ВЫСШЕЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАНИЕ

ПРОГРАММНЫЕ СТАТИСТИЧЕСКИЕ КОМПЛЕКСЫ

Допущено
Учебно-методическим объединением вузов по университетскому политехническому образованию в качестве учебного пособия для студентов высших учебных заведений, обучающихся по специальности «Стандартизация и сертификация»

ACADEMIA

Москва
Издательский центр «Академия»
2011

УДК 83.77.31(075.8)
ББК 22.172я73
П784

Авторы:
О. С. Логунова, Е. Г. Филиппов, В. В. Павлов, Е. А. Ильина, В. В. Королева

Рецензенты:
декан факультета информатики Магнитогорского университета, канд. пед. наук, проф. *Г. Н. Чусавитина*;
начальник лаборатории экономических исследований управления экономики ОАО «ММК-МЕТИЗ», канд. экон. наук *В. П. Обломец*

Программные статистические комплексы : учеб. пособие
П784 для студ. высш. учеб. заведений / О.С.Логунова, Е.Г.Филиппов, В.В.Павлов и др. — М. : Издательский центр «Академия», 2011. — 240 с.
ISBN 978-5-7695-6297-6

Изложен теоретический и практический материал для проведения занятий по дисциплине «Программные статистические комплексы. Теория и практика». Данное учебное пособие способствует наилучшему освоению студентами практики использования новых методов обработки данных, а также решению задач с использованием универсального пакета STATISTICA 6.0.

Для студентов высших учебных заведений.

УДК 83.77.31(075.8)
ББК 22.172я73

ISBN 978-5-7695-6297-6

ВВЕДЕНИЕ

За последние годы благодаря целенаправленной государственной политике произошел заметный скачок в качестве информации, доступной для нужд науки и образовании. Анализ процессов показал, что существует возможность сбора чрезвычайно больших объемов статистической информации разного рода. Для обработки этой информации необходимы сокращение размерности получаемых данных и построение моделей функционирования за счет использования статистических методов, например методов главных компонент и регрессионного анализа. Это, в свою очередь, требует разработки специализированного программного пакета для прикладных статистических исследований, учитывающих специфику структуры данных и задач.

Термин «статистика» произошел от латинского слова «статус» (status), что означает «определенное положение вещей». Употреблялся он первоначально в значении слова «государствоведение»; впервые был введен в обиход в 1749 г. немецким ученым Г. Ахенвалем, выпустившим книгу о государствоведении.

В настоящее время термин «статистика» употребляется в трех значениях. Во-первых, под статистикой понимают особую отрасль практической деятельности людей, направленную на сбор, обработку и анализ данных, характеризующих социально-экономическое развитие страны, ее регионов, отраслей экономики, отдельных предприятий. Во-вторых, статистикой называют науку, занимающуюся разработкой теоретических положений и методов, используемых статистической практикой. Между статистической наукой и статистической практикой существует тесная связь. Статистическая практика применяет правила, выработанные наукой; в свою очередь, статистическая наука опирается на материалы практики и, обобщая опыт практики, разрабатывает новые положения. В-третьих, статистикой часто называют статистические данные, представленные в отчетности предприятий, организаций, отраслей экономики, а также публикуемые в сборниках, справочниках, периодической прессе, которые представляют собой результат статистической работы.

Особенность статистики заключается в том, что статистические данные сообщаются в количественной форме, т. е. статистика говорит языком цифр, отображающих общественную жизнь во всем многообразии ее проявлений. При этом статистику прежде всего интересуют те выводы, которые можно сделать на основе анализа надлежащим образом собранных и обработанных цифровых данных.

Статистика имеет многовековую историю, уходя своими корнями в глубокую древность. С образованием государств появилась необходимость в статистической практике, т. е. в сборе сведений о наличии земель, численности населения, о его имущественном положении. Несколько тысячелетий назад такой учет проводился в Китае, Древнем Риме и Египте.

