\mathbf{z}$  its empirical probability

$$\text{Prob} \{z \in A\} = \frac{\#z_i \in A}{N}$$

that is the proportion of the observed samples in  $D_N$  which occur in  $A$ . It can be proved that the vector of observed frequencies in  $\hat{F}$  is a sufficient statistic for the true distribution  $F$ , i.e. all the information about  $F$  contained in  $D_N$  is also contained in  $\hat{F}$ .

# Empirical distribution function

Consider now the specific case of the **distribution function**  $F_{\mathbf{z}}(z)$  of a continuous rv  $\mathbf{z}$  in terms of a set of  $N$  samples  $D_N = z_1, \dots, z_N$ .  
Since

$$F_{\mathbf{z}}(z) = \text{Prob} \{ \mathbf{z} \leq z \}$$

we define  $N(z)$  as the number of samples in  $D_N$  that do not exceed  $z$ .  
We obtain then the empirical estimate of  $F$

$$\hat{F}_{\mathbf{z}}(z) = \frac{N(z)}{N} = \frac{\#z_i \leq z}{N}$$

This function is a staircase function with discontinuities at the points  $z_i$ . It can be shown that

$$E_{D_N}[\hat{\mathbf{F}}_{\mathbf{z}}(z)] = F_{\mathbf{z}}(z)$$

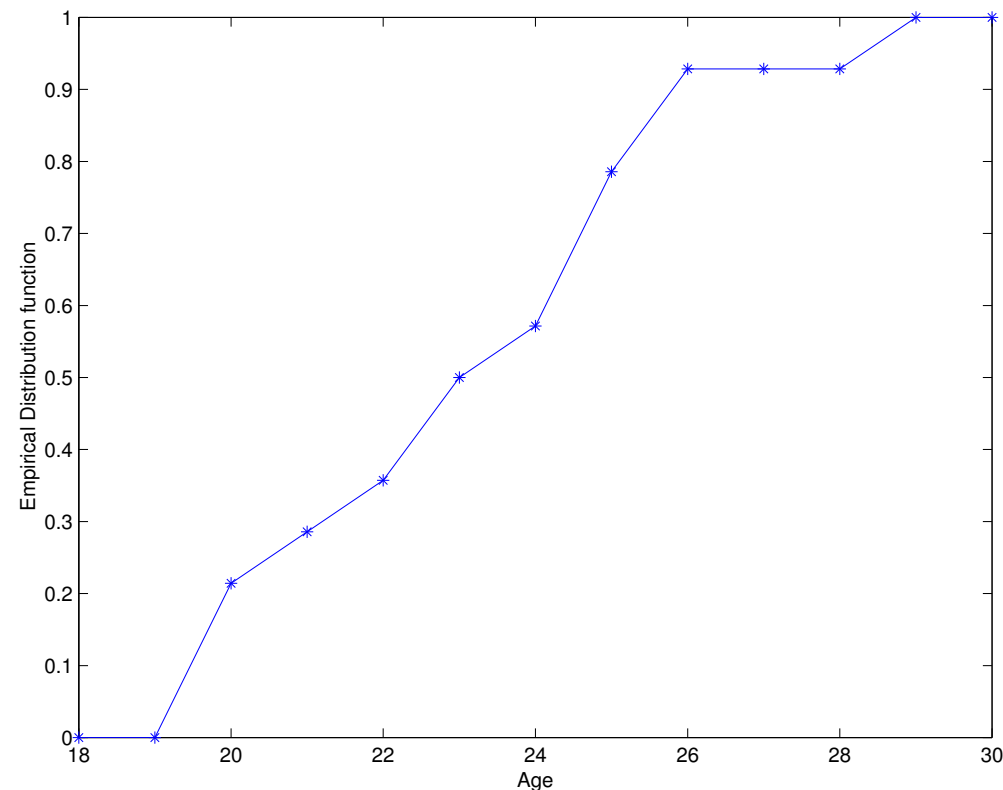
$$\hat{\mathbf{F}}_{\mathbf{z}}(z) \rightarrow F_{\mathbf{z}}(z) \quad \text{for } N \rightarrow \infty$$

# TP MATLAB: empirical distribution

- Suppose that our dataset of observations of the age is made of the following  $N = 14$  samples

$$D_N = \{20, 21, 22, 20, 23, 25, 26, 25, 20, 23, 24, 25, 26, 29\}$$

- Here it is the empirical distribution function  $\hat{F}_Z$  (cumdis.m)



