Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ "МЭИ"

Институт информационных и вычислительных технологий

Кафедра математического и компьютерного моделирования

Отчёт по лабораторной работе №3 "Оценки"

Студент: Симаков А.М.
Преподаватель: Шевченко О.В.

Москва 2023

1 Постановка задачи оценивания

Пусть (x_n) - выборка, т.е. n независимых испытаний случайной величины X, имеющей функцию распределения $F(x/a)$, зависящую от параметра a, значение которого неизвестно. Требуется оценить значение параметра a.

Оценкой $\widehat{a} = \varphi(x_1, ..., x_n)$ называется функция наблюдений, используемая для приближенного определения неизвестного параметра. Значение оценки является **случайной величиной**, поскольку $(x_1, ..., x_n)$ - случайная величина (многомерная).

Свойства оценок

1) Оценка $\widehat{a} = \varphi(x_1, ..., x_n)$ называется **состоятельной**, если при $n \to \infty$ $\widehat{a} \to a$ по вероятности $\forall a$.
2) Оценка $\widehat{a} = \varphi(x_1, ..., x_n)$ называется **несмещенной**, если $\forall a$ $M\widehat{a} = M\varphi(x_1, ..., x_n) = a$.
3) Оценка φ^* называется **оптимальной**, если для неё средний квадрат ошибки $M[\widehat{a} - a]^2 = M[\varphi^*(x_1, ..., x_n) - a]^2 = \min M[\varphi(x_1, ..., x_n) - a]^2$ минимален среди всех оценок (φ).

Метод моментов

Пусть (ξ_n) - выборка, т.е. n независимых наблюдений над случайной величиной ξ_0. $F(x; a_1, ..., a_r)$ - **функция распределения**, зависящая от неизвестных параметров $a = (a_1, ..., a_r)$. Требуется оценить их.
Идея: *неизвестные параметры выразить через начальные моменты, а затем вместо моментов подставить несмещенные и состоятельные оценки моментов.*

$$m_1 = \int_{-\infty}^{\infty} x dF(x; a_1, ..., a_r) = f_1(a_1, ..., a_r)$$
$$...$$
$$m_r = \int_{-\infty}^{\infty} x^r dF(x; a_1, ..., a_r) = f_r(a_1, ..., a_r)$$

Отсюда

$$a_j = g_j(m_1, ..., m_r), \quad j \in [1, r] \cap \mathbb{N}$$

Подставив вместо моментов $m_1, ..., m_r$ их оценки $\widehat{m_1}, ..., \widehat{m_r}$ получим

$$\widehat{a_j} = g_j(\widehat{m_1}, ..., \widehat{m_r}), \quad \widehat{m_k} = \frac{1}{n}\sum_{i=1}^{n}\xi_i^k, \quad j,k \in [1,r] \cap \mathbb{N}$$

Метод максимального правдоподобия

Пусть (ξ_n) - выборка. $q(x_i, a)$ - **распределение** i-ого наблюдения, $a = (a_1, ..., a_r)$ - неизвестный параметр. $p_\xi(x;a) = \prod_{i=1}^n q(x_i, a)$ - **распределение** выборки $x = (x_n)$. Функция $p_\xi(x;a)$, как функция параметра a, при фиксированном x, называется **функцией правдоподобия**.

Оценкой максимального правдоподобия a^* параметра a называется такое значение, при котором функция правдоподобия $p_\xi(x;a)$ достигает максимума:

$$a^* : p_\xi(x;a) = \max_a p_\xi(x;a)$$

Если максимум достигается во внутренней точке области определения функции, то a^* удовлетворяет системе уравнений:

$$\frac{\partial \log p_\xi(x;a)}{\partial a_i}\Big|_{a=a^*} = 0, \quad i \in [1,r] \cap \mathbb{N}$$

Использование логарифма не изменяет точки максимума, но упрощает выкладки при независимых наблюдениях. Оценка $a^* = a^*(x)$ является функцией наблюдений x. Чтобы подчеркнуть случайность аргумента, напишем $a^*(\xi)$.

Метод порядковых статистик

Метод основан на оценках ζ_p при разных p. Пусть (ξ_n) - выборка с $F(x;a)$, зависящей от параметра a, значение которого требуется оценить. Выберем p так, чтобы квантиль x_p зависела от параметра:

$$x_p = f(a)$$

Тогда получим

$$a = g(x_p)$$

Вместо x_p подставим выборочную квантиль $\zeta_p = \xi_{[np]+1}$, в результате чего получим состоятельную оценку

$$\widehat{a} = g(\xi_{[np]+1})$$

Таким же образом можно построить оценки и для многомерного параметра.

