ЭКОНОМИКА
И МАТЕМАТИЧЕСКИЕ МЕТОДЫ

1966
том II, вып. 2

ОБРАБОТКА НЕОДНОРОДНОГО РАСПРЕДЕЛЕНИЯ РАБОТАЮЩИХ ПО ЗАРАБОТНОЙ ПЛАТЕ

Б. М. БОГДАНОВ

(Москва)

Применение ЭВМ в экономических исследованиях требует использования совершенной информации. Данная статья является попыткой определить одно из возможных направлений в обработке экономической информации, задаваемой в виде рядов распределений. Суждения построены на анализе фактических рядов распределения рабочих и служащих по размеру заработной платы *, а также данных бюджетной статистики.

Один из способов обработки рядов распределений — их статистическое выравнивание. Теоретическая кривая при правильном выборе закона распределения не должна искажать выравниваемый эмпирический ряд, а теоретическое распределение в основном не нарушает наблюденной структуры. Поэтому ведущее место в выборе выравнивающей функции должно быть отведено экономическому анализу причин формирования данного эмпирического ряда и его внутренних закономерностей.

Экономико-статистический анализ рядов распределения работающих по заработной плате показывает, что наиболее приемлема для их выравнивания функция логарифмически-нормального распределения. Выравнивание рядов распределения по логарифмически-нормальной функции дает определенный эффект, если ряд имеет одну вершину (одномодален) и *качественно* однороден. Последнее, впрочем, относится к любой другой функции распределения.

Распределение работающих по заработной плате в СССР можно рассматривать как неоднородное распределение, в котором смешаны по меньшей мере две однородные совокупности. Понятие «качественная однородность или неоднородность» совокупности в данном случае нами истолковывается не с точки зрения принадлежности работников к разным отраслям промышленности или народного хозяйства и объединения их в одном распределении, а с точки зрения *причин роста* заработной платы. Ведь в советской экономике можно наблюдать две формы роста заработной платы: рост заработной платы, обусловленный ростом производительности труда, и рост заработной платы, обусловленный политикой государства по уменьшению дифференциации в уровне жизни населения.

Рост заработной платы, обусловленный ростом производительности труда, назовем естественным. Рост заработной платы, обусловленный «упорядочением» заработной платы, назовем искусственным. Искусственный рост заработной платы наблюдается у определенной категории населения, которую называют низкооплачиваемой.

---

* Ряд распределения работающих по размеру заработной платы показывает структуру определенной совокупности населения с точки зрения получаемой заработной платы.

Таким образом, всех работающих можно разделить на две совокупности: а) работающие, заработная плата которых имела только естествен-

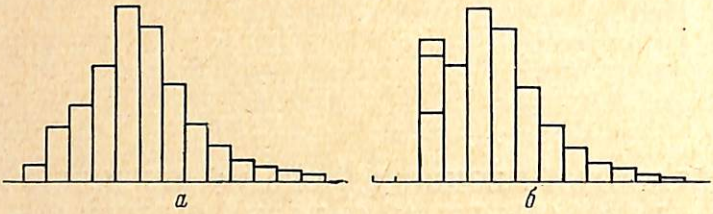

Рис. 1

ный прирост, и б) работающие, заработная плата которых повышалась в основном искусственно.

А так как ЦСУ СССР дает информацию о распределении работающих по заработной плате общей строчкой, то совокупности смешаны подобно тому, как в рядах распределения работающих по заработной плате в угольной промышленности смешаны рабочие, занятые на подземных работах, с рабочими и служащими наземных служб. При графическом изображении эмпирического ряда подобное смешение неоднородных распределений приводит к появлению второго вполне закономерного «горба», ибо категория низкооплачиваемых была как бы перенесена законом об упорядочении заработной платы в следующие группы (интервалы) заработной платы (рис. 1). «Перенесение» повлекло за собой увеличение асимметрии кривой распределения.

Для многих плановых расчетов, связанных с уровнем жизни, необходим анализ двух указанных совокупностей в отдельности, ибо на основе информации о генеральном распределении работающих по заработной плате на перспективу и данных бюджетной статистики строится генеральное распределение семей по доходу на перспективу, которое является в свою очередь основой плановых расчетов по потреблению, спросу, денежному обращению, товарообороту и т. д. Удовлетворение спроса явится в конечном счете критерием планирования производства продуктов питания и товаров широкого потребления. Ошибка (выравнивание указанного распределения по общепринятым методам), привнесенная в первоначальные расчеты, может вызвать за собой «лавину» накопленных ошибок в последней «инстанции», что приведет не только к результатам, порочащим экономико-математические методы планирования, но и к нарушению соответствия между спросом и возможностью его удовлетворения, затовариванию и другим народнохозяйственным затруднениям.

В случае с распределением по заработной плате перед исследователем встает задача выделить из неоднородного (смешанного) распределения два составляющих его неизвестных распределения, т. е. задача определить их параметры, а следовательно, и теоретические частоты (структуру).

Условимся называть этот процесс разбиением неоднородного распределения на два однородных.

Если исследователь имеет дело с достаточно полной информацией о распределении, т. е. если данных много, то можно провести разбиение графически [1, стр. 138]. Но поскольку графический метод, несмотря на всю его простоту, довольно приближенный и неприемлем при использовании электронно-вычислительных машин, представляется важным отыскание аналитических методов разбиения неоднородных распределений.

Так как исследователю приходится считать в большинстве случаев на счетно-клавишной машине, то методы должны быть просты и рассчитаны на реальную его математическую подготовку.

Задача аналитического разбиения неоднородных распределений в теории статистики ставилась и решалась К. Пирсоном [2], К. Шарлье [3], К. Бёро [4], Б. Стрёмгреном [5], В. Урбахом [6—8]. Наблюдается последовательное упрощение методов разбиения неоднородных распределений в указанной последовательности. Так, Пирсон предусматривал определение корней уравнения девятой степени, Шарлье — совместное решение двух кубических уравнений. В этом случае при большом объеме информации задача практически неосуществима. Бёро и Стрёмгрен пользовались специально составленной таблицей, аргумент которой — определенная комбинация семи-инвариантов Тиле. Эта методика также громоздка и трудоемка. В. Урбах значительно упростил и детализировал задачу разбиения, сведя ее к решению системы уравнений начальных моментов

$$nM^j = \sum_{j=1} n_{1,2} M_{1,2}{}^j,$$

где j — порядок момента. Однако использование моментов пятого и шестого порядка, а также их вариации, значительно снижает точность параметров искомых распределений. К тому же расчет не связан единой схемой, что делает его труднодоступным для экономистов.

