Е.В. Ермолицкая, ***соискатель***
Белорусский государственный экономический университет
УДК 519.23

Инструментально-ориентированный подход при использовании методов статистического анализа

В статье наряду со стандартными подходами рассмотрены альтернативные методы (инструментально-ориентированные) работы со статистической информацией, получения статистических группировок с целью более углубленного анализа. Реальность и актуальность предлагаемого подхода с наличием эффективных инструментальных средств продиктованы прежде всего перестройкой в сфере информационных технологий, связанной с "выдавливанием" из этой сферы малоквалифицированных манипуляторов компьютерными процедурами. Особо важным такой подход видится в сфере управления в связи с постоянным и актуальным совершенствованием управленческих структур и переобучением современного управленческого персонала.

Along with standard approaches the article considers alternative (instrument oriented) methods of working with statistics, making statistical grouping for the more advanced analysis. Such approach would be most important for the managerial sphere, keeping in mind constant and up-to-date improvement and retraining of the managerial staff.

Предметом статистики вообще являются массовые явления любой природы, в частности, их количественная сторона. Одной из важнейших ее отраслей считается экономическая статистика, предмет которой составляет количественная характеристика массовых явлений и процессов в экономике. Она же есть вид "…практической деятельности органов государственной статистики" [1]. В данном аспекте можно говорить об основной цели статистического исследования – **выявления закономерностей в экономике государства.** Поскольку экономические данные имеют статистическую природу, то для их анализа и обработки применяются специальные методы – статистические.

В зависимости от объекта и цели статистического исследования, а также его этапа можно выделить несколько групп статистических методов. На этапе сбора данных – это статистическое наблюдение. При обобщении данных и первичной обработке информации – это группировки, сводки, ряды распределения. На этапе представления данных – это статистические таблицы, графики, карты.

Наиболее сложным и заключительным этапом любого статистического исследования является анализ и интерпретация данных. Среди методов статистического анализа можно выделить метод обобщающих статистических показателей, выборочный метод, корреляционный и регрессионный анализ, метод динамических рядов, индексный метод и др.

Среди вышеперечисленных важнейшую роль играют методы математической статистики, так как закономерности в экономике выражаются в виде зависимостей экономических показателей и математических моделей их поведения. Более того, эконометрический анализ служит основой для экономического анализа и прогнозирования.

Но математика начинает работать там, где есть модель экономического процесса или явления и исходные данные, организованные определенным образом. В дальнейшем будет сделан акцент на инструментально-ориентированные методы организации статистических данных, доступа к ним и анализа.

На сегодняшний день первичная статистическая отчетность, регулярно предоставляемая субъектами хозяйствования территориальным органам государственной статистики, является основой при формировании статистических баз данных, а статистические группировки и сводки, получаемые органами государственного управления всех уровней, – информационной основой для принятия того или иного управленческого решения.

Схема информационного наполнения и работы стандартного статистического комплекса представлена на рисунке 1.

Основу информационного наполнения любой системы составляют базы первичных данных. Как правило, это набор плоских таблиц (статистических совокупностей большого объема), строками которых являются сами наблюдения, а столбцы содержат признаки, характеризующие каждую единицу совокупности. Такая информация жестко структурирована и организована средствами систем управления базами данных (СУБД) в виде списочных структур, пополнение и доступ к которым осуществляется программистом при помощи программного кода. Опять же формирование выходных таблиц производится программистом согласно жесткому алгоритму, прописанному в техническом задании. Другими словами, группировка статистических данных имеет строго регламентированный вид. Чаще всего основанием группировки является атрибутивный признак (территория, форма собственности, вид экономической деятельности) и группы располагаются в подлежащем статистической таблицы. В сказуемом располагаются обобщающие статистические показатели. Благодаря группировке данные приобретают систематизированный вид. На ее основе рассчитываются