# Plug-in principle to define an estimator

- Consider a r.v.  $\mathbf{z}$  and sample dataset  $D_N$  drawn from the parametric distribution  $p(\theta, \mathbf{z})$ .
- How to define an estimate of  $\theta$ ? How to choose  $t$ ?
- A possible solution is given by the **plug-in principle**, that is a simple method of estimating parameters from sample.
- The **plug-in estimate** of a parameter  $\theta$  is defined to be

$$\hat{\theta} = t(\hat{F}(\mathbf{z}))$$

obtained by replacing the distribution function with the empirical distribution in the analytical expression of the parameter

- The sample average is an example of plug-in estimate

$$\hat{\mu} = \int \mathbf{z} d\hat{F}(\mathbf{z}) = \frac{1}{N} \sum_{i=1}^N z_i$$

# Sampling distribution

- Given a dataset  $D_N$ , we have a point estimate

$$\hat{\theta} = g(D_N)$$

which is a specific value.

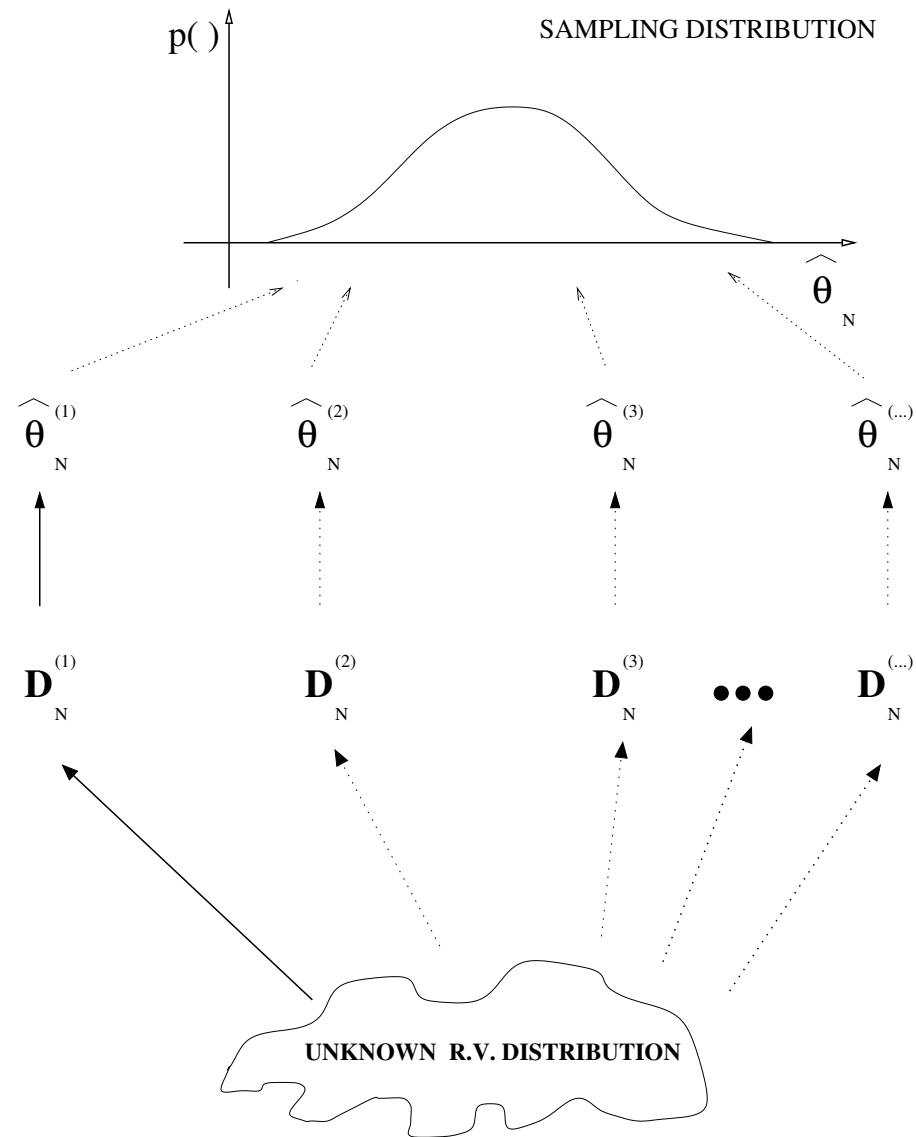
- However it is important to remark that  $D_N$  is the outcome of the sampling of a r.v.  $\mathbf{z}$ . As a consequence  $D_N$  can be considered as realization of a random variable  $\mathbf{D}_N$ .
- Applying the transformation  $g$  to the random variable  $\mathbf{D}_N$  we obtain another random variable

$$\hat{\theta} = g(\mathbf{D}_N)$$

which is called the *point estimator* of  $\theta$ .

- The probability distribution of the r.v.  $\hat{\theta}$  is called the **sampling distribution**.

# Sampling distribution



See the MATLAB file `sampl_distr.m`.

