ISSN 1814-1196
Научный вестник НГТУ
том 80, № 4, 2020, с. 121–144

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 80, No. 4, 2020, pp. 121–144

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND CONTROL

УДК 330.564 + 519.234 DOI: 10.17212/1814-1196-2020-4-121-144

Теоретические и эмпирические функции Лоренца, индексы Джини и их свойства*

Д.А. СЕМЕНОВ[1,a], В.Ю. ЩЕКОЛДИН[2,b]

[1] *630128, РФ, г. Новосибирск, ул. Кутателадзе, 4Г, Федеральная кадастровая палата по Новосибирской области*
[2] *630073, РФ, г. Новосибирск, пр. К. Маркса, 20, Новосибирский государственный технический университет*

[a] *miktor.semenov@gmail.com* [b] *raix@mail.ru*

Вопросы оценивания справедливости и эффективности распределения совокупного дохода общества между различными группами населения привлекали внимание ученых с давних времен. Наиболее актуальными они стали в конце XIX – начале XX века в связи с расслоением стран с разнообразным политическим и социальным устройством, вызванным интенсивным развитием экономики, науки и техники. Функция и кривая Лоренца, а также индекс Джини обычно используются для теоретических исследований и приложений в экономических и социальных науках. Первоначально эти инструменты были введены для описания и изучения неравенства распределения дохода и благосостояния среди определенной популяции населения. В последние годы они нашли широкое применение в таких отраслях знания как демография, страхование, здравоохранение, теории риска и надежности, а также и в других областях деятельности человека. В настоящей работе приводятся свойства функции Лоренца и различные представления индекса Джини, систематизируются аналитические результаты для равномерного, экспоненциального, степенного (типа I и II), логнормального распределений, а также распределения Парето (типа I и II). Дополнительно изучен вопрос об оценивании неравенства на основе индекса Пьетра и его связи с функцией Лоренца. Рассматриваются непараметрические оценки функции Лоренца и индекса Джини на основе выборки из соответствующего распределения. Показана строгая состоятельность и асимптотическая несмещенность этих оценок при определенных условиях на исходное распределение при увеличении объема выборки. На основе метода линеаризации оценок установлена асимптотическая нормальность эмпирической функции Лоренца и эмпирического индекса Джини.

Ключевые слова: оценка неравенства, функция Лоренца, кривая Лоренца, индекс Джини, индекс Пьетра, линеаризация оценок, строгая состоятельность, асимптотическая несмещенность, нормальность

* *Статья получена 06 мая 2020 г.*

ВВЕДЕНИЕ

Вопросы распределения доходов и благосостояния и связанные с ними концепции экономического неравенства и социального благосостояния восходят к кодексу Хаммурапи, трудам Аристотеля, Фомы Аквинского, Жан-Жака Руссо и других философов прошлых веков.

С переходом к рыночной экономике во многих странах резко обострился процесс расслоения общества по уровню доходов. Неравенство распределения совокупного дохода общества между различными группами населения стало объектом изучения экономистов и статистиков в конце XIX – начале XX века. Основной проблемой изучения являлась оценка справедливости и эффективности распределения доходов и богатств.

Интенсивно происходящее социальное расслоение общества требовало активного вмешательства государства в процесс перераспределения доходов. Измерение степени неравенства доходов и оценивание уровня бедности стало необходимым для стран с самым разнообразным политическим и социальным устройством.

Для описания и изучения неравенства доходов были предложены различные модели распределения, такие как, например, логнормальное, Парето и другие, применение которых на практике требует соблюдения определенных условий. Для рассмотрения общих ситуаций необходимо наличие более широкого класса инструментов анализа неравенства, наиболее распространенным из которых является кривая Лоренца.

В 1905 г. американский экономист и статистик Макс Отто Лоренц [15] предложил метод анализа распределения доходов и благосостояния населения с помощью кривой на плоскости, получившей впоследствии его имя. Эмпирическая кривая Лоренца строится на основе совокупности n упорядоченных по возрастанию выборочных данных $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ следующим образом: в опорных точках с абсциссами $\frac{i}{n}$, $i = 0,\ldots,n$, полагается $L_n(0) = 0$, $L_n\left(\frac{i}{n}\right) = \frac{S_i}{S_n}$, где $S_i = x_{(1)} + x_{(2)} + \ldots + x_{(i)}$.

Эмпирическая кривая Лоренца $L_n(p)$ определяется для всех $p \in [0,1]$ линейной интерполяцией по опорным точкам, а ее сглаживание некоторой аналитической зависимостью представляет собой *теоретическую функцию Лоренца* $L(p)$, график которой называется *кривой Лоренца* (рис. 1).

Кривая Лоренца располагается в первом квадранте между началом координат $(0,0)$ и точкой $(1,1)$. При этом точка на кривой с координатами $(p, L(p))$ означает, что доля p населения анализируемой территории обладает долей $L(p)$ совокупного дохода. Диагональ единичного квадрата, т. е. прямая $L(p) = p$, называемая *эгалитарной линией*, определяет ситуацию абсолютного равенства распределения доходов. Отличие кривой Лоренца от эгалитарной линии определяет дифференциацию доходов: чем больше кривая Лоренца отклоняется от линии абсолютного равенства, тем больше неравенство в распределении доходов.

Одним из количественных показателей степени дифференциации общества по отношению к какому-либо признаку является индекс Джини G, предложенный в 1912 г. итальянским экономистом, статистиком и демографом Коррадо Джини [11]. В экономических расчетах в качестве изучаемого признака часто рассматривается величина годового дохода общества. *Индекс Джини* основывается на кривой Лоренца и определяется как отношение площади S_{A} фигуры A, ограниченной кривой Лоренца и эгалитарной линией (рис. 1), к площади треугольника под эгалитарной линией $\left(S_{\Delta} = \frac{1}{2} \text{ на рис. 1}\right)$, т. е. $G = 2S_A$. Эта величина принимает значения от нуля до единицы и показывает, насколько распределение доходов отличается от абсолютного равенства, при котором $G = 0$. Чем больше значение индекса Джини отличается от нуля, тем в большей степени доходы сконцентрированы в руках отдельных (небольших) групп населения. Предельное значение $G = 1$ говорит об абсолютном неравенстве, при котором все доходы сосредоточены в руках одного индивида или одной группы населения.

Как кривая Лоренца, так и индекс Джини обычно используются в экономических и социальных науках. Однако они могут отражать неравенство в распределении самых различных величин. Поэтому методы, основанные на этих показателях, в последние годы нашли применение в таких областях знания, как демография, страхование, здравоохранение, теория надежности и др.

