Statistical foundations of machine learning

INFO-F-422

Gianluca Bontempi

Machine Learning Group
Computer Science Department
mlg.ulb.ac.be

Approaches to parametric estimation

There are two main approaches to parametric 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.

Parametric estimation

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

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

$$F_{\mathbf{z}}(z) = F_{\mathbf{z}}(z, \theta)$$

   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 θ so that the parametrized distribution $F_{\mathbf{z}}(z, \hat{\theta})$ 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 $F_{\mathbf{z}}(z, \theta)$.

I.I.D. samples

- Consider a set of N i.i.d. random variables $\mathbf{z}_i$.
- *I.I.D.* means Identically and Independently Distributed.
- Identically distributed means that all the observations have been sampled from the same distribution, that is

$$\text{Prob}\left\{\mathbf{z}_i = z\right\} = \text{Prob}\left\{\mathbf{z}_j = z\right\} \qquad \text{for all } i, j = 1, \dots, N \text{ and } z \in \mathcal{Z}$$

- Independently distributed means that the fact that we have observed a certain value $\mathbf{z}_i$ does not influence the probability of observing the value $\mathbf{z}_j$, that is

$$\text{Prob}\left\{\mathbf{z}_j = z | \mathbf{z}_i = z_i\right\} = \text{Prob}\left\{\mathbf{z}_j = z\right\}$$

Some estimation problems

1. Let $D_N = \{20, 31, 14, 11, 19, \dots\}$ be the times in minutes spent the last 2 weeks to go home. How much does it take in average to reach my house from ULB?
2. 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 mean inter-arrival time?
3. Consider the students of the last year of Computer Science. What is the variance of their grades?

Parametric estimation (II)

Parametric estimation is a **mapping** from the space of the sample data to the space of parameters Θ. Two are the possible outcomes

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

Point estimation

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

$$\theta = t(F)$$

  This corresponds to the fact that θ is a characteristic of the population described by $F_{\mathbf{z}}(\cdot)$.
- 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.

Methods of constructing estimators

We will discuss:

- Plug-in principle
- Maximum likelihood

Empirical distribution function

Suppose we have observed a i.i.d. random sample of size N from a **distribution function** $F_{\mathbf{z}}(z)$ of a continuous rv $\mathbf{z}$

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

where

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

Let $N(z)$ be the number of samples in D_N that do not exceed z. The empirical distribution function is

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

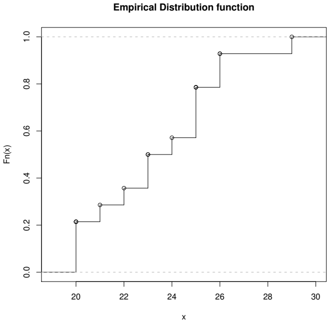
This function is a staircase function with discontinuities at the points z_i.

TP R: 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 $\hat{F}_{\mathbf{z}}$ (`cumdis.R`)



Plug-in principle to define an estimator

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

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

  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

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

Sample average

- Consider a r.v. $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$ such that

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

  with θ unknown.

- Suppose we have available the sample $F_{\mathbf{z}} \rightarrow D_N$, made of N observations.
- The *plug-in* point estimate of θ 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 r.v. $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$ where the mean μ and the variance σ^2 are unknown.
- Suppose we have available the sample $F_{\mathbf{z}} \to D_N$.
- Once we have the sample average $\hat{\mu}$, the *plug-in* estimate of σ^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 for all z_i

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

Other plug-in estimators

- Skewness estimator:

$$\hat{\gamma} = \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{\mu})^3}{\hat{\sigma}^3}$$

- Upper critical point estimator:

$$\hat{z}_\alpha = \sup\{z : \hat{F}(z) \leq 1 - \alpha\}$$

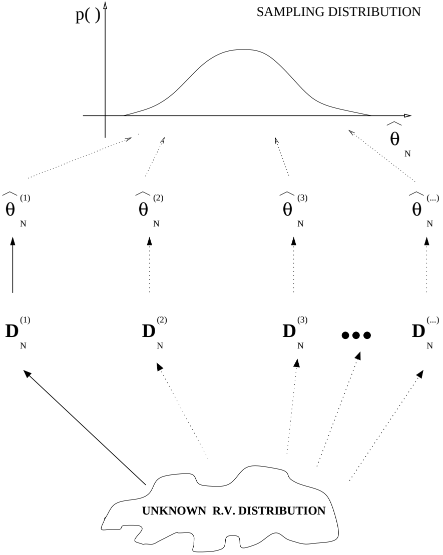
- Sample correlation:

$$\hat{\rho}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \hat{\mu}_x)^2}\sqrt{\sum_{i=1}^{N}(y_i - \hat{\mu}_y)^2}}$$

