Математическое моделирование

Mathematical modeling

УДК 519.224.22
https://doi.org/10.32362/2500-316X-2021-9-3-88-97

НАУЧНАЯ СТАТЬЯ

Достаточная статистика для параметра распределения Парето

И.С. Пулькин [@], **А.В. Татаринцев** [@]

МИРЭА – Российский технологический университет, Москва, 119454 Россия
[@] *Авторы для переписки, e-mail: pulkin@mirea.ru, tatarintsev@mirea.ru*

Резюме. Актуальной является задача оценки параметров распределения Парето, в первую очередь, показателя этого распределения, по заданной выборке. В настоящей статье устанавливается, что для этой оценки достаточно знать значение произведения элементов выборки. Доказано, что это произведение является достаточной статистикой для показателя распределения Парето. На основании метода максимального правдоподобия вычислена оценка показателя степени распределения. Доказано, что эта оценка – смещенная, и обоснована формула, устраняющая смещение. Для произведения элементов выборки, рассматриваемого как случайная величина, найдены функция распределения, плотность вероятности, вычислены математическое ожидание, старшие моменты и дифференциальная энтропия. Построены соответствующие графики. Кроме того, отмечается, что достаточной статистикой является любая функция от этого произведения, в частности, среднее геометрическое. Для среднего геометрического, также рассматриваемого как случайная величина, найдены функция распределения, плотность вероятностей, также вычислены математическое ожидание, старшие моменты и дифференциальная энтропия и построены соответствующие графики. Кроме того, обосновано то, что среднее геометрическое выборки является более удобной достаточной статистикой с практической точки зрения, чем произведение элементов выборки. Также, на основании теоремы Рао – Блекуэлла – Колмогорова построены эффективные оценки параметра распределения Парето. В заключение в качестве примера развитая здесь техника применена к показательному распределению. Для него показано, что в качестве достаточной статистики для оценки неизвестного параметра этого распределения могут быть использованы как сумма, так и среднее арифметическое выборки.

Ключевые слова: распределение Парето, достаточная статистика, эффективная оценка, функция распределения, моменты

• **Поступила:** 20.01.2021 • **Доработана:** 27.01.2021 • **Принята к опубликованию:** 05.04.2021

Для цитирования: Пулькин И.С., Татаринцев А.В. Достаточная статистика для параметра распределения Парето. *Российский технологический журнал*. 2021;9(3):88–97. https://doi.org/10.32362/2500-316X-2021-9-3-88-97

Прозрачность финансовой деятельности: Никто из авторов не имеет финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

RESEARCH ARTICLE

Sufficient statistics for the Pareto distribution parameter

Igor S. Pulkin [@], **Andrey V. Tatarintsev** [@]

MIREA – Russian Technological University, Moscow, 119454 Russia
[@] *Corresponding authors, e-mail: pulkin@mirea.ru, tatarintsev@mirea.ru*

Abstract. The task of estimating the parameters of the Pareto distribution, first of all, of an indicator of this distribution for a given sample, is relevant. This article establishes that for this estimate, it is sufficient to know the product of the sample elements. It is proved that this product is a sufficient statistic for the Pareto distribution parameter. On the basis of the maximum likelihood method the distribution degree indicator is estimated. It is proved that this estimate is biased, and a formula eliminating the bias is justified. For the product of the sample elements considered as a random variable the distribution function and probability density are found; mathematical expectation, higher moments, and differential entropy are calculated. The corresponding graphs are built. In addition, it is noted that any function of this product is a sufficient statistic, in particular, the geometric mean. For the geometric mean also considered as a random variable, the distribution function, probability density, and the mathematical expectation are found; the higher moments, and the differential entropy are also calculated, and the corresponding graphs are plotted. In addition, it is proved that the geometric mean of the sample is a more convenient sufficient statistic from a practical point of view than the product of the sample elements. Also, on the basis of the Rao–Blackwell–Kolmogorov theorem, effective estimates of the Pareto distribution parameter are constructed. In conclusion, as an example, the technique developed here is applied to the exponential distribution. In this case, both the sum and the arithmetic mean of the sample can be used as sufficient statistics.

Keywords: Pareto distribution, sufficient statistics, effective estimation, distribution function, moments

• **Submitted:** 20.01.2021 • **Revised:** 27.01.2021 • **Accepted:** 05.04.2021

For citation: Pulkin I.S., Tatarintsev A.V. Sufficient statistics for the Pareto distribution parameter *Rossiiskii tekhnologicheskii zhurnal = Russian Technological Journal.* 2021;9(3):88–97 (in Russ.). https://doi.org/10.32362/2500-316X-2021-9-3-88-97

Financial disclosure: No author has a financial or property interest in any material or method mentioned.

The authors declare no conflicts of interest.

ВВЕДЕНИЕ

Данная работа является продолжением исследований по оценке параметра распределения Парето, проведенных авторами в работах [1–3]. Рассмотрим классическое распределение Парето с функцией распределения

$$F(x)=1-\left(\frac{\theta}{x}\right)^{\alpha},\ x\geq\theta$$

и плотностью вероятности

$$\rho(x)=\frac{\alpha\theta^{\alpha}}{x^{\alpha+1}},\ x\geq\theta.$$

Математическое ожидание и дисперсия случайной величины, распределенной с такой плотностью вероятности, зависят от параметров распределения следующим образом:

$$M\left(\frac{x}{\theta}\right)=\frac{\alpha}{\alpha-1};\ M\left(\frac{x}{\theta}\right)^{2}=\frac{\alpha}{\alpha-2};$$
$$D\left(\frac{x}{\theta}\right)=\frac{\alpha}{(\alpha-1)^{2}(\alpha-2)}.$$

Исследования, связанные с изучением свойств распределения Парето и его обобщений, вызывают пристальный интерес во всем мире [4–8]. Это

связано, в частности, с применимостью этого распределения к анализу интенсивностей природных и техногенных катастроф [9–11]. В большинстве случаев практического применения ставится задача оценить для некоторого распределения значение его параметров по случайной выборке величины, распределенной с заданной плотностью вероятности. Пусть $\{x_1, ..., x_n\}$ – выборка, подчиняющаяся распределению Парето с параметрами α, θ и объемом выборки n. Как отмечалось в [1], для практических нужд обычно можно считать известным параметр θ и оценивать только параметр α.

