INFO-F-422: Statistical foundations of machine learning

Parametric estimation

Gianluca Bontempi

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

Parametric estimation

Given a r.v. $\mathbf{z}$ suppose that

1. we do not know $F_{\mathbf{z}}(z)$ but we can write

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

   where $\theta \in \Theta$ is a constant (or time-invariant) parameter,
2. sample D_N of N i.i.d. measurements of $\mathbf{z}$

Estimation: find a value $\hat{\theta}$ of the parameter θ so that the parametrized distribution $F_{\mathbf{z}}(z, \hat{\theta})$ closely matches the distribution $F_{\mathbf{z}}(z)$.

I.I.D. samples

- *I.I.D.*: Identically and Independently Distributed.
- Identically distributed: all observations sampled from the same distribution
- N.B.: data is random but the underlying distribution is fixed: this is the regularity we want to learn!

$$\text{Prob}\{\mathbf{z}_i = z\} = \text{Prob}\{\mathbf{z}_j = z\} \quad \forall \quad i, j = 1, \ldots, N \text{ and } z \in \mathcal{Z}$$

- Independently distributed: the fact that we observed a certain z_i does not influence the probability of observing the value z_j

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

Parametric estimation (II)

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

Two outcomes:

1. **point estimation**: specific value of Θ
2. **interval of confidence**: region of Θ.

Point estimation

- $\mathbf{z}$ with a parametric distribution $F_{\mathbf{z}}(z, \theta)$, $\theta \in \Theta$.
- parameter θ as function(al) of F

$$\theta = t(F)$$

- N i.i.d. observations $D_N = \{z_1, z_2, \ldots, z_N\}$.
- *Point estimator*: function

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

of dataset D_N. The value of the function is called the estimate.

Methods of constructing estimators

How to define h?

We will focus on:

1. Plug-in principle
2. Maximum likelihood

Empirical distribution function

Given a i.i.d. sample

$$F_{\mathbf{z}} \to \{z_1, z_2, \ldots, z_N\}$$

where

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

the *empirical distribution* function is

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

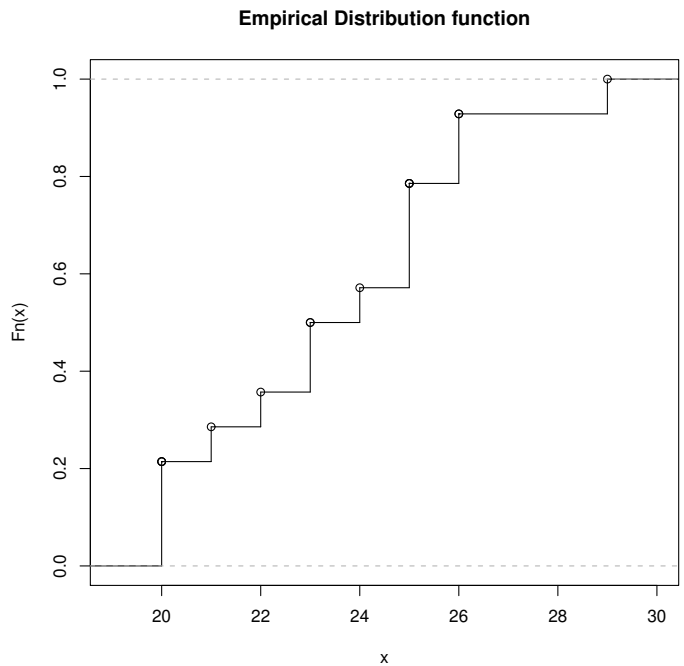
where $N(z)$ is the number of samples in D_N that do not exceed z.

TP R: empirical distribution

- Dataset of $N = 14$ observed ages

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

- Empirical distribution function $\hat{F}_{\mathbf{z}}$ is a staircase function with discontinuities at the points z_i.



