# Statistical Inference

S.M. Iacus

Department of Economics, Management and Quantitative Methods, University of Milan
CREST JST

June 25, 2019

**YSS2019**

# What is an estimator?

Suppose to have a sample of i.i.d. observations $X_i$, $i = 1, \ldots, n$, with common distribution indexed by some parameter $\theta \in \Theta$, say $\{P_\theta . \theta \in \Theta\}$.

An estimator $T_n$ of $\theta$ is any function solely of the data $X$ and not $\theta$, i.e. $T_n = f(X_1, \ldots, X_n)$.

An estimator is said to be Unbiased if its **Bias** is zero, i.e.

$$\mathrm{Bias}_\theta(T_n) = \mathbf{E}_\theta(T_n) - \theta, \quad \forall \theta \in \Theta$$

We can also study the variance of an estimator, define din the usual way

$$\mathrm{Var}_\theta(T_n) = \mathbf{E}_\theta(T_n - \mathbf{E}\,T_n)^2$$

# Mean Squared Error (MSE)

As always in statistics, there is a tradeoff between mean and variance. One measure of the quality of an estimator is given by the **Mean Squared Error** or simply MSE, defined as follows

$$\mathrm{MSE}_\theta(T_n) = \mathbf{E}_\theta(T_n - \theta)^2, \quad \forall \theta \in \Theta$$

Which can be decomposed as

$$\begin{aligned}
\mathrm{MSE}_\theta(T_n) &= \mathbf{E}_\theta(T_n - \theta \pm \mathbf{E}T_n)^2 \\
&= \mathbf{E}_\theta(T_n - \mathbf{E}T_n)^2 + \mathbf{E}_\theta(\mathbf{E}T_n - \theta)^2 + 2\mathbf{E}_\theta\{(T_n - \mathbf{E}T_n)(\mathbf{E}T_n - \theta)\} \\
&= \mathrm{Var}_\theta(T_n) + (\mathrm{Bias}_\theta(T_n))^2
\end{aligned}$$

# Relative Efficiency

Given two estimators $T_n^{(1)}$ and $T_n^{(2)}$ of $\theta$ we prefer $T_n^{(1)}$ to $T_n^{(2)}$ it

$$\mathrm{MSE}_\theta\left(T_n^{(1)}\right) \leq \mathrm{MSE}_\theta\left(T_n^{(2)}\right), \quad \forall \theta \in \Theta$$

or, equivalently, if the **relative efficiency** behaves as follows:

$$e_\theta\left(T_n^{(1)}, T_n^{(2)}\right) = \frac{\mathrm{MSE}_\theta\left(T_n^{(1)}\right)}{\mathrm{MSE}_\theta\left(T_n^{(2)}\right)} \leq 1, \quad \forall \theta \in \Theta$$

quite often too optimistic! Often the best estimator is searched within the class of unbiased estimators (e.g. OLS), i.e. only variance is compared.

# Asymptotic properties of an estimator

An estimator $T_n$ can be biased for finite $n$ but **asymptotically unbiased** if:

$$\mathrm{Bias}_\theta(T_n) \to 0, \quad \text{as} \quad n \to \infty$$

An estimator is said to **consistent** if:

$$T_n \xrightarrow{p} \theta, \quad \text{as} \quad n \to \infty$$

# The likelihood function

Suppose to have a sample of i.i.d. observations $X_i$, $i = 1, \ldots, n$, with common distribution indexed by some parameter $\theta \in \Theta$. Seen as a random vector, the sample $(X_1, X_2, \ldots, X_n)$ has its own probability.

So, for a given set of observed values $(x_1, x_2, \ldots, x_n)$ from the random vector $(X_1, X_2, \ldots, X_n)$, we might wonder about which is the probability that these data come from a given model specified by $\theta$.