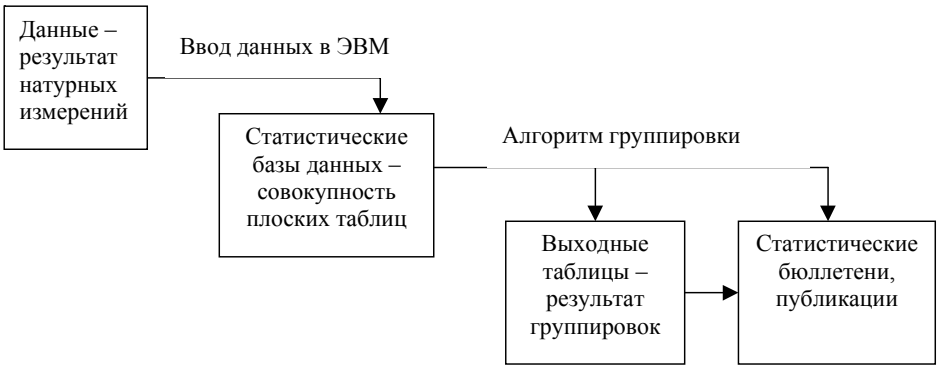

Рис. 1. Схема работы статистического комплекса

№ группы	*Значение группировочного признака*	*Число единиц совокупности в группе*	*Обобщающий показатель 1*	*Обобщающий показатель 2*	…
1 2 … m					
ИТОГО					

Рис. 2. Статистическая группировка

сводные показатели по группам, появляется возможность их сравнения. Макет таблицы типичной статистической группировки представлен на рисунке 2.

Как было указано выше, на этапах первичной обработки и представления данных такой подход к агрегированию и анализу данных является актуальным. Однако нет необходимости подробно описывать цепь взаимодействия "пользователь – программист", если требуется получить группировку, несколько отличную от регламентной. Часто эти "итерации" носят продолжительный и не всегда полностью согласованный характер. Другими словами, доступ пользователя-аналитика к статистическим данным организуется в виде информации, предоставляемой ему сквозь "матрицу" жестко прописанной постановки. Подобная парадигма приводит к ряду проблем. Некоторые из них приведены ниже.

Во-первых. Анализ должен быть оперативным, группировки – многоальтернативными, позволяющими увидеть проблему с различных точек зрения и выявить связи между изучаемыми признаками. Но в то время как "обслуживающий" аналитика персонал в области информационных технологий (ИТ) "общается" с компьютером, сам аналитик "пребывает" в информационном вакууме, тогда как время отклика на запрос, порожденный в его голове, должно быть настолько мало, "…чтобы не успели разомкнуться ассоциативные связи, породившие данный запрос" [2].

Во-вторых. Основой выборочного метода при анализе статистических совокупностей является закон больших чисел, "..суть которого состоит в следующем: с увеличением объема выборки вероятность появления больших ошибок и пределы максимально возможной ошибки уменьшаются (чем больше обследуется единиц, тем меньше будет величина расхождений выборочных и генеральных характеристик)" [3]. Центральная предельная теорема закона больших чисел гласит, что при увеличении объема выборки распределение выборочной средней приближается к нормальному. А получить выборку достаточно большого объема можно только из первичных массивов.

В третьих. В последнее время статистические бюро многих стран мира все чаще используют технологии ГИС (географическая информационная система) для распространения статистических данных в виде тематических карт. Однако при помощи ГИС решаются и другие, более сложные задачи, носящие аналитический характер. При построении пространственной модели определяющим является уровень агрегации данных по территориальному признаку. Очевидно, что более детальный и точный пространственный анализ возможен при как можно большей дезинтеграции изучаемого территориального объекта (например, региона) на составляющие части (например, населенные пункты). Понятно, что построение и выбор варианта пространственной модели, к которой присоединятся статистические данные с целью их дальнейшего анализа, – прерогатива аналитика.

Единственным и естественным выходом из создавшейся ситуации видится исключение на этапе анализа статистических совокупностей из схемы **данные – программист – информация – аналитик – знания** "передаточного механизма" в лице программиста и передача полномочий доступа к данным и информации самого аналитика. В данном случае можно говорить об **инструментальном комплексе** – совокупности инструментальных компонент, каждая из которых, предоставляя пользователю свой набор инструментов, делает его центральным и активным элементом информационной аналитической системы и независящим от специалистов по различным аспектам ИТ.

Среди основных видов ИТ, применяемых при организации статистических совокупностей, визуализации и анализа данных выделим следующие:

• технология СУБД (системы управления базами данных);

• технология статистической обработки и анализа данных;

• технология OLAP (On Line Analytical Processing, оперативный анализ данных);

• ГИС-технология.

Не будем перечислять наиболее современные программные продукты, представляющие эти технологии. Оговорим лишь, что каждый из них имеет встроенный интерфейс, интуитивно понятный пользователю-непрограммисту, а также инструментарий (как правило, представленный в виде вкладок меню и панелей инструментов), позволяющие аналитику в зависимости от решаемых задач строить различного рода информационные модели и на их основе "манипулировать" информационным ресурсом системы.