Plug-in principle to define an estimator

- Sample D_N from $F_{\mathbf{z}}(z, \theta)$ where $\theta = t(F(z))$. .
- **Plug-in estimate** of θ:

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

where distribution function is replaced by empirical distribution.

Sample average

- Consider the expectation of $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$

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

with θ unknown.

- Sample $F_{\mathbf{z}} \to D_N$
- *Plug-in* point estimate of θ is the **sample average**

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

which is a function h of the dataset D_N.

Sample variance

- $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$ whose mean μ and variance σ^2 are unknown.
- Sample $F_{\mathbf{z}} \to D_N$.
- *Plug-in* estimate of σ^2 is the **sample variance**

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

  where $\hat{\mu}$ is the plug-in estimate of μ.
- Why $N-1$ instead of N at the denominator?

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:

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

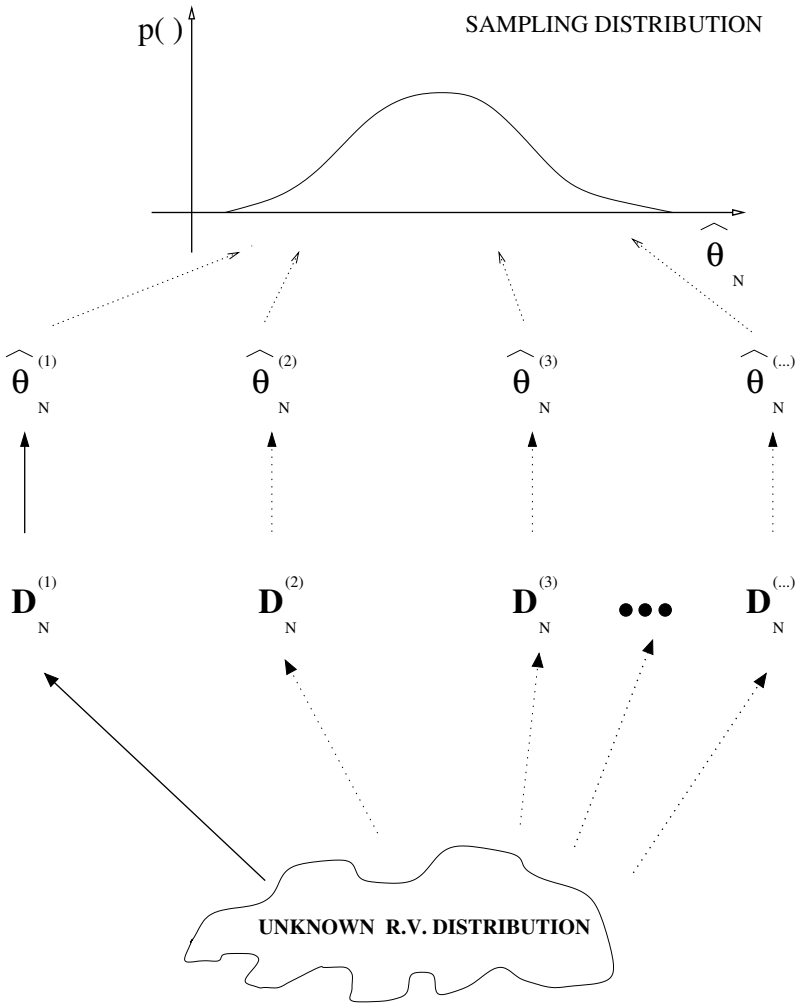
- Sample correlation:

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

Sampling distribution

- Point estimate is
$$\hat{\theta} = h(D_N)$$
where D_N is the realisation of a random variable $\mathbf{D}_N$.
- If $\mathbf{D}_N$ is random, the *point estimator*
$$\hat{\boldsymbol{\theta}} = h(\mathbf{D}_N)$$
is random as well and its probability distribution is the sampling distribution.
- Sampling distribution is a **theoretical** notion (cannot be directly observed with a single dataset).

Sampling or finite-sample distribution



See the Python scripts `Estimation/samdis.py` and `Estimation/est_step.py`.

