Methods of estimation

Probability theory. Statistics

Anton Afanasev

Higher School of Economics

DSBA 211
December 3, 2022

Seminar Overview

Quiz

Two measurements of the side of the square were produced. Suppose the two measurements X_1 and X_2 are two independent random variables with mean a and variance σ^2. The true length of the side of the square is a. Find MSE for the following estimator of the area of the square: $X_1 X_2$.

Moments of a random variable

- Moment of a random variable – quantitative measure, which describes the shape of p.d.f.'s graph independently of translation.
- k^{th} raw moment of a random variable X:

$$\mu_k = \mathsf{E}\left(X^k\right).$$

- k^{th} central moment of a random variable X:

$$\widetilde{\mu}_k = \mathsf{E}\left((X - \mathsf{E}(X))^k\right).$$

Example

- $\mu_1 = \mathsf{E}(X)$ – expected value,
- $\widetilde{\mu}_2 = \mathsf{V}(X)$ – variance.

- If the p.d.f. of X represents density, then $\mathsf{E}(X)$ is the center of mass, and $\mathsf{V}(X)$ is the rotational inertia.

Problem 1

Come up with any linear p.d.f. (non-uniform)

1. Can you conclude which parameter is greater without calculations – mean or median?
2. Calculate them explicitly.
3. Discuss the result. Why p.d.f. is unbalanced in terms of “mass” around its median?

Higher-order moments

- Skewness shows the asymmetry of a p.d.f.:

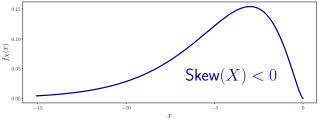
$$\mathsf{Skew}(X) = \frac{\widetilde{\mu}_3}{\sigma^3}.$$

- Excess kurtosis shows the sharpness of a p.d.f. peak:

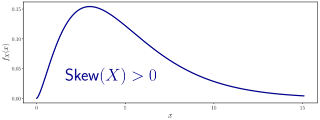
$$\mathsf{Kurt}(X) = \frac{\widetilde{\mu}_4}{\sigma^4} - 3.$$

- Shift -3 is used to manipulate the excess kurtosis of standard normal distribution to be 0, since the sharpness of a peak is estimated with a reference to that of $\mathcal{N}(0, 1)$.
- In order to make correct comparisons of two distributions with excess kurtosis, their variance should be identical.
- If $\mathsf{Kurt}(X) > 0$ – distribution peak is sharper than standard normal's one, if $\mathsf{Kurt}(X) < 0$ – distribution peak is smoother.

Higher-order moments



(a) Negatively skewed.



(b) Positively skewed.

Figure: P.d.f.-s of random variables with opposite skewness.

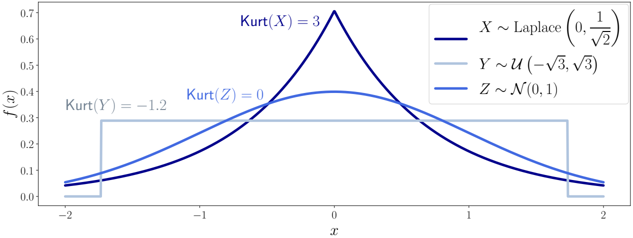


Figure: Comparison of excess kurtosis for Laplace X, uniform Y and normal Z distributions with zero mean and variance 1.