Sampling distribution

- Given a dataset D_N of N samples, 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
$$\hat{\boldsymbol{\theta}} = g(\mathbf{D}_N)$$
which is called the *point estimator* of θ.
- The probability distribution of the r.v. $\hat{\boldsymbol{\theta}}$ is called the sampling distribution.
- In practical situations only one dataset is observed but the theoretical notion of sampling distribution is crucial to understand the estimation process.

Sampling or finite-sample distribution



See the R scripts `sam_dis.R` and `est_step.R`.

Monte Carlo illustration of a sampling distribution

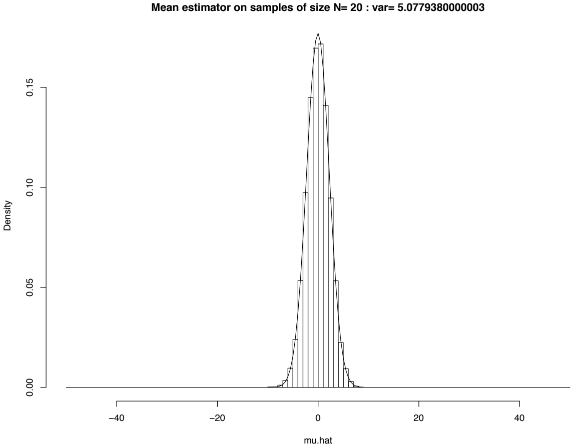
In order to illustrate the theoretical notion of sampling distribution, we need a random number generator from a distribution $F(z, \theta)$.

```
$S = \{\}$
for $r = 1$ to $R$ do
    $F_{\mathbf{z}} \rightarrow D_N = \{z_1, z_2, \dots, z_N\}$  // sample dataset
    $\hat{\theta} = g(D_N)$  // compute estimate
    $S = S \cup \{\hat{\theta}\}$
end for
Plot histogram of $S$
Compute statistics of $S$ (mean, variance)
Study distribution of $S$ with respect to $\theta$
```

R script

```
mu<-0  # parameter
R<-10000  # number trials
N<-20   # size dataset
mu.hat<-numeric(R)
for (r in 1:R){
     D<-rnorm(N,mean=mu,sd=10)    #  random generator
     mu.hat[r]<-mean(D)      #  estimator
}
hist(mu.hat)   #  histogram
```

Histogram



Suppose $\theta = 0$. What could you say about this estimator? And if $\theta = 1$

Bias and variance

How accurate is $\hat{\theta}$? This lead us to the definition of bias, variance and standard error of an estimator.

Definition

An estimator $\hat{\boldsymbol{\theta}}$ of θ is said to be *unbiased* if and only if

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

Otherwise, it is called *biased* with bias

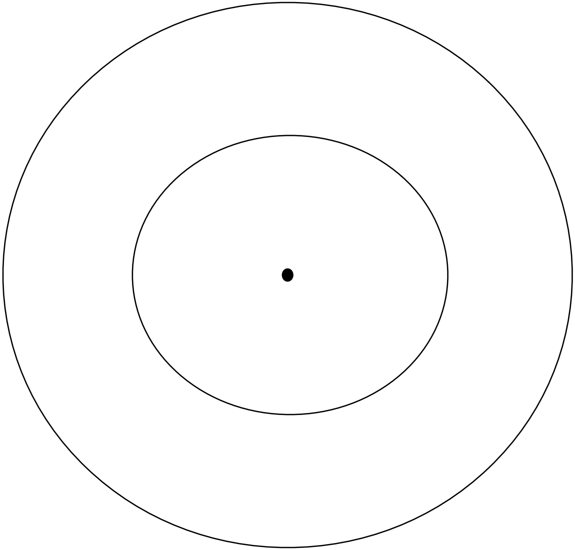
$$\text{Bias}[\hat{\boldsymbol{\theta}}] = E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \theta$$

