Likelihood theory: Maximum likelihood estimation

(An overview)

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

The likelihood function[1]

Maximum likelihood estimation: theory

Some numerical aspects

[1] Agresti, Kateri: sec 4.2

The likelihood function

The likelihod function

IS A FUNDAMENTAL IDEA IN STATISTICS USED TO ESTIMATE PARAMETERS OF A PROBABILITY DISTRIBUTION BASED ON OBSERVED DATA. IT QUANTIFIES HOW LIKELY IT IS TO OBSERVE THE GIVEN DATA, ASSUMING THE MODEL AND ITS PARAMETERS ARE CORRECT

Introduced by Sir Ronald Fisher, the **likelihood function** for a certain statistical model $f_\theta(\mathbf{y})$ for the data $\mathbf{y}$ is given by the following function of the parameter $\boldsymbol{\theta}$

WE HAVE TO ASSUME A PARAMETRIC MODEL BECAUSE THE LIKELIHOOD IS A FUNCTION OF THE PARAMETER

$$\begin{aligned} L \quad : \quad & \boldsymbol{\Theta} \to \mathbb{R}^+ \\ & \boldsymbol{\theta} \to c(\mathbf{y})\, f_\theta(\mathbf{y}) \,, \end{aligned}$$

THE MAIN CONCEPT:
IMMAGINE YOU HAVE A STATISTICAL MODEL $P(X,\theta)$ WHERE X RAPPRESENT THE OBSERVED DATA AND θ THE PARAMETERS OF THE MODEL
=> THE LIKELIHOOD FUNCTION IS DEFINED AS:

$$L(\theta, X) = P(X, \theta)$$

IT'S THE PROBABILITY (OR DENSITY) OF THE DATA X, VIEWED AS A FUNCTION OF θ

where $c(\mathbf{y}) > 0$ is an arbitrary constant of proportionality.

We may write $L(\boldsymbol{\theta}; \mathbf{y})$ to stress the fact that the data enter the function, though its argument is given by $\boldsymbol{\theta}$.

IF WE PLOT $L(\theta/y)$, THE SHAPE GIVES INSIGHT INTO HOW "PLAUSIBLE" EACH PARAMETER VALUE θ IS.
A SHARP PEAK INDICATES HIGH CONFIDENCE IN $\hat{\theta}_{MLE}$, WHILE A FLAT LIKELIHOOD SUGGEST UNCERTAINTY.

Interpreting the likelihood function

The likelihood function assigns support (*credibility*) to possible values of $\boldsymbol{\theta}$, meaning that if $L(\boldsymbol{\theta}_1) > L(\boldsymbol{\theta}_2)$ then $\boldsymbol{\theta}_1$ is more supported by the observed data than $\boldsymbol{\theta}_2$. THE HIGHER THE VALUE, THE HIGHER IS THE TRUST IN THE PARAMETER

So the *likelihood ratio* $L(\boldsymbol{\theta}_1)/L(\boldsymbol{\theta}_2)$ allows for the comparison between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$; note that the constant $c(\mathbf{y})$ cancels out.

A mathematical justification for the above interpretation is given by the **Wald inequality**: if $\boldsymbol{\theta}_t$ is the **true parameter value**, then

$$E_{\boldsymbol{\theta}_t}\{\log L(\boldsymbol{\theta}_t; \mathbf{Y})\} > E_{\boldsymbol{\theta}_t}\{\log L(\boldsymbol{\theta}; \mathbf{Y})\} \qquad \boldsymbol{\theta} \neq \boldsymbol{\theta}_t\,.$$

The above fact can be proven by straightforward application of the Jensen's inequality.

LET'S SAY θ_2 IS MORE LIKE TO BE THE TRUE PARAMETER, THEN THE RATIO $L(\theta_1)/L(\theta_2)$ SHOULD ASSUME A VALUE HIGHER THAN 1

The log likelihood function

In the previous slide the **log likelihood function** has been introduced, which is simply the logarithm of $L(\boldsymbol{\theta})$, namely

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) .$$

The log likelihood function carries the same information of the likelihood function, but it is much more manageable. Indeed, for a random sample

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(y_i)$$

but

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(y_i) .$$

$\log(L(\theta)) = \log\left(\prod_{i=1}^{n} f_\theta(y_i)\right)$

OFTEN THE LIKELIHOOD INVOLVES PRODUCTS (FOR MULTIPLE DATA POINTS), WHICH CAN BE COMPUTATIONALLY CHALLENGING AND NUMERICALLY UNSTABLE. THE LOG-LIKELIHOOD SIMPLIFIES THIS WHILE PRESERVING THE MAXIMA AND IS OFTEN EASIER TO INTERPRET IN TERMS OF ADDITIVE CONTRIBUTIONS OF INDIVIDUAL DATA POINTS.

Notice that $\ell(\boldsymbol{\theta})$ is defined up to an additive constant, depending only on the data **y**.