Laplace distribution

- Laplace distribution of $X \iff X \sim \text{Laplace}(\mu, b)$, where μ is a location of the peak, and b is a scale parameter.
- P.d.f.:
$$f_X(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}.$$
- C.d.f.:
$$F_X(x) = \frac{1}{2} + \frac{1}{2} \cdot \text{sign}(x-\mu) \cdot \left(1 - e^{-\frac{|x-\mu|}{b}}\right).$$
- Mean:
$$\mathsf{E}(X) = \mu.$$
- Variance:
$$\mathsf{V}(X) = 2b^2.$$

Moment-generating function

Definition

Moment-generating function of a random variable X:

$$\mathsf{M}_X(t) = \mathsf{E}\left(e^{tX}\right).$$

- In continuous case it's a bilateral Laplace transform of a p.d.f. $f_X(x)$ with parameter $-t$:

$$\mathsf{M}_X(t) = \int\limits_{-\infty}^{\infty} f_X(x) e^{-(-t)x} dx.$$

- $\mathsf{M}_X(t)$ is used to acquire raw moments μ_k:

$$\mu_k = \mathsf{E}\left(X^k\right) = \frac{d^k}{dt^k} \mathsf{M}_X(t) \bigg|_{t=0}.$$

Moment-generating function

- If $X_1, \ldots, X_n$ are independent random variables, then:

$$\mathsf{M}_{\sum\limits_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} \mathsf{M}_{X_i}(t).$$

- For linear transformation $\alpha X + \beta$, where $\alpha, \beta \in \mathbb{R}$:

$$\mathsf{M}_{\alpha X + \beta}(t) = e^{\beta t} \cdot \mathsf{M}_X(\alpha t).$$

- Collection of μ_k from $k = 0$ to $k = \infty$ uniquely determines the distribution, if p.d.f. is defined on bounded range. Please, refer to
  1. Hausdorff moment problem – support on $[0, 1]$. Always unique.
  2. Stieltjes moment problem – support on $[0, \infty)$. Requires sufficient condition for uniqueness.
  3. Hamburger moment problem – support on $(-\infty, \infty)$. Requires sufficient condition for uniqueness.

Method of moments

- Let $X_1, \ldots, X_n$ be a random sample from $F_X(x;\ \theta_1, \ldots, \theta_m)$.
- k^{th} sample moment:

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \overline{X^k}.$$

- Method of moments (MM) to find estimators of $\theta_1, \ldots, \theta_m$:

$$\begin{aligned} \mu_1\Big|_{\theta_1=\widehat{\theta}_1,\ldots,\theta_m=\widehat{\theta}_m} &= M_1, \\ \mu_2\Big|_{\theta_1=\widehat{\theta}_1,\ldots,\theta_m=\widehat{\theta}_m} &= M_2, \\ &\ldots \\ \mu_m\Big|_{\theta_1=\widehat{\theta}_1,\ldots,\theta_m=\widehat{\theta}_m} &= M_m. \end{aligned}$$

Problem 2

Let $\{X_1, \ldots, X_n\}$ be a random sample from a Bin (m, π) distribution, with both m and π unknown. Find the method of moments estimators for m – the number of trials, and π – the probability of success.

Problem 3

Suppose that we have a random sample $\{X_1, \ldots, X_n\}$ from a uniform distribution. Find the method of moments estimator of θ if

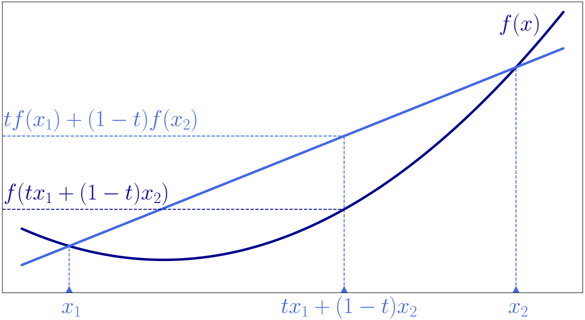
1. $X \sim \mathcal{U}(0, \theta)$,
2. $X \sim \mathcal{U}(-\theta, \theta)$.