Definition

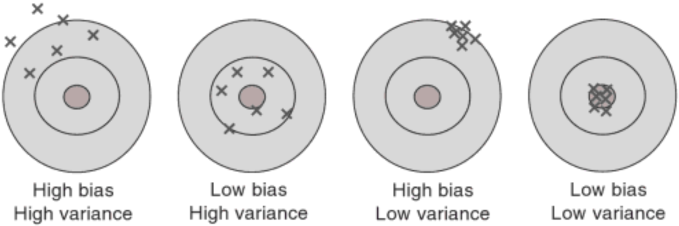
The variance of an estimator $\hat{\boldsymbol{\theta}}$ of θ is the variance of its sampling distribution

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

Estimation and darts



Bias/variance issue



Some consideration

- An unbiased estimator is an estimator that takes on average the right value.
- Many unbiased estimators may exist for a parameter θ.
- If $\hat{\theta}$ is an unbiased estimator of θ, it may happen that $f(\hat{\theta})$ be a BIASED estimator of $f(\theta)$.
- A biased estimator with a known bias (not depending on θ) 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 $\hat{\boldsymbol{\mu}}$ and the sample variance $\hat{\boldsymbol{\sigma}}^2$ are unbiased estimators of the mean $E[\mathbf{z}]$ and the variance $\text{Var}[\mathbf{z}]$, respectively.
- In general $\hat{\boldsymbol{\sigma}}$ is not an unbiased estimator of σ even if $\hat{\boldsymbol{\sigma}}^2$ is an unbiased estimator of σ^2.

Useful relationships

- $$E[a\mathbf{x} + b\mathbf{y}] = aE[\mathbf{x}] + bE[\mathbf{y}]$$

- $$\begin{aligned}\mathrm{Var}\,[a\mathbf{x} + b\mathbf{y}] = a^2\mathrm{Var}\,[\mathbf{x}]+b^2\mathrm{Var}\,[\mathbf{y}]+2ab\,(E[\mathbf{xy}] - E[\mathbf{x}]E[\mathbf{y}]) =\\ = a^2\mathrm{Var}\,[\mathbf{x}] + b^2\mathrm{Var}\,[\mathbf{y}] + 2ab\mathrm{Cov}[\mathbf{x}, \mathbf{y}]\end{aligned}$$

  where

  $$\mathrm{Cov}[\mathbf{x}, \mathbf{y}] = E\left[(\mathbf{x} - E[\mathbf{x}])(\mathbf{y} - E[\mathbf{y}])\right] = E[\mathbf{xy}] - E[\mathbf{x}]E[\mathbf{y}]$$

  is the **covariance**.

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

- Consider a random variable $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$.
- Let μ and σ^2 the mean and the variance of $F_{\mathbf{z}}(\cdot)$, respectively.
- Suppose we have observed the i.i.d. sample $D_N \leftarrow F_{\mathbf{z}}$.
- The following relation holds

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

- This means that the *sample average estimator* is not biased ! This holds for whatever distribution $F_{\mathbf{z}}(\cdot)$.
- Since $\mathrm{Cov}[\mathbf{z}_i, \mathbf{z}_j] = 0$, for $i \neq j$, the variance of the *sample average estimator* is

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

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

What is the bias of the estimator of the variance?
Given an i.i.d. $D_N \leftarrow \mathbf{z}$ it can be shown that

$$E_{\mathbf{D}_N}[\hat{\boldsymbol{\sigma}}^2] = \sigma^2$$

- Sample variance (with $N-1$ at denominator) is not biased !
- Question? Is $\hat{\boldsymbol{\mu}}^2$ an unbiased estimator of μ^2? Try to answer first in analytical terms, then use a Monte Carlo simulation to validate the answer.