# Bias and variance

Summary statistics as  $\hat{\theta}$  are often the first outputs of a data analysis. The next thing we want to know is the accuracy of  $\hat{\theta}$ . This lead us to the definition of bias, variance and standard error of an estimator.

**Definition 1.** *An estimator  $\hat{\theta}$  of  $\theta$  is said to be unbiased if and only if*

$$E_{\mathbf{D}_N}[\hat{\theta}] = \theta$$

*Otherwise, it is called biased with bias*

$$\text{Bias}[\hat{\theta}] = E_{\mathbf{D}_N}[\hat{\theta}] - \theta$$

**Definition 2.** *The variance of an estimator  $\hat{\theta}$  of  $\theta$  is the variance of its sampling distribution*

$$\text{Var}[\hat{\theta}] = E_{\mathbf{D}_N}[(\hat{\theta} - E[\hat{\theta}])^2]$$

# Some consideration

- An unbiased estimator is an estimator that takes on average the right value.
- Many unbiased estimators may exist for a parameter  $\theta$ .
- If  $\hat{\theta}$  is unbiased for  $\theta$ ,  $f(\hat{\theta})$  is in generally NOT unbiased for  $f(\theta)$
- A biased estimator with a known bias (not depending on  $\theta$ ) is equivalent to an unbiased estimator since we can easily compensate for the bias.
- Given a r.v.  $\mathbf{z}$  and the set  $D_N$ , it can be shown that the sample average  $\hat{\mu}$  and the sample variance  $\hat{\sigma}^2$  are unbiased estimators of the mean  $E[\mathbf{z}]$  and the variance  $\text{Var}[\mathbf{z}]$ , respectively.
- In general  $\hat{\sigma}$  is not an unbiased estimator of  $\sigma$  even if  $\hat{\sigma}^2$  is an unbiased estimator of  $\sigma^2$ .

# Bias and variance of $\hat{\mu}$

$$E_{\mathbf{D}_N}[\hat{\mu}] = E_{\mathbf{D}_N} \left[ \frac{1}{N} \sum_{i=1}^N \mathbf{z}_i \right] = \frac{N\mu}{N} = \mu$$

Sample average is not biased !

$$\text{Var}[\hat{\mu}] = \text{Var} \left[ \frac{1}{N} \sum_{i=1}^N \mathbf{z}_i \right] = \frac{1}{N^2} \text{Var}[\mathbf{z}_i] = \frac{1}{N^2} N \sigma^2 = \frac{\sigma^2}{N}$$

# Bias of $\hat{\sigma}^2$

$$\begin{aligned} E_{\mathbf{D}_N}[\hat{\sigma}^2] &= E_{\mathbf{D}_N} \left[ \frac{1}{N-1} \sum_{i=1}^N (\mathbf{z}_i - \hat{\boldsymbol{\mu}})^2 \right] = \frac{N}{N-1} E_{\mathbf{D}_N} \left[ \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \hat{\boldsymbol{\mu}})^2 \right] = \\ &= \frac{N}{N-1} E_{\mathbf{D}_N} \left[ \frac{1}{N} \sum_{i=1}^N \mathbf{z}_i^2 - \hat{\boldsymbol{\mu}}^2 \right] \end{aligned}$$

Since  $E[\mathbf{z}^2] = \mu^2 + \sigma^2$ , we have

$$\begin{aligned} E_{\mathbf{D}_N}[\hat{\sigma}^2] &= \frac{N}{N-1} \left( \frac{1}{N} (\mu^2 + \sigma^2) - (\mu^2 + \sigma^2/N) \right) = \\ &= \frac{N}{N-1} \left( \frac{N-1}{N} \sigma^2 \right) = \sigma^2 \end{aligned}$$

Sample variance (with  $N - 1$  at denominator) is not biased !

# The distribution of the sample statistics

Let  $\mathbf{z}_1, \dots, \mathbf{z}_N$  be i.i.d.  $\mathcal{N}(\mu, \sigma^2)$  and let us consider the following sample statistics

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^N \mathbf{z}_i, \quad \widehat{\mathbf{SS}} = \sum_{i=1}^N (\mathbf{z}_i - \hat{\boldsymbol{\mu}})^2, \quad \hat{\sigma}^2 = \frac{\widehat{\mathbf{SS}}}{N-1}$$

