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_1X_2.

Solution:

Since measurements are independent:

$$\mathsf{E}(X_1X_2) = \mathsf{E}(X_1)\mathsf{E}(X_2) = a^2,$$

which means that X_1X_2 is an unbiased estimator of a^2:

$$\text{Bias}(X_1X_2) = \mathsf{E}(X_1X_2) - a^2 = a^2 - a^2 = 0.$$

Also, we will need an expectation of X_i^2. From variance and mean identity:

$$\mathsf{E}\left(X_i^2\right) = \mathsf{V}(X_i) + \mathsf{E}(X_i)^2 = \sigma^2 + a^2.$$

MSE is a sum of bias squared and variance. Let's calculate a variance of X_1X_2 (using independence of X_1 and X_2):

$$\begin{aligned}\mathsf{V}(X_1X_2) &= \mathsf{E}\left(X_1^2X_2^2\right) - \mathsf{E}(X_1X_2)^2 = \mathsf{E}\left(X_1^2\right)\mathsf{E}\left(X_2^2\right) - \mathsf{E}(X_1)^2\mathsf{E}(X_1)^2 = \\ &= \left(\sigma^2 + a^2\right)^2 - \left(a^2\right)^2 = \sigma^4 + 2a^2\sigma^2 + a^4 - a^4 = \sigma^4 + 2a^2\sigma^2.\end{aligned}$$

MSE of X_1X_2 then:

$$\text{MSE}(X_1X_2) = \text{Bias}^2(X_1X_2) + \mathsf{V}(X_1X_2) = 0^2 + \sigma^4 + 2a^2\sigma^2 = \boxed{\sigma^4 + 2a^2\sigma^2}.$$

Problem 1

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

(a) Can you conclude which parameter is greater without calculations – mean or median?

(b) Calculate them explicitly.

(c) Discuss the result. Why p.d.f. is unbalanced in terms of "mass" around its median?

Solution:

Let's consider random variable X with p.d.f.:

$$f(x) = \left(1 - \frac{x}{2}\right) \cdot I_{\{0 \le x \le 2\}}.$$

(a) The distribution $f(x)$ is continuous unimodal, and its right tail is longer than the left one (skewed to the right), which means that

$$\mathsf{E}(X) > \text{median}(X).$$

See the illustration in the fig. 1.

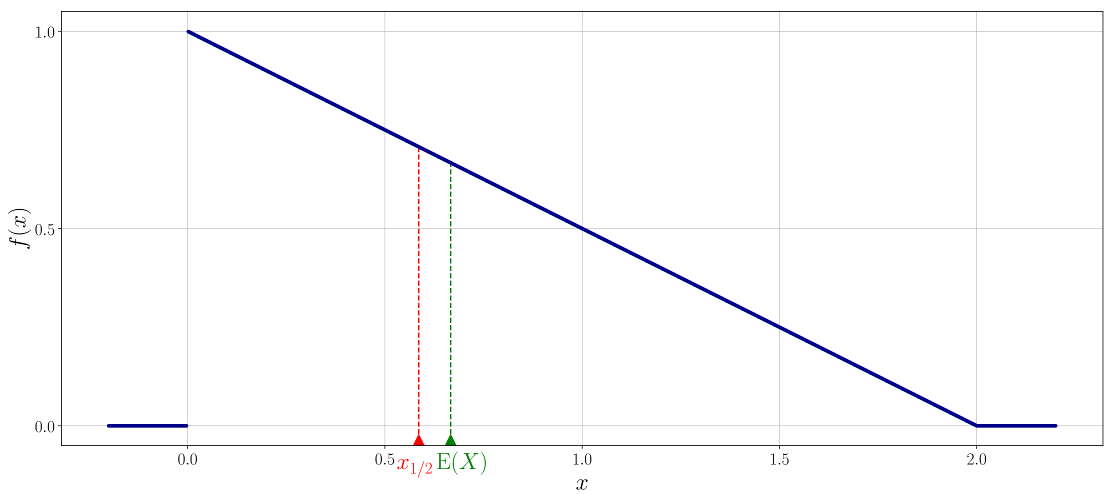


Figure 1: P.d.f. $f(x) = \left(1 - \frac{x}{2}\right) \cdot I_{\{0 \le x \le 2\}}$ with mean $\mathsf{E}(X)$ and median $x_{1/2}$.

(b) Mean:

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{2} x\left(1 - \frac{x}{2}\right) dx = \frac{x^2}{2}\Big|_0^2 - \frac{x^3}{6}\Big|_0^2 = \boxed{\frac{2}{3}}.$$

Let $x_{1/2} = \text{median}(X)$. Then:

$$\int\limits_0^{x_{1/2}} \left(1 - \frac{x}{2}\right) dx = \frac{1}{2},$$
$$x\Big|_0^{x_{1/2}} - \frac{x^2}{4}\Big|_0^{x_{1/2}} = \frac{1}{2},$$
$$x_{1/2}^2 - 4x_{1/2} + 2 = 0.$$

$$x_{1/2} = \boxed{2 - \sqrt{2}}.$$

(c) It's often unclear why mean and median of some random variable are different quantities. From a physical analogy we know that expected value is the center of mass of a distribution. Let's try to balance the distribution around its median.

We have a lever with an fulcrum at the point $x_{1/2} = 2 - \sqrt{2}$. Masses on both sides from fulcrum are identical and equal to 1/2. Now we need to find their centers of mass.

The left one:

$$x_1 = \frac{\int\limits_0^{x_{1/2}} xf(x)dx}{\int\limits_0^{x_{1/2}} f(x)dx} = \frac{\int\limits_0^{2-\sqrt{2}} \left(x - \frac{x^2}{2}\right) dx}{\frac{1}{2}} = \frac{2}{3}\left(\sqrt{2} - 1\right).$$

The right one:

$$x_2 = \frac{\int\limits_{x_{1/2}}^{2} xf(x)dx}{\int\limits_{x_{1/2}}^{2} f(x)dx} = \frac{\int\limits_{2-\sqrt{2}}^{2} \left(x - \frac{x^2}{2}\right) dx}{\frac{1}{2}} = \frac{2}{3}\left(3 - \sqrt{2}\right).$$

Distances from median to those points:

$$x_{1/2} - x_1 = \frac{8 - 5\sqrt{2}}{3} \approx 0.31,$$
$$x_2 - x_{1/2} = \frac{\sqrt{2}}{3} \approx 0.47,$$

which means that the right part produces bigger moment and a lever will rotate clockwise. If the fulcrum had been in a point of expected value, lever would have been stable.

Mean and median are such values that minimize following expressions:

$$\mathsf{E}(X) = \arg\min_{M_2} \int_X (x - M_2)^2 \cdot f(x)dx,$$
$$\text{median}(X) = \arg\min_{M_1} \int_X |x - M_1| \cdot f(x)dx,$$

and you can see that in contrast to median, mean considers distance to the point unevenly – with quadratic penalty. So the further point from expected value, the higher its "cost". This is reflected in mechanical analogy – we consider not only a mass, but also a distance to that mass.

Problem 2

Let $\{X_1, \dots, X_n\}$ be a random sample from a $\text{Bin}(m, \pi)$ 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.

Solution:

Moment is a quantitative measure, related to the shape of p.d.f.'s graph.