Considerations

- The results so far are **independent** of the form $F(\cdot)$ of the distribution.
- The variance of $\hat{\boldsymbol{\mu}}$ is $1/N$ times the variance of $\mathbf{z}$. This is a reason for collecting several samples: the larger N, the smaller is $\mathrm{Var}\left[\hat{\boldsymbol{\mu}}\right]$, so bigger N means a better estimate of μ.
- According to the central limit theorem, under quite general conditions on the distribution $F_{\mathbf{z}}$, the distribution of $\hat{\boldsymbol{\mu}}$ will be approximately normal as N gets large, which we can write as

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

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

Exercise

Let $\mathbf{z}$ such that $E[\mathbf{z}] = \mu$ and $\text{Var}[\mathbf{z}] = \sigma^2$. Suppose we want to estimate from i.i.d. dataset D_N the parameter $\theta = \mu^2$.
Let us consider three estimators:

1. $$\hat{\theta}_1 = \left(\frac{\sum_{i=1}^{N} z_i}{N} \right)^2$$

2. $$\hat{\theta}_2 = \frac{\sum_{i=1}^{N} z_i^2}{N}$$

3. $$\hat{\theta}_3 = \frac{(\sum_{i=1}^{N} z_i)^2}{N}$$

Are they unbiased? Compute analytically the bias and verify the result by Monte Carlo simulation for different values of N.

Hint: $\sigma^2 = E[\mathbf{z}^2] - \mu^2$
Solution in the file `gbcode/exercises/Exercise1.pdf` in the R package.

Bias/variance decomposition of MSE

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

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

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

  $$\text{MSE} = (E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \theta)^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right] = \left[\text{Bias}[\hat{\boldsymbol{\theta}}]\right]^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right]$$

  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.
- See R script `mse_bv.R`.

Bias/variance decomposition of MSE (II)

$$
\begin{aligned}
\text{MSE} &= E_{\mathbf{D}_N}[(\theta - \hat{\boldsymbol{\theta}})^2] = \\
&= E_{\mathbf{D}_N}[(\theta - E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] + E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] = \\
&= E_{\mathbf{D}_N}[(\theta - E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}])^2] + E_{\mathbf{D}_N}[(E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] + \\
&\quad + E_{\mathbf{D}_N}[2(\theta - E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}])(E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})] = \\
&= E_{\mathbf{D}_N}[(\theta - E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}])^2] + E_{\mathbf{D}_N}[(E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] + \\
&\quad + 2(\theta - E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}])(E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}]) = \\
&= (E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \theta)^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right] = \\
&= \left[\text{Bias}[\hat{\boldsymbol{\theta}}]\right]^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right]
\end{aligned}
$$

Question

- Suppose $z_1, \ldots, z_N$ is a random sample of observations from a distribution with mean θ and variance σ^2.
- Study the unbiasedness of the three estimators of the mean μ:

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

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

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

Efficiency

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

Definition (Relative efficiency)

Let us consider two unbiased estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$. If

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

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

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

Sampling distributions for Gaussian r.v

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

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

It can be shown that the following relations hold

1. $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\mu, \sigma^2/N)$ and $\frac{\hat{\boldsymbol{\mu}}-\mu}{\frac{\sigma}{\sqrt{N}}} \sim \mathcal{N}(0,1)$
2. $\sqrt{N}(\hat{\boldsymbol{\mu}} - \mu)/\hat{\boldsymbol{\sigma}} \sim \mathcal{T}_{N-1}$ or $\frac{\hat{\boldsymbol{\mu}}-\mu}{\frac{\hat{\boldsymbol{\sigma}}}{\sqrt{N}}} \sim \mathcal{T}_{N-1}$ where $\mathcal{T}_{N-1}$ denotes the Student distribution with $N-1$ degrees of freedom.
3. if $E[|\mathbf{z} - \mu|^4] = \mu_4$ then $\mathrm{Var}\left[\hat{\boldsymbol{\sigma}}^2\right] = \frac{1}{N}\left(\mu_4 - \frac{N-3}{N-1}\sigma^4\right)$.