В общем виде задача разбиения неоднородного распределения на два (однородных) может быть сформулирована следующим образом.

Пусть $\tilde{f}(x)$ есть заданное эмпирическое распределение, которое может быть аппроксимировано некоторым теоретическим распределением $f(x)$ с параметрами $m = E(x^i)$, $\sigma^2 = E[(x^i - m)^2]$ и объемом эмпирической совокупности n.

Требуется найти два таких нормальных теоретических распределения $f_1(x)$ и $f_2(x)$ с параметрами m_1, $\sigma_1{}^2$; m_2, $\sigma_2{}^2$ и объемами n_1 и n_2, чтобы выпуклая линейная комбинация (сумма) их, равная некоторому расчетному распределению $\hat{f}(x)$, аппроксимировала распределение $\tilde{f}(x)$ наилучшим образом, а $n_1 + n_2 = n$.

В настоящей статье предлагаются два аналитических метода разбиения неоднородного распределения. Так как они предназначены для исследователей-нематематиков, то краткое теоретическое изложение вопроса иллюстрируется расчетами на двух примерах. Это позволит экономисту творчески применить метод, сообразуясь с поставленной перед ним задачей и имеющимися у него материалом и техникой.

1. РАЗБИЕНИЕ ДВУХВЕРШИННОГО НЕОДНОРОДНОГО РАСПРЕДЕЛЕНИЯ НА ДВА НОРМАЛЬНЫХ

При анализе имеющихся в распоряжении исследователя сведений о процессе, который приводит к наблюдаемым явлениям, часто можно разбить наблюдения на две или большее число групп, каждая из которых соответствует нормальному закону распределения.

Пусть информация о неоднородном распределении $\tilde{f}(x)$ задана табл. 1, где x^i — середины i-х интервалов, $\tilde{f}(x^i)$ — частоты в процентах. Рис. 2, построенный на данных табл. 1, показывает ярко выраженное двухвершинное распределение (непрерывная линия).

Выдвигаем гипотезу о том, что данное распределение может быть аппроксимировано нормальным распределением с параметрами $m = 13{,}31$ и $\sigma^2 = 27{,}70$ ($\sigma = 5{,}26$) *.

Выдвинутую гипотезу приходится отвергнуть, так как, во-первых, нами не выделены таким образом распределения $f_1(x)$ и $f_2(x)$, и, во-вторых, кривая плотности распределения $f(x)$ не отвечает требованию поставленной задачи (см. рис. 2, точечная линия).

Таблица 1

i	x^i	$\tilde{f}(x^i)$	$f_1(x^i)$	$f_2(x^i)$
1	1,25	0,76	0,76	—
2	3,50	4,00	3,72	0,28
3	5,50	8,51	7,52	0,99
4	7,50	8,75	5,94	2,81
5	9,50	8,09	1,84	6,25
6	12,00	18,14	0,22	17,92
7	15,50	30,55	—	30,55
8	20,00	18,85	—	18,85
9	25,00	2,35	—	2,35
Итого:		100,00	20,00	80,00

Таблица 1а

i	$x^i \geqslant [m_2]^0$	$f(x^i \geqslant [m_2]^0)$	$(x^i - [m_2]^0) \cdot f(x^i \geqslant [m_2]^0)$
7	15,50	15,28	0
8	20,00	18,85	181,69
9	25,00	2,35	212,04
Σ		36,48	593,80

Выдвигаем вторую гипотезу: распределение $\tilde{f}(x)$ лучше аппроксимируется суммой двух нормальных распределений $\hat{f}(x)$.

Следовательно, для того чтобы принять или отвергнуть эту гипотезу, сначала необходимо вычислить распределения $f_{1,2}(x)$.

Поскольку параметры распределения $f(x)$ равно принадлежат и распределению $\tilde{f}(x)$, то первый этап поставленной задачи сводится к определению связи между параметрами распределения $f(x)$ и параметрами распределений $f_1(x)$ и $f_2(x)$, каждое из которых, еще раз подчеркиваем, имеет нормальный закон распределения.

Легко показать, что в этом случае

$$n_1 + n_2 = n, \tag{1}$$

$$n_1 m_1 + n_2 m_2 = nm, \tag{2}$$

$$n_1\sigma_1^2 + n_2\sigma_2^2 + n_1(m_1 - m)^2 + n_2(m_2 - m)^2 = n\sigma^2. \tag{3}$$

Следовательно, для определения параметров $n_{1,2}$, $m_{1,2}$, $\sigma^2_{1,2}$ искомых распределений $f_1(x)$ и $f_2(x)$ необходимо решить систему из трех уравнений (1) — (3) с шестью неизвестными.

Так как данная система недоопределена, то имеем множество решений. Для практических расчетов предлагаем воспользоваться методом последовательного приближения.

Дадим геометрическую интерпретацию этого метода для нашего случая (рис. 3).

Пусть прямая a схематично изображает значения правых частей уравнений системы (1) — (3), а прямая b — их левых частей.

Предположим, что в точке M отложены полученные грубые значения параметров распределений $f_1(x)$ и $f_2(x)$, а следовательно, и левых частей уравнений (1) — (3) в начальном приближении $[x]^0$ **. Расстояние

* Параметры вычислены обычным образом на основе информации из табл. 1: $m = \frac{1}{n}\Sigma x^i$, $\sigma^2 = \frac{1}{n}\Sigma(x^i - m)^2$, $i = 1, 2, \ldots, n$.

** В данном написании здесь и далее цифра 0 за прямыми скобками [ ] указывает на начальное значение, 1 — на значение в первом приближении, 2 — во втором, k — в k-м.

