Вопросы статистики. 2005. № 10

Статистика интервальных данных в обследовании заработной платы

(Оценка характеристик функции распределения в интервале)

С.В. Степанов, кандидат социологических наук,
консалтинговая компания ПЛАНОВА-Консалтинг
E-mail: ssv@mipkgk.msk.ru

В статье рассматривается случай обследования, проводимого измерением значений признаков в интервалах. В складывающихся в этом случае условиях отсутствия точечных данных внутри интервала предлагается модель оценивания характеристик объективно неизвестной закономерности распределения, основанная на конструктивном подходе метаматематической теории. Цель такой оценки – получить вычислимое по Тьюрингу правило оперирования субинтервальными диапазонами. В качестве эмпирического материала рассматривается методика обследования по Форме № 1 «Распределение численности работников по размерам заработной платы» (данные 2004 года).

* * *

При проведении статистических обследований часто возникает ситуация, когда инструментарий обследования (измерения) данных по объективным обстоятельствам способен регистрировать данные исследуемого явления только в виде интервалов значений признака, а не его точечных характеристик. Так происходит, к примеру, при измерениях некоторых показателей потока частиц из реактора. Каждый регистратор частиц в силу своих конструктивных особенностей способен регистрировать суммарный импульс, энергию и количество частиц только строго определённого диапазона энергий или скоростей. Необходимое количество регистраторов охватывают весь возможный диапазон и таким образом проводится измерение. Такое свойство некоторых компонент инструментария может объясняться не только объективными обстоятельствами, но и соображениями экономического или организационного характера, в целях сокращения ресурсных затрат при проведении обследования.

Явление регистрируется и измеряются его характеристики в заданных заранее интервалах. Таким образом, в результате проведённого обследования есть данные по единицам наблюдения, но *отсутствуют данные по аналитическим единицам*, на основании которых можно было бы определить функцию распределения случайной величины, ответственной за вариацию исследуемого явления. В связи с отсутствием представления о *функции распределения в интервалах* затруднены не только аналитические задачи, но

также и прогнозные гипотезы о поведении исследуемого явления как за пределами измеренных диапазонов, так и во временной перспективе.

Существенные трудности возникают при определении и расчёте статистических показателей для подмножеств аналитических единиц внутри диапазонов измерения и *для подмножеств, охватывающих больше одного интервала*, среди которых есть не целые. Описанные трудности объективной неполноты определения вероятностных характеристик исследуемого явления присущи всем статистическим обследованиям, использующим при получении данных наблюдения, идентификацию события, подлежащего регистрации, на основании принадлежности значений некоторого (некоторых) признака (признаков) этого события к заранее заданным интервалам.

Именно такой случай мы встречаем при обработке данных государственного статистического наблюдения по Форме № 1 «Распределение численности работников по размерам начисленной заработной платы», проводимого ежегодно на основании Федеральной программы статистических работ и Производственного плана статистических работ на текущий год. Отчётность предоставляют предприятия, регистрируя данные по численности в заданных интервалах заработных плат и суммы, начисленные работникам, попавшим в соответствующие интервалы.

Таблица 1. Фрагмент Формы № 1 «Распределение численности работников по размерам начисленной заработной платы».

Коды по ОКЕИ: человек-792; рублей-383; тысяча рублей -384

Размер начисленной заработной платы за отчетный месяц, рублей	№ строки	Численность работников всего, человек	Суммы, начисленные работникам, учтенным в графе 3, рублей
1	2	3	4
до 600,0	01		
от 600,1 до 800,0	02		
от 800,1 до 1000,0	03		
от 1000,1 до 1400,0	04		
от 1400,1 до 1800,0	05		
от 1800,1 до 2200,0	06		
от 2200,1 до 2600,0	07		
от 2600,1 до 3000,0	08		
от 3000,1 до 3400,0	09		
от 3400,1 до 4200,0	10		
от 4200,1 до 5000,0	11		
от 5000,1 до 5800,0	12		
от 5800,1 до 7400,0	13		
от 7400,1 до 9000,0	14		
от 9000,1 до 10600,0	15		
от 10600,1 до 13800,0	16		
от 13800,1 до 17000,0	17		
от 17000,1 до 20200,0	18		
от 20200,1 до 25000,0	19		
от 25000,1 до 35000,0	20		
от 35000,1 до 50000,0	21		
от 50000,1 до 75000,0	22		

Свыше 75000,0	23		
Всего работников (стр. с 01 по 23)	24		

Данные по единицам наблюдения (предприятиям) агрегируются в виде регламентной отчётности. Аналитической единицей в этом обследовании является работник, однако данных по отдельному работнику в результатах обследования не предусматривает ни методология обследования ни его инструментарий. Сводные данные представлены в разрезах по формам собственности, отраслям промышленности (виду деятельности) и субъектам административно-территориального деления. Фрагмент сводных данных наблюдения по Карачаево-Черкесской республике приведён в Таблице 2.

Таблица 2. Фрагмент сводной таблицы «Сводные данные о распределении численности, полученные по результатам обследования за апрель 2004 г.» по Карачаево-Черкесской республике.