Статистическая задача

По выборке (x_n) для случайной величины $X \sim \mathcal{R}[0, a]$ оценить параметр a. Сравним три оценки

Метод моментов

$$\widehat{a_1} = \frac{2}{n}\sum_{i=1}^{n} x_i$$

Метод максимального правдоподобия (исправлена смещённость)

$$\widehat{a_2} = \frac{n+1}{n}\max x_i$$

Метод порядковых статистик

$$\widehat{a_3} = 2 \cdot \widehat{x}_{0.5} = x_k + x_{k+1},$$

где $\widehat{x}_{0.5} = \frac{x_k + x_{k+1}}{2}$- выборочная квантиль порядка 0.5, т.е. **выборочная медиана**. x_k - член вариационного ряда (полагаем $n = 2k$).

Точность этих оценок можно сравнить теоретически и экспериментально (статистически).

Все три оценки несмещённые, что можно проверить методами теории вероятностей. Определим дисперсии оценок.

$$D\widehat{a_1} = D\frac{2}{n}\sum_{i=1}^{n} x_i = \frac{4}{n^2}\sum_{i=1}^{n} Dx_i = \frac{4}{n^2}nDx_i = \frac{4na^2}{12n^2} = \frac{a^2}{3n}$$

Определим функцию распределения статистики $\max x_i$

$$F(t) = P\left\{\max x_i < t\right\} = \prod_{i=1}^{n} P\left\{x_i < t\right\} = \left(\frac{t}{a}\right)^n, \quad t \in [0, a]$$

Тогда плотность распределения

$$p(t) = \frac{dF}{dt} = \frac{n}{a}\left(\frac{t}{a}\right)^{n-1}, \quad t \in [0, a]$$

Далее

$$M\widehat{a_2} = M\left[\frac{n+1}{n}\max x_i\right] = \frac{n+1}{n}\int_0^a t\frac{n}{a}\left(\frac{t}{a}\right)^{n-1} dt = a$$

$$M\widehat{a_2}^2 = M\left[\frac{n+1}{n}\max x_i\right]^2 = \left(\frac{n+1}{n}\right)^2\int_0^a t^2\frac{n}{a}\left(\frac{t}{a}\right)^{n-1} dt = \left(\frac{n+1}{n}\right)^2\frac{n}{n+2}a^2$$

Тогда

$$D\widehat{a_2} = M\widehat{a_2}^2 - M^2\widehat{a_2} = \frac{a^2}{n(n+1)}$$

Для вычисления $D\widehat{a_3}$ используем теорему Крамера, которая говорит, что выборочная p-квантиль имеет дисперсию, приблизительно равную $\frac{1}{n}\frac{p(1-p)}{f^2(x_p)}$, где x_p - истинная p-квантиль, $f(x)$ - плотность распределения выборки. При $n = 2k$ статистика

$$\frac{x_k + x_{k+1}}{2} \equiv m$$

является выборочной медианой ($p = 0.5$), $f(x_{0.5}) = 1/a$, $\widehat{a_3} = 2m$. Следовательно,

$$D\widehat{a_3} = Dm = 4\frac{1}{n}\frac{\frac{1}{2}\frac{1}{2}}{\left(\frac{1}{a}\right)^2} = \frac{a^2}{n}$$

Имеем

$$D\widehat{a_1} = \frac{a^2}{3n}, \quad D\widehat{a_2} = \frac{a^2}{n(n+1)}, \quad D\widehat{a_3} = \frac{a^2}{n}$$

Следовательно, $\widehat{a_2}$ - наиболее точная оценка, $\widehat{a_3}$ - наименее.

Статистическое сравнение оценок

Далеко не всегда удается аналитически вычислить дисперсию оценки. Как экспериментально определить, какой из оценок пользоваться?

Характеристиками разброса значений (a_k) оценки $\widehat{a}$ будем считать размах

$$w = \max a_i - \min a_i$$

и среднеквадратичное отклонение (СКО)

$$S_a = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(a_i - \overline{a})^2}, \qquad \overline{a} = \frac{1}{k}\sum_{i=1}^{k} a_i$$

2 Результаты экспериментов

Приведу характеристики оценок $\widehat{a_1}$, $\widehat{a_2}$, $\widehat{a_3}$ для выборки $(x_n) : x_n \sim \mathcal{R}[0,1]$ разного объёма ($n \in \{10, 40, 160\}$) в виде таблиц.
Таблица для $\widehat{a_1}$

n	СКО	MIN	MAX	Разброс
10	0.162	0.598	1.226	0.628
40	0.121	0.725	1.207	0.482
160	0.042	0.941	1.117	0.176