$LM = [\xi]^0$ покажет величину отклонений (невязок) левых частей уравнений (1)—(3) от правых в начальном приближении. Привнесение некоторых поправок в значения параметров распределений $f_1(x)$ и $f_2(x)$ уменьшит величину неявок. На рисунке поправки изображены прямой MN, па-

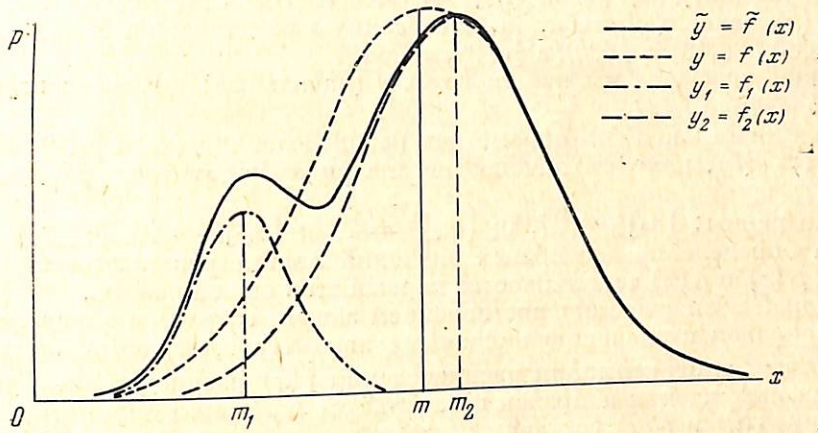

Рис. 2

раллельной оси абсцисс. Очевидно, что абсцисса точки N совпадает со значением искомых параметров в первом приближении $[x]^1$ ($NO = [\xi]^1$). Далее прием повторяется ($PR = [\xi]^2$, $ST = [\xi]^3$) до тех пор, пока величина невязок ξ не будет величиной минимальной, на рисунке равной нулю,

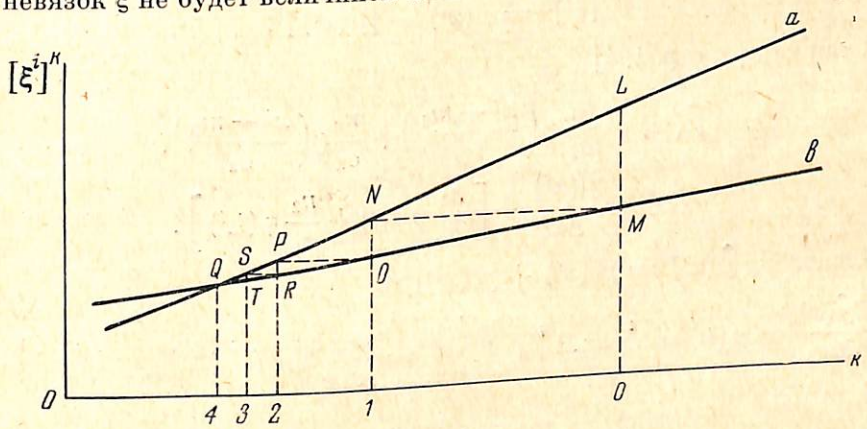

Рис. 3

т. е. пока прямые a и b не пересекутся. Тогда в точке пересечения Q получим истинные значения параметров двух искомых распределений.

Для определения начальных значений параметров сделаем допущение следующего порядка: правая (нисходящая) ветвь распределения $\tilde{f}(x)$ принадлежит только одному из искомых нормальных распределений $f_2(x)$ (рис. 2, непрерывная линия).

Если это предположение принято, то мода распределения $f_2(x)$ равна 15,50 (см. табл. 1). А так как для нормального распределения мода и средняя равны, то $[m_2]^0 = 15{,}50$.

Теперь, зная среднюю и одну из ветвей нормального распределения $f_2(x)$, можно определить объем его совокупности и дисперсию.

Так как из нормальности распределения $f_2(x)$ следует, что левая ветвь его симметрична правой с центром симметрии, проходящим через $[m_2]^0 = 15{,}50$, то численность половины совокупности распределения $f_2(x)$ равна 36,48 (см. табл. 1 : ½·30,55 + 18,85 + 2,35). Объем всей совокупности распределения $f_2(x)$ равен $[n_2]^0 = 72{,}96 \approx 73{,}00$. Дисперсия распределения $f_2(x)$ равна (см. табл. 1а, построенную на данных табл. 1): $[\sigma_2^2]^0 = 593{,}80 / 36{,}48 = 16{,}28$, а $[\sigma_2]^0 = 4{,}04$.

Таким способом мы нашли грубые (начальные) значения параметров распределения $f_2(x)$.

Подставив значения параметров распределений $f(x)$ и $f_2(x)$ в уравнения (1)—(3), находим начальные значения параметров распределения $f_1(x)$.

Они равны: $[n_1]^0 = 27{,}00$; $[m_2]^0 = 7{,}35$; $[\sigma_1^2]^0 = 10{,}09$; $[\sigma_1]^0 = 3{,}17$.

На определении начальных значений параметров искомых распределений $f_1(x)$ и $f_2(x)$ заканчивается первый этап нашей работы.

Второй этап решения поставленной задачи состоит в определении таких поправок к значению найденных параметров, которые приведут к наилучшей аппроксимации распределения $\tilde{f}(x)$ распределением $\hat{f}(x)$.

Для нахождения поправок к начальным значениям параметров распределений $f_1(x)$ и $f_2(x)$ составим систему условных уравнений [9, стр. 119], свободными членами которой являются невязки $[\xi^i]$ $(i = 1, 2, \ldots, n)$, а коэффициентами при неизвестных — первые производные от $\hat{p}(x^i)$ по всем значениям искомых параметров n_1, n_2; m_1, m_2; σ_1, σ_2.

Легко видеть, что $[\xi^i]^0$ равно разности соответствующих плотностей расчетного и эмпирического распределений, т. е.

$$[\xi^i]^0 = \tilde{p}(x^i) - [\hat{p}(x^i)]^0. \tag{4}$$

А так как [1, стр. 134]

$$\hat{p}(x^i) = \frac{n_1}{\sigma_1}\varphi\left(\frac{x^i - m_1}{\sigma_2}\right) + \frac{n_2}{\sigma_2}\varphi\left(\frac{x^i - m_2}{\sigma_2}\right), \tag{5}$$

то

$$[\xi^i]^0 = \frac{\tilde{f}(x^i)}{\Delta x^i} - \left[\frac{n_1}{\sigma_1}\varphi(t_1^i) + \frac{n_2}{\sigma_2}\varphi(t_2^i)\right]. \tag{6}$$

Поскольку $\hat{p}(x^i) = p_1(x^i) + p_2(x^i)$, то

$$\left.\begin{aligned} &\frac{\partial \hat{p}(x)}{\partial n_{1,2}} = \varphi(t_{1,2}^i)\,dt_{1,2}^i \equiv A_{1,2}^i, \quad \frac{\partial \hat{p}(x)}{\partial m_{1,2}} = \frac{n_{1,2}}{\sigma_{1,2}} t_{1,2}^i \varphi(t_{1,2}^i)\,dt_{1,2}^i \equiv B_{1,2}^i \\ &\frac{\partial \hat{p}(x)}{\partial \sigma_{1,2}} = \frac{n_{1,2}}{\sigma_{1,2}}[(t_{1,2}^i)^2 - 1]\varphi(t_{1,2}^i)\,dt_{1,2}^i \equiv C_{1,2}^i \end{aligned}\right\} \tag{7}$$