Размер начисленной заработной платы за отчетный месяц, рублей	№ строки	Численность работников всего, человек	Суммы, начисленные работникам, учтенным в графе 3, тыс. руб.	Средняя заработная плата, рублей
1	2	3	4	5
до 600,0	01	3160.00	1235714.00	391.10
от 600,1 до 800,0	02	3320.00	2348463.00	707.30
от 800,1 до 1000,0	03	2999.00	2786625.00	929.20
от 1000,1 до 1400,0	04	5006.00	5975929.00	1193.70
от 1400,1 до 1800,0	05	4756.00	7560672.00	1589.70
от 1800,1 до 2200,0	06	4810.00	9566648.00	1988.80
от 2200,1 до 2600,0	07	4627.00	11059946.00	2390.20
от 2600,1 до 3000,0	08	4818.00	13388403.00	2778.60
от 3000,1 до 3400,0	09	8977.00	28437282.00	3167.80
от 3400,1 до 4200,0	10	7548.00	28707707.00	3803.20
от 4200,1 до 5000,0	11	5782.00	26738466.00	4624.60
от 5000,1 до 5800,0	12	3966.00	21508411.00	5422.60
от 5800,1 до 7400,0	13	6908.00	45637542.00	6606.30
от 7400,1 до 9000,0	14	2768.00	22332655.00	8068.60
от 9000,1 до 10600,0	15	2044.00	20233797.00	9897.80
от 10600,1 до 13800,0	16	985.00	11725323.00	11907.20
от 13800,1 до 17000,0	17	240.00	3611545.00	15034.20
от 17000,1 до 20200,0	18	170.00	3079813.00	18116.50
от 20200,1 до 25000,0	19	102.00	2333304.00	22825.80
от 25000,1 до 35000,0	20	29.00	859534.00	29639.10
от 35000,1 до 50000,0	21	14.00	576469.00	42701.40
от 50000,1 до 75000,0	22	8.00	475301.00	59412.60
Свыше 75000,0	23	20.00	2091658.00	104582.90
Всего работников (стр. с 01 по 23)	24	**73058.00**	**272271206.00**	**3726.80**

Первичные данные, получаемые от предприятий, как это видно из формы отчётности, представляют собой не точечные значения признаков объекта наблюдения, а суммы признака, относящиеся к регистрируемым интервалам и только к ним. По таким интегрированным данным получить функцию распределения, её параметры, к примеру, численности для уровней заработных плат, нельзя без определенных условных допущений. А значит,

нельзя достоверно рассчитать статистические показатели по любым разрезам численности.

В составе отчётных таблиц обследования есть «Таблица № 3…», отражающая важную информацию о социально-экономическом состоянии распределения вознаграждения за труд. В этой таблице рассчитываются статистические показатели в иных интервалах, а именно в интервалах десятипроцентных долей (децилей) численности. Пример такой таблицы, точнее, её фрагмент, по Карачаево-Черкесской республике приведён ниже.

Таблица 3. Фрагмент Таблицы № 3 «Общие сведения, полученные по результатам обследования за апрель 2004 г.»

Территория 91 Карачаево-Черкесская Республика[1]											
Все формы собственности В том числе по 10-ти процентным группам работников											
	Всего	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Общая сумма средств, направленных на оплату труда, тыс.руб	272271	4352	8196	12682	16818	21480	24008	28705	35896	46698	73437
Удельный вес средств, направленных на оплату труда, %	100.00	1.60	3.01	4.66	6.18	7.89	8.82	10.54	13.18	17.15	26.97
Средняя заработная плата, руб.	3726.8	595.7	1121.9	1735.9	2302.0	2940.1	3286.2	3929.0	4913.3	6391.9	10052.0

На примере практической расчетной задачи получения вычисляемых данных для таблиц, аналогичных Таблице № 3, требующей преобразования измеренных данных к иным интервалам, рассмотрим необходимое доказательство корректности применяемых расчетных процедур и достаточных допущений, делающих такой расчёт возможным. Достоверное преобразование данных, полученных на основе измерений одних интервалов, в другие, без знания функции распределения и её параметров, представляет собой нетривиальную задачу. После завершения статистического наблюдения, сбора и агрегирования данных по всем предприятиям можно оценить распределение численности по всей территории в целом и сделать суждение о форме функции распределения, или, как это принято в параметрической традиции, определить принадлежность эмпирического распределения к параметрическому семейству кривых Пирсона. Эту функцию распределения с позиций нашей, интервальной задачи, следует назвать *глобальной* функцией распределения (ГФР), чтобы отличить её от функций распределения в интервалах измерений, которые логично называть локальными ФР. Их взаимосвязь окажется для нас важной с позиций формулирования вычислительной процедуры расчета межинтервальных показателей.

[1] Данные, показанные в Таблице 3, являются условными, но адекватными реальности структурно.

Учитывая интервальный характер исходных данных, для построения графика эмпирического распределения численности по размерам заработной платы, воспользуемся в качестве точечной интерпретации интервальных данных показателем средней заработной платы в интервале. Это превращает исходные интервальные измерения в эмпирический ряд из 23 точек.