Likelihood

Let us consider

1. a density distribution $p_{\mathbf{z}}(z, \theta)$ which depends on a parameter θ
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

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

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

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 θ 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 $\hat{\boldsymbol{\theta}}$ is called the maximum likelihood estimator (m.l.e.).
- It is usual to consider the log-likelihood $l_N(\theta)$ since being $\log(\cdot)$ a monotone function, we have

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

Example: maximum likelihood

- Let us observe $N = 10$ realizations of a continuous variable $\mathbf{z}$:

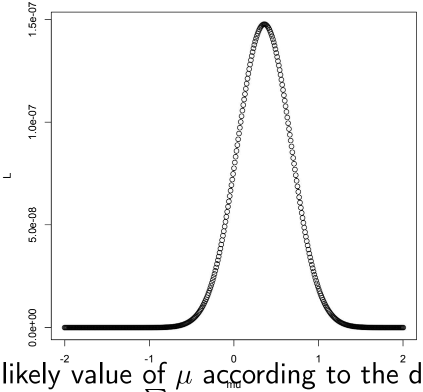
$$D_N = \{z_1, \dots, z_{10}\} = \{1.263, \dots, 2.405\}$$

- Suppose that the probabilistic model underlying the data is Gaussian with an unknown mean μ and a known variance $\sigma^2 = 1$.
- The likelihood $L_N(\mu)$ is a function of (only) the unknown parameter μ.
- By applying the maximum likelihood technique we have

$$\hat{\mu} = \arg\max_{\mu} L(\mu) = \arg\max_{\mu} \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z_i-\mu)^2}{2\sigma^2}} \right)$$

Example: maximum likelihood (II)

By plotting $L_N(\mu)$, $\mu \in [-2, 2]$ we have



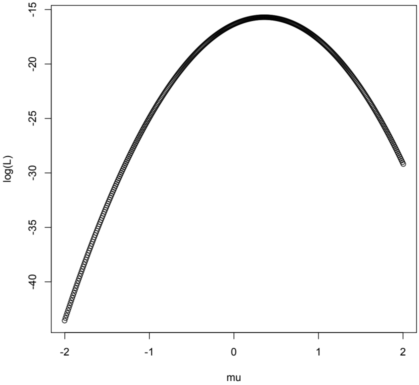
Then the most likely value of μ according to the data is $\hat{\mu} \approx 0.358$.
Note that in this case $\hat{\mu} = \frac{\sum z_i}{N}$.
R script `ml_norm.R`

Some considerations

- The likelihood measures 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 the highest probability of arising must be intuitively the best estimator of θ.
- M.l. can be considered a measure of how plausible the parameter values are in light of the data.
- The likelihood function is a function of the parameter θ.
- According to the classical approach to estimation, since θ is not a r.v., the likelihood function is NOT the probability function of θ.
- $L_N(\theta)$ is rather the conditional probability of observing the dataset D_N for a given θ.
- In other terms the likelihood is the probability of the data given the parameter and not the probability of the parameter given the data.

Example: log likelihood

Consider the previous example.
The behaviour of the log-likelihood for this model is



M.l. estimation

If we the take a parametric approach, the analytical form of the log-likelihood $l_N(\theta)$ is known. In many cases the function $l_N(\theta)$ is well behaved in being continuous with a single maximum away from the extremes of the range of variation of θ.

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

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

subject to

$$\left.\frac{\partial^2 l_N(\theta)}{\partial \theta^2}\right|_{\hat{\theta}_{\mathrm{ml}}} < 0$$

to ensure that the identified stationary point is a maximum.

Gaussian case: ML 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

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

- The log-likelihood is

$$l_N(\mu, \sigma^2) = \log L_N(\mu, \sigma^2) = \log \left[\prod_{i=1}^{N} p_{\mathbf{z}}(z_i, \mu, \sigma^2) \right] =$$

$$= \sum_{i=1}^{N} \log p_{\mathbf{z}}(z_i, \mu, \sigma^2) = -\frac{\sum_{i=1}^{N} (z_i - \mu)^2}{2\sigma^2} + N \log \left(\frac{1}{\sqrt{2\pi\sigma}} \right)$$

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

Gaussian case: ML estimators (II)

- Taking the derivatives with respect to μ and σ^2 and setting them equal to zero, we obtain

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

- 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.

Questions

1. - Let $\mathbf{z} \sim \mathcal{U}(0, M)$ and $F_{\mathbf{z}} \rightarrow D_N = \{z_1, \ldots, z_N\}$.
   - Find the maximum likelihood estimator of M.
2. - Let $\mathbf{z}$ have a Poisson distribution, i.e.

$$p_{\mathbf{z}}(z, \lambda) = \frac{e^{-\lambda}\lambda^{z}}{z!}$$

   - If $F_{\mathbf{z}}(z, \lambda) \rightarrow D_N = \{z_1, \ldots, z_N\}$, find the m.l.e. of λ

M.l. estimation with numerical methods

Computational difficulties may arise if

1. No analytical solution in explicit form exists for $\partial l_N(\theta)/\partial\theta = 0$. Iterative numerical methods must be used (see a numerical analysis class). This is particularly serious for a vector of parameters θ 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.