Обозначив для удобства $\operatorname{const}\frac{n_{1,2}}{\sigma_{1,2}} = k_{1,2}$, $\frac{n_{1,2}}{\sigma_{1,2}} t_{1,2}^i = k_{1,2}\, t_{1,2}^i = l_{1,2}^i$, $[(t_{1,2}^i)^2 - 1] = g_{1,2}^i$, $\frac{n_{1,2}}{\sigma_{1,2}}[(t_{1,2}^i)^2 - 1] = k_{1,2}\, g_{1,2}^i = j_{1,2}^i$ получим

$$\begin{aligned} A_{1,2}^i &\approx \varphi(t_{1,2}^i)\Delta t_{1,2}^i, \\ B_{1,2}^i &\approx l_{1,2}^i \varphi(t_{1,2}^i)\Delta t_{1,2}^i. \\ C_{1,2}^i &\approx j_{1,2}^i \varphi(t_{1,2}^i)\Delta t_{1,2}^i. \end{aligned} \tag{8}$$

Если $a_{1,2} \equiv \Delta n_{1,2}$, $b_{1,2} \equiv \Delta m_{1,2}$, $c_{1,2} \equiv \Delta\sigma_{1,2}$, то система условных уравнений (линейная) имеет следующий вид:

$$[A_1^i]^0 a_1 + [A_2^i]^0 a_2 + [B_1^i]^0 b_1 + [B_2^i]^0 b_2 + [C_1^i]^0 c_1 + [C_2^i]^0 c_2 = [\xi^i]^0,$$

$$i = 1, 2, \ldots, n. \qquad (9)$$

Задаваясь начальными значениями вычисленных ранее параметров распределений $f_1(x)$ и $f_2(x)$, находим коэффициенты системы уравнений (9), частоты этих распределений $f_{1,2}(x^i)$ и их плотности $p_{1,2}(x^i)$. Сначала вычисляются названные искомые по одному из распределений, например первому, затем — по другому. Вычисления производятся по следующей табличной схеме:

Таблица 2

i	x^i	$x^i - m_{1,2}$	Δx^i	$\Delta t^i_{1,2} = \frac{\Delta x^i}{\sigma_{1,2}}$	$t^i_{1,2} = \frac{x^i - m_{1,2}}{\sigma_{1,2}}$	$\varphi(t^i_{1,2})$	$g^i_{1,2} = (t^i_{1,2})^2 - 1$
0	1	2	3	4	5	6	7

$l^i_{1,2} = k_{1,2} t^i_{1,2}$	$j^i_{1,2} = k_{1,2} g^i_{1,2}$	$l^i_{1,2} \cdot \varphi(t^i_{1,2})$	$j^i_{1,2} \cdot \varphi(t^i_{1,2})$	$n_{1,2}\varphi(t^i_{1,2})$	$A^i_{1,2}$
8	9	10 (6·8)	11 (6·9)	12	13 (4·6)

$B^i_{1,2}$	$C^i_{1,2}$	$f_{1,2}(x^i)$	$p_{1,2}(x^i)$
14 (4·10)	15 (4·11)	16 (4·12)	17 (16:3)

Предложенная схема не только проста, но и довольно удобна при вычислениях на счетно-клавишной машине, ибо предусматривает наиболее экономичный вариант для исследователя, решающего такую задачу.

Для определения невязок $[\xi^i]^0$ необходимо поинтервально просуммировать значения $[p_1(x^i)]^0$ и $[p_2(x^i)]^0$, вычислив таким образом $[\hat{p}(x^i)]^0$, и вычесть из каждого значения $\tilde{p}(x^i)$ соответствующее $[\hat{p}(x^i)]^0$.

Решая полученную систему n уравнений методом наименьших квадратов, получаем значение поправок. В нашем примере они равны:

$$[a_1]^0 = -8{,}00, \qquad [a_2]^0 = 8{,}00,$$
$$[b_1]^0 = -1{,}18, \qquad [b_2]^0 = -0{,}53,$$
$$[c_1]^0 = -0{,}50, \qquad [c_2]^0 = 0{,}27.$$

Следовательно,

$$[n_1]^1 = 27 - 8 = 19, \qquad [n_2]^1 = 73 + 8 = 81,$$
$$[m_1]^1 = 7{,}35 - 1{,}18 = 6{,}17, \qquad [m_2]^1 = 15{,}50 - 0{,}53 = 14{,}97,$$
$$[\sigma_1]^1 = 3{,}17 - 0{,}50 = 2{,}67, \qquad [\sigma_2]^1 = 4{,}04 + 0{,}27 = 4{,}31.$$

Частоты и плотности распределений в первом приближении находятся по той же схеме (см. табл. 2, колонки 0—6, 12, 16, 17).

Для получения более точных значений параметров распределений $f_1(x)$ и $f_2(x)$ по сравнению с их значением в первом приближении весь процесс необходимо повторить, но уже на основе значений параметров в первом приближении.

Процесс повторений производится до тех пор, пока величина

$$S^k = \sum_{i=1}^{n} ([\xi^i]^k)^2 \tag{10}$$

не станет величиной минимальной.

Уже в первом приближении она равна 0,0036, тогда как в начальном она составляет 2,4898. Получив столь малую величину, процесс вычислений параметров искомых распределений можно прекратить.

Уже во втором приближении были получены (с некоторым округлением) следующие значения параметров распределений $f_1(x)$ и $f_2(x)$:

$$n_1 = 20{,}00, \quad m_1 = 6{,}00, \quad \sigma_1 = 2{,}00,$$

$$n_2 = 80{,}00, \quad m_2 = 15{,}00, \quad \sigma_2 = 4{,}00.$$

Вычисленные частоты записаны в табл. 1 — $f_1(x^i)$ и $f_2(x^i)$, а соответствующие кривые нанесены на рис. 2 (прерывистая и пунктирная линии). В данном примере кривые $\tilde{p}(x^i)$ и $\hat{p}(x^i)$ совпадают, так как для контроля за вычислениями и их наглядностью распределение $\tilde{f}(x)$ было получено нами искусственно как сумма распределений 20(6; 2) и 80(15; 4).

2. РАЗБИЕНИЕ НЕОДНОРОДНОГО АСИММЕТРИЧНОГО РАСПРЕДЕЛЕНИЯ НА ДВА АСИММЕТРИЧНЫХ

При изложении второго метода разбиения неоднородных распределений оставим те же обозначения, что и в первом методе. Кроме того, пусть распределение $\tilde{f}(x)$ имеет центральные моменты M^j, которые равно принадлежат и распределению $f(x)$, а некоторое расчетное распределение $\hat{f}(x)$ — центральные моменты Q^j $(j = 2, 3, 4)$.