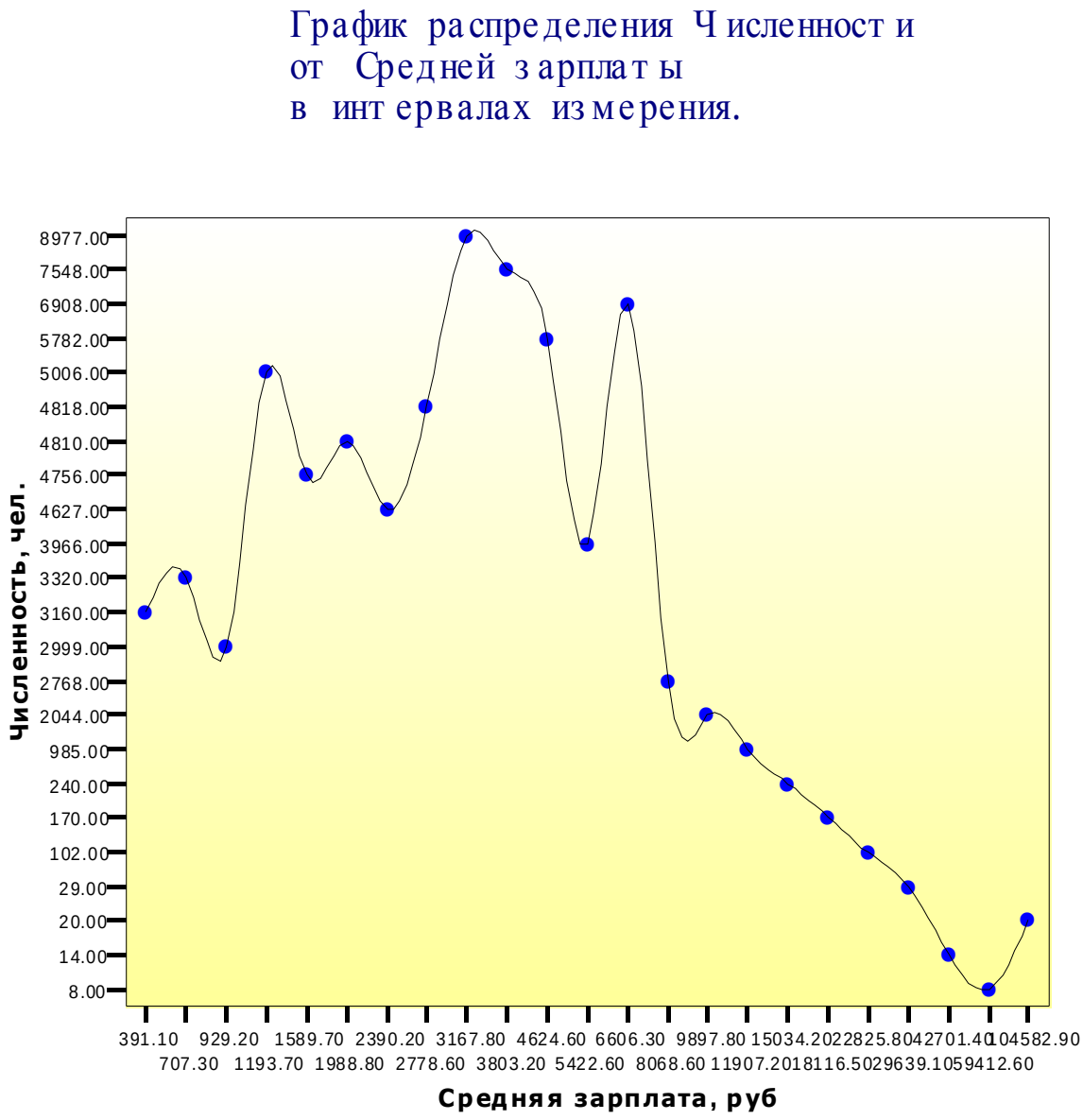

Рисунок 1. Эмпирическое глобальное распределение численности от средней зарплаты.

Эмпирическое распределение, наблюдаемое на Рис. 1 можно оценить как логарифмическое нормальное распределение, с учётом того, что по оси X расположены наблюдаемые события (всего числом 23), то есть – интервалы наблюдения, а средние зарплаты есть лишь метки этих событий.

Для расчета показателей зарплаты по десятипроцентным интервалам численности удобнее представить эту зависимость при ином расположении осей, кроме того, по оси зарплаты отложены не события (измерения), а значения зарплаты.