Совместная плотность вероятности для распределения выборки имеет вид:

$$G(x_1, \ldots, x_n \mid \alpha) = \rho(x_1) \cdot \ldots \cdot \rho(x_n) = \frac{\alpha^n \cdot \theta^{\alpha n}}{(x_1 \cdot \ldots \cdot x_n)^{\alpha+1}}.$$

Как известно [1], несмещенная оценка $\hat{\alpha}$ для степенного параметра α распределения Парето определяется выражением:

$$\hat{\alpha} = \frac{n-1}{\ln(x_1 \cdot \ldots \cdot x_n) - n\ln\theta} = \frac{n-1}{\ln X - n\ln\theta},$$

$$M(\hat{\alpha}) = \alpha, \quad D(\hat{\alpha}) = \frac{\alpha^2}{n-2}, \quad n > 2$$

и является функцией от $X = x_1 \cdot \ldots \cdot x_n$. Таким образом, для оценки параметра $\hat{\alpha}$ достаточно знать не независимые значения x_i, $i = 1, ..., n$, а только их произведение. Совместная плотность вероятности также зависит от произведения элементов выборки:

$$G(x_1, \ldots, x_n \mid \alpha) = G(x_1 \cdot \ldots \cdot x_n \mid \alpha).$$

Функцию $G(x_1, \ldots, x_n \mid \alpha)$ принято называть функционалом правдоподобия. Обычно при этом оговаривается, что она рассматривается как функция от неизвестных параметров α, θ при заданной выборке.

В силу известной теоремы о факторизации [12] статистика $X = x_1 \cdot \ldots \cdot x_n$ является достаточной статистикой для определения неизвестного параметра α. Действительно, функционал правдоподобия $G(x_1, \ldots, x_n \mid \alpha)$ представим в виде

$$G(x_1, \ldots, x_n \mid \alpha) = g(X) \cdot h(x),$$

где

$$g(x_1, \ldots, x_n \mid \alpha) = \alpha^n \cdot \theta^{\alpha n} \cdot X^{-(n+1)}$$

и $h = 1$ не зависит от α.

ДОСТАТОЧНАЯ СТАТИСТИКА ДЛЯ ОЦЕНКИ ПАРАМЕТРА ПАРЕТО

Введем в рассмотрение статистику Y

$$Y = \frac{x_1 \cdot \ldots \cdot x_n}{\theta^n}.$$

Эта статистика, так же, как и сама оценка $\hat{\alpha}$, является достаточной статистикой для оценки параметра распределения Парето, то есть содержит всю необходимую информацию для оценки значения α, имеющуюся в выборке $\{x_1, ..., x_n\}$. Это позволяет построить оптимальную оценку, то есть оценку неизвестного параметра α, имеющую минимальную дисперсию. Именно, теорема Рао – Блекуэлла – Колмогорова [12, 13] утверждает следующее. Пусть T – несмещенная оценка параметра распределения. Тогда условное математическое ожидание $T_1 = M(T \mid Y = y)$ для достаточной статистики Y также является несмещенной оценкой степенного параметра распределения Парето. При этом для дисперсий этих оценок справедливо неравенство

$$D_\alpha T_1 \le D_\alpha T.$$

Такая связь оцениваемого параметра и достаточной статистики позволяет предложить другой способ для вычисления оцениваемого параметра α.

Рассмотрим случайную величину Y, равную, как было сказано ранее, произведению всех элементов выборки $\{x_1, ..., x_n\}$, нормированных на пороговое значение распределения. Для нее можно найти функцию распределения. Она равна

$$F_n(y) = P(Y < y) = \int_{\Delta_n(y)} G(x_1, \ldots, x_n \mid \alpha)\, dx_1 \ldots dx_n = \alpha^n \cdot \theta^{\alpha n} \int_{\Delta_n(y)} \frac{dx_1 \ldots dx_n}{(x_1 \cdot \ldots \cdot x_n)^{\alpha+1}},$$

где интегрирование ведется по объему $\Delta_n(y)$: $\{x_i > \theta;\ (x_1 \cdot \ldots \cdot x_n)/\theta^n < y\}$ в n-мерном пространстве. Для устранения большого количества параметров сделаем замену переменных $x_i = \theta \cdot \exp(\xi_i / \alpha)$, $i = 1, 2, ..., n$, после которой интеграл примет вид:

$$F_n(y) = \int_{\Delta_n(\xi)} e^{-(\xi_1 + \ldots + \xi_n)} d\xi_1 \ldots d\xi_n.$$

Область интегрирования $\Delta_n(\xi)$ будет следующей: $\{\xi_i > 0,\ \xi_1 + \ldots + \xi_n < b\}$. Единственный параметр в интеграле $b = \alpha \cdot \ln y$ содержит комбинацию исходных параметров задачи. Похожий интеграл встречался уже в работе [1]. Вычисление его

связано со свойствами соответствующей рекурсии. Действительно, расставив пределы интегрирования по симплексу $\Delta_n(\xi)$, получим:

$$F_n(y) \equiv I_n(b) = \int_0^b e^{-\xi_1}\, d\xi_1 \int_0^{b-\xi_1} e^{-\xi_2}\, d\xi_2 \ldots \int_0^{b-\xi_1-\ldots-\xi_{n-1}} e^{-\xi_n}\, d\xi_n.$$

Для интеграла $I_n(b)$ имеем рекурсивное равенство:

$$I_n(b) = \int_0^b e^{-\xi} \cdot I_{n-1}(b-\xi)\, d\xi = e^{-b}\int_0^b e^{\chi} \cdot I_{n-1}(\chi)\, d\chi,$$