Так как по условию распределения $f_1(x)$ и $f_2(x)$ нормальны, то

$$\mu^{2j} = 1 \cdot 3 \cdot 5, \ldots \cdot (2j - 1)\,(\mu^2)^j \tag{11}$$

и

$$\mu^{2j-1} = 0 \tag{12}$$

или

$$\mu^2 = \sigma^2; \quad \mu^1 = \mu^3 = 0, \quad \mu^4 = 3\sigma^4. \tag{13}$$

Для дискретных распределений центральные моменты равны

$$\mu^j = \frac{1}{n} \sum_{i=1}^{n} (x^i - m)^j. \tag{14}$$

Отсюда

$$n\mu^j = \sum_{i=1} (x^i - m)^j. \tag{15}$$

Предположим, исходя из анализа причин формирования эмпирического ряда распределения или по каким-либо другим соображениям, мы пришли к выводу о его неоднородности. Как и в первом методе, выдвигаем две гипотезы об аппроксимации распределения $\tilde{f}(x)$. По тем же соображениям отказываемся от первой из них.

Задача определения связи параметров искомых нормальных распределений $f_1(x)$ и $f_2(x)$ с параметрами распределения $f(x)$ в данном случае

ставится шире, ибо, кроме этого, необходимо найти еще связь искомых параметров с центральными моментами распределения $f(x)$.

Так как из публикуемой информации о распределении работающих по заработной плате не представляется возможным сделать вывод о величине дисперсий искомых распределений, а бюджетные данные в конечном счете мало репрезентативны, то приходится принять одно из трех возможных допущений о соотношении дисперсий распределений $f_1(x)$ и $f_2(x)$:

или дисперсии равны друг другу:

$$\sigma_1^2 = \sigma_2^2 \equiv \delta^2, \tag{16}$$

или дисперсия одного распределения составляет заданную k-ю часть второго:

$$k\sigma_1^2 = \sigma_2^2, \tag{17}$$

или отношение дисперсий равно отношению квадратов средних, или, что то же самое, коэффициент вариации первого распределения равен коэффициенту вариации второго; тогда

$$m_2^2\sigma_1^2 = m_1^2\sigma_2^2. \tag{18}$$

Уравнения (16) и (17) по внешнему виду кажутся частными случаями уравнения (18), но не являются таковыми по их внутреннему (экономическому) содержанию.

Например, при равных m_1 и m_2 из уравнения (18) мы должны были бы получить и равные дисперсии (16). На поверку такой формальный подход к соотношению дисперсий не оправдывается. Так, средняя заработная плата в химической промышленности и машиностроении примерно одинакова, но дисперсии распределений работающих по заработной плате в этих отраслях промышленности значительно отличаются друг от друга. Это происходит потому, что разница между минимальной и максимальной градациями заработной платы в химии относительно мала (распределение сжато), а в машиностроении указанная разница довольно велика (распределение растянуто).

Такое положение объясняется прежде всего тем, что в химии 70—85% работников находятся на повременной оплате труда. При стабильности и автоматизации технологического процесса величина заработка не зависит от ловкости, квалификации и прочих качеств работника. В машиностроении при довольно большом количестве вспомогательных рабочих доля работников с повременной оплатой труда гораздо меньше. Заработки же сдельщиков имеют широкий диапазон в зависимости от квалификации, рода работ, качества сырья, работоспособности, производительности труда работника и многих других причин, носящих экономический и психофизиологический характер. Поэтому дисперсия в распределении работающих по заработной плате в химической промышленности меньше примерно в 1,7 раза.

Кстати, если мы сложим эти два распределения, то получим одновершинное неоднородное распределение, по внешнему виду которого невозможно сделать какой-либо вывод о его неоднородности. Следовательно, процессу разбиения неоднородных распределений по заработной плате должен предшествовать глубокий экономико-статистический анализ. Он же необходим и при выборе одной из возможных формул связи дисперсий распределений $f_1(x)$ и $f_2(x)$.

Следующим этапом нашей работы будет отыскание в общем виде связи параметров распределений $f_1(x)$ и $f_2(x)$ с вторым, третьим и четвертым центральными моментами распределения $f(x)$.

Она находится следующим образом. Произведем небольшие преобразования правой части уравнения (15):

$$\sum_{i=1}^{n}(x^i-m)^2=\sum_{i=1}^{n_1}(x^i-m)^2+\sum_{i=1}^{n_2}(x^i-m)^2=\sum_{i=1}^{n_1}(x^i-m_1+m_1-m)^2+$$

$$+\sum_{i=1}^{n_2}(x^i-m_2+m_2-m)^2=\sum_{i=1}^{n_1}\{(x^i-m_1)+(m_1-m)\}^2+$$

$$+\sum_{i=1}^{n_2}\{(x^i-m_2)+(m_2-m)\}^2.$$

Раскрыв скобки и применяя (13) и (15), имеем

$$\sum_{i=1}^{n}(x^i-m)^2=n_1\sigma_1^2+n_2\sigma_2^2+n_1(m_1-m)^2+n_2(m_2-m)^2.$$

Если $m_1-m=d_1$, а $m_2-m=d_2$, то

$$nM^2=n_1\sigma_1^2+n_2\sigma_2^2+n_1d_1^2+n_2d_2^2. \tag{19}$$

Аналогично (19) находим

$$nM^3=3(n_1\sigma_1^2d_1+n_2\sigma_2^2d_2)+n_1d_1^3+n_2d_2^3, \tag{20}$$

$$nM^4=3(n_1\sigma_1^4+n_2\sigma_2^4)+6(n_1\sigma_1^2d_1^2+n_2\sigma_2^2d_2^2)+n_1d_1^4+n_2d_2^4. \tag{21}$$