It can be shown that the following relations hold

1.  $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\mu, \sigma^2/N)$  and  $N(\hat{\boldsymbol{\mu}} - \mu)^2 \sim \sigma^2 \chi_1^2$ .
2.  $\mathbf{z}_i - \mu \sim \mathcal{N}(0, \sigma^2)$ , so  $\sum_{i=1}^N (\mathbf{z}_i - \mu)^2 \sim \sigma^2 \chi_N^2$
3.  $\sum_{i=1}^N (\mathbf{z}_i - \mu)^2 = \widehat{\mathbf{SS}} + N(\hat{\boldsymbol{\mu}} - \mu)^2$
4.  $\hat{\sigma}^2$  is an unbiased estimator of  $\sigma^2$ .
5.  $\widehat{\mathbf{SS}} \sim \sigma^2 \chi_{N-1}^2$  or equivalently  $\frac{(N-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{N-1}^2$
6. if  $E[|\mathbf{z} - \mu|^4] = \mu_4$  then  $\text{Var}[\hat{\sigma}^2] = \frac{1}{N} \left( \mu_4 - \frac{N-3}{N-1} \sigma^4 \right)$ .

# Considerations

- The variance of  $\hat{\mu}$  is  $1/N$  times the variance of  $\mathbf{z}$ . This is a reason for taking averages: the larger  $N$ , the smaller is  $\text{Var}[\hat{\mu}]$ , so bigger  $N$  means a better estimate of  $\mu$ .
- According to the central limit theorem, under quite general conditions on the distribution  $F_{\mathbf{z}}$ , the distribution of  $\hat{\mu}$  will be approximately normal as  $N$  gets large, which we can write as

$$\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/N) \quad \text{for } N \rightarrow \infty$$

- Standard error  $\sqrt{\text{Var}[\hat{\mu}]}$  is a common way of indicating statistical accuracy. Roughly speaking we expect  $\hat{\mu}$  to be less than one standard error away from  $\mu$  about 68% of the time, and less than two standard errors away from  $\mu$  about 95% of the time .

# Bias/variance decomposition of MSE

When  $\hat{\theta}$  is a biased estimator of  $\theta$ , its accuracy is usually assessed by its **mean-square error** (MSE) rather than by its variance.

The MSE is defined by

$$\text{MSE} = E_{\mathbf{D}_N} [(\theta - \hat{\theta})^2]$$

- The MSE of an unbiased estimator is its variance.
- For a generic estimator it can be shown that

$$\text{MSE} = \text{Var} [\hat{\theta}] + (E[\hat{\theta}] - \theta)^2$$

i.e., the mean-square error is equal to the sum of the variance and the squared bias. This decomposition is typically called the **bias-variance decomposition**.

# Consistency

Suppose that the sample data contains  $N$  independent observations  $z_1, \dots, z_N$  of a univariate random variable. Let the estimator of  $\theta$  based on  $N$  samples be denoted  $\hat{\theta}_N$ .

As  $N$  becomes larger, we might reasonably expect that  $\hat{\theta}_N$  improves as estimator of  $\theta$  (in other terms it gets closer to  $\theta$ ). The notion of consistency introduces this concept.

**Definition 3.** *The estimator  $\hat{\theta}_N$  is said weakly consistent if  $\hat{\theta}_N$  converge to  $\theta$  in probability.*

$$\lim_{N \rightarrow \infty} \text{Prob} \{ |\hat{\theta}_N - \theta| \leq \epsilon \} = 1$$

**Definition 4.** *The estimator  $\hat{\theta}_N$  is said strongly consistent if  $\hat{\theta}_N$  converge to  $\theta$  with probability 1 (or almost surely).*

$$\text{Prob} \left\{ \lim_{N \rightarrow \infty} \hat{\theta}_N = \theta \right\} = 1$$

# Some considerations

- Suppose an estimator of the mean that takes into consideration only the first 10 samples, whatever the number  $N$  of samples: its variance stays constant as  $N \rightarrow \infty$ . It is evident that this estimator is not consistent.
- For a scalar  $\theta$  the property of convergence guarantee that the sampling distribution of  $\hat{\theta}_N$  becomes less disperse as  $N \rightarrow \infty$ .
- A sufficient condition for weak consistency of unbiased estimators  $\hat{\theta}_N$  is that  $\text{Var} \left[ \hat{\theta}_N \right] \rightarrow 0$  as  $N \rightarrow \infty$ .
- A consistent estimator is asymptotically unbiased.
- Property of unbiasedness and consistency are largely unrelated.

# TP: example

- Suppose  $z_1, \dots, z_N$  is a random sample of observations from a distribution with mean  $\theta$  and variance  $\sigma^2$ .
- Study the unbiasedness and the consistency of the three estimators of the mean  $\mu$ :

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum_{i=1}^N z_i}{N}$$

$$\hat{\theta}_2 = \frac{N\hat{\theta}_1}{N+1}$$

$$\hat{\theta}_3 = z_1$$

# Efficiency

Suppose we have two **unbiased and consistent** estimators. How to choose between them?