Example 1: the Poisson model

For a random sample $y_1, \ldots, y_n$, with $Y_i \sim \mathcal{P}(\lambda)$ i.i.d., we readily get

$$L(\lambda) = \frac{\lambda^{\sum_{i=1}^n y_i} \exp\{-n\,\lambda\}}{\prod_{i=1}^n y_i!}\,,$$

so that

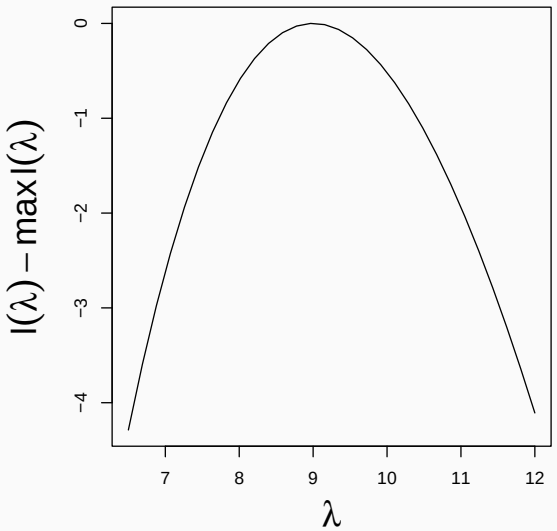
$$\ell(\lambda) = \log(\lambda) \sum_{i=1}^n y_i - n\,\lambda\,,$$

neglecting the term which does not depend on λ.

R lab: the Poisson log likelihood

Assume that for a sample $n = 10$ we observe $\sum_i y_i = 90$.

```
lik_pois <- function(lam, n, sumy) log(lam) * sumy - n * lam
xx <- seq(6.5, 12, l = 30)
ll <- sapply(xx, lik_pois, sumy = 90, n = 10)
par(pty = "s")
plot(xx, ll - max(ll), type = "l", xlab = expression(lambda),
     ylab = expression(l(lambda)-max(l(lambda))), cex.lab = 2)
```



Example 2: the normal model

For a random sample $y_1, \ldots, y_n$, with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left\{ -\frac{(y_i - \mu)^2}{2\,\sigma^2} \right\},$$

and then with some simple algebra

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 .$$

Sufficient statistics

The definition of sufficient statistic, given in the probability part, can be re-interpreted for the log likelihood function: $t(\mathbf{y})$ is **sufficient** for θ if $L(\theta)$ can be written as

FUNCTION THAT ONLY DEPENDS ON THE DATA

FUNCTION THAT DEPENDS ON THE PARAMETER AND ON THE DATA THROUGHT THE SUFFICIENT STATISTIC (SUM y, SUM y^2, ECC..)

$$L(\theta) = h(\mathbf{y})\, g_\theta\{t(\mathbf{y})\}.$$

The **minimal sufficient statistic** allows for the maximal reduction of dimensionality, in the sense that a minimal sufficient statistic is a function of every other sufficient statistic.

MINIMAL SUFFICIENT STATISTIC BECAUSE YOU CAN'T REDUCE IT MORE

For the Poisson model, the $\sum_i y_i$ (or, equivalently, the sample mean $\overline{y}$) is sufficient for λ, whereas for the normal model the sufficient statistic is given by the pair $(\sum_i y_i, \sum_i y_i^2)$ (or, equivalently, by the pair $(\overline{y}, s^2)$).

THEY CAN BE WRITE IN THE FORM ABOVE

These two statistical models are an instance of an **exponential family**, an important model class that includes also other important elements, such as the binomial distribution. They play an important role in the theory of *generalized linear models*. AND HAVE A LOT OF USEFULL DISTRIBUTION

WHAT ARE SUFFICIENT STATISTICS?

A SUFFICIENT STATISTIC IS A SUMMARY OF THE DATA THAT CONTAINS ALL THE INFORMATION NEEDED TO COMPUTE THE LIKELIHOOD OF A PARAMETER θ.

IN SIMPLE TERMS:

> ONCE YOU HAVE THE SUFFICIENT STATISTIC, YOU DON'T NEED RAW DATA ANYMORE TO ESTIMATE θ.

FORMALLY, A STATISTIC $T(y)$ IS SUFFICIENT FOR θ IF THE CONDITIONAL DISTRIBUTION OF THE DATA y, GIVEN $T(y)$, DOES NOT DEPEND ON θ.

MATHEMATICALLY:

$$P(y \mid T(y), \theta) = P(y \mid T(y))$$

EXAMPLE: FOR A NORMAL DISTRIBUTION $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$:

THE SAMPLE MEAN $\bar{X}$ AND THE SAMPLE VARIANCE S^2 ARE SUFFICIENT FOR μ AND σ^2.

FOR A BINOMIAL $X \sim Bin(n, p)$:

THE NUMBER OF SUCCESSES $\sum X_i$ IS SUFFICIENT FOR p.

SUFFICIENT STATISTIC ARE CRITICAL TO MLE BECAUSE THEY REDUCE THE COMPLEXITY OF THE ESTIMATION PROBLEM:

- INSTEAD OF WORKING WITH THE FULL DATA X, WE CAN FOCUS ONLY ON $T(X)$, THE SUFFICIENT STATISTICS.

