Maximum Likelihood estimation

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

jan.graffelman@upc.edu

September 19, 2022

Contents

Models

Scientists use models to describe and understand the phenomena they study.

We generally distinguish:

- Deterministic models.
- Stochastic models, also called statistical models.

Some examples:

- $V = I \times R$ (Ohm's law)
- $Y_i \sim N(\mu, \sigma^2)$ (Normal distribution)
- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (Linear regression model)
- $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_i$ (Time series model)
- ...

- We will focus on statistical models, which have a probabilistic nature
- Probability theory is the foundation of statistics.

Statistics and probability

Probability

Population:

Infinitely many balls, 75% black, 25% white

deduction

$\Downarrow$

Sample:

P("Observing 5 black balls in sample of 10") = ??

Statistics

Population:

What's the % of black balls in the population?

$\Uparrow$

induction

Sample:

We observe 5 black balls in a sample of 10

Population and sample: an opinion poll

We wish to know what percentage of all adult people in Spain favor legalization of Marijuana.

1000 Spanish adults are interviewed and their opinion is registered.

- Population: all adult people in Spain.
- Sample: 1000 Spanish adults.

Note: in statistics, populations are often assumed infinite.

Population and sample: a laboratory experiment in Physics

We wish to know the speed of light.

We measure the speed of light in a laboratory experiment, and we repeat the measurements many times, say 100 times. We could use the mean of all these measurements to estimate the speed of light.

- Population: all experiments we could possibly perform (conceptual, and infinite!)
- Sample: 100 speed measurements.

Parameters, estimators and statistics.

- Parameters are fixed, unknown quantities that specify the population.
- Any number you compute using some sample of data is called a statistic.
- Statistics are random variables whereas parameters are not (unless you are a Bayesian).
- Estimators are statistics that are used to estimate the unknown parameters.

Notation:

- $\hat{\mu} = \overline{x}$ with μ the true population mean.
- $\hat{\sigma} = s$ with σ the true population standard deviation.

Point estimate and interval estimate

Statistical models have unknown parameters that need to be estimated.

- Estimating a population parameter with a single value computed from a sample is called point estimation.
- Estimating a population parameter with a range of plausible values computed from a sample is called interval estimation.

Methods for obtaining point estimators

- Maximum likelihood (ML) method.
- Method of moments (MM).
- Bayesian methods.

Maximum likelihood estimators

- Let $X_1, \ldots, X_n$ be a random sample from a distribution $f(x|\theta_1, \ldots, \theta_k)$.
- The likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ is defined as

$$L(\boldsymbol{\theta}|\mathbf{x}) = L(\theta_1, \ldots, \theta_k | x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k)$$

- This is in fact, the joint density function, considering the data as given.
- We will often work with the log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{x})$, defined correspondingly as

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \ln\left(L(\boldsymbol{\theta}|\mathbf{x})\right) = \ln\left(L(\theta_1, \ldots, \theta_k | x_1, \ldots, x_n)\right) = \sum_{i=1}^{n} \ln\left(f(x_i|\theta_1, \ldots, \theta_k)\right)$$

Example: Bernoulli distribution

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim Bern(p)$

$$P\left(X_1 = x_1 \mid p\right) = p^{x_1} \left(1 - p\right)^{1 - x_1}$$

$$\begin{aligned} P\left(X_1 = x_1, \ldots, X_n = x_n \mid p\right) &= \prod_{i=1}^{n} p^{x_i} \left(1 - p\right)^{1 - x_i} \\ &= p^{\sum_{i=1}^{n} x_i} \left(1 - p\right)^{n - \sum_{i=1}^{n} x_i} \end{aligned}$$

$$L(p \mid x_1, \ldots, x_n) = p^{\sum_{i=1}^{n} x_i} \left(1 - p\right)^{n - \sum_{i=1}^{n} x_i}$$

Example: exponential distribution

Let $X_1, \ldots, X_n$ be a random sample with $X \sim exp(\, \mu \,)$.