Population k^{th} raw moments of a random variable X are calculated as follows:

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

Moments about mean $\mathsf{E}(X)$ are called central and calculated the following way:

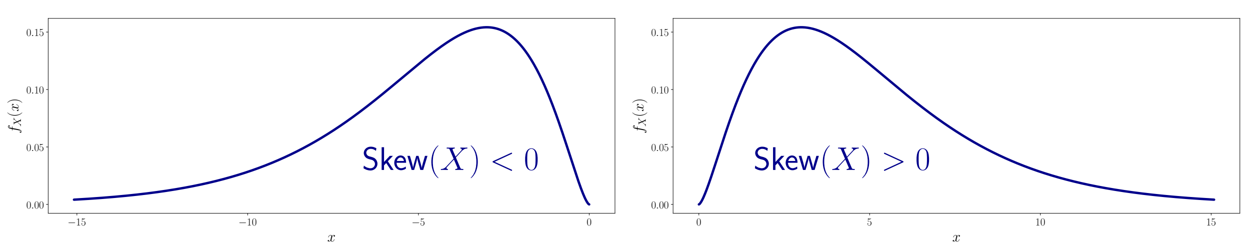
$$\widetilde{\mu}_k = \mathsf{E}\left((X - \mathsf{E}(X))^k\right).$$

They describe the shape of the function, independently of translation. The best known moments are: the 1^{st} raw one – mean, and the 2^{nd} central – variance σ^2. If the function represents mass, then the 1^{st} moment is the center of the mass, and the 2^{nd} moment is the rotational inertia.

Let's discuss a few of higher-order moments. The 3^{rd} standardized central moment is called skewness and shows the asymmetry of a function:

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

Signs of skewness and their effect on p.d.f. are reflected in the fig. 2.



(a) Negatively skewed.

(b) Positively skewed.

Figure 2: Probability density functions of random variables with opposite skewness.

The 4^{th} standardized central moment is called excess kurtosis and shows the sharpness of distribution peak:

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

where -3 shift 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. The example is illustrated in the fig. 3.

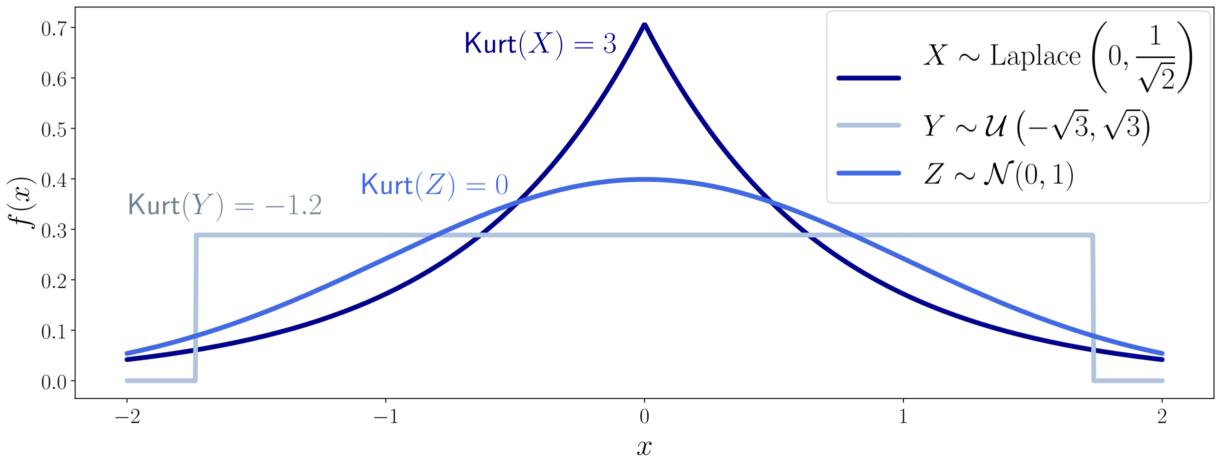


Figure 3: Comparison of excess kurtosis for Laplace X, uniform Y and normal Z distributions with zero mean and variance 1.

Population moments can be acquired from moment-generating function $\mathsf{M}_X(t)$ of a random variable X, which is defined as

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

which is basically two-sided Laplace transform with parameter $-t$ of a p.d.f. $f_X(x)$ in continuous case. We can get moments μ_k as follows:

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

Let's note some important properties of moment-generation functions. 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 $\alpha, \beta \in \mathbb{R}$:

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

In addition to population moments μ_k, which we can get from population distribution of a random variable X, there are also sample moments M_k, which are calculated from a sample of n i.i.d. observations $X_1, \ldots, X_n$:

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

The method of moments (MM) is to equate population and sample moments of the same degree k:

$$\mu_k = M_k,$$

and out of this system to get an estimation of a required parameter. Number of equations depends on the number of unknown parameters.

In our problem we have 2 unknown parameters – m and π. Thus, we need to make equations for the 1st and the 2nd moments:

$$\begin{cases} \mu_1 = M_1, \\ \mu_2 = M_2, \end{cases} \implies \begin{cases} \mathsf{E}(X)\Big|_{m=\widehat{m},\pi=\widehat{\pi}} = \overline{X}, \\ \mathsf{E}\left(X^2\right)\Big|_{m=\widehat{m},\pi=\widehat{\pi}} = \overline{X^2}, \end{cases} \implies \begin{cases} \widehat{m}\widehat{\pi} = \overline{X}, \\ \widehat{m}\widehat{\pi}\left(1-\widehat{\pi}\right) + \left(\widehat{m}\widehat{\pi}\right)^2 = \overline{X^2}. \end{cases}$$

Here we used the fact that for $X \sim \text{Bin}(m,\pi)$: $\mathsf{E}(X) = m\pi$ and $\mathsf{E}\left(X^2\right) = \mathsf{V}(X) + \mathsf{E}(X)^2 = m\pi(1-\pi) + (m\pi)^2$. Also, when we equated population and sample moments, we have made an estimation for required parameters m and π (true parameter shall not change, depending on the values from sample), thus putting hats over them.

Substituting $\overline{X}$ into the second equation:

$$\widehat{m}\widehat{\pi}\left(1-\widehat{\pi}\right) + \overline{X}^2 = \overline{X^2},$$

$$\widehat{m}\widehat{\pi}\left(1-\widehat{\pi}\right) = \overline{X^2} - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \widehat{\sigma}^2,$$

where $\widehat{\sigma}^2$ is a biased sample variance. In the end we have two simple equations:

$$\begin{cases} \widehat{m}\widehat{\pi} = \overline{X}, \\ \widehat{m}\widehat{\pi}\left(1-\widehat{\pi}\right) = \widehat{\sigma}^2, \end{cases} \implies \begin{cases} \widehat{m} = \dfrac{\overline{X}}{\widehat{\pi}}, \\ 1-\widehat{\pi} = \dfrac{\widehat{\sigma}^2}{\overline{X}}, \end{cases} \implies \boxed{\begin{cases} \widehat{m} = \dfrac{\overline{X}^2}{\overline{X}-\widehat{\sigma}^2}, \\ \widehat{\pi} = 1 - \dfrac{\widehat{\sigma}^2}{\overline{X}}. \end{cases}}$$

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

(a) $X \sim \mathcal{U}(0, \theta)$,

(b) $X \sim \mathcal{U}(-\theta, \theta)$.

Solution:

Let's remember that for a random variable $X \sim \mathcal{U}(a, b)$, population moments are:

$$\mathsf{E}(X) = \frac{a+b}{2} \quad \text{and} \quad \mathsf{V}(X) = \frac{(b-a)^2}{12}.$$

(a) Equation on first moments gives:

$$\mathsf{E}(X)\Big|_{\theta=\widehat{\theta}} = \overline{X}, \qquad \Longrightarrow \qquad \frac{0+\widehat{\theta}}{2} = \overline{X}, \qquad \Longrightarrow \qquad \widehat{\theta} = \boxed{2\overline{X}}.$$

(b) Let's try to repeat previous calculations, but for symmetric distribution:

$$\mathsf{E}(X)\Big|_{\theta=\widehat{\theta}} = \overline{X}, \qquad \Longrightarrow \qquad \frac{-\widehat{\theta}+\widehat{\theta}}{2} = \overline{X}, \qquad \Longrightarrow \qquad 0 = \overline{X}.$$

We end up with degenerate result, which can be "approximately correct" for large sample sizes.