Принимая во внимание (1), (2) и (16)—(18), имеем три системы уравнений, каждую из которых можно рассматривать в качестве возможной статической модели распределения работающих по заработной плате при выбранной гипотезе соотношения σ_1^2 и σ_2^2:

$$\left.\begin{array}{l} n_1+n_2=n,\ n_1m_1+n_2m_2=nm \\ \sigma_1^2=\sigma_2^2=\delta^2 \\ n\delta^2+n_1d_1^2+n_2d_2^2=nM^2 \\ 3\delta^2(n_1d_1+n_2d_2)+n_1d_1^3+n_2d_2^3=nM^3 \\ 3n\delta^4+6\delta^2(n_1d_1^2+n_2d_2^2)+n_1d_1^4+n_2d_2^4=nM^4 \end{array}\right\} \tag{22}$$

$$\left.\begin{array}{l} n_1+n_2=n,\ n_1m_1+n_2m_2=nm,\ k\sigma_1^2=\sigma_2^2 \\ \sigma_2^2\left(\dfrac{n_1}{k}+n_2\right)+n_1d_1^2+n_2d_2^2=nM^2 \\ 3\sigma_2^2\left(\dfrac{n_1}{k}d_1+n_2d_2\right)+n_1d_1^3+n_2d_2^3=nM^3 \\ 3\sigma_2^4\left(\dfrac{n_1}{k}+n_2\right)+6\sigma_2^2\left(\dfrac{n_1}{k}d_1^2+n_2d_2^2\right)+n_1d_1^4+n_2d_2^4=nM^4 \end{array}\right\} \tag{23}$$

$$\left.\begin{array}{l} n_1+n_2=n,\ n_1m_1+n_2m_2=nm,\ \ m_2^2\sigma_1^2=m_1^2\sigma_2^2 \\ n_1\sigma_1^2+n_2\sigma_2^2+n_1d_1^2+n_2d_2^2=nM^2 \\ 3(n_1\sigma_1^2d_1+n_2\sigma_2^2d_2)+n_1d_1^3+n_2d_2^3=nM^3 \\ 3(n_1\sigma_1^4+n_2\sigma_2^4)+6(n_1\sigma_1^2d_1^2+n_2\sigma_2^2d_2^2)+n_1d_1^4+n_2d_2^4=nM^4 \end{array}\right\} \tag{24}$$

Вычисление искомых параметров n_1, n_2; m_1, m_2; σ_1^2, σ_2^2 производится в следующей последовательности: 1) определяются центральные моменты эмпирического распределения; 2) выбирается после предварительного анализа или по каким-либо другим соображениям о характере связи σ_1^2 и σ_2^2 система уравнений из (22)—(24); 3) как и в первом методе, опреде-

ляются начальные значения параметров сначала одного из искомых распределений, затем другого; 4) находятся поправки к найденным значениям параметров.

Разберем второй метод разбиения неоднородных распределений на примере разбиения смешанного распределения женщин и мужчин по величине месячной заработной платы на севере США [10, стр. 67].

Информация о распределении дана в табл. 3. Для решения выбираем систему (23), так как предварительный анализ данных американской ста-

Таблица 3

i	Интервалы зарплаты (в долларах)	$\tilde{f}(x^i)$	$f_1(x^i)$	$f_2(x^i)$	$\hat{f}(x^i)$
1	60—65	0,4	0,5	0,1	0,6
2	65—70	2,4	2,4	0,4	2,8
3	70—75	7,9	5,3	1,5	6,8
4	75—80	12,6	7,7	3,7	11,4
5	80—85	14,2	7,2	7,0	14,2
6	85—90	9,1	4,9	9,9	14,8
7	90—95	17,3	2,5	11,5	14,0
8	95—100	13,0	1,0	10,3	11,8
9	100—105	10,2	0,3	8,8	9,1
10	105—110	4,3	0,1	6,2	6,3
11	110—115	3,1	0,0	3,9	3,9
12	115—120	3,1		2,2	2,2
13	120—125	1,2		1,2	1,2
14	125—130	0,8		0,6	0,6
15	130—135	0,4		0,3	0,3
Итого:		100,0	31,9	68,1	100,0

тистики показал, что дисперсия в распределении женщин примерно в 1,5 раза меньше, чем в аналогичном распределении мужчин. Коэффициенты вариации этих распределений не равны.

Вычисляем параметры и моменты данного эмпирического распределения как параметры и моменты логарифмически-нормального распределения*. Они равны: $m = 4{,}4995$, $\sigma^2 = M^2 = 0{,}020668$, $M^3 = 0{,}000274$, $M^4 = 0{,}001093$.

Начальные значения параметров искомых распределений определяются по тем же принципам, что и в первом методе, причем при определении дисперсии распределения $f_1(x)$ $k = 1{,}5$.

Начальные значения параметров для данного примера равны:

$$[n_2]^0 = 89{,}4, \qquad [m_2]^0 = 4{,}5263, \qquad [\sigma_2^2]^0 = 0{,}0175,$$
$$[n_1]^0 = 10{,}6, \qquad [m_1]^0 = 4{,}2694, \qquad [\sigma_1^2]^0 = 0{,}0117.$$

На втором этапе наших вычислений сделаем следующее: подставим начальные значения параметров распределений $f_1(x)$ и $f_2(x)$ в последние три уравнения системы (23) и определим отклонение левой части каждого из этих уравнений от их правых частей, т. е. определим отклонение nQ^j от nM^j в начальном приближении. Разность $nM^j - nQ^j$ обозначим че-

* Фактически данное распределение, как и все распределения работающих по заработной плате, положительно асимметрично. Но замена масштаба оси абсцисс на логарифмический преобразует такие распределения в близкие к нормальному.

Таблица 4

j	$n_2d_2^j$	$n_1d_1^j$	$n_2\sigma_2^2d_2^{j-2}$	$n_1\sigma_1^2d_1^{j-2}$	$n_2\sigma_2^4$	$n_1\sigma_1^4$	$[nQ^j]^k$	$[nM^j]^k$	$[n\xi^j]^k$
	$[n_2]^0 = 89{,}4$	$[m_2]^0 = 4{,}5268$	$[d_2]^0 = 0{,}0273$	$[\sigma_2^2]^0 = 0{,}0175$	$[n_1]^0 = 10{,}6$	$[m_1]^0 = 4{,}2694$	$[d_1]^0 = -0{,}2301$	$[\sigma_1^2]^0 = 0{,}0117$	начальное приближение
1	2,4406	—2,4391							
2	0,0666	0,5612	1,5645	0,1240			2,3163	2,0668	—0,2495
3	0,0018	—0,1291	0,0427	—0,0285			—0,0847	0,0274	0,1121
4	0,0000	0,0297	0,0012	0,0666	0,0274	0,0014	0,1621	0,1093	—0,0536

рез $n\xi^j$; $[n\xi^j]^k$ будет показывать нам величину невязок в k-м приближении ($k = 1, 2, \ldots, n$).

Для более быстрого и удобного определения невязок предлагаем таблицу-схему (табл. 4), в которой приведен расчет $[n\xi^j]^0$.

На третьем этапе решения необходимо найти такие поправки a, b, c к значению найденных параметров распределений $f_{1,2}(x)$, чтобы величина

$$S^k = \sum_{j=2}^{4} \{[n\xi^j]^k\}^2 \qquad (25)$$

была величиной минимальной.