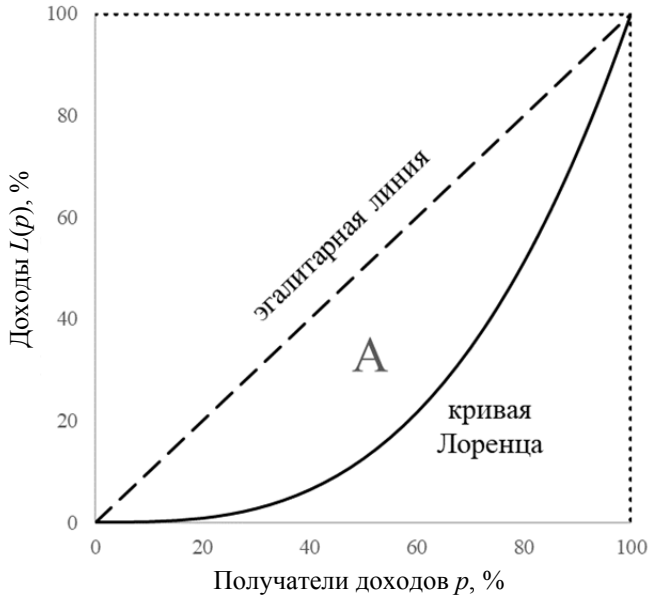

Рис. 1. Геометрическая интерпретация кривой Лоренца и индекса Джини

Fig. 1. Geometric interpretation of the Lorenz curve and the Gini index

1. ФУНКЦИЯ ЛОРЕНЦА И ЕЕ СВОЙСТВА

Пусть X – неотрицательная случайная величина (с.в.) с функцией распределения (ф.р.) $F(x)=P\{X<x\}$, $x\in R$, и математическим ожиданием, или средним значением,

$$\mu \equiv \mathrm{E}(X)=\int_0^\infty x\,dF(x). \tag{1}$$

В дальнейшем всюду будем полагать, что $0<\mu<\infty$. Первоначально определение функции Лоренца, соответствующей ф.р. $F(x)$ с плотностью распределения (п.р.) $f(x)$, было сформулировано в параметрическом виде при помощи системы уравнений (см., например, [3, с. 75]):

$$\begin{cases} p=F(x)=\int_0^x f(t)\,dt, \\ L(p)=L(F(x))=\dfrac{1}{\mu}\int_0^x t\,f(t)\,dt. \end{cases} \tag{2}$$

Для унифицированного определения функции Лоренца, соответствующего произвольному распределению, в том числе и дискретному, в работе [9] использовалось квантильное преобразование $F^{-1}(p)$, $0\le p\le 1$, функции $F(x)$:

$$F^{-1}(p)=\sup\{x\colon F(x)<p\}=\inf\{x: F(x)\ge p\},\ 0<p\le 1,$$

$F^{-1}(0)=\inf\{x: F(x)>0\}$ есть левая граница носителя распределения $F(x)$.

Сделав в (1) и (2) замену переменной $p=F(x)$ и воспользовавшись теоремой о замене переменной под знаком интеграла, получим следующее выражение для функции Лоренца:

$$L(p)=\frac{1}{\mu}\int_0^p F^{-1}(u)\,du=\frac{\int_0^p F^{-1}(u)\,du}{\int_0^1 F^{-1}(u)\,du},\ 0\le p\le 1. \tag{3}$$

Если $\mu=0$ или $\mu=+\infty$, то функция Лоренца не определена. Заметим, что в математической статистике величину $x_p=F^{-1}(p)$ называют *квантилью порядка* p или *p-квантилью*, $0\le p\le 1$. Очевидно, что эта величина имеет смысл для любых $F(x)$. Если ф.р. $F(x)$ непрерывна, то $F^{-1}(p)$ есть минимальное решение уравнения $F(x)=p$, причем решение будет един-

ственным, если $F(x)$ строго монотонна. Нетрудно также видеть, что при непрерывности $F(x)$ верно

$$F(x_p) = p, \quad 0 \le p \le 1. \tag{4}$$

Функция Лоренца обладает рядом полезных свойств. Рассмотрим некоторые из них.

1º. Функция $L(p)$ непрерывна для всех $p \in [0,1]$, $L(0) = 0$ и $L(1) = 1$.

Доказательство. Как ф.р. $F(x)$, так и обратная к ней $F^{-1}(x)$ являются непрерывными слева. Функция $L(p)$ как функция верхнего предела интегрирования p (см. (3)) в силу свойств интеграла непрерывна для всех $p \in [0,1]$. Равенства $L(0) = 0$ и $L(1) = 1$ очевидны из определения.

2º. Если ф.п.р. $f(x_p) > 0$, то существует производная

$$L'(p) = \frac{x_p}{\mu}, \quad 0 \le p \le 1. \tag{5}$$

Доказательство. Дифференцируя равенство (4), получаем

$$\frac{dF(x_p)}{dp} = \frac{dF(x_p)}{dx_p}\frac{dx_p}{dp} = 1,$$

откуда

$$\frac{dx_p}{dp} = \frac{1}{f(x_p)}. \tag{6}$$

В силу (2) и (6) имеем

$$\frac{dL(p)}{dp} = \frac{1}{\mu}\frac{d\left[\int_0^{x_p} x f(x)\,dx\right]}{dx_p}\frac{dx_p}{dp} = \frac{1}{\mu}\frac{x_p f(x_p)}{f(x_p)} = \frac{x_p}{\mu}, \quad 0 \le p \le 1.$$

Поскольку $L'(p) = \frac{x_p}{\mu} > 0$, $0 \le p \le 1$, то из геометрического смысла производной очевидно, что имеет место следующее.

Следствие 1.1. Функция Лоренца $L(p)$ возрастает при $0 \le p \le 1$.

3º. Если ф.п.р. $f(x_p) > 0$, то существует вторая производная

$$L''(p) = \frac{1}{\mu f(x_p)}, \quad 0 \le p \le 1. \tag{7}$$

Доказательство. Из формул (5) и (6) непосредственно вытекает, что

$$\frac{d^2L(p)}{dp^2} = \frac{d}{dp}\left(\frac{x_p}{\mu}\right) = \frac{1}{\mu f(x_p)}, \quad 0 \le p \le 1.$$

Так как $L''(p) = \dfrac{1}{\mu f(x_p)} > 0$, $0 \le p \le 1$, то из геометрического смысла второй производной вытекает, что

Следствие 1.2

Функция Лоренца выпукла вниз при $0 \le p \le 1$.

Свойство функции Лоренца в п.1) означает, в свою очередь, что эгалитарная линия мажорирует функцию Лоренца на этом интервале, т. е.

$$0 \le L(p) \le p, \quad 0 \le p \le 1.$$

4º. Функция $L(p)$ инвариантна относительно положительного масштабирования: с.в. X и cX, где $c > 0$ – произвольная константа, имеют одну и ту же функцию Лоренца.

Доказательство. Рассмотрим с.в. $Y = cX, c > 0$. Ниже индексами X и Y будем обозначать характеристики, относящиеся к с.в. X и Y соответственно. Очевидно, что

$$\mu_X = c\mu_Y \tag{8}$$

и

$$F_Y(y) = P\{cX < y\} = P\{X < y/c\} = F_X(y/c). \tag{9}$$

Для с.в. Y p-квантиль y_p есть решение уравнения $F_Y(y_p) = p$, или, в силу (9), уравнения

$$F_X(y_p/c) = p. \tag{10}$$

Сравнение соотношений (4) и (10) показывает, что

$$y_p/c = x_p, \text{ или } y_p = c\,x_p. \tag{11}$$

Тогда из (8) и (11) вытекает, что

$$L_Y(p) = \frac{1}{\mu_Y}\int_0^p y_u\,du = \frac{1}{c\mu_X}\int_0^p c\,x_u\,du = L_X(p).$$

Следствие 1.3. Любой ф.р. $F(x)$ с конечным средним μ соответствует единственная функция Лоренца $L(p)$.

Обратное утверждение в общем случае в силу свойства 4º неверно. Однако имеет место следующая

Теорема 1.1 ([17]). Пусть $L(p)$ – непрерывная функция, определенная на отрезке $[0,1]$, со второй производной $L''(p)$. Тогда функция $L(p)$ есть функция Лоренца, соответствующая некоторому распределению $F(x)$, тогда и только тогда, когда $L(0)=0,\ L(1)=1,\ L'(p)>0,\ L''(p)>0,\ 0\le p\le 1$.