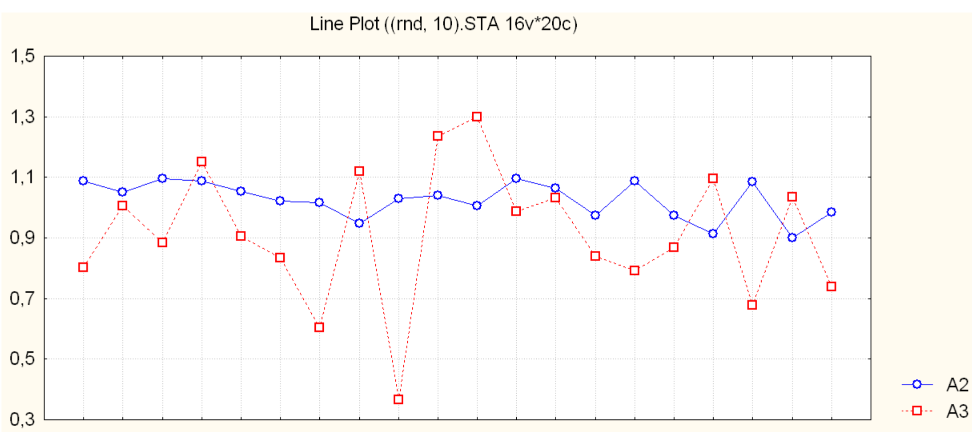
Таблица для $\widehat{a_2}$

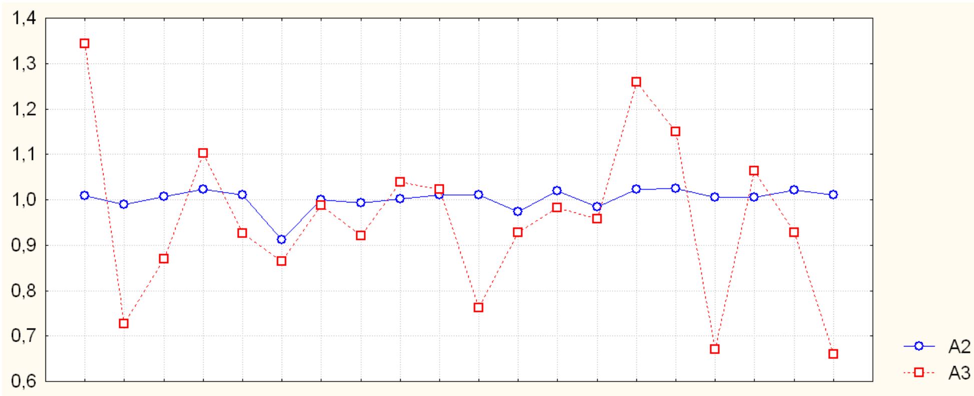
n	СКО	MIN	MAX	Разброс
10	0.061	0.901	1.097	0.196
40	0.025	0.912	1.025	0.113
160	0.006	0.984	1.006	0.022

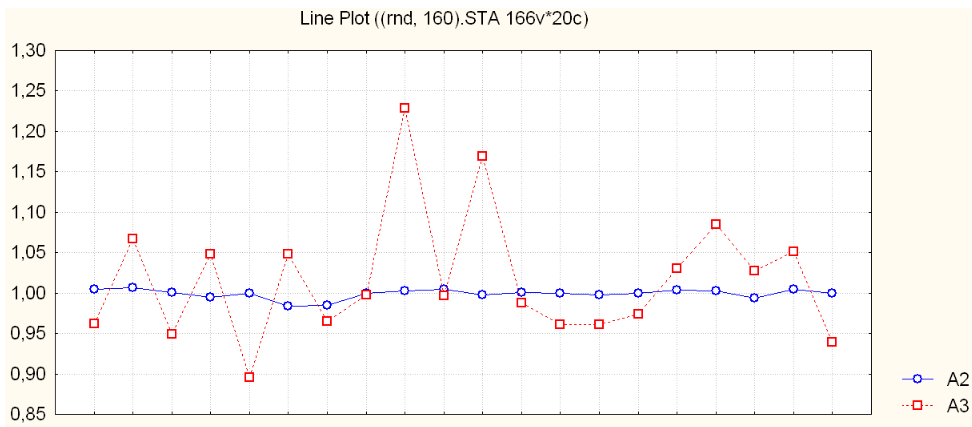
Таблица для $\widehat{a_3}$

n	СКО	MIN	MAX	Разброс
10	0.224	0.366	1.299	0.933
40	0.178	0.661	1.344	0.683
160	0.079	0.896	1.228	0.332

