Point Estimators

Rosario Barone

Tor Vergata University of Rome

Statistical tools for decision making

Undergraduate Degree in Global Governance

A.Y. 2023/2024

Three famous frequentist approaches for point estimation

- **Method of Moments**
- Least Squares Estimation
- Maximum Likelihood Estimation

Method of Moments

- The method of moments starts by equating the sample moments (e.g., mean, variance) with their corresponding population moments.
- We derive equations by setting the sample moments equal to the corresponding theoretical moments.
- The resulting system of equations can be solved to estimate the parameters of the distribution.
- The number of equations needed depends on the number of parameters to be estimated.

Gaussian model: Mean Estimator

- Let's consider a random sample $X_1, X_2, \ldots, X_n$ from a Gaussian distribution with mean μ and variance σ^2.
- The first population moment is the mean, $\mu_1 = E(X) = \mu$.
- The first sample moment is the sample mean, $m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Equating the population and sample moments, we have $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$, which gives us the mean estimator $\hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Derivation of Variance Estimator

Definition of the r-th moment

- if X is discrete $E(X^r) = \sum_i p(x_i) x_i^r$
- if X is continuous in A $E(X^r) = \int_A f(x) x^r dx$

- The second population moment is $\mu_2 = E(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.
- We know that: $Var(X) = E(X^2) - (E(X))^2$
- Equating the population and sample moments, we have $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}(X_i)\right)^2$.
- Simplifying the equation, we obtain the variance estimator $\hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Advantages and Limitations

- Advantages of the method of moments include its simplicity, ease of implementation, and intuitive interpretation
- Is fairly simple and yields consistent estimators (under very weak assumptions)
- It may not always yield the most efficient estimators, especially in small samples or complex models
- It often yelds biased estimators
- the suitability of the method depends on the specific problem and the available data

Likelihood Function

- The likelihood function is a function of the parameters of a statistical model, given the observed data.
- It measures the likelihood or plausibility of the observed data for different parameter values.
- For a random sample of independent and identically distributed (i.i.d.) observations, the likelihood function is the product of the probability density function (pdf) or probability mass function (pmf) for each observation.

Likelihood Function

Likelihood Function

The likelihood function, denoted as $\mathcal{L}(\theta)$, measures the probability of observing the data given the parameter values θ.

- For independent and identically distributed (i.i.d.) observations of a statistical model with density function $f(\cdot;\theta)$, the likelihood function is the product of the individual densities.
- If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, the likelihood function is given by:

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

Properties of the Likelihood Function

- The likelihood function possesses several important properties:
  - **Non-negativity**: The likelihood function is non-negative for all parameter values.
  - **Monotonicity**: As the parameter values move away from the true parameter values, the likelihood function decreases.
  - **Scale invariance**: Multiplying the likelihood function by a constant factor does not change the relative likelihoods of different parameter values.
  - **Likelihood principle**: The likelihood function contains all the information about the unknown parameters that is available in the data.

Log-Likelihood Function

Log-Likelihood Function

The log-likelihood function, denoted as $\ell(\theta)$, is the natural logarithm of the likelihood function. It is often easier to work with than the likelihood function itself.

- Taking the logarithm helps simplify calculations and does not change the location of the maximum point.
- The log-likelihood function is given by:

$$\ell(\theta) = \sum_{i=1}^{n} \ln(f(X_i; \theta))$$

Score Function

- The score function measures the sensitivity of the log-likelihood function with respect to the parameters of interest.
- It provides information about the direction and magnitude of the parameter effects.
- Mathematically, the score function is defined as:

$$S(\theta) = \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta|\mathbf{x})$$

  where $\mathcal{L}(\theta|\mathbf{x})$ is the likelihood function, θ is the parameter of interest, and $\mathbf{x}$ is the observed data.

Fisher Information

- The Fisher information quantifies the amount of information provided by the data about the parameters of interest.
- It measures the precision or uncertainty of the parameter estimates.
- The Fisher information matrix is defined as:

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \ln \mathcal{L}(\theta|\mathbf{x})\right]$$

where $\mathbb{E}$ denotes the expectation operator.

Point Estimators

Three famous frequentist approaches:

- Method of Moments
- Least Squares Estimation
- **Maximum Likelihood Estimation**

The Maximum Likelihood Estimator

- Maximum Likelihood Estimation (MLE) is a method used to estimate the parameters of a statistical model based on observed data.
- It involves finding the parameter values that maximize the likelihood function, which measures the probability of observing the data given the parameter values.
- MLE is widely used in various fields, including statistics, econometrics, and machine learning.

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation involves finding the parameter values that maximize the likelihood function (or equivalently, the log-likelihood function).

- The MLE estimates, denoted as $\hat{\theta}_{\text{MLE}}$, are obtained by solving the equation $\frac{\partial \ell(\theta)}{\partial \theta} = 0$.
- In some cases, it may be easier to maximize the log-likelihood function numerically using optimization algorithms.