TP: Numerical optimization in R

- Suppose we known the analytical form of a one dimensional function $f(x) : I \to \mathbb{R}$.
- We want to find the value of $x \in I$ that minimizes the function.
- If no analytical solution is available, numerical optimization methods can be applied (see course "Calcul numérique").
- In the R language these methods are already implemented
- Let $f(x) = (x - 1/3)^2$ and $I = [0, 1]$. The minimum is given by

```
f <- function (x,a) (x-a)^2
xmin <- optimize(f, c(0, 1), tol = 0.0001, a = 1/3)
xmin
```

TP: Numerical max. of likelihood in R

- Let D_N be a random sample from the r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$.
- The minus log-likelihood function of the N samples can be written in R by

```
eml <- function(m,D,var) {
         N<- length(D)
         Lik<-1
         for (i in 1:N)
          Lik<-Lik*dnorm(D[i],m,sqrt(var))
         -log(Lik)
     }
```

- The numerical minimization of $-l_N(\mu, s^2)$ for a given $\sigma = s$ in the interval $I = [-10, 10]$ can be written in R in this form

```
xmin<-optimize( eml,c(-10,10),D=DN,var=s)
```

- Script `emp_ml.R`.

Properties of m.l. estimators

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

- $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is *asymptotically unbiased* but usually biased in small samples (e.g. $\hat{\sigma}^2_{\text{ml}}$).
- the Cramer-Rao theorem establishes a lower bound to the variance of an estimator
- $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is consistent.
- $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is asymptotically normally distributed around θ.

Interval estimation

- Unlike point estimation which is based on a one-to-one mapping from the space of data to the space of parameters, interval estimation maps D_N to an interval of Θ.
- A point estimator is a function which, given a dataset D_N generated from $F_{\mathbf{z}}(z, \theta)$, returns an estimate of θ.
- An **interval estimator** is a transformation which, given a dataset D_N, returns an interval estimate of θ.
- While an estimator is a random variable, an interval estimator is a random interval.
- Let $\underline{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ be the lower and the upper bound respectively.
- While an interval either contains or not a certain value, a random interval has a certain probability of containing a value.

Horseshoe game

From sampling distribution to confidence interval

Let us suppose we have an unbiased estimator $\hat{\boldsymbol{\theta}}$ with variance $\sigma^2_{\hat{\boldsymbol{\theta}}}$ and that its distribution is normal $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\theta, \sigma^2_{\hat{\boldsymbol{\theta}}})$
For instance, we can write

$$\text{Prob}\left\{\theta - 1.96\sigma^2_{\hat{\boldsymbol{\theta}}} \leq \hat{\boldsymbol{\theta}} \leq \theta + 1.96\sigma^2_{\hat{\boldsymbol{\theta}}}\right\} = 0.95$$

which entails

$$\text{Prob}\left\{\hat{\boldsymbol{\theta}} - 1.96\sigma^2_{\hat{\boldsymbol{\theta}}} \leq \theta \leq \hat{\boldsymbol{\theta}} + 1.96\sigma^2_{\hat{\boldsymbol{\theta}}}\right\} = 0.95$$

Interval estimation (II)

- Suppose that our interval estimator satisfies

  $$\text{Prob}\left\{\underline{\boldsymbol{\theta}} \leq \theta \leq \bar{\boldsymbol{\theta}}\right\} = 1 - \alpha \qquad \alpha \in [0, 1]$$

  then the random interval $[\underline{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}]$ is called a $100(1-\alpha)\%$ confidence interval of θ.
- Notice that θ is a fixed unknown value and that at each realization D_N the interval either does or does not contain the true θ.
- If we repeat the procedure of sampling D_N and constructing the confidence interval many times, then our confidence interval will contain the true θ at least $100(1-\alpha)\%$ of the time (i.e. 95% of the time if $\alpha = 0.05$).
- While an estimator is characterized by bias and variance, an interval estimator is characterized by its **endpoints** and **confidence**.