электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.67

Статистические методы анализа данных

Смирнов Р. М., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Автоматизированные системы обработки информации и управления»

Гарина И.О., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Автоматизированные системы обработки информации и управления»

Научный руководитель: Тоноян С. А., к.т.н, доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
tonoyansl@mail.ru

Статистика как наука

Полная и достоверная статистическая информация является тем необходимым основанием, на котором базируется процесс управления во всех массовых явлениях. Принятие управленческих решений на всех уровнях – от общегосударственного или регионального и до уровня отдельной корпорации или частной фирмы – невозможно без должного статистического обеспечения.

Статистика – это наука, изучающая количественную сторону массовых явлений и процессов в неразрывной связи с их качественной стороной, количественное выражение закономерностей общественного развития в конкретных условиях места и времени.

Для современного уровня статистической науки характерно то, что наряду с развитием статистических и экономико-математических методов анализа социально-экономических явлений все более широко используется компьютерная техника. Это не только значительно расширяет возможности сбора информации различных объемов и ее предварительной обработки, но и совершенствует систему статистического анализа.

Общей методологией изучения статистических совокупностей является использование основных принципов, которыми руководствуются в любой науке. К этим принципам относятся следующие:

1. объективность изучаемых явлений и процессов;
2. выявление взаимосвязи и системности, в которых проявляется содержание изучаемых факторов;

3. достижение поставленных целей со стороны исследователя, изучающего соответствующие статистические данные.

Это выражается в получении сведений о тенденциях, закономерностях и возможных последствиях развития изучаемых процессов. Знание закономерностей развития социально-экономических процессов, интересующих общество, имеет важное практическое значение.

Основные задачи статистического анализа

Главной задачей статистики является получение и обработка статистической информации для принятия решений, направленных на достижение желаемого результата во всех сферах деятельности общества. Статистика призвана способствовать выявлению наиболее острых проблем в экономической и социально-политической сферах, а также обоснованию путей достижения многообразных целей развития общества.

Статистика дает сигналы о неблагополучии в отдельных частях механизма управления, показывая, таким образом необходимость обратной связи - управляющих решений. Общие принципы и методы научного познания служат фундаментом для понимания и правильного использования статистической методологии. Итак, основной задачей статистики является сбор, учет, обработка и хранение данных (информации), отображающих ход общественного развития.

Таким образом, статистика выступает важнейшим инструментом познания и использования экономических и других законов общественного развития.

Собранные в процессе статистического наблюдения сведения подвергаются в дальнейшем сводке (первичной научной обработке), в процессе которой из всей совокупности обследованных единиц выделяются характерные части (группы). Выделение групп и подгрупп единиц из всей обследованной массы называется в статистике группировкой. Группировка в статистике является основой обработки и анализа собранной информации. Осуществляется она на основе определенных принципов и правил. В процессе обработки статистической информации совокупность обследованных единиц и выделенные ее части на основе применения метода группировок характеризуются системой цифровых показателей: абсолютных и средних величин, относительных величин, показателей динамики и т.д.

Основные этапы статистического исследования

Процесс статистического исследования состоит из шести этапов.

Этап 1. Определение проблемы

Первый этап любого статистического исследования заключается в выяснении проблемы. При ее определении исследователь должен принимать во внимание цель исследования, соответствующую исходную информацию, какая информация необходима и как она будет использована при принятии решения. Определение проблемы включает в себя ее обсуждение с лицами, принимающими решения, анализ вторичных данных и, возможно, проведение отдельных качественных исследований. Как только проблема точно установлена, можно разрабатывать план статистического исследования и приступать к его проведению.

Этап 2. Разработка подхода к решению проблемы

Разработка подхода к решению проблемы включает в себя формулировку теоретических рамок исследования, аналитических моделей, поисковых вопросов, гипотез, а также определение факторов, которые могут влиять на план исследования. Этот этап характеризуется следующими действиями: обсуждение с руководством компании-клиента и экспертами по данной сфере, изучение ситуаций и моделирование, анализ вторичных данных, качественные исследования и прагматические соображения.