**Definition 5 (Relative efficiency).** . Let us consider two unbiased estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$ . If

$$\text{Var} \left[ \hat{\theta}_1 \right] < \text{Var} \left[ \hat{\theta}_2 \right]$$

we say that  $\hat{\theta}_1$  is more efficient than  $\hat{\theta}_2$ .

If the estimators are biased, typically the comparison is done on the basis of the mean square error.

# Sufficiency

**Definition 6 (Sufficiency.).** *An estimator  $\hat{\theta}$  is said to be sufficient for  $\theta$  if the conditional distribution of  $\mathbf{z}$  given  $\hat{\theta}$  does not depend on  $\theta$ .*

**Property 1 (Fisher-Neyman factorization criterion).** *An estimator  $\hat{\theta}$  is sufficient for  $\theta$  if and only if*

$$P_{\mathbf{z}}(z, \theta) = g(\hat{\theta}, \theta)h(z)$$

**Definition 7 (Minimal sufficiency.).** *If  $\hat{\theta}$  is sufficient and no statistic of lower dimension is sufficient the  $\hat{\theta}$  is said to be minimal sufficient*

# Some considerations

- If an estimator  $\hat{\theta}$  is sufficient for  $\theta$ , this means that all the information about  $\theta$  contained in the data  $D_N$  is obtained from  $\hat{\theta}$  alone.
- Usually  $\hat{\theta}$  will be of lower dimension than  $D_N$ .

Example: consider a sample data  $D_N$  from  $\mathcal{N}(\mu, \sigma^2)$  where  $\sigma^2$  is known. The estimator  $\hat{\mu}$  is a sufficient estimator. This means that only  $\hat{\mu}$  and no other element of the data provides information about  $\mu$ .

# Linear estimators

**Definition 8.** *An estimator  $\hat{\theta}$  is said to be a linear estimator of  $\theta$  if it is a linear function of the observation vector  $\mathbf{D}_N$ , that is  $\hat{\theta} = B^T \mathbf{D}_N$  where  $B$  is a vector of coefficients not dependent on  $D_N$ .*

Example: the estimator sample average  $\hat{\mu}$  is a linear estimator of  $\mu$ .

Property: the estimator  $\hat{\mu}$  is the linear unbiased estimator of  $\mu$  with the smallest variance.

# Likelihood

Let us consider

1. a density distribution  $p_{\mathbf{z}}(z, \theta)$  which depends on a parameter  $\theta$
2. a sample data  $D_N = \{z_1, z_2, \dots, z_N\}$  drawn independently from this distribution.

The joint probability density of the sample data is

$$p_{\mathbf{D}_N}(D_N, \theta) = \prod_{i=1}^N p_{\mathbf{z}}(z_i, \theta) = L_N(\theta) \quad (1)$$

where for a fixed  $D_N$ ,  $L_N$  is a function of  $\theta$  and is called the *empirical likelihood* of  $\theta$  given  $D_N$ .

# Cramer-Rao lower bound (I)

Define the **log-likelihood**  $l_N(\theta)$  by

$$l_N(\theta) = \log_e [L_N(\theta)]$$

Assume that

1.  $\theta$  is a scalar parameter,
2. the first two derivatives of  $L_N(\theta)$  with respect to  $\theta$  exist for all  $\theta$ .
3. certain operations of integration and differentiation may be interchanged.

Let  $\hat{\gamma}$  be an unbiased estimator of the function  $\gamma(\theta)$

We are interested in giving a lower bound to the variance of the estimator  $\hat{\gamma}$ .

# Cramer-Rao lower bound (II)

Once put  $l_N(\theta) = \log_e[L_N(\theta)]$  it can be shown that

$$E \left[ \frac{\partial l_N(\theta)}{\partial \theta} \right] = 0$$
$$E \left[ \left( \frac{\partial l_N(\theta)}{\partial \theta} \right)^2 \right] = -E \left[ \frac{\partial^2 l_N(\theta)}{\partial^2 \theta} \right]$$

where the quantity  $\partial l(\theta)/\partial \theta$  is called **score**.

Hence, it is possible to show that

$$\text{Var} [\hat{\gamma}] \geq - \frac{(\gamma'(\theta))^2}{E \left[ \frac{\partial^2 l_N(\theta)}{\partial^2 \theta} \right]}$$

An estimator having as variance the right term is called the **Minimum Variance Bound (MVB) estimator**.

# Information

If the sample data consists of  $N$  independent observations  $z_i$ , we have

$$l_N(\theta) = \log L_N(\theta) = \log \prod_{i=1}^N p(z_i, \theta) = \sum_{i=1}^N \log p(z_i, \theta)$$

and

$$E \left[ \left( \frac{\partial l_N}{\partial \theta} \right)^2 \right] = N E \left[ \left( \frac{\partial \log p(\mathbf{z})}{\partial \theta} \right)^2 \right] = I_N(\theta)$$

where  $I_N(\theta)$  is called the amount of **information** in the sample and  $I = I_N/N$  is the amount of information in a single observation.