Jensen's inequality

Jensen's inequality for 2 points x_1 and x_2

If f is a convex function, then

$$\forall t \in [0,1]: \qquad f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2),$$



Jensen's inequality

- Increasing number of splits, such that $\sum_{i=1}^{n} t_i = 1$:

$$f(t_1 x_1 + t_2 x_2 + \ldots + t_n x_n) \leq t_1 f(x_1) + t_2 f(x_2) + \ldots + t_n f(x_n),$$

- Giving t_i meaning of probabilities, Jensen's inequality becomes following:

Jensen's inequality (probabilistic statement)

If f is a convex function, then for any random variable X

$$f\left(\mathsf{E}(X)\right) \leq \mathsf{E}\left(f(X)\right).$$

- If f is concave, inequality is inverted.
- Equality is achieved when f is not strictly convex or concave, in other words only when f is linear.

Problem 3

- Results of Part (2):
  - $\widehat{\theta^2} = 3\overline{X^2}$,
  - $\widehat{\theta} = \sqrt{3\overline{X^2}}$.
- While $\widehat{\theta^2}$ is unbiased:

$$\mathsf{E}\left(\widehat{\theta^2}\right) = \mathsf{E}\left(3\overline{X^2}\right) = 3 \cdot \frac{1}{n} \cdot \mathsf{E}\left(n \cdot X^2\right) = 3 \cdot \frac{\theta^2}{3} = \theta^2,$$

  $\widehat{\theta}$ is NOT unbiased.
- Since square root $f(x) = \sqrt{x}$ is a strictly concave function, according to Jensen's inequality:

$$\mathsf{E}\left(\sqrt{3\overline{X^2}}\right) < \sqrt{\mathsf{E}\left(3\overline{X^2}\right)} = \sqrt{\theta^2} = \theta.$$

- Thus, $\widehat{\theta}$ underestimates true parameter θ.

Ordinary least squares

- Linear regression model of n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

  where $\mathbf{x}_i = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}^\top$ – vector of p regressors (independent variables), y_i – regressand (response variable), ε_i – noise / error / other influences on y_i.
- In matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

  where $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}^\top$ – vector of regressands,
  $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_p \end{pmatrix}^\top$ – vector of unknown parameters,
  $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{pmatrix}^\top$ – vector of noise / errors,
  $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$ – matrix of regressors.

Ordinary least squares

- Goal of regression: to find such estimate $\widehat{\boldsymbol{\beta}}$ that it will be as close as possible to the real vector $\boldsymbol{\beta}$.
- Ordinary least squares (OLS) is a method to find $\widehat{\boldsymbol{\beta}}$ via quadratic minimization of $\|\boldsymbol{\varepsilon}\|_2^2 = \varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2$.

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\varepsilon}\|^2 = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

- This minimization problem has known solution:

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

Theorem (Gauss-Markov)

$\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ has the lowest sample variance within the class of linear unbiased estimators, if $\forall i, j \in \overline{1, n} \quad \mathsf{E}(\varepsilon_i) = 0, \mathsf{V}(\varepsilon_i) = \sigma^2$ *and* $\mathsf{Cov}(\varepsilon_i, \varepsilon_j) \overset{i \neq j}{=} 0$.

Problem 4

Suppose that you are given observations y_1, y_2, y_3 and y_4 such that:

$$\begin{aligned}
y_1 &= \alpha + \beta + \varepsilon_1, \\
y_2 &= -\alpha + \beta + \varepsilon_2, \\
y_3 &= \alpha - \beta + \varepsilon_3, \\
y_4 &= -\alpha - \beta + \varepsilon_4.
\end{aligned}$$

The variables $\varepsilon_i, i \in \{1, 2, 3, 4\}$, are independent and normally distributed with mean 0 and variance σ^2.

1. Find the least squares estimators of the parameters α and β.
2. Verify that the least squares estimators in (a) are unbiased.
3. Find the variance of the least squares estimator of the parameter α.