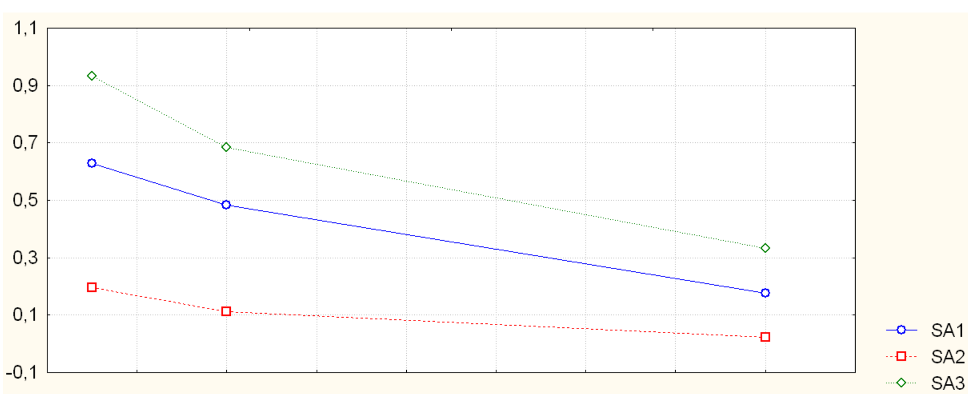
По таблицам можно видеть, что $\widehat{a_2}$ - наиболее точная оценка, $\widehat{a_3}$ - наименее. Также приведу графики.







По графикам видно, что у $\widehat{a_3}$ разброс больше, чем у $\widehat{a_2}$. Графически сравним СКО трёх наших оценок.



Очевидна правильность наших предыдущих слов.

3 Задание для самостоятельной работы

Сравнить статистически на выборках объема $n = 10$ две оценки: оценку максимального правдоподобия и медианную оценку
1) среднего нормального распределения
2) параметра показательного распределения.

Оценка среднего нормального распределения
Оценка максимального правдоподобия

$$\widehat{a_2} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Медианная оценка

$$\widehat{a_3} = \widehat{x}_{0.5} = \frac{x_k + x_{k+1}}{2}$$

Рассмотрим 20 выборок размера 10, распределенных по закону $\mathcal{N}(0,1)$. Получим следующие характеристики оценок:

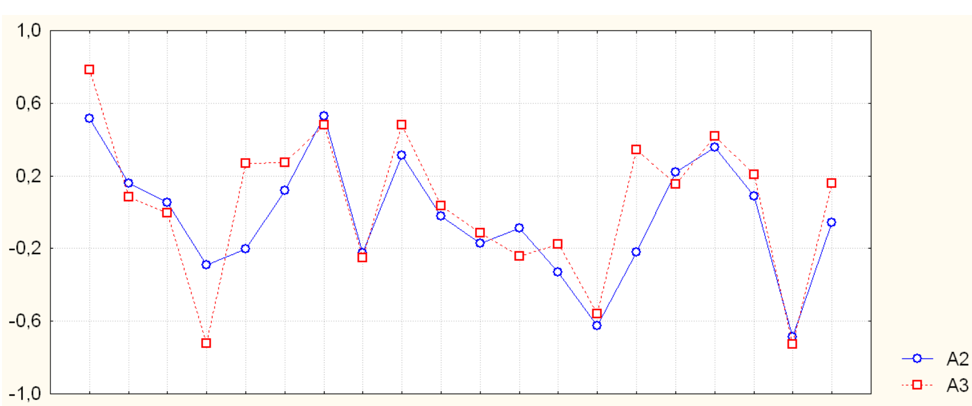
Таблица для $\widehat{a_2}$

n	СКО	MIN	MAX	Разброс
10	0.332	-0.689	0.527	1.216

Таблица для $\widehat{a_3}$

n	СКО	MIN	MAX	Разброс
10	0.404	-0.725	0.783	1.508

Также приведу график



Сравнивая СКО и разбросы делаем вывод, что оценка $\widehat{a_2}$ более точная, чем $\widehat{a_3}$.