Assume that the $X_i$'s are discrete random variables with probability mass function $p(x; \theta) = P_\theta(X = x)$. Let us construct the probability of the observed sample as

$$P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

# The likelihood function for discrete r.v.'s

$$P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

Seen only as a function of $\theta \in \Theta$ and given the observed values $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, this quantity if called the

"*likelihood* of $\theta$ given the sample data"

and we write

$$L_n(\theta) = L_n(\theta | x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

# The likelihood function for continuous r.v.'s

In case of continuous random variables with density function $f(x;\theta)$ we denote the likelihood as

$$L_n(\theta) = L_n(\theta|x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i;\theta).$$

Now recall that $f(x) \neq P(X = x) = 0$ for continuous random variables (but $f(x)dx \simeq P\{X \in [x, x+dx)\}$), so it is important to interpret $L_n(\theta)$ as the likelihood of $\theta$, rather than the "probability of the sample".

Indeed, $L_n(\theta)$ weights different values of $\theta \in \Theta$ on the basis of the observed (and given) data. This allows to define a general approach in the search of estimators of the unknown parameter $\theta$ as we will see shortly.

# How to use the likelihood?

Consider again the case of discrete random variables with our data $(x_1, \ldots, x_n)$ at hands/given/not changeable/etc

$$L_n(\theta) = L_n(\theta | x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

Suppose that for some value $\theta = \theta_1$ we observe that

$$L_n(\theta_1) > L_n(\theta_2)$$

where $\theta_2$ is another value of $\theta$.

# How to use the likelihood?

Consider again the case of discrete random variables with our data $(x_1, \ldots, x_n)$ at hands/given/not changeable/etc

$$L_n(\theta) = L_n(\theta | x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

Suppose that for some value $\theta = \theta_1$ we observe that

$$L_n(\theta_1) > L_n(\theta_2)$$

where $\theta_2$ is another value of $\theta$.

In this case, we do think that it is more likely that the observed data $(x_1, \ldots, x_n)$ come from the model $P_{\theta_1}$ rather than the model $P_{\theta_2}$.

# How to use the likelihood?

Consider again the case of discrete random variables with our data $(x_1, \ldots, x_n)$ at hands/given/not changeable/etc

$$L_n(\theta) = L_n(\theta | x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

Suppose that for some value $\theta = \theta_1$ we observe that

$$L_n(\theta_1) > L_n(\theta_2)$$

where $\theta_2$ is another value of $\theta$.

In this case, we do think that it is more likely that the observed data $(x_1, \ldots, x_n)$ come from the model $P_{\theta_1}$ rather than the model $P_{\theta_2}$.

Is there a particular value of $\theta \in \Theta$ which makes $L_n(\theta)$ the highest ? If so, we can take this value as an estimate of $\theta$. Indeed!

# How to use the likelihood?

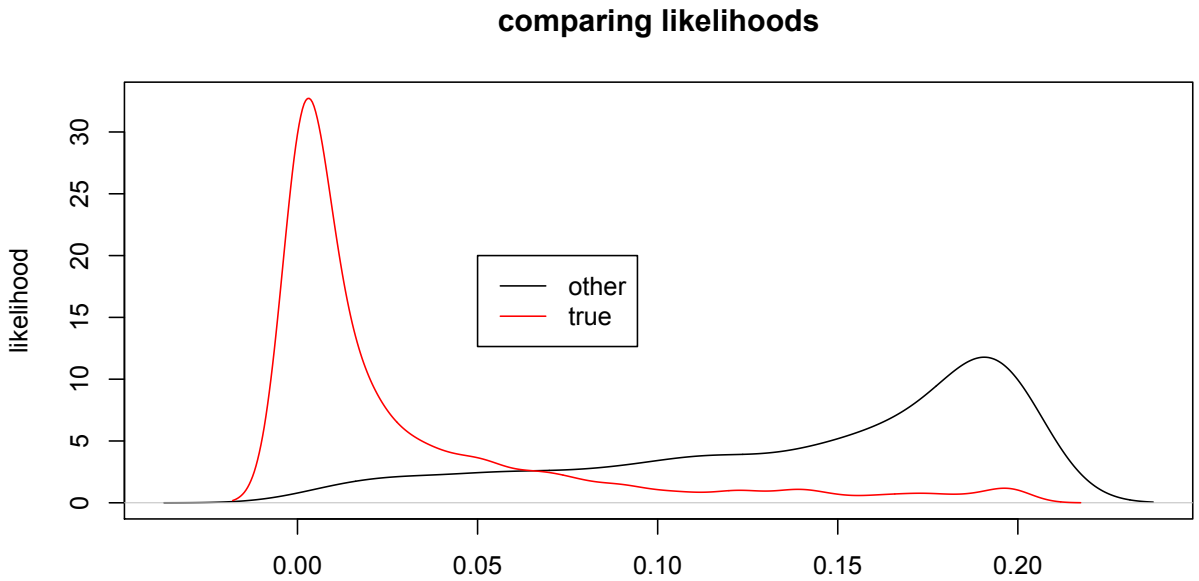
```
R> set.seed(123)
R> library("stats4")
R> x <- rnorm(1000, mean = 5, sd = 2)

R> lik1 <- dnorm(x, mean = 5, sd = 2)
R> lik2 <- dnorm(x, mean = 10, sd = 2)
R> lik1[1:10]
   [1] 0.17047746 0.19425636 0.05919776 0.19897593 0.19781098 0.04583031
   [7] 0.17936946 0.08961114 0.15755682 0.18061402
R> lik2[1:10]
   [1] 0.0018448795 0.0048005716 0.1280799331 0.0104275863 0.0120074975
   [6] 0.1465853830 0.0249466166 0.0001665977 0.0012431569 0.0026044131
R> lik1[1:10] > lik2[1:10] # most of the times true
   [1]  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE

R> sum(lik1 > lik2)/1000*100
  [1] 89.9
```