где второе представление получено после замены переменной интегрирования $\chi = b - \xi$. Вычисляя первые значения последовательности интегралов:

$$I_1 = 1 - e^{-b};\ I_2 = 1 - (1+b)e^{-b};$$
$$I_3 = 1 - \left(1 + b + \frac{b^2}{2}\right)e^{-b};\ \ldots,$$

приходим к гипотезе, которую доказываем по индукции:

$$F_n(x) = 1 - e^{-b}\sum_{k=0}^{n-1}\frac{b^k}{k!};\ b = \alpha \ln y,\quad y \geq 1.$$

Очевидно, что сумма в этом выражении является частичной суммой ряда для показательной функции. Плотность вероятности случайной величины Y может быть получена дифференцированием функции распределения по аргументу y:

$$\rho_n(y) = \frac{dF}{db}\cdot\frac{db}{dy} \Rightarrow \rho_n(y) = \frac{\alpha^n}{\Gamma(n)}\cdot\frac{(\ln y)^{n-1}}{y^{\alpha+1}}$$

для $y \geq 1$ и $\rho_n(y) = 0$ для $y < 1$.

В соответствии с теоремой Рао – Блекуэлла – Колмогорова, несмещенная оценка для параметра распределения Парето получается как условное математическое ожидание по полученному распределению от величины

$$X = \frac{n-1}{\ln Y} \Rightarrow M(X|Y=y) = \int_1^{+\infty}\frac{n-1}{\ln y}\cdot\rho_n(y)\,dy.$$

Нетрудно вычислить, что $M(X|Y=y) = \alpha$ и $D(X|Y=y) = \alpha^2/(n-2)$. Таким образом, оценка параметра и ее дисперсия для достаточной статистики в точности повторяют соответствующий результат для оценки максимального правдоподобия, полученной в работах [1, 2].

СВОЙСТВА РАСПРЕДЕЛЕНИЯ ДОСТАТОЧНОЙ СТАТИСТИКИ

В данном пункте рассмотрим свойства распределения случайной величины $Y = (x_1 \cdot \ldots \cdot x_n)/\theta^n$, заданной плотностью вероятности:

$$\rho_n(y) = \frac{\alpha^n}{\Gamma(n)}\cdot\frac{(\ln y)^{n-1}}{y^{\alpha+1}},\quad y \geq 1.$$

Легко проверить, что полученное распределение удовлетворяет (как и должно быть) условию нормировки:

$$\int_1^{+\infty}\rho_n(y)\,dy = 1.$$

Плотность вероятности имеет максимум в точке $y_n^{(0)}$, являющейся нулем производной функции $\rho_n(y)$. Точка максимума, называемая модой распределения, задается выражением:

$$y_n^{(0)} = \exp\left(\frac{n-1}{\alpha+1}\right).$$

Далее приведен график распределения случайной величины Y.

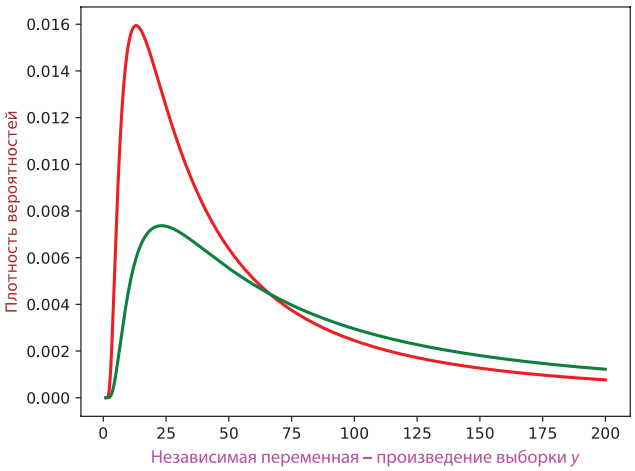

Рис. 1. Графики плотности вероятностей распределения случайной величины Y с $n = 10$ (красный) и $n = 12$ (зеленый), $\alpha = 2.5$

При увеличении n мода распределения экспоненциально возрастает, то есть точка максимума распределения сдвигается в сторону больших значений y. Значение плотности вероятности в точке максимума стремится, очевидно, к нулю:

$$\max\rho_n = \frac{\alpha^n}{\Gamma(n)}\cdot\left(\frac{n-1}{\alpha+1}\right)^{n-1}\cdot e^{-n+1} \sim$$
$$\sim\left(\frac{\alpha}{\alpha+1}\right)^n\cdot\frac{\alpha+1}{\sqrt{2\pi(n-1)}} \to 0, n \to +\infty.$$

Кроме того, легко посчитать и моменты данной случайной величины. Для момента произвольного p-го порядка имеем выражения:

$$M_n\left(Y^p\right)=\int_1^{+\infty} y^p\cdot\rho_n(y)\,dy=\frac{\alpha^n}{\Gamma(n)}\cdot\int_1^{+\infty}\frac{(\ln y)^{n-1}}{y^{\alpha-p+1}}\,dy.$$

Интеграл для моментов сходится для $p < \alpha$, то есть случайная величина Y имеет конечное число низших моментов. После замены $y = e^t$ получим окончательный вид интеграла для их вычисления:

$$M_n\left(Y^p\right)=\frac{\alpha^n}{\Gamma(n)}\cdot\int_0^{+\infty} t^{n-1}e^{-(\alpha-p)t}\,dt=\left(\frac{\alpha}{\alpha-p}\right)^n.$$

Выражение для моментов зависит от объема выборки n как показательная функция. В частности, математическое ожидание Y, вследствие того, что данная величина является произведением независимых величин, образующих случайную выборку, равно произведению математических ожиданий каждого множителя:

$$M(Y)=M\left(\frac{x_1}{\theta}\right)\cdot\ldots\cdot M\left(\frac{x_n}{\theta}\right)=\left(\frac{\alpha}{\alpha-1}\right)^n.$$

Значения математического ожидания, дисперсии и всех существующих моментов неограниченно возрастают, так же как и мода распределения, при возрастании n. Скорость возрастания этих величин тем больше, чем больше порядок момента p.

