Национальный исследовательский университет
Высшая школа экономики
Московский институт электроники и математики

Департамент прикладной математики
кафедра компьютерной безопасности

Домашнее задание №3 по математической статистике Построение точечных оценок параметра распределения

Дискретное распределение: *дискретное равномерное I*
Неизвестный параметр: $\theta = 121$
Непрерывное распределение: *распределение Парето*
Неизвестный параметр: $\theta = 12$

Выполнила
Мазитова Е.А.

Проверил
Богданов Д.С.

Москва 2025

Содержание

1. Получение оценок методом моментов и методом максимального прав доподобия

Дискретное распределение (дискретное равномерное I)

$$P(x) = \theta^{-1}, \quad x \in \{1, .., \theta\}$$

Метод моментов

Математическое ожидание было найдено в ДЗ1:

$$E[X] = \sum_{x=1}^{\theta} x \cdot \theta^{-1} = \frac{1}{\theta}\sum_{x=1}^{\theta} x = \frac{1}{\theta} \cdot \frac{\theta(\theta+1)}{2} = \frac{\theta+1}{2}$$

Приравниваем теоретический первый момент к выборочному:

$$E[X] = \overline{X}$$

$$\frac{\theta+1}{2} = \overline{X}$$

$$\theta + 1 = 2\overline{X}$$

Получили:

$$\hat{\theta}_{\text{ММ}} = 2\overline{X} - 1$$

Метод максимального правдоподобия

Функция правдоподобия для выборки $x_1, \ldots, x_n$:

$$L(\theta) = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} \frac{1}{\theta} = \theta^{-n},$$

при условии, что все $x_i \in \{1, 2, \ldots, \theta\}$

Это условие выполняется тогда и только тогда, когда

$$\max_{1 \le i \le n} x_i \le \theta$$

Таким образом,

$$L(\theta) = \begin{cases} \theta^{-n}, & \theta \ge \max\limits_i x_i, \\ 0, & \theta < \max\limits_i x_i. \end{cases}$$

Функция θ^{-n} убывает по θ, поэтому её максимум на области $\theta \ge \max x_i$ достигается при наименьшем возможном θ, то есть при

$$\theta = \max_{1 \le i \le n} x_i.$$

Следовательно, оценка максимального правдоподобия:

$$\hat{\theta}_{\text{ММП}} = \max_{1 \le i \le n} x_i.$$

Непрерывное распределение (распределение Парето)

$$f(x) = \theta \cdot x^{-(\theta+1)}, \quad x \in [1, +\infty), \quad \theta > 0$$

Метод моментов

Математическое ожидание было найдено в ДЗ1:

$$E[X] = \frac{\theta}{\theta - 1}$$

Приравниваем теоретический первый момент к выборочному:

$$E[X] = \overline{X}$$

$$\frac{\theta}{\theta - 1} = \overline{X}$$

$$\theta = \overline{X}(\theta - 1)$$

$$\theta = \overline{X}\theta - \overline{X}$$

$$\overline{X}\theta - \theta = \overline{X}$$

$$\theta(\overline{X} - 1) = \overline{X}$$

Получили:

$$\hat{\theta}_{\text{ММ}} = \frac{\overline{X}}{\overline{X} - 1}$$

Метод максимального правдоподобия

Функция правдоподобия для выборки $x_1, x_2, \ldots, x_n$:

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \theta x_i^{-(\theta+1)}.$$

$$L(\theta) = \theta^n \cdot \left(\prod_{i=1}^{n} x_i \right)^{-(\theta+1)}.$$

$$\ln L(\theta) = \ln \left[\theta^n \cdot \left(\prod_{i=1}^{n} x_i \right)^{-(\theta+1)} \right].$$

$$\ln L(\theta) = n \ln \theta - (\theta + 1) \ln \left(\prod_{i=1}^{n} x_i \right).$$

$$\ln\left(\prod_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \ln x_i.$$

$$\ln L(\theta) = n\ln\theta - (\theta+1)\sum_{i=1}^{n} \ln x_i.$$

Дифференцируем по θ:

$$\frac{d}{d\theta}\ln L(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i.$$

Приравниваем к нулю:

$$\frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i = 0.$$

$$\frac{n}{\theta} = \sum_{i=1}^{n} \ln x_i,$$

$$\theta = \frac{n}{\sum\limits_{i=1}^{n} \ln x_i}.$$

Получили:

$$\hat{\theta}_{\text{ММП}} = \frac{n}{\sum\limits_{i=1}^{n} \ln x_i}$$

Для каждой выборки, сгенерированной в пункте 2.1, приведем значения полученных оценок. Значения ниже были вычислены на основе сгенерированных в ДЗ2 выборок с помощью написания программы на языке программирования Python (см. приложение):