Monte Carlo illustration of a sampling distribution

To illustrate the theoretical notion of sampling distribution, we need to generate random samples from $F_{\mathbf{z}}(z, \theta)$.

1: $S = \{\}$
2: **for** $r = 1$ to R **do**
3: $\quad F_{\mathbf{z}} \rightarrow D_N = \{z_1, z_2, \ldots, z_N\}$ *// sample dataset*
4: $\quad \hat{\theta} = h(D_N)$ *// compute estimate*
5: $\quad S = S \cup \{\hat{\theta}\}$
6: **end for**
7: Plot histogram of S
8: Compute statistics of S (mean, variance)
9: Study distribution of S with respect to θ

Python script (Gaussian **z**)

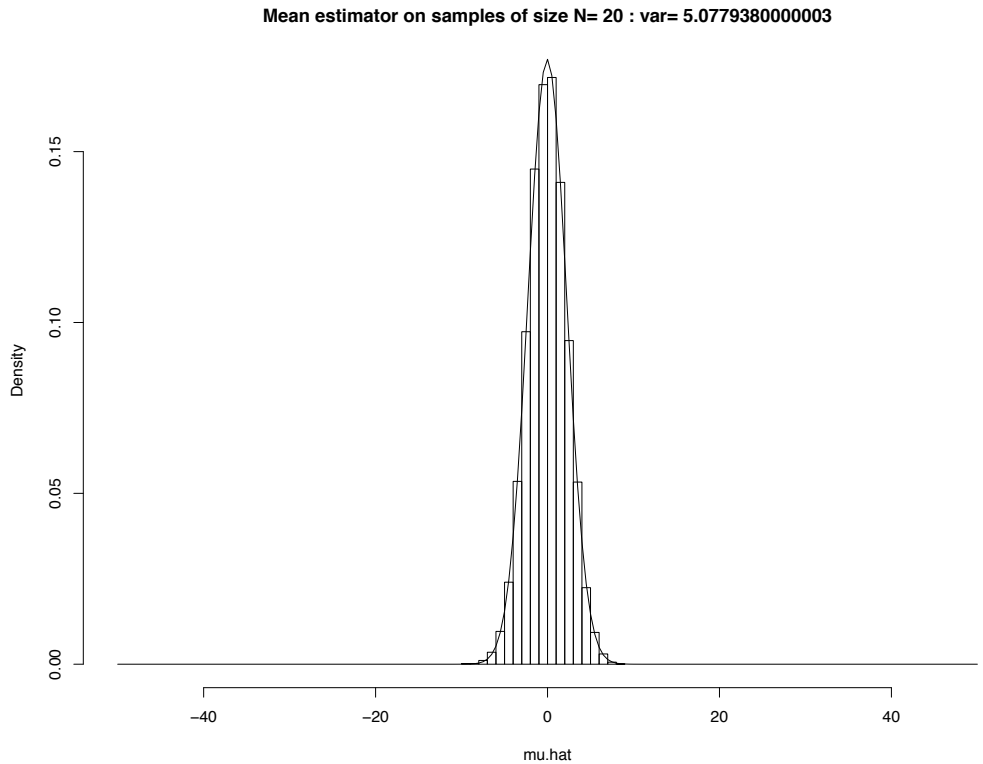
```
import numpy as np
import matplotlib.pyplot as plt

mu = 0          # parameter
R = 10000       # number trials
N = 20          # size dataset

mu_hat = np.zeros(R)
for r in range(R):
    D = np.random.normal(loc=mu, scale=10, size=N)
    # random generator
    mu_hat[r] = np.mean(D)
    # estimator

plt.hist(mu_hat)      # histogram
plt.show()
```

Histogram



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

Bias and variance

How accurate is $\hat{\theta}$?

