Time Series

Peter Lukianchenko

4 March 2023

Point Estimators

Definition

Estimators are random variables and therefore have probability distributions, known as sampling distributions. Two important properties of probability distributions are the mean and variance.

The main objective is to create a formal criterion which combines both of these properties to assess the relative performance of different estimators.

Let $\hat{\theta}$ be an estimator of the population parameter θ. The bias of an estimator could be defined as:

$$Bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$$

Point Estimators

Properties

An estimator is:

- *Positively* biased estimator means the estimator would systematically overestimate the parameter by the size of the bias, on average:

$$E\left(\hat{\theta}\right) - \theta > 0$$

- *Negatively* means the estimator would systematically underestimate the parameter by the size of the bias, on average:

$$E\left(\hat{\theta}\right) - \theta < 0$$

- *Unbiased* means the estimator would estimate the parameter correctly, on average:

$$E\left(\hat{\theta}\right) - \theta = 0$$

Point Estimators

Definition

The **variance of an estimator**, denoted *Var(θ)*, is obtained directly from the estimator's sampling distribution.

The **mean squared error (MSE)** of an estimator is the average squared error. Formally, this is defined as:

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

It is possible to decompose the MSE into components involving the bias and variance of an estimator:

$$Var(\hat{\theta}) = E(X^2) - (E(X))^2$$

$$E(X^2) = Var(\hat{\theta}) + (E(X))^2$$

Point Estimators

Definition

Also, note that for any constant k, $Var(X \pm k) = Var(X)$, that is adding or subtracting a constant has no effect on the variance of a random variable. Noting that the true parameter θ is some (unknown) constant, it immediately follows, by setting $X = (\hat{\theta} - \theta)$, that:

$$
\begin{aligned}
MSE(\hat{\theta}) &= E\left((\hat{\theta} - \theta)^2\right) \\
&= Var(\hat{\theta} - \theta) + \left(E(\hat{\theta} - \theta)\right)^2 \\
&= Var(\hat{\theta} - \theta) + \left(Bias(\hat{\theta})\right)^2
\end{aligned}
$$

Point Estimators

Important notes

i. $\hat{\mu} = \bar{X}$ is the better estimator than X_1 :

$$MSE(\hat{\mu}) = \frac{\sigma^2}{n} < MSE(\bar{X}) = \sigma^2$$

Point Estimators

Important notes

ii. As n → ∞, MSE($\bar{X}$) → 0, i.e. when the sample size goes to infinity, the error in estimation goes to 0. Such an estimator is called a (mean-square) **consistent estimator**.

Consistency is a reasonable requirement. It may be used to rule out some silly estimators.

For $\tilde{\mu} = \frac{X_1+X_4}{2}, MSE\ (\tilde{\mu}) = \frac{\sigma^2}{2}$ which does not converge to 0 as n → ∞.
This is due to the fact that only a *small portion of information* (i.e. X_1 and X_4) was used in the estimation.

Maximum likelihood estimation

Definition

Let $f(x_1, x_2, \ldots x_n; \theta)$ be the joint probability density function (or probability function) for random variables $X_1, X_2, \ldots X_n$. Then the maximum likelihood estimator (MLE) of θ based on the observations $X_1, X_2, \ldots X_n$ is defined as the $\hat{\theta}$ for which:

$$f(x_1, x_2, \ldots x_n; \theta) = \max_{\theta} f(X_1, X_2, \ldots X_n; \theta)$$

Maximum likelihood estimation

Definition

The likelihood function is defined as:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

- the likelihood function is the function of θ, while $X_1, X_2, \dots X_n$ are treated as constants (as given observations);
- the likelihood function reflects the information about the unknown parameter θ in the data $X_1, X_2, \dots X_n$.

P.	α	4α	$1-5\alpha$
x	1	2	3

SRS → $P(A \cap B) = PA \cdot PB$

$\hat{\alpha}_{ML}$ - ? : set: 3; 3; 2; 1; 1; 2; 3

$L(\alpha) = (1-5\alpha)^3 \cdot (4\alpha)^2 \cdot \alpha^2$

$\ell(\alpha) = \ln L(\alpha) = 3\ln(1-5\alpha) + \ln 16 + 4\ln\alpha \to \max_{\alpha}$

F.O.C. $\frac{\partial \ell}{\partial \alpha} = \frac{-5 \cdot 3}{1-5\alpha} + \frac{4}{\alpha} = \frac{-15\alpha + 4 - 20\alpha}{\alpha(1-5\alpha)} = 0$ $\quad \hat{\alpha}_{ML} = \frac{4}{35}$

point-estimate

~~point estimator~~

S.O.C. $\frac{\partial^2 \ell}{\partial \alpha^2} = < 0$

Maximum likelihood estimation SRS

Important notes

- The likelihood function is a function of the parameter. It is defined up to positive constant factors. A likelihood function is not a probability density function. It contains all the information about the unknown parameter from the observations.

- The MLE is $\hat{\theta} = \arg\max_{\theta} L(\theta)$, i.e. $L(\hat{\theta}) = \arg\max_{\theta} L(\theta)$

- It is often more convenient to use the log-likelihood function denoted as

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log(f(X_i; \theta))$$

as it transforms product into a sum

Maximum likelihood estimation

Standard Normal Distribution

Assume that sample is made of first n terms of an IID sequence $\{X_n\}$ of normal random variables having mean μ and variance σ^2 and pdf of the following form:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

where μ and σ^2 are parameters to be estimated.

In that case likelihood function would be the following:

$$L(\mu, \sigma^2; x_1, x_2 \ldots x_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right]$$

Consistency

- Throughout, we assume that we have i.i.d. data.
- We let $\mathbf{X}_n = (X_1, \ldots, X_n)$, where the X_i's are i.i.d. with density $p(x;\theta_0) \in \mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$.
- We are interested in estimating $g(\theta_0)$, where $g(\cdot)$ is some function of θ_0.
- We focus on the large sample properties of the proposed estimators.

Consistency

- At the very least, we would like our estimator to be consistent.
- That is, as the sample size grows, we would like our estimator to get arbitrarily close to $g(\theta_0)$.

Let $T_n(\mathbf{X}_n)$ be a sequence of estimators. This sequence is said to be consistent for $g(\theta)$ if for all $\theta \in \Theta$ and all $\epsilon > 0$,

$$P_\theta[\|T_n(\mathbf{X}_n) - g(\theta)\| > \epsilon] \to 0$$

as $n \to \infty$.

Consistency

To prove consistency, choose $\epsilon > 0$. We know that

$$\begin{aligned} P_f[\hat{m}_n - m > \epsilon] &= P_f[\hat{m}_n > m + \epsilon] \\ &= P_f[\text{At least (n+1)/2 of the } X_i\text{'s exceeds } m + \epsilon] \end{aligned}$$

Let V_n denote the number of X_i's in a sample of size n that exceed $m + \epsilon$. So, $V_n \sim Binomial(n, \phi)$, where $\phi = P_f[X > m + \epsilon] < 0.5$.
So,

Markov's Inequality !

Chebyshev's Inequality !

$$\begin{aligned} P_f[\hat{m}_n - m > \epsilon] &= P[V_n \geq (n+1)/2] \\ &= P[V_n - n\phi \geq (n+1)/2 - n\phi] \\ &= P[V_n - n\phi \geq n(1/2 - \phi) + 1/2] \\ ! \quad &< P[V_n - n\phi \geq n(1/2 - \phi)] \\ &< \frac{\phi(1-\phi)}{n(1/2-\phi)^2} \xrightarrow[n\to\infty]{} 0 \end{aligned}$$

$\phi \in \mathbb{R}$

Consistency

- Most often, it is the case that

$$\sqrt{n}(T_n(\mathbf{X}_n) - g(\theta)) \xrightarrow{D} \text{Normal Distribution}$$

- We refer to this property as *asymptotic normality*.
- Also, the mean of the limiting normal distribution is usually zero. When this happens, we say that the sequence of estimators is "*asymptotically unbiased*".
- If two consistent (competitive) estimators are both asymptotically normal, then we can compare the resulting limiting normal distributions.
- The normal distribution which is "closest" to zero, on average, will be better asymptotically.

Consistency

- Let's return to the problem of estimating the center of a continuous, symmetric distribution. We want to compare the sample mean and sample median.
- By the central limit theorem for i.i.d. random variables, we know that $\sqrt{n}(\bar{\mathbf{X}}_n - \xi) \overset{D(f)}{\rightarrow} N(0, \sigma^2)$, where σ^2 is the variance associated with f.

Consistency

- So,

$$
\begin{aligned}
P_f[\hat{m}_n \leq \xi + a/\sqrt{n}] &= P_f[V_n \geq ((n+1)/2))] \\
&= P_f[\frac{V_n - n\phi_n}{(n\phi_n(1-\phi_n))^{1/2}} \geq \frac{((n+1)/2) - n\phi_n}{(n\phi_n(1-\phi_n))^{1/2}}
\end{aligned}
$$

- Let $H_n = \frac{V_n - n\phi_n}{(n\phi_n(1-\phi_n))^{1/2}}$ and $h_n = \frac{((n+1)/2) - n\phi_n}{(n\phi_n(1-\phi_n))^{1/2}}$.
- We want to evaluate $P_f[H_n \geq h_n]$.
- By the CLT for triangular arrays, we know that $P_f[H_n \geq u] \to \Phi(u)$, where $\Phi(\cdot)$ is the survivor function of a Normal(0,1) random variable.
- This implies that $|P_f[H_n \geq h_n] - \Phi(h_n)| \to 0$.

$\bar{X} = E(X)$
$\overline{X^2} = E(X^2)$
$\overline{X^3} = E(X^3)$

Method of moments

- **Advantage:** simplest approach for constructing an estimator

- **Disadvantage:** usually are not the "best" estimators possible

- **Principle:**

Equate the kth population moment $E[X^k]$ with the kth sample moment $\frac{1}{n}\sum_{n} X_i^k$ and solve for the unknown parameter

$u(x;\theta) = \frac{\partial \log p(x;\theta)}{\partial \theta}$

Proposition The Efficient Score Function has the following properties:

$$\begin{aligned} E[u(X;\theta_0) \mid \theta = \theta_0] &= 0. \\ Var[u(X;\theta_0) \mid \theta = \theta_0] &= E([u(X;\theta_0)]^2 \mid \theta = \theta_0) = I(\theta_0). \end{aligned}$$

$I(\theta)$ is the *Fisher information* about θ contained in X which satisfies the following identity

$$I(\theta_0) = Var[(u(X;\theta_0) \mid \theta_0] = E\left[-\frac{\partial^2 \log p(X \mid \theta_0)}{\partial \theta^2} \mid \theta_0\right]$$

Proof:

$$\begin{aligned} \frac{\partial}{\partial\theta}\int p(x \mid \theta)dx &= \frac{\partial}{\partial\theta}1 \\ \implies \int \frac{\partial p(x \mid \theta)}{\partial \theta}dx &= \frac{\partial}{\partial \theta}(1) = 0 \\ \implies \int [\frac{\partial p(x \mid \theta)}{\partial \theta}/p(x \mid \theta)]p(x \mid \theta)dx &= 0 \\ \implies \int [\frac{\partial log[p(x \mid \theta)]}{\partial \theta}p(x \mid \theta)dx &= 0 \\ \implies E[u(X;\theta) \mid \theta] &= 0 \end{aligned}$$

$$
\begin{aligned}
E[u(X;\theta) \mid \theta] &= 0 \\
\iff \int [\frac{\partial log[p(x \mid \theta)]}{\partial \theta} p(x \mid \theta) dx &= 0 \\
\frac{\partial}{\partial \theta} \left(\int [\frac{\partial log[p(x \mid \theta)]}{\partial \theta} p(x \mid \theta) dx \right) &= \frac{\partial}{\partial \theta}(0) \\
\int \left(\frac{\partial^2 log[p(x \mid \theta)]}{\partial \theta^2} p(x \mid \theta) + \frac{\partial log[p(x \mid \theta)]}{\partial \theta} (\frac{\partial p(x \mid \theta)}{\partial \theta}) \right) dx &= 0
\end{aligned}
$$

The last line can be written as:

$$
\int [\frac{\partial^2 log[p(x \mid \theta)]}{\partial \theta^2} p(x \mid \theta) dx] + \int [\frac{\partial log[p(x \mid \theta)]}{\partial \theta}]^2 p(x \mid \theta) dx = 0
$$

I.e.,

$$
E\left[\frac{\partial^2 log[p(x \mid \theta)]}{\partial \theta^2} \mid \theta\right] + E\left[\left(\frac{\partial log[p(x \mid \theta)]}{\partial \theta}\right)^2 \mid \theta\right] = 0
$$

So we have

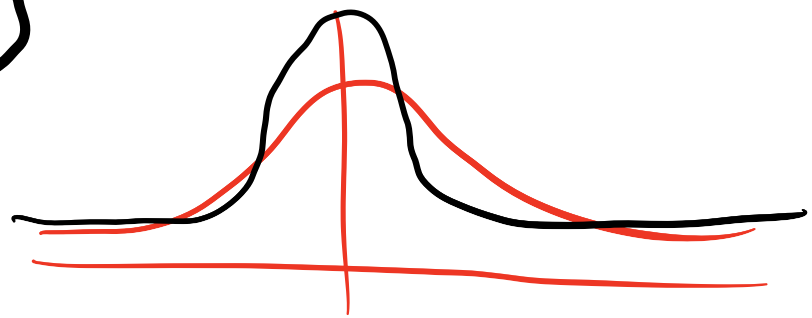
$$
\begin{aligned}
I(\theta) &= E[(u(X;\theta))^2 \mid \theta] = -E\left[\frac{\partial^2 log[p(x \mid \theta)]}{\partial \theta^2} \mid \theta\right] . \\
&= Var[u(X;\theta) \mid \theta]
\end{aligned}
$$

unbiased p.e. $\quad Var(T(X);\theta) \geq \dfrac{1}{I(\theta)}$

Theorem 3.4.1. Information Inequality

For a regular problem, let $T(X)$ be any statistic such that

$$E[T(X) \mid \theta] = \psi(\theta).$$
$$Var[T(X) \mid \theta] < \infty, \text{ for all } \theta.$$



Then for all θ:

- $Var[T(X) \mid \theta] \geq \boxed{\dfrac{[\psi'(\theta)]^2}{I(\theta)}}$,

($\psi(\theta)$ is differentiable and $I(\theta) =$ Fisher Information of P_θ).

Proof: By the conditions of a regular problem:

$$\begin{aligned}
\psi'(\theta) &= \frac{\partial}{\partial\theta}\left(\int T(x)p(x \mid \theta)dx\right) \\
&= \int \left(T(x)\frac{\partial}{\partial\theta}[p(x \mid \theta)]\right) dx \\
&= \int \left(T(x)\frac{\partial}{\partial\theta}[\log p(x \mid \theta)]p(x \mid \theta)\right) dx \\
&= E[T(X)U(X;\theta) \mid \theta] = Cov[T(X), U(X;\theta) \mid \theta]
\end{aligned}$$

(the last equation follows since $E[U(X;\theta) \mid \theta] = 0$)