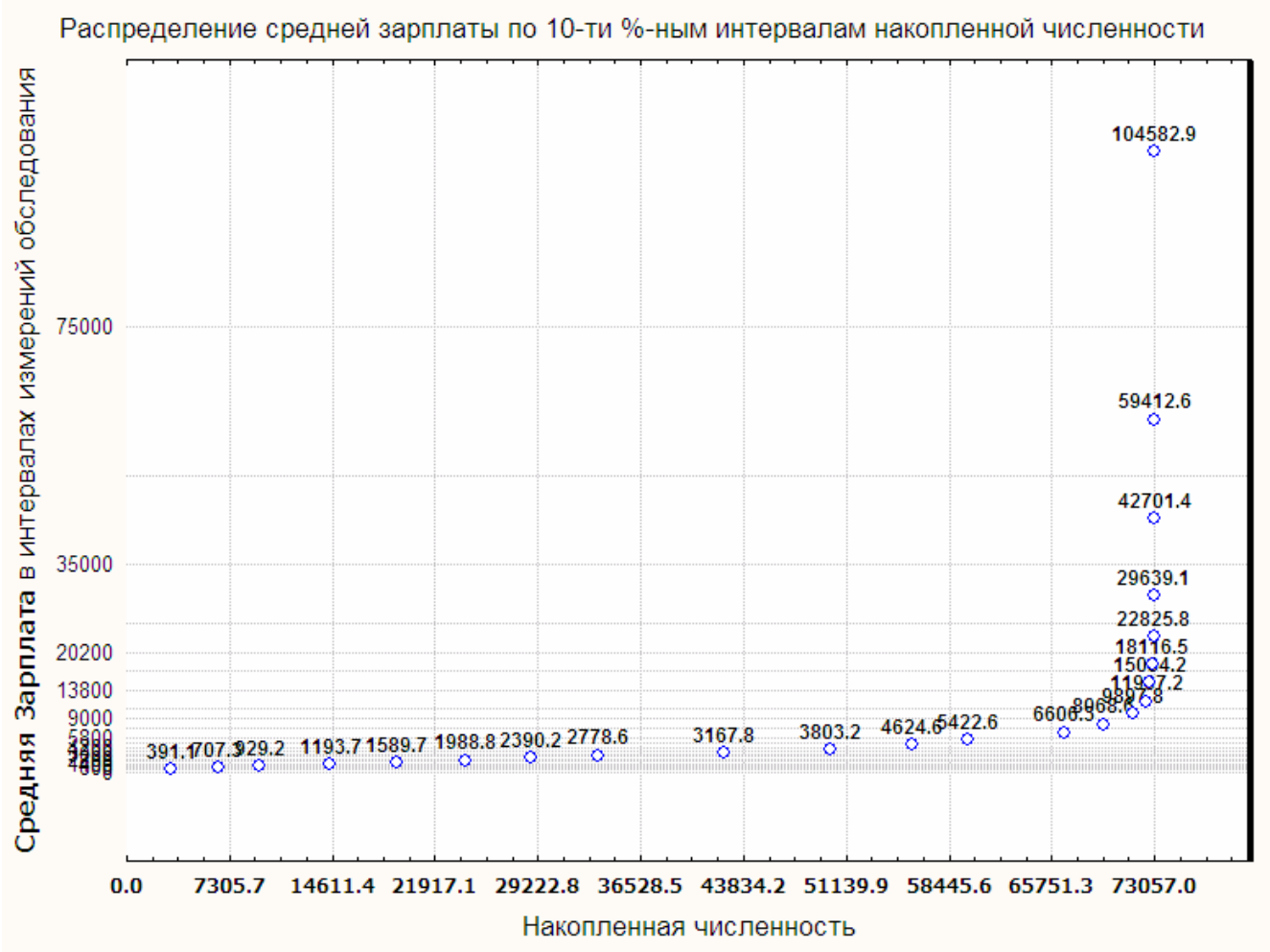

Рисунок 2. Зависимость средней зарплаты от численности.

Средняя зарплата – вычисляемая величина из начисленного фонда зарплаты и численности в интервалах наблюдения. По оси X можно отложить численность в виде гипотетического ряда отдельных работников, отсортированного по величине их заработной платы, это соответствует накопленной численности, что соответствует данным задачи. Исходные данные связаны так:

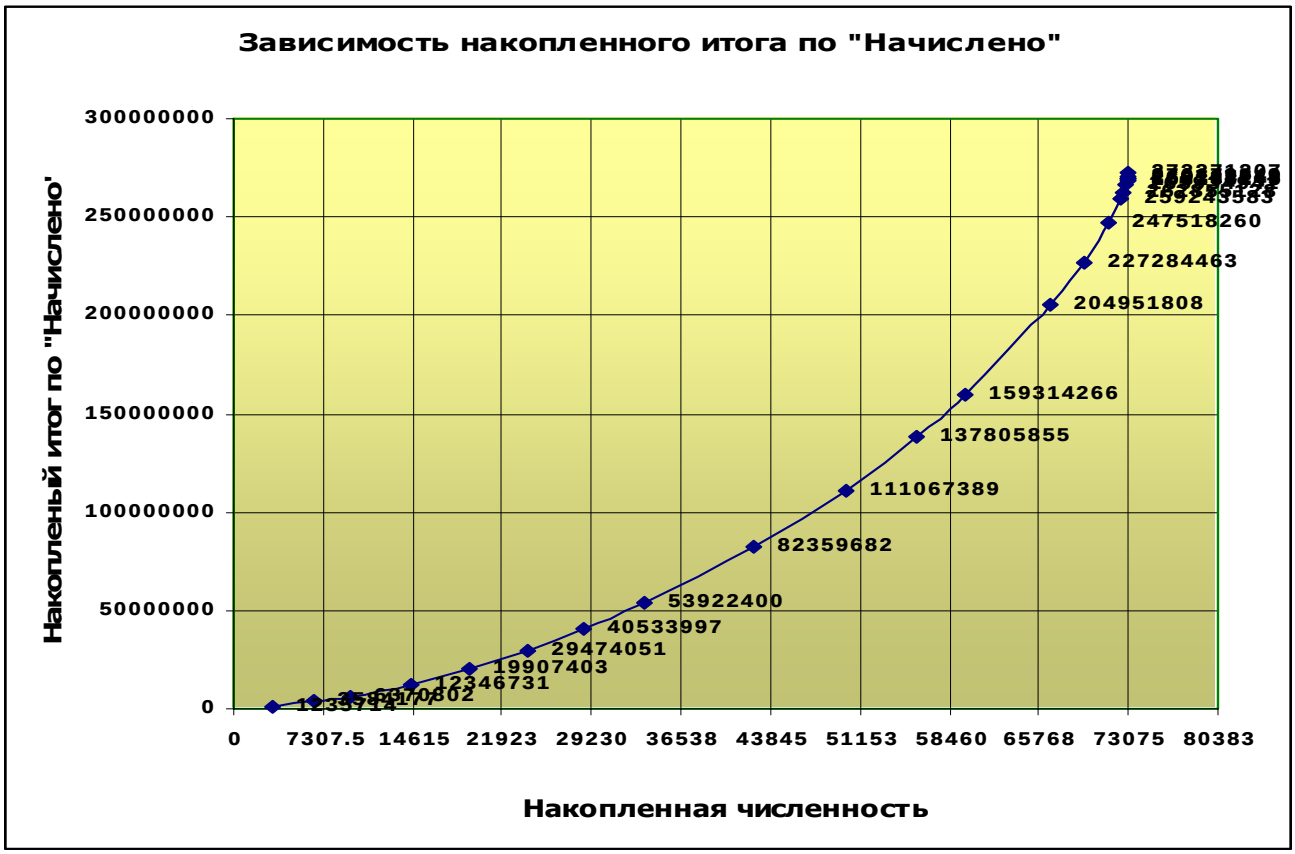

Рисунок 3. Зависимость начисленного фонда заработной платы от численности из наблюдений по 10%-ным интервалам численности.

Или то же в логарифмической шкале и по той же численности и средней зарплатой в интервале измерения. Интервалы измерения и интервалы расчёта (10%-ые численности) не совпадают:

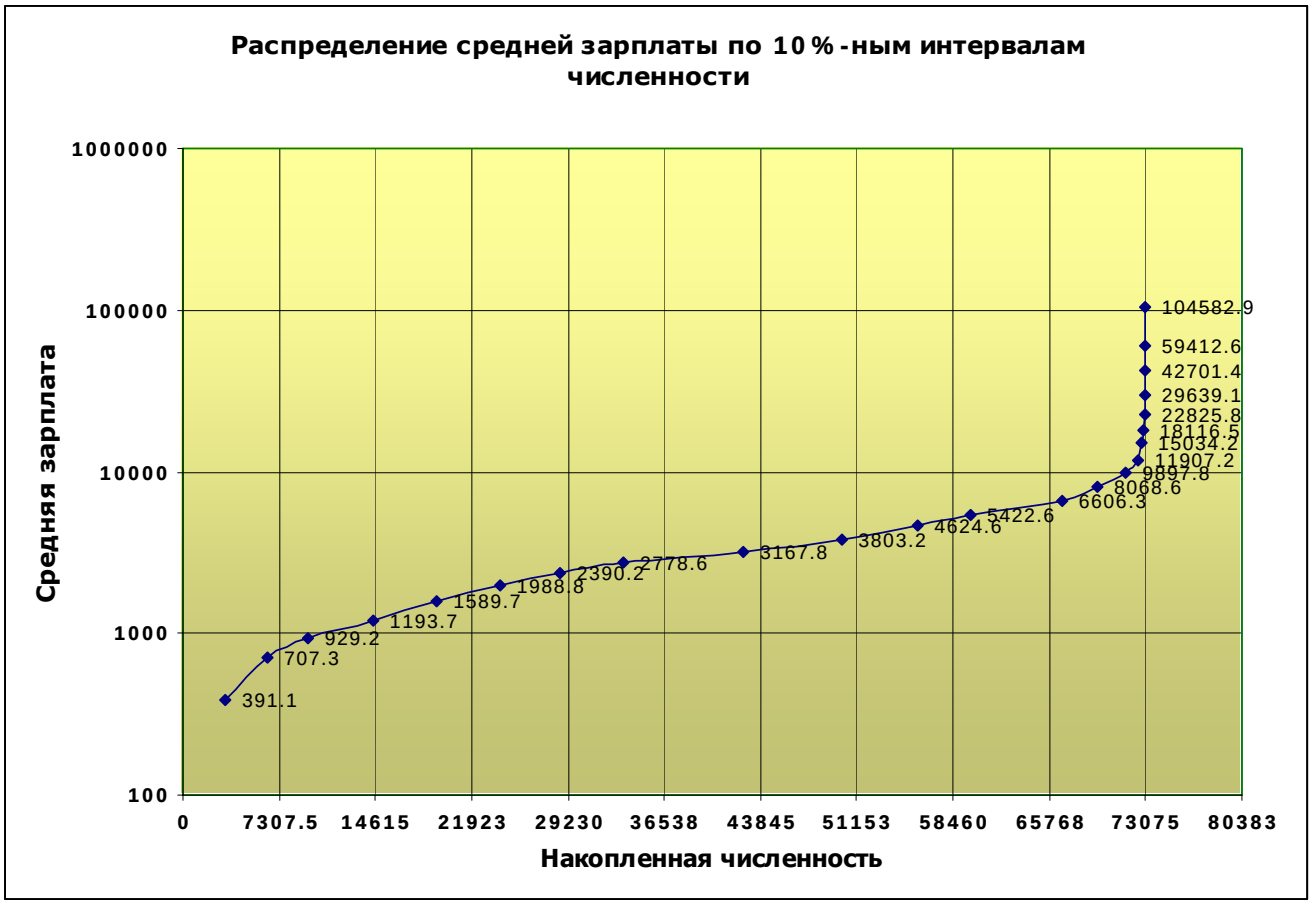

Рисунок 4. Зависимость средней зарплаты от численности в логарифмической шкале.

Конструктивным путём решения задачи создания алгоритма пересчёта данных из одних интервалов в другие можно было бы считать получение таких аппроксимирующих функций по ВСЕМ исходным интервалам, *ошибка аппроксимации* по которым *не превосходила бы*, или, в лучшем случае, была сравнима с заданной ошибкой выборки. Возможность перерасчёта интервалов возможна, если посчитать, что, в отсутствии информации о динамике изменения зарплаты, в пределах интервала измерения все зарплаты работников одинаковы и равны средней. Однако такое заведомое упрощение не может считаться нами удовлетворительным приближением реальному положению. Наша задача - попытаться достоверно и доказательно всё-таки учесть *неравномерность* распределения в интервале. *Конструктивность* здесь понимается в *интуитивистском* смысле, то есть - в возможности получения *эффективного процесса* исчисления интересующего нас показателя, поэтому строго – это требование конструктивности модели представления - задано нами ограничено [**5**, стр. 68, **7**, стр. 235]. Неконструктивное решение задачи перерасчёта внутри и между интервалами измерения возможно, например, в случае использования нейронной сети (НС), обученной на предыдущих наблюдениях, в том числе, на иных интервалах. Численные значения внутри интервалов такая обученная НС будет давать, однако, такое решение не будет являться в полной мере *конструктивным*, так как у нас не будет возможности записать способ этого решения в каком-либо последовательном формализме

для передачи, к примеру, его на реализацию в виде алгоритма, разрешимого по Тьюрингу.

Несмотря на то, что мы не знаем точного вида и параметров зависимости внутри интервалов, мы можем определить несколько ограничивающих правил, которые вытекают из сущности явления и механизма наблюдения. Это позволит максимально возможно ограничить класс функций, пригодных для аппроксимаций и достижения *эффективного механизма* расчёта показателей фрагмента интервала. Такой механизм позволит производить любые, ограниченные допущениями, перерасчёты внутри и между интервалами. В рассматриваемых нами условиях измерений нельзя даже строго говорить об аппроксимации, так как точечные значения (траектория) внутри интервала нам неизвестны и недоступны. Мы можем сделать лишь некоторые заключения о свойствах класса возможных форм зависимостей, которые позволяет содержание исследуемого явления.

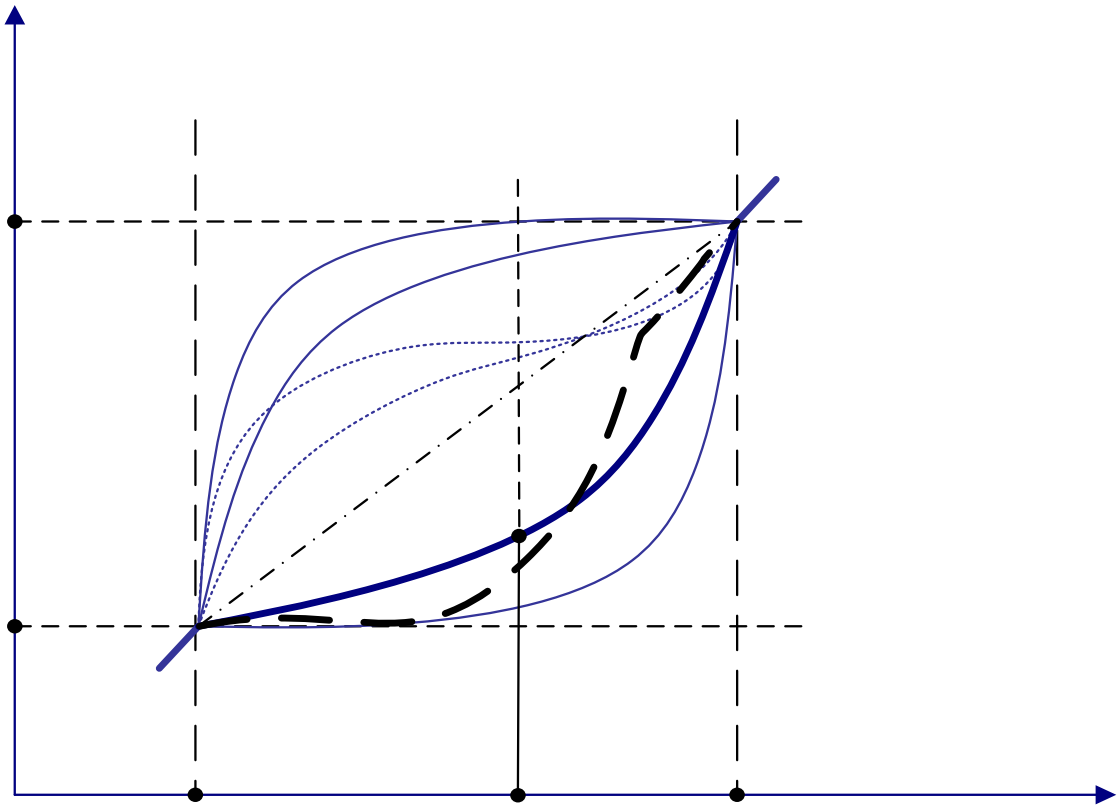

Рисунок 5. Класс функций, допустимых для интерпретации зависимости <<Численность-Зарплата>>.

На **Рисунок** 5 показана обобщённая картина зависимости в пределах одного интервала, где:

x_k , x_{k+1} — границы интервала накопленной численности (ряд номеров работников, упорядоченный по размеру заработной платы), в который попали работники, имеющие заработную плату в пределах интервала y_m , y_{m+1} ;

y_m , y_{m+1} — границы интервала заработной платы, заданные в инструментарии наблюдения, таких интервалов 23;

x_l — точка внутри рассматриваемого интервала, произвольно его делящая, конкретно – граница 10%-го интервала численности;

ψ_i — функции, допустимые для описания зависимости из класса, который *требует определения*.

Исследуемая реальная зависимость может быть представлена линиями функций ψ_i из некоего обобщённого класса, помеченными латинскими буквами $a,b,c,d,e,f.$ Исследуемая нами расчётная модель должна позволять вычислять площади фигур, аналогичных фигурам $\{x_k PT x_l\}$ или $\{x_l TQ x_{k+1}\}$.

Запишем формально определение класса функций $\aleph$, такого, что он вполне удовлетворяет описанию исследуемых нами зависимостей, без ограничений тривиальности.

1. По оси X мы отложили накопленную численность, или, что то же самое, перенумеровали работников, расположенных по возрастанию размера заработной платы. Это корректное допущение, несмотря на то, что реально такие данные нам не доступны по условиям проведения наблюдения по Форме № 1. Мы в действительности имеем только интегрированные (суммарные) данные как по численности, так и по накопленной заработной плате в интервале измерения, но при этом имеем право не забывать, откуда и как эти данные формируются предприятием, составляющим отчёт по Форме № 1. Следовательно, функции, определяющие зависимость, могут быть только *неубывающими*, то есть производная такой функции в интервале всегда неотрицательна:

$$\forall x(x \in [x_k, x_{k+1}]) \quad \exists \Psi : \frac{d\Psi}{dx} \geq 0 \qquad (1)$$

2. Интегральные показатели измерений фиксированы по интервалам.

$$\forall i \forall j (i, j \in \Re) \;\; \exists \Psi_i \exists \Psi_j : \int_{x_k}^{x_{k+1}} \Psi_i dx = \int_{x_k}^{x_{k+1}} \Psi_j dx = C_k \qquad (2)$$

где C_k – накопленная заработная плата в интервале (фонд заработной платы).

3. Конструктивное ограничение. Никакая информация о наблюдаемом социально-экономическом феномене не позволяет нам ещё более сузить определение класса допустимых функций $\aleph$. По условиям проведения наблюдения такой информации нет. Таким образом, мы попадаем в ситуацию, когда про утверждение, что «такая-то функция, отвечающая требованиям (**1**) и (**2**), описывает исследуемую зависимость» мы не можем однозначно сказать, истинно оно или ложно. В самом деле, выбранная нами функция зависимости, удовлетворяющая (**1**) и (**2**), не противоречит эмпирическим данным, и, стало быть, утверждение истинно, в то же время, мы можем указать сколь угодно много функций, удовлетворяющих тем же условиям, но дающих существенно отличающиеся результаты при расчёте внутри интервала площадей фигур вида $\{x_k P T x_l\}$ или $\{x_l T Q x_{k+1}\}$ из *Рисунок 5,* как это видно для кривых e и g**.** Обоснование доказательств в формальных системах без закона «исключённого третьего» не входит в задачу этой статьи, так как это существенно затруднило бы формулирование простого расчётного правила или алгоритма, к которому мы стремимся. Пришлось бы дополнительно обосновывать критерии приемлемости, к примеру, энтропийные, или рассчитывать аттрактор траекторий, что поставило бы нас перед требованием больших рядов наблюдений, которое не всегда достижимо. Поэтому мы пойдём на введение *дополнительного ограничения*, имеющего внешний характер в отношении рассматриваемой проблемы и не связанного с логикой, описывающей исследуемое явление, так как мы показали, что она не полна для решения нашей, сугубо расчетной, в конечном итоге, задачи.