Этап 3. Разработка плана исследования

План статистического исследования детализирует ход выполнения процедур, необходимых для получения нужной информации. Он необходим для того, чтобы разработать план проверки гипотез, определить возможные ответы на поисковые вопросы и выяснить, какая информация необходима для принятия решения. Проведение поискового исследования, точное определение переменных и определение соответствующих шкал для их измерения — все это тоже входит в план статистического исследования.

Этап 4. Сбор данных

Сбор данных осуществляется персоналом по проведению полевых работ, которые работают либо в полевых условиях, как в случае личного интервьюирования, либо из офиса с помощью телефона, либо по почте, либо с помощью электронных средств. Надлежащий отбор, обучение, контроль и оценка сотрудников минимизирует ошибки при сборе данных.

Этап 5. Подготовка данных и их анализ

Подготовка данных включает в себя редактирование, кодирование, расшифровку и проверку данных. Каждая анкета или форма наблюдения проверяются или редактируются и, если необходимо, корректируются. Для анализа данных используются одномерные методы статистического анализа в том случае, если элементы выборки измеряются по

одному показателю, или когда имеется несколько показателей, но каждая переменная анализируется отдельно. С другой стороны, если имеется два или более измерений каждого элемента выборки, а переменные анализируются одновременно, то для анализа данных используются многомерные методы.

Этап 6. Подготовка отчета и его презентация

Ход и результаты статистических исследований должны быть изложены письменно в виде отчета, в котором четко обозначены конкретные вопросы исследования, описан метод и план исследования, процедуры сбора данных и их анализа, результаты и выводы. Полученные выводы должны быть представлены в виде, удобном для использования при принятии управленческих решений. Кроме того, руководству компании-клиента должна быть сделана и устная презентация с использованием таблиц, цифр и диаграмм, чтобы повысить доходчивость и воздействие на аудиторию.

Основные статистические методы

Корреляционный анализ — статистический метод анализа данных, предназначенный для исследования взаимозависимости выборок. Корреляционный анализ является составной частью любого статистического исследования. Основной параметрический показатель корреляции является ковариация (или корреляционный момент). Для совместного распределения двух случайных ковариация определяется как математическое ожидание произведения отклонений случайных величин:

$$cov_{XY} = M\big[\big(X - M(X)\big)\big(Y - M(Y)\big)\big] = M(XY) - M(X)M(Y)$$

Дисперсионный анализ — статистический метод анализа данных, предназначенный для исследования степени влияния независимых переменных на зависимые. Математическая модель дисперсионного анализа представляет собой частный случай основной линейной модели. Пусть с помощью методов $A_j, (1 \le j \le m)$ производится измерение нескольких параметров $x_i, (1 \le i \le n)$, чьи точные значения — $\mu_i, (1 \le i \le n)$. В таком случае, результаты измерений различных величин различными методами можно представить как:

$$x_{i,j} = \mu_i + a_{i,j} + e_{i,j}$$

где:

$x_{i,j}$ — результат измерения i-го параметра по методу A_j;

μ_i — точное значение i-го параметра;

$a_{i,j}$ — систематическая ошибка измерения i-го параметра в группе по методу A_j;

$e_{i,j}$ — случайная ошибка измерения i-го параметра по методу A_j.

Тогда дисперсии следующих случайных величин:

$$x_{*,j} = \frac{1}{n}\sum_i x_{i,j}$$

$$x_{i,*} = \frac{1}{m}\sum_j x_{i,j}$$

$$x_{*,*} = \frac{1}{nm}\sum_{i,j} x_{i,j}$$

выражаются как:

$$S^2 = \frac{1}{nm}\sum_i\sum_j (x_{i,j} - x_{*,*})^2$$

$${s_0}^2 = \frac{1}{nm}\sum_i\sum_j (x_{i,j} - x_{i,*} - x_{*,j} + x_{*,*})^2$$

$${s_1}^2 = \frac{1}{n}\sum_i (x_{i,*} - x_{*,*})^2$$

$${s_2}^2 = \frac{1}{m}\sum_j (x_{*,j} - x_{*,*})^2$$

и удовлетворяют тождеству:

$$S^2 = {s_0}^2 + {s_1}^2 + {s_2}^2$$

Процедура дисперсионного анализа состоит в определении соотношения систематической (межгрупповой) дисперсии к случайной (внутригрупповой) в измеряемых данных.