$$f(x_i \mid \mu) = \frac{1}{\mu} \exp\left(\frac{-x_i}{\mu} \right)$$

$$f(x_1, \ldots, x_n \mid \mu) = \frac{1}{\mu^n} \exp\left(-\frac{\sum_{i=1}^n x_i}{\mu} \right)$$

The likelihood function is:

$$L(\mu \mid x_1, \ldots, x_n) = \frac{1}{\mu^n} \exp\left(\frac{-\sum_{i=1}^n x_i}{\mu} \right)$$

Example: the normal distribution

Let $X_1, \ldots, X_n$ be a random sample with $X \sim N(\mu, \sigma^2)$.
The joint density function is

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\,\pi\,\sigma^2}} \right)^n \exp\left\{ -\frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

and the likelihood function is

$$L(\mu, \sigma^2 \mid x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\,\pi\,\sigma^2}} \right)^n \exp\left\{ -\frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

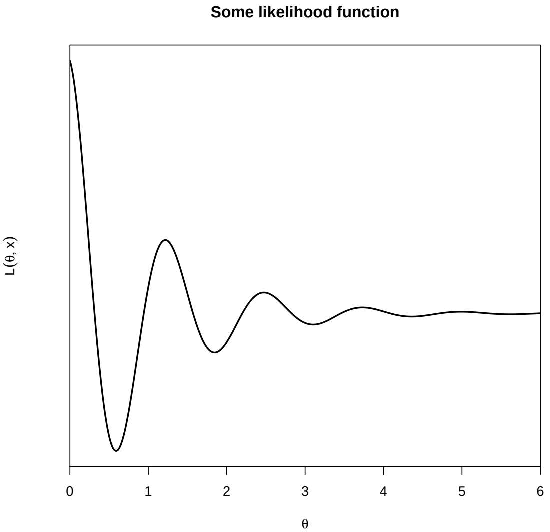
Maximum likelihood estimator (MLE)

- The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta}|x)$ as a function of $\boldsymbol{\theta}$.
- The method selects a value for $\boldsymbol{\theta}$ such that the sample is most likely.
- Obtaining a maximum likelihood estimator is an optimization problem.
- In practice, it is often easier (and equivalent) to maximize the natural logarithm of the likelihood function, thus maximize $\ell(\boldsymbol{\theta}|x)$.
- In general, MLE's have good properties.

Some problems in maximum likelihood estimation

- How can we find a global maximum, or verify that a global maximum has been found?
- Numerical sensitivity: how sensitive is the estimate when the data is slightly perturbed?

Candidates for MLE



Candidates for MLE

- Interior points where

$$\frac{\partial}{\partial \theta} L(\theta|\mathbf{x}) = 0, \qquad i = 1, \ldots, k. \text{ and } \quad \frac{\partial^2}{\partial \theta^2} L(\theta|\mathbf{x})|_{\theta=\hat{\theta}} < 0$$

- Boundary points

Example: exponential distribution

Let X be a random variable with $X \sim exp(\lambda)$.

$$f(x \mid \lambda) = \lambda\, e^{-\lambda\, x}$$

We observe $x = 3$ (sample size $n = 1$)

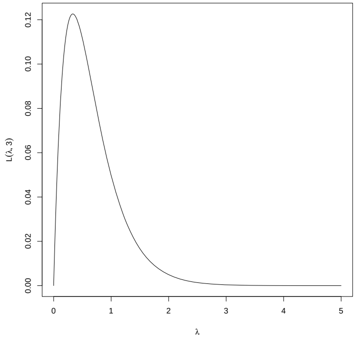
$$L(\lambda \mid x = 3) = \lambda\, e^{-3\,\lambda}$$

$$L'(\lambda \mid x = 3) = e^{-3\,\lambda}\,(1 - 3\,\lambda)$$

$$\hat{\lambda} = \frac{1}{3} \qquad L''(\lambda = 1/3 \mid x = 3) < 0$$

$$\lim_{\lambda \to 0} \lambda\, e^{-\lambda\, x} = 0$$

$$\lim_{\lambda \to \infty} \lambda\, e^{-\lambda\, x} = 0$$



Example: Bernoulli's distribution

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim Bern(p)$, and $\Theta = [0, 1]$.