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 *biased* with bias

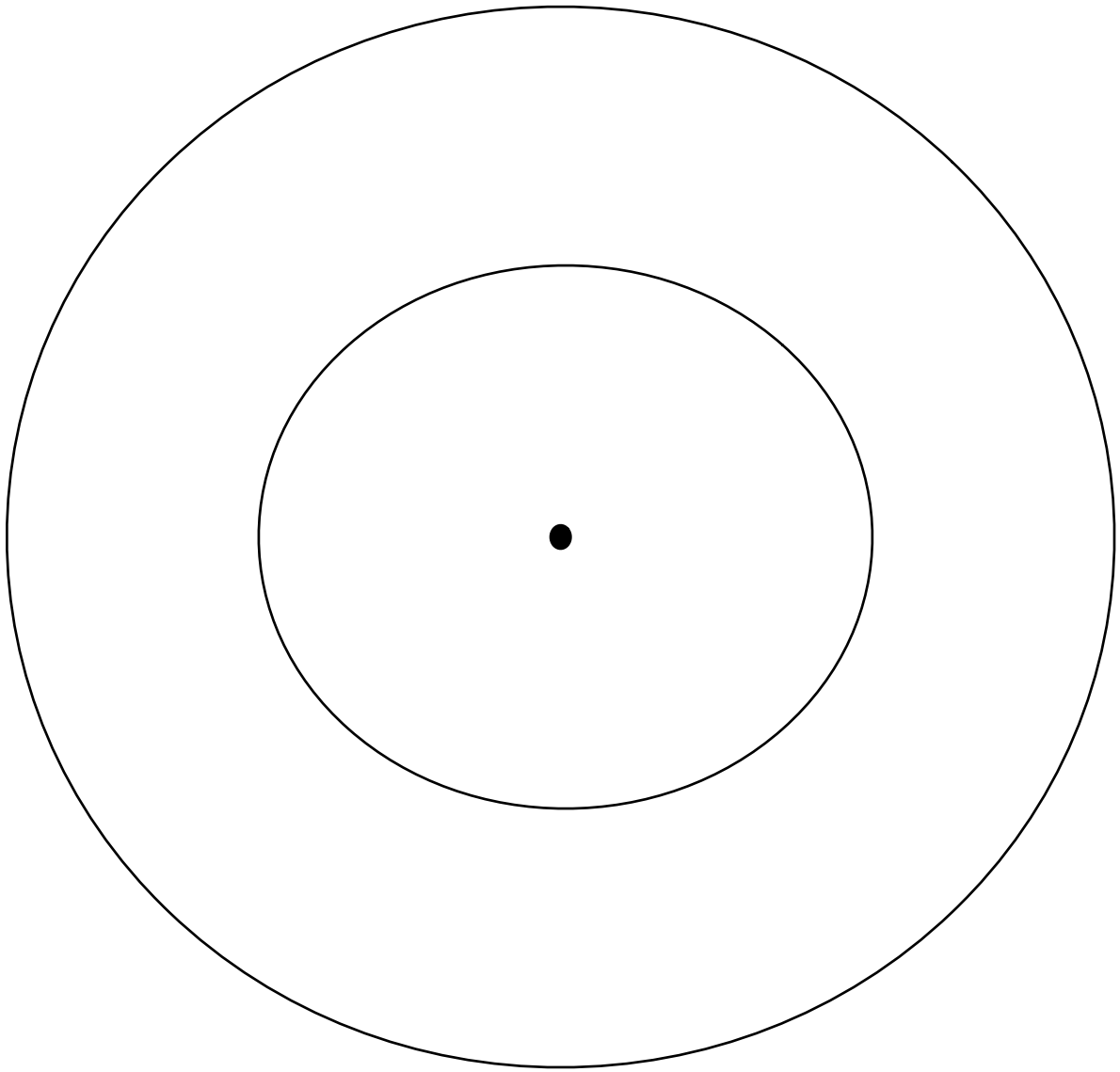
$$\text{Bias}[\hat{\boldsymbol{\theta}}] = E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \theta$$