Let's try to use an equation of another order, the 2^{nd} for instance:

$$\mathsf{E}\left(X^2\right)\Big|_{\theta=\widehat{\theta}} = \mathsf{V}\left(X\right)\Big|_{\theta=\widehat{\theta}} + \mathsf{E}\left(X\right)^2\Big|_{\theta=\widehat{\theta}} = \overline{X^2}, \qquad \Longrightarrow \qquad \frac{\left(\widehat{\theta} - \left(-\widehat{\theta}\right)\right)^2}{12} + 0^2 = \overline{X^2},$$

$$\Longrightarrow \qquad \widehat{\theta}^2 = 3\overline{X^2}, \qquad \Longrightarrow \qquad \widehat{\theta} = \boxed{\sqrt{3\overline{X^2}}}.$$

Let's find an estimator for θ^2 and calculate its bias:

$$\mathsf{E}\left(X^2\right)\Big|_{\theta^2=\widehat{\theta^2}} = \overline{X^2}, \qquad \Longrightarrow \qquad \frac{\left(\sqrt{\widehat{\theta^2}} - \left(-\sqrt{\widehat{\theta^2}}\right)\right)^2}{12} = \overline{X^2}, \qquad \Longrightarrow \qquad \widehat{\theta^2} = 3\overline{X^2}.$$

Estimator $\widehat{\theta^2}$ is clearly unbiased:

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

Does it mean that the estimator $\widehat{\theta}$ is also unbiased? We can't calculate bias directly, but the answer to the question is NO, which can be seen from Jensen's inequality:

$$\mathsf{E}\left(\sqrt{3\overline{X^2}}\right) \leq \sqrt{\mathsf{E}\left(3\overline{X^2}\right)},$$

which holds since square root $f(x) = \sqrt{x}$ is a concave function. Moreover, equality is achieved when function f is not strictly convex or convex, in other words when it's linear. Since we have strictly concave function, in our case:

$$\mathsf{E}\left(\widehat{\theta}\right) < \sqrt{\mathsf{E}\left(\widehat{\theta}^2\right)} = \sqrt{\theta^2} = \theta, \qquad \Longrightarrow \qquad \text{Bias}\left(\widehat{\theta}\right) < 0.$$

Thus, the estimator $\widehat{\theta}$ always underestimates true parameter θ.

The standard formulation of Jensen's inequality for 2 points x_1 and x_2 is following:

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

where f is a convex function. The inequality formalizes the statement that the secant line of a convex function lies above the graph of the function, illustrated in the fig. 4.

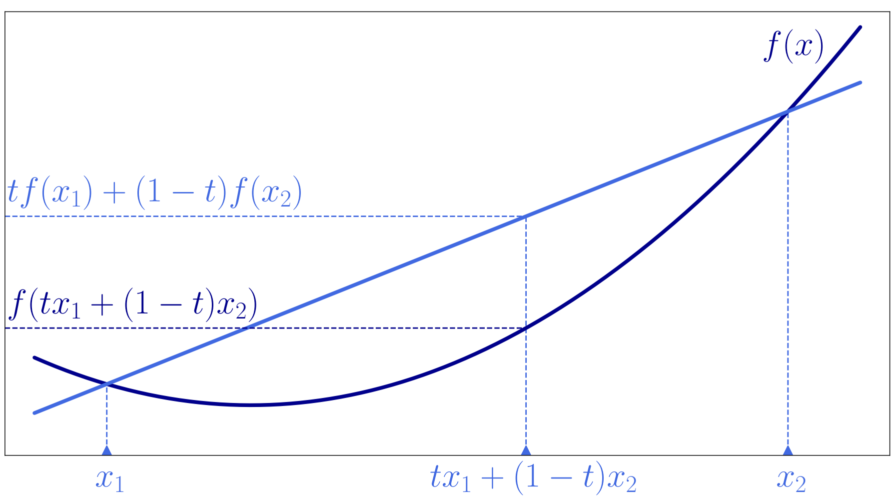


Figure 4: Jensen's inequality for 2 points.

If we would split the interval (x_1, x_2) with not one t, but with several t_i, such that $\sum\limits_{i=1}^{n} t_i = 1$, number of which could tend to infinity, Jensen's inequality can be expressed in probabilistic statement, where t_i have physical meanings of probabilities:

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

where f is a convex function. Inequality inverts for a concave function.

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.

(a) Find the least squares estimators of the parameters α and β.

(b) Verify that the least squares estimators in (a) are unbiased.

(c) Find the variance of the least squares estimator of the parameter α.

Solution:

(a) **I. Matrices.** Let's use a known solution for OLS in matrix notation:

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

where $\mathbf{X}$ is a matrix of regressors, $\mathbf{y}$ is a vector of regressands, $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is vector of OLS parameters estimates:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \\ 1 & -1 \\ -1 & -1 \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}}_{\text{OLS}} = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix}.$$

Pseudoinverse matrix:

$$\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top = \left(\begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \\ 1 & -1 \\ -1 & -1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{pmatrix} =$$

$$= \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{pmatrix}.$$

Thus, OLS estimates:

$$\boxed{\begin{cases} \widehat{\alpha} = \dfrac{y_1 - y_2 + y_3 - y_4}{4}, \\ \widehat{\beta} = \dfrac{y_1 + y_2 - y_3 - y_4}{4}. \end{cases}}$$

II. Direct calculation. Let's minimize square of the second norm of error vector $\boldsymbol{\varepsilon}$ explicitly. Error vector is:

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 \end{pmatrix}^\top,$$

and quantity to minimize:

$$\text{ESS} = \|\boldsymbol{\varepsilon}\|^2 = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2,$$

$$\widehat{\alpha}, \widehat{\beta} = \arg\min_{\alpha,\beta} \text{ESS}.$$

Expressing ESS via regressands and parameters:

$$\text{ESS} = (y_1 - \alpha - \beta)^2 + (y_2 + \alpha - \beta)^2 + (y_3 - \alpha + \beta)^2 + (y_4 + \alpha + \beta)^2.$$

By necessary condition of extremum:

$$\begin{cases} \left.\dfrac{\partial \text{ESS}}{\partial \alpha}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} = -2(y_1 - \widehat{\alpha} - \widehat{\beta}) + 2(y_2 + \widehat{\alpha} - \widehat{\beta}) - 2(y_3 - \widehat{\alpha} + \widehat{\beta}) + 2(y_4 + \widehat{\alpha} + \widehat{\beta}) = 0, \\ \left.\dfrac{\partial \text{ESS}}{\partial \beta}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} = -2(y_1 - \widehat{\alpha} - \widehat{\beta}) - 2(y_2 + \widehat{\alpha} - \widehat{\beta}) + 2(y_3 - \widehat{\alpha} + \widehat{\beta}) + 2(y_4 + \widehat{\alpha} + \widehat{\beta}) = 0, \end{cases}$$

$$\begin{cases} 4\widehat{\alpha} = y_1 - y_2 + y_3 - y_4, \\ 4\widehat{\beta} = y_1 + y_2 - y_3 - y_4, \end{cases} \implies \boxed{\begin{cases} \widehat{\alpha} = \dfrac{y_1 - y_2 + y_3 - y_4}{4}, \\ \widehat{\beta} = \dfrac{y_1 + y_2 - y_3 - y_4}{4}. \end{cases}}$$

Here we need to prove that values of $\widehat{\alpha}$ and $\widehat{\beta}$ are indeed arguments of minimum. Second derivatives:

$$\left.\frac{\partial^2 \text{ESS}}{\partial \alpha^2}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} = 8, \qquad \left.\frac{\partial^2 \text{ESS}}{\partial \alpha \partial \beta}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} = \left.\frac{\partial^2 \text{ESS}}{\partial \beta \partial \alpha}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} = 0, \qquad \left.\frac{\partial^2 \text{ESS}}{\partial \beta^2}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} = 8.$$

By sufficient condition of the minimum, let's see if the Hessian matrix $\mathcal{H}$ is positive-definite (all minors are positive):

$$\mathcal{H} = \begin{pmatrix} \left.\dfrac{\partial^2 \text{ESS}}{\partial \alpha^2}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} & \left.\dfrac{\partial^2 \text{ESS}}{\partial \alpha \partial \beta}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} \\ \left.\dfrac{\partial^2 \text{ESS}}{\partial \beta \partial \alpha}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} & \left.\dfrac{\partial^2 \text{ESS}}{\partial \alpha^2}\right|_{\alpha=\widehat{\alpha},\beta=\widehat{\beta}} \end{pmatrix} = \begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix} \succ 0.$$

Q.E.D.