Несомненно, пользователем такой системы может быть человек, имеющий четкое представление о методах построения информационных моделей и основных понятиях реляционной алгебры, умеющий пользоваться стандартными средствами импорта и экспорта данных из одной среды в другую. И, что является главным (в данном случае идет речь о статистике), иметь навыки в работе с основными методами организации и анализа статистических совокупностей, в широком спектре представленными такими мощными программными продуктами как пакеты статистической обработки и анализа данных (SPSS, STATISTICA). Следует отметить, что с вышеуказанными понятиями знакомят курсы высшей математики и статистики, экономической информатики и эконометрики, преподаваемые студентам многих экономических специальностей.

На практике подобный подход был использован автором при анализе данных в области демографической статистики. Для работы использовался инструментарий СУБД Access97 пакета статистической обработки и анализа данных STATISTICA 6, а также информационный ресурс банка данных переписи населения Республики Беларусь 1999 г.

Банк данных переписи населения включает в себя несколько информационных компонент. Основной является база данных первичной информации. Это две генеральные статистические совокупности. Первая представляет собой более 10 млн. наблюдений по каждому жителю республики, вторая – почти 4 млн. по каждому домашнему хозяйству. Обе организованы в виде плоских таблиц, строками которых являются сами наблюдения, а столбцами – признаки. Среди основных признаков наблюдений первой отметим пол, возраст, состояние в браке, национальность, уровень образования каждого человека, а наблюдений второй – тип жилого помещения, вид благоустройства, вид собственности жилья и др. Основным атрибутивным признаком каждого наблюдения является территория – десятизначный код СОАТО (система обозначений административно-территориальных образований).

Второй важнейший компонент информационного ресурса – это совокупность выходных таблиц (рис. 1) – результат определенных техническим заданием группировок. На их основе были сформированы и опубликованы сборники по итогам переписи населения, а также произведен анализ основных показателей, отражающих динамику процессов, происходящих в области демографического развития Республики Беларусь за ряд лет. Для ввода первичной информации в компьютер и получения выходных таблиц был разработан статистический комплекс, позволяющий получить любую, прописанную в постановке группировку, по любому, предусмотренному ею же, разрезу.

Таким образом, был строго соблюден принцип регламентации, отраженный на рисунке 1, который на этапах обработки, обобщения и представления данных является основой при проектировании и внедрении статистического вычислительного комплекса.

А теперь рассмотрим в пошаговом режиме технологию доступа к информации, когда наряду с работой с данными регламентных таблиц применены альтернативные методы (инструментально-ориентированные) использования первичных статистических массивов с целью более углубленного анализа.

1. Выявление проблемной ситуации в области демографического развития территорий.

Практически по всей территории республики на селе в репродуктивных возрастах число мужчин превышает количество женщин, тогда как в городских поселениях – наоборот, что явно сказывается на интегральных показателях демографической безопасности как республики в целом, так и отдельных ее территориальных образований. На рисункне 3 представлена гистограмма, наглядно демонстрирующая дисбаланс в этой области. Одним из факторов, порождающих его, является миграция молодых женщин из села в город.

Гистограмма – результат визуализации одной из выходных регламентных таблиц. Для ее построения было достаточно использования инструментария пакета MS Excel.

2. Выдвижение статистических гипотез.

Можно предположить, что одной из причин миграции молодых женщин из села в город является низкий уровень благоустройства жилья. Опять же для визуализации сложившейся ситуации с жильем достаточно использования регламентированных группировок. Так, например, в таблице 1 представлена сравнительная характеристика жилищного фонда городских и сельских населенных пунктов с точки зрения благоустройства.

Из таблицы видно, что по благоустройству жилые помещения сельской местности сильно уступают городским. Кроме того, 66,9% жилья на селе построены из дерева, а в городских поселениях 84,5% – из кирпича, бетона, железобетона, блока, панели.

Итак, выдвинуты две альтернативные гипотезы:

H_0 – структура полов в сельской местности не зависит от благоустройства жилья;

H_1 – структура полов в сельской местности зависит от благоустройства жилья.

Для принятия той или иной гипотезы необходимо количественно оценить вышеуказанную зависимость. Для этого применим методы математической статистики с использованием инструментария пакета STATISTICA.