На Руси уже в X—XII вв. собирались различного рода сведения, связанные с налогообложением. Петровские реформы, затронувшие все стороны общественной жизни, потребовали значительно большего числа точных статистических данных: вводится учет цен на хлеб, городов и городского населения, внешней торговли, осуществляется регистрация новых фабрик и заводов. В этот же период зарождается текущий учет численности населения — проводимая церковью регистрация браков, рождений, смертей. По мере усложнения общественной жизни все более расширялся круг учитываемых явлений.

В период становления капитализма рост общественного производства, расширение торговых и международных отношений послужили стимулом развития учета и статистики. Наряду с простой бухгалтерией в Италии (приблизительно с начала XIV в.) появляется система двойной бухгалтерии, при которой операция фиксируется дважды — в дебете и в кредите. Значительно возрастает потребность в анализе экономической конъюнктуры, поэтому объем статистической информации особенно резко увеличивается; требуются сведения о размерах и размещении промышленного и сельскохозяйственного производства, рынках сбыта товаров, рынках труда, сырьевых ресурсов и т. д.

Расширение практики учетно-статистических работ в различных странах способствовало формированию статистики. Как наука она стала развиваться с середины XVII в. по двум направлениям: описательному и математическому.

Важнейшими представителями описательной школы государствоведения были немецкие ученые Г. Конринг (1606—1681) и Г. Ахенваль (1719—1772). Первой отличительной чертой этого направления было то, что задачей статистики его представители считали описание «государственных достопримечательностей». К их числу относили территорию государства, государственное устройство, население, религию, внешнюю политику и т. п. Таким

образом, в этом случае предмет статистики не ограничивался теми явлениями, которые имеют числовую характеристику. Более того, ранние представители описательной школы вообще избегали пользоваться числовыми данными, и лишь позднее (в середине XVIII в.) числовые данные постепенно завоевали право быть включенными в работы описательной статистики. Вторая особенность описательного направления статистики заключалась в том, что в этих работах отсутствовал анализ закономерностей и взаимосвязей, присущих общественным процессам. Следовательно, то, что представители описательного направления называли статистикой, было весьма далеко от действительной статистики в ее современном понимании.

В ходе исторического развития статистической науки в ее составе обособился ряд самостоятельных статистических дисциплин; это объясняется наличием конкретного предмета исследования и особой системы статистических показателей для его характеристики. Тем не менее ряд методов статистики остается единообразным для всех возможных отраслей.

Программные статистические комплексы являются мощным инструментом проведения статистического анализа собранных данных. В настоящее время появился ряд программных продуктов, которые могут быть использованы в учебном процессе при изучении таких дисциплин, как «Статистика», «Эконометрика», «Программные статистические комплексы», «Обработка экспериментальных данных на ЭВМ» и т. п.

Авторы не преследовали цель полностью изложить статистические методы или дать полное описание возможностей программного продукта STATISTICA 6.0. Цель работы — показать возможность использования этого программного обеспечения для реализации стандартных статистических задач, с которыми чаще всего сталкивается студент высшего учебного заведения и инженер, занимающийся обработкой наблюдений в ходе своей профессиональной деятельности.

Авторы благодарят рецензентов за проявленный интерес к рукописи. Учебное пособие подготовлено в рамках Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» 2009 — 2013 гг. по государственному контракту П2402.

Глава 1

СТАТИСТИЧЕСКИЕ ДАННЫЕ: ПОИСК, ДОБЫЧА И ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ

1.1. Понятие о статистическом изучении данных. Возможные подходы к изучению данных

Статистика, изучая самые разнообразные явления, в своих выводах опирается на числовые данные, полученные в конкретных условиях места и времени. Результаты статистического наблюдения регистрируются прежде всего в форме первичных абсолютных величин, отражающих уровень развития данного явления. Основная масса абсолютных показателей фиксируется в первичных учетных документах.

С развитием рыночных отношений роль информационной базы возрастает, поскольку усложняются связи субъектов рынка, возникает все более настоятельная потребность в изучении влияния различных факторов на результаты деятельности, социальные последствия, а также в прогнозировании, обобщениях как на макро-, так и на микроуровне. Важнейшим ресурсом в управлении становится статистическая информация.