Example: MLE for Bernoulli Distribution

Step 1: Likelihood Function

Consider a Bernoulli distribution with parameter p. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with the following probability mass function:

$$P(X_i = x_i) = \begin{cases} p & \text{if } x_i = 1 \\ 1 - p & \text{if } x_i = 0 \end{cases}$$

The likelihood function, denoted by $\mathcal{L}(p)$, can be expressed as the joint probability mass function of the observations:

$$\mathcal{L}(p) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p^{\sum_{i=1}^{n} x_i}(1 - p)^{n - \sum_{i=1}^{n} x_i}$$

Example: MLE for Bernoulli Distribution

Step 2: Log-Likelihood Function

To simplify the calculations, we take the logarithm of the likelihood function to obtain the log-likelihood function, denoted by $\ell(p)$:

$$\ell(p) = \log \mathcal{L}(p) = \sum_{i=1}^{n} x_i \log(p) + (n - \sum_{i=1}^{n} x_i) \log(1 - p)$$

Example: MLE for Bernoulli Distribution

Step 3: Maximizing the Log-Likelihood

$$\frac{d\ell(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{p} = \frac{n - \sum_{i=1}^{n} x_i}{1-p}$$

$$\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i = pn - p\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i = pn$$

$$\hat{p}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example: MLE for Normal Distribution

Step 1: Likelihood Function

Consider a normal distribution with unknown mean μ and variance σ^2. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables following a normal distribution. The likelihood function, denoted by $\mathcal{L}(\mu, \sigma^2)$, can be expressed as the joint density function of the observations:

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}}$$

Step 2: Log-Likelihood Function

To simplify the calculations, we take the logarithm of the likelihood function to obtain the log-likelihood function, denoted by $\ell(\mu, \sigma^2)$:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(X_i - \mu)^2}{2\sigma^2} \right)$$

Example: MLE for Normal Distribution

Step 3: Maximizing the Log-Likelihood

To find the maximum likelihood estimator, we differentiate the log-likelihood function with respect to μ and σ^2 and set them equal to zero:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{\sum_{i=1}^{n} (X_i - \mu)}{\sigma^2} = 0$$

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} = 0$$

Step 4: Solving for Maximum Likelihood Estimators

Solving the equations, we obtain the maximum likelihood estimators:

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu}_{MLE})^2$$

Asymptotic properties of MLE

- The MLE possesses several desirable properties:
  - Consistency: The MLE converges to the true parameter value as the sample size increases.
  - Asymptotic Normality: The MLE is asymptotically normally distributed.
  - Efficiency: The MLE achieves the Cramér-Rao lower bound for the variance of an unbiased estimator.

Consistency of the MLE

- Let $\hat{\theta}_{\text{MLE}}$ be the Maximum Likelihood Estimator for parameter θ.
- We say that $\hat{\theta}_{\text{MLE}}$ is consistent if:

$$\lim_{n \to \infty} \Pr(|\hat{\theta}_{\text{MLE}} - \theta| > \epsilon) = 0 \quad \text{for all } \epsilon > 0.$$

- In simpler terms, the probability that $\hat{\theta}_{\text{MLE}}$ deviates from θ by more than ϵ approaches 0 as the sample size n increases.

Asymptotic normality of the MLE

- Let $\hat{\theta}_{\text{MLE}}$ be the Maximum Likelihood Estimator for the parameter θ.
- It can be proved that $\hat{\theta}_{\text{MLE}}$ is asymptotically normal. Namely, as the sample size n approaches infinity:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} \mathcal{N}(\theta, \mathcal{I}(\theta)^{-1})$$

  where $\xrightarrow{d}$ denotes convergence in distribution and $\mathcal{I}(\theta)$ is the Fisher Information matrix.

Note that we can calculate the standard error of the MLE as $\mathcal{I}(\theta)^{-1}$

Example: Asymptotic normality of the Bernulli MLE

The log-likelihood function for a sample of n independent and identically distributed (i.i.d.) observations $X_1, X_2, \ldots, X_n$ from the Bernoulli distribution is:

$$\ell(p) = \log \prod_{i=1}^{n} f(X_i; p) = \sum_{i=1}^{n} \log f(X_i; p)$$

Taking the derivative of the log-likelihood function with respect to p, we get:

$$\frac{d\ell(p)}{dp} = \frac{1}{p} \sum_{i=1}^{n} X_i - \frac{1}{1-p} \sum_{i=1}^{n} (1 - X_i)$$

Taking the derivative of the score function we get:

$\frac{d^2\ell(p)}{dp^2} = -\frac{1}{p^2} \sum_{i=1}^{n} X_i - \frac{1}{(1-p)^2} \sum_{i=1}^{n} (1 - X_i)$
Since X_i follows a Bernoulli distribution with parameter p, we have:

$$\mathbb{E}[X_i] = p \quad \text{and} \quad \mathbb{E}[1 - X_i] = 1 - p$$

Substituting these values, we get:

$$\mathbb{E}\left[\frac{d^2\ell(p)}{dp^2}\right] = -\frac{1}{p^2}\sum_{i=1}^{n} p - \frac{1}{(1-p)^2}\sum_{i=1}^{n}(1-p)$$