THIS IS ESPECIALLY USEFUL WHEN:

- THE DATA IS LARGE, AND SUMMARIZING IT REDUCES COMPUTATIONAL OVERHEAD
- THE SUFFICIENT STATISTIC MAKES IT EASIER TO WRITE DOWN AND SOLVE THE LIKELIHOOD EQUATIONS.

Maximum likelihood estimation

Given the interpretation of the (log) likelihood, the maximum of $\ell(\boldsymbol{\theta})$ is the value of the parameter which is most supported by the data.

A natural step is to take it as the point estimate, the **maximum likelihood estimate** (MLE) of $\boldsymbol{\theta}$

IDENTIFY THE PARAMETER θ THAT MAXIMISE THE LIKELIHOOD

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta})$$

Notice that since $\ell(\theta)$ is also a function of $\mathbf{y}$, the MLE is a statistic.

The MLE in the two examples

For the Poisson model, simple calculus gives

$$\widehat{\lambda} = \frac{1}{n} \sum_{i=1}^{y} y_i = \overline{y}\,.$$

For the normal model, we need to maximize a function of two variables, and we get

$$\begin{cases} \widehat{\mu} = \overline{y} \\ \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2\,. \end{cases}$$

MLE: comments

Maximum likelihood estimation has a **central role** in modern statistics (and machine learning). There are several reasons for this:

1. The MLE algorithm is **automatic**: given a parametric statistical model for the data, the MLE follows from the chosen model.
2. The MLE of a function of a parameter $\psi = g(\boldsymbol{\theta})$ is defined by the simple plug-in rule $\widehat{\psi} = g(\widehat{\boldsymbol{\theta}})$, which is very convenient for applications.
3. The MLE has **excellent properties**, which we illustrate in what follows.

RECAP

WE HAVE DEFINED THE LIKELIHOOD FUNCTION FOR A PARAMETER θ, ASSUMING AN iid SAMPLE, AS THE PRODUCT OF UNIVARIATE PROBABILITY DENSITY FUNCTION:

$$L(\theta) = \prod_{i=0}^{n} f_\theta(y_i)$$

AND THUS COME FROM THE JOINT PROBABILITY DENSITY FUNCTION.

WHEN WE OBSERVED OUR SAMPLE AND INSERT THEM INTO THE LIKELIHOOD FUNCTION, WE CONSIDER IT LIKE A FUNCTION OF THE PARAMETER AND WE CAN SEE WHICH VALUES OF THE PARAMETER SPACE ARE THE MORE "SUPPORTED".

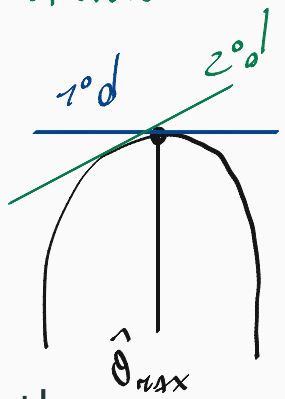
THINK OF THE LIKELIHOOD AS TRYING TO LOCATE A "SWEET SPOT" FOR A MACHINE SETTING (θ) THAT MAKES IT PRODUCE THE OBSERVED OUTCOMES (X) MOST CONSISTENLY.
THE SUFFICIENT STATISTIC IS LIKE A DASHBOARD SUMMARY THAT CONDENSES ALL RELEVANT DATA, SO YOU DON'T HAVE TO INSPECT EVERY PIECE OF THE MACHINE'S HISTORY TO FIGURE OUT THE OPTIMAL SETTING.

Maximum likelihood estimation: theory

Likelihood quantities

THE FIRST DERIVATIVE IDENTIFY THE MAX
THE SECOND DERIVATE IDENTIFY THE RATE AT WHICH WE OBSERVE INFORMATION

The first two derivatives of $\ell(\boldsymbol{\theta})$ play an important role.



The vector of first derivatives is called the **score function**

↳ THE GRADIENT

$$U(\boldsymbol{\theta}) = U(\boldsymbol{\theta}; \mathbf{y}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$U(\hat{\theta}_{MLE}) = 0$

The matrix of second derivatives, with negative sign, is called the **observed information matrix**:

A STEPPER CURVATURE INDICATES MORE CONFIDENCE IN THE PARAMETER ESTIMATE

$$J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}; \mathbf{y}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

IF θ IS A VECTOR, WE OBTAIN A VECTOR OF DERIVATIVES. IF SO, WE HAVE TO COMPUTE A SYSTEM TO OBTAIN THE MAX

$\theta = (\mu, \sigma^2)$

$$\begin{cases} \frac{\partial \ell(\theta)}{\partial \mu} = 0 \\ \frac{\partial \ell(\theta)}{\partial \sigma^2} = 0 \end{cases}$$

$$U(\theta) = \begin{pmatrix} \frac{\partial \ell(\theta)}{\partial \mu} \\ \frac{\partial \ell(\theta)}{\partial \sigma^2} \end{pmatrix}$$

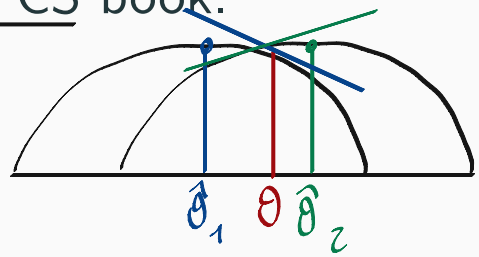
IF WE WANT THE 2° DERIVATIVES OF A VECTOR, WE OBTAIN A MATRIX THAT DEPENDS ON THE SIZE

Some properties

THIS PROPRERITES ARE ASSUMED IF WE CAN REPRODUCE THE TEST A LARGE AMMOUNT OF TIME

The derivatives of the log likelihood function satisfy some important properties, provided that some **regularity conditions** hold (we shall return on them later on).