Для дифференциальной энтропии распределения

$$H_n(Y)=-\int_1^{+\infty}\ln\rho_n(y)\cdot\rho_n(y)\,dy,$$

подставив плотность вероятности $\rho_n(y)$ в определение, получим интеграл:

$$H_n(Y)=$$
$$=\int_1^{+\infty}\left(\ln\Gamma(n)-n\ln\alpha+(\alpha+1)\ln y-(n-1)\ln\ln y\right)\cdot\rho_n(y)\,dy.$$

После замены переменной интегрирования $y = e^t$ перейдем к выражению, связанному с интегральным представлением для гамма-функции Эйлера

$$H_n(Y)=$$
$$=\frac{\alpha^n}{\Gamma(n)}\int_0^{+\infty}\left(\ln\Gamma(n)-n\ln\alpha+(\alpha+1)t-(n-1)\ln t\right)\cdot t^{n-1}\cdot e^{-\alpha t}\,dt.$$

Как известно, дифференциальная энтропия определяется с точностью до аддитивной константы. Поэтому при вычислении этого интеграла были отброшены слагаемые, не зависящие от n. В результате вычисление этого интеграла дает следующее значение энтропии:

$$H_n(Y)=\ln\Gamma(n)-(n-1)\psi(n)+n\left(1-\ln\alpha+\frac{1}{\alpha}\right).$$

Здесь используется обозначение:

$$\psi(x)=\frac{\Gamma'(x)}{\Gamma(x)}=\frac{d}{dx}\ln\Gamma(x)$$

– так называемая дигамма-функция Эйлера (см. [14]). В частности, для натуральных аргументов пси-функция выражается следующим образом:

$$\psi(n)=-\gamma+\sum_{k=1}^{n-1}\frac{1}{k},\quad n>2.$$

Здесь $\gamma = 0.5772...$ – постоянная Эйлера.

График зависимости дифференциальной энтропии от n изображен на рисунке 2.

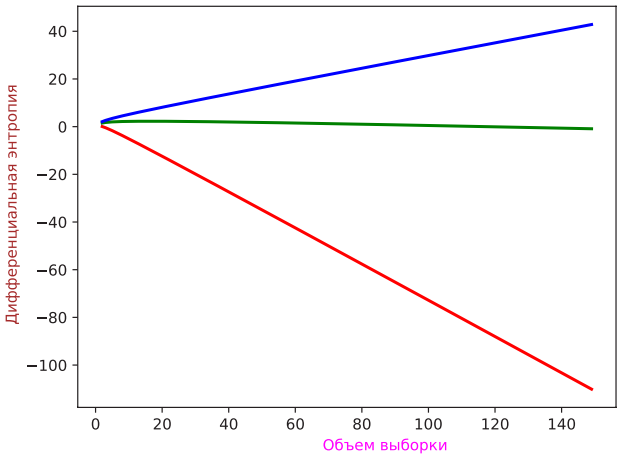

Рис. 2. Дифференциальная энтропия распределения произведения выборки с $\alpha = 3$ (красный), $\alpha = 1.8$ (зеленый) и $\alpha = 1.5$ (синий)

Зависимость дифференциальной энтропии от объема выборки n асимптотически линейная. Асимптотика для дифференциальной энтропии при больших объемах выборки $n \to +\infty$ имеет вид:

$$H_n(Y)=n\left(\frac{1}{\alpha}-\ln\alpha\right)+\frac{1}{2}\ln n+O(1).$$

Критическое значение параметра α_0, при котором коэффициент линейного члена обращается в нуль, находится как корень уравнения $\alpha^{-1}=\ln\alpha$ и составляет приближенно $\alpha_0 = 1.7632...$.

Таким образом, полученное распределение обладает некоторыми особенностями, которые следует скорее отнести к недостаткам, чем к его

достоинствам. Это, прежде всего, существование только конечного числа моментов, зависящего от параметра распределения, возрастание дисперсии и моментов с ростом объема выборки n, что приводит к размыванию пика распределения и его смещению в бесконечность. Все это делает работу с этим распределением не очень удобной. Авторами работы найден оригинальный способ избавиться от недостатков, добавив положительных черт распределению для достаточной статистики.

СРЕДНЕЕ ГЕОМЕТРИЧЕСКОЕ ВЫБОРКИ КАК ДОСТАТОЧНАЯ СТАТИСТИКА

Вместо случайной величины Y, являющейся достаточной статистикой для оценки степенного параметра Парето, рассмотрим другую случайную величину, равную среднему геометрическому элементов выборки $\{x_1, ..., x_n\}$:

$$Y_g = \frac{\sqrt[n]{x_1 \cdot \ldots \cdot x_n}}{\theta}.$$

Преимуществом данной величины является то, что она ограничена и при увеличении n не стремится неограниченно к бесконечности, в отличие от значения Y. Действительно, ранее было установлено, что оценка

$$\hat{\alpha} = \frac{n-1}{\ln X - n \ln \theta}$$

является несмещенной и состоятельной оценкой неизвестного параметра α. Поэтому предел правой части при $n \to \infty$ равен α. Это возможно (при $\alpha \neq 1$) только в том случае, когда $\ln X \sim Cn$, то есть когда $\ln X$ растет на бесконечности пропорционально n. Но тогда

$$\ln Y_g = \frac{\ln X}{n} - \ln \theta$$

имеет конечный предел при $n \to \infty$.