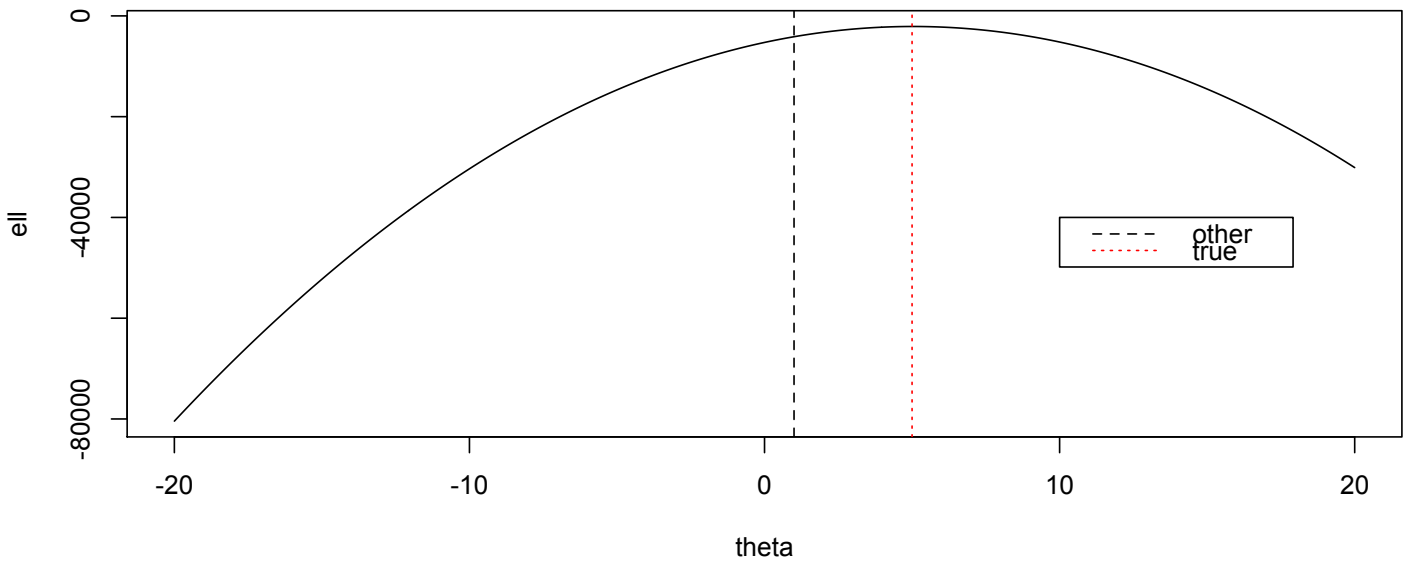
# How to use the likelihood?

```
R> d1 <- density(lik1)
R> d2 <- density(lik2)
R> xlim = range(c(d1$x, d2$x))
R> ylim = range(c(d1$y, d2$y))
R> plot(d1, xlim=xlim, ylim=ylim,
R+      main = "comparing likelihoods",
R+      xlab="", ylab="likelihood") # likely for most points
R> lines(density(lik2),col="red") # very unlikely for most points
R> legend(0.05, 20, legend=c("other", "true"), col=c("black","red"),lty=1)
```