The proofs are simple, and they are reported in the CS book.



1. *Zero expected score*

$$E_{\boldsymbol{\theta}}\left\{U(\boldsymbol{\theta};\mathbf{Y})\right\} = \mathbf{0}$$

2. *2nd Bartlett identity*

ASSUMING ϑ THE TRUE VALUE, IF WE REPEAT SAMPLING AND EVALUATE THE SCORE FUNCTION, ON AVERAGE THE EXPECTATION IS EQUAL TO 0.

$$\mathrm{cov}_{\boldsymbol{\theta}}\left\{U(\boldsymbol{\theta};\mathbf{Y})\right\} = E_{\boldsymbol{\theta}}\left\{J(\boldsymbol{\theta};\mathbf{Y})\right\} = \mathcal{I}(\boldsymbol{\theta})$$

The expected value $\mathcal{I}(\boldsymbol{\theta})$ of the observed information matrix is called the *Fisher information matrix* (or just the *expected information matrix*).

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial \ln f_\theta(y)}{\partial \theta}\right)^2\right]$$

The Cramér-Rao lower bound

The third property is important, and we first state it for a one-parameter model (scalar θ).

3. *The Cramér-Rao lower bound*: the variance of *any unbiased estimator* $\tilde{\theta}$ cannot be smaller than the reciprocal of the expected information:

$$\mathrm{var}_\theta\{\tilde{\theta}(\mathbf{Y})\} \geq \frac{1}{\mathcal{I}(\theta)}.$$

IT PROVIDES A THEORETICAL LOWER BOUND FOR THE VARIANCE OF ANY UNBIASED ESTIMATOR $\hat{\theta}$

Actually, by differentiation of the unbiasedness condition with respect to θ it follows that $\mathrm{cov}_\theta\{\tilde{\theta}, U(\theta; \mathbf{Y})\} = 1$, which readily implies the Cramér-lower bound.

THE MLE IS EFFICENT IN LARGE SAMPLES, MEANING IT ACHIVES THIS LOWER BOUND ASYMPTOTICALLY

The extension to multiparameter models is given by the condition that the matrix $\mathrm{cov}(\tilde{\boldsymbol{\theta}}) = \mathcal{I}(\boldsymbol{\theta})^{-1}$ is positive semi-definite.

DIM OF PROPERTIES 3

$$|\mathrm{CORR}(\tilde{\vartheta}(y), U(\theta))| = \frac{|\mathrm{Cov}(\tilde{\vartheta}(y),\ U(\theta))|}{\sqrt{\mathrm{VAR}(\tilde{\vartheta}(y))\ \mathrm{VAR}(U(\theta))}} \leq 1$$

$$\frac{|\mathrm{Cov}(\tilde{\vartheta}(y),\ U(\theta))|^2}{\mathrm{VAR}(\tilde{\vartheta}(y))\ \mathrm{VAR}(U(\theta))} \leq 1 \quad \Rightarrow \quad \mathrm{VAR}(\tilde{\vartheta}(y)) \geq \frac{\mathrm{Cov}(\tilde{\vartheta}(y),\ U(\theta))^2}{\mathrm{VAR}(U(\theta))} = \frac{1}{\mathcal{I}(\theta)}$$

$$\mathrm{Cov}(\tilde{\vartheta}(y), U(\theta))^2 = \mathbb{E}(\tilde{\vartheta}(y)\, U(\theta)) - \mathbb{E}(\tilde{\vartheta}(y))\, \mathbb{E}(U(\theta)) \quad (\mathbb{E}(U(\theta)) = 0)$$

$$= \mathbb{E}(\tilde{\vartheta}(y)\, U(\theta)) = \int \tilde{\vartheta}(y)\, U(\theta) \cdot f_\theta(y)\, dy$$

$$= \int \tilde{\vartheta}(y)\, \frac{\partial \log f_\theta(y)}{\partial \theta}\, f_\theta(y)\, dy$$

$$= \int \tilde{\vartheta}(y)\, \frac{1}{f_\theta(y)}\, \frac{\partial f_\theta(y)}{\partial \theta}\, f_\theta(y)\, dy$$

$$= \frac{\partial}{\partial \theta} \int \tilde{\vartheta}(y)\, f_\theta(y)\, dy = \frac{\partial \mathbb{E}(\tilde{\vartheta}(y))}{\partial \theta} = \frac{\partial \theta}{\partial \theta} = 1$$

$$\mathrm{VAR}(\tilde{\vartheta}(y)) \geq \frac{1}{\mathcal{I}(\theta)}$$

Consistency of MLE

EX OF NOT CONSISTENT: THE NUMBER OF PARAMETER INCREASES WITH THE SAMPLE SIZE

We are ready to state the first crucial property of the MLE:

Maximum likelihood estimators are usually consistent, that is if the sample size tends to infinity $\widehat{\boldsymbol{\theta}}$ tends to $\boldsymbol{\theta}_t$, the **true** parameter value.