Данное значение в моем варианте для дискретного распределения: $\theta = 121$

Данное значение в моем варианте для распределения Парето: $\theta = 12$

Таблица 1: Оценки $\hat{\theta}_{\text{ММ}}$ и $\hat{\theta}_{\text{ММП}}$ для дискретного равномерного распределения

n	Серия 1	Серия 2	Серия 3	Серия 4	Серия 5	Среднее
Метод моментов ($\hat{\theta}_{\text{ММ}}$)						
5	142.600	176.600	102.600	148.200	108.200	135.640
10	122.200	153.000	87.400	127.600	98.800	117.800
100	134.280	128.960	109.420	117.040	114.620	120.864
200	130.780	124.830	116.060	115.580	121.770	121.804
400	130.035	119.895	115.790	117.790	121.110	120.924
600	126.470	120.533	118.047	118.963	122.623	121.327
800	124.763	118.478	119.473	119.883	124.135	121.346
1000	123.310	119.080	119.918	119.702	124.572	121.316
Метод максимального правдоподобия ($\hat{\theta}_{\text{ММП}}$)						
5	113	116	103	106	120	111.600
10	120	116	103	106	120	113.000
100	121	121	120	121	121	120.800
200	121	121	121	121	121	121.000
400	121	121	121	121	121	121.000
600	121	121	121	121	121	121.000
800	121	121	121	121	121	121.000
1000	121	121	121	121	121	121.000

Таблица 2: Оценки $\hat{\theta}_{\text{ММ}}$ и $\hat{\theta}_{\text{ММП}}$ для распределения Парето

n	Серия 1	Серия 2	Серия 3	Серия 4	Серия 5	Среднее
Метод моментов ($\hat{\theta}_{\text{ММ}}$)						
5	11.34785	7.20198	9.80484	9.64941	10.84038	9.76889
10	12.29277	9.39223	9.68814	10.73614	13.42842	11.10754
100	12.22700	13.58764	12.53103	11.28500	12.77170	12.48048
200	11.65621	13.52800	11.68906	11.59474	13.56687	12.40697
400	11.23749	12.49350	12.12724	11.43871	12.84091	12.02757
600	11.49692	12.03521	12.09116	11.44901	12.70387	11.95523
800	11.55334	12.05871	12.14945	11.69274	12.52485	11.99582
1000	11.74966	12.16773	11.86630	11.77453	12.56133	12.02391
Метод максимального правдоподобия ($\hat{\theta}_{\text{ММП}}$)						
5	11.43898	6.86817	9.42499	9.75338	10.75380	9.64786
10	12.21969	9.12632	9.50079	10.80638	13.52990	11.03662
100	12.25889	13.58282	12.57037	11.36169	12.68442	12.49164
200	11.63142	13.47814	11.62652	11.66821	13.50097	12.38105
400	11.22229	12.47301	12.06763	11.43875	12.83756	12.00785
600	11.50300	12.00316	12.02775	11.42588	12.67077	11.92611
800	11.56329	12.04773	12.10698	11.67834	12.50793	11.98085
1000	11.76056	12.15582	11.83473	11.75980	12.52434	12.00705

2. Поиск оптимальных оценок

Дискретное равномерное распределение

Для дискретного равномерного распределения $P_\theta(x) = \theta^{-1}, \quad x \in \{1, 2, \ldots, \theta\}$, где $\theta \in \mathbb{N}$ — неизвестный параметр, построим оптимальную оценку методом, основанным на полной достаточной статистике.

1. Достаточная статистика. Функция правдоподобия для выборки $\mathbf{X} = (X_1, \ldots, X_n)$:

$$L(\mathbf{x}; \theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{I}_{\{1 \le x_i \le \theta\}} = \theta^{-n}\, \mathbb{I}_{\{x_{(n)} \le \theta\}}\, \mathbb{I}_{\{x_{(1)} \ge 1\}},$$

где $x_{(n)} = \max\{x_1, \ldots, x_n\}$. По критерию факторизации Неймана–Фишера статистика $T(\mathbf{X}) = X_{(n)}$ является достаточной для θ.

