

Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 19 - Maximum likelihood estimation

Salvatore Ruggieri

Department of Computer Science

University of Pisa, Italy

salvatore.ruggieri@unipi.it

Example: number of German tanks



- Tanks' ID drawn at random without replacement from $1, \dots, N$. Objective: estimate N .

Example: number of German tanks

- Let x_1, \dots, x_n be the observed ID's
- E.g., 61, 19, 56, 24, 16 with $n = 5$
- They are realizations of X_1, \dots, X_n draws without replacement from $1, \dots, N$
 - ▶ X_1, \dots, X_n is **not a random sample**, as they are not independent!
 - ▶ The marginal distribution is $X_i \sim U(1, N)$ [prove it, or see Sect. 9.3 of [T]]

- **Estimator based on the mean**

- ▶ Since:

$$E[\bar{X}_n] = E[X_i] = \frac{N+1}{2}$$

we can define an estimator:

$$T_1 = 2\bar{X}_n - 1$$

- ▶ T_1 is unbiased:

$$E[T_1] = 2E[\bar{X}_n] - 1 = N$$

- ▶ E.g., $t_1 = 2(61 + 19 + 56 + 24 + 16)/5 - 1 = 69.4$

Example: number of German tanks

- Let x_1, \dots, x_n be the observed ID's
- E.g., 61, 19, 56, 24, 16 with $n = 5$
- **Estimator based on the maximum**
 - ▶ Let $M_n = \max\{X_1, \dots, X_n\}$
 - ▶ Since:

[see Sect. 20.1 of [T]]

$$E[M_n] = n \frac{N+1}{n+1}$$

we can define an estimator:

$$T_2 = \frac{n+1}{n} M_n - 1$$

- ▶ T_2 is also unbiased:

$$E[T_2] = \frac{n+1}{n} E[M_n] - 1 = N$$

- ▶ E.g., $t_2 = 6/5 \max\{61, 19, 56, 24, 16\} - 1 = 72.2$

See R script

Estimators

- So far, estimators were derived from parameter definition through the plug-in method
- A general principle to derive estimators will be shown today
- Example

Table 21.1. Observed numbers of cycles up to pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

- Assume that the data is generated from geometric distributions:

$$P(X_i = k) = (1 - p)^{k-1}p$$

where p is distinct for smokers and non smokers.

- What is an estimator for p ?

[parametric inference]

- ▶ E.g., since $p = P(X_i = 1)$, we could use $S = \frac{|\{i \mid X_i=1\}|}{n}$, and show $E[S] = p$
- ▶ $p = 29/100$ for smokers, and $p = 198/486 = 0.41$ for non-smokers
- ▶ But we did not use all of the available data!

The maximum likelihood principle

The maximum likelihood principle

Given a dataset, choose the parameter(s) of interest in such a way that the data are most likely.

Table 21.1. Observed numbers of cycles up to pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

- For $k = 1, \dots, 12$, $P(X_i = k) = (1 - p)^{k-1}p$. Moreover, $P(X_i > 12) = (1 - p)^{12}$
- Since the X_i 's are independent, we can write the probability of observing the smokers as:
$$L(p) = C \cdot P(X_i = 1)^{29} \cdot P(X_i = 2)^{16} \cdot \dots \cdot P(X_i = 12)^3 \cdot P(X_i > 12)^7 = Cp^{93}(1 - p)^{322}$$
 - ▶ C is the number of ways we can assign 29 ones, 16 twos, \dots , 3 twelves, and 7 numbers larger than 12 to 100 smokers
- ML principle: choose $\hat{p} = \arg \max_p L(p)$

Example

- ML principle: choose $\hat{p} = \arg \max_p L(p) = \arg \max_p Cp^{93}(1-p)^{322}$
- $L'(p) = C(93p^{92}(1-p)^{322} - 322p^{93}(1-p)^{321}) = Cp^{92}(1-p)^{321}(93 - 415p)$
- $L'(p) = 0$ for $p = 0$ or $p = 1$ or $p = 93/415 = 0.224$
- ML estimate is $\arg \max_p L(p) = 0.224 < 0.41$ (estimate using S)
- Equivalent formulation for maximization:

$$\arg \max_p L(p) = \arg \max_p \log L(p)$$

- $\log L(p) = \log C + 93 \log p + 322 \log (1 - p)$
- $\log' L(p) = \frac{93}{p} - \frac{322}{1-p}$
- $\log' L(p) = 0$ for $322p = 93(1 - p)$, i.e., $p = 93/(322 + 93) = 0.224$

See R script

Likelihood and log-likelihood

Likelihood and log-likelihood functions

Let x_1, \dots, x_n be a dataset, i.e., realizations of a random sample X_1, \dots, X_n where the density/p.m.f of X_i 's is $f_\theta(\cdot)$, parametric on θ . The likelihood function is:

$$L(\theta) = \prod_{i=1}^n f_\theta(x_i)$$

and the log-likelihood function is:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$$

MAXIMUM LIKELIHOOD ESTIMATES. The *maximum likelihood estimate* of θ is the value $t = h(x_1, x_2, \dots, x_n)$ that maximizes the likelihood function $L(\theta)$. The corresponding random variable

$$T = h(X_1, X_2, \dots, X_n)$$

is called the *maximum likelihood estimator* for θ .

Example: MLE of exponential distribution

- Random sample of $Exp(\lambda)$
- Since $f_\lambda(x) = \lambda e^{-\lambda x}$ for $x \geq 0$:

$$E[X] = 1/\lambda$$

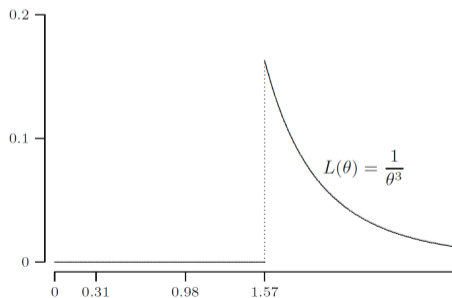
$$\ell(\lambda) = \sum_{i=1}^n (\log \lambda - \lambda x_i) = n \log \lambda - \lambda(x_1 + \dots + x_n) = n(\log \lambda - \lambda \bar{x}_n)$$

- $\ell'(\lambda) = 0$ iff $n(1/\lambda - \bar{x}_n) = 0$ iff $\lambda = 1/\bar{x}_n$
- $T = 1/\bar{X}_n$ is the MLE of λ for a $Exp(\lambda)$ -distributed random sample
- It is biased!: $E[T] \geq 1/E[\bar{X}_n] = \lambda$ *[Jensen's inequality]*
- **Exercise at home**
 - ▶ show that \bar{X}_n is an unbiased MLE of θ for a $Exp(1/\theta)$ -distributed random sample

Example: upper point of a uniform distribution

- Dataset: $x_1 = 0.98, x_2 = 1.57, x_3 = 0.31$ from $U(0, \theta)$ for unknown $\theta > 0$
- $f_\theta(x) = 1/\theta$ for $0 \leq x \leq \theta$ and $f_\theta(x) = 0$ otherwise

$$L(\theta) = f_\theta(x_1)f_\theta(x_2)f_\theta(x_3) = \begin{cases} \frac{1}{\theta^3} & \text{if } \theta \geq \max\{x_1, x_2, x_3\} = 1.57 \\ 0 & \text{otherwise} \end{cases}$$



- In general, MLE estimator is $\max\{X_1, \dots, X_n\}$

Example: MLE of normal distribution

- Random sample of $N(\mu, \sigma^2)$
- MLE of $\theta = (\mu, \sigma^2)$ where $f_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ [we work on σ^2 , not on σ]

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

- Partial derivatives:

$$\frac{d}{d\mu} \ell(\mu, \sigma) = \frac{n}{\sigma^2} (\bar{x}_n - \mu) \qquad \frac{d}{d\sigma^2} \ell(\mu, \sigma) = \frac{1}{2\sigma^2} \left(\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - n \right)$$

- Partial derivatives at 0 for $\mu = \bar{x}_n$ and $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ **[prove it is a maximum]**
- MLE estimators $\hat{\mu} = \bar{X}_n$ (*unbiased*) and $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ (*biased*)

See R script

Loss functions (to be minimized)

- Negative log-likelihood (nLL)

$$nLL(\theta) = -\ell(\theta)$$

- How to compare estimators that use more parameters?

- ▶ T_1 assuming a $Ber(p)$ vs T_2 assuming $Bin(n, p)$
- ▶ Neural network with 10 nodes vs with 100 nodes

- Akaike information criterion (AIC), balances model fit against model simplicity

$$AIC(\theta) = 2|\theta| - 2\ell(\theta)$$

- Bayesian information criterion (BIC), stronger balances over model simplicity

$$BIC(\theta) = |\theta| \log n - 2\ell(\theta)$$

See R script

Properties of MLE estimators

- MLE estimators can be biased, but under mild assumptions, they are asymptotically unbiased! *[Asymptotic unbiasedness]*

$$\lim_{n \rightarrow \infty} E[T_n] = \theta$$

- If T is the MLE estimator of θ and $g(\cdot)$ is an invertible function, then $g(T)$ is the MLE estimator of $g(\theta)$ *[Invariance principle]*

- ▶ E.g., MLE of σ for normal data is $\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2}$
- ▶ but, $E[T] = \theta$ does **NOT** necessarily imply $E[g(T)] = g(\theta)$
- ▶ See also Exercise at home

- Under mild assumptions, MLE estimators have asymptotically the smallest variance among unbiased estimators *[Asymptotic minimum variance]*

Score function and Fisher information

- Consider a density function $f_\theta(x)$

Score function and Fisher information

The *score function* is the random variable:

$$S(\theta) = \frac{\partial}{\partial \theta} \ell(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_\theta(X_i)$$

The **Fisher information** is the variance of it:

$$I(\theta) = \text{Var}(S(\theta))$$

- $I(\theta)$ quantifies the sensitivity of X w.r.t. θ : if small changes in θ result in large changes in the density values (high variance of $I(\theta)$), then data easily provides information on the correct θ .
 - Recall that $H(X) = E[-\log f(X)]$ is the entropy of X [see Lesson 09]
 - See **this introduction**.
- For $N(\mu, \sigma^2)$, we calculated: $S(\mu) = \frac{d}{d\mu} \ell(\mu, \sigma) = \frac{n}{\sigma^2} (\bar{X}_n - \mu)$. Hence:

$$I(\mu) = \text{Var}(S(\mu)) = \frac{n^2}{\sigma^4} \frac{\sigma^2}{n} = \frac{n}{\sigma^2}$$

Minimum Variance Unbiased Estimators (MVUE)

Score function and Fisher information

The *score function* is the random variable:

$$S(\theta) = \frac{\partial}{\partial \theta} \ell(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_{\theta}(X_i)$$

The **Fisher information** is the variance of it:

$$I(\theta) = \text{Var}(S(\theta))$$

- Since $E[S(\theta)] = 0$, $I(\theta) = E[S(\theta)^2]$ [prove it or see *sdsln.pdf* Chpt. 1]
- Since X_i 's are i.i.d, $I(\theta) = E[S(\theta)^2] = nE[(\frac{\partial}{\partial \theta} \log f_{\theta}(X))^2]$ [prove it or see *sdsln.pdf* Chpt. 1]
- **Cramér-Rao's bound** for unbiased estimator T (under some assumptions):

$$\text{Var}(T) \geq \frac{1}{I(\theta)}$$

- An unbiased estimator T such that $\text{Var}(T) = 1/I(\theta)$ is a *MVUE*

Example

- Normal distribution and μ parameter: $f_{\mu}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- Unbiased MLE estimator of μ is $T = \bar{X}_n = (X_1 + \dots + X_n)/n$.
- The Fisher information is:

$$I(\mu) = \frac{n}{\sigma^2} = \frac{1}{\text{Var}(\bar{X}_n)}$$

where the last equality follows because for i.i.d. random variables $\text{Var}(\bar{X}_n) = \sigma^2/n$.

- By taking the reciprocals: $\text{Var}(\bar{X}_n) = 1/I(\mu)$
- Hence \bar{X}_n is a MVUE of μ

Fisher information and MLE standard error

- The standard deviation of the sampling distribution is called the *standard error* (*se*)
- An MLE estimator T is asymptotically unbiased
- An MLE estimator T has asymptotic minimum variance
- By Cramér-Rao's bound, asymptotically we have:

$$se(T) = \sqrt{\text{Var}(T)} = \frac{1}{\sqrt{I(\theta)}}$$

- E.g., for the normal distribution and the MLE estimator $\hat{\mu}$ of μ :

$$se(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

but because σ is unknown, we plug-in its estimate $\hat{\sigma}$

$$se(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

See R script