A justification for the result is given by the fact that in regular situations $\ell(\boldsymbol{\theta})/n \to E_{\boldsymbol{\theta}}\{\ell(\boldsymbol{\theta})\}/n$ as $n \to \infty$, so that eventually the maximum of $\ell(\boldsymbol{\theta})$ and $E\{\ell(\boldsymbol{\theta})\}$ must coincide at $\boldsymbol{\theta}_t$ by the Wald inequality.

The formal proof (typically) employs the law of large numbers.

Consistency can fail if the number of parameters increases with the sample size.

Large-sample distribution of MLE

We establish it by a Taylor expansion for the score function:

$$U(\widehat{\boldsymbol{\theta}}) \doteq U(\boldsymbol{\theta}_t) - (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t)\, J(\boldsymbol{\theta}_t)\,,$$

with equality when $n \to \infty$ since $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \to \mathbf{0}$. (TUX TO CONSISTENCY)

From the definition of $\widehat{\boldsymbol{\theta}}$, we get $U(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. Under mild assumptions

$$\frac{J(\boldsymbol{\theta}_t)}{n} \to \frac{\mathcal{I}(\boldsymbol{\theta}_t)}{n}\,,$$

whereas $U(\boldsymbol{\theta}_t)$ is a random vector with mean vector $\mathbf{0}$ and covariance matrix $\mathcal{I}(\boldsymbol{\theta}_t)$.

In the large sample limit

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \,\dot{\sim}\, \mathcal{I}(\boldsymbol{\theta}_t)^{-1}\, U(\boldsymbol{\theta}_t; \mathbf{y})\,,$$

implying that $E(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t) = \mathbf{0}$ and $\mathrm{cov}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t) = \mathcal{I}(\boldsymbol{\theta}_t)^{-1}$.

Large-sample normality of MLE

ON THE AGRESTY BOOK SECTION 2.3.1

In the case when the sample is formed by independent observations, it follows that the log likelihood is the sum of independent contributions: under mild conditions the central limit theorem applies, and in the large sample limit

$$\widehat{\boldsymbol{\theta}} \dot{\sim} \mathcal{N}\{\boldsymbol{\theta}_t, \mathcal{I}(\boldsymbol{\theta}_t)^{-1}\}.$$

Notice that whenever this holds, it would be possible (and recommendable, in some sense) to use $J(\boldsymbol{\theta}_t)$ in place of $\mathcal{I}(\boldsymbol{\theta}_t)$.

Again, since $\boldsymbol{\theta}_t$ is unknown, we replace it by $\widehat{\boldsymbol{\theta}}$, obtaining the following estimated standard error for the k-th component of $\boldsymbol{\theta}$

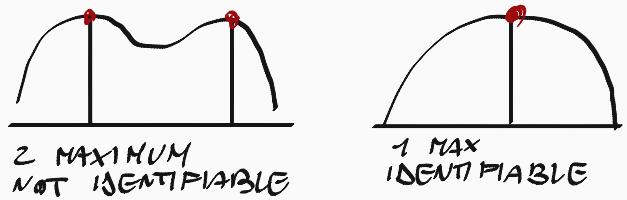
$$\mathrm{SE}(\widehat{\boldsymbol{\theta}}_k) = \sqrt{\left[J(\widehat{\boldsymbol{\theta}})^{-1}\right]_{kk}}$$

→ VALUE ON THE DIAGONAL

Note: for *regular models* (see next slide), the observed information is positive definite at $\widehat{\boldsymbol{\theta}}$, so that the SE above is well defined.

Regularity conditions

We end the summary of the theory by mentioning the **regularity conditions**, which are some assumptions on the statistical model, required for the previous results to be valid.



The CS book lists the following ones:

1. The pdf of $\mathbf{y}$ defined by different values of $\boldsymbol{\theta}$ are distinct, namely the model is *identifiable*. (IDENTIFIABILITY)
2. The true parameter value $\boldsymbol{\theta}_t$ is interior to $\boldsymbol{\Theta}$. (COMPACTNESS)
3. Within some neighbourhood of $\boldsymbol{\theta}_t$, the first three derivatives of $\ell(\boldsymbol{\theta})$ exist and are bounded, while the expected information satisfies the 2nd Bartlett identity, is positive definite and finite. (SMOOTHNESS)

These are mild conditions, which are generally valid in most cases.

Winding up

The previous results have illustrated that

1. The MLE is a **consistent estimator**.
2. The MLE is **asymptotic efficient**, since its asymptotic variance attains the Cramér-Rao lower bound.
3. The large sample distribution (aka the approximate distribution) of the MLE is **multivariate normal**, with standard error that can be estimated by the observed information evaluated at the parameter estimate.