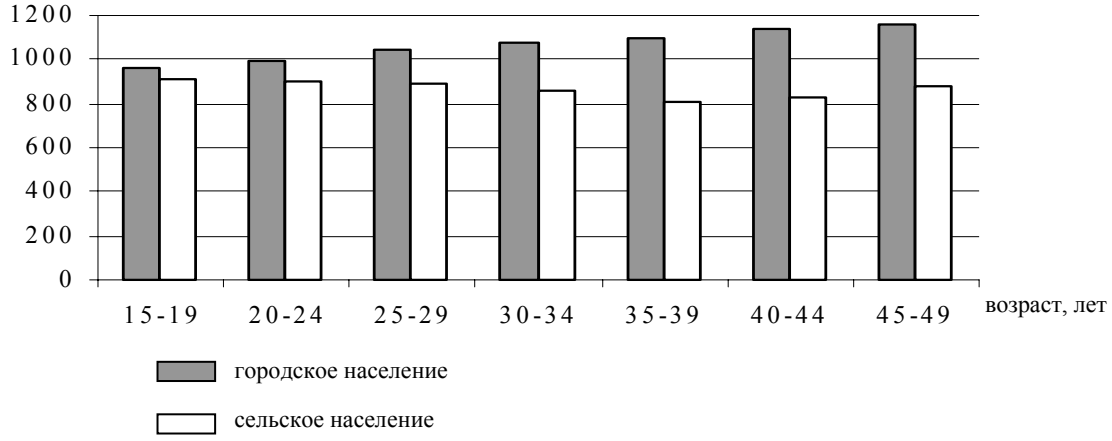

Рис. 3. Число женщин на 1000 мужчин по Республике Беларусь

Таблица 1. Доля жилищ с отдельными видами благоустройства в общей численности жилищ, %

	Телефон	*Газ сетевой*	*Центральное отопление*	*Централизованное горячее водоснабжение*	*Водопровод*	*Канализация*	*Ванная или душ*
Город	71,6	79,0	72,4	87,4	84,5	81,1	71,9
Село	13,7	11,2	6,1	24,2	20,7	17,2	25,7

Аналитические группировки на основе первичных массивов для изучения взаимосвязей между признаками строились при помощи инструментария СУБД Access.

3. Количественная оценка взаимосвязи структуры полов от благоустройства жилья. Выделим 3 этапа.

3.1. Для количественной оценки взаимосвязи признаков последние должны быть выражены количественно, поэтому введем некоторые коэффициенты:

Кж/м – коэффициент соотношения полов в репродуктивных возрастах (сколько женщин приходится на одного мужчину);

Кпериод – удельный вес жилых помещений, построенных в 1960 г. и позднее;

Кматер – удельный вес помещений из кирпича, бетона и железобетона;

Кблаг – коэффициент благоустройства, представляющий среднюю арифметическую удельных весов помещений с отдельными видами благоустройства в общей численности жилищ.

Теперь задача сводится к оценке взаимосвязи Кж/м от всех остальных.

3.2. Нам необходим массив информации (выборка) достаточно большого объема. Используя инструменты СУБД для построения запросов и первичные массивы, получим группировку из 264 строк (группировочный признак – код территории, условие выборки – сельские Советы Витебской области). Далее для каждого наблюдения ставится в соответствие значение переменных (Кж/м – зависимая переменная, Кпериод, Кматер, Кблаг – факторные). Для их расчета применялись стандартные математические функции MS Access. В конечном итоге получена аналитическая группировка, представленная на рисунке 4.

3.3. Статистический анализ с использованием пакета STATISTICA.

Для импорта данных использовался инструментарий стандарта ODBC (Open Database Connectivity, открытый доступ к базам данных), в результате чего таблица, представленная на рисунке 4, была размещена в среде STATISTICA. Принимая во внимание большой объем

№ наблюдения	*Код территории по СОАТО*	*Кж/м*	*Кпериод*	*Кматер*	*Кблаг*
1.	2205762	0,79	0,30	0,15	0,023
2.	2205805	0,81	0,48	0,13	0,069
…					
264					

Рис. 4. Аналитическая группировка

выборки, тест на нормальность распределения не проводился. При помощи гистограмм установлено, что формы распределения переменных приближены к нормальному распределению и, следовательно, для оценки взаимосвязи переменных возможно применение линейной корреляции Пирсона. Обозначив Кж/м – зависимой переменной, Кпериод, Кматер, Кблаг – факторными, получим следующую корреляционную матрицу:

	Кпериод	Кматер	Кблаг
Кж/м	0,37	0,31	0,39

Уровень статистической значимости p=0,05.