2. Полнота статистики $X_{(n)}$. Распределение максимального значения выборки:

$$P_\theta\big(X_{(n)} = k\big) = \frac{k^{\,n} - (k-1)^{\,n}}{\theta^{\,n}}, \qquad k = 1, 2, \ldots, \theta.$$

Пусть $g : \{1, 2, \ldots\} \to \mathbb{R}$ такова, что $E_\theta\big[g(X_{(n)})\big] = 0$ для всех $\theta \in \mathbb{N}$. Тогда

$$\sum_{k=1}^{\theta} g(k)\big[k^{\,n} - (k-1)^{\,n}\big] = 0 \qquad \forall \theta \in \mathbb{N}.$$

Последовательно подставляя $\theta = 1, 2, \ldots$, получаем $g(1) = 0$, затем $g(2) = 0$ и т.д. Следовательно, $g(k) \equiv 0$, (доказывается по математической индукции) что означает полноту статистики $X_{(n)}$.

3. Построение оптимальной оценки.

Поскольку $X_{(n)}$ — полная достаточная статистика, оптимальная оценка (UMVUE) для θ имеет вид $\hat{\theta} = h(X_{(n)})$, где функция h удовлетворяет условию несмещённости:

$$E_\theta[h(X_{(n)})] = \theta \quad \forall \theta \in \mathbb{N}.$$

$$\sum_{k=1}^{\theta} h(k) \frac{k^n - (k-1)^n}{\theta^n} = \theta.$$

Умножим на θ^n:

$$\sum_{k=1}^{\theta} h(k)[k^n - (k-1)^n] = \theta^{n+1} \quad \forall \theta.$$

Запишем для $\theta = m$ и $\theta = m + 1$:

$$\sum_{k=1}^{m} h(k)[k^n - (k-1)^n] = m^{n+1},$$

$$\sum_{k=1}^{m+1} h(k)[k^n - (k-1)^n] = (m+1)^{n+1}.$$

$$h(m+1)[(m+1)^n - m^n] = (m+1)^{n+1} - m^{n+1}.$$

$$h(m+1) = \frac{(m+1)^{n+1} - m^{n+1}}{(m+1)^n - m^n}.$$

$$h(k) = \frac{k^{n+1} - (k-1)^{n+1}}{k^n - (k-1)^n}.$$

Следовательно, точная оптимальная несмещённая оценка параметра θ имеет вид

$$\boxed{\hat{\theta}_{\text{opt}} = \frac{X_{(n)}^{n+1} - (X_{(n)} - 1)^{n+1}}{X_{(n)}^n - (X_{(n)} - 1)^n}}.$$

Распределение Парето

Рассмотрим распределение Парето с плотностью

$$f_\theta(x) = \theta x^{-(\theta+1)}, \quad x \geq 1,\ \theta > 0,$$

где θ — неизвестный параметр. Имеем выборку $X_1, X_2, \ldots, X_n$ из этого распределения.

1. Достаточная статистика.

Функция правдоподобия:

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \theta x_i^{-(\theta+1)} = \theta^n \exp\left\{-(\theta+1)\sum_{i=1}^{n} \ln x_i\right\}.$$

Перепишем в виде:

$$L(x_1, \ldots, x_n; \theta) = \exp\left\{n \ln\theta - \theta \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} \ln x_i\right\}.$$

Обозначим $T = \sum_{i=1}^{n} \ln X_i$. Тогда

$$L(x_1, \ldots, x_n; \theta) = \underbrace{\exp\{n\ln\theta - \theta T\}}_{g(T;\theta)} \cdot \underbrace{\exp\{-T\}}_{h(x_1,\ldots,x_n)}.$$

По критерию факторизации статистика $T = \sum_{i=1}^{n} \ln X_i$ является достаточной для θ.

2. Полнота статистики T.

Представим совместную плотность выборки в канонической форме экспоненциального семейства:

$$L(x_1, \dots, x_n; \theta) = \exp\{-\theta T + n \ln \theta\} \cdot \exp\{-T\}.$$

Это соответствует общей форме:

$$f(\mathbf{x}; \theta) = \exp\{A(\theta)B(\mathbf{x}) + C(\theta) + D(\mathbf{x})\},$$

где $A(\theta) = -\theta$, $B(\mathbf{x}) = T$, $C(\theta) = n \ln \theta$, $D(\mathbf{x}) = -T$.

Так как функция $A(\theta) = -\theta$ непрерывна и непостоянна на $\Theta = (0, \infty)$, и множество Θ содержит интервал, то по теореме о полноте экспоненциальных семейств статистика T является полной.

3. Построение оптимальной оценки.