Example 1: Poisson model

$Y_1 \dots Y_n \overset{iid}{\sim} Pois(\lambda)$ $\quad \hat{\lambda} = \frac{\sum y_i}{n} \Rightarrow \sum y_i \sim Pois(n\lambda)$

Here $\widehat{\lambda} = \overline{y}$, and consistency follows from the law of large numbers, in agreement with likelihood theory.

$\hat{\lambda}_{ML} = \frac{\sum y_i}{n}$

WE HAVE TO CHEK IF WE OBTAINE THE SAME RESULT AS THE ONE OBTAINED FROM THE CENTRAL LIMIT THEOREM USING THE RESULTS FROM THE LIKELIHOOD THEORY

Furthermore, the CLT states that for large n

$$\widehat{\lambda} \dot{\sim} \mathcal{N}(\lambda, \lambda/n)\,. \quad (4)$$

$\Rightarrow E[\hat{\lambda}_{ML}] = \lambda$ (2)

$VAR(\hat{\lambda}_{ML}) = E[J(\lambda)]^{-1}$

This result can be obtained also from likelihood theory. Indeed, we get

$$\boxed{J(\lambda) = \frac{\sum_i y_i}{\lambda^2}}$$

so that $\mathcal{I}(\lambda) = n/\lambda$ and $\mathcal{I}(\lambda)^{-1} = \lambda/n$.

$U(\lambda) = \frac{\sum y_i}{\lambda} - n$

$J(\lambda) = -\frac{\partial^2 \ell(\lambda)}{\partial \lambda} \overset{!}{=} -\frac{\partial U(\lambda)}{\partial \lambda} \overset{!}{=} \frac{\partial}{\partial \lambda}\left[\frac{\sum y_i}{\lambda} - n\right] \overset{!}{=} -\left[-\frac{\sum y_i}{\lambda^2}\right] \overset{!}{=} \frac{\sum y_i}{\lambda^2}$

NOW, WE WANT THE FISHER INFORMATION MATRIX:

$\mathcal{I}(\lambda) = E[J(\lambda)] \overset{!}{=} E\left[\frac{\sum y_i}{\lambda^2}\right] \overset{!}{=} \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$

FROM THIS $VAR(\hat{\lambda}_{ML}) = \frac{\lambda}{n}$ (3)

IF WE COMBINE (2) AND (3) WE GET:

$\hat{\lambda}_{ML} \dot{\sim} \mathcal{N}(\lambda, \frac{\lambda}{n})$ OBTAINING THE SAME RESULT GIVEN BY (1)

Example 2: normal example

Here we get

$$J(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{n}{\sigma^4}(\bar{y} - \mu) \\ \frac{n}{\sigma^4}(\bar{y} - \mu) & -\frac{n}{2\,\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2 \end{pmatrix}$$

and therefore

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

The implication is that $\widehat{\mu}$ and $\widehat{\sigma}^2$ are (asymptotically) uncorrelated, and the two estimated standard errors are (DUE TO THE UNCORRELATION)

$$\mathrm{SE}(\widehat{\mu}) = \frac{\widehat{\sigma}}{\sqrt{n}}, \qquad \mathrm{SE}(\widehat{\sigma}^2) = \frac{\sqrt{2}\,\widehat{\sigma}^2}{\sqrt{n}}.$$

Some numerical aspects

SOME TIMES, IN COMPLEX MODEL, YOU CAN SEE THE MAXIMISATION PROBLEM AS AN OPTIMIZATION PROBLEM AND APPLY SOME ALGORITHM

Numerical optimisation

The algorithmic nature of the MLE estimation method translates the statistical model into an optimisation problem: once a (sensible) statistical model has been specified for the data, we obtain parameter estimates with excellent properties by maximizing the log likelihood.

In some simple settings, like in the examples above, it is possible to find the analytical expression for the MLE, but in general we must resort to **numerical optimisation** of the log likelihood.

There are indeed several methods available for the task. Some knowledge of the most important issues related to it turns out particularly useful even for the application of off-the-shelf routines in `R` (or other environments).

Newton's method

Newton's method for optimisation is commonly used for minimization, in this case of the objective function $f(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta})$.

The theory is well described in the CS book, here we mention the most important aspects. The idea is to locally approximate $f(\boldsymbol{\theta})$ as a quadratic function, which is repeatedly minimised.