Несмещенная оценка параметра распределения α в этом случае будет иметь слегка модифицированный вид:

$$X = \frac{n-1}{n \cdot \ln Y_g}.$$

Статистика Y_g – также достаточная статистика, но свойства ее существенно отличаются от свойств Y. Действительно, функция распределения $F_g(y) = P(Y_g < y)$ вычисляется во многом аналогично предыдущему случаю:

$$F_g(y) = \int\limits_{\Delta_g(y)} G(x_1,\ldots,x_n|\alpha)\, dx_1 \ldots dx_n =$$

$$= \alpha^n \cdot \theta^{\alpha n} \int\limits_{\Delta_g(y)} \frac{dx_1 \ldots dx_n}{(x_1 \cdot \ldots \cdot x_n)^{\alpha+1}},$$

где интегрирование ведется по объему $\Delta_g(y)$: $\{x_i > \theta;\ (x_1 \cdot \ldots \cdot x_n)/\theta^n < y^n\}$. Приведем конечный результат вычислений для $F_g(y)$:

$$F_g(y) = 1 - e^{-d} \sum_{k=0}^{n-1} \frac{d^k}{k!};\ d = \alpha n \ln y;\ y \geq 1.$$

Плотность вероятности Y_g получается дифференцированием функции распределения по аргументу y:

$$\rho_g(y) = \frac{(\alpha n)^n}{\Gamma(n)} \cdot \frac{(\ln y)^{n-1}}{y^{\alpha n+1}},\ y \geq 1$$

и $\rho_g(y) = 0$ для $y < 1$.

На рис. 3 приведен график распределения случайной величины Y_g.

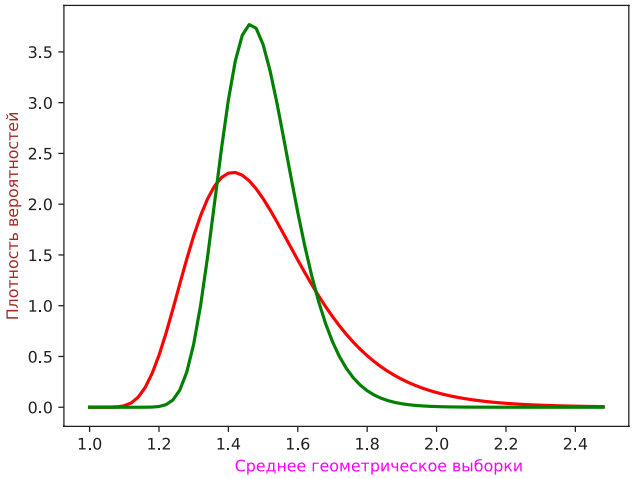

Рис. 3. Графики плотности вероятностей распределения случайной величины Y с $n = 10$ (красный) и $n = 30$ (зеленый), $\alpha = 2.5$

Свойства данного распределения существенно отличаются от предыдущего $\rho_n(y)$. Так, например, мода распределения теперь имеет вид:

$$y_g^{(0)} = \exp\left(\frac{n-1}{\alpha n+1}\right)$$

и при увеличении объема выборки n стремится к постоянному значению $e^{1/\alpha}$. Кроме того, моменты данной случайной величины также легко вычисляются:

$$M_n\left(Y_g^p\right) = \int\limits_1^{+\infty} y^p \cdot \rho_{n,g}(y)\, dy = \frac{(\alpha n)^n}{\Gamma(n)} \cdot \int\limits_1^{+\infty} (\ln y)^{n-1} \frac{dy}{y^{\alpha n - p + 1}}.$$

Интеграл для моментов сходится для $\alpha n > p$. При увеличении n число моментов, существующих у данной случайной величины, неограниченно возрастает.

$$M_n\left(Y_g^p\right)=\frac{(\alpha n)^n}{\Gamma(n)}\cdot\int_0^{+\infty}t^{n-1}e^{-(\alpha n-p)t}\,dt=\left(\frac{\alpha n}{\alpha n-p}\right)^n.$$

Для моментов существует предельное выражение:

$$\lim_{n\to+\infty}M_n\left(Y_g^p\right)=e^{p/\alpha}.$$

Дисперсия среднего геометрического стремится к нулю при увеличении n:

$$D_n\left(Y_g\right)=\left(\frac{\alpha n}{\alpha n-2}\right)^n-\left(\frac{\alpha n}{\alpha n-1}\right)^{2n},\ \lim_{n\to+\infty}D_n\left(Y_g\right)=0.$$

Асимптотики математического ожидания и дисперсии при больших n имеют вид:

$$M_n\left(Y_g\right)=e^{1/\alpha}\left(1+\frac{1}{2\alpha^2 n}+\frac{8\alpha+3}{24\alpha^4 n^2}+\bar{o}\left(\frac{1}{n^2}\right)\right),$$

$$D_n\left(Y_g\right)=e^{2/\alpha}\left(\frac{1}{\alpha^2 n}+\frac{4\alpha+3}{2\alpha^4 n^2}+\bar{o}\left(\frac{1}{n^2}\right)\right),$$

и неплохо аппроксимируют функции на всем интервале допустимых значений $\alpha n > 2$.

Дифференциальная энтропия для этого распределения:

$$H_n\left(Y_g\right)=-\int_1^{+\infty}\ln\rho_g(y)\cdot\rho_g(y)dy=$$

$$=\int_1^{+\infty}\left((\alpha n+1)\ln y+\ln\Gamma(n)-n\ln(\alpha n)-(n-1)\ln\ln y\right)\cdot\rho_g(y)dy$$

после замены переменной интегрирования $y = e^t$ приводится к интегралу:

$$H_n\left(Y_g\right)=$$

$$=\frac{(\alpha n)^n}{\Gamma(n)}\int_0^{+\infty}\left((\alpha n+1)t+\ln\Gamma(n)-n\ln(\alpha n)-(n-1)\ln t\right)\cdot t^{n-1}\cdot e^{-\alpha n t}dt.$$

Вычисление этого интеграла дает значение энтропии:

$$H_n\left(Y_g\right)=\ln\Gamma(n)-(n-1)\psi(n)+n-n\ln(n\alpha)+\frac{1}{\alpha}.$$

Последнее слагаемое не зависит от n, однако сохранено для большей общности. Данное выражение для энтропии среднего геометрического получается из полученного ранее выражения для энтропии произведения заменой α на $(n\alpha)$.