В качестве показателя изменчивости используется сумма квадратов отклонения значений параметра от среднего: SS, которая раскладывается на межгрупповую сумму квадратов SS_{BG} и внутригрупповую сумму квадратов SS_{WG}:

$SS = SS_{BG} + SS_{WG}$

Пусть точное значение каждого параметра есть его математическое ожидание, равное среднему генеральной совокупности E(X) = M. При отсутствии систематических ошибок групповое среднее и среднее генеральной совокупности тождественны: $M_j = M$. Тогда случайная ошибка измерения есть разница между результатом измерения $x_{i,j}$ и средним группы: $x_{i,j} - M_j$. Если же метод A_j оказывает систематическое воздействие, то систематическая ошибка при воздействии этого фактора есть разница между средним группы M_j и средним генеральной совокупности: $M_j - M$.

$$SS = \sum_{i=1}^{n_j} (x_{i,j} - M)^2$$

$$SS_{BG} = \sum_{i=1}^{n_j} (M_j - M)^2$$

$$SS_{WG} = \sum_{i=1}^{n_j} (x_{i,j} - M_j)^2$$

Степени свободы:

$$df = df_{BG} + df_{WG}$$

$$df_{BG} = J - 1$$

$$df_{WG} = N - J$$

где N – объём полной выборки, J – количество групп.

Тогда дисперсия каждой части, именуемая в модели дисперсионного анализа как «средний квадрат», или MS, есть отношение суммы квадратов к числу их степеней свободы:

$$MS = \frac{SS}{N - 1}$$

$$MS_{BG} = \frac{SS_{BG}}{J - 1}$$

$$MS_{WG} = \frac{SS_{WG}}{N - J}$$

Соотношение межгрупповой и внутригрупповой дисперсий имеет F – распределение (распределение Фишера) и определяется при помощи F-критерия Фишера:

$$F_{df_{BG}, df_{WG}} = \frac{MS_{BG}}{MS_{WG}}$$

Примеры задач, решаемых с помощью дисперсионного анализа.

- Влияет ли упаковка на уровень объема сбыта?
- Имеет ли влияние выбор каналов сбыта на объем сбыта?

Методы прогнозирования

Методы прогнозирования классифицируются по различным критериям:

- по форме предоставления результата прогнозы делятся на количественные и качественные; первые базируются на численных, математических процедурах, а вторые на использовании имеющихся опыта, знаний и интуиции исследователя;
- по величине периода упреждения выделяют краткосрочные (1 год и менее), среднесрочные (2-5 лет), долгосрочные (свыше 5 лет);

- по охвату прогнозированием объекта исследования прогнозы бывают общими (прогноз общего развития народного хозяйства) и частные (прогноз для отдельных отраслей, инфраструктуры, отдельных показателей).

Классификация методов, используемых при прогнозировании в системах маркетинга и эффективность их применения на практике.

Количественные методы:

- экстраполяция трендов;
- метод скользящей средней;
- регрессионный анализ;
- экспоненциальное сглаживание;
- моделирование.

Качественные методы:

- оценки коммерсантов и технического руководства;
- опрос потребителей;
- тестирование товара;
- методы аналогии;
- результаты тестирования рынка;
- экспертные оценки методом "Дельфи";
- сценарии.

Каждый из возможных методов прогнозирования обладает определенными достоинствами и недостатками. Их применение более эффективно в краткосрочном прогнозировании. Они сильно упрощают реальные процессы, чтобы можно было рассчитывать на получение с их помощью результатов, выходящих за рамки представлений сегодняшнего дня. Практически невозможно отразить в моделях долгосрочного прогнозирования структурные сдвиги, постоянно происходящие в изменяющемся мире.