(b) Estimator is unbiased, if its expected value is equal to estimated parameter.

$$\mathsf{E}\left(\widehat{\alpha}\right) = \mathsf{E}\left(\frac{y_1 - y_2 + y_3 - y_4}{4}\right) = \frac{1}{4}\left(\mathsf{E}(y_1) - \mathsf{E}(y_2) + \mathsf{E}(y_3) - \mathsf{E}(y_4)\right).$$

Regressands y_i are random variables, since they are linear combination of parameters (constants) and errors ε_i, which are random variables themselves. Since ε_i are zero-mean, expected value of y_i is defined by terms of α and β:

$$\mathsf{E}\left(\widehat{\alpha}\right) = \frac{1}{4}\left((\alpha + \beta) - (-\alpha + \beta) + (\alpha - \beta) - (-\alpha - \beta)\right) = \boxed{\alpha}.$$

$$\mathsf{E}\left(\widehat{\beta}\right) = \mathsf{E}\left(\frac{y_1 + y_2 - y_3 - y_4}{4}\right) = \frac{1}{4}\left(\mathsf{E}(y_1) + \mathsf{E}(y_2) - \mathsf{E}(y_3) - \mathsf{E}(y_4)\right) =$$
$$= \frac{1}{4}\left((\alpha + \beta) + (-\alpha + \beta) - (\alpha - \beta) - (-\alpha - \beta)\right) = \boxed{\beta}.$$

(c) Variance of the regressand y_i is equal to the variance ε_i, since constant terms do not affect variance. Since ε_i are independent, y_i are also independent.

$$\mathsf{V}\left(\widehat{\alpha}\right) = \mathsf{V}\left(\frac{y_1 - y_2 + y_3 - y_4}{4}\right) = \frac{1}{4^2}\big(\mathsf{V}(y_1) + (-1)^2 \cdot \mathsf{V}(y_2) + \mathsf{V}(y_3) + (-1)^2 \cdot \mathsf{V}(y_4)\big) =$$
$$= \frac{1}{16}\left(\sigma^2 + \sigma^2 + \sigma^2 + \sigma^2\right) = \boxed{\frac{\sigma^2}{4}}.$$

Problem 5

A coin was tossed 10 times. Faces of a coin turned out as follows:

$$H, H, H, H, T, H, H, H, T, H.$$

What is the most likely probability of getting heads after one toss?

Solution:

Let the result of one throw of a coin be a Bernoulli random variable with probability of success (getting heads) p:

$$X \sim \text{Bernoulli}(p).$$

The result of an experiment with 10 throws has the following probability (considering outcomes in sample collectively independent):

$$\begin{aligned} &\mathsf{P}(X_1 = 1, X_2 = 1, \ldots, X_9 = 0, X_{10} = 1) = \\ &= \mathsf{P}(X_1 = 1) \cdot \mathsf{P}(X_2 = 1) \cdot \ldots \cdot \mathsf{P}(X_9 = 0) \cdot \mathsf{P}(X_{10} = 1) = p^8(1-p)^2. \end{aligned}$$

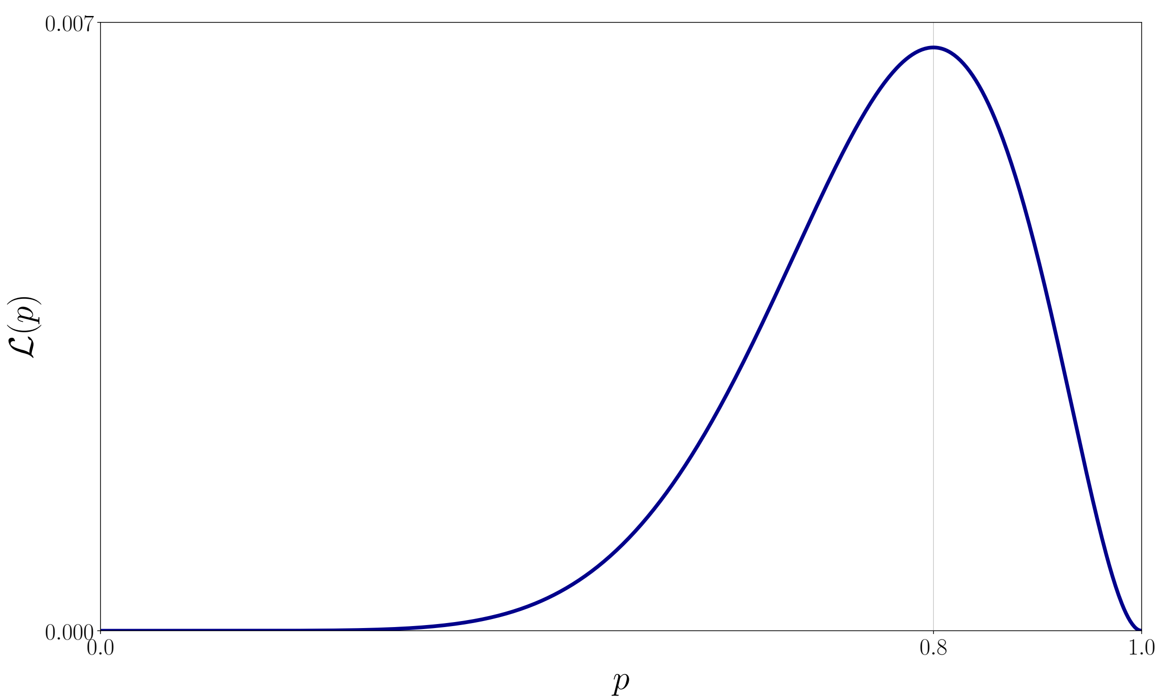


Figure 5: Likelihood function $\mathcal{L}(p)$.

Now the joint probability is not the function of random variable, but the function of parameter p, which shows the likelihood of getting that sample $\{X_1, \dots, X_{10}\}$ with a particular value of parameter p.

This function is called likelihood function $\mathcal{L}(p)$ and achieves its maximum in a value of $\widehat{p}$, called maximum likelihood estimator of p:

$$\mathcal{L}(p) = p^8(1-p)^2.$$

So we need to find the argument of maximum of $\mathcal{L}(p)$:

$$\widehat{p} = \arg\max_p \mathcal{L}(p).$$

The intuition suggests that the correct value is $\widehat{p} = 0.8$, and it's validated by the graph of $\mathcal{L}(p)$, shown in the fig. 5, but let's derive it explicitly.

According to the FOC of extrema, we should find a derivative of $\mathcal{L}(p)$ and equalize it to zero. In those points the MLE $\widehat{p}$ should be found. Additionally, we also should verify that the extremum is indeed maximum with the SOC (second-order derivative is less than zero in a point $\widehat{p}$).