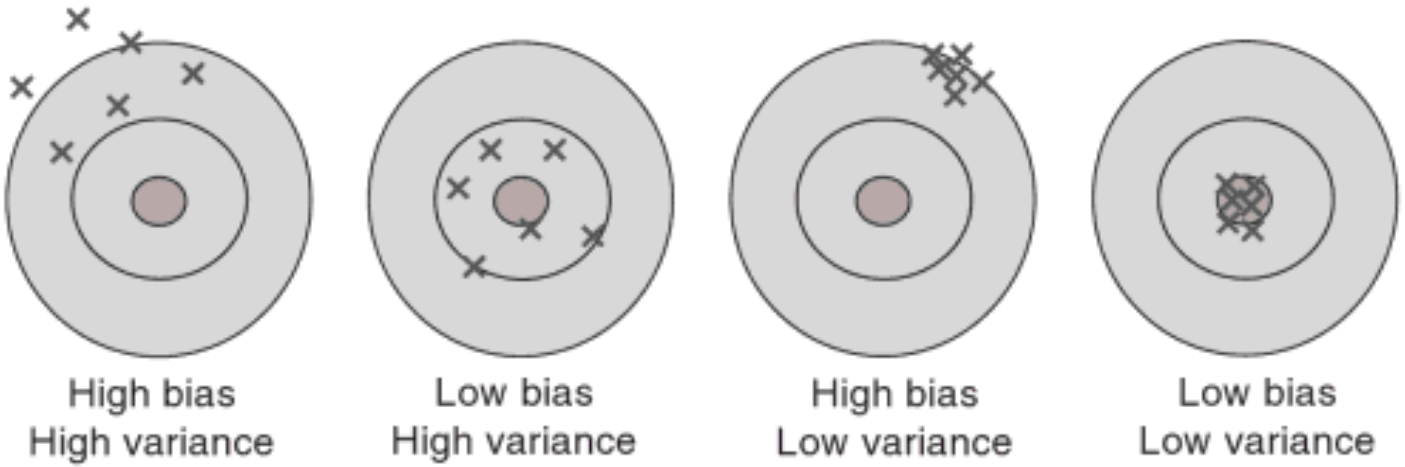
Definition

The variance of $\hat{\boldsymbol{\theta}}$ is the variance of the sampling distribution

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

Estimation and darts




High bias
High variance
Low bias
High variance
High bias
Low variance
Low bias
Low variance

Some considerations

- unbiased estimator takes on average the right value.
- many unbiased estimators may exist for θ.
- if $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of θ, it may happen that $f(\hat{\boldsymbol{\theta}})$ be a BIASED estimator of $f(\theta)$.
  - in general $\hat{\boldsymbol{\sigma}}$ is a biased estimator of σ though $\hat{\boldsymbol{\sigma}}^2$ is an unbiased estimator of σ^2.
- a biased estimator with a known bias (not depending on θ) can be easily made unbiased (e.g. compensating for the bias).
- sample average $\hat{\boldsymbol{\mu}}$ and sample variance $\hat{\boldsymbol{\sigma}}^2$ are unbiased estimators of the mean $E[\mathbf{z}]$ and the variance $\mathrm{Var}[\mathbf{z}]$, respectively.

Useful relationships

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

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

  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 $\hat{\mu}$

- Let μ and σ^2 be mean and variance of $F_{\mathbf{z}}(\cdot)$.
- i.i.d. sample $D_N \leftarrow F_{\mathbf{z}}$.
- $$E_{\mathbf{D}_N}[\hat{\mu}] = E_{\mathbf{D}_N}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\right] = \frac{\sum_{i=1}^{N}E[\mathbf{z}_i]}{N} = \frac{N\mu}{N} = \mu$$
- *sample average estimator* is not biased whatever is the 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

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

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

What is the bias of the estimator of the variance?

