Statistics

2022 Lectures
Part 8 - Point Estimation

Institute of Economic Studies
Faculty of Social Sciences
Charles University

Statistical inference and data

- What can we say about specific aspects of the stochastic mechanisms that govern the occurrence of our data?
- Whatever inference we make from the actual data, it is subject of error. This error (not mistake!) is the central concept of statistics.
- Using terminology of decision theory, we are in one of the situations labeled as "world θ_1" $\ldots$ "world θ_n" and we do not know at which one and assume that there is one action appropriate for each "world". To find out at which "world θ_j" we are, we conduct experiments.
- Let X be symbol for the results of such experiments and we assume that X follows distribution that depends on θ_j.
- Most often the outcome of X is possible under several θ_j's (with different probabilities depending on θ_j).
- From now on: data are observations collected via simple random sampling method.

Estimation

- Estimation is a process of extracting information about the value of a certain population parameter θ.
- Estimator of θ is then a rule that allows us to calculate an approximation of θ based on sample $X_1, \ldots, X_n$.
- There may be more then one estimator of the same parameter.
- We observe independent rv's $X_1, \ldots X_n$ sampled from distribution depending on θ which can have values from a parameter space Θ
- E.g. $X_1, \ldots, X_n$ are normally distributed with an unknown mean θ and known standard deviation.
- In simple scenarios θ is a single number, so Θ is a subset of the real line, but in general, Θ may be a set from a multidimensional space.

From now on, unless stated otherwise, θ is scalar; Θ is an interval of the real line.

Estimators

Definition 39: A statistics is called an estimator if it is used to estimate θ. The value of an estimator, obtained from a particular sample, is called an estimate of θ.

Example 82: Let $X_1, \ldots, X_n \sim U[0, \theta]$. Estimate θ!

- $T_1 = X_{n:n}$ ... the largest value should always satisfy $X_{n:n} \leq \theta$
- $T_2 = \frac{n+1}{n} T_1$ ... $X_1, \ldots X_n$ divides $[0, \theta]$ into $n + 1$ intervals of “even” length
- $T_3 = (n + 1) X_{1:n}$
- $T_4 = 2\bar{X}$ ... average should be close to midpoint
- and the list could go on

Desired properties of estimators

Definition 40: Let $T_n = T_n(X_1, \dots, X_n)$ be the estimator. Then it is called (weakly) consistent if $T_n \xrightarrow{P} \theta$. It is called strongly consistent if $T_n \xrightarrow{a.s.} \theta$.

- "one gets closer to the true value of the parameter by increasing sample size"

Definition 41: Let $T_n = T_n(X_1, \dots, X_n)$ be the estimator. Then it is called unbiased if $E_\theta(T_n) = \theta$ for every n. Otherwise it is called biased. The difference $B_\theta(T_n) = E_\theta(T_n) - \theta$ is called bias of T_n. It is called asymptotically unbiased if $\lim_{n\to\infty} E_\theta(T_n) = \theta$.

- "if we repeat sampling then the estimate is on average the true value"
- we can require many other properties, e.g. we may be interested in estimator with the lowest variance ("highest precision")

Example cont.

Example 82 cont.:

- $P(T_1 \leq t) = P(X_1 \leq t, \ldots, X_n \leq t) = \left(\frac{t}{\theta}\right)^n$ and thus
$$P(|T_1 - \theta| < \varepsilon) = 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n \to 1$$
and so T_1 is consistent.
- $T_2 = \frac{n+1}{n} T_1$ and so $T_2 \xrightarrow{P} \theta$. Also T_2 is consistent.
- $P(|T_3 - \theta| < \varepsilon) = P\left(X_{1:n} < \frac{\theta+\varepsilon}{n+1}\right) - P\left(X_{1:n} < \frac{\theta-\varepsilon}{n+1}\right) =$
$= \left(1 - \frac{\theta-\varepsilon}{\theta(n+1)}\right)^n - \left(1 - \frac{\theta+\varepsilon}{\theta(n+1)}\right)^n \to e^{-\frac{\theta-\varepsilon}{\theta}} - e^{-\frac{\theta+\varepsilon}{\theta}} < 1$
Thus T_3 is not consistent estimator of θ.
- By LLN $2\bar{X} \xrightarrow{P} 2E(X) = 2\frac{\theta}{2} = \theta$.

Mean square error

Definition 42: Let T be an estimator of θ. Then

$$MSE_\theta(T) = E_\theta[(T(X_1,\dots,X_n)-\theta)^2]$$

is called a mean squared error of T.

Theorem 57: For any estimator T of parameter θ we have

$$MSE_\theta(T) = Var_\theta T + (B_\theta(T))^2.$$

Example 82 cont.:

$E(T_1) = \frac{n}{n+1}\theta \neq \theta,\ E(T_2) = E(T_3) = E(T_4) = \theta$,

$Var(X_{1:n}) = Var(X_{n:n}) = \frac{n\theta^2}{(n+1)^2(n+2)}$.