Так как T — полная достаточная статистика, оптимальная оценка (UMVUE) для θ имеет вид $\hat{\theta} = h(T)$, где функция h удовлетворяет условию несмещённости:

$$E_\theta[h(T)] = \theta \quad \text{для всех } \theta > 0.$$

Плотность распределения T:

$$f_T(t; \theta) = \frac{\theta^n}{\Gamma(n)} t^{n-1} e^{-\theta t}, \quad t > 0.$$

$$\int_0^\infty h(t) \frac{\theta^n}{\Gamma(n)} t^{n-1} e^{-\theta t} \, dt = \theta.$$

$$\int_0^\infty h(t) t^{n-1} e^{-\theta t} \, dt = \frac{\Gamma(n)}{\theta^{n-1}}.$$

Теперь продифференцируем обе части равенства по θ слева и справа. Левая часть после дифференцирования под знаком интеграла:

$$\frac{d}{d\theta} \int_0^\infty h(t) t^{n-1} e^{-\theta t} \, dt = -\int_0^\infty h(t) t^n e^{-\theta t} \, dt.$$

Правая часть:

$$\frac{d}{d\theta} \left(\frac{\Gamma(n)}{\theta^{n-1}} \right) = -(n-1) \frac{\Gamma(n)}{\theta^n}.$$

$$-\int_0^\infty h(t) t^n e^{-\theta t} \, dt = -(n-1) \frac{\Gamma(n)}{\theta^n}.$$

$$\int_0^\infty h(t) t^n e^{-\theta t} \, dt = (n-1) \frac{\Gamma(n)}{\theta^n}.$$

Заметим, что правая часть равна $(n-1)$ умноженное на преобразование Лапласа функции $t^{n-1}/\Gamma(n)$ с параметром θ:

$$(n-1)\frac{\Gamma(n)}{\theta^n} = (n-1)\int_0^\infty \frac{t^{n-1}}{\Gamma(n)} e^{-\theta t}\, dt.$$

Таким образом:

$$\int_0^\infty h(t) t^n e^{-\theta t}\, dt = \int_0^\infty (n-1)\frac{t^{n-1}}{\Gamma(n)} e^{-\theta t}\, dt \cdot \Gamma(n) = \int_0^\infty (n-1) t^{n-1} e^{-\theta t}\, dt.$$

Полученное равенство интегралов должно выполняться для всех $\theta > 0$. Это возможно только если подынтегральные функции совпадают почти всюду.

Поэтому:

$$h(t)t^n = (n-1)t^{n-1} \quad \text{для почти всех } t > 0.$$

Отсюда:

$$h(t) = \frac{n-1}{t}.$$

Таким образом, оптимальная несмещённая оценка параметра θ:

$$\boxed{\hat{\theta}_{\text{opt}} = \frac{n-1}{\sum_{i=1}^n \ln X_i}}.$$

Для каждой выборки, сгенерированной в пункте 2.1, приведем значения полученных оптимальных оценок. Значения ниже были вычислены на основе сгенерированных в ДЗ2 выборок с помощью написания программы на языке программирования Python (см. приложение):

Таблица 3: Оптимальные оценки $\hat{\theta}_{\text{opt}}$ для дискретного равномерного распределения

n	Серия 1	Серия 2	Серия 3	Серия 4	Серия 5	Среднее
5	135.003555	138.603463	123.003902	126.603791	143.403347	133.323612
10	132.000000	127.600000	113.300000	116.600000	132.000000	124.300000
100	122.210000	122.210000	121.200000	122.210000	122.210000	122.008000
200	121.605000	121.605000	121.605000	121.605000	121.605000	121.605000
400	121.302500	121.302500	121.302500	121.302500	121.302500	121.302500
600	121.201667	121.201667	121.201667	121.201667	121.201667	121.201667
800	121.151250	121.151250	121.151250	121.151250	121.151250	121.151250
1000	121.121000	121.121000	121.121000	121.121000	121.121000	121.121000

Данные значения в моем варианте: для дискретного равномерного распределения $\theta = 121$, для распределения Парето $\theta = 12$.

Таблица 4: Оптимальные оценки $\hat{\theta}_{\text{opt}}$ для распределения Парето

n	Серия 1	Серия 2	Серия 3	Серия 4	Серия 5	Среднее
5	9.151183	5.494533	7.539992	7.802704	8.603039	7.718290
10	10.997720	8.213690	8.550712	9.725740	12.176909	9.932954
100	12.136304	13.446987	12.444669	11.248069	12.557579	12.366721
200	11.573264	13.410753	11.568388	11.609865	13.433464	12.319147
400	11.194230	12.441824	12.037457	11.410155	12.805465	11.977826
600	11.483823	11.983150	12.007700	11.406833	12.649654	11.906232
800	11.548839	12.032667	12.091841	11.663742	12.492299	11.965877
1000	11.748797	12.143668	11.822900	11.748035	12.511820	11.995044

3. Приложения

1. https://github.com/faisvire/mathstat-hw