But the likelihood $\mathcal{L}(p)$ is a product (by definition) and is kinda hard to take derivative from. That's why log-likelihood $l(p)$ is introduced:

$$l(p) = \ln \mathcal{L}(p).$$

In log-transformation the product converts into the sum, which is pretty easy to differentiate. Also, the argument of maximum doesn't change after the transform, since the logarithm is a monotonic function:

$$\widehat{p} = \arg\max_p \mathcal{L}(p) = \arg\max_p \ln \mathcal{L}(p) = \arg\max_p l(p).$$

The log-likelihood function:

$$l(p) = 8\ln p + 2\ln(1-p).$$

According to the FOC of extrema:

$$\left.\frac{dl(p)}{dp}\right|_{p=\widehat{p}} = \frac{8}{\widehat{p}} - \frac{2}{1-\widehat{p}} = 0,$$

$$\boxed{\widehat{p} = 0.8}.$$

According to the SOC of extrema:

$$\left.\frac{d^2l(p)}{dp^2}\right|_{p=\widehat{p}} = -\frac{8}{\widehat{p}^2} - \frac{2}{(1-\widehat{p})^2} = -62.5 < 0,$$

which means that $\widehat{p}$ is indeed an argument of maximum.

Problem 6

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

(a) Derive the MLE of λ.

(b) State the MLE of $\theta = \lambda^3$.

Solution:

(a) The p.d.f. of a random variable with distribution $\text{Exp}(\lambda)$ is

$$f(x) = \lambda e^{-\lambda x} \cdot I_{\{x \geq 0\}}.$$

The likelihood function:

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} f(X_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} \cdot I_{\{X_i \geq 0\}} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} X_i\right) \cdot \prod_{i=1}^{n} I_{\{X_i \geq 0\}}.$$

If at least one X_i is less than 0, then we are in the wrong to use $\text{Exp}(\lambda)$ model. If the sample is fine, we can drop indicators from consideration.

The log-likelihood function:

$$l(\lambda) = \ln \mathcal{L}(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} X_i.$$

According to the FOC of extrema, possible values of the MLE $\widehat{\lambda}$ are found as follows:

$$\left.\frac{dl(\lambda)}{d\lambda}\right|_{\lambda=\widehat{\lambda}} = \frac{n}{\widehat{\lambda}} - \sum_{i=1}^{n} X_i = 0.$$

$$\boxed{\widehat{\lambda} = \frac{1}{\overline{X}}}.$$

According to the SOC of extrema:

$$\left.\frac{d^2 l(\lambda)}{d\lambda^2}\right|_{\lambda=\widehat{\lambda}} = -\frac{n}{\widehat{\lambda}^2} = -n\overline{X}^2 < 0,$$

which means that $\widehat{\lambda}$ is indeed an argument of maximum.

(b) According to the invariance principle of the MLE:

$$\widehat{\theta} = \widehat{\lambda}^3 = \boxed{\left(\frac{1}{\overline{X}}\right)^3}.$$

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 $0 \le \theta \le 1$ 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.

Solution:

Since the sample is $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$, the likelihood is

$$\begin{aligned}\mathcal{L}(\theta) = \mathsf{P}(X=3)\cdot\mathsf{P}(X=0)\cdot\mathsf{P}(X=2)\cdot\mathsf{P}(X=1)\cdot\mathsf{P}(X=3)\,\cdot \\ \cdot\,\mathsf{P}(X=2)\cdot\mathsf{P}(X=1)\cdot\mathsf{P}(X=0)\cdot\mathsf{P}(X=2)\cdot\mathsf{P}(X=1).\end{aligned}$$

Substituting from the probability distribution given above, we have

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathsf{P}_X(x_i;\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

The log-likelihood function is following:

$$\begin{aligned}l(\theta) &= \log\mathcal{L}(\theta) = \sum_{i=1}^{n} \log\mathsf{P}(x_i;\theta) = \\ &= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) + \\ &+ 2\left(\log\frac{1}{3} + \log(1-\theta)\right) = 5\log\frac{1}{3} + 5\log\frac{2}{3} + 5\log\theta + 5\log(1-\theta),\end{aligned}$$

For MLE $\widehat{\theta}$ the derivative of $l(\theta)$ with respect to θ equals zero:

$$\left.\frac{dl(\theta)}{d\theta}\right|_{\theta=\widehat{\theta}} = \frac{5}{\widehat{\theta}} - \frac{5}{1-\widehat{\theta}} = 0,$$

and the solution gives us the MLE, which is $\boxed{\widehat{\theta} = 0.5}$. The method of moments estimation is $\widehat{\theta}_{\text{MM}} = \frac{5}{12}$, which is different from MLE.

Problem 8

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

Solution:

The p.d.f. of a random variable with distribution $\mathcal{U}(0, \theta)$ is

$$f(x) = \frac{1}{\theta} \cdot I_{\{0 \le x \le \theta\}}.$$

The likelihood function:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(X_i; \theta) = \prod_{i=1}^{n} \frac{1}{\theta} \cdot I_{\{0 \le X_i \le \theta\}} = \left(\frac{1}{\theta}\right)^n \cdot \prod_{i=1}^{n} I_{\{0 \le X_i \le \theta\}}.$$

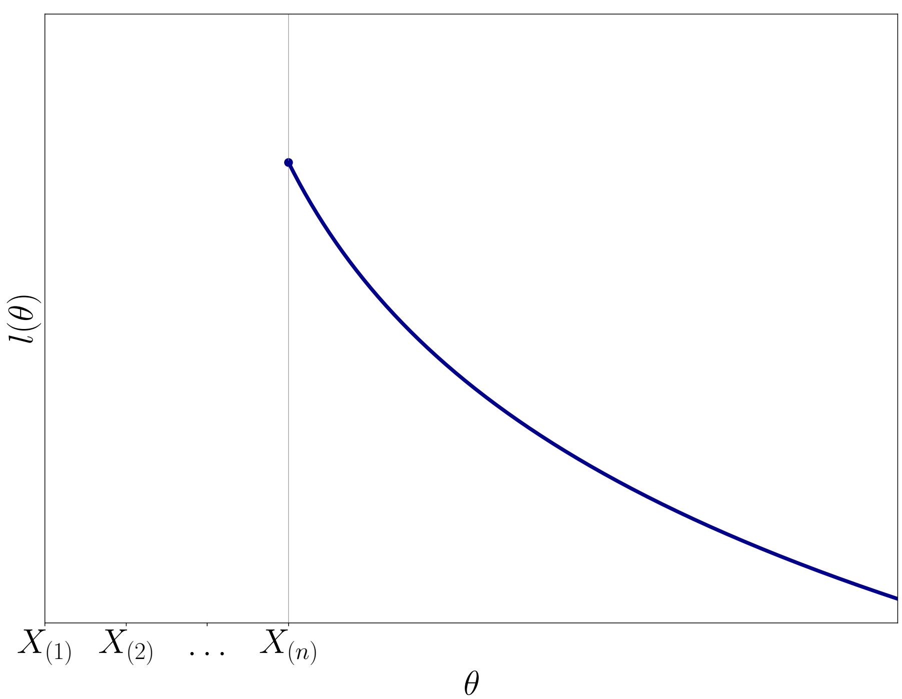


Figure 6: Log-likelihood function $l(\theta)$.

The log-likelihood function:

$$l(\theta) = \ln \mathcal{L}(\theta) = -n \ln \theta + \begin{cases} 0, & X_i \in [0, \theta] \quad \forall i \in \overline{1, n}, \\ -\infty, & \text{otherwise.} \end{cases}$$

The log-likelihood is a function of θ, but the condition in cases is built around X_i. Let's rearrange it.

Having in mind that $\forall i \in \overline{1,n} : X_i \geq 0$ (otherwise the presupposition to use $\mathcal{U}(0,\theta)$ model is wrong):

$$l(\theta) = -n \ln \theta + \begin{cases} 0, & \theta \geq X_i \quad \forall i \in \overline{1,n}, \\ -\infty, & \text{otherwise.} \end{cases}$$

The condition of θ being greater than all X_i means that it's greater than the maximal X_i. In notation of order statistics: $\theta \geq X_{(n)}$. This cutoff is shown in the fig. 6.