Для того чтобы выполнить статистическое исследование, необходима научно обоснованная информационная база. Она формируется в результате статистического наблюдения, которое является начальной стадией экономико-статистического исследования.

Статистическим наблюдением называется планомерный научно обоснованный сбор данных или сведений о социально-экономических явлениях и процессах.

Статистические данные представляют собой составную часть глобальной информационной системы, которая формируется в соответствии с концепцией информатизации, разработанной в Российской Федерации. Информационная база статистики призвана обеспечить поддержку формирующегося рынка, дать всестороннюю и объективную информацию для разработки вариантов, обоснования и принятия управленческих решений. В этих целях специальный статистический аппарат занимается систематическим сбором данных, их обработкой и предоставлением ре-

зультатов в виде статистической информации государственным и другим органам, коммерческим пользователям. Владея информацией, предприятия могут эффективнее решать постановленные задачи.

При работе с некоторым набором наблюдений сталкиваются с понятием «интеллектуальный анализ», или Data Mining. Data Mining переводится как «добыча» или «раскопка данных». Нередко рядом с Data Mining встречаются слова «обнаружение знаний в базах данных» (knowledge discovery in databases) и «интеллектуальный анализ данных». Их можно считать синонимами Data Mining. Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных.

До начала 1990-х гг., казалось, не было особой нужды переосмысливать ситуацию в этой области. Все шло своим чередом в рамках направления, называемого прикладной статистикой. Теоретики проводили конференции и семинары, писали внушительные статьи и монографии, изобиловавшие аналитическими выкладками. Вместе с тем практики всегда знали, что попытки применить теоретические экзерсисы для решения реальных задач в большинстве случаев оказываются бесплодными. Но на озабоченность практиков до поры до времени можно было не обращать особого внимания — они решали главным образом свои частные проблемы обработки небольших локальных баз данных (БД).

В связи с совершенствованием технологий записи и хранения данных на людей обрушились колоссальные потоки информационной руды в самых различных областях. Деятельность любого предприятия (коммерческого, производственного, медицинского, научного и т.д.) теперь сопровождается регистрацией и записью всех подробностей его деятельности. Что делать с этой информацией? Стало ясно, что без продуктивной переработки потоки «сырых» данных образуют никому не нужную свалку.

Специфика современных требований к такой переработке такова:

- данные имеют неограниченный объем;
- данные являются разнородными (количественными, качественными, текстовыми);
- результаты должны быть конкретны и понятны;
- инструменты для обработки «сырых» данных должны быть просты в использовании.

Традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, откровенно спасовала перед лицом возникших проблем. Главная причина — концепция усреднения по выборке, приводящая к операциям над фиктивными величинами (типа средней температуры пациентов по больнице, средней высоты дома на улице,

состоящей из дворцов и лачуг, и т. п.). Методы математической статистики оказались полезными главным образом для проверки заранее сформулированных гипотез (verification-driven data mining) и для «грубого» разведочного анализа, составляющего основу оперативной аналитической обработки данных (on line analytical processing — OLAP).

В основу современной технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей. Примеры заданий на такой поиск при использовании Data Mining приведены в табл. 1.1.

Важное положение Data Mining — нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные (unexpected) регулярности в данных, составляющие так называемые скрытые знания (hidden knowledge). К обществу пришло понимание того, что «сырые» данные (raw data) содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие самородки.

Таблица 1.1

Примеры формулировок задач при использовании методов OLAP и Data Mining

OLAP	Data Mining
Каковы средние показатели травматизма для курящих и некурящих?	Встречаются ли точные шаблоны в описаниях людей, подверженных повышенному травматизму?
Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов (отказавшихся от услуг телефонной компании)?	Имеются ли характерные портреты клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?
Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке?	Существуют ли стереотипные схемы покупок для случаев мошенничества с кредитными карточками?