# How to use the likelihood?

```
R> ell <- NULL
R> theta <- seq(-20,20, length=500)
R> for(mu in theta){
R>  ell <- c(ell, sum(dnorm(x, mean = mu, sd = 2, log=TRUE)))
R> }
R> plot(theta, ell, type="l")
R> abline(v=1, lty=2)
R> abline(v=5, lty=3, col="red")
R> legend(10, -40000,  legend=c("other", "true"), col=c("black","red"),lty=c
    (2,3))
```



# Maximum Likelihood Estimators (MLE)

If we study the likelihood $L_n(\theta)$ as a function of $\theta$ given the $n$ numbers $(X_1 = x_1, \ldots, X_n = x_n)$ and we find that this function has a maximum, we can use this maximum value as an estimate of $\theta$.

In general we define *maximum likelihood estimator* of $\theta$, and we abbreviate this with *MLE*, the following estimator

$$\begin{aligned}\hat{\theta}_n &= \arg\max_{\theta \in \Theta} L_n(\theta) \\ &= \arg\max_{\theta \in \Theta} L_n(\theta | X_1, X_2, \ldots, X_n)\end{aligned}$$

provided that the maximum exists.

**Remark:** MLE's are often biased but asymptotically unbiased (see later example)

# Log-likelihood function

We do normally use the log-likelihood function $\ell_n(\theta) = \log L_n(\theta)$ rather than the likelihood function $L_(\theta)$ because, especially in the i.i.d. setting, transforms the product into a sum

$$\ell_n(\theta) = \log L_n(\theta) = \sum \log p(x_i; \theta)$$

or

$$\ell_n(\theta) = \log L_n(\theta) = \sum \log f(x_i; \theta)$$

Moreover, being the log a monotonic function, for numerical purposes, the problem is transformed from

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} L_n(\theta) \quad \text{into} \quad \hat{\theta}_n = \arg\min_{\theta \in \Theta} -\ell_n(\theta)$$

# Derived quantities from the likelihood

The quantity $\mathcal{I}_n(\theta)$ defined below is called the *Fisher information* (for the sample)

$$\begin{aligned}\mathcal{I}_n(\theta) &= \mathbf{E}_\theta \left\{ \frac{\partial}{\partial\theta} \ell_n(\theta) \right\}^2 = -\mathbf{E}_\theta \left\{ \frac{\partial^2}{\partial\theta^2} \ell_n(\theta) \right\} \\ &= n\mathcal{I}(\theta)\end{aligned}$$

(where $\mathcal{I}(\theta)$ is the Fisher information for one observation). The quantity

$$\frac{\partial}{\partial\theta} \ell_n(\theta)$$

is called the **score function** of the model.

**Cramér-Rao's theorem:** for any estimator $T_n$ of $\theta$

$$\mathrm{Var}_\theta(T_n) \geq \frac{\left(1 + \frac{\partial}{\partial\theta}\mathrm{Bias}_\theta(T_n)\right)^2}{\mathcal{I}_n(\theta)}$$

which means: no matter $T_n$ the lower bound of the variance is controlled by the Fisher information of the model.

# MLE and optimality

Clearly, if $\hat{\theta}_n$ is the MLE of $\theta$, then

$$\frac{\partial}{\partial \theta} \ell_n(\theta) \Big|_{\theta=\hat{\theta}_n} = 0$$

because the above represents the normal equation in optimization.
If $T_n$ is the ML estimator, we have that

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1})$$

Therefore MLE, being asymptotically unbiased, are also **asymptotically optimal** because they reach the asymptotic lower bound of Cramér-Rao.