Для определения поправок достаточно решить следующую систему линейных уравнений:

$$[A^j]^0 a + [B^j]^0 b + [C^j]^0 c = [n\xi^j]^k, \; j = 2, 3, 4. \qquad (26)$$

Вычисление коэффициентов $[A^j]^0$, $[B^j]^0$, $[C^j]^0$ производится в следующей последовательности.

Определим, как изменится величина nQ^j, если произвольно (но в рамках разумного) изменить поочередно каждый из трех найденных параметров распределения $f_2(x)$ при условии, что два оставшихся не изменяются. Для этого:

а) изменим величину параметра $[n_2]^0$ на Δn_2 и вычислим значения nQ^j, которые обозначим через $nP^j{}_{\Delta n_2}$ (величина параметров $[m_2]^0$ и $[\sigma_2^2]^0$ не изменяется). Затем находим разность $nQ^j - nP^j{}_{\Delta n_2}$, которую обозначим через $n\xi^j{}_{\Delta n_2}$;

б) изменим параметр $[m_2]^0$ на Δm_2, вычислим значения $nP^j{}_{\Delta m_2}$ и $n\xi^j{}_{\Delta m_2}$ (параметры $[n_2]^0$ и $[\sigma_2^2]^0$ не изменяются);

в) изменим параметр $[\sigma_2^2]^0$ на $\Delta\sigma_2^2$, вычислим $nP^j{}_{\Delta\sigma_2^2}$ и $n\xi^j{}_{\Delta\sigma_2^2}$ (параметры $[n_2]^0$ и $[m_2]^0$ не изменяются).

При вычислениях а) — в), конечно, нужно помнить, что с изменением параметров распределения $f_2(x)$ изменяются и параметры распределения $f_1(x)$ (см. в (23) первые три уравнения).

Таблица 5

j	$n\xi^j{}_{\Delta n_2}$	$n\xi^j{}_{\Delta m_2}$	$n\xi^j{}_{\Delta\sigma_2^2}$	$[A^j]^0$	$[B^j]^0$	$[C^j]^0$
	$\Delta n_2 = 0{,}6000$	$\Delta m_2 = 0{,}0032$	$\Delta\sigma_2^2 = 0{,}0125$	$\Delta n_2 = 1{,}0000$	$\Delta m_2 = 0{,}0001$	$\Delta\sigma_2^2 = 0{,}0001$
2	0,0455	0,1561	1,2055	0,0757	0,0049	0,0096
3	—0,0186	—0,0455	0,0306	—0,0320	—0,0014	0,0002
4	0,0096	0,0275	0,2001	0,0160	0,0009	0,0017

В нашем примере изменим $[n_2]^0$ на 0,6, $[m_2]^0$ — на 0,0032, $[\sigma_2^2]^0$ — на 0,0125. Вычисленные значения $n\xi^j_{\Delta n_2}$, $n\xi^j_{\Delta m_2}$, $n\xi^j_{\Delta\sigma_2^2}$ приведены в табл. 5.

Если увеличение $[n_2]^0$ на 0,6 изменило величину nQ^2 на 0,0455, nQ^3 — на $-0{,}0186$, nQ^4 — на 0,0096; увеличение $[m_2]^0$ на 0,0032 изменило nQ^2 на 0,1561 и т. д., то возникает следующий вопрос: каково будет изменение nQ^j, если $[n_2]^0$ изменится на 1,0, $[m_2]^0$ — на 0,0001, $[\sigma_2^2]^0$ — на 0,0001?

Очевидно, что для этого нужно $n\xi^j_{\Delta n_2}$ разделить на Δn_2, $n\xi^j_{\Delta m_2}$ — на $\Delta m_2 \cdot 10^4$, $n\xi^j_{\Delta\sigma_2^2}$ — на $\Delta\sigma_2^2 \cdot 10^4$. Эти значения и будут коэффициентами системы (26), причем *

$$[A^j]^0 \equiv n\xi^j_{\Delta n_2} : \Delta n_2; \quad [B^j]^0 \equiv n\xi^j_{\Delta m_2} : (\Delta m_2 \cdot 10^4); \quad [C^j]^0 \equiv n\xi^{j_2}_{\Delta\sigma_2^2} : (\Delta\sigma_2^2 \cdot 10^4).$$

В нашем примере для получения коэффициентов $[A^j]^0$, $[B^j]^0$, $[C^j]^0$ нужно $n\xi^j_{\Delta n_2}$, $n\xi^j_{\Delta m_2}$, $n\xi^j_{\Delta\sigma_2^2}$ разделить соответственно на 0,6, на 32, на 125, ибо за единицу изменения n_2 нами принята величина 1,0, за единицу изменения m_2 — 0,0001, за единицу изменения σ_2^2 — 0,0001. Полученные значения коэффициентов выписаны в табл. 5.

Решая систему нормальных уравнений

$$\begin{aligned} 0{,}0757a + 0{,}0049b + 0{,}0096c &= -0{,}2495, \\ -0{,}0320a - 0{,}0014b + 0{,}0002c &= \quad 0{,}1121, \\ 0{,}0160a + 0{,}0009b + 0{,}0017c &= -0{,}0536, \end{aligned}$$

получаем поправки a, b, c к начальным значениям параметров распределения $f_2(x)$: $[a]^0 = -3{,}5794$, $[b]^0 = 1{,}9201$, $[c]^0 = 1{,}2562$.

Для того чтобы найти значения параметров искомых распределений в первом приближении, сначала определяем их для распределения $f_2(x)$: $[n_2]^1 = 89{,}4 - 3{,}6 = 85{,}8$, $[m_2]^1 = 4{,}5268 + 0{,}0002 = 4{,}5270$, $[\sigma_2^2]^1 = 0{,}0175 + 0{,}0001 = 0{,}0176$. Подставляем найденные значения в первые три уравнения системы (23) и определяем параметры распределения $f_1(x)$: $[n_1]^1 = 14{,}2$, $[m_1]^1 = 4{,}3335$, $[\sigma_1^2]^1 = 0{,}0117$.

Подставляя значения параметров в первом приближении распределений $f_1(x)$ и $f_2(x)$ в оставшиеся три уравнения системы (23), определяем невязки $[n\xi^j]^1$, которые принимаются за значения правых частей системы уравнений (26). Коэффициенты системы (26) все время остаются неизменными. Решение новой системы даст нам значения поправок, привнося которые в параметры в первом приближении, получим их значение во втором приближении.