The log-likelihood function $l(\theta)$ reaches its maximum in $\theta = X_{(n)}$, which means that it's the maximum likelihood estimator of θ:

$$\boxed{\widehat{\theta} = X_{(n)} \equiv \max(X_1, \ldots, X_n)}.$$

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.

Note: The probability density function of a random variable that follows Gamma(α, β) distribution is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

where $\Gamma(\alpha)$ is a function, which represents generalization of the factorial to non-integer numbers:

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

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

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

In particular, in case of α being a positive integer:

$$\Gamma(\alpha) = (\alpha - 1)!$$

(a) Find the maximum likelihood estimator (MLE) of η.

(b) Show that the MLE of η is unbiased and find its variance.

(c) Compare the MLE with the alternative estimator

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

Which one is better?

Solution:

(a) Let's construct the likelihood function $\mathcal{L}(\eta)$, which is the product of p.d.f.s with known outcomes X and Y and population parameters $\left(a, \dfrac{1}{\eta}\right)$ and $\left(b, \dfrac{1}{\eta}\right)$ respectively:

$$\begin{aligned}\mathcal{L}(\eta) &= f\left(X; a, \frac{1}{\eta}\right) \cdot f\left(Y; b, \frac{1}{\eta}\right) = \\ &= \left(\frac{1}{\eta}\right)^a \frac{1}{\Gamma(a)} X^{a-1} e^{-\frac{X}{\eta}} \cdot \left(\frac{1}{\eta}\right)^b \frac{1}{\Gamma(b)} Y^{b-1} e^{-\frac{Y}{\eta}} = \\ &= \left(\frac{1}{\eta}\right)^{a+b} e^{-\frac{X+Y}{\eta}} \frac{X^{a-1}Y^{b-1}}{\Gamma(a)\Gamma(b)}.\end{aligned}$$

Building log-likelihood $l(\eta)$, and using the property of the logarithm of product:

$$l(\eta) = \ln \mathcal{L}(\eta) = -(a+b)\ln\eta - \frac{X+Y}{\eta} + \ln\left(\frac{X^{a-1}Y^{b-1}}{\Gamma(a)\Gamma(b)}\right).$$

Let's find the maximum of the function $l(\eta)$. Firstly, we have to find all possible extrema via Fermat's theorem of stationary points – in those points the derivative of the function $l(\eta)$ equals zero. The derivative of $l(\eta)$ is:

$$\frac{dl}{d\eta} = -\frac{a+b}{\eta} + \frac{X+Y}{\eta^2}.$$

Thus, stationary points are:

$$\frac{dl}{d\eta}(\widehat{\eta}) = 0 \quad \iff \quad -\frac{a+b}{\widehat{\eta}} + \frac{X+Y}{\widehat{\eta}^2} = 0 \quad \iff \quad \boxed{\widehat{\eta} = \frac{X+Y}{a+b}}.$$

We have to prove that $\widehat{\eta}$ is the point of maximum. To do that, we can find the sign of the second derivative of $l(\eta)$ in this point. The second derivative of $l(\eta)$ is:

$$\frac{d^2l}{d\eta^2} = \frac{a+b}{\eta^2} - \frac{2(X+Y)}{\eta^3}.$$

Substituting $\widehat{\eta}$ gives:

$$\frac{d^2l}{d\eta^2}(\widehat{\eta}) = \frac{a+b}{\widehat{\eta}^2} - \frac{2(X+Y)}{\widehat{\eta}^3} = -\frac{(a+b)^3}{(X+Y)^2} < 0.$$

Since the second derivative of $l(\eta)$ in $\widehat{\eta}$ is negative, $\widehat{\eta}$ is the point of maximum. Thus, we have proven that $\widehat{\eta}$ is the maximum likelihood estimator of η.

(b) If the expected value of the estimator equals exactly the estimated parameter, this estimator is considered to be unbiased. In other words, we have to prove that $\mathsf{E}\left(\widehat{\eta}\right) = \eta$.

Using the linearity of expected value:

$$\mathsf{E}\left(\widehat{\eta}\right) = \mathsf{E}\left(\frac{X+Y}{a+b}\right) = \frac{1}{a+b}\left(\mathsf{E}(X) + \mathsf{E}(Y)\right).$$

The expected value of outcome X by definition is the expected value of the variable from its population. X was taken from the population with p.d.f. $f\left(x; a, \dfrac{1}{\eta}\right)$:

$$\mathsf{E}(X) = \int_{-\infty}^{+\infty} x f\left(x; a, \frac{1}{\eta}\right) dx = \int_{0}^{\infty} \left(\frac{1}{\eta}\right)^a \frac{1}{\Gamma(a)} x^a e^{-\frac{x}{\eta}} dx.$$

This integral looks very similar to the definition of gamma function from the note – $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$. In order to get the exponent without scaling parameter $\dfrac{1}{\eta}$ let's switch to the new variable y:

$$y = \frac{x}{\eta}, \qquad dy = \frac{dx}{\eta}, \qquad (0, \infty)_x \to (0, \infty)_y.$$

Substituting y into the previous integral and putting constants out of integral:

$$\mathsf{E}(X) = \left(\frac{1}{\eta}\right)^a \frac{1}{\Gamma(a)} \int_0^{\infty} \eta^a y^a e^{-y} dy \cdot \eta = \frac{\eta}{\Gamma(a)} \int_0^{\infty} y^a e^{-y} dy = \frac{\eta}{\Gamma(a)} \cdot \Gamma(a+1).$$

Using the gamma function's property of recurrence – $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$:

$$\mathsf{E}(X) = \eta \frac{\Gamma(a+1)}{\Gamma(a)} = \eta a.$$

Since Y is the outcome from identical population, but with parameter b – $f\left(x; b, \dfrac{1}{\eta}\right)$, the result of its expected value will be same, but with respective replacement $a \to b$:

$$\mathsf{E}(Y) = \eta b.$$

Thus, the expected value of the estimator $\widehat{\eta}$:

$$\mathsf{E}\left(\widehat{\eta}\right) = \frac{1}{a+b}\left(\mathsf{E}(X) + \mathsf{E}(Y)\right) = \frac{1}{a+b}\left(\eta a + \eta b\right) = \boxed{\eta}.$$

It means that the estimator $\widehat{\eta}$ is unbiased.

Now let's find the variance of $\widehat{\eta}$. Since variance is the quadratic function, and outcomes X and Y are independent:

$$\mathsf{V}\left(\widehat{\eta}\right) = \mathsf{V}\left(\frac{X+Y}{a+b}\right) = \frac{1}{(a+b)^2}\left(\mathsf{V}(X) + \mathsf{V}(Y)\right).$$

By definition of variance – $\mathsf{V}(X) = \mathsf{E}\left(X^2\right) - \mathsf{E}(X)^2$. The expected value of X^2 is calculated identically to the $\mathsf{E}(X)$:

$$\mathsf{E}\left(X^2\right) = \int\limits_{-\infty}^{+\infty} x^2 f\left(x; a, \frac{1}{\eta}\right) dx = \int\limits_{0}^{\infty} \left(\frac{1}{\eta}\right)^a \frac{1}{\Gamma(a)} x^{a+1} e^{-\frac{x}{\eta}} dx.$$