Given an i.i.d. $D_N \leftarrow F_{\mathbf{z}}$ it can be shown that

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

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

Considerations

- Previous results are **independent** of the form $F(\cdot)$ of the distribution.
- Variance of $\hat{\boldsymbol{\mu}}$ is $1/N$ times the variance of $\mathbf{z}$. Rationale for collecting several samples: the larger N, the smaller is $\mathrm{Var}\,[\hat{\boldsymbol{\mu}}]$, 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

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

- **Standard error** $\sqrt{\mathrm{Var}\,[\hat{\boldsymbol{\mu}}]}$ indicates 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 .

Homework

Let $\mathbf{z}$ such that $E[\mathbf{z}] = \mu$ and $\mathrm{Var}[\mathbf{z}] = \sigma^2$. 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

Mean-square error (MSE)

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

- 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 (**bias-variance** decomposition).
- See R script `Estimation/mse_bv.py`.

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 i.i.d. sample of observations from a distribution with mean μ and variance σ^2.
- Study the unbiasedness of the three estimators of the mean μ:

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

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

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

Sampling distributions for Gaussian r.v

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

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

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 $E[|\mathsf{z} - \mu|^4] = \mu_4$ 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 i.i.d. $D_N = \{z_1, z_2, \ldots, z_N\}$ from such 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

- 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 r.v $\hat{\boldsymbol{\theta}}$ is 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 $\mathbf{z}$:

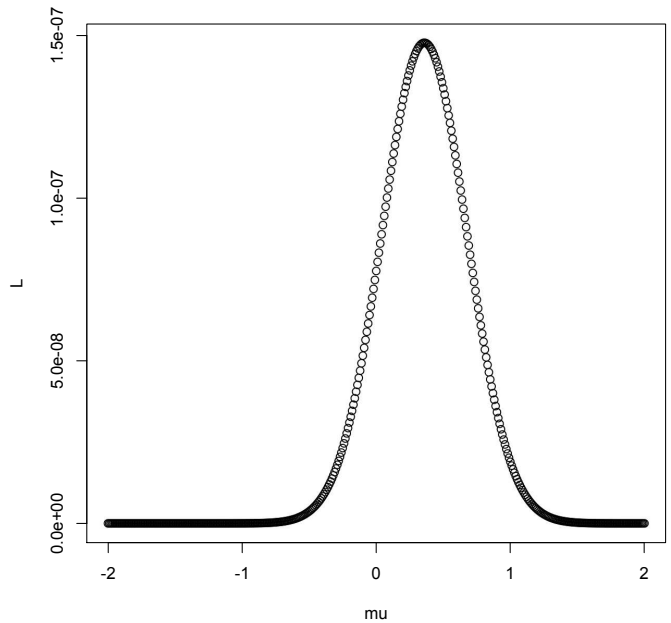
$$D_N = \{z_1, \ldots, z_{10}\} = \{1.263, \ldots, 2.405\}$$