Асимптотика данного выражения при больших $n\to+\infty$ имеет всегда отрицательный коэффициент при старшей степени:

$$H_n\left(Y_g\right)=-n\ln(\alpha n)+\frac{1}{2}\ln n+\underline{O}(1).$$

Таким образом, зависимость от объема выборки при любом параметре α имеет аналогичный характер и функция монотонно убывает в отрицательной области значений.

График зависимости дифференциальной энтропии от n изображен на рисунке 4.

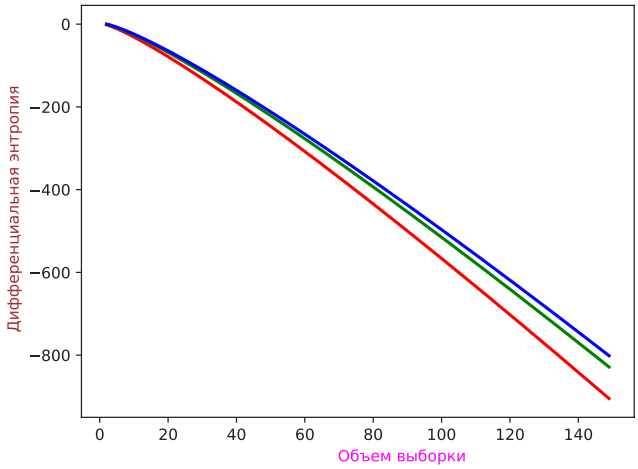

Рис. 4. Дифференциальная энтропия распределения среднего геометрического выборки с α = 3 (красный), α = 1.8 (зеленый) и α = 1.5 (синий)

Как видно, график энтропии монотонно спадает с ростом n в отрицательную область и очень похож на аналогичный для распределения параметра Парето [2]. Бесконечное убывание энтропии в отрицательной области значений, как известно, свидетельствует о том, что с ростом объема выборки происходит сжатие распределения в δ-образную форму, что является следствием увеличения детерминированности и точности определения значения параметра распределения α.

ОЦЕНКА ПАРАМЕТРА ДЛЯ ПОКАЗАТЕЛЬНОГО РАСПРЕДЕЛЕНИЯ

Показательное распределение является одним из известнейших распределений математической статистики и широко используется в теории массового обслуживания и многих других прикладных задачах. Показательное распределение также содержит параметр, оценка которого имеет важное значение в практических целях. Плотность вероятности распределения имеет вид:

$$\rho(x)=\lambda\cdot e^{-\lambda x},\ x\ge 0$$

и $\rho(x)=0,\ x<0$. Математическое ожидание и дисперсия случайной величины, распределенной по показательному закону, определяются, как известно, следующим образом:

$$M(X)=\frac{1}{\lambda};\ D(X)=\frac{1}{\lambda^2}.$$

Для выборки $\{x_1, ..., x_n\}$ случайных величин, распределенных по показательному закону распределения с заданными параметрами λ и n, совместная плотность вероятности выборки имеет вид:

$$G(x_1, \ldots, x_n|\lambda) = \rho(x_1)\cdot\ldots\cdot\rho(x_n) = \lambda^n \cdot e^{-\lambda(x_1+\ldots+x_n)}.$$

Плотность зависит только от суммы элементов выборки, следовательно, для определения параметра достаточно знать значение $Y = x_1 + ... + x_n$. Данная случайная величина и является достаточной статистикой для определения параметра распределение λ. Логарифмируя функцию G, получаем так называемый функционал правдоподобия:

$$L(\alpha) \equiv \ln G(x_1,\ldots,x_n|\alpha) = n\ln\lambda - \lambda(x_1+\ldots+x_n).$$

Дифференцируя его по параметру распределения λ и приравнивая производную к нулю, получаем оценку для параметра распределения методом максимального правдоподобия:

$$\hat{\lambda} = \frac{n}{x_1+\ldots+x_n}.$$

Так же как и для распределения Парето, данная оценка является смещенной. Нетрудно вычислить математическое ожидание этой случайной величины:

$$M(\hat{\lambda}) = \int_0^{+\infty}\ldots\int_0^{+\infty}\hat{\lambda}\cdot G(x_1,\ldots,x_n|\lambda)dx_1\cdot\ldots\cdot dx_n = \frac{n}{n-1}\cdot\lambda.$$

Таким образом, несмещенной оценкой параметра λ является величина X, равная:

$$X = \frac{n-1}{x_1+\ldots+x_n} = \frac{n-1}{Y}.$$

Математическое ожидание и дисперсия этой величины оценки параметра распределения равны соответственно:

$$M(X) = \lambda;\ D(X) = \frac{\lambda^2}{n-2}.$$

Можно также получить и функцию распределения для случайной величины X:

$$F(x) = P(X<x) = 1-\lambda^n\int_{\Delta(x)} e^{-\lambda(x_1+\ldots+x_n)}dx_1\cdot\ldots\cdot dx_n,$$

где интегрирование осуществляется по n-мерному симплексу $\Delta(x)$: $\{x_i > 0, x_1+\ldots+x_n < (n-1)/x\}$. Вычисляя интеграл, приходим к выражению:

$$F(x) = e^{-a_1/x}\sum_{k=1}^{n-1}\frac{(a_1/x)^k}{k!}.$$

Здесь $a_1 = (n-1)\lambda$ – параметр. Плотность вероятности, получаемая дифференцированием функции распределения, будет иметь вид, похожий на соответствующую плотность распределения оценки показателя Парето:

$$\rho(x) = \frac{1}{a_1\Gamma(n)}\cdot\left(\frac{a_1}{x}\right)^{n+1}e^{-a_1/x}.$$