# A simple example: $N(\theta, \sigma^2)$

Let $X_i$, $i = 1, \ldots, n$, be an i.i.d. sample extracted from the Gaussian distribution $\mathrm{N}(\theta, \sigma^2)$. For simplicity, assume $\sigma^2$ is known. We want to find the MLE of $\theta$. Hence

$$L_n(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\theta)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(X_i-\theta)^2}{2\sigma^2}}.$$

Instead of maximizing $L_n(\theta)$ we maximize the log-likelihood $\ell_n(\theta) = \log L_n(\theta)$

$$\ell_n(\theta) = n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{n} \frac{(X_i - \theta)^2}{2\sigma^2}$$

# A simple example: $N(\theta, \sigma^2)$

The first term of $\ell_n(\theta)$ contains only constants so it does not affect optimization, hence we just need to solve

$$\hat{\theta}_n = \arg\min_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \theta)^2$$

but this minimum is exactly $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ by the properties of the arithmetic mean.[1] Hence the ML estimator of $\theta$ is $\hat{\theta}_n = \bar{X}_n$. Explicit calculations lead to the same result

$$\frac{\partial}{\partial \theta} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \theta)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \theta) = 0 \Leftrightarrow$$

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

[1] It is easy to show that: $\min_a \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$.

# Example: continues

Consider the setup of the previous example. Find the maximum likelihood estimator of $\theta = (\mu, \sigma^2)$.

We minimize minus the log-likelihood function as a function of $\mu$ and $\sigma$

$$h(\mu, \sigma^2) = -\ell_n(\mu, \sigma^2) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\sigma^2 + \sum_{i=1}^{n}\frac{(X_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\mu}h(\mu, \sigma^2) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0$$

$$\frac{\partial}{\partial\sigma^2}h(\mu, \sigma^2) = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

From the first equation we get $\hat{\mu} = \bar{X}_n$ and pluggin-in this value into the second equation we obtain $\hat{\sigma}^2 = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_1 - \bar{X}_n)^2$.

# Example: continues

We now need to verify that at least one of the two second derivatives at point $(\hat{\mu}, \hat{\sigma}^2)$ is positive and the determinant of the Hessian matrix of second-order partial derivatives of $h(\mu, \sigma^2)$ evaluated at the point $(\hat{\mu}, \hat{\sigma}^2)$ is positive. So we calculate partial derivatives first.

$$\begin{aligned}
\frac{\partial^2}{\partial \mu^2} h(\mu, \sigma^2) &= \frac{n}{\sigma^2}, \\
\frac{\partial^2}{\partial (\sigma^2)^2} h(\mu, \sigma^2) &= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^{n} (X_i - \mu)^2, \\
\frac{\partial^2}{\partial \mu \sigma^2} h(\mu, \sigma^2) &= +\frac{1}{\sigma^4} \sum_{i=1}^{n} (X_i - \mu).
\end{aligned} \tag{1}$$

# Example: continues

Now we recall that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu})^2$ and $\sum_{i=1}^n (X_i - \hat{\mu}) = 0$, hence

$$\frac{\partial^2}{\partial \mu^2} h(\mu, \sigma^2)\Big|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} = \frac{n}{\hat{\sigma}^2} > 0,$$

$$\frac{\partial^2}{\partial (\sigma^2)^2} h(\mu, \sigma^2)\Big|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^4} + \frac{n}{\hat{\sigma}^4} = \frac{n}{2\hat{\sigma}^4} > 0,$$

$$\frac{\partial^2}{\partial \mu \sigma^2} h(\mu, \sigma^2)\Big|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^4}\sum_{i=1}^n (X_i - \hat{\mu}) = 0.$$