Substituting $y = \dfrac{x}{\eta}$ gives:

$$\mathsf{E}\left(X^2\right) = \left(\frac{1}{\eta}\right)^a \frac{1}{\Gamma(a)} \int\limits_{0}^{\infty} \eta^{a+1} y^{a+1} e^{-y} dy \cdot \eta = \frac{\eta^2}{\Gamma(a)} \int\limits_{0}^{\infty} y^{a+1} e^{-y} dy = \frac{\eta^2}{\Gamma(a)} \cdot \Gamma(a+2) =$$
$$= \eta^2 (a+1) \frac{\Gamma(a+1)}{\Gamma(a)} = \eta^2 (a+1) a.$$

Thus, the variance $\mathsf{V}(X)$:

$$\mathsf{V}(X) = \mathsf{E}\left(X^2\right) - \mathsf{E}(X)^2 = \eta^2 (a+1) a - (\eta a)^2 = \eta^2 a.$$

Similarly for Y (same explanation as in $\mathsf{E}(Y)$):

$$\mathsf{V}(Y) = \eta^2 b.$$

Finally, the variance of the estimator $\widehat{\eta}$:

$$\mathsf{V}\left(\widehat{\eta}\right) = \frac{1}{(a+b)^2}\left(\mathsf{V}(X) + \mathsf{V}(Y)\right) = \frac{1}{(a+b)^2}\left(\eta^2 a + \eta^2 b\right) = \boxed{\frac{\eta^2}{a+b}}.$$

(c) Let's find out whether the estimator $\widehat{H}$ is unbiased. Using linearity of the expected value:

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

Since $\mathsf{E}\left(\widehat{H}\right) = \eta$, the estimator is unbiased.

The efficiency of estimators is compared via their Mean-Squared Error (MSE), which is by definition: $\text{MSE}\left(\widehat{\theta}\right) = \mathsf{V}\left(\widehat{\theta}\right) + \text{Bias}\left(\widehat{\theta}\right)^2$. Since both estimators ($\widehat{\eta}$ and $\widehat{H}$) have zero bias, we should compare their variances.

Let's find the variance of $\widehat{H}$. Since variance is the quadratic function, and outcomes X and Y are independent:

$$\mathsf{V}\left(\widehat{H}\right) = \mathsf{V}\left(\frac{1}{2}\left(\frac{X}{a} + \frac{Y}{b}\right)\right) = \frac{\mathsf{V}(X)}{4a^2} + \frac{\mathsf{V}(Y)}{4b^2} = \frac{\eta^2 a}{4a^2} + \frac{\eta^2 b}{4b^2} = \frac{\eta^2}{4}\left(\frac{1}{a} + \frac{1}{b}\right).$$

Building the ration of variances gives:

$$\frac{\mathsf{V}\left(\widehat{\eta}\right)}{\mathsf{V}(\widehat{H})} = \frac{\eta^2}{a+b} \cdot \frac{4ab}{\eta^2(a+b)} = \frac{4ab}{(a+b)^2}.$$

Let's compare numerator and denominator of the last ratio:

$$\begin{array}{rcl} (a+b)^2 & \vee & 4ab \\ a^2 + 2ab + b^2 & \vee & 4ab \\ a^2 - 2ab + b^2 & \vee & 0 \\ (a-b)^2 & \geq & 0 \\ & \Downarrow & \\ (a+b)^2 & \geq & 4ab \end{array}$$

It means that
$$\frac{\mathsf{V}\left(\widehat{\eta}\right)}{\mathsf{V}(\widehat{H})} = \frac{4ab}{(a+b)^2} \leq 1 \qquad \Longrightarrow \qquad \mathsf{V}\left(\widehat{\eta}\right) \leq \mathsf{V}(\widehat{H}).$$

Since $4ab = (a+b)^2$ only when $a = b$, estimators $\widehat{\eta}$ and $\widehat{H}$ are identical in the case of $a = b$. When $a \neq b$, the MLE $\widehat{\eta}$ is better than $\widehat{H}$ (in terms of MSE).

Thus, the answer is following:

$$\boxed{\begin{cases} \widehat{\eta} \text{ is more efficient than } \widehat{H}, & a \neq b, \\ \widehat{\eta} \equiv \widehat{H}, & a = b. \end{cases}}$$

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.:

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

Solution:

The likelihood function:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(X_i; \theta) = \prod_{i=1}^{n} \frac{1}{2}e^{-|X_i-\theta|} = \left(\frac{1}{2}\right)^n \exp\left(-\sum_{i=1}^{n} |X_i - \theta|\right).$$

The log-likelihood function:

$$l(\theta) = \ln \mathcal{L}(\theta) = -n\ln 2 - \sum_{i=1}^{n} |X_i - \theta|.$$

Let $g(\theta) = \sum\limits_{i=1}^{n} |X_i - \theta|$. Since $-n\ln 2$ is a constant:

$$\widehat{\theta} = \arg\max_{\theta} l(\theta) = \arg\min_{\theta} g(\theta).$$

Due to a nature of absolute value function, let's explore $g(\theta)$ without differentiation. To arrange the sample, let's use order statistics $X_{(i)}$ instead of common X_i.

- If $n = 1$, $g(\theta) = \left|X_{(1)} - \theta\right|$, see fig. 7. The minimum is clearly in $X_{(1)}$.
- If $n = 2$, $g(\theta) = \left|X_{(1)} - \theta\right| + \left|X_{(2)} - \theta\right|$, see fig. 8. The minimum is in the interval $\left[X_{(1)}, X_{(2)}\right]$.
- If $n = 3$, $g(\theta) = \left|X_{(1)} - \theta\right| + \left|X_{(2)} - \theta\right| + \left|X_{(3)} - \theta\right|$, see fig. 9. The minimum is in $X_{(2)}$.
- ...
- If n is arbitrary, $g(\theta) = \sum\limits_{i=1}^{n} |X_i - \theta|$, see fig. 10. The minimum by induction will be in the median MED of the sample.

Thus, the result:

$$\boxed{\widehat{\theta} = \mathsf{MED}(X_1, \ldots, X_n)}.$$

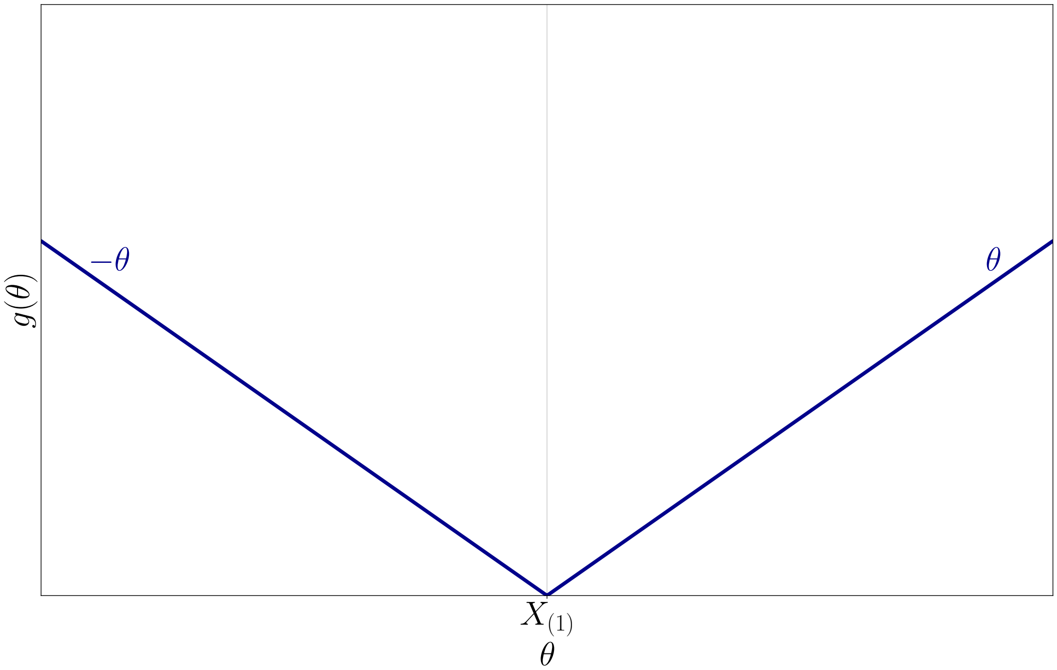


Figure 7: $g(\theta) = \sum_{i=1}^{n} |X_i - \theta|$ for $n = 1$.