$\theta_{n+1} = \theta_n - \frac{U(\theta_n)}{H(\theta_n)}$

The resulting method consists in an **iterative algorithm**, which is started with $k = 0$ and a *guesstimate* $\boldsymbol{\theta}^{[0]}$, and iterates the following steps:

$k = 0 \ldots n$

1. Evaluate $\ell(\boldsymbol{\theta}^{[k]})$, $U(\boldsymbol{\theta}^{[k]})$ and $J(\boldsymbol{\theta}^{[k]})$.
2. If $U(\boldsymbol{\theta}^{[k]}) \doteq \mathbf{0}$ and $J(\boldsymbol{\theta}^{[k]})$ is positive definite then stop.
3. If $\mathbf{H} = J(\boldsymbol{\theta}^{[k]})$ is not positive definite, perturb it so that it is.
4. Solve $\mathbf{H}\,\boldsymbol{\delta} = U(\boldsymbol{\theta}^{[k]})$ for the search direction $\boldsymbol{\delta}$.
5. If $\ell(\boldsymbol{\theta}^{[k]} + \boldsymbol{\delta})$ is not $> \ell(\boldsymbol{\theta}^{[k]})$, repeatedly halve $\boldsymbol{\delta}$ until it is (*this is the step-length control*).
6. Set $\boldsymbol{\theta}^{[k+1]} = \boldsymbol{\theta}^{[k]} + \boldsymbol{\delta}$, increment k by one and return to step 1.

Fisher scoring and Quasi-Newton.

THIS MEASURES THE CURVATURE OF THE ACTUAL OBSERVED $\ell(\theta)$

IS THE NEGATIVE OF THE HESSIAN $I_{obs}(\theta) = -H(\theta)$

Whenever available, it is always a good idea to replace the observed information with the expected information $\mathcal{I}(\boldsymbol{\theta}^{[k]})$ in the Newton's method.

The resulting algorithm has a long successful tradition in statistics, it is called **Fisher scoring** and, indeed, it has better convergence properties.

→ • MORE STABLE
• MORE CONSISTENCY

Another variant avoids the computation of either $J(\boldsymbol{\theta}^{[k]})$ or $\mathcal{I}(\boldsymbol{\theta}^{[k]})$, by building an approximation to the second derivative of $\ell(\boldsymbol{\theta})$ as the optimization proceeds. This is the approach of the **Quasi-Newton** methods, such as the widely used BFGS algorithm.

$\theta_{n+1} = \theta_n + \frac{U(\theta_n)}{I(\theta_n)}$

Quasi-Newton methods are implemented in several `R` functions and packages; see the `CRAN Task View` for *Optimisation* (https://cran.r-project.org/web/views/Optimization.html).

Method	Uses observed information	Uses expected information	Computational complexity
Newton's Method	Yes	No	High (requires Hessian)
Fisher Scoring	No	Yes	Moderate
Quasi-Newton	No	Approximation only	Low (Hessian's approx.)

Stability
Less stable (Hessian can be ill-conditioned)
More stable
Stable, efficient

An example: logistic regression

We follow the MASS book for a simple example on a dose-response model.

Namely, we assume that y_i is the number of dead budworms (out of 20) for a dose of insecticide x_i^*. In particular, the statistical model is

$$Y_i \sim \mathcal{B}_i(20, \pi_i) \qquad i = 1, \ldots, 12, \text{ independent}$$

with

$$\pi_i(\alpha, \beta) = \frac{e^{\alpha + \beta\, x_i}}{1 + e^{\alpha + \beta\, x_i}}$$

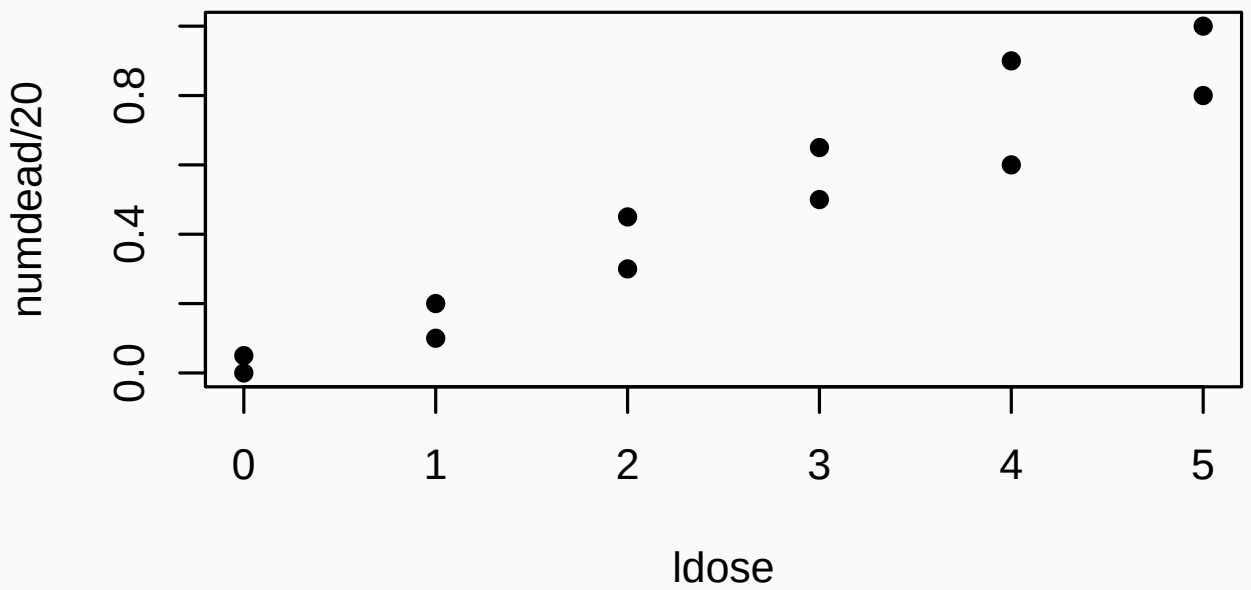
with $x_i = \log(x_i^*)$.