Оценка параметра показательного распределения

Оценка максимального правдоподобия

$$\widehat{a_2} = \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^{-1}$$

Медианная оценка

$$\widehat{a_3} = \frac{\log 2}{x_k + x_{k+1}}$$

Рассмотрим 20 выборок размера 10, распределенных по закону $\mathcal{E}(5)$. Получим следующие характеристики оценок:

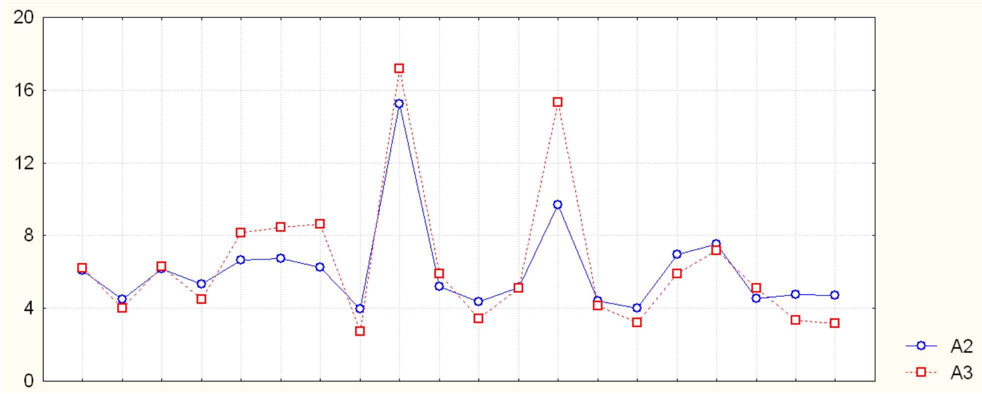
Таблица для $\widehat{a_2}$

n	СКО	MIN	MAX	Разброс
10	2.574	3.947	15.251	11.304

Таблица для $\widehat{a_3}$

n	СКО	MIN	MAX	Разброс
10	3.834	2.724	17.172	14.448

Также приведу график



Сравнивая СКО и разбросы делаем вывод, что оценка $\widehat{a_2}$ более точная, чем $\widehat{a_3}$.

4 Конкретная индивидуальная задача

Количество производственных травм за время T является случайной величиной с распределением $\mathcal{P}(\lambda T)$. Поквартальные данные за k лет оказались (x_n), $n = 4k$. Если за год число травм оказывается равным N и более, завод подвергается штрафу и проверке. Оценить вероятность этого события в следующем году. Поквартальные данные получить моделированием; среднее значение принять m.

$$k = 5, \quad N = 40, \quad m = 8, \quad n = 4k = 20$$

1 год = 4 квартала Так как мы рассматриваем количество травм за год, то $T = 4$.

Решение

По наблюдениям (x_n) $(n = 20)$ над случайной величиной X с распределением $\mathcal{P}(\lambda T)$ оценить вероятность

$$P\left\{\xi \geq N\right\} = 1 - P\left\{\xi < N\right\} = 1 - P\left\{\xi < 40\right\}$$

Так как ξ- случайная величина с распределением $\mathcal{P}(4\lambda)$, то

$$P\left\{\xi < 40\right\} = \sum_{i=1}^{40} \frac{(4\lambda)^k}{k!} e^{-4\lambda}$$

Построим оценку максимального правдоподобия для параметра λ. Для величины с распределением $\mathcal{P}(\lambda)$ имеем

$$P\left\{x = k\right\} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Тогда для выборки получим

$$p_\xi(x;\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-\lambda n}$$

Найдём максимум функции правдоподобия:

$$\log\left[\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-\lambda n}\right] = -\lambda n + \log\lambda \sum_{i=1}^{n} x_i - \log\left[\prod_{i=1}^{n} x_i!\right]$$

$$\implies \frac{d}{d\lambda}\left(-\lambda n + \log\lambda \sum_{i=1}^{n} x_i - \log\left[\prod_{i=1}^{n} x_i!\right]\right) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

$$\implies \widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Получим значение $\widehat{\lambda}$ по исходным данным

$$X = (4, 4, 13, 7, 2, 9, 7, 5, 8, 10, 2, 5, 9, 11, 6, 6, 7, 8, 12, 9) \implies \widehat{\lambda} = \frac{144}{20} = 7.2$$

Тогда получим следующую оценку для вероятности

$$\widehat{p} = P\left\{\xi \geq 40\right\} = 1 - \sum_{i=1}^{40} \frac{(4 \cdot 7.2)^k}{k!} e^{-4 \cdot 7.2} \approx 0.028$$

В случае, если бы мы искали вероятность по известному значению $m = 8$, то получили бы следующее значение

$$\tilde{p} = P\left\{\xi \geq 40\right\} = 1 - \sum_{i=1}^{40} \frac{(4 \cdot 8)^k}{k!} e^{-4 \cdot 8} \approx 0.096$$