На самом деле все эти методы являются взаимодополняющими. Эффективная прогнозная система должна обеспечивать возможность использования любого из этих методов.

Примером сложной задачи прогнозирования, которая не решается с помощью какого-то одного метода, является прогнозирование объёма продаж нового товара. При проведении маркетинговых исследований оцениваются объемы продаж нового товара в течение первых лет после выпуска. Для этой цели могут быть применены экспертные методы, методы опросов, проведение продаж на контрольном рынке.

Ясно, что в условиях турбулентной внешней среды интуиция и воображение способны стать важными инструментами восприятия реальности, дополняя количественные подходы, которые, по определению опираются только на наблюдаемые факторы. С другой стороны, понятно, что чисто качественному методу также присущи значительные погрешности и что интуиция должна в возможно большей степени проверяться с помощью доступных фактов и знаний. Таким образом, следую обеспечить совместное использование этих двух подходов.

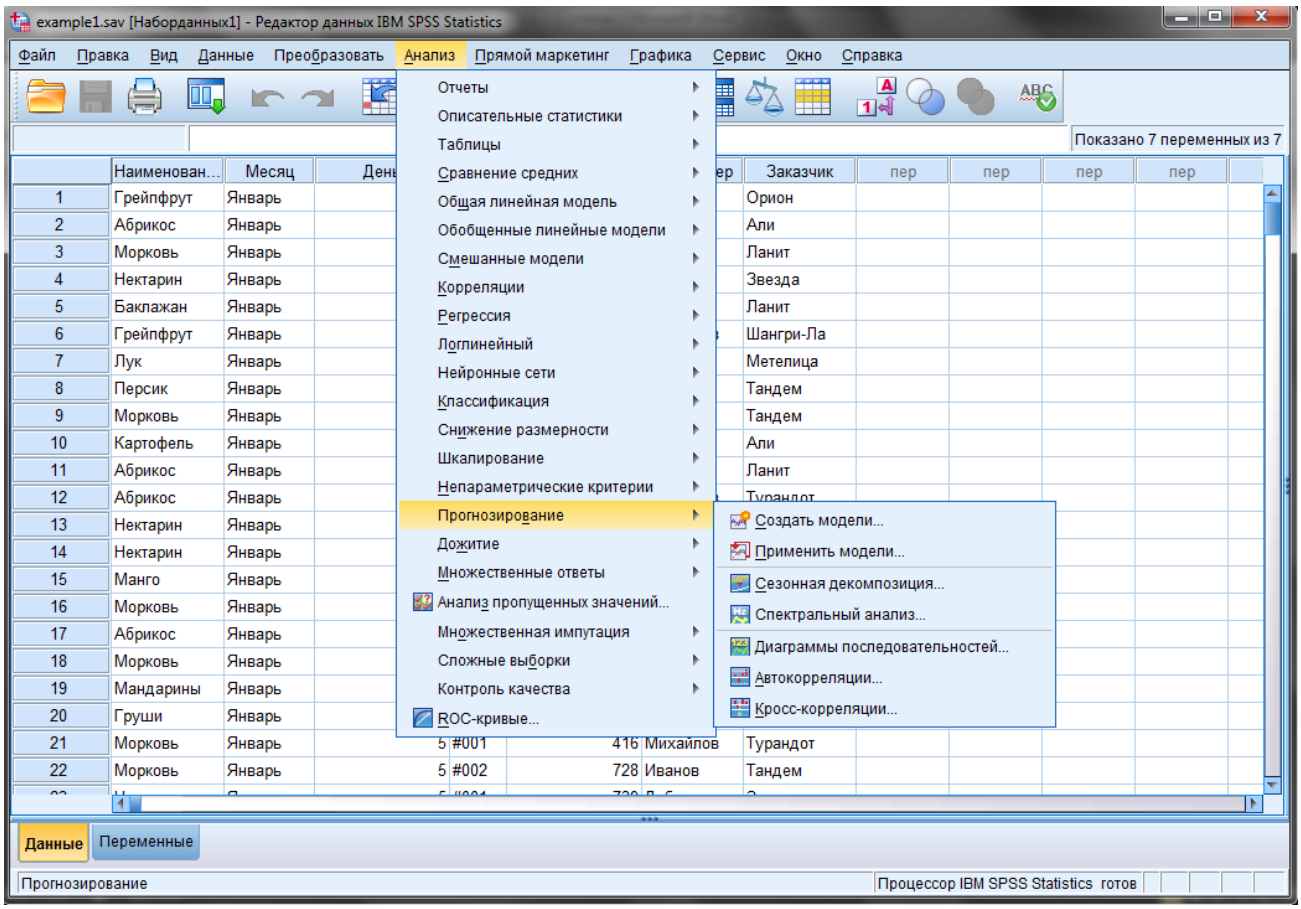
Использование статистических методов на примере программы SPSS Statistics