Maximum likelihood estimator

- Let $X_1, \dots, X_n$ be a random sample from $F_X(x;\ \theta)$.
- Probability that all those observations happened is given by joint p.d.f. (or p.m.f. in discrete case):

$$f(X_1, X_2, \dots, X_n;\ \theta).$$

- Probability that we have sample $X_1, \dots, X_n$ given some particular θ is the same with probability that we have the same θ given that sample $X_1, \dots, X_n$.
  The latter is called likelihood function $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta;\ X_1, \dots, X_n) = f(X_1, X_2, \dots, X_n;\ \theta).$$

- Due to independence of X_i in sample:

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

Maximum likelihood estimator

- Maximum likelihood estimator (MLE) is an estimate of parameter θ, which maximizes likelihood function $\mathcal{L}(\theta)$:

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathcal{L}(\theta).$$

- Likelihood function is built as a product $\prod$, which is pretty hard to take a derivative from.
  Log-likelihood $l(\theta)$ is introduced:

$$l(\theta) = \ln \mathcal{L}(\theta),$$

  which converts products to sums.
  Since the logarithm is a monotonic function:

$$\arg\max_{\theta} l(\theta) = \arg\max_{\theta} \mathcal{L}(\theta).$$

- Invariance principle of the MLE:

$$\phi = g(\theta) \qquad \Longrightarrow \qquad \widehat{\phi}_{\text{MLE}} = g\left(\widehat{\theta}_{\text{MLE}}\right).$$

Problem 5

Let $\{X_1, \ldots, X_n\}$ be a random sample from $\text{Exp}(\lambda)$ distribution. Find the MLE of λ.

Problem 6

A random sample $\{X_1, X_2, \ldots, X_n\}$ is drawn from the following probability distribution:

$$p(x; \lambda) = \frac{\lambda^{2x} e^{-\lambda^2}}{x!}$$

and $p(x; \lambda) = 0$ for all other values of x, with $\lambda > 0$.

1. Derive the maximum likelihood estimator of λ.
2. State the maximum likelihood estimator of $\theta = \lambda^3$.

Problem 7

Suppose that X is a discrete random variable with the following probability mass function:

x	0	1	2	3
$\mathsf{P}_X(x)$	$\frac{2\theta}{3}$	$\frac{\theta}{3}$	$\frac{2(1-\theta)}{3}$	$\frac{1-\theta}{3}$

where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution:

$$(3, 0, 2, 1, 3, 2, 1, 0, 2, 1).$$

What is the maximum likelihood estimate of θ.

Problem 8

Let $\{X_1, \ldots, X_n\}$ be a random sample from $\mathcal{U}(0, \theta)$ distribution. Find the MLE of θ.

Gamma distribution

- Gamma distribution of $X \iff X \sim \text{Gamma}(\alpha, \beta)$, where α is a shape parameter, and β is a rate parameter.
- P.d.f.:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \cdot I_{\{x>0\}},$$

where $\Gamma(\alpha)$ is a generalization of the factorial to non-integers:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \qquad \Re(\alpha) > 0.$$

Its factorial-like behaviour is expressed via property of recurrence:

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha).$$

- Mean: $\mathsf{E}(X) = \dfrac{\alpha}{\beta}$.
- Variance: $\mathsf{V}(X) = \dfrac{\alpha}{\beta^2}$.

Problem 9

Suppose that independent observations X and Y are taken from distributions Gamma $\left(a, \frac{1}{\eta}\right)$ and Gamma $\left(b, \frac{1}{\eta}\right)$ respectively, where both a and b are known and positive.

1. Find the maximum likelihood estimator (MLE) of η.
2. Show that the MLE of η is unbiased and find its variance.
3. Compare the MLE with the alternative estimator

$$\widehat{H} = \frac{1}{2}\left(\frac{X}{a} + \frac{Y}{b}\right).$$

   Which one is better?

Problem 10

Find maximum likelihood estimator of parameter θ from sample $\{X_1, \ldots, X_n\}$ with Laplace$(\theta, 1)$ distribution and p.d.f.:

$$f(x) = \frac{1}{2}e^{-|x-\theta|}.$$

Look at the time!