The Cramer-Rao bound becomes

$$\text{Var} [\hat{\gamma}] \geq \frac{(\gamma'(\theta))^2}{N E \left[ \left( \frac{\partial \log p(\mathbf{z})}{\partial \theta} \right)^2 \right]} = \frac{(\gamma'(\theta))^2}{I_N}$$

# TP example

Show that for a random sample from  $\mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  known

$$I = \frac{1}{\sigma^2}$$

# MVB estimators

It can be shown that an estimator  $\hat{\gamma}$  is the MVB estimator of  $\gamma(\theta)$  if and only if

$$\frac{\partial l}{\partial \theta} = k(\theta)(\hat{\gamma} - \gamma(\theta))$$

- Since  $E[\partial \mathbf{l}_N / \partial \theta] = 0$  the MVB estimator is unbiased.
- The variance of the MVB estimator takes the value  $\gamma'(\theta)/k(\theta)$ .
- $\hat{\gamma}$  is a sufficient statistic for  $\theta$ .
- a MVB estimator may exist for one function  $\gamma(\theta)$  but not for another function  $\psi(\theta)$ .

# TP: MVB example

Let  $D_N = \{z_1, \dots, z_N\}$  be a random sample from  $\mathcal{N}(0, \theta)$ .

Show that

$$\frac{1}{N} \sum_{i=1}^N z_i^2$$

is the MVB estimator of the variance  $\theta$ .

# The Rao-Blackwell theorem

**Theorem 1.** *Let  $\hat{\theta}$  be an unbiased estimator of  $\theta$  and  $\theta^*$  be a sufficient statistic for  $\theta$ . Then the estimator*

$$\hat{\hat{\theta}} = E[\hat{\theta} | \theta^*]$$

*is also unbiased and*

$$\text{Var} [\hat{\hat{\theta}}] \leq \text{Var} [\hat{\theta}]$$

In other terms, if we have an unbiased estimator, and a sufficient statistic, then the expected value of the unbiased estimator conditional on the sufficient statistic is unbiased, is a function of the sufficient statistic and has variance no larger than that of the unbiased estimator.

# Methods of constructing estimators

For a generic probability distribution, the plug-in principle cannot be applied. Alternative methods have to be employed. Examples are:

- Maximum likelihood
- Least squares
- Minimum Chi-Squared

# Maximum likelihood

The principle of maximum likelihood was first used by Lambert around 1760 and by D. Bernoulli about 13 years later. It was detailed by Fisher in 1920.

Idea: given an unknown parameter  $\theta$  and a sample data  $D_N$ , the **maximum likelihood estimate**  $\hat{\theta}$  is the value for which the likelihood  $L_N(\theta)$  has a maximum

$$\hat{\theta}_{\text{ml}} = \arg \max_{\theta \in \Theta} L_N(\theta)$$

The estimator  $\hat{\theta}$  is called the **maximum likelihood estimator (m.l.e.)**. It is usual to consider the log-likelihood  $l_N(\theta)$  since

$$\hat{\theta}_{\text{ml}} = \arg \max_{\theta \in \Theta} L_N(\theta) = \arg \max_{\theta \in \Theta} \log(L_N(\theta)) = \arg \max_{\theta \in \Theta} l_N(\theta)$$