$MSE_\theta(T_1) = \frac{2\theta^2}{(n+1)(n+2)}$

$MSE_\theta(T_2) = \frac{\theta^2}{n(n+2)}$

$MSE_\theta(T_3) = \frac{n\theta^2}{n+2}$

$MSE_\theta(T_4) = \frac{\theta^2}{3n}$

Why $T = \bar{X}_n$ as estimate of the mean?

Assume that the observations $X_1, X_2, \dots$ are iid random variables, with $E(X_i) = \theta$ and $Var(X_i) = \sigma^2 < \infty$, σ assumed known.

- T is an unbiased estimate. Moreover, it is consistent and $MSE_\theta(T) = \frac{\sigma^2}{n}$.
- $MSE_\theta(T)$ for a fixed n does not depend on the parameter θ.
- It decreases only in proportion to $\frac{1}{n}$! Decrease in proportion to $\frac{1}{n^2}$ is more desirable. However, the first case is much more common.
- T is linear!

Test of consistency

Test of consistency (based on Chebyschev inequality)

- check whether the estimator T is unbiased or not
- calculate $VarT$ and $B(T)$, the bias of T
- an unbiased estimator is consistent if $VarT \to 0$ as $n \to \infty$
- a biased estimator is consistent if both $VarT \to 0$ and $B(T) \to 0$ as $n \to \infty$

Example 83: Let $X_1, \ldots, X_n$ be a random sample with true mean μ and finite variance. Then, the sample mean $\bar{X}$ is a consistent estimator of the population mean μ.

Fisher information

Assume $X_1, X_2, \ldots$ with distribution such that the set of points at which $f(x,\theta) > 0$ does not depend on θ. Hence no single observation can rule out some values of θ. (This excludes the situation in the previous example!). We call this regularity assumption. What is the amount of information about θ in the event $\{X = x\}$?

Definition 43: Let X be a random variable with twice differentiable function $f(x,\theta)$ determining the distribution of X such that the set of x with $f(x,\theta) > 0$ is the same for all θ. Then the Fisher information about θ in a single observation X is defined by

$$I(\theta) = E_\theta[(J(X,\theta))^2], \text{ where } J(X,\theta) = \frac{\partial}{\partial\theta} \log f(X,\theta)$$

provided the expectation exists.

Examples

Example 84: $X \sim N(\theta, \sigma^2)$

$$J(X,\theta) = \frac{\partial}{\partial\theta}\left(-\log\sigma\sqrt{2\pi} - \frac{(X-\theta)^2}{2\sigma^2}\right) = \frac{X-\theta}{\sigma^2}$$

$$I(\theta) = \frac{1}{\sigma^4}E[(X-\theta)^2] = \frac{1}{\sigma^2}.$$

Example 85: Bernoulli trial with probability θ
$P(X=1|\theta)=\theta,\ P(X=0|\theta)=1-\theta$
$f(x,\theta)=\theta^x(1-\theta)^{1-x},\ x=0,1$

$$J(X,\theta) = \begin{cases} -\frac{1}{1-\theta} & X=0; \\ \frac{1}{\theta} & X=1; \end{cases} \quad I(\theta) = \frac{1}{\theta(1-\theta)}.$$

So $I(\theta)$ has minimal value 4 at $\theta=\frac{1}{2}$ and for θ approaching 0 or 1 the information goes to infinity.

Properties of Fisher information

Theorem 58: Under regularity assumptions,

a) $E_\theta(J(X,\theta)) = 0$;

b) $Var_\theta J(X,\theta) = I(\theta)$;

c) $I(\theta) = -E_\theta\left(\frac{\partial}{\partial\theta} J(X,\theta)\right)$;

d) The information $I_n(\theta)$ in a random sample of n observations is

$$I_n(\theta) = nI(\theta).$$

Theorem 59: (Rao - Cramér)
Under regularity assumptions, for any unbiased estimator T_n of a parametric function $m(\theta)$ we have

$$Var_\theta T_n \geq \frac{(m'(\theta))^2}{nI(\theta)}.$$

For $m(\theta) = \theta$, i.e. T_n unbiased estimator of θ, $Var_\theta T_n \geq \frac{1}{nI(\theta)}$.

Efficiency of estimators

Definition 44: Any unbiased estimator T that satisfies the regularity assumption and whose variance attains the Rao-Cramér bound is called efficient. The ratio

$$\frac{\left(\frac{1}{nI(\theta)}\right)}{Var_\theta T_n} \leq 1$$

is called efficiency.
If $\tilde{T}_n$ and $\hat{T}_m$ are two unbiased estimators of θ,

$$\frac{Var_\theta \tilde{T}_n}{Var_\theta \hat{T}_m}$$

is called relative efficiency of $\hat{T}_m$ with respect to $\tilde{T}_n$.

Method of Moments Estimators

- a method of estimation of population parameters such as mean, variance, median, etc. (which need not be moments), by equating sample moments with unobservable population moments
- estimator is then found as a solution of the resulting equation with respect to the (unknown) parameter
- suitable for estimating also several parameters at once
- advantages: quickly and easily computable by hand
- disadvantages: non-unique based on chosen equations
- simple rule: take moments of the lowest orders

Method of Moments Estimators

Example 86: $X_1, \ldots, X_n \sim EXP(\theta)$
then $E(X_i) = \frac{1}{\theta}$ and since its sample counterpart is $\bar{X}_n$ then logically the estimator is

$$T_1 = \frac{1}{\bar{X}_n}.$$

or $E(X_i^2) = \frac{2}{\theta^2}$ and since its sample counterpart is $\frac{1}{n}\sum_{i=1}^n X_i^2$ then the estimator can also be

$$T_2 = \sqrt{\frac{2n}{\sum_{i=1}^n X_i^2}}.$$

If we want to estimate, e.g., $p = P(X \geq 3) = e^{-3\theta}$ we can use either $p_1 = \exp\left\{-\frac{3}{\bar{X}_n}\right\}$ or $p_2 = \exp\left\{-3 \cdot \sqrt{\frac{2n}{\sum X_i^2}}\right\}$.

Method of Moments Estimators

Example 87: In general, consider $\theta = (\mu, \sigma^2)^\top$ (both paramaters are unknown). Let $X_1, \ldots, X_n$ be random sample with $E(X) = \mu$ and $VarX = \sigma^2$.

Since $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$ then we have to solve

$$\begin{aligned} \bar{X} &= \hat{\mu} \\ \frac{1}{n} \sum X_i^2 &= \hat{\sigma}^2 + \hat{\mu}^2 \end{aligned}$$

which leads to $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

- moment estimators are consistent under mild assumptions: if parameter θ is a continuous function of moments (usually it is so)
- often they coincide with maximum likelihood estimator or they are inferior to them

Maximum Likelihood Estimators

If $X_1 = x_1, \ldots, X_n = x_n$ then $f(x_1, \theta) \cdots f(x_n, \theta)$, the probability of this sample or joint density of the random sample, regarded as a function of θ, can be understood as the likelihood function of the sample:

$$L(\theta, x) = \prod_{i=1}^{n} f(x_i, \theta).$$

Definition 45: Given the sample $x = (x_1, \ldots, x_n)$, the value $\hat{\theta}(x)$ maximizing $L(\theta, x)$, is called the maximum likelihood estimate (MLE) of θ.

- We often define $\ell(\theta, x) = \log L(\theta, x)$ and maximize $\ell(\theta, x)$ instead $L(\theta, x)$.

Maximum Likelihood Estimators

Example 88: Suppose we observed three successes and two failures in five Bernoulli trials with probability of success θ. We have $f(x,\theta) = \theta^x(1-\theta)^{1-x}$ where $x = 0$ represents failure and $x = 1$ success

$L(\theta, x) = \theta^3(1-\theta)^2$ for $0 \leq \theta \leq 1$ and L attains maximum at $\hat{\theta} = 3/5$.

Example 89: Suppose two observations $x_1 = 3, x_2 = -2$ from a $N(0, \theta^2)$ distribution

$\ell(\theta, x) = -\log 2\pi - 2\log\theta - \frac{13}{2\theta^2}$ and $\hat{\theta}_{MLE} = \sqrt{\frac{13}{2}}$.

Maximum Likelihood Estimators

- likelihood function can be regarded as random function (randomness from the sample)
- Invariance principle: If $\hat{\theta}$ is the MLE of parameter θ, then $h(\hat{\theta})$ is the MLE of parameter $h(\theta)$.
- application: if $\hat{\theta}$ is the MLE of the variance σ^2 then $\sqrt{\hat{\theta}}$ is the MLE of the standard deviation σ.
- Likelihood principle: Consider two sets of data sampled from the same population. If $L_1(\theta, x)/L_2(\theta, y)$ does not depend on θ, then both data sets contain the same information about θ and should provide the same estimate of θ.

Least Squares Estimators

- now a slightly different setup: data are independent but possibly not from the same distribution
- we observe Y and U and assume that

$$Y = Q(u) + \varepsilon$$

  called regression of Y on u, where u are the observations of U, $Q(u)$ is some function of u and ε is a random “error” such that $E(\varepsilon) = 0, Var\varepsilon = \sigma^2$.
- sometimes we can have several observations of Y for a given value u, in some cases we have exactly one
- function Q is usually called the regression of Y on u
- in general $Q(u) = \varphi(u, \theta_1, \ldots, \theta_r)$, often linear in θ, e.g. $Q(u) = \alpha + \beta u$. In such a case we speak of linear regression model

Least Squares Estimators

- the method of least squares is based on finding values $\hat{\theta}_1, \ldots, \hat{\theta}_n$, called least squares estimate (LSE), minimizing

$$S(\theta_1, \ldots, \theta_r) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \varphi(u_i, \theta_1, \ldots, \theta_r))^2,$$

  where y_{ij} is the jth observation of Y for the value u_i of U.
- usual way to find LS-estimators of $\theta_1, \ldots, \theta_r$ is by solving the set of so called normal equations

$$\frac{\partial S}{\partial \theta_i} = 0, \;\; i = 1, \ldots, r.$$