Итак, в результате статистического анализа гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1.

Несмотря на "кустарность" вышеприведенных процедур, автору удалось, применяя инструментальные компоненты популярных программных продуктов, работать с большими объемами информации, не прибегая к написанию программных кодов, помощи программиста и средствам громоздких вычислительных комплексов.

Отдельно следует сказать о таких современных информационных технологиях, как OLAP-и ГИС- технологии.

С возрастанием количества группировочных признаков наглядность таблиц, построенных при помощи комбинационных группировок (последовательное агрегирование), ухудшается. Поэтому наряду с последними часто используют методы многомерных группировок. Цель этих методов – классификация данных на основе множества признаков, а более обоснованным методом многомерной классификации является кластерный анализ. Каждая единица совокупности в кластерном анализе рассматривается как точка в заданном признаковом пространстве. С развитием вычислительной техники и внедрением программных продуктов, ориентированных на конечного пользователя-аналитика, все чаще на практике применяется OLAP-технология, 12 основных признаков которой были изложены Коддом [4]. Основанная на принципах геометрического представления информации и на общих закономерностях человеческого мышления, она призвана "…повысить эффективность информационно-аналитической и управленческой деятельности руководящего персонала"[2]. Если говорить о статистике, технологию применения OLAP-машин можно выразить следующим образом: статистическая совокупность (первичный массив) – выбор группировочных признаков – выбор измерений, соответствующих каждому из них – группировка – получение многомерного OLAP-куба – многомерный анализ.

Удобный инструментарий для построения OLAP – кубов содержит компонент MS Query распространенного интегрированного пакета MS Office. Используя опять же стандарт ODBC, можно подключиться к любому источнику данных (например, базам данных MS Access, Visual FoxPro, SQL – серверам и пр.) для создания запроса и построения OLAP – куба, а результат в виде многомерной сводной таблицы разместить в среде MS Excel. Отметим, что последний содержит набор инструментов, позволяющий проводить как статистический анализ, так и строить различные графики и диаграммы.

Основным информационным компонентом любой ГИС являются цифровые карты. А каждая цифровая карта описывается атрибутивной таблицей, основной атрибут которой – это территориальный код элемента карты, по которому и привязывается информация. Для подключения к пространственной основе статистических данных они должны быть сгруппированы по территориальному признаку. Используя инструментарий любой реляционной СУБД и конструктор запросов, пользователь может построить перекрестную таблицу, строками которой будут коды территорий, а в столбцах разместятся необходимые показатели. Каждый ГИС-пакет в своем составе имеет встроенный ODBC – протокол, позволяющий импортировать большие массивы информации из различных сред, в том числе и вышеупомянутые сгруппированные статистические данные, которые по общему атрибуту можно связать с картой. Кроме того, статистический пакет SPSS содержит в своем составе инструментарий, позволяющий наряду с организацией статистических совокупностей и их сложнейшего анализа создавать геосеты (наборы слоев карты) и привязывать к ним статистические данные с целью их визуализации.

Из всего вышесказанного можно сделать следующие выводы:

1. Наряду с эксплуатацией вычислительных комплексов предлагается делегировать полномочия распорядителя информационными ресурсами пользователю-аналитику, предоставляя ему неограниченный доступ к информации и данным. Это продиктовано прежде всего перестройкой в сфере информационных технологий, связанной с "выдавливанием" из этой сферы малоквалифицированных манипуляторов компьютерными процедурами.

2. Реальность и актуальность предлагаемого подхода с наличием эффективных инструментальных средств с рассмотренными функциями очевидны.

3. Особо важным такой подход видится **в сфере управления** в связи с постоянным и актуальным совершенствованием управленческих структур и переобучением современного управленческого персонала. Современные менеджеры могут иметь историческую перспективу лишь как специалисты, обученные и владеющие указанными типами инструментальных средств компьютерных технологий, выступающие как системные администраторы их и распорядители ресурсов компенсации риска намеченных к управлению параметров.

Литература

1. Экономическая статистика: Учебник / Под. ред. Ю.Н. Иванова. – Москва: ИНФРА-М, 1999.

2. Архипенков С. ORACLE Express OLAP. – Москва: ДИАЛОГ – МИФИ, 2000.

3. Елисеева И.Е., Юзбашев М.М. Общая теория статистики. – Москва: Финансы и статистика, 2002.

4. ORACLE 8: Энциклопедия пользователя. – Москва: Изд. "ДиаСофт", 1998.