Лабораторная работа №2
Выполнил: Кузнецов Павел M3207

Задача 2. Методом моментов найти оценку квадрата масштабирующего параметра θ распределения Лапласа (сдвиг считать нулевым). Найти смещение оценки, дисперсию, среднеквадратическую ошибку. Эксперимент для $\theta = 0.5$.

Решение:

Распределение Лапласа (θ, β) имеет функцию плотности:

$$f(x|\theta, \beta) = \frac{1}{2\theta} exp(-\frac{|x-\beta|}{\theta})$$

, где $\beta = 0$ (сдвиг, по условию).

Воспользуемся методом моментов, приравнивая теоретические характеристики распределения к эмприческим хар-кам выборки:

$$\nu_1 = E(X) \approx \bar{X}_n \text{ (при больших n)}$$
$$\hat{E}(X) = \bar{X}_n \text{ (оценка мат. ожидания методом моментов)}$$

При этом первый момент для нашего распределения Лапласа равен нулю, так как параметр сдвига равен нулю: $E(X) = \beta = 0$

Найдём оценку квадрата параметра через второй момент:

$$\nu_2 = E(X^2)$$
$$E(X^2) = D(X) + E^2(X), E(X) = 0$$
$$E(X^2) = D(X)$$
$$\hat{E}(X^2) = S_n^2 \text{ (оценка равна выборочной дисперсии)}$$

при этом $\nu_2 = 2\theta^2$
Следовательно, $\hat{\theta}^2 = \frac{S_n^2}{2}$

Найти смещение оценки, дисперсию и среднеквадратическую ошибку:

Смещение оценки находится из формулы:

$$bias(\hat{\theta}) = E(\hat{\theta}^2) - \theta^2 = E(\frac{S_n^2}{2}) - \theta^2 = E(\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{2n}) - \theta^2 =$$

$$E(\frac{\sum_{i=1}^{n}(X_i)^2}{2n}) - \theta^2 = \frac{\sum_{i=1}^{n}E(X_i^2)}{2n} - \theta^2 = \frac{nE(X_i^2)}{2n} - \theta^2$$

$$E(X_i^2) = 2\theta^2 \Longrightarrow bias(\hat{\theta}) = 0$$

Оценка несмещена. (тут непонятно, думал на защите спросить: выборочная дисперсия - это оценка для дисперсии, есть смещенный и несмещенный вариант, сверху расписываю по смещенному, следовательно оценка параметра θ автоматически тоже смещенная? Берём тогда несмещенную оценку и доказываем, что оценка смещенная, цепочка тождеств такая же.)

Дисперсия: $var(\hat{\theta}^2) = var(\frac{S_n^2}{2}^2) = var(\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{2n}) =$

$$\frac{var(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2)}{4n^2} = \frac{var(\sum_{i=1}^{n}(X_i)^2)}{4n^2} = \frac{var(X^2)}{4n^2} = \frac{E(X^4) - E^2(X^2)}{4n^2}$$

$$E^2(X^2) = D^2(X) = 4\theta^4$$

$$E(X^4) = \nu_4 = \frac{1}{2}\sum_{k=0}^{4}(\frac{4!}{(4-k)!}\theta^k 0^{4-k}(1 + (-1)^k)) = 24\theta^4$$

$$var(\hat{\theta}^2) = \frac{E(X^4) - E^2(X^2)}{4n^2} = \frac{24\theta^4 - 4\theta^4}{4n^2} = \frac{5\theta^4}{n^2}$$

Среднеквадратическая ошибка:

$$MSE(\hat{\theta}^2) = var(\hat{\theta}^2) = \frac{5\theta^4}{n^2}, (bias(\theta^2) = 0)$$

Проведём эксперимент для $\theta = 0.5$:

```python
import numpy
import matplotlib.pyplot as plt
sizes = [100, 500, 1000, 10000, 100000, 1000000]
for i in sizes:
    sample_array = numpy.random.laplace(0, 0.5, i)
    # we can generate several samples of each size for better demonsration,
    # but that is enough
    sample_mean = numpy.mean(sample_array)
    sample_var = numpy.var(sample_array)
    print ("mean: ", sample_mean, "var: ", sample_var, "with size: ", i)
```

Вывод:

mean: -0.00758196491827923 var: 0.35240663019179036 with size: 100

mean: -0.03537322587695742 var: 0.5813357854759184 with size: 500

mean: -0.03535078067779483 var: 0.5068039470484109 with size: 1000

mean: -0.0015312248937861322 var: 0.5125913151286939 with size: 10000

mean: 0.001784203927182552 var: 0.5045930291531003 with size: 100000

mean: -0.0009742731843855747 var: 0.49967599520005107 with size: 1000000

Видим, что при увеличении n, оценка параметра θ^2 (выборочная дисперсия выборки) стремится к 0.5 (теоритическому значению)

Задача 3. Методом максимального правдоподобия найти оценку параметра θ биномиального распределения $Bin(n, \theta)$, считая n известным. Найти смещение оценки, дисперсию, среднеквадратическую ошибку. Является ли найденная оценка эффективной? Эксперимент при $n = 4, \theta = 1/5$.