$$L(p \,|\, x) = p^{\sum_{i=1}^n x_i} \, (1 - p)^{n - \sum_{i=1}^n x_i}$$

$$log \; L(p \,|\, x) = \left(\sum_{i=1}^n x_i \right) log \; p + \left(n - \sum_{i=1}^n x_i \right) log(1 - p)$$

$$\frac{d}{d\,p} \, log \; L(p \,|\, x) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1 - p} = 0 \Leftrightarrow \widehat{p} = \frac{\sum_{i=1}^n x_i}{n}$$

$\frac{\sum_{i=1}^n x_i}{n}$ is the only stationary point in $\Theta = [0, 1]$.

$$\left. \frac{d^2}{d\,p^2} \, log \; L(p \,|\, x) \right|_{p = \widehat{p}} = \left. - \frac{\sum_{i=1}^n x_i}{p^2} + \frac{\sum_{i=1}^n x_i - n}{(1 - p)^2} \right|_{p = \widehat{p}} =$$

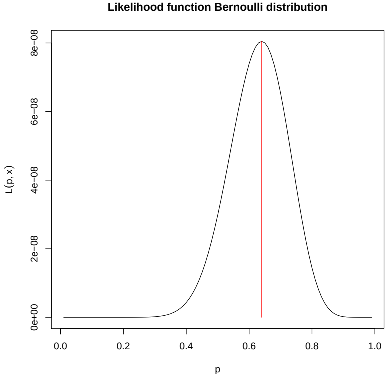
$$- \frac{n \, \widehat{p}}{\widehat{p}^2} - \frac{n \, (1 - \widehat{p})}{(1 - \widehat{p})^2} = - \left(\frac{n}{\widehat{p}} + \frac{n}{1 - \widehat{p}} \right) < 0$$

Boundary points: $L(0 \,|\, x) = 0$ and $L(1 \,|\, x) = 0$

Example: Bernoulli distribution

Exercise (in R)

- Simulate 100 flips of a fair coin ($P(\text{"Heads"}) = P(\text{"Tail"}) = 0.50$)
- Calculate the value of the ML estimator, $\hat{p}_{ML}$
- Write a function that calculates the ML estimator as a function of p
- Make a plot of the likelihood function
- Verify graphically that $\hat{p}$ maximizes the likelihood function



Example: normal distribution

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim N(\mu, 1)$.

$$L(\mu \,|\, \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \mu)^2} = \frac{1}{(2\,\pi)^{n/2}}\, e^{-\frac{1}{2}\,\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ell(\mu, \mathbf{x}) = -\frac{n}{2} \ln{(2\pi)} - \frac{1}{2} \sum (x_i - \mu)^2$$

$$\frac{d}{d\mu}\,\ell(\mu \,|\, x) = 0 \rightarrow \sum_{i=1}^{n} (x_i - \mu) = 0 \rightarrow \widehat{\mu} = \overline{x}$$

$$\frac{d^2}{d\,\mu^2}\,\ell(\mu \,|\, x)|_{\mu=\overline{x}} < 0$$

$$\lim_{\mu \to +\infty} L(\mu \,|\, x) = \lim_{\mu \to -\infty} L(\mu \,|\, x) = 0$$

Additional examples

- Find the MLE for parameter λ of the Poisson distribution.
- Find the MLE for parameter p of the Geometric distribution.

Mutation rate in DNA

Sequence	Mutations
GACACGTATAAGGCATAACATACACTGCGGTTCGTTCCGATTATGAATCC...	—
GACACGTATAAGGCATAACATACACTGCGGTTCGTTCCGATTATGAATCC...	0
GACACGTATAAGGCATAACATACACTGCGGTTCGTTCCGATTATGAATCC...	0
GACACGTATAAGGCATAACATACACTGCGGTTCGTTCCGATTATGAATCC...	0
GACACGTAGAAGGCATAACATACACTGCGGTTCGTTCCGATTATGAATCC...	1
GACACGTATAAGGCATAACATACACTGCGGTTCGTTCCGATTATGAATCC...	0
GACACGTATAAGGCATAACATACACTGCGGTTCGTTCCGACTATGAATCC...	1
GACACGTATATGGCATAACATACACTGCGGTTCGTTCCGACTATGAATCC...	2
GACACGTATAAGGCATAACATACACTGCGGTTCGTTCCGATTATGAATCC...	0
⋮	