Simplifying, we have:

$$\mathbb{E}\left[\frac{d^2\ell(p)}{dp^2}\right] = -\frac{n}{p} - \frac{n}{1-p}$$

Finally, the Fisher Information Matrix is the negative expected value of the second derivative of the log-likelihood function:

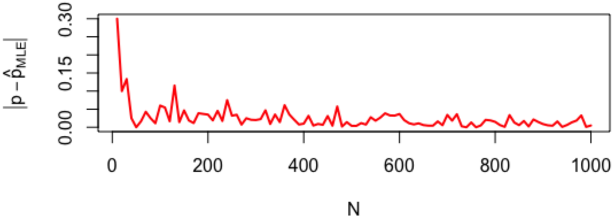
$$\mathcal{I}(p) = -\mathbb{E}\left[\frac{d^2\ell(p)}{dp^2}\right] = \frac{n}{p(1-p)}.$$

Therefore,

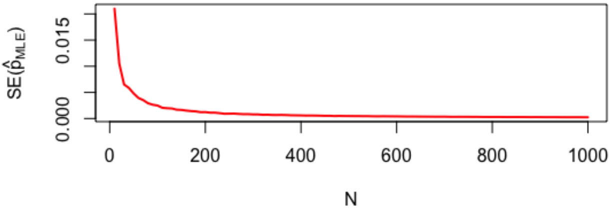
$$\hat{p}_{MLE} \xrightarrow{d} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Consistency: Proof in R

```
library(latex2exp)
N <- seq(10,1000,length.out=100)
p0 <- 0.4

# Generate many random samples of increasing size
and compute MLE.
samples <- lapply(1:100,function(j) rbinom(N[j],
size = 1, prob = p0) )

mles<-sapply(1:100, function(j)
mean(unlist(samples[j])))
# Compute the standard error.
se<-(mles*(1-mles))*(N^(-1))

# Plot the standard error values olf the MLEs as
the sample size increases.
plot(N,abs(mles-p0), type = "l", lwd=2, col="red",
xlab="N", ylab = TeX("$|p-\\hat{p}_{MLE}|$"))
plot(N,se, type = "l", lwd=2, col="red",xlab="N",
ylab =  TeX("SE($\\hat{p}_{MLE}$)"))
```



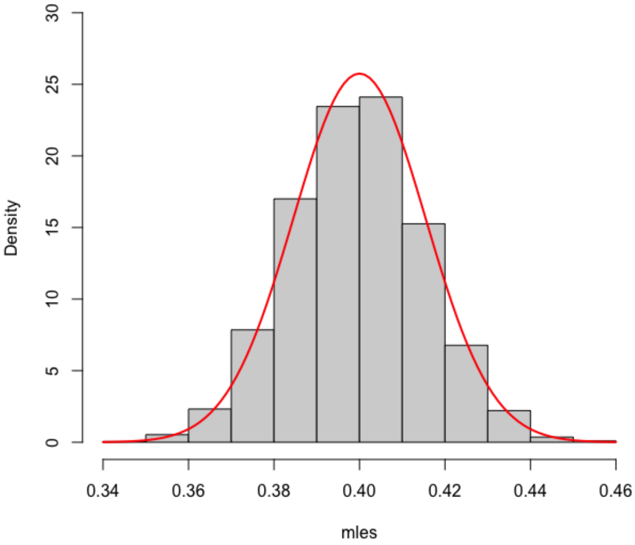
$|p - \hat{p}_{MLE}|$
0.00
0.15
0.30
0
200
400
600
800
1000
N



$SE(\hat{p}_{MLE})$
0.000
0.015
0
200
400
600
800
1000
N

AsymptoticNormality: Proof in R

```
# Plot the asymptotically normal distribution.
N <- 1000
p0 <- 0.4

# Generate many random samples of size N and
compute MLE.
mles <- replicate(10000, mean(rbinom(N, size = 1,
prob = p0)))

# Plot histogram of MLEs.
hist(mles, freq = FALSE, ylim=c(0,30))
curve(dnorm(x, mean = p0, sd = sqrt((p0 * (1 - p0)) / N)),
 add=TRUE, col="red", lwd=2)
```



Density
mles
0
5
10
15
20
25
30
0.34
0.36
0.38
0.40
0.42
0.44
0.46

Efficiency of the MLE

- The CRLB gives a lower bound on the variance of any unbiased estimator.
- By asymptotic Normality of the MLE:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} \mathcal{N}(\theta, \mathcal{I}(\theta)^{-1})$$

we conclude that the variance of the MLE is equal to the CRLB, proving its efficiency.

Summary on the MLE

- Maximum Likelihood Estimation is a powerful method for estimating parameters in statistical models.
- It involves maximizing the likelihood function (or log-likelihood function) to obtain parameter estimates.
- MLE estimators possess desirable properties, such as consistency and asymptotic normality.
- These properties make the MLE a reliable and powerful tool for parameter estimation.