   Обобщение форм зависимостей, которые мы наблюдаем на ***Рисунок*** *5,* позволяют нам выбирать без потери общности между классами гипербол вида

$$Y = \frac{mx+n}{px+q} \qquad (3)$$

и экспонентами вида:

$$Y = \frac{a}{p} e^{(bx+c)} \qquad (4)$$

Однако экспоненциальная форма зависимости представляется более предпочтительной. Экспонента – самая быстрая функция, и, кроме того, в вычислительном смысле, упрощается интегрирование, которое нас особенно интересует. В пользу экспоненты, как формы зависимости, наиболее радикально отражающей скорость изменения (дифференциации), можно привести и совсем не математическое обоснование предпочтения. В рассматриваемой задаче речь идет, в частности, о объемах фонда заработной платы, потребляемых группами персонала, существенно отличающихся по степени дифференциации заработной платы между ними. Другими словами, одной из целей проводимого наблюдения является определение степени разброса зарплаты между наименее и наиболее оплачиваемыми группами работников. Это важный фактор социального напряжения. Экспоненциальная интерпретация позволяет легко определять по графику зависимости: какова степень и тенденция этого фактора социального напряжения, к какому полюсу ближе ситуация. Чем ближе полученная в наблюдении зависимость к кривой вида a из ***Рисунка*** *5*, тем, очевидно, дифференциация зарплат мягче. Зарплаты быстро возрастают с минимума низкооплачиваемого персонала к среднеоплачиваемому и медленно продолжают расти к группе начальников. Напротив, если наблюдённая зависимость больше похожа на кривую вида f, можно говорить о возможной тенденции роста социальной напряжённости. В этом случае почти весь фонд заработной платы рассматриваемого интервала потребляет высокооплачиваемая группа работников, при этом весьма немногочисленная, а разница в зарплатах от среднеоплачиваемых к начальникам возрастает экспоненциально, то есть очень быстро. Отношение к тому или иному виду экспоненциальной функции определяется параметрами выражения (**4**), которые вычисляются на основе реальных данных обследования.

Итак, удобство визуальной оценки и простота расчётного правила (алгоритма) дополнительно склоняет нас выбрать для оценки зависимости в интервалах функции класса **(4)**. Интегральное уравнение **(3)** сокращённо можно конкретизировать так:

$$\int_{x_k}^{x_{k+1}} \frac{a}{p} e^{(bx+c)} dx = C_k \qquad (5)$$

Решение этого уравнения в конкретных параметрах по всем интервалам x_k и даст нам возможность пересчёта на любые диапазоны, не совпадающие с интервалами обследования.

Эта методика оценки параметров распределения корректно обобщается на любые интервалы и позволяет получить если не реальную форму распределения, то как минимум его «динамические» характеристики, что при отсутствии данных внутри интервала существенно увеличивает наше представление о характере исследуемого процесса и нашу возможность оперировать данными субинтервальных диапазонов.

* * *

Необходимо упомянуть, что при подготовке материалов этой статьи существенная помощь в виде обсуждения многих аспектов процедуры обследования и свойств вычисляемых результатов была получена от главного специалиста технологического отдела территориального комитета Карачаево-Черкесской республики Олега Владимировича Бобрышева. Отдельная благодарность заместителю начальника отдела статистики численности и оплаты труда работников организаций Управления статистики труда, образования, науки и культуры Росстата Жихаревой Ольге Борисовне за системное формулирование существенных вопросов к исследуемой проблематике.

Литература

1. Вощинин А.П. Метод оптимизации объектов по интервальным моделям целевой функции. - М.: МЭИ, 1987.
2. Гуссерль Э. Логические исследования. – М.: АСТ 2000.
3. Дубров А.М., Лагоша Б.А., Хрусталев Е.Ю. Моделирование рисковых ситуаций в экономике и бизнесе: Учебное пособие. – М.: Финансы и статистика, 1999.
4. Иванов Ю.Н., Токарев В.В., Уздемир А.П. Математическое описание элементов экономики. – М.: Физматлит, 1994.
5. Карри Х. Основания математической логики. – М.: Мир, 1969.
6. Кендалл М., Стюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976.
7. Клини С. Математическая логика. – М.: Мир, 1973.
8. Краснощеков П.С., Петров А.А. Принципы построения моделей. – М.: Фазис, 2000.
9. Курбатов В.И., Угольницкий Г.А. Математические методы социальных технологий. - М.: Вузовская книга, 1998.
10. Прохнов Ю.В., Розанов Ю.А. Теория вероятностей. – М.: Наука, 1973.
11. Пригожин И. Конец определённости. – Москва-Ижевск, R&C Dynamies, 2001.
12. Шокин Ю.И. Интервальный анализ. – Новосибирск: Наука, 1981.