ДОСТАТОЧНАЯ СТАТИСТИКА ДЛЯ ПОКАЗАТЕЛЬНОГО РАСПРЕДЕЛЕНИЯ

Как было указано ранее, одной из достаточных статистик для показательного распределения при оценке значения параметра этого распределения по случайной выборке $\{x_1, ..., x_n\}$ является статистика случайной величины

$$Y = x_1+\ldots+x_n.$$

Сумма элементов выборки содержит полную информацию об оцениваемом параметре распределения. Получим теперь закон распределения и плотность вероятности для величины Y.

$$F(y) = P(Y<y) = \lambda^n\int_{\Delta(y)} e^{-\lambda(x_1+\ldots+x_n)}\ dx_1\cdot\ldots\cdot dx_n.$$

Область интегрирования $\Delta(x)$: $\{x_i > 0, x_1 + ... + x_n < y\}$. Функция распределения для оценки параметра

$$F(y) = 1-e^{-\lambda y}\sum_{k=0}^{n-1}\frac{(\lambda y)^k}{k!}$$

и плотность вероятности

$$\rho(y) = \frac{\lambda^n}{\Gamma(n)}\cdot y^{n-1}\cdot e^{-\lambda y}$$

получаются обычным образом. Математическое ожидание и дисперсия случайной величины могут быть получены и без использования плотности вероятности. Используя свойства математического ожидания независимых величин, получим:

$$M(Y) = M(x_1)+\ldots+M(x_n) = \frac{n}{\lambda}.$$

Аналогично для дисперсии:

$$D(Y) = D(x_1)+\ldots+D(x_n) = \frac{n}{\lambda^2}.$$

Свойства данного распределения обладают такими же недостатками, как и статистика произведения элементов выборки для оценки параметра

распределения Парето. А именно, математическое ожидание и дисперсия неограниченно возрастают с ростом объема выборки. Это делает работу с этой статистикой неудобной в случае больших значений n. Если использовать достаточную статистику, связанную со средним арифметическим элементов выборки:

$$Y_a = \frac{x_1 + \ldots + x_n}{n},$$

то такой подход позволит избежать бесконечных моментов распределения. Используя свойства моментов суммы независимых величин, нетрудно получить:

$$M(Y_a) = \frac{1}{\lambda};\ D(Y_a) = \frac{1}{n\lambda^2} \to 0,\ n \to +\infty.$$

Плотность распределения случайной величины Y_a равна:

$$\rho(y) = \frac{(n\lambda)^n}{\Gamma(n)} \cdot y^{n-1} \cdot e^{-n\lambda y}.$$

Несмещенная оценка для параметра показательного распределения имеет вид:

$$X = \frac{n-1}{n \cdot Y_a};\ M(X) = \lambda.$$

Для этих введенных случайных величин – суммы и среднего арифметического выборки из показательного распределения – также можно вычислить дифференциальную энтропию.

Как было получено ранее – случайная величина $Y = x_1 + \ldots + x_n$ распределена с плотностью вероятности

$$\rho(y) = \frac{\lambda^n}{\Gamma(n)} \cdot y^{n-1} \cdot e^{-\lambda y}.$$

Нетрудно посчитать, что для такого распределения дифференциальная энтропия

$$H_n(Y) = \ln\Gamma(n) - (n-1)\psi(n) + n - \ln\lambda$$

монотонно возрастает независимо от значения λ. Асимптотика этой функции

$$H_n(Y) = \frac{1}{2}\ln n + \underline{O}(1).$$

Для достаточной статистики, связанной со средним арифметическим элементов выборки $Y_a = (x_1 + \ldots + x_n)/n$, распределение вероятности, как было указано, получается заменой параметра λ на $n\lambda$:

$$\rho(y) = \frac{(n\lambda)^n}{\Gamma(n)} \cdot y^{n-1} \cdot e^{-n\lambda y}.$$

Выражение для энтропии и ее асимптотика будут монотонно убывать в отрицательную область значений:

$$H_n(Y) = \ln\Gamma(n) - (n-1)\psi(n) + n - \ln(n\lambda);$$

$$H_n(Y) = -\frac{1}{2}\ln n + \underline{O}(1).$$

Это свидетельствует о том, что особенности выбора достаточной статистики, исследованные для базового распределения Парето, не являются случайными, но повторяются и для показательного распределения.

ЗАКЛЮЧЕНИЕ

Таким образом, использование достаточных статистик позволяет строить несмещенные и эффективные оценки параметров распределения, как для распределения Парето, так и для показательного распределения. Полезным приемом, улучшающем качество оценивания, оказалось построение такой функции от достаточной статистики, которая ограничена, если объем выборки возрастает. В частности, для распределений Парето таким свойством обладает среднее геометрическое выборки, а для показательного распределения – среднее арифметическое.

Рассматривая это среднее геометрическое как случайную величину, мы получаем инструмент для построения эффективных оценок параметров распределений, поскольку ее дисперсия стремится к нулю с ростом объема выборки.

Вклад авторов. Все авторы в равной степени внесли свой вклад в исследовательскую работу.

Authors' contribution. All authors equally contributed to the research work.

СПИСОК ЛИТЕРАТУРЫ

1. Пулькин И.С., Татаринцев А.В. Свойства оценки максимального правдоподобия показателя распределения Парето. *Российский технологический журнал*. 2018;6(6):77–83. https://doi.org/10.32362/2500-316X-2018-6-6-74-83

REFERENCES

1. Pulkin I.S., Tatarintsev A.V. Properties of the maximum likelihood estimates of the exponent of Pareto distribution. *Rossiiskii tekhnologicheskii zhurnal = Russian Technological Journal*. 2018;6(6):74–83 (in Russ.). https://doi.org/10.32362/2500-316X-2018-6-6-74-83