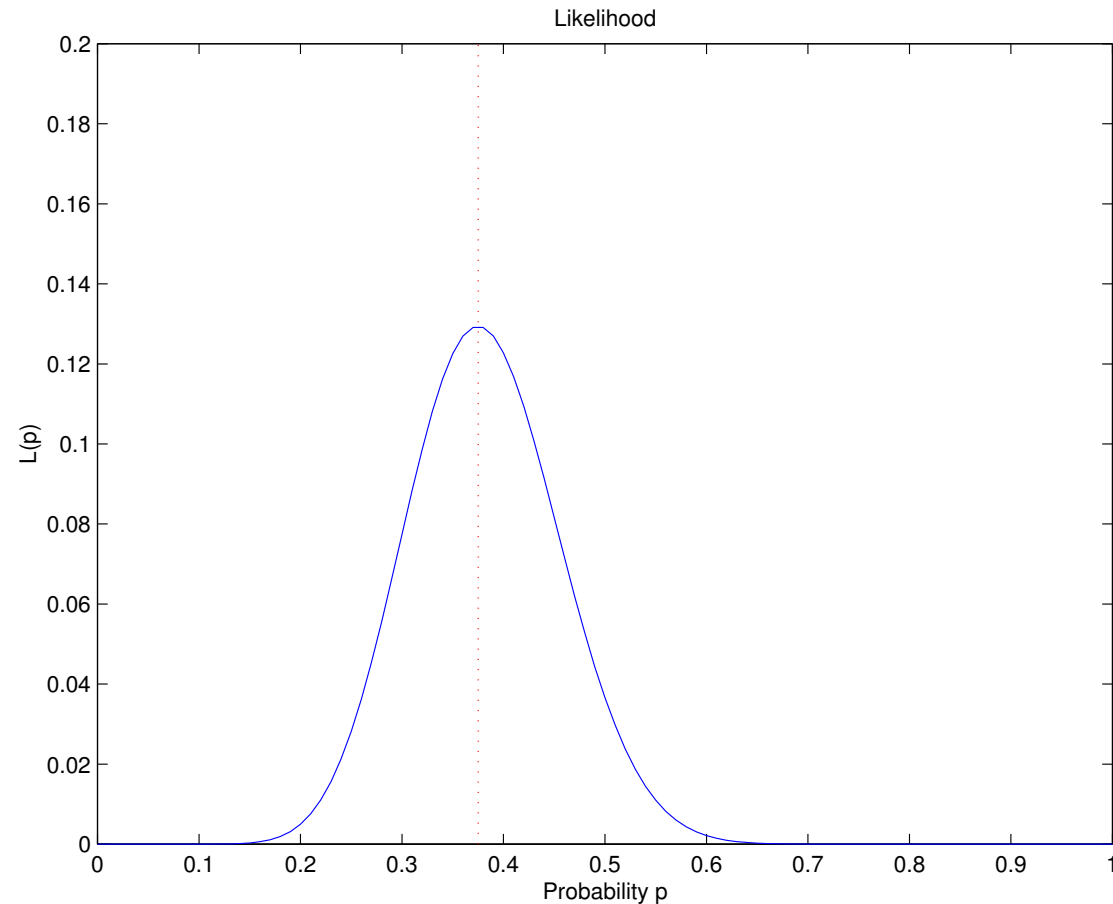
# Example: maximum likelihood

- Consider a binary variable (e.g. a coin) which takes  $z = 15$  times the value 1 in  $N = 40$  trials.
- Suppose that the probabilistic model underlying the data is Binomial with an unknown probability  $\theta = p$ .
- The likelihood  $L(p)$  is a function of (only) the unknown parameter  $p$ .
- By applying the maximum likelihood technique we have

$$\hat{p} = \arg \max_p L(p) = \arg \max_p \binom{N}{z} p^z (1 - p)^{N-z}$$

# Example: maximum likelihood (II)

By plotting  $L(p)$ ,  $p \in [0, 1]$  we have



Then the most likely value of  $p$  according to the data is  $\hat{p} \approx 0.4$ . Note that in this case  $\hat{p} = z/N$ .

MATLAB script `ml_bin.m`

# Some considerations

- The likelihood the relative abilities of the various parameter values to *explain* the observed data.
- The principle of m.l. is that the value of the parameter under which the obtained data would have had highest probability of arising must be **intuitively** our best estimator of  $\theta$ .
- M.l. can be considered a measure of how plausible the parameter values are in light of the data.
- The likelihood function is NOT a probability function.

# Example: log likelihood

Consider the previous example: binomial distribution with  $N = 40$ ,  $z = 15$ .

The log-likelihood for this model is

$$\begin{aligned}\log L(p) &= \log \binom{N}{z} + z \log(p) - (N - z) \log(1 - p) = \\ &= \log \binom{40}{15} + 15 \log p - 25 \log(1 - p)\end{aligned}$$

# M.I. estimation

In many situations the log-likelihood  $l_N(\theta)$  is particularly well behaved in being continuous with a single maximum away from the extremes of the range of variation of  $\theta$ .

Then  $\hat{\theta}$  is obtained simply as the solution of

$$\frac{\partial l_N(\theta)}{\partial \theta} = 0$$

subject to

$$\frac{\partial^2 l_N(\theta)}{\partial \theta^2} \Big|_{\hat{\theta}_{\text{ml}}} = 0$$

to ensure that the identified stationary point is a maximum.

# Maximum likelihood estimators

- Let  $D_N$  be a random sample from the r.v.  $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ .
- The likelihood of the  $N$  samples is given by

$$L_N(\mu, \sigma^2) = \prod_{i=1}^N p_{\mathbf{z}}(z_i, \mu, \sigma^2) = \left( \frac{1}{\sigma^N (2\pi)^{N/2}} \right) \exp \left[ \frac{-\sum_{i=1}^N (z_i - \mu)^2}{2\sigma^2} \right]$$

- The log-likelihood is

$$l_N(\mu, \sigma^2) = -\frac{C}{2\sigma^2} \sum_{i=1}^N (z_i - \mu)^2$$

- Note that, for a given  $\sigma$ , maximizing the log-likelihood is equivalent to minimize the sum of squares of the difference between  $z_i$  and the mean.

