Statistics and learning

Statistical estimation

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Friday 11^{th} January 2013

Statistical estimation

Steps in estimation procedure:

- Consider a population (size N) described by a random variable X (known or unknown distribution) with parameter θ,

Statistical estimation

Steps in estimation procedure:

- Consider a population (size N) described by a random variable X (known or unknown distribution) with parameter θ,
- a sample with $n \leq N$ independent observations $(x_1 \dots x_n)$ is extracted,

Statistical estimation

Steps in estimation procedure:

- Consider a population (size N) described by a random variable X (known or unknown distribution) with parameter θ,
- a sample with $n \leq N$ independent observations $(x_1 \dots x_n)$ is extracted,
- θ is estimated through a **statistic** ($=$function of X_i's): $\hat{\theta} = T(X_1 \dots X_n)$.

Statistical estimation

Steps in estimation procedure:

- Consider a population (size N) described by a random variable X (known or unknown distribution) with parameter θ,
- a sample with $n \leq N$ independent observations $(x_1 \ldots x_n)$ is extracted,
- θ is estimated through a **statistic** (=function of X_i's): $\hat{\theta} = T(X_1 \ldots X_n)$.

Note: independence is true only if drawing is made with replacement.
Without replacement, the approximation is ok if $n << N$.

Statistical estimation

Steps in estimation procedure:

- Consider a population (size N) described by a random variable X (known or unknown distribution) with parameter θ,
- a sample with $n \leq N$ independent observations $(x_1 \dots x_n)$ is extracted,
- θ is estimated through a **statistic** (=function of X_i's): $\hat{\theta} = T(X_1 \dots X_n)$.

Note: independence is true only if drawing is made with replacement. Without replacement, the approximation is ok if $n << N$.

Mean estimation

Estimate the average life span of a bulb...

Point estimation of a parameter

Recall

At our disposal: n realisations of random variables $(X_1 \dots X_n)$ iid. Some parameters can be of interest. Direct computation not feasible so estimation needed. **Objective** here: tools and maths grounds for estimation.

Point estimation of a parameter

Recall

At our disposal: n realisations of random variables $(X_1 \dots X_n)$ iid. Some parameters can be of interest. Direct computation not feasible so estimation needed. **Objective** here: tools and maths grounds for estimation.

Definitions

- Statistical model: definition of a probability distribution P_θ (joint if discrete rv and density if continuous rv), with θ is a (vector of) unknown parameter(s).
- **Statistic**: $T : \mathbb{R}^n \to \mathbb{R}, (x_i)_{i=1\dots n} \mapsto T(x_1 \dots x_n)$. Examples: empirical mean or variance (known/unknown mean).

Estimator, bias, comparison

Exercice

Lift can bear $1,000\,kg$. User weight $\sim \mathcal{N}(75, 16^2)$.

- Max. number of people allowed in it if $P(\text{lift won't take off}) = 10^{-6}$?
- Lift manufacturer allows 11 people inside. $P(\text{overweight}) =??$

Estimator, bias, comparison

Exercice

Lift can bear $1,000\,kg$. User weight $\sim \mathcal{N}(75, 16^2)$.

- Max. number of people allowed in it if $P(\text{lift won't take off}) = 10^{-6}$?
- Lift manufacturer allows 11 people inside. $P(\text{overweight}) = ??$

Definitions

- **Estimator** of an unknown parameter θ: a statistic denoted $\hat{\theta}$ (observed values are approximations of θ). The **bias** associated to $\hat{\theta}$ is $E[\hat{\theta}] - \theta$ (if nul, $\hat{\theta}$ is said to be unbiased). Ex: (exercices) (i) the empirical mean is an unbiaised estimator for the (theoretical) mean. $S_n^2 := \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ is a biased estimator for σ^2.
- $\hat{\theta}$ is **asymptotically unbiased** if $\lim_{n\to\infty} E[\hat{\theta}] = \theta$.
- $\hat{\theta}_1$ and $\hat{\theta}_2$: 2 unbiased estimator for θ; $\hat{\theta}_1$ is better than $\hat{\theta}_2$ if $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$; in practice, $\hat{\theta}_1$ converges faster than $\hat{\theta}_2$.

Convergence of estimators

Def: $\hat{\theta}$ converges in probability towards θ if $\forall \epsilon > 0$, $P(|\hat{\theta} - \theta| < \epsilon) \rightarrow_n 1$.

Convergence of estimators

Def: $\hat{\theta}$ converges in probability towards θ if $\forall \epsilon > 0$, $P(|\hat{\theta} - \theta| < \epsilon) \rightarrow_n 1$.

Theorem

An (asymptotically) unbiased estimator s.t. $\lim_n Var(\hat{\theta}) = 0$ *converges in probability towards* θ*.*

Convergence of estimators

Def: $\hat{\theta}$ converges in probability towards θ if $\forall \epsilon > 0$, $P(|\hat{\theta} - \theta| < \epsilon) \rightarrow_n 1$.

Theorem

An (asymptotically) unbiased estimator s.t. $\lim_n Var(\hat{\theta}) = 0$ *converges in probability towards* θ.

Theorem

An unbiased estimator $\hat{\theta}$ *with the following technical regularity hypotheses (H1-H5) verifies* $Var(\hat{\theta}) > V_n(\theta)$, *with the Cramer-Rao bound* $V_n(\theta) := (-E[\frac{\partial^2 \log f(X_1 \dots X_n;\theta)}{\partial \theta^2}])^{-1}$ *(inverse of Fisher information).*

(H1) *the support* $D := \{X, f(x;\theta) > 0\}$ *does not depend upon* θ.

(H2) θ *belongs to an open interval* I.

(H3) *on* $I \times D$, $\frac{\partial f}{\partial \theta}$ *and* $\frac{\partial^2 f}{\partial \theta^2}$ *exist and are integrable over* x.

(H4) $\theta \mapsto \int_A f(x;\theta)dx$ *has a second order derivative* ($x \in I, A \in B(\mathbb{R})$)

(H5) $(\frac{\partial \log f(X;\theta)}{\partial \theta})^2$ *is integrable.*

Application to the estimation of a $|\mathcal{N}|$

Definition

An unbiased estimator $\hat{\sigma}$ for θ is **efficient** if its variance is equal to the Cramer-Rao bound. It is the best possible among unbiased estimators.

Application to the estimation of a $|\mathcal{N}|$

Definition

An unbiased estimator $\hat{\sigma}$ for θ is **efficient** if its variance is equal to the Cramer-Rao bound. It is the best possible among unbiased estimators.

Exercice

Let $(X_i)_{i=1\dots n}$ iid rv $\sim \mathcal{N}(m,\sigma^2)$. $Y_i := |X_i - m|$ is observed.

- Density of Y_i ? Compute $E[Y_i]$? Interpretation compared to σ ?
- Let $\hat{\sigma} := \sum_i a_i Y_i$. If we want $\hat{\sigma}$ to be unbiased, give a constraint on (a_i)'s. Under this constraint, show that $Var(\hat{\sigma})$ is minimum iif all a_i are equal. In this case, give the variance.
- Compare the Cramer-Rao bound to the above variance. Is the built estimator efficient ?

Likelihood function

Definition

The likelihood of a rv $\mathbf{X} = (X_1 \dots X_n)$ is the function L:

$$L : \mathbb{R}^n \times \Theta \longrightarrow \mathbb{R}^+$$

$$(x, \theta) \longmapsto L(x;\theta) := \begin{cases} f(x;\theta), \text{ the density of } \mathbf{X} \\ \text{or} \\ P_\theta(X_1 = x_1 \dots X_n = x_n), \text{ if } \mathbf{X} \text{ discrete} \end{cases}$$

Likelihood function

Definition

The likelihood of a rv $\mathbf{X} = (X_1 \dots X_n)$ is the function L:

$$L : \mathbb{R}^n \times \Theta \longrightarrow \mathbb{R}^+$$

$$(x, \theta) \longmapsto L(x;\theta) := \begin{cases} f(x;\theta), \text{ the density of } \mathbf{X} \\ \text{or} \\ P_\theta(X_1 = x_1 \dots X_n = x_n), \text{ if } \mathbf{X} \text{ discrete} \end{cases}$$

Examples

- X_i Gaussian iid rv:

$$L(x;\theta) = \prod_i f(x_i;\theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_i \left(\frac{x_i - m}{\sigma}\right)^2\right]$$

- X_i Bernouilli iid rv: $L(x;\theta) = p^{\sum x_i}(1-p)^{n-\sum x_i}$

Maximum likelihood estimation (MLE)

Definition

$$\hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} (\log) L(x_1 \dots x_n; \theta)$$

Interpretation: $\hat{\theta}_{MLE}$ is the parameter value that gives maximum probability to the observed values or random variables...

Maximum likelihood estimation (MLE)

Definition

$$\hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} (\log) L(x_1 \dots x_n; \theta)$$

Interpretation: $\hat{\theta}_{MLE}$ is the parameter value that gives maximum probability to the observed values or random variables...

Remark: the EMV does not always exists (possible alternatives: least square or moments). When it exists, it is not necessarily unique, can be biased or not efficient. However...

Maximum likelihood estimation (MLE)

Definition

$$\hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} (\log) L(x_1 \dots x_n; \theta)$$

Interpretation: $\hat{\theta}_{MLE}$ is the parameter value that gives maximum probability to the observed values or random variables...

Remark: the EMV does not always exists (possible alternatives: least square or moments). When it exists, it is not necessarily unique, can be biased or not efficient. However...

Theorem

- $\hat{\theta}_{MLE}$ *is asymptotically unbiased and efficient.*
- $\frac{\hat{\theta}_{MLE} - \theta}{V_n(\theta)} \longrightarrow_n \mathcal{N}(0,1)$, *where* $V_n(\theta)$ *is the Cramer-Rao bound.*
- $\hat{\theta}_{MLE}$ *converges to* θ *in squared mean.*

Maximum likelihood estimation (MLE)

Definition

$$\hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} (\log) L(x_1 \dots x_n; \theta)$$

Interpretation: $\hat{\theta}_{MLE}$ is the parameter value that gives maximum probability to the observed values or random variables...

Remark: the EMV does not always exists (possible alternatives: least square or moments). When it exists, it is not necessarily unique, can be biased or not efficient. However...

Theorem

- $\hat{\theta}_{MLE}$ *is asymptotically unbiased and efficient.*
- $\frac{\hat{\theta}_{MLE}-\theta}{V_n(\theta)} \longrightarrow_n \mathcal{N}(0,1)$, *where* $V_n(\theta)$ *is the Cramer-Rao bound.*
- $\hat{\theta}_{MLE}$ *converges to* θ *in squared mean.* *'MLE for a proportion' exercice ?*

Sufficient statistic

Remark/definition

Any realisation (x_i) of a rv X, unknown distribution but parameterised by θ, from a sample contains information on θ. If the statistic summarises all possible information from the sample, it is sufficient. In other words "no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter" (Fisher 1922)
In mathematical terms: $P(X = x|T = t, \theta) = P(X = x|T = t)$

Sufficient statistic

Remark/definition

Any realisation (x_i) of a rv X, unknown distribution but parameterised by θ, from a sample contains information on θ. If the statistic summarises all possible information from the sample, it is sufficient. In other words "no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter" (Fisher 1922)
In mathematical terms: $P(X = x|T = t, \theta) = P(X = x|T = t)$

Theorem (Fisher-Neyman)

T(X) is sufficient if there exist 2 functions g and h s.t.
$L(x; \theta) = g(t; \theta)h(x)$

Implication: in the context of MLE, 2 samples yielding the same value for T yield the same inferences about θ. (dep. on θ is only in conjunction with T).

Quantiles

Definition

The cumulative distribution function F ($F(x) = \int_{-\infty}^{x} f(t)dt$, with f density of X) is a non-decreasing function $\mathbb{R} \rightarrow [0;1]$. Its inverse F^{-1} is called the quantile function. $\forall \beta \in]0;1[$, the β-quantile is defined by $F^{-1}(\beta)$.

Quantiles

Definition

The cumulative distribution function F ($F(x) = \int_{-\infty}^{x} f(t)dt$, with f density of X) is a non-decreasing function $\mathbb{R} \rightarrow [0;1]$. Its inverse F^{-1} is called the quantile function. $\forall \beta \in]0;1[$, the β-quantile is defined by $F^{-1}(\beta)$.

In particular: $P(X \leq F^{-1}(\beta)) = \beta$ and $P(X \geq F^{-1}(\beta)) = 1 - \beta$

Quantiles

Definition

The cumulative distribution function F ($F(x) = \int_{-\infty}^{x} f(t)dt$, with f density of X) is a non-decreasing function $\mathbb{R} \rightarrow [0;1]$. Its inverse F^{-1} is called the quantile function. $\forall \beta \in]0;1[$, the β-quantile is defined by $F^{-1}(\beta)$.

In particular: $P(X \leq F^{-1}(\beta)) = \beta$ and $P(X \geq F^{-1}(\beta)) = 1 - \beta$

In practice, either quantile are read from tables: either F or F^{-1} (old-fashioned) or they are computed using statistics softwares on computers.
Quantile for the Gaussian distribution will be denoted z_β.

Interval estimation

$\hat{\theta}$: a point estimation of θ; even in favourable situations, it is very unlikely that $\hat{\theta} = \theta$. How close is it ? Could an interval that contains the true value of θ with say a high probability (low error) be built ? Not too big (informative), but not too restricted neither (for the true value has a great chance of being in it).

Interval estimation

$\hat{\theta}$: a point estimation of θ; even in favourable situations, it is very unlikely that $\hat{\theta} = \theta$. How close is it ? Could an interval that contains the true value of θ with say a high probability (low error) be built ? Not too big (informative), but not too restricted neither (for the true value has a great chance of being in it).

Typically, a $1/\sqrt{n}$-neighbourhood of $\hat{\theta}$ will do the job. It is much more useful than many digits in an estimator to give an interval with the associated error.

Interval estimation

$\hat{\theta}$: a point estimation of θ; even in favourable situations, it is very unlikely that $\hat{\theta} = \theta$. How close is it ? Could an interval that contains the true value of θ with say a high probability (low error) be built ? Not too big (informative), but not too restricted neither (for the true value has a great chance of being in it).
Typically, a $1/\sqrt{n}$-neighbourhood of $\hat{\theta}$ will do the job. It is much more useful than many digits in an estimator to give an interval with the associated error.

Definition

1. A confidence interval $\hat{I}_n$ is defined by a couple of estimators: $\hat{I}_n = [\hat{\theta}_1; \hat{\theta}_2]$.
2. its associated confidence level $1 - \alpha$ ($\alpha \in [0; 1]$) is s.t. $P(\theta \in \hat{I}_n) \geq 1 - \alpha$.
3. $\hat{I}_n$ is asymptotically of level at least $1 - \alpha$ if $\forall \epsilon > 0$, $\exists N_e$ s.t. $P(\theta \in \hat{I}_n) \geq 1 - \alpha - \epsilon$ for $n \geq N_e$.

Confidence intervals you need to know

a partial typology

- $X_i \sim \mathcal{N}(m, \sigma^2)$, with σ^2 known, then $I(m) = [\bar{x} + / - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$.
- when σ^2 is unknown, it becomes $I(m) = [\bar{x} + / - t_{n-1;1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}}]$, with $s_{n-1}^2 := \frac{\sum(x_i-\bar{x})^2}{n-1}$ and $t_{n-1;1-\alpha/2}$ the quantile of a Student distribution with $n-1$ degrees of freedom (df). Note that $t_{n-1;1-\alpha/2} \simeq_n z_{1-\alpha/2}$.
- if Gaussianity is lost, we can only derive asymptotic confidence intervals.
- as for σ^2: if m is known $I_\alpha = [\frac{n\widehat{\sigma^2}}{\chi^2_{n;1-\alpha/2}}; \frac{n\widehat{\sigma^2}}{\chi^2_{n;\alpha/2}}]$
- when m is unknown: $I_\alpha = [\frac{(n-1)S^2_{n-1}}{\chi^2_{n-1;1-\alpha/2}}; \frac{(n-1)S^2_{n-1}}{\chi^2_{n;\alpha/2}}]$
- confidence interval for a proportion: exercices (if time permits)
- for other distributions: use the Cramer-Rao bound !

Next time

Multivariate descriptive statistics !

Next time

Multivariate descriptive statistics !

Some notions of (advanced) algebras wil be needed. *E.g.* Matrices, operations, inverse, rank, projection, metrics, scalar product, eigenvalues/vectors, matrix norm, matrix approximation . . . .