получаются n (по числу значений атрибута) подмножеств и, соответственно, создаются n потомков корня, каждому из которых поставлено в соответствие свое подмножество, полученное при разбиении множества T. Затем эта процедура рекурсивно применяется ко всем подмножествам (потомкам корня) и т.д. В результате получаем правила вида:

$$Если\ y_1 > 85\ \ и\ y_2 > 80\ и\ y_3 > 87\ и\ y_4 > 81 \quad то\ y = высокий.$$

Построенное дерево решений используется для распознавания нового объекта. Обход дерева решений начинается с корня дерева. На каждом внутреннем узле проверяется значение объекта по атрибуту, который соответствует проверке в данном узле, и, в зависимости от полученного ответа, находится соответствующее ветвление, и по этой дуге двигаемся к узлу, находящему на уровень ниже и т.д. Обход дерева заканчивается, как только встретится узел решения, который и дает название класса объекта.

Такая же методика применяется, когда дерево используется для классификации новых примеров. Если на каком-то узле дерева при выполнении проверки выясняется, что значение соответствующего атрибута классифицируемого примера пропущено, то алгоритм исследует все возможные пути вниз по дереву и определяет, с какой вероятностью пример относится к различным классам. В этом случае, «классификация» — это скорее распределение классов. Как только распределение классов установлено, то класс, имеющий наибольшую вероятность появления, выбирается в качестве ответа дерева решений. Процедура построения деревьев решений была проведена с использованием аналитической платформы Deductor (см.: http://basegroup.ru/).

В результате система правил продукции позволила исследовать возможности оценки и анализ компонентов профессиональной компетентности студента. Преимуществом построенной модели является то, что после каждой процедуры контроля оценки уровня сформированности профессиональных компетенций преподаватель и студент может получать индивидуальную диаграмму уровней сформированности компетенций.

Градуировка коэффициента Джини
(Памяти В.И. Арнольда (1937–2010))[1]

Шмерлинг Дмитрий Семенович, *НИУ ВШЭ*

Проблема неравенства в доходах хорошо известна, по крайней мере, с работы Макса Лоренца об измерении концентрации богатства[2]. Один из наиболее распространенных методов измерения неравенства стал расчет коэффициента (индекса) Джини:

$$\Delta_1 = \frac{1}{N(N-1)} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \left| x_j - x_k \right| f(x_j) f(x_k) \qquad (1)$$

(для дискретного случая, без учета совпадений), где x_1, x_2, … — величина доходов, $f(x_1)$, $f(x_2)$ — вероятность (или частота по выборке) людей с доходами x_1, x_2, … соответственно.

Величину Δ_1 обычно нормируют так, чтобы $\Delta_1^* = \Delta_1 / \Delta_1^{max} \in [0,1]$, при этом, чем она больше (ближе к 1), тем значительней неравенство населения.

Как известно, удвоенная площадь между диагональю и кривой рассеяния равно коэффициенту Джини. Заметим, что площадь над кривой рассеяния (Лоренца) и под диагональю[3] равна $\Delta_1/4$

[1] Автор благодарит В.И. Арнольда, А.Я. Кируту, Я.Ю. Никитина, А.И. Орлова, Ю.Н. Толстову, Ю.Н. Тюрина, В.В. Ульянова за содействие и обсуждение.

[2] Lorenz M.D. Methods of Measuring the Concentration of Wealth // Publ. Amer. Statist. Ass. 1905. Vol. 9. No. 70. P. 209–219; Кендалл М.Дж., Стьюарт А. Теория распределения. М.: Наука, 1966. § 2.25. Впрочем, о неравенстве писал и В.И. Ленин в работе «Развитие капитализма в России» (1899).

[3] Кендалл М.Дж., Стьюарт А. Теория распределения. М.: Наука, 1966. С. 75. Рис. 2.2.

μ_1, где $\mu_1 = \int_{-\infty}^{\infty}(x-a)f(x)dx$, f(x) — плотность распределения, обычно организуют a=0[1]. Собственно кривая рассеяния есть «неполный первый момент распределения»[2]:

$$\phi(x) = \frac{1}{\mu_1} \int_{-\infty}^{x} xf(x)dx \quad (2)$$

Коэффициент Джини вычисляется и публикуется для большинства стран уже десятки лет, в т.ч. со группированными данными[3]. К примеру в Норвегии он составлял 0,25 (2008 год), во Франции 0,32 (2008 год), в России 0,42 (2008 год), в США 0,45 (2007 год), в Мексике 0,48 (2008 год), в Южной Африке 0,65 (2005 год), в Намибии 0,71 (2003 год)[4].

Коэффициент Джини измеряет величину дифференциации доходов населения, «богатств», расходов, ВВП регионов или стран и тому подобных показателей, которые по-английски все вместе называются «Size» (эквивалентный русский термин отсутствует). Значения, большие, чем 0,3–0,4, по мнению большинства специалистов, свидетельствуют о высоком неравенстве и приводят к замедлению темпов развития стран, например, вследствие «ловушки бедности» [5]. Читатель-экономист, может быть, уже привык к таким данным, но насколько обществом понят смысл значений коэффициента Джини?

Существует обширная литература о вреде высокого (> 0,3) коэффициента Джини[6]. Однако граница 0,3 выбрана достаточно произвольно и представляет что-то около среднего общеевропейского коэффициента Джини. И было бы полезно поискать за значениями нормированного коэффициента G', $0 \le G' \le 1$, какой-нибудь «предметный» (например, экономический) смысл.

Рассмотрим следующую модель. Пусть x_i — доход лиц, относящихся к i-му уровню иерархии применительно к компании, населению территории и т.п. Модель P такова, что доход на i-м уровне (i = 1 — лица с наименьшим, а i = n — с наибольшими доходами) равен $x_i = ki^m$, m = 1, 2, 3, …, k > 0.

Теорема: коэффициент Джини $G_m'(n)$ для модели P равен асимптотически при $n \to \infty$

$$G_m'(n) = \frac{m}{m+2} \quad (1a)$$

Набросок доказательства:

$$G'(n) = \frac{"\sigma"(n)}{\max\limits_x "\sigma"(n)}, \quad (2a)$$

где maximum берется по всем возможным $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$, таким, что

$$\sum_{1 \le i \le n} x_{(i)} = C(n), \quad (3)$$

$$"\sigma"'(n) = \frac{2\sqrt{\pi}}{n(n-1)} \sum_{1 \le i \le n} (i - \frac{n+1}{n}) x_{(i)}, \quad (4)$$

---

[1] Кендалл М.Дж., Стьюарт А. Теория распределения. М.: Наука, 1966. § 2.3.

[2] Кендалл М.Дж., Стьюарт А. Теория распределения. М.: Наука, 1966. С. 75–77. § 2.31.

[3] Gastwirth Y.L. The Estimation of the Lorenz Curve and Gini Index // Rev. of Econ. Statistics. 1972. Vol. 52. No. 3. P. 306–316; Moderres R., Gastwirth J.L. A Cautionary Note on Estimating the Standard Error of the Gini Index of Inequality // Oxford Bull. of Econ. Statist. 2006. Vol. 68. P. 385-390; Morgan J. The Anatomy of Income Distribution // Rev. of Econ.Statist. 1962. Vol. 44. No 3. P. 270–283.

[4] Wikipedia. List of Countries by income equality // http://en.wikipedia.org/wiki/List_of_countries_by_income_equality

[5] См., например: Atkinson A.B., Bourguignon F. (ed.) Handbook of Income Distribution. Amsterdam et al.: Elsevier, 2000. Vol. 1.

[6] См., например: Atkinson A.B., Bourguignon F. (ed.) Handbook of Income Distribution. Amsterdam et al.: Elsevier, 2000. Vol. 1.

$X_{(i)}$ – i-ая порядковая статистика[1]

$$"\sigma"\ (n) = \frac{1}{2}\sqrt{\pi}G(n), \quad (5)$$

$$G(n) = \frac{1}{n(n-1)} \sum_{1 \le i.j \le n} \left| x_i - x_j \right|. \quad (6)$$

Используя другую форму "σ"(n)= "σ"

$$"\sigma" = \frac{\widetilde{\kappa} 2 * \sqrt{\pi}}{n(n-1)} \left\{ \sum_{1}^{n} {}_i x_{(i)} - \frac{n+1}{n} \sum_{1}^{n} x_{(i)} \right\} \quad (7)$$

и вычисляя

$$\max_x "\sigma" = \frac{\sqrt{\pi} * \widetilde{\kappa}}{n} S_m(n) \quad (8)$$

где $S_m(n) = \sum_{k=0}^{n-1} k^m$ [2]

сумма целых чисел в степени m=1, 2, 3,…, k = 0, 1, 2,…., n–1, можно получить выражение для (3). Именно, при

$$\widetilde{S}_m(n) = S_m(n) + n^m,$$

$$"\sigma" = \frac{2\widetilde{\kappa}\sqrt{\pi}}{n(n-1)} \left\{ \widetilde{S}_{m+1}(n) - \frac{n+1}{n} \widetilde{S}_m(n) \right\} / \frac{\widetilde{\kappa}\sqrt{\pi}}{n} \widetilde{S}_m(n) \quad (9)$$

Откуда (здесь m – напоминает о степени многочлена в формуле $x_i = ki^m$, m = 1, 2, 3, …, k > 0)

$$G'_m(n) = \frac{2}{n-1} \left\{ \frac{\widetilde{S}_{m+1}(n)}{\widetilde{S}_m(n)} - \frac{n+1}{2} \right\} \quad (10)$$

Теперь нам понадобятся выражения $S_i(n)$, приведенные в удобной форме в уже цитировавшейся великолепной книге Р. Грехэма с соавторами[3]:

$$\widetilde{S}_m(n) = \frac{1}{m+1} \sum_{0 \le k \le n} \binom{m+1}{k} B_k n^{m+1-k} \quad (11),$$

где B_k, k = 0, 1, 2,…., числа Якова Бернулли, а именно:

k	1	2	3	4	5	6	7	8	9	10	11	12	…
B_k	1	$-\frac{1}{2}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$	0	$\frac{5}{66}$	0	$-\frac{691}{2730}$	…

[1] См.: Дейвид Г. Порядковые статистики: Пер. с англ. М.: Наука, 1978. С. 187–189 (§7.4), 214 (§9.6), где обсуждается асимптотическая нормальность "σ". При этом $E"\sigma" = 2\sqrt{\pi \int_{-\infty}^{+\infty} x\left[P_{(x)} - \frac{1}{2}\right] dP_{(x)}}$, "σ" — несмещенная оценка для σ в случае нормальных выборок, $P_{(x)}$ — функция распределения.

[2] См.: Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики / Пер. с англ.; 3-е изд. М.: БИНОМ; Лаборатория знаний; СМР, 2009.

[3] Там же. § 61.78.

Формулы Sm(n), m = 0, 1, 2,…, 10 см. в упомянутой книге с. 314. Из (10) легко получается при n→∞

$$G'_m(n) \cong \frac{2}{n-1}\left\{\frac{m+1}{m+2}n - \frac{n+1}{2}\right\} \approx \frac{m}{m+2}, QED.$$

Приведем таблицу 2 для $G_m'(n)$[1]

m	1	2	3	4	5	6	7	8	9	10	…
$G_m'(n)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{5}$	$\frac{2}{3}$	$\frac{5}{7}$	$\frac{3}{4}$	$\frac{7}{9}$	$\frac{4}{5}$	$\frac{9}{11}$	$\frac{5}{6}$	…

Теперь проинтерпретируем наши результаты. Леопольд Кронекер (1823–1891) не зря говорил, что целые числа придумал Бог, а остальное — люди. В нашем случае для модели P с помощью целых значений *m* любые совокупности сообществ, компаний, стран, регионов можно аналитически (условно) разделить на линейные (m≈1), квадратичные (m≈2), кубические (m≈3), «тетричные» (m≈4), «пентальные» (m≈5) и т.д.

Таким образом, Норвегию, у которой G'=0,25 и m<1, по распределению доходов можно отнести к сублинейным странам, Францию при G'=0,327 и *m*≈1 — к линейным, Мексику при G'=0,482 и m≈2 — к квадратичным; Россия с G'=0,423 попадает между Францией и Мексикой (1<*m*<2), существенно отставая, например, от Гаити (G'=0.538, 2<*m*<3), Сьерра-Леоне (G'=0,629, 3<*m*<4) и Намибии (5<*m*<6, по разным данным 0,707< G'<0,750).

В этой градуировке Москва может претендовать на кубический тип распределения доходов, поскольку, по данным официальной статистики, G' доходил до 0,62, а по мнению многих специалистов, реальные значения G' в Москве находятся в интервале 0,60–0,70.

Аналогичные расчеты по ведущим российским компаниям, проведенные по данным годовых отчетов, публикуемым газетой «Ведомости»[2], указывают на величину 2<m<3 в 2009 г. При величине ежемесячной зарплаты топ-менеджера в $1,5–3,0 тыс. внутрикорпоративная *m* может указывать на линейность в распределении доходов, однако с учетом бонусов порядка в $1–3 млн. в год *m* может достигать и 4.

Здесь требуется обсуждение. Можно увязать обсуждаемую модель с традиционными статистическими распределениями. Для распределения Парето с таким же G^, как в нашей модели P, лишь степень m<1 обеспечивает конечную дисперсию, а для лог-логистического распределения для того же требуется m<2.

Что касается лог-нормального распределения, то дисперсия не стремится к бесконечности, а лишь медленно растет при росте m. Заметим, что распределения Парето и лог-логистическое хорошо описывают правый (верхний) хвост распределения доходов, а лог-нормальное хорошо описывает не слишком большие доходы, но плохо описывает правый хвост.

Интерпретация может быть следующей: при высокой степени неравенства в модели P малая (богатая) часть общества стремится увеличить свои доходы, так что верхние хвосты распределения утяжеляются и дисперсия стремится к бесконечности. В тоже время средняя по доходам часть общества медленно реагирует на рост степени модели m.

[1] При m=1 выражении для индекса Джини точное, при всех n = 1, 2, 3….

[2] См.: Милек О. Изучение распределения дохода с помощью распределения с тяжелыми хвостами: магистерская диссертация / ГУ–ВШЭ. М., 2010.