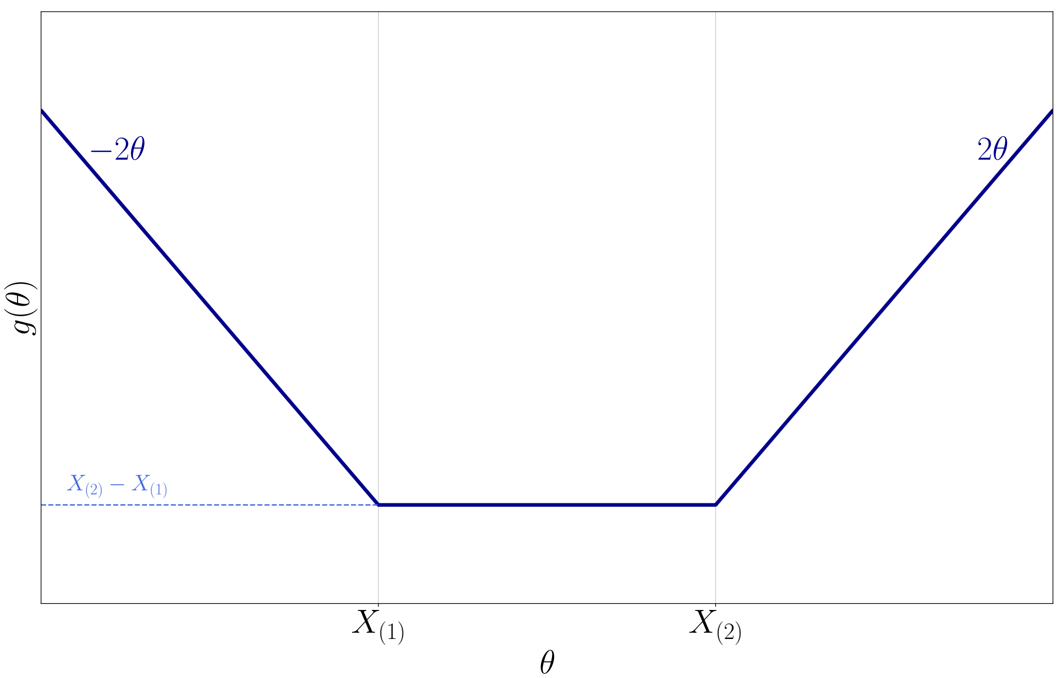


Figure 8: $g(\theta) = \sum_{i=1}^{n} |X_i - \theta|$ for $n = 2$.

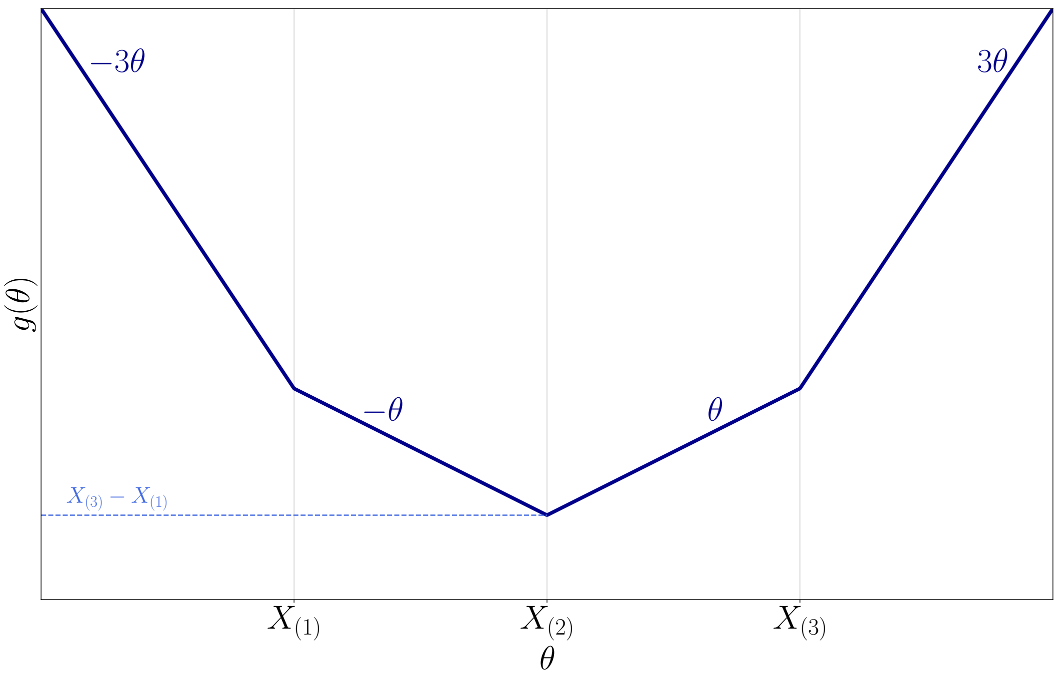


Figure 9: $g(\theta) = \sum_{i=1}^{n} |X_i - \theta|$ for $n = 3$.

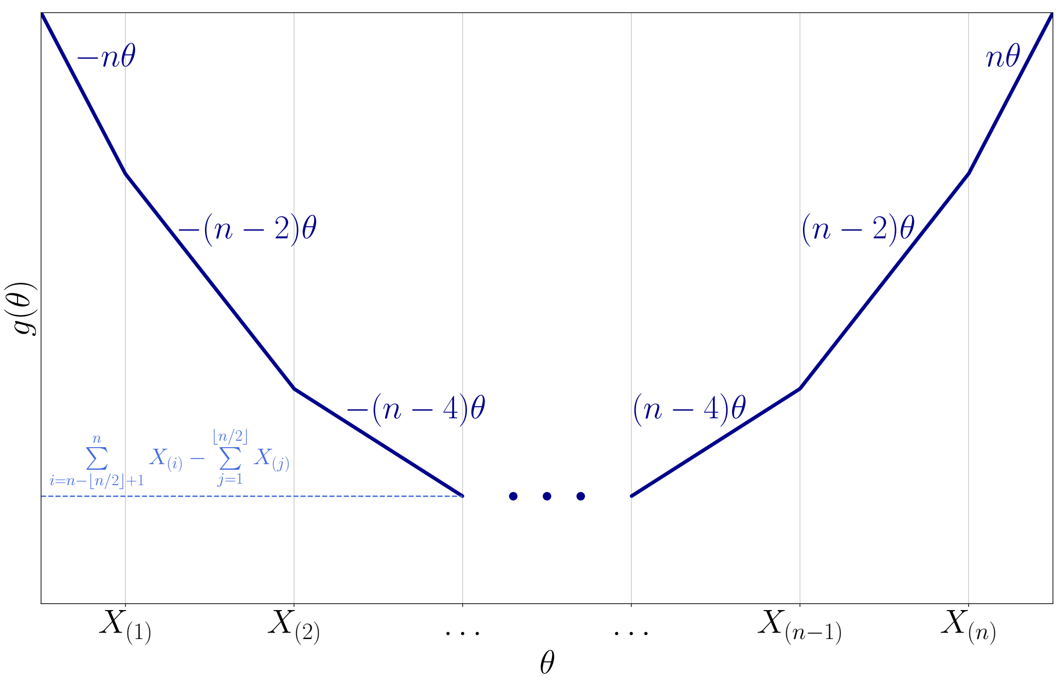


Figure 10: $g(\theta) = \sum_{i=1}^{n} |X_i - \theta|$ for arbitrary n.