Решение:

Биномиальное распределение $Bin(n, \theta)$ имеет функцию плотности:

$$f(x|n, \theta) = P(X = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Функция правдоподобия для X:

$$L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=0}^{N} f(x_i, \theta)$$

Оценим параметр θ методом максимального правдоподобия:

Оценка вычисляется из уравнения максимального правдоподобия, выходящего из равенства:

$$L(X_l, \theta) = (max) \prod_{i=0}^{N} \binom{n}{x_i} \theta^{x_i} (1-\theta)^{n-x_i}$$

Само уравнение:

$$\frac{\partial}{\partial \theta} \sum_{l=1}^{N} lnf(X_l, \theta) = 0$$

$$\sum_{i=1}^{N} \ln(\binom{n}{x_i} \theta^{x_i} (1-\theta)^{n-x_i}) =$$

$$\sum_{i=1}^{N} \ln \binom{n}{x_i} + \ln\theta \sum_{i=1}^{N} x_i + nN\ln(1-\theta) - \ln(1-\theta)\sum_{i=1}^{N} x_i$$

(логарифм произведения равен сумме логарифмов)

$$\frac{\partial}{\partial\theta}\sum_{l=1}^{N} lnf(X_l,\theta) = \frac{\sum_{i=1}^{N} x_i}{\theta} - \frac{nN}{1-\theta} + \frac{\sum_{n=1}^{N} x_i}{1-\theta}$$

Приравняем к нулю, так как производная равна нулю в максимальной точке:

$$\frac{\sum_{i=1}^{N} x_i}{\hat{\theta}} - \frac{nN}{1-\hat{\theta}} + \frac{\sum_{n=1}^{N} x_i}{1-\hat{\theta}} = 0$$

Домножаем на знаменатели:

$$(1-\hat{\theta})\sum_{i=1}^{N} x_i - nN\hat{\theta} + \hat{\theta}\sum_{i=1}^{N}(n - x_i) = 0$$

$$\hat{\theta} = \frac{\sum\limits_{i=1}^{N} x_i}{nN}$$

Смещение оценки:

$$bias(\theta^2) = E(\hat{\theta}) - \theta = E(\frac{\sum\limits_{i=1}^{N} x_i}{nN}) - \theta =$$

$$= \frac{\sum\limits_{i=1}^{N} E(x_i)}{nN} - \theta = \frac{nN\theta}{nN} - \theta = 0$$

, $E(X) = n\theta$, оценка несмещена.

Дисперсия оценки:

$$Var(\hat{\theta}) = Var(\frac{\sum_{i=1}^{N} x_i}{nN}) = \frac{1}{n^2 N} Var(x_i) = \frac{n\theta(1-\theta)}{n^2 N} = \frac{\theta(1-\theta)}{nN}$$

Среднеквадратическая ошибка:

$$MSE(\hat{\theta}) = var(\hat{\theta}) = \frac{\theta(1-\theta)}{nN}, (bias(\theta) = 0)$$

Проверим эффективность оценки: оценка несмещена, проверим выполняется ли рав-во Рао-Крамера (выходящее из нер-ва). Информация Фишера:

$$I(\theta) = E(\tfrac{\partial \ln(P(X=x_i))}{\partial\theta})^2 =$$
$$E(\tfrac{\partial(\ln\binom{n}{x_i} + x_i\ln(\theta) + (n-x_i)\ln(1-\theta))}{\partial\theta})^2 = E(\tfrac{x_i}{\theta} - \tfrac{n-x_i}{1-\theta})^2 =$$
$$E(\tfrac{x_i - n\theta}{\theta(1-\theta)})^2 = \tfrac{E(x_i-n\theta)^2}{\theta^2(1-\theta^2)} = \tfrac{Var(x_i)}{\theta^2((1-\theta)^2)} = \tfrac{n\theta(1-\theta)}{\theta^2((1-\theta)^2)} = \tfrac{n}{\theta(1-\theta)}$$
$$\tfrac{1}{I(\theta)} = var(\hat{\theta}) \Longrightarrow$$

$\Longrightarrow$ оценка эффективна.

Проведём эксперимент для $n = 4, \theta = \frac{1}{5}$:

```
sizes = [100, 500, 1000, 10000, 100000, 1000000]
for i in sizes:
    sample_array = numpy.random.binomial(4, 0.2, size = (100, i))
    sample_mean = numpy.mean(sample_array, axis=1)
    sample_theta_estimation = sample_mean/4
    sample_theta_estimation = numpy.mean(sample_theta_estimation)
    print ("mean of theta estimation for samples of size", i, " = ", sample_theta_estimation)
```

Вывод:

Среднее оценки параметра(эмпирических значений) для выборок размера 100 : – 0.198725

Среднее оценки параметра(эмпирических значений) для выборок размера 500 : – 0.19977

Среднее оценки параметра(эмпирических значений) для выборок размера 1000 : – 0.20036

Среднее оценки параметра(эмпирических значений) для выборок размера 10000 : – 0.20004600000000003

Среднее оценки параметра(эмпирических значений) для выборок размера 100000 : – 0.20001625

Среднее оценки параметра(эмпирических значений) для выборок размера 1000000 : – 0.2000397675

Видим, что при увеличении n, оценка параметра θ (выборчная дисперсия выборки) стремится к 0.2 (теоритическому значению), следовательно является эффективной