$$X = \text{Number of mutations} \sim \text{Pois}(\lambda)$$

Invariance property of the MLE

Let $\widehat{\theta}$ be the MLE of θ. Then for any function $\tau(\theta)$ the MLE of $\tau(\theta)$ is $\tau(\widehat{\theta})$.
Example:

- Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim Bern(p)$.
- We wish to estimate $\ln\left(\frac{p}{1-p}\right)$ (the log odds).
- We know $\widehat{p} = \frac{\sum_{i=1}^n x_i}{n}$
- The MLE of $\ln\left(\frac{p}{1-p}\right)$ is $\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right)$

Two parameters: the normal distribution

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim N(\theta, \sigma^2)$.

$$L(\theta, \sigma^2 \,|\, \boldsymbol{x}) = \frac{1}{(2\,\pi\,\sigma^2)^{n/2}}\, e^{-\frac{1}{2\,\sigma^2}\,\sum_{i=1}^{n}(x_i-\theta)^2}$$

$$\log L(\theta, \sigma^2 \,|\, \boldsymbol{x}) = -\frac{n}{2}\log 2\,\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2$$

$$\frac{\partial}{\partial\theta}\log L(\theta, \sigma^2 \,|\, \boldsymbol{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) = 0$$

$$\frac{\partial}{\partial(\sigma^2)}\log L(\theta, \sigma^2 \,|\, \boldsymbol{x}) = -\frac{n}{2\,\sigma^2} + \frac{1}{2\,\sigma^4}\sum_{i=1}^{n}(x_i-\theta)^2 = 0$$

$$\widehat{\theta} = \overline{x} \quad \text{i} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Two (and more) parameters: some general comments

- With two (or more) parameters, ML estimation amounts to the maximization of a function that depends on two (or more) variables (the parameters in this case).
- Such maximization in multiple variables is mathematically more involved, and explained in detail in the MDO course.
- In many advanced ML estimation problems, explicit (closed form) solutions do often not exist or are hard to find. In those cases, we maximize the likelihood iteratively, with numerical methods.
- The numerical methods typically require some initial estimate or first guess of the maximum.
- A sensible initial estimate can often be obtained by alternative estimation methods.

Two parameters: gamma distribution

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim \Gamma(\alpha, \lambda)$.

$$f(x \mid \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha - 1} e^{-\lambda x} \quad \text{for } 0 \leq x < \infty$$

$$L(\theta) = L(\alpha, \lambda) = \left(\frac{1}{\Gamma^n(\alpha)} \right) \lambda^{n \alpha} \left(\prod_{i=1}^{n} x_i \right)^{\alpha - 1} e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ell(\theta) = \log L(\theta) = -n \log \Gamma(\alpha) + n \alpha \log \lambda + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial \ell}{\partial \alpha} = -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + n \log \lambda + \sum_{i=1}^{n} \log x_i$$

$$\frac{\partial \ell}{\partial \lambda} = n \alpha \frac{1}{\lambda} - \sum_{i=1}^{n} x_i$$

$$n \alpha \frac{1}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Leftrightarrow \widehat{\lambda} = \frac{n \widehat{\alpha}}{\sum_{i=1}^{n} x_i} = \frac{\widehat{\alpha}}{\overline{x}_n}$$

$$-n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + n \log \frac{\hat{\alpha}}{\overline{x}_n} + \sum_{i=1}^{n} \log x_i = 0$$

There is no explicit solution.