Подобный процесс повторений требуется проводить до тех пор, пока S^{k+1} не будет больше S^k (см. (25)). В нашем примере это случилось в восьмом приближении. Так, $S^7 = 1{,}9273 \cdot 10^{-4}$; $S^8 = 1{,}9673 \cdot 10^{-4}$.

Значения параметров в седьмом приближении равны:

$$\begin{aligned} n_1 &= 31{,}9, \qquad m_1 = 4{,}3830, \qquad \sigma_1^2 = 0{,}0106, \\ n_2 &= 68{,}1, \qquad m_2 = 4{,}5541, \qquad \sigma_2^2 = 0{,}0159, \end{aligned}$$

а отклонения в значении моментов M^j и Q^j ничтожны. Так: $\xi^2 = 0{,}000006$; $\xi^3 = 0{,}000114$; $\xi^4 = -0{,}000079$.

Частоты искомых распределений $f_1(x)$ и $f_2(x)$ находятся обычно — как частоты лог-нормального распределения. Их значения, а также частоты распределений $\tilde{f}(x)$ и $\hat{f}(x)$ приведены в табл. 3. Кривая плотности каждо-

---

* Произвольный выбор значений Δn_2, Δm_2, $\Delta\sigma_2^2$ не оказывает влияния на величину полученных коэффициентов.

го из этих распределений изображена на рис. 4, причем нами произведен возврат от логарифмического масштаба по оси абсцисс к обычному.

В табл. 3 и на рис. 4 (точечная линия) приведено распределение $\hat{f}(x)$. В данном примере кривая этого распределения не «улавливает» впа-

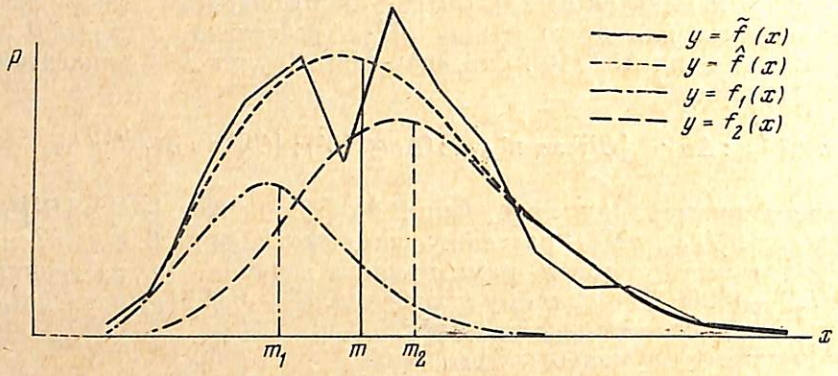

Рис. 4

дины эмпирического распределения. Это произошло потому, что разность между вершинами в масштабе σ равна

$$D = \frac{m_2 - m_1}{\sigma} = \frac{4{,}5541 - 4{,}3830}{0{,}1439} = 1{,}2. \qquad (27)$$

Если бы эта разность превышала 2, то кривая распределения $\hat{f}(x)$ была бы двухвершинной [1, стр. 134].

3. НЕКОТОРЫЕ ОБЩИЕ ЗАМЕЧАНИЯ

Опыт разбиения неоднородных распределений показывает, что формальный подход к неоднородному распределению часто приводит к абсурдным выводам. Например, неоднородность может быть вызвана двумерностью или многомерностью распределения, а мы его обрабатываем как одномерное. Поэтому, прежде чем приступить к разбиению, необходимо произвести глубокий экономический анализ причин неоднородности. К неудовлетворительным результатам может привести также неудачный выбор соотношения дисперсий искомых распределений. Например, модель (24) для разобранного примера даст результаты хуже, чем модель (23). Так, $S^6 = 4{,}7421 \cdot 10^{-4}$, а

$$n_1 = 34{,}8, \quad m_1 = 4{,}4209, \quad \sigma_1^2 = 0{,}0163,$$
$$n_2 = 65{,}2, \quad m_2 = 4{,}5467, \quad \sigma_2^2 = 0{,}0173.$$

Поэтому, если исследователь старается получить наиболее точную модель явления, следует проверить хотя бы две гипотезы из (22) — (24) и выбрать ту, в которой S^h меньше.

Отличие предложенных методов от уже существующих [1—8] состоит в их простоте, точности и универсальности.

Многочисленные проведенные нами расчеты показали, что первый метод разбиения дает лучшие результаты, если неоднородность распределения привела к явной его двухвершинности, второй метод — если двухвершинность неявная, расстояние между вершинами мало или неоднородное распределение одновершинно. Тогда $m_1 = m_2 = m$. Применяя свойства логарифмически-нормального распределения, получаем из асиммет-

ричного неоднородного распределения два также асимметричных распределения.

Разбиение неоднородных распределений, нам кажется, позволит по-новому подойти к измерению дифференциации в оплате труда. Мы предлагаем измерять ее как следующие отношения:

$$U = \frac{m}{m_1}, \quad V = \frac{m_2}{m_1}, \quad W = \frac{m_2}{m}. \tag{28}$$

Являясь модификацией известных статистических отношений Обухова, они лучше отражают экономический смысл дифференциации, хотя численное значение квартильного отношения в рядах распределения работающих по заработной плате в СССР отличается от численного значения V (28) на довольно малую величину (в среднем 7%).

Автор пользуется возможностью выразить благодарность В. Н. Перегудову за советы во время разработки обоих методов.

ЛИТЕРАТУРА

1. А. Хальд. Математическая статистика с техническими приложениями. М., Изд-во иностр. лит., 1956.
2. K. Pearson. Contributions to the Mathematical Theory of Evolution. Philos. Trans. Roy. Soc., 1894, № 185-A.
3. C. V. L. Charlier. Researches into the Theory of Probability. Lund, 1906.
4. C. Burrau. The Half-Invariants of Two Typical Lows of Errors, with an Application to the Problem of Dissecting a Frequency Curve into Components. Skandinavisk Aktuarietidskrift, 1934, № 17.
5. B. Strömgren. Tabls and Diagrams for Dissecting a Frequency into Components by the Half-Invariant Method (там же).
6, 7. В. Урбах. К вопросу о разложении отклоняющихся от нормального статистических распределений на два нормальных распределения. Биофизика, 1965, т. 6, вып. 1, 3.
8. В. Урбах. Биометрические методы. М., «Наука», 1964.
9. Б. Демидович, И. Марон, Э. Шувалова. Численные методы анализа. М., Физматгиз, 1963.
10. E. Dubois. Essential Methods in Business Statistics. McGraw-Hill Company, Inc., 1963.

Поступила в редакцию
27 IV 1965