5º. Максимальное расхождение по вертикали между кривой Лоренца $L(p)$ и прямой абсолютного равенства $L(p)=p$ достигается в точке $p^*=F(\mu)$, и эта величина, называемая *индексом Пьетра* [16], равна

$$P\equiv F(\mu)-L(F(\mu))=\frac{E\left(|x-\mu|\right)}{2\mu}. \tag{12}$$

Доказательство. Поскольку $L(p)$ выпукла вниз, то функция $l(p)=p-L(p)$ выпукла вверх и $l(0)=l(1)=0$. Поэтому существует единственная точка $p^*\in[0,1]$ максимума функции $l(p)$, которая определяется из уравнения $l'(p^*)=0$. Дифференцируя $l(p)$ и используя равенство (5), находим

$$l'(p^*)=1-L'(p^*)=1-F^{-1}(p^*)/\mu=0,$$

откуда $p^*=F(\mu)$. Тогда индекс Пьетра равен

$$P=\max_{0\le p\le 1} l(p)=F(\mu)-L(F(\mu)).$$

Далее в силу уравнений (2) имеем

$$\mu P=\mu\int_0^{\mu} dF(x)-\int_0^{\mu} x\,dF(x)=\int_0^{\mu}(\mu-x)\,dF(x).$$

Но поскольку $\int_0^{\infty}(\mu-x)\,dF(x)=0$, то

$$\mu P=\frac{1}{2}\left[\int_0^{\mu}(\mu-x)\,dF(x)+\int_{\mu}^{\infty}(x-\mu)\,dF(x)\right]=\frac{1}{2}\int_0^{\infty}|x-\mu|\,dF(x)=\frac{1}{2}E\left(|x-\mu|\right),$$

откуда получаем (12).

Индекс Пьетра P показывает, какая доля совокупного дохода (богатства) общества должна быть перераспределена в пользу беднейшего населения. На практике чаще используется другой показатель степени неравенства в распределении доходов – индекс Джини.

2. ИНДЕКС ДЖИНИ И ЕГО ПРЕДСТАВЛЕНИЯ

Мера неравенства распределения некоторого неотрицательного признака X (в том числе и дохода) – *индекс Джини* G – определяется на основе функции Лоренца $L(p)$ формулой

$$G = 1 - 2\int_0^1 L(p)\,dp\,. \tag{13}$$

Это определение согласуется с геометрическим, приведенным во введении. Действительно, из рис. 1 видно, что площадь фигуры A есть

$$S_A = \int_0^1 (p - L(p))\,dp = \frac{1}{2} - \int_0^1 L(p)\,dp = \frac{1}{2}G\,,$$

откуда $G = 2S_A$.

Из формулы (13) вытекают другие представления индекса Джини, основанные на ф.р. $F(x)$, ковариации и средней абсолютной разности.

1º. Исходя из определения (13) и используя интегрирование по частям, находим

$$G = 1 - 2\int_0^1 L(p)\,dp = 1 - 2[pL(p)]\Big|_0^1 + 2\int_0^1 pL'(p)\,dp = 2\int_0^1 pL'(p)\,dp - 1\,.$$

В последнем выражении сделаем замену переменной $p = F(x)$ и воспользуемся формулой (5), что дает нам

$$G = \frac{2}{\mu}\int_0^\infty xF(x)\,dF(x) - 1\,. \tag{14}$$

2º. Рассмотрим интеграл в (14):

$$I = \int_0^\infty xF(x)\,dF(x) = \int_0^\infty x(F(x)-1)\,dF(x) + \int_0^\infty x\,dF(x) = \mu - \int_0^\infty x(1-F(x)\,dF(x)\,.$$

Проинтегрируем по частям второй член в последнем выражении и воспользуемся тем фактом, что если с.в. X имеет конечное среднее

$$\mu = E(X) = \int_0^\infty (1-F(x))\,dx\,,\ \text{то}\ \lim_{x\to\infty} x(1-F(x)) = 0\,.$$

В дальнейшем мы воспользуемся этими соотношениями. Тогда

$$I = \mu - x(1-F(x)F(x)\Big|_0^\infty + \int_0^\infty F(x)\,d[x(1-F(x))] = \mu + \int_0^\infty F(x)(1-F(x)dx - I\,.$$

Из последнего равенства получаем

$$I = \frac{\mu}{2} + \frac{1}{2}\int_0^\infty F(x)(1-F(x)dx\,,$$

что вместе с (14) дает

$$G = \frac{1}{\mu}\int_0^\infty F(x)(1-F(x))\,dx\,. \tag{15}$$

3º. Интеграл в (15) можно переписать как

$$\int_0^\infty F(x)(1-F(x))\,dx = \int_0^\infty (F(x)-1)(1-F(x))\,dx + \int_0^\infty (1-F(x))\,dx =$$

$$= \mu - \int_0^\infty (1-F(x))^2\,dx,$$

откуда

$$G = 1 - \frac{1}{\mu}\int_0^\infty (1-F(x))^2\,dx\,. \tag{16}$$

4º. Напомним, что

$$\operatorname{cov}(X,F(X)) = E(X\,F(X)) - E(X)E(F(X))\,,$$

где $E(X)=\mu$, $E(F(X))=\frac{1}{2}$. Действительно,

$$E(F(X)) = \int_0^\infty F(x)\,dF(x) = \int_0^\infty (F(x)-1)\,dF(x) + \int_0^\infty dF(x) = 1 - \int_0^\infty (1-F(x))\,dF(x)\,.$$

Интегрирование по частям последнего выражения показывает, что

$$E(F(X)) = 1 - (1-F(x))\,F(x)\Big|_0^\infty - \int_0^\infty F(x)\,dF(x) = 1 - E(F(X))\,.$$

Это доказывает, что $E(F(X))=\frac{1}{2}$.

Так как

$$\operatorname{cov}(X,F(X)) = \int_0^\infty x\,F(x)\,dF(x) - \frac{\mu}{2}\,,$$

то последнее выражение вместе с (14) дает

$$G = \frac{2}{\mu}\operatorname{cov}(X,F(X)) = \frac{2}{\mu}\int_0^\infty x\,F(x)\,dF(x) - 1\,. \tag{17}$$

5º. Первоначально Джини [11] ввел индекс G с помощью коэффициента рассеяния

$$\Delta = \int_0^\infty \int_0^\infty |x - y| dF(x) dF(y)$$

следующим образом:

$$G = \frac{\Delta}{2\mu}. \tag{18}$$

Средняя абсолютная разность Δ характеризует разброс значений случайного признака X друг относительно друга, однако прямое ее вычисление сопряжено с известными трудностями.

Определения (13) и (18) являются эквивалентными. Вывод формулы (18) из (13) приведен, например, в [3]. Получим определение (13) из (18).

Пусть X и Y – независимые с.в. с одной и той же ф.р. $F(x)$, т. е. «копии» друг друга. Тогда

$$\Delta = E(|X - Y|) = E[X + Y - 2\min(X, Y)]. \tag{19}$$

Очевидно, что

$$P\{\min(X,Y) < x\} = 1 - P\{\min(X,Y) \ge x\} = 1 - P\{X \ge x, Y \ge x\} =$$

$$= 1 - P\{X \ge x\} P\{Y \ge x\} = 1 - (1 - F(x))^2. \tag{20}$$

Тогда из (19), в силу (20), следует, что

$$\Delta = 2\mu + 2\int_0^\infty x\, d(1 - F(x))^2 = 2\mu - 4\int_0^\infty x(1 - F(x))\, dF(x) =$$

$$= 2\mu - 4\int_0^\infty x\, dF(x) + 4\int_0^\infty x F(x)\, dF(x) = 4\int_0^\infty x F(x)\, dF(x) - 2\mu. \tag{21}$$

Сделав в (21) замену переменной $p = F(x)$, получим

$$\Delta = 4\int_0^1 pF^{-1}(p)dp - 2\mu\,. \tag{22}$$

Так как в силу равенства (5) $F^{-1}(p) = \mu L'(p)$, то из (22) следует, что

$$\Delta = 4\mu\int_0^1 pL'(p)dp - 2\mu = 4\mu\int_0^1 p\, dL(p) - 2\mu\,. \tag{23}$$

Вычисляя интеграл в (23) по частям, находим

$$\Delta = 4\mu\left[pL(p)\Big|_0^1 - \int_0^1 L(p)\, dp \right] - 2\mu = 2\mu - 4\mu\int_0^1 L(p)\, dp = 2\mu\left[1 - 2\int_0^1 L(p)\, dp \right],$$

или согласно определению (13) $\Delta / 2\mu = G$, что завершает вывод (18).

В таблице приведены функции Лоренца и индексы Джини, соответствующие некоторым наиболее распространенным в эконометрическом анализе распределениям. Более подробные сведения можно найти, например, в [12].

3. ЭМПИРИЧЕСКАЯ ФУНКЦИЯ ЛОРЕНЦА, ЭМПИРИЧЕСКИЙ ИНДЕКС ДЖИНИ И ИХ СВОЙСТВА

Пусть $X_1, X_2,..., X_n$ – независимая выборка объема n из генеральной совокупности с ф.р. $F(x)$ и пусть далее $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$ – вариационный ряд, построенный по этой выборке. Заметим, что с.в. $X_{(1)}, X_{(2)},..., X_{(n)}$, называемые *порядковыми статистиками*, уже не являются ни независимыми, ни одинаково распределенными.

Функции Лоренца и индексы Джини для ряда распределений

Lorenz functions and Gini indices for some distributions

№ п/п	Распределение	Функция распределения, $F(x)$	Функция Лоренца, $L(p)$	Индекс Джини, G
1	Равномерное на отрезке $[a,b]$, $0 \le a < b < \infty$	$\frac{x-a}{b-a}$, $x \in [a,b]$	$\frac{2ap+(b-a)p^2}{a+b}$	$\frac{b-a}{3(a+b)}$
2	Экспоненциальное с параметром $\lambda > 0$	$1-e^{-\lambda x}$, $x \ge 0$	$p+(1-p)\ln(1-p)$	$\frac{1}{2}$
3	Степенное I с параметром $\alpha > 0$	x^{α}, $x \in [0,1]$	$p^{1+1/\alpha}$	$\frac{1}{2\alpha+1}$
4	Степенное II с параметром $\beta > 0$	$1-(1-x)^{\beta}$, $x \in [0,1]$	$1-(1+\beta)(1-p)+$ $+\beta(1-p)^{1+1/\beta}$	$\frac{\beta}{2\beta+1}$
5	Парето I с параметрами $\alpha > 0$, $c > 0$	$1-\left(\frac{c}{x}\right)^{\alpha}$, $x \ge c$	$1-(1-p)^{1-1/\alpha}$	$\frac{1}{2\alpha-1}$
6	Парето II с параметрами $\alpha > 0$, $c > 0$	$1-\left(\frac{c}{c+x}\right)^{\alpha}$, $x \ge c$	$\alpha(1-p)^{1-1/\alpha}-\alpha-$ $-p(1-\alpha)$	$2+\alpha-\frac{2\alpha^2}{2-\alpha}$
7	Логнормальное с параметрами $\mu > 0$, $\sigma > 0$	$\Phi\left(\frac{\ln x-\mu}{\sigma}\right)$, $x > 0$	$\Phi(\Phi^{-1}(p)-\sigma)$	$2\Phi\left(\frac{\sigma}{\sqrt{2}}\right)-1$

Примечание. $\Phi(x)$ – функция стандартного нормального распределения.

Пусть $F_n^{-1}(x)$ – функция, обратная к эмпирической ф.р. вида

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{X_i < x\}, \quad x \in R,$$

где $\mathrm{I}\{\mathrm{A}\}$ – индикатор события A. В терминах порядковых статистик определение функции $F_n^{-1}(x)$, $x \in [0,1]$, выглядит так:

$$F_n^{-1}(x) = X_{(k)}, \text{ если } x \in \left[\frac{k-1}{n}, \frac{k}{n}\right], k = 1,...,n.$$

Это следует из того, что эмпирическая ф.р. $F_n(x)$ возрастает скачками величины $\frac{1}{n}$ в точках $X_{(k)}, k = 1,...,n$. Следовательно, функция $F_n^{-1}(x)$ полностью определяется порядковыми статистиками.

Выборочной квантилью порядка p называется величина

$$x_p^* = F_n^{-1}(p) = X_{(k)},$$

где $k = \begin{cases} np, & \text{если } np - \text{целое}, \\ [np]+1 & \text{иначе}, \end{cases}$ и $[.]$ – операция взятия целой части числа.

Теорема 3.1. Если с.в. X имеет строго монотонную ф.р. $F(x)$, то при $n \to \infty$ выборочная квантиль x_p^*, $0 \le p \le 1$, является:

a) строго состоятельной, т. е. сходится почти наверное (п. н.), или с вероятностью единица: $x_p^* \xrightarrow{\text{п.н.}} x_p$;

b) асимптотически несмещенной: $\mathrm{E}(x_p^*) \to x_p$;

c) если ф.п.р. $f(x)$ и ее производная $f'(x)$ непрерывны в некоторой окрестности точки $x_p = F^{-1}(p)$ и $f(x_p) > 0$, то выборочная квантиль x_p^* является асимптотически нормальной с параметрами x_p и σ_p^2 / n, где $\sigma_p^2 = p(1-p)/f^2(x_p)$, т. е. сходится слабо или по распределению к с.в., имеющей стандартное нормальное распределение: $\sqrt{n}(x_p^* - x_p)/\sigma_p \xRightarrow{d} N(0,1)$.

Доказательство. *a*) Поскольку ф.р. $F(x)$ строго монотонна, то $x_p = F^{-1}(p)$ – единственное решение уравнения $F(x) = p$. Тогда $F(x_p - \varepsilon) < p < F(x_p + \varepsilon)$ для произвольного $\varepsilon > 0$. Согласно теореме Гливенко–Кантелли [1], $F_n(x) \xrightarrow{\text{п.н.}} F(x)$ при $n \to \infty$, поэтому при $n \to \infty$

$$F_n(x_p \pm \varepsilon) \xrightarrow{\text{п.н.}} F(x_p \pm \varepsilon),$$

откуда

$$P\{F_m(x_p-\varepsilon)<p<F_m(x_p+\varepsilon),\forall m\ge n\}\to 1.$$

Очевидно, что для любой ф.р. $F(x)$ имеем

$$F(x)>p \text{ тогда и только тогда, когда } x>F^{-1}(p). \tag{24}$$

Поэтому при $n\to\infty$

$$P\{x_p-\varepsilon<F_m^{-1}(p)<x_p+\varepsilon,\ \forall m\ge n\}\to 1,$$

откуда

$$P\left\{\sup_{m\ge n}|F_m^{-1}(p)-x_p|>\varepsilon\right\}\to 0,$$

Что завершает доказательство п. *a*) теоремы.

b) Для ф.р. с.в. x_p^* в силу (24) имеем

$$P\{x_p^*<x\}=P\{F_n^{-1}(p)<x\}=P\{F_n(x)>p\}. \tag{25}$$

Так как $F_n(x)\overset{\text{п.н.}}{\to}F(x)$ при $n\to\infty$, то из (25) и (24) следует, что

$$P\{x_p^*<x\}\to P\{F(x)>p\}=P\{x>F^{-1}(p)\}=I\{x>x_p\}. \tag{26}$$

Отсюда получаем, что при $n\to\infty$

$$E(x_p^*)=\int_0^1(1-P\{x_p^*<x\})dx\to\int_0^1(I-I\{x>x_p\})dx=\int_0^1 I\{x\le x_p\}dx=\int_0^{x_p}dx=x_p.$$

c) В условиях теоремы 3.1 имеет место разложение Бахадура [6, 10] для x_p^* в виде

$$x_p^*=x_p+y_n(p)+o_p(n^{-1/2}). \tag{27}$$

Здесь

$$y_n(p)=\frac{1}{n}\sum_{i=1}^{n}\left(p-I\{X_i<x_p\}\right)/f(x_p), \tag{28}$$

а символ $o_p(n^{-1/2})$ означает такую с.в. r_n, что при $n\to\infty$

$$r_n/n^{-1/2}=\sqrt{n}\,r_n\overset{P}{\to}0 \text{ (сходится по вероятности).} \tag{29}$$

Индикатор события $I\{X_i < x_p\}$ в (28) – это с.в., имеющая распределение Бернулли с параметрами

$$E(I\{X_i < x_p\}) = P\{X_i < x_p\} = F(F^{-1}(p)) = p \tag{30}$$

и

$$D(I\{X_i < x_p\}) = E(I\{X_i < x_p\}^2) - (E(I\{X_i < x_p\}))^2 = p - p^2 = p(1-p) \tag{31}$$

для всех $i = 1,...,n$. Тогда из равенств (28)–(31) вытекает, что

$$E(y_n(p)) = 0, \quad D(y_n(p)) = \frac{p(1-p)}{n f^2(x_p)} = \frac{\sigma_p^2}{n}. \tag{32}$$

С.в. $y_n(p)$ представляет собой нормированную сумму независимых одинаково распределенных с.в. и согласно центральной предельной теореме асимптотически нормальна с параметрами, определяемыми формулами (32).

В представлении (27) величина x_p – не с.в., а при $n \to \infty$ верно, что $\sqrt{n} y_n(p) / \sigma_p \overset{d}{\Rightarrow} N(0,1)$ с остаточным членом порядка $o_p(n^{-1/2})$. Следовательно, в силу теоремы Слуцкого [4] получаем, что выборочная квантиль x_p^* асимптотически нормальна с параметрами x_p и σ_p^2 / n.

Замечание 3.1. Метод, использованный при доказательстве асимптотической нормальности выборочной квантили x_p^*, называют *линеаризацией оценки*. Этот подход мы будем применять также при изучении асимптотики эмпирических функции Лоренца и индекса Джини. Существуют и другие методы установления асимптотической нормальности оценок, основанные на *принципе инвариантности*, *функции влияния*, U-*статистиках* и др. С ними можно ознакомиться, например, в работах [1–5, 7, 8, 13, 14]; дополнительную литературу можно найти там же.

Поскольку на практике распределение исследуемого признака *X*, как правило, неизвестно, для оценки предельной дисперсии σ_p^2 необходимо оценить неизвестную ф.п.р. $f(x)$. Один из современных подходов к решению данной задачи основан на использовании *ядерных оценок* [1]. В этом случае в качестве статистического аналога теоретической ф.п.р. $f(x)$ рассматривают случайную функцию

$$f_n(x) = \frac{1}{n\alpha_n} \sum_{i=1}^{n} \mathrm{K}\left(\frac{x - X_i}{\alpha_n} \right)$$

при соответствующем выборе функции ядра $K(x)$ и последовательности чисел $\alpha_n > 0$ (диаметров ядра).

Следствие 3.1. При использовании ядерной оценки ф.п.р. $f(x)$ оценка дисперсии σ_p^2 в виде

$$(\sigma_p^*)^2 = \frac{p(1-p)}{f_n^2(x_p^*)}$$

будет состоятельной.

Обозначим $Q(p) \equiv x_p = F^{-1}(p)$ и $Q_n(p) \equiv x_p^* = F_n^{-1}(p)$, $0 \le p \le 1$. Естественной оценкой функции Лоренца $L(p)$ является *эмпирическая функция Лоренца*

$$L_n(p) = \frac{1}{\mu_n}\int_o^p Q_n(t)\,dt = \frac{\frac{1}{n}\sum_{i=1}^{k} X_{(i)}}{\frac{1}{n}\sum_{i=1}^{n} X_{(i)}} = \frac{\sum_{i=1}^{k} X_{(i)}}{\sum_{i=1}^{n} X_{(i)}}, \qquad (33)$$

где $\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ – выборочное среднее, $k = \begin{cases} np, & \text{если } np - \text{целое}, \\ [np]+1 & \text{иначе.} \end{cases}$

Далее нам понадобятся следующие леммы.

Лемма 3.1. Пусть $\{X_n, n \ge 1\}$ – последовательность с.в. такая, что $X_n \xrightarrow{\text{п.н.}} X$ при $n \to \infty$, где X – некоторая с.в., а $\{Y_n, n \ge 1\}$ – другая последовательность с.в., такая, что $Y_n \xrightarrow{\text{п.н.}} C$ при $n \to \infty$, где $C \ne 0$ – некоторая постоянная. Тогда при $n \to \infty$ верно $X_n / Y_n \xrightarrow{\text{п.н.}} X / C$.

Доказательство см., например, в [4] с незначительными модификациями для сходимости почти наверное.

Лемма 3.2. Если выполнены условия пункта *c* теоремы 3.1, то

$$L_n(p) = L(p) + l_n(p) + o_p(n^{-1/2}), \qquad (34)$$

где

$$l_n(p) = \frac{1}{\mu}\frac{1}{n}\sum_{i=1}^{n}\left[X_i \mathrm{I}\{X_i < x_p\} - x_p \mathrm{I}\{X_i < x_p\} - X_i L(p) + p x_p\right]. \qquad (35)$$

Доказательство. Элементарные выкладки показывают, что

$$L_n(p) - L(p) = \frac{1}{\mu_n}\int_0^p [Q_n(t) - Q(t)]\,dt - \frac{\mu - \mu_n}{\mu_n} L(p). \qquad (36)$$

Из равенств (27), (28) и (6) вытекает, что

$$\int_0^p [Q_n(t) - Q(t)]\,dt = \frac{1}{n}\sum_{i=1}^{n}\int_0^p [t - \mathrm{I}\{X_i < x_p\}] \;/\; f(x_t)\,dt + o_p(n^{-1/2}) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_0^p [t - \mathrm{I}\{X_i < x_p\}]\,dQ(t) + o_p(n^{-1/2}). \qquad (37)$$

Вычисляя последний интеграл в (37) по частям, находим

$$\int_0^p [t - \mathrm{I}\{X_i < x_t\}]\,dQ(t) = [t - \mathrm{I}\{X_i < x_t\}]\,x_t\Big|_0^p - \int_0^p Q(t)\,d[t - \mathrm{I}\{X_i < x_t\}] =$$

$$= [p - \mathrm{I}\{X_i < x_p\}]\,x_p - \mu L(p) + \int_0^p Q(t)\,d[\mathrm{I}\{X_i < x_t\}]. \qquad (38)$$

Далее имеем

$$\int_0^p Q(t)\,d[\mathrm{I}\{X_i < x_t\}] = \int_0^p F^{-1}(t)\,d[\mathrm{I}\{X_i < F^{-1}(t)\}] =$$

$$= \int_0^p F^{-1}(t)\,d[\mathrm{I}\{F(X_i) < t\}]. \qquad (39)$$

Заметим, что в (39) дифференциал $d[\mathrm{I}\{F(X_i) < t\}]$ – это дельта-функция, и в силу ее свойств можно записать

$$\int_0^p F^{-1}(t)\,d[\mathrm{I}\{F(X_i) < t\}] = F^{-1}(F(X_i))\,\mathrm{I}\{F(X_i) < p\} = X_i\,\mathrm{I}\{X_i < x_p\} \qquad (40)$$

для всех $i = 1,...,n$. Кроме того, очевидно, что

$$(\mu_n - \mu)L(p) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu L(p)). \qquad (41)$$

Наконец, согласно усиленному закону больших чисел

$$\mu_n \overset{п.н.}{\to} \mu,\ n \to \infty. \qquad (42)$$

Собирая теперь вместе соотношения (36)–(42) и используя лемму 3.1, получим требуемое представление эмпирической функции Лоренца $L_n(p)$.

Теорема 3.2. Если с.в. X имеет строго монотонную ф.р. $F(x)$, то эмпирическая функция Лоренца $L_n(p)$ при $n \to \infty$ является:

a) строго состоятельной;

b) асимптотически несмещенной;

c) если дисперсия $\sigma^2 \equiv D(X)$ с.в. X конечна, то $L_n(p)$ асимптотически нормальна с параметрами $L(p)$ и σ_L^2, где σ_L^2 определяется формулой (43).

Доказательство. *a*) Это утверждение есть непосредственное следствие представления (36), строгой состоятельности оценок μ_n (см. (42)) и $Q_n(t)$ (см. пункт *a*) теоремы 3.1 и леммы 3.1).

b) Так как $0 \le L_n(p) \le 1, n \ge 1$, с вероятностью единица, $L_n(p) \xrightarrow{\text{п.н.}} L(p)$, то, в силу теоремы Лебега о мажорируемой сходимости $E(L_n(p)) \to L(p)$ при $n \to \infty$.

c) Поскольку

$$E(X_i I\{X_i < x_p\}) = \int_0^{x_p} x\, dF(x) = \mu L(p),$$

$$E(x_p I\{X_i < x_p\}) = x_p P\{X_i < x_p\} = p\, x_p, \quad E(X_i L(p)) = \mu L(p), \ i = 1,...,n,$$

то из этих равенств и (35) вытекает, что $E(l_n(p)) = 0.$ Слагаемые в (35) независимы и одинаково распределены, поэтому

$$D(l_n(p)) = \frac{1}{\mu^2}\frac{1}{n^2}\sum_{i=1}^{n}[D(X_i I\{X_i < x_p\}) + x_p^2 D(I\{X_i < x_p\}) + L^2(p) D(X_i)] =$$

$$= \frac{1}{\mu^2}\frac{1}{n^2}\sum_{i=1}^{n}[E(X_i^2 I\{X_i < x_p\}) - \mu^2 L^2(p) + p(1-p)x_p^2 + \sigma^2 L^2(p)] = \sigma_L^2 / n,$$

где

$$\sigma_L^2 = \frac{1}{\mu^2}[E(X^2 I\{X < x_p\}) + L^2(p)(\sigma^2 - \mu^2) + p(1-p)x_p^2]. \qquad (43)$$

С.в. $l_n(p)$ представляет собой нормированную сумму независимых одинаково распределенных с.в. и согласно центральной предельной теореме асимптотически нормальна с параметрами 0 и σ_L^2 / n. Следовательно, в силу теоремы Слуцкого [4] из разложения (34) вытекает требуемое утверждение.

Следствие 3.2. Для оценки дисперсии σ_L^2 предельного распределения надо в выражении (43) заменить теоретические характеристики их состоятельными оценками, которые будем обозначать символом $(*)$:

$$\mu^* = \mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad (\sigma^*)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad x_p^* = F_n^{-1}(p) = X_{(k)},$$

$$L^*(p) = L_n(p) = \frac{\sum_{i=1}^{k} X_{(i)}}{\sum_{i=1}^{n} X_i}, \quad E^*(X^2 I\{X < x_p\}) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 I\{X_i < X_{(k)}\},$$

где, как и раньше, $k = \begin{cases} np, & \text{если } np - \text{целое}, \\ [np]+1 & \text{иначе.} \end{cases}$

Используя формулы (13), (14) и порядковые статистики для представления эмпирической ф.р. $F_n(x)$, получим следующую оценку индекса Джини:

$$G_n = 1 - 2\int_0^1 L_n(p)\,dp = \frac{1}{\mu_n}\int_0^\infty x\,d(F_n(x))^2 - 1 =$$

$$= \frac{1}{\bar{x}}\sum_{i=1}^{n} X_{(i)}\left[\left(\frac{i}{n}\right)^2 - \left(\frac{i-1}{n}\right)^2\right] - 1 = \frac{2}{n^2\bar{x}}\sum_{i=1}^{n} X_{(i)}(i - \tfrac{1}{2}) - 1, \qquad (44)$$

где $\bar{x} \equiv \mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ – выборочное арифметическое среднее.

С помощью формул (15)–(20) можно получить другие эквивалентные представления эмпирического индекса Джини G_n.

Лемма 3.3 ([8]). Если выполнены условия пункта *c* теоремы 3.1, то

$$G_n = G + g_n + o_p(n^{-1/2}), \qquad (45)$$

где

$$g_n = \frac{2}{\mu}\frac{1}{n}\sum_{i=1}^{n}\left[-\frac{I}{\mu}(X_i - \mu) + X_i F(X_i) - m(X_i) - 2I + \mu\right], \qquad (46)$$

$$I = \int_0^\infty xF(x)\,dF(x), \; m(x) = \int_0^x t\,dF(t). \qquad (47)$$

Теорема 3.3. Если ф.р. $F(x)$ с.в. X строго монотонна, то эмпирический индекс Джини G_n при $n \to \infty$ является:

a) строго состоятельным;

b) асимптотически несмещенным;

c) если дисперсия $D(X) < \infty$, то G_n асимптотически нормален с параметрами G и σ_G^2 / n, где σ_G^2 определяется формулой (48).

Доказательство. Утверждения *a* и *b* непосредственно вытекают из первой части формулы (44) и аналогичных утверждений пунктов *a* и *b* теоремы 3.2. Для *c* имеем

$$E(X_i F(X_i)) = \int_0^\infty xF(x)\,dF(x) \equiv I\,,$$

$$E(m(X_i)) = \int_0^\infty m(x)\,dF(x) = \int_0^\infty \int_0^x t\,dF(t)\,dF(x) = \int_0^\infty t\left(\int_t^\infty dF(x)\right)dF(t) =$$

$$= \int_0^\infty t(1-F(t))\,dF(t) = E(X(1-F(X)) = \mu - I,\ i = 1,...,n.$$

Вычисляя математическое ожидание от обеих частей равенства (46) и используя полученные выше соотношения, находим, что $E(g_n) = 0$.

Из формул (15) и (25) следует, что $2I/\mu = G$. Так как слагаемые в (46) независимы и одинаково распределены, то

$$D(g_n) = \frac{1}{\mu^2 n^2}\sum_{i=1}^n D\{-(G+1)X_i + 2[X_i F(X_i) - m(X_i)]\} =$$

$$= \frac{1}{n}\frac{1}{\mu^2}\left(\sigma^2(G+1)^2 + 4D[XF(X) - m(X)]\right) = \frac{\sigma_G^2}{n},$$

где

$$\sigma_G^2 = \frac{1}{\mu^2}\left\{\sigma^2(G+1)^2 + 4D[XF(X) - m(X)]\right\}. \qquad (48)$$

С.в. g_n представляет собой нормированную сумму независимых одинаково распределенных с.в. и в силу центральной предельной теоремы асимптотически нормальна с параметрами 0 и σ_G^2/n. Поэтому из теоремы Слуцкого [4] и представления (45) следует требуемое утверждение.

Следствие 3.3. Для нахождения состоятельной оценки предельной дисперсии σ_G^2, как и в следствии 3.2, теоретические характеристики в (48) заменим их состоятельными оценками. Величины μ, σ^2 и G оценены в следствии 3.2:

$$\mu^* = \mu_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad (\sigma^*)^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2,$$
$$G^* = \frac{2}{\mu^* n^2}\sum_{i=1}^n X_{(i)}\left(i - \tfrac{1}{2}\right) - 1. \qquad (49)$$

Для оценивания дисперсии $D[XF(X)-m(X)]$ выпишем

$$D[XF(X)-m(X)]=E[XF(X)-m(X)]^2-(E[XF(X)-m(X)])^2=I_1-I_2^2.$$

Имеем

$$I_1=E[X^2F^2(X)]-2E[XF(X)m(X)]+E[m^2(X)]=$$

$$=\int_0^\infty x^2F^2(x)dF(x)-2\int_0^\infty xF(x)\left\{\int_0^x ydF(y)\right\}dF(x)+\int_0^\infty\left\{\int_0^x ydF(y)\right\}^2dF(x)=$$

$$=\frac{1}{3}\int_0^\infty x^2dF^3(x)-\int_0^\infty x\left\{\int_0^x ydF(y)\right\}dF^2(x)+\int_0^\infty\left\{\int_0^x ydF(y)\right\}^2dF(x).$$

Аналогично

$$I_2=\frac{1}{2}\int_0^\infty xdF^2(x)-\int_0^\infty\left\{\int_0^x ydF(y)\right\}dF(x).$$

Следовательно, состоятельная оценка для I_1 будет

$$I_1^*=\frac{1}{3}\sum_{i=1}^n X_{(i)}^2\left[\left(\frac{i}{n}\right)^3-\left(\frac{i-1}{n}\right)^3\right]-\sum_{j=1}^n\left\{\sum_{i=1}^j X_{(i)}\right\}X_{(j)}\left[\left(\frac{j}{n}\right)^3-\left(\frac{j-1}{n}\right)^3\right]+$$

$$+\frac{1}{n}\sum_{k=1}^n\left\{\frac{1}{n}\sum_{i=1}^k X_{(i)}\right\}\left\{\frac{1}{n}\sum_{j=1}^k X_{(j)}\right\}=\frac{1}{3n^2}\sum_{i=1}^n X_{(i)}^2(3i^2-3i+1)-$$

$$-\frac{1}{n^3}\sum_{j=1}^n\left\{\sum_{i=1}^j X_{(i)}\right\}X_{(j)}(2j-1)+\frac{1}{n^3}\sum_{k=1}^n\left\{\sum_{i=1}^k X_{(i)}\right\}\left\{\sum_{j=1}^k X_{(j)}\right\}. \qquad (50)$$

Состоятельная оценка I_2 задается посредством

$$I_2^*=\frac{1}{2}\sum_{i=1}^n X_{(i)}\left[\left(\frac{i}{n}\right)^2-\left(\frac{i-1}{n}\right)^2\right]-\frac{1}{n^2}\sum_{j=1}^n\left\{\sum_{i=1}^j X_{(i)}\right\}=$$

$$=\frac{1}{2n^2}\sum_{i=1}^n X_{(i)}(2i-1)-\frac{1}{n^2}\sum_{j=1}^n\left\{\sum_{i=1}^j X_{(i)}\right\}. \qquad (51)$$

Собирая теперь оценки (49)–(51) в формулу (48), получим состоятельную оценку $(\sigma_G^*)^2$ предельной дисперсии σ_G^2 оценки G_n индекса Джини G.

Замечание. Состоятельность и асимптотическую несмещенность x_p^*, $L_n(p)$ и G_n можно установить непосредственно из асимптотической нормальности этих оценок. Однако для этого необходимо усилить требования на распределение с.в. X: вместо строгой монотонности ф.р. $F(x)$ потребовать выполнение условия *c* теорем 3.1–3.3.

ЗАКЛЮЧЕНИЕ

На сегодняшний день задачи оценивания степени неравенства возникают в самых разнообразных областях научного знания, связанных с экономикой, информатикой, медициной, биологией и т. д. Наличие хорошо развитого аппарата анализа проблемы неравенства на основе кривых Лоренца дает исследователям эффективный инструмент, имеющий как качественное, так и количественное обоснование. В работе рассмотрены как традиционные, так и специфические характеристики кривых Лоренца, изучены свойства количественных показателей неравенства типа индексов Джини и Пьетра, сформулированы и доказаны полезные в статистическом смысле свойства асимптотической несмещенности, асимптотической нормальности и строгой состоятельности для соответствующих теоретическим эмпирических кривых Лоренца и индекса Джини. Большинство доказательств носят конструктивный характер, что позволяет использовать схожие подходы для исследования более широкого класса задач.

СПИСОК ЛИТЕРАТУРЫ

1. *Боровков А.А.* Математическая статистика. – Новосибирск: Наука: Изд-во Ин-та математики, 1997. – 772 с.

2. *Дэйвид Г.* Порядковые статистики. – М.: Наука, 1979. – 336 с.

3. *Кендалл М., Стьюарт А.* Теория распределений. – М.: Наука, 1966. – 588 с.

4. *Рао С.Р.* Линейные статистические методы и их применения. – М.: Наука, 1968. – 548 с.

5. *Уилкс С.* Математическая статистика. – М.: Наука, 1967. – 632 с.

6. *Bahadur R.R.* A note on quantiles in large samples // The Annals of Mathematical Statistics. – 1966. – Vol. 37 (3). – P. 577–580.

7. *Bhattacharya D.* Inference on inequality from household survey data // Journal of Econometrics. – 2007. – Vol. 137 (2). – P. 674–707.

8. *Davidson R.* Reliable inference for the Gini index // Journal of Econometrics. – 2009. – Vol. 150 (1). – P. 30–40.

9. *Gastworth J.L.* A general definition of the Lorenz curve // Econometrica. – 1971. – Vol. 39 (6). – P. 1037–1039.

10. *Ghost J.K.* A new proof of the Bahadur representation of quantities and an application // The Annals of Mathematical Statistics. – 1971. – Vol. 42. – P. 1957–1961.

11. *Gini C.W.* Variabilita emutabilita. – Bologna: P. Cuppini, 1912.

12. *Giorgi G.M., Nadarajah S.* Bonferroni and Gini indicies for various parametric families of distributions // METRON. – 2010. – Vol. 68. – P. 23–46.

13. *Goldie C.M.* Convergence theorems for empirical Lorenz curves and their inverses // Advances in Applied Probability. – 1977. – Vol. 9. – P. 756–791.

14. *Hoeffding W.A.* A class of statistics with asymptotically normal distribution // Annals of Mathematical Statistics. – 1948. – Vol. 19. – P. 293–325.

15. *Lorenz M.O.* Methods of measuring the concentration of wealth // Publications of the American Statistical Association. – 1905. – Vol. 9 (70). – P. 209–219.

16. *Pietra G.* Delle relazioni tra gli indici di variabilita // Atti del Regio Istituto veneto di scienze, lettere ed arti. – 1915. – Vol. 74. – P. 775–792.

17. *Sarabia J.M.* Parametric Lorenz curves: models and applications // Modeling income distributions and Lorenz Curves / ed. by D. Chotikapanich. – New York: Springer, 2008. – P. 167–190.

Семенов Дмитрий Александрович, информатик-экономист по специальности «Прикладная математика (в экономике)», ведущий специалист отдела развития Федеральной кадастровой палаты по Новосибирской области. Области научных интересов: финансовая и актуарная математика, информационные технологии, теория риска. Имеет одну научную публикацию. E-mail: miktor.semenov@gmail.com

Щеколдин Владислав Юрьевич, кандидат технических наук, доцент кафедры маркетинга и сервиса Новосибирского государственного технического университета. Основные направления научных исследований: экономико-математическое моделирование, статистика, планирование оптимальных экспериментов, логистика, эконометрика, маркетинговые исследования. Автор более 75 научных статей. E-mail: raix@mail.ru

Semenov Dmitry A., informatics-economist specializing in applied mathematics (in economics), leading specialist of the development department of the Federal Cadastral Chamber in the Novosibirsk region. His research interests include financial and actuarial mathematics, information technology, and risk theory. He has 1 scientific publication. E-mail: miktor.semenov@gmail.com

Shchekoldin Vladislav Yu., PhD (Eng.), associate professor, department of marketing and service, Novosibirsk State Technical University. The main areas of his research are economic and mathematical modeling, statistics, planning of optimal experiments, logistics, econometrics, and marketing research. He is the author of over 75 scientific articles. E-mail: raix@mail.ru

DOI: 10.17212/1814-1196-2020-4-121-144

Theoretical and empirical Lorenz functions, Gini indices, and their properties[*]

D.A. SEMENOV[1,a], *V.Y. SHCHEKOLDIN*[2,b]

[1] *Federal Cadastral Chamber of the Novosibirsk Region, 4G. Kutateladze Street, Novosibirsk, 630128, Russian Federation*

[2] *Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation*

[a] *miktor.semenov@gmail.com* [b] *raix@mail.ru*

Abstract

The issues of assessing the fairness and efficiency of the distribution of the total income of society between different groups of the population have attracted attention of scientists for a long time. They became most relevant at the end of the 19th – beginning of the 20th centuries in connection with the intensive stratification of countries with various political and social sys-

[*] *Received 06 May 2020.*

tems caused by the intensive development of the economy, science and technology. The Lorenz function and the Lorenz curve, as well as the Gini index, are commonly used for theoretical research and applications in the economic and social sciences. These tools were originally introduced to describe and study the inequality in the incomes and wealth distribution among a given population. Nowadays they have found wide application in such fields as demography, insurance, healthcare, the risk and reliability theory, as well as in other areas of human activities. In this paper we present the properties of the Lorentz function and various representations of the Gini index, systematize the analytical results for uniform, exponential, power-law (types I and II) and lognormal distributions, as well as for the Pareto distribution (types I and II). Additionally, the issue of estimating inequality based on the Pietra index and its relationship with the Lorentz function was studied. Nonparametric estimates of the Lorentz function and the Gini index based on a sample from the corresponding distribution are considered. Strict consistency and asymptotic unbiasedness of these estimates are shown under certain conditions for the initial distribution with an increase in the sample size. On the basis of the method of linearization of estimates, the asymptotic normality of the empirical Lorentz function and the empirical Gini index is determined.

Keywords: inequality estimation, Lorentz function, Lorentz curve, Gini index, Pietra index, linearization of estimates, strict consistency, asymptotic unbiasedness, normality

REFERENCES

1. Borovkov A.A. *Matematicheskaya statistika* [Mathematical statistics]. Novosibirsk, Nauka Publ., 1997. 772 p.

2. David H.A. *Order statistics*. New York, Wiley and Sons, 1970. 272 p. (Russ. ed.: Deivid G. *Poryadkovye statistiki*. Moscow, Nauka Publ., 1979. 336 p.

3. Kendall M., Stuart A. *Advanced theory of statistics*. Vol. 1. *Distribution theory*. 2nd ed. London, Griffin, 1963. 433 p. (Russ. ed.: Kendall M., St'yuart A. *Teoriya raspredelenii*. Moscow, Nauka Publ., 1966. 588 p.).

4. Rao C.R. Linear statistical inference and its applications. 2nd ed. New York, Wiley and Sons, 1965. 522 p. (Russ. ed.: Rao S.R. *Lineinye statisticheskie metody i ikh primeneniya*. Moscow, Nauka Publ., 1968. 548 p.).

5. *Wilks C.C. Mathematical statistics*. New York, Wiley and Sons, 1962. 644 p. (Russ. ed.: Uilks S. *Matematicheskaya statistika*. Moscow, Nauka Publ., 1967. 632 p.).

6. Bahadur R.R. A note on quantities in large samples. *The Annals of Mathematical Statistics*, 1966, vol. 37 (3), pp. 577–580.

7. Bhattacharya D. Inference on inequality from household survey data. *Journal of Econometrics*, 2007, vol. 137 (2), pp. 674–707.

8. Davidson R. Reliable inference for the Gini index. *Journal of Econometrics*, 2009, vol. 150 (1), pp. 30–40.

9. Gastworth J.L. A general definition of the Lorenz curve. *Econometrica*, 1971, vol. 39 (6), pp. 1037–1039.

10. Ghost J.K. A new proof of the Bahadur representation of quantities and an application. *The Annals of Mathematical Statistics*, 1971, vol. 42, pp. 1957–1961.

11. Gini C.W. *Variabilita emutabilita*. Bologna, P. Cuppini, 1912.

12. Giorgi G.M., Nadarajah S. Bonferroni and Gini indicies for various parametric families of distributions. *METRON*, 2010, vol. 68, pp. 23–46.

13. Goldie C.M. Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability*, 1977, vol. 9, pp. 756–791.

14. Hoeffding W.A. A class of statistics with asymptotical normal distribution. *Annals of Mathematical Statistics*, 1948, vol. 19, pp. 293–325.

15. Lorenz M.O. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 1905, vol. 9 (70), pp. 209–219.

16. Pietra G. Delle relazioni tra gli indici di variabilita. *Atti del Regio Istituto veneto di scienze, lettere ed arti*, 1915, vol. 74, pp. 775–792.

17. Sarabia J.M. Parametric Lorenz curves: models and applications. *Modeling income distribution and Lorenz Curves*. Ed. by D. Chotikapanich. New York, Springer, 2008, pp. 167–190.

Для цитирования:

Семенов Д.А., Щеколдин В.Ю. Теоретические и эмпирические функции Лоренца, индексы Джини и их свойства // Научный вестник НГТУ. – 2020. – № 4 (80). – С. 121–144. – DOI: 10.17212/1814-1196-2020-4-121-144.

For citation:

Semenov D.A., Shchekoldin V.Yu. Teoreticheskie i empiricheskie funktsii Lorentsa, indeksy Dzhini i ikh svoistva [Theoretical and empirical Lorenz functions, Gini indices, and their properties]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta = Science bulletin of the Novosibirsk state technical university*, 2020, no. 4 (80), pp. 121–144. DOI: 10.17212/1814-1196-2020-4-121-144.