2. Пулькин И.С., Татаринцев А.В. Статистические свойства показателя распределения Парето. *Cloud of Science*. 2020;7(3):498–509.
3. Пулькин И.С., Татаринцев А.В. Инерция формы оценки показателя распределения Парето. *Cloud of Science*. 2020;7(4):790–800.
4. Afify A.Z., Yousof H.M., Butt N.S., Hamedani G.G. The transmuted Weibull-Pareto distribution. *Pak. J. Statist*. 2016;32(3):183–206.
5. Dixit U.J., Nooghabi M.J. Comments on the estimate for Pareto distribution. *Stat. Methodology*. 2010;7:687–691.
6. Ekpenyong E.J., Njoku O.J., Akpan V.M. Efficiency of some estimation methods of the parameters of a two-parameter Pareto distribution. *Am. J. Mathem. Stat*. 2018;8(5):105–110.
7. Gui W. Modified inverse moment estimation: its principle and applications. *Comm. Stat. Appl. Meth.* 2016;23(6):479–496. https://doi.org/10.5351/CSAM.2016.23.6.479
8. Hussain S., Bhatti S.H., Ahmad T., Aftab M., Tahir M. Parameter estimation of Pareto distribution: some modified moment estimators. *Maejo Int. J. Sci. Technol*. 2018;12(01):11–27.
9. Langousis A., Mamalakis A., Puliga M., Deidda R. Threshold detection for the generalized Pareto distribution: review of representative methods and application to the NOAA NCDC daily rainfall database. *Water Resour. Res*. 2016;52(4):2659–2681. https://doi.org/10.1002/2015WR018502
10. Mansoor R.M., Devendra K. Generalized Pareto distribution based on generalized order statistics and associated inference. *Statistics in Transition. New series*. 2019;20(3):57–79. https://doi.org/10.21307/stattrans-2019-024
11. Pu C., Pan X. On the actuarial simulation of the general Pareto distribution of catastrophe loss. In: *Lecture Notes in Electrical engineering. Book series*. 2013. V. 242. P. 1153–1164. https://doi.org/10.1007/978-3-642-40081-0_97
12. Ивченко Г.И., Медведев Ю.И. *Введение в математическую статистику*. М.: Изд-во ЛКИ; 2010. 600 с.
13. Козлов М.В., Прохоров А.В. *Введение в математическую статистику*. М.: Изд-во МГУ; 1987. 264 с.
14. Абрамовиц М., Стиган И. *Справочник по специальным функциям*: пер. с англ. М.: Наука; 1979. 832 с.

2. Pulkin I.S., Tatarintsev A.V. Statistical properties of the Pareto distribution indicator. *Cloud of Science*. 2020;7(3):498–509 (in Russ.).
3. Pulkin I.S., Tatarintsev A.V. Shape inertion of the Pareto distribution parameter estimating. *Cloud of Science*. 2020;7(4):790–800 (in Russ.).
4. Afify A.Z., Yousof H.M., Butt N.S., Hamedani G.G. The transmuted Weibull-Pareto distribution. *Pak. J. Statist*. 2016;32(3):183–206.
5. Dixit U.J., Nooghabi M.J. Comments on the estimate for Pareto distribution. *Stat. Methodology*. 2010;7:687–691.
6. Ekpenyong E.J., Njoku O.J., Akpan V.M. Efficiency of some estimation methods of the parameters of a two-parameter Pareto distribution. *Am. J. Mathem. Stat*. 2018;8(5):105–110.
7. Gui W. Modified inverse moment estimation: its principle and applications. *Comm. Stat. Appl. Meth.* 2016;23(6):479–496. https://doi.org/10.5351/CSAM.2016.23.6.479
8. Hussain S., Bhatti S.H., Ahmad T., Aftab M., Tahir M. Parameter estimation of Pareto distribution: some modified moment estimators. *Maejo Int. J. Sci. Technol*. 2018;12(01):11–27.
9. Langousis A., Mamalakis A., Puliga M., Deidda R. Threshold detection for the generalized Pareto distribution: review of representative methods and application to the NOAA NCDC daily rainfall database. *Water Resour. Res*. 2016;52(4):2659–2681. https://doi.org/10.1002/2015WR018502
10. Mansoor R.M., Devendra K. Generalized Pareto distribution based on generalized order statistics and associated inference. *Statistics in Transition. New series*. 2019;20(3):57–79. https://doi.org/10.21307/stattrans-2019-024
11. Pu C., Pan X. On the actuarial simulation of the general Pareto distribution of catastrophe loss. In: *Lecture Notes in Electrical engineering. Book series*. 2013. V. 242. P. 1153–1164. https://doi.org/10.1007/978-3-642-40081-0_97
12. Ivchenko G.I., Medvedev Yu.I. *Vvedenie v matematicheskuyu statistiku* (*An introduction to mathematical statistics*). Moscow: LKI; 2010. 600 p. (in Russ.).
13. Kozlov M.V., Prokhorov A.V. *Vvedenie v matematicheskuyu statistiku* (*An introduction to mathematical statistics*). Moscow: MGU; 1987. 264 p. (in Russ.).
14. Abramovits M., Stigan I. *Spravochnik po spetsial'nym funktsiyam* (*Handbook of special functions*). Moscow: Nauka; 1979. 832 p. (in Russ.).
[*Handbook of mathematical functions with formular, graphs and mathematical tables*. (Eds.). Abramowitz M., Stegun I. N.Y.: Dover Publications; 1964. 1046 p.]

Об авторах

Пулькин Игорь Сергеевич, к.ф.-м.н., доцент кафедры высшей математики Института кибернетики ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: pulkin@mirea.ru.

Татаринцев Андрей Владимирович, к.ф.-м.н., доцент кафедры высшей математики-2 Физико-технологического института ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: tatarintsev@mirea.ru. Scopus Autor ID: 57221996001, 7004076246.

About the authors

Igor S. Pulkin, Cand. Sci. (Phys.–Math.), Associate Professor, Higher Mathematics Department, Institute of Cybernetics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: pulkin@mirea.ru.

Andrey V. Tatarintsev, Cand. Sci. (Phys.–Math.), Associate Professor, Department of Mathematics, Institute of Physics and Technology, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: tatarintsev@mirea.ru. Scopus Autor ID: 57221996001, 7004076246.