Finally, we calculate the determinant of the Hessian matrix evaluated at point $(\hat{\mu}, \hat{\sigma}^2)$ to check if it is positive

$$H(\hat{\mu}, \hat{\sigma}^2) = \begin{vmatrix} \frac{\partial^2}{\partial \mu^2} h(\mu, \sigma^2) & \frac{\partial^2}{\partial \mu \sigma^2} h(\mu, \sigma^2) \\ \frac{\partial^2}{\partial \mu \sigma^2} h(\mu, \sigma^2) & \frac{\partial^2}{\partial (\sigma^2)^2} h(\mu, \sigma^2) \end{vmatrix}_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2}$$

$$= \begin{vmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{vmatrix} = \frac{1}{2}\frac{n^2}{\hat{\sigma}^6} > 0.$$

# Example: continues - Fisher information

From (1) we can obtain the Fisher information matrix of the Gaussian model. Indeed

$$-\mathbf{E}_\theta\left(\frac{\partial^2}{\partial\mu^2}\ell_n\right) = \frac{n}{\sigma^2}$$

$$-\mathbf{E}_\theta\left(\frac{\partial^2}{\partial(\sigma^2)^2}\ell_n\right) = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} = \mathbf{E}_\theta\left(\sum_{i=1}^{n}(X_i-\mu)^2\right)$$

$$= -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} = \frac{n}{2\sigma^4}$$

$$-\mathbf{E}_\theta\left(\frac{\partial^2}{\partial\mu\sigma^2}\ell_n\right) = +\frac{1}{\sigma^4}\mathbf{E}_\theta\left(\sum_{i=1}^{n}(X_i-\mu)\right) = 0$$

$$\mathcal{I}_n(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} = n\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} = n\mathcal{I}(\theta)$$

# MLE with R

It is not always the case, that maximum likelihood estimators can be obtained in explicit form.

For what concerns applications to real data, it is important to know if mathematical results about optimality of MLE estimators exist and then find the estimators numerically.

R offers a prebuilt generic function called `mle()` in the package package `stats4` which can be used to maximize a likelihood.

The `mle()` function actually minimizes the negative log-likelihood $-\ell(\theta)$ as a function of the parameter $\theta$.

# MLE with R

For example, consider a sample of $n = 1000$ observations from a Gaussian law with $N(\mu = 5, \sigma^2 = 4)$, and let us estimate the parameters numerically:

```
R> set.seed(123)
R> library("stats4")
R> x <- rnorm(1000, mean = 5, sd = 2)
R> log.lik <- function(mu = 1, sigma = 1) -sum(dnorm(x, mean = mu,
+     sd = sigma, log = TRUE))
R> fit <- mle(log.lik, lower = c(0, 0), method = "L-BFGS-B")
R> fit
Call:
mle(minuslogl = log.lik, method = "L-BFGS-B", lower = c(0, 0))

Coefficients:
      mu    sigma
5.032256 1.982398
```

# MLE with R

Using explicit estimators for $\mu$ and $\sigma^2$ we get:

```
R> mean(x)
[1] 5.032256
R> sd(x)
[1] 1.98339
```

which almost coincides numerically.

What is worth knowing is that the output of the `mle()` function is an object which contains several informations, including the value of $\ell(\theta)$ at the point of its maximum

```
R> logLik(fit)
'log Lik.' -2103.246 (df=2)
```

# MLE with R

The variance-covariance matrix of the estimators, which is obtained inverting the Hessian matrix at the point $\theta$ corresponding to the maximum likelihood estimate

```
R> vcov(fit)
                  mu         sigma
mu    3.929901e-03 8.067853e-10
sigma 8.067853e-10 1.964946e-03
```

# MLE with R

Similarly, approximate confidence intervals and the complete summary of the estimated parameters can be obtained using respectively the functions `confint()` and `summary()`

```
R> confint(fit)
Profiling...
         2.5 %   97.5 %
mu    4.909269 5.155242
sigma 1.898595 2.072562
R> summary(fit)
Maximum likelihood estimation

Call:
mle(minuslogl = log.lik, method = "L-BFGS-B", lower = c(0, 0))

Coefficients:
      Estimate Std. Error
mu    5.032256 0.06268893
sigma 1.982398 0.04432771

-2 log L: 4206.492
```

**R code**: MLE.R