SPSS (аббревиатура англ. «Statistical Package for the Social Sciences», «статистический пакет для социальных наук») — компьютерная программа для статистической обработки данных, один из лидеров рынка в области коммерческих статистических продуктов, предназначенных для проведения прикладных исследований в социальных науках. По мнению некоторых авторов, SPSS «занимает ведущее положение среди программ, предназначенных для статистической обработки информации».

Продукт SPSS Statistics предоставляет широкие возможности для анализа данных. Интуитивно понятный интерфейс программного обеспечения включает в себя все функции управления данными, статистические процедуры и средства создания отчетов для проведения анализа любой степени сложности.

Пакет включает в себя команды определения данных, преобразования данных, команды выбора объектов. Как видно из рисунка, в SPSS Statistics реализованы следующие методы статистической обработки информации:

- суммарные статистики по отдельным переменным;
- частоты, суммарные статистики и графики для произвольного числа переменных;
- построение N-мерных таблиц сопряженности и получение мер связи;
- средние, стандартные отклонения и суммы по группам;
- дисперсионный анализ и множественные сравнения;
- корреляционный анализ;
- дискриминантный анализ;
- однофакторный дисперсионный анализ и т.д.



Собирая данные, исследователь руководствуется определенными гипотезами. Информация относится к избранным предмету и теме исследования, но нередко она представляет собой сырой материал, в котором необходимо изучить структуру показателей, характеризующих объекты, а также выявить однородные группы объектов. Полезно представить эту информацию в геометрическом пространстве, лаконично отразить ее особенности в классификации объектов и переменных.

Наглядная статистика

Данные опции позволяют получить описательные статистики выборки. Слева направо, начиная со второго столбца: N – объем выборки; минимальное значение; максимальное значение; среднее значение; стандартное отклонение. Первая и вторая ячейки нижней строки указывают на число значений, пригодных для расчетов.

Выборочные таблицы

Здесь предоставляются возможности по снабжению выборки большим набором различных характеристик.

После щелчка на строке («Простые таблицы») соответствующего раскрывающегося списка появится окон, нажатие в котором на кнопку «Статистика…» вызывает на экран второе окно, содержащее большой список статистических понятий

(медиана, мода, средние, среднеквадратическое отклонение, процент или минимальное и максимальное значение и т.д.) из которых пользователь может выбрать нужные.

Средства сравнения

Здесь располагаются критерии для сравнения среднего, различные варианты t-критерия Стьюдента для одной выборки (сравнение среднего значения с неким задаваемым числом); для двух независимых выборок. В результате чего в файле результатов будут даны показатель t-критерия, уровень статистической значимости, стандартная ошибка, значения доверительного интервала и т.д.

Корреляция

Здесь располагаются опции, предназначенные для проведения корреляционного анализа.

Дисперсионный анализ

С помощью дисперсионного анализа исследуют влияние одной или нескольких независимых переменных на одну зависимую переменную (одномерный анализ) или на несколