В. А. Силаева, А. М. Силаев

Национальный исследовательский университет
Высшая школа экономики – Нижний Новгород

АЛГОРИТМЫ КОРРЕКЦИИ БАЛЛОВ ЕГЭ

1. Введение

Статистические данные по результатам сдачи Единых государственных экзаменов (ЕГЭ) в Российской Федерации демонстрируют неоднородность оценок – различие статистических характеристик для баллов по разным предметам и для разных лет. Например, в табл. 1 представлены средние тестовые баллы по разным предметам в 2014 и 2015 гг. со страницы официального информационного портала ЕГЭ ege.edu.ru/ru/main/satistics-ege/.

Таблица 1. Средний тестовый балл ЕГЭ в 2014-2015 гг.

Общеобразовательный предмет	**2014**	**2015**
Русский язык	62,5	65,9
Математика (профильный уровень)	46,4	45,4
Физика	45,4	51,2
Химия	55,3	56,3
Информатика и ИКТ	57,1	53,6
Биология	54,1	53,2
История	45,3	46,7
География	52,9	52,9
Английский язык	62,8	64,8
Обществознание	55,4	53,3
Литература	53,6	56,9

Из табл. 1 видно, что разница между средними баллами для английского языка и истории в 2014 г. составляла 17,5 баллов, а для русского языка и математики в 2015 г. составляла 20,5 баллов. Средние баллы ЕГЭ (в расчете на один предмет) зачисленных в вуз студентов рассматриваются в качестве одного из показателей для оценки эффективности вузов Министерством образования и науки РФ, по средним баллам поступивших студентов составляются рейтинги вузов и оценивается качество приема на различные образовательные программы. Поскольку баллы по разным предметам имеют разные вероятностные распределения, использование только средних баллов при сравнении результатов ЕГЭ не вполне корректно. Возникает задача коррекции баллов ЕГЭ, чтобы сделать оценки более однородными, и использования более точных инструментов для сопоставления качества подготовки выпускников школ и абитуриентов, зачисленных на разные программы обучения в вузах.

Данная задача актуальна, так как баллы ЕГЭ характеризуют уровень образования выпускников школ и используются в эконометрических исследованиях проблем экономики образования в нашей стране. В статьях (Польдин, 2011; Пересецкий, Давтян, 2011; Замков, Пересецкий, 2013) изучается связь ЕГЭ и успеваемость студентов на первых курсах бакалавриата. В статье (Андрущак и др., 2012) исследуется проявление эффектов сообучения в студенческой группе, при этом результаты ЕГЭ по русскому языку и математике используются для оценки способностей студентов. Оценке результатов приема на образовательные программы в вузах по результатам ЕГЭ посвящены статьи (Польдин, Силаев, 2011; Польдин и др., 2014), в которых предлагается использовать для анализа кривые спроса, полученные в результате сортировки по убыванию суммы баллов ЕГЭ студентов, зачисленных на программы обучения, а также рассчитываемые на их основе количественные характеристики.

2. Уравнения алгоритмов коррекции баллов

Предположим, что известны баллы $x_1, x_2, \dots, x_m$, упорядоченные по возрастанию из интервала $[0;100]$, и количество учащихся $n_1, n_2, \dots, n_m$, набравших соответствующие баллы, для какой-либо отдельной дисциплины ЕГЭ в России. Можно вычислить эмпирические оценки вероятностей получения того или иного балла

$$p_i = n_i / n\,,\ i = 1, 2, \dots, m\,. \tag{1}$$

Здесь $n = \sum_{i=1}^{m} n_i$ – общее число учащихся, сдававших данный предмет.

Введем вспомогательную случайную величину X такую, что она имеет непрерывную интегральную функцию распределения $F(x)$ со значениям в точках $x_1, x_2, \dots, x_m$ равными

$$F(x_k) = \sum_{i=1}^{k} p_i\,,\ k = 1, 2, \dots, m\,. \tag{2}$$

Функцию распределения $G(y)$ случайной величины Y можно выбрать на интервале баллов $0 \le y \le 100$ произвольной, но с условием, чтобы существовала обратная функция $y = G^{-1}(t)$ при $0 \le t \le 1$. Метод нелинейной коррекции баллов для каждой отдельной дисциплины основан на следующем известном свойстве нелинейных преобразований случайных величин: если две случайные величины X и Y имеют функции распределения вероятностей $F(x)$ и $G(y)$, то третья случайная величина Z, полученная с помощью нелинейного преобразования $Z = G^{-1}(F(X))$, будет распределена как случайная величина Y.

Используем данное нелинейное преобразование для коррекции баллов в соответствии с формулой:

$$x_k^c = G^{-1}(F(x_k)) = G^{-1}\left(\sum_{i=1}^{k} p_i\right),\ k = 1, 2, \dots, m\,. \tag{3}$$

Эта формула дает для каждой фактической оценки ЕГЭ x_k по предмету в интервале от 0 до 100 соответствующую скорректированную оценку x_k^c. В целом для всех школьников в России оценки пересчитываются так, что значения $x_1^c, x_2^c, \dots, x_m^c$ и их вероятности $p_1, p_2, \dots, p_m$ описывают случайную величину Z, функция распределения которой в точках x_k^c совпадает с функцией распределения случайной величины Y:

$$P\left(Z \le x_k^c\right) = G\left(x_k^c\right) = \sum_{i=1}^{k} p_i\ ,\ k = 1, 2, \dots, m. \tag{4}$$

Если пересчитывать оценки ЕГЭ по разным предметам или для разных лет, применяя одну и ту же функцию распределения $G(y)$, то скорректированные оценки для разных предметов будут однородными, то есть описываться одинаковыми вероятностными распределениями. Поэтому их можно более обоснованно сравнивать между собой, сопоставляя результаты ЕГЭ.

Рассмотрим, некоторые возможные функции распределения $G(y) = \int_0^y g(y)dy$, где введено обозначение $g(y)$ для плотности вероятности случайной величины Y.

Равномерное распределение вероятности:

$$g_u(y) = 0{,}01,\ G_u(y) = 0{,}01y,\ 0 \le y \le 100. \tag{5}$$

В этом случае согласно (3) скорректированные баллы вычисляются по правилу

$$x_k^c = G_u^{-1}\left(\sum_{i=1}^{k} p_i\right) = 100F\left(x_k\right) = 100\sum_{i=1}^{k} p_i\ ,\ k = 1, 2, \dots, m. \tag{6}$$

Здесь p_i – вероятность получения балла x_i по рассматриваемому предмету. Количественные значения баллов x_i и их вероятности p_i по каждому предмету можно найти на сайтах сопровождающих ЕГЭ для каждого конкретного года.

Среднее значение и среднеквадратичное отклонение равномерно распределенной на интервале $[0;100]$ случайной величины Y равны $E(Y) = 50$, $\sigma_Y = 50/\sqrt{3} \approx 28{,}868$. Если учитывать целочисленность случайной величины Y и равномерность вероятностей значений в интервале $[0;100]$, то получим $E(Y) = 50$, $\sigma_Y = \sqrt{850} \approx 29{,}155$. Можно сделать вывод, что учет целочисленности баллов незначительно влияет на величину среднеквадратичного отклонения, поэтому допустимо использовать непрерывнозначную аппроксимацию равномерного распределения величины Y.

Треугольное распределение вероятности (симметричное относительно значения $y = 50$):

$$g_t(y) = \begin{cases} y/2500, & 0 \le y \le 50; \\ (100 - y)/2500, & 50 \le y \le 100; \end{cases}$$

$$G_t(y) = \begin{cases} y^2/5000, & 0 \le y \le 50; \\ 1 - (100 - y)^2/5000, & 50 \le y \le 100; \end{cases} \tag{7}$$

В этом в соответствии с формулой (3) скорректированные баллы вычисляются по правилу

$$x_k^c = G_t^{-1}(F(x_k)) = \begin{cases} 100\sqrt{\dfrac{F(x_k)}{2}}, & 0 \le F(x_k) \le 0{,}5; \\ 100\left(1 - \sqrt{\dfrac{1 - F(x_k)}{2}}\right), & 0{,}5 \le F(x_k) \le 1; \end{cases} \tag{8}$$

где $F(x_k) = \sum_{i=1}^{k} p_i$, $k = 1, 2, \ldots, m$.

Среднее значение и среднеквадратичное отклонение случайной величины Y с треугольным распределением плотности вероятности (7) равны $E(Y) = 50$, $\sigma_Y = 50/\sqrt{6} \approx 20{,}412$. Отметим, что, если учесть целочисленность Y в интервале $[0;100]$ и предположить треугольное распределения вероятностей симметричное относительно значения $y = 50$, тогда получим $E(Y) = 50$, $\sigma_Y = \sqrt{1300/3} \approx 20{,}817$. Снова можно сделать вывод, что учет целочисленности баллов незначительно влияет на величину среднеквадратичного отклонения. Поэтому можно пренебречь дискретнозначностью оценок и для вычислений использовать непрерывнозначную аппроксимацию треугольного распределения величины Y.

Бета распределение с плотностью вероятности:

$$g_b(y) = \frac{y^{\alpha-1}(100 - y)^{\beta-1}}{100^{\alpha+\beta-1} B(\alpha, \beta)}, \; 0 \le y \le 100, \tag{9}$$

где α, β – неотрицательные коэффициенты, $B(\alpha, \beta)$ – бета функция Эйлера, и интегральной функцией распределения

$$G_b(y) = \int_0^y g_b(y)dy, \quad 0 \le y \le 100.$$

Согласно (2), (3) скорректированные баллы будут вычисляться по правилу

$$x_k^c = G_b^{-1}(F(x_k)) = G_b^{-1}\left(\sum_{i=1}^{k} p_i\right). \tag{10}$$

Среднее значение и дисперсия бета распределенной случайной величины Y с плотностью вероятности (9) выражаются через параметры следующим образом

$$E(Y) = \frac{100\alpha}{\alpha+\beta}, \ \sigma_Y^2 = \frac{\alpha\beta 10^4}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{11}$$

Предположим, что $\alpha = \beta$. Тогда плотность вероятности (9) будет симметричной относительно точки $y = 50$. При этом среднее значение и среднеквадратичное отклонение величины Y будут равны

$$E(Y) = 50, \ \sigma_Y = \frac{50}{\sqrt{2\alpha+1}}. \tag{12}$$

Распределение вероятности для дискретнозначных величин:

$$g_d(y) = \sum_{i=1}^{r} q_i \delta(y - y_i), \ G_d(y) = \sum_{i=1}^{r} q_i I(y - y_i), \tag{13}$$

где параметры ограничены значениями $q_i > 0$, $\sum_{i=1}^{r} q_i = 1$, а также введены обозначения: $\delta(y - y_i)$ – дельта-функция Дирака; $I(y-y_i) = \begin{cases} 0, & y - y_i < 0 \\ 1 & y - y_i \ge 0 \end{cases}$ – индикаторная функция. В этом случае формула (3) для вычисления скорректированных баллов не работает, так как не существует однозначной обратной функции распределения $G_d^{-1}(y)$. Необходимо специально доопределить значения скорректированных оценок.

Для примера рассмотрим частный случай:

$$G_d(y) = 0.25(I(y-2) + I(y-3) + I(y-4) + I(y-5)). \tag{14}$$

Тогда правило вычисления скорректированных оценок может быть следующим:

$$x_k^c = \begin{cases} 2, & F(x_k) < 0.25; \\ 3, & 0.25 \le F(x_k) < 0.5; \\ 4, & 0.5 \le F(x_k) < 0.75; \\ 5, & 0.75 \le F(x_k) \le 1; \end{cases} \tag{15}$$

где $F(x_k) = \sum_{i=1}^{k} p_i$, $k = 1, 2, \dots, m$. Таким образом, все учащиеся, сдававшие экзамен, разбиваются на 4 квартиля каждый по 25% общей численности. Попадание в тот или иной квартиль зависит от набранного балла ЕГЭ и приводит соответствующей оценке – в данном примере от 2 до 5.

3. Сравнение результатов ЕГЭ по математике и русскому языку в 2013 г.

Применим предложенные алгоритмы коррекции баллов для анализа результатов ЕГЭ по математике и русскому языку в России в 2013 г. Статистика результатов по округам и субъектам Российской федерации представлена на сайте ФГБУ «Федеральный центр тестирования» www.rustest.ru/ege/statistics/results/.

В табл. 2 и 3 приведены значения фактических и скорректированных баллов по русскому языку и математике, вычисленных для равномерного, треугольного, бета распределения с параметрами $\alpha = \beta = 1{,}5$ (при этом $\sigma_Y = 25$) и дискретнозначного стандартных распределений вероятностей по формулам (6), (8), (10), (15).

Таблица 2. Фактические и скорректированные баллы ЕГЭ по русскому языку в 2013 г.

Балл	Равн.	Треуг.	Бета	Дискр.	Балл	Равн.	Треуг.	Бета	Балл
0	0	0	0	2	**52**	20,5	32,0	25,8	3
3	0,02	1,0	0,2	2	**53**	22,5	33,5	27,6	3
5	0,02	1,1	0,3	2	**54**	24,5	35,0	29,4	3
7	0,03	1,2	0,3	2	**55**	26,7	36,6	31,3	3
9	0,04	1,4	0,4	2	**56**	29,0	38,1	33,2	3
11	0,06	1,7	0,5	2	**57**	31,4	39,6	35,1	3
13	0,09	2,1	0,6	2	**58**	33,8	41,1	37,1	3
15	0,13	2,6	0,8	2	**59**	36,3	42,6	39,2	3
17	0,2	3,2	1,1	2	**60**	38,9	44,1	41,3	3
20	0,3	3,9	1,5	2	**61**	41,6	45,6	43,4	3
22	0,4	4,7	1,9	2	**62**	44,3	47,1	45,5	3
24	0,6	5,4	2,3	2	**63**	47,1	48,5	47,7	3
26	0,7	6,1	2,7	2	**64**	49,9	50,0	49,9	4
28	0,9	6,9	3,2	2	**65**	52,9	51,5	52,2	4
30	1,2	7,6	3,6	2	**66**	55,9	53,0	54,6	4
32	1,4	8,4	4,1	2	**67**	58,9	54,7	57,0	4
34	1,6	9,1	4,6	2	**68**	62,0	56,4	59,5	4
36	1,9	9,8	5,1	2	**69**	65,1	58,2	62,0	4
37	2,5	11,1	6,0	2	**70**	68,3	60,2	64,6	4
38	3,1	12,4	7,0	2	**71**	71,5	62,2	67,2	4
39	3,7	13,6	7,9	2	**72**	74,7	64,4	69,9	5
40	4,4	14,9	8,9	2	**73**	77,8	66,7	72,6	5
41	5,2	16,1	10,0	2	**76**	80,9	69,1	75,4	5
42	6,1	17,4	11,1	2	**79**	83,8	71,6	78,1	5
43	7,1	18,8	12,3	2	**82**	86,7	74,2	80,9	5
44	8,1	20,2	13,6	2	**84**	89,4	77,0	83,7	5
45	9,3	21,6	14,9	2	**87**	91,9	79,9	86,4	5
46	10,6	23,0	16,3	2	**90**	94,1	82,9	89,2	5
47	12,0	24,5	17,7	2	**92**	96,1	86,1	91,8	5
48	13,5	25,9	19,2	2	**95**	97,7	89,3	94,3	5
49	15,1	27,5	20,8	2	**98**	99,0	92,8	96,6	5
50	16,8	29,0	22,4	2	**100**	100	100	100	5
51	18,6	30,5	24,1	2					

Таблица 3. Фактические и скорректированные баллы ЕГЭ по математике в 2013 г.

Балл	Равн.	Треуг.	Бета	Дискр.	Балл	Равн.	Треуг.	Бета	Балл
0	0	0	0	2	**68**	86,5	74,0	80,7	5
5	0,6	5,3	2,2	2	**70**	89,7	77,3	84,0	5
10	2,1	10,1	5,3	2	**72**	92,0	80,0	86,6	5
15	3,5	13,2	7,7	2	**74**	93,8	82,4	88,8	5
20	4,9	15,6	9,6	2	**77**	95,2	84,5	90,5	5
24	6,2	17,6	11,3	2	**79**	96,2	86,3	92,0	5
28	11,0	23,5	16,7	2	**81**	97,1	87,9	93,2	5
32	16,8	28,9	22,4	2	**83**	97,7	89,4	94,3	5
36	22,8	33,8	27,9	3	**85**	98,3	90,7	95,3	5
40	29,1	38,1	33,2	3	**87**	98,7	91,9	96,1	5
44	35,4	42,1	38,4	3	**90**	99,0	93,0	96,8	5
48	41,7	45,7	43,5	3	**92**	99,3	94,1	97,4	5
52	48,4	49,2	48,7	4	**94**	99,6	95,3	98,1	5
56	55,8	53,0	54,6	4	**96**	99,7	96,3	98,6	5
60	64,6	57,9	61,6	4	**98**	99,9	97,3	99,1	5
63	74,7	64,4	70,0	5	**100**	100	100	100	5
66	80,9	69,1	75,4	5					

Чтобы сравнить между собой результаты экзаменов в разных регионах удобно перейти к ненормированным эмпирическим функциям распределения баллов ЕГЭ по выборке учащихся из рассматриваемого региона. При этом количество учащихся, сдававших экзамен в регионе, будем откладывать по горизонтальной оси, а полученные баллы за экзамен, отсортированные по убыванию от 100 до 0, по вертикальной оси. В итоге получим графики кривых спроса (Польдин, Силаев, 2011; Польдин и др., 2014). Данное название отражает отрицательный наклон кривых и возможность сложения по горизонтали, как при агрегировании спроса на однородные товары при совершенной конкуренции.

На рис. 1, 2 приведены кривые спроса для учащихся различных федеральных округов по русскому языку и математике с использованием коррекции баллов к равномерному стандартному распределению по формуле (6). На рис. 3 представлены суммарные кривые спроса для всех федеральных округов, полученные путем "горизонтального суммирования" кривых спроса по русскому языку и математике для отдельных округов с использованием равномерного распределения для коррекции баллов. Можно констатировать, что, чем больше учащихся в регионе, тем, как правило, выше и правее на рисунках расположены кривые спроса. В этом случае для каждого учащегося, результат которого представлен на более низкой кривой спроса, найдется учащийся с тем же порядковым номером на более высокой кривой спроса, имеющий более высокий балл ЕГЭ, то есть имеет место доминирование первой степени (Польдин, Силаев, 2011).

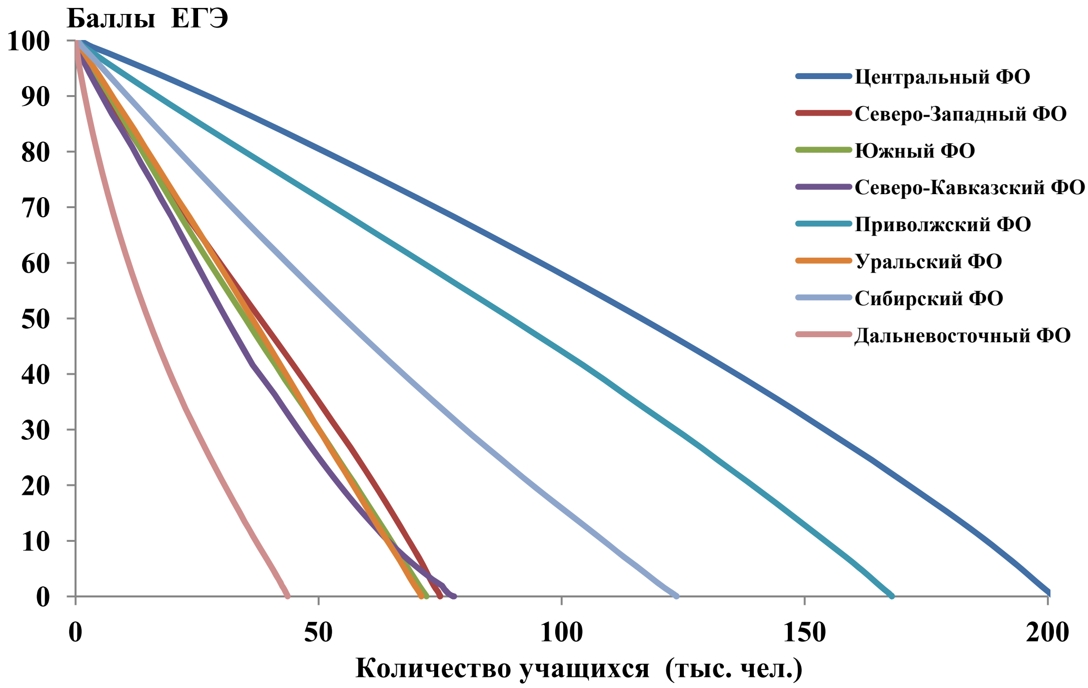

Рис. 1. Кривые спроса для экзаменов по русскому языку с использованием равномерного распределения для коррекции баллов.

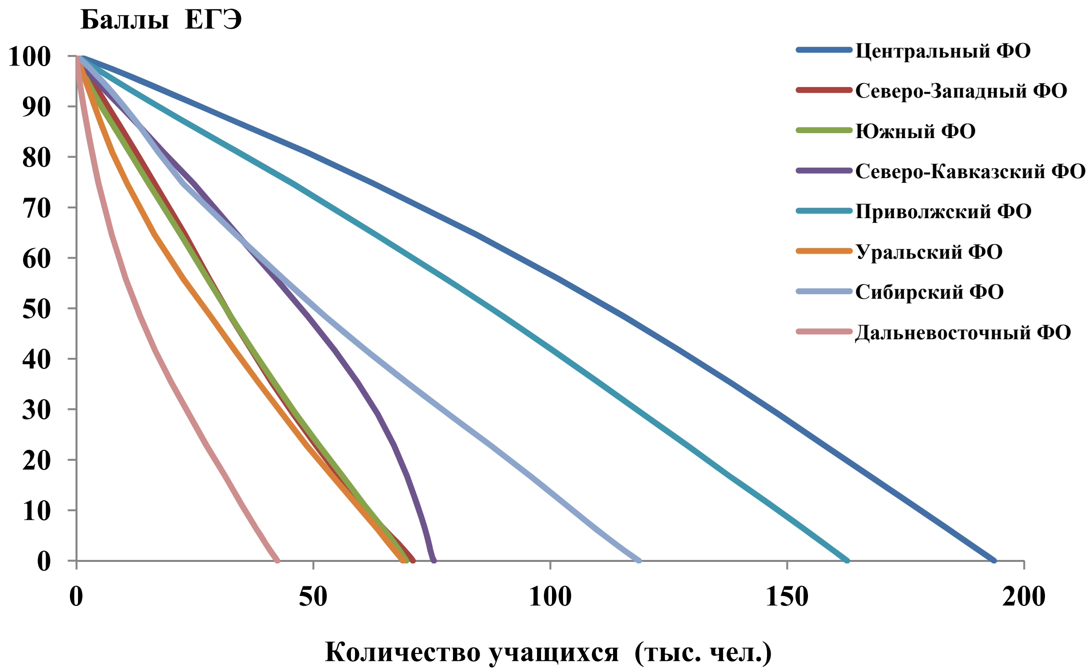

Рис. 2. Кривые спроса для экзаменов по математике с использованием равномерного распределения для коррекции баллов.

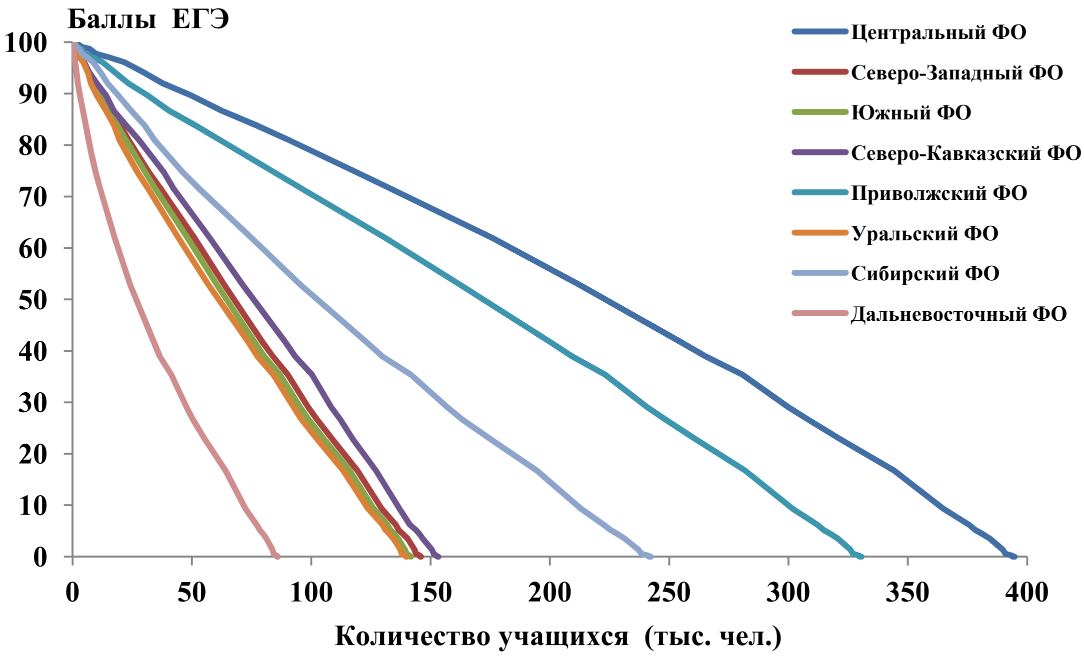

Рис. 3. Суммарные кривые спроса для экзаменов по русскому языку и математике в федеральных округах в 2013 г., найденные с использованием равномерного распределения для коррекции баллов.

Кривые спроса на рис. 3 ранжируются в порядке, который совпадает порядком при сортировке суммарного количества учащихся, сдававших экзамены в каждом федеральном округе. В Центральном федеральном округе было наибольшее количество сдававших экзамены по русскому языку и математике, и соответствующая кривая спроса на рис. 3 расположена выше других. Затем следуют Приволжский федеральный округ, Сибирский федеральный округ, Северо-Кавказский федеральный округ, Северо-Западный федеральный округ, Южный федеральный округ, Уральский федеральный округ и Дальневосточный федеральный округ.

Несмотря на существенное различие экзаменов по разным предметам, с помощью алгоритмов коррекции баллов удается выявить некоторые общие для этих предметов закономерности, которые характеризуют ЕГЭ в России в 2013 г. Графики на рис. 1 – 3 демонстрируют, что количество как хороших, так и плохих оценок по ЕГЭ в том или ином регионе в основном зависит от количества учащихся в этом регионе. Чем больше было учащихся, сдававших экзамен, тем больше было в этом округе высоких оценок, но также тем больше и низких оценок. При этом количество высоких оценок в регионах растет несколько быстрее, а количество низких оценок несколько медленнее, чем число учащихся, сдававших экзамен в каждом регионе.

Список литературы

1. Андрущак Г.В., Польдин О.В., Юдкевич М.М. (2012). Эффекты сообучения в административно формируемых студенческих группах. // Прикладная эконометрика. № 2 (26). С. 3–16.
2. Замков О.О., Пересецкий А.А. (2013). ЕГЭ и академические успехи студентов бакалавриата МИЭФ НИУ ВШЭ. // Прикладная эконометрика. №2 (30). С. 93–114.
3. Пересецкий А.А., Давтян М.А. (2011). Эффективность ЕГЭ и олимпиад как инструмента отбора абитуриентов. // Прикладная эконометрика. №3 (23). С. 41–56.
4. Польдин О.В. (2011). Прогнозирование успеваемости в вузе по результатам ЕГЭ. // Прикладная эконометрика. № 1 (21). С. 56–69.
5. Польдин О.В., Силаев А.М. (2011). Сравнение образовательных программ по результатам ЕГЭ зачисленных студентов. // Вопросы образования. № 3. С. 192–209.
6. Польдин О.В., Силаева В.А., Силаев А.М. (2014). Сравнение приема на образовательные программы в вузе по результатам олимпиад и баллов ЕГЭ. // Прикладная эконометрика. №4 (36). С. 118–132.