The Newton-Raphson method

- We compute $\hat{\alpha}$ iteratively (Newton-Raphson method)
- Roots of the function $f(\alpha) = 0$ can be found by:

$$\hat{\alpha}_{n+1} = \hat{\alpha}_n + h_n \qquad h_n = -\frac{f(\hat{\alpha}_n)}{f'(\hat{\alpha}_n)}$$

- For our problem:

$$f(\hat{\alpha}) = -n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + n \log \frac{\hat{\alpha}}{\overline{x}} + \sum_{i=1}^{n} \log x_i = -n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + n \log \hat{\alpha} - n \log \overline{x} + \sum_{i=1}^{n} \log x_i$$

and

$$f'(\hat{\alpha}) = -n\frac{d}{d\hat{\alpha}}\left(\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}\right) + \frac{n}{\hat{\alpha}},$$

- The fraction $\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}$ is known as the **digamma** function.
- Its derivative $\frac{d}{d\hat{\alpha}}\left(\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}\right)$ is known as the **trigamma** function.
- An initial value α_0 is needed. We could use $\alpha_0 = 1$ or take the value of the estimator obtained by the method of moments.

Two parameters: gamma distribution

Consider a sample of 10.000 observations of a $\Gamma(\alpha = 2, \lambda = 3)$ distribution. The mean of the sample is 0.659269. We find the value of the MLE iteratively

i	α	$f(\alpha)$	$f'(\alpha)$	h
0	1.000000	3.102326e+03	-6449.341	4.810300e-01
1	1.481030	1.071487e+03	-2755.763	3.888170e-01
2	1.869847	2.364494e+02	-1672.638	1.413632e-01
3	2.011210	1.764700e+01	-1432.216	1.232146e-02
4	2.023532	1.141968e-01	-1413.740	8.077639e-05
5	2.023612	4.844877e-06	-1413.620	3.427285e-09
6	2.023612	-1.818989e-12	-1413.620	-1.286760e-15

$\hat{\alpha} = 2.023612$. Using $\hat{\lambda} = \frac{\hat{\alpha}}{\overline{x}}$ we find $\hat{\lambda} = \frac{2.023612}{0.659269} = 3.069479$.
By the method of moments, we find:

$$\hat{\alpha}_{MM} = \frac{\overline{x}^2}{\frac{1}{n}(\sum_{i=1}^{n} x_i^2) - \overline{x}^2} = 2.045233$$

this is a better initial point, from which we converge faster to the maximum.

i	α	$f(\alpha)$	$f'(\alpha)$	h
0	2.045233	-3.022058e+01	-1382.049	-2.186650e-02
1	2.023367	3.471806e-01	-1413.984	2.455335e-04
2	2.023612	4.477239e-05	-1413.620	3.167216e-08
3	2.023612	-1.818989e-12	-1413.620	-1.286760e-15

Some special cases

- Sometimes the support of the density depends on the parameter of interest.
- It then makes sense to use indicator variables that account for this.
- Examples: uniform distribution, distributions with a translation parameter, ...

Example: uniform distribution

$$X \sim U[0, \theta]$$

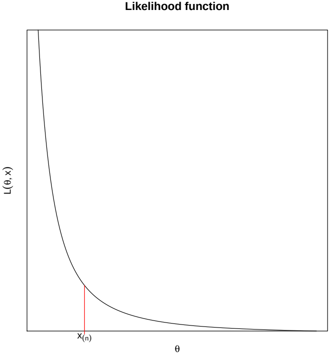
$$f(x, \theta) = \frac{1}{\theta} \cdot I_{0 \leq x \leq \theta}$$

$$L(\theta | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} \prod_{i=1}^{n} I_{x_i \leq \theta}$$

$$= \frac{1}{\theta^n} I_{x_{(n)} \leq \theta}$$

$$= \frac{1}{\theta^n} I_{\theta \geq x_{(n)}}$$

$$\hat{\theta}_{ML} = X_{(n)}$$



Exercise:

$$X \sim U(\alpha, 1) \qquad f(x) = \frac{1}{1 - \alpha} \qquad 0 < \alpha < x < 1$$

Find the ML estimator for α

Precision of the ML estimator

- A point estimate obtained by ML is, by itself, not very informative.
- We need to specify its precision.
- The precision depends on the variance or the Fisher information of the ML estimator.