- 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



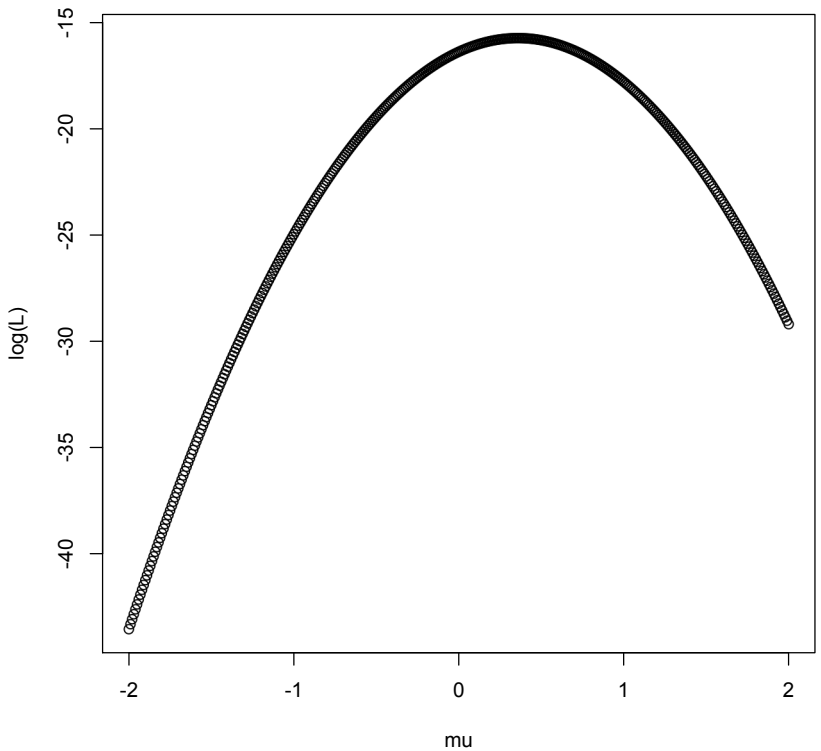
Then the most likely value of μ according to the data is $\hat{\mu} \approx 0.358$.
Note that in this case $\hat{\mu} = \frac{\sum z_i}{N}$.
R script `Estimation/ml_norm.py`

Some considerations

- Likelihood measures the relative abilities of the various parameter values to *explain* the observed data.
- Rationale: the value of the parameter under which the observed data have the highest probability of arising is the best estimator of θ.
- Likelihood function is a function of the parameter θ.
- Likelihood function is NOT the probability function of θ (θ is constant).
- $L_N(\theta)$ is rather the conditional probability of observing the dataset D_N for a given θ.
- 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 the parametric distribution is given, 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

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

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

$$\hat{\mu}_{\text{ml}} = \frac{\sum_{i=1}^{N} z_i}{N} = \hat{\mu}$$

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

- M.l. estimator of the mean coincides with the sample average
- 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 $F_{\mathbf{z}} \to D_N = \{z_1, \ldots, z_N\}$.
   - 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 $F_{\mathbf{z}}(z, \lambda) \to D_N = \{z_1, \ldots, z_N\}$, find the m.l.e. of λ

M.I. 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.

Numerical optimization

- 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 Python language these methods are already implemented, e.g. function `minimize` in the package `scipy.optimize`.

Python example: Numerical max. likelihood

- 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 Python by

```
def eml(m, D, var):   # emp likelihood function (1 argument)
    N = len(D)
    LLik = 0
    for i in range(N):
        LLik += np.log(norm.pdf(D[i], m, np.sqrt(var)))
    return -LLik
```

- The numerical minimization of $-l_N(\mu, s^2)$ for a given $\sigma = s$ in the interval $I = [-1, 1]$ can be written in Python in this form

```
xmin = minimize_scalar(eml, args=(DN, 1),
bounds=(-1, 1), method='bounded')
```

- Script `Estimation/emp_ml.py`.

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 asymptotically normally distributed around θ.

Interval estimation

- Unlike point estimation, interval estimation maps D_N to an interval 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 $\underline{\theta}$ and $\bar{\theta}$ 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 estimator $\hat{\boldsymbol{\theta}}$ of the parameter θ which is

- unbiased
- with variance $\sigma^2_{\hat{\boldsymbol{\theta}}}$
- with a Normal sampling distribution $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\theta, \sigma^2_{\hat{\boldsymbol{\theta}}})$

We can write

$$\text{Prob}\left\{\theta - 1.96\sigma_{\hat{\boldsymbol{\theta}}} \leq \hat{\boldsymbol{\theta}} \leq \theta + 1.96\sigma_{\hat{\boldsymbol{\theta}}}\right\} = 0.95$$

which entails

$$\text{Prob}\left\{\hat{\boldsymbol{\theta}} - 1.96\sigma_{\hat{\boldsymbol{\theta}}} \leq \theta \leq \hat{\boldsymbol{\theta}} + 1.96\sigma_{\hat{\boldsymbol{\theta}}}\right\} = 0.95$$

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**.