This is a simple instance of a *logistic regression model*.

R lab: budworm data

There are two observations at each dose (M/F budworms), but here for the sake of simplicity we ignore the different sex.

```
ldose <- rep(0:5, 2)
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
plot(ldose, numdead / 20, pch=16)
```



Logistic regression: likelihood quantities

With some simple algebra we get:

$$\ell(\alpha, \beta) = \sum_i \left\{ y_i \left(\alpha + \beta\, x_i\right) - 20\, \log(1 + e^{\alpha + \beta\, x_i}) \right\}$$

$$U(\alpha, \beta) = \begin{pmatrix} \sum_i \{y_i - 20\, \pi_i(\alpha, \beta)\} \\ \sum_i \{y_i - 20\, \pi_i(\alpha, \beta)\}\, x_i \end{pmatrix}$$

$$\mathcal{I}(\alpha, \beta) = \begin{pmatrix} \sum_i 20\, \pi_i(\alpha, \beta)\, \{1 - \pi_i(\alpha, \beta)\} & \sum_i 20\, \pi_i(\alpha, \beta)\, \{1 - \pi_i(\alpha, \beta)\}\, x_i \\ \sum_i 20\, \pi_i(\alpha, \beta)\, \{1 - \pi_i(\alpha, \beta)\}\, x_i & \sum_i 20\, \pi_i(\alpha, \beta)\, \{1 - \pi_i(\alpha, \beta)\}\, x_i^2 \end{pmatrix}$$

Notice that for this model $J(\alpha, \beta) = \mathcal{I}(\alpha, \beta)$.

R lab: likelihood and score functions

```
loglik <- function(theta, data){
           eta <- theta[1] + theta[2] * data$x
           out <- sum(data$y * eta - 20 * log(1+exp(eta)))
           return(out)
           }

score <- function(theta, data){
          prob <- plogis(theta[1] + theta[2] * data$x)
          out <- c(sum(data$y - prob * 20),
                   sum((data$y - prob * 20) * data$x))
          return(out)
        }
```

R lab: information function

```
info <- function(theta, data){
        prob <- plogis(theta[1] + theta[2] * data$x)
        info11 <- sum(20 * prob * (1-prob))
        info12 <- sum(20 * prob * (1-prob) * data$x)
        info22 <- sum(20 * prob * (1-prob) * data$x^2)
        out <- matrix(c(info11, info12, info12, info22), 2, 2)
        return(out)
        }
```

R lab: starting point

Let's start from $\alpha = \beta = 0$: we obtain

```
theta0 <- c(0, 0); budw <- data.frame(y = numdead, x = ldose)

loglik(theta0, budw)
```

```
## [1] -166.3553
```

```
score(theta0, budw)
```

```
## [1]  -9 105
```

```
info(theta0, budw)
```

```
##      [,1] [,2]
## [1,]   60  150
## [2,]  150  550
```

R lab: first step

```
H <- info(theta0, budw)
u0 <- score(theta0, budw)
delta <- solve(H, u0)
theta1 <- theta0 + delta
```

```
theta1
```

```
## [1] -1.9714286  0.7285714
```

```
loglik(theta1, budw)
```

```
## [1] -114.7219
```

which is clearly an improvement.

R lab: first 10 steps

```
## k = 1 theta= -1.971429 0.7285714 loglik= -114.7219
## k = 2 theta= -2.621436 0.9572079 loglik= -111.8192
## k = 3 theta= -2.760585 1.004947 loglik= -111.734
## k = 4 theta= -2.766079 1.006804 loglik= -111.7339
## k = 5 theta= -2.766087 1.006807 loglik= -111.7339
## k = 6 theta= -2.766087 1.006807 loglik= -111.7339
## k = 7 theta= -2.766087 1.006807 loglik= -111.7339
## k = 8 theta= -2.766087 1.006807 loglik= -111.7339
## k = 9 theta= -2.766087 1.006807 loglik= -111.7339
## k = 10 theta= -2.766087 1.006807 loglik= -111.7339
```

The algorithm converges quickly, and actually after 10 iterations

```
cat(score(theta10, budw), det(info(theta10, budw)),
    sqrt(diag(solve(info(theta10, budw)))))
```

```
## 1.776357e-15 5.329071e-15 2361.462 0.3701342 0.1235889
```

R lab: glm analysis

```
budworm.lg0 <- glm(cbind(y, 20-y) ~ x, binomial, budw)
summary(budworm.lg0, cor = FALSE)

##
## Call:
## glm(formula = cbind(y, 20 - y) ~ x, family = binomial, data =
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.7661 α̂  0.3701  -7.473 7.82e-14 ***
## x            1.0068 β̂  0.1236   8.147 3.74e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 124.876  on 11  degrees of freedom
## Residual deviance:  16.984  on 10  degrees of freedom
```