Fisher information of a sample

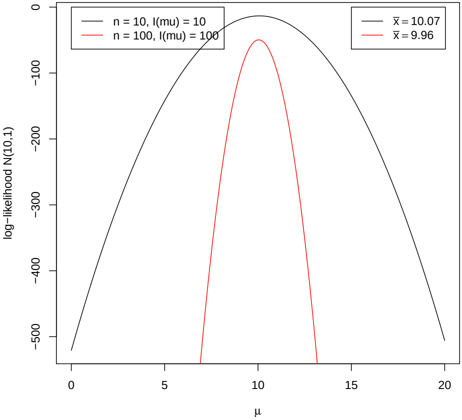
Let $X_1, \ldots, X_n$ be a random sample with

$$f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

The Fisher information about θ contained in $\boldsymbol{x}$ is defined by

$$I_{\boldsymbol{x}}(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta} \ln\left(f(\boldsymbol{x} \mid \theta)\right)\right)^2\right]$$

Interpretation of Fisher Information



Fisher information relates to the curvature of the likelihood function

Cramér-Rao lower bound

- For any unbiased estimator ($E\left(\hat{\theta}\right) = \theta$), there exists a lower bound on its variance.
- This bound equals the reciprocal of the Fisher information.

$$V\left(\hat{\theta}\right) \geq \frac{1}{I_{\mathbf{x}}(\theta)}$$

- An unbiased estimator that attains the Cramér-Rao lower bound is called efficient.

Asymptotic distribution of the ML estimator

Let $X_1, \ldots, X_n$ be i.i.d. with density $f(x|\theta)$, and let $\hat{\theta}$ be the MLE of θ. Under regularity conditions we have

$$\hat{\theta}_n \rightarrow N\left(\theta, \frac{1}{I_{\mathbf{x}}(\theta)}\right)$$

where $1/I_{\mathbf{x}}(\theta)$ is the Cramér-Rao lower bound.

Thus, MLE are asymptotically (for large samples)

- unbiased,
- efficient,
- and normally distributed.

Interval estimation with maximum likelihood estimators

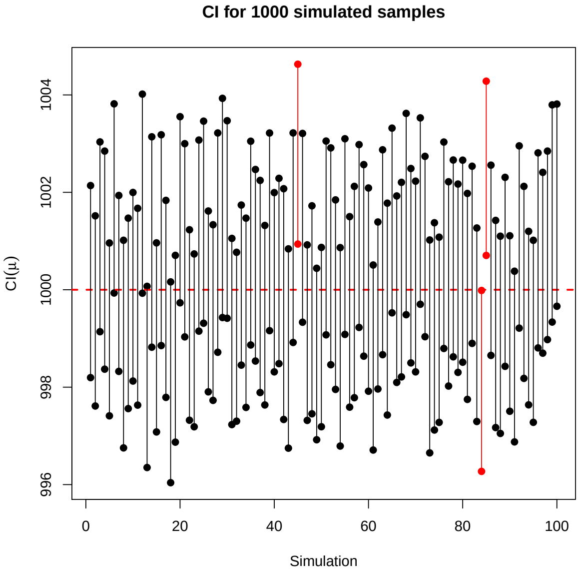
- Having the variance and the distribution of the ML estimator, we can now say something about uncertainty.
- A confidence interval is an expression of the uncertainty of the estimate.
- A classical result, with $X_i \sim N(\mu, \sigma^2)$, is

$$CI(\mu)_{1-\alpha} = \overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (1)$$

  where $\overline{X} = \hat{\mu}_{ML}$, and $\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{V(\hat{\mu})}$.
- Term $\frac{\sigma}{\sqrt{n}}$ (σ estimated by s) is called the standard error of the mean.
- Term $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \approx 2 \frac{\sigma}{\sqrt{n}}$ when $\alpha = 0.05$ is the error margin.
- Equation (1) holds in general for ML estimators:

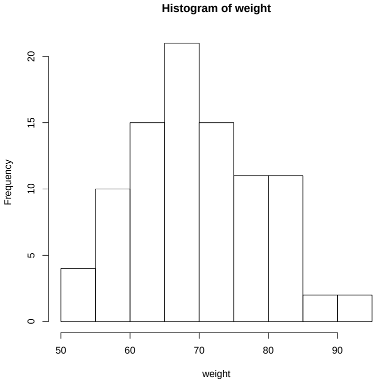
$$CI(\theta)_{1-\alpha} = \hat{\theta} \pm z_{\alpha/2} \sqrt{V\left(\hat{\theta}\right)} \quad (2)$$

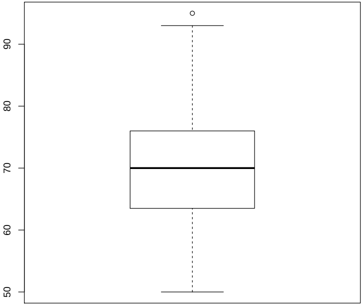
Frequentist interpretation of a confidence interval



A practical example of ML estimation

In a study on physical characteristics of students, data on the weight (in kg) of $n = 91$ students is collected.

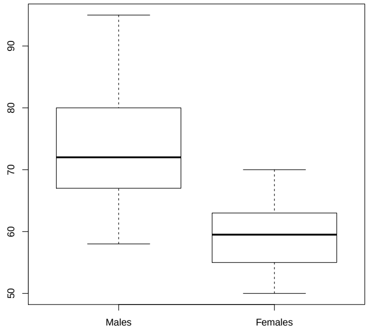




$$\hat{\mu}_{ML} = \overline{X} = 70.58 \qquad CI(\mu)_{0.95} = (68.65, 72.52)$$

Any problem?

Stratifying by gender



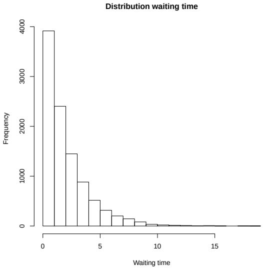
$\hat{\mu}_{males} = \overline{X} = 72.62 \qquad CI(\mu)_{0.95} = (70.73, 74.52)$

$\hat{\mu}_{females} = \overline{X} = 59.36 \qquad CI(\mu)_{0.95} = (56.23, 62.48)$

R instructions

```
X <- read.table(
"http://www-eio.upc.es/~jan/data/StudentWeight.txt",
header=TRUE)

weight <- X[,2]
sex <- X[,4]

hist(weight,breaks=12)
boxplot(weight)

boxplot(weight~sex,names=c("Males","Females"))

mean(weight[sex==0])
mean(weight[sex==1])

t.test(weight[sex==0]) % for obtaining the confidence
t.test(weight[sex==1]) % intevals
```

Example: ML estimation of the rate of an exponential distribution



- What is the rate of decay?
- What is the precision of a rate estimate?

```
> fitdistr(x,"exponential")
      rate
  0.498116487
 (0.004981165)
```

Density and likelihood:

$$f(x|\lambda) = \lambda e^{-\lambda x} \quad L(\lambda|\mathbf{x}) = \lambda^n e^{-\lambda \sum x_i}$$

With some algebra, it follows that

$$\hat{\lambda} = 1/\bar{x}, \qquad I_n(\lambda) = n/\lambda^2 \qquad V(\hat{\lambda}) = \lambda^2/n$$

$$CI_{1-\alpha}(\lambda) = \hat{\lambda} \pm z_{\alpha/2}\sqrt{V\left(\hat{\lambda}\right)} = \hat{\lambda} \pm z_{\alpha/2}\frac{\hat{\lambda}}{\sqrt{n}}$$

Descriptive statistics of a sample of $n = 10.000$ waiting times

	N	N*	Mean	Stdev	Med	Q1	Q3	Min	Max
X	10000	0	2.0075	2	1.397	0.579	2.768	0.001	18.163

$$\hat{\lambda} = 1/2.0075 = 0.49812$$

$$CI_{0.95}(\lambda) = 0.49812 \pm 1.96\frac{0.49812}{\sqrt{10000}} = (0.4884; 0.5079)$$

Other methods

We give a brief account of

- The method of moment
- Bayesian methods

Moments

- The k^{th} moment of a r.v. X is given by

$$\mu_k = E\left(X^k\right)$$

- The k^{th} central moment of a r.v. is given by

$$\mu_k = E\left(X - \mu_1\right)^k$$

- E.g. the variance of X is the second central moment

$$V\left(X\right) = E\left(X - E\left(X\right)\right)^2$$

Method of moments

Sample	Population
$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu_1 = E\left(X\right)$
$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$	$\mu_2 = E\left(X^2\right)$
$m_3 = \frac{1}{n}\sum_{i=1}^{n} X_i^3$	$\mu_3 = E\left(X^3\right)$
$\vdots$	$\vdots$
$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$	$\mu_k = E\left(X^k\right)$

- Equate sample moments to population moments.
- Use as many moments as the number of parameters you need to estimate.
- Write the parameters as a function of the sample moments.

Method of moments

Let $X \sim U(0, \theta)$. We take a simple random sample of size n.

$$m_1 = \overline{X} \qquad E(X) = \frac{\theta}{2}$$

$$\frac{\hat{\theta}_{MM}}{2} = \overline{X} \rightarrow \hat{\theta}_{MM} = 2\overline{X}$$

Exercise:

Let $X \sim Exp(\lambda)$. We take a simple random sample of size n.

$$f(x, \lambda) = \lambda e^{-\lambda x}$$

Find an estimator $\hat{\lambda}_{MM}$ for λ by using the method of moments.

Method of moments (Normal distribution)

Let $X_1, X_2, \ldots X_n$ be random sample of size n from a $N(\mu, \sigma^2)$ distribution.

$$\mu_1 = E\left(X\right) = \mu \qquad \mu_2 = E\left(X^2\right) = V\left(X\right) + E\left(X\right)^2 = \sigma^2 + \mu^2$$

$$m_1 = \overline{X} \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\hat{\mu}_{MM} = \overline{X}$$

$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

The Bayesian approach

- In classical, frequentist statistics, θ is assumed to be an unknown, fixed quantity.
- In the Bayesian approach, θ is a random variable, and its variation is described by a distribution, the prior distribution, $\pi(\theta)$.
- The prior distribution, $\pi(\theta)$, is subjective, and chosen by the investigator.
- A sample $X_1, X_2, \ldots X_n$ is observed, and in the light of this data the distribution of θ is updated.
- The newly obtained distribution is called the posterior distribution, $\pi(\theta|\mathbf{x})$.
- The posterior distribution is

$$\pi(\theta|\mathbf{x}) = \frac{f(\theta, \mathbf{x})}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}$$

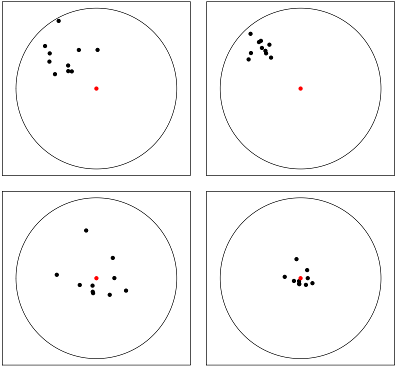
with

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$$

- A point estimate for θ is obtained by calculating the expectation (or the median) of the posterior distribution.
- The posterior distribution is proportional to the likelihood function and the prior

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

Some criteria for comparing estimators



- Bias $= E\left(\hat{\theta}\right) - \theta$
- Variance $V\left(\hat{\theta}\right)$ (or Precision $= \frac{1}{V(\hat{\theta})}$)
- Mean squared error $MSE(\hat{\theta}) \equiv E\left((\hat{\theta}-\theta)^2\right) = V\left(\hat{\theta}\right) + (Bias(\hat{\theta}))^2$

References on ML estimation

- Casella, R. & Berger, R. L. (2002) Statistical Inference. Duxbury, Pacific Grove, CA, USA. Second edition, Chapter 7.
- DeGroot, M. H. & Schervish, M.J. (2002) Probability and Statistics. Addison-Wesley. Third edition. Chapter 6.
- Ewens, W. J. & Grant, G. R. (2005) Statistical Methods in Bioinformatics. An Introduction. Springer. Second Edition.