# Maximum likelihood estimators (II)

- Taking the derivatives with respect to  $\mu$  and  $\sigma^2$  and setting them equal to zero, we obtain

$$\hat{\mu}_{\text{ml}} = \frac{\sum_{i=1}^N z_i}{N} = \hat{\mu}$$
$$\hat{\sigma}_{\text{ml}}^2 = \frac{\sum_{i=1}^N (z_i - \hat{\mu}_{\text{ml}})^2}{N} \neq \hat{\sigma}^2$$

- Note that the m.l. estimator of the mean coincides with the sample average but that the m.l. estimator of the variance differs from the sample variance for the different denominator.

# Problems with m.l. estimation

Computational difficulties may arise if

1. No explicit solution exists for  $\partial l_N(\theta)/\partial\theta = 0$ . Iterative numerical methods must be used (see *Calcul numérique*). This is particularly serious for a vector of parameters  $\theta$  or when there are several relative maxima of  $l_N$  .
2.  $l_N(\theta)$  may be discontinuous, or have a discontinuous first derivative, or a maximum at an extremal point.

# Properties of m.l. estimators (I)

Under the (strong) assumption that the probabilistic model structure is known, the maximum likelihood technique features the following properties:

- $\hat{\theta}_{ml}$  is consistent.
- $\hat{\theta}_{ml}$  is *asymptotically biased* but usually biased in small samples.
- if a MVB estimator exists, this is the same as  $\hat{\theta}_{ml}$ . In this case  $\hat{\theta}_{ml}$  is optimum.
- the variance of  $\hat{\theta}_{ml}$  is often difficult to determine. For large samples we can use as approximation

$$\left( -E \left[ \frac{\partial^2 l_N}{\partial \theta^2} \right] \right)^{-1} \quad \text{or} \quad \left( -\frac{\partial^2 l_N}{\partial \theta^2} \Big|_{\hat{\theta}_{ml}} \right)^{-1}$$

# Properties of m.l. estimators (II)

- If  $\hat{\theta}_{\text{ml}}$  is the m.l.e. of  $\theta$ ,  $\gamma(\hat{\theta}_{\text{ml}})$  is the m.l.e. of  $\gamma(\theta)$ .
- $\hat{\theta}_{\text{ml}}$  is asymptotically fully efficient, that is

$$\text{Var} \left[ \hat{\theta}_{\text{ml}} \right] \rightarrow [I_N(\theta)]^{-1} = [NI(\theta)]^{-1}$$

- $\hat{\theta}_{\text{ml}}$  is asymptotically normally distributed, that is

$$\hat{\theta}_{\text{ml}} = \mathcal{N}(\theta, [I_N(\theta)]^{-1})$$

- the score is asymptotically normally distributed

$$\frac{\partial \mathbf{l}_N}{\partial \theta} = \mathcal{N}(0, I_N(\theta))$$

# Least squares

Suppose that the sample data  $D_N$  is generated by the linear model

$$D_N = H\theta + w$$

where  $D_N$  is an  $[N \times 1]$  vector,  $H$  is a  $[N \times n]$  matrix ( $N > n$ ),  $\theta$  is a  $[n \times 1]$  parameter vector and  $w$  is the realization of a  $[N \times 1]$  vector  $\mathbf{w}$  whose components are the errors or *noise* associated with the observations. We assume that  $E[\mathbf{w}] = 0$  and  $\text{Var}[\mathbf{w}] = V\sigma^2$  where the  $(N \times N)$  symmetric matrix is known precisely..

**Definition 9.** *The linear estimator  $\hat{\theta}_{ls} = B\mathbf{D}_N$  is called a least-squares (LS) estimator if*

$$B = (H^T H)^{-1} H^T \quad (2)$$

# Some properties

The LS estimator has a number of interesting properties that account for its widespread use in estimation problems.

- It is simple to construct
- it is unbiased
- it has variance  $\text{Var} \left[ \hat{\boldsymbol{\theta}}_{\text{ls}} \right] = (H^T V^{-1} H)^{-1} \sigma^2$
- it has a minimum variance property among all linear unbiased estimator (BLUE or Best Linear Unbiased Estimator).
- if the multivariate distribution of  $\mathbf{w}$  is normal,  $\hat{\boldsymbol{\theta}}_{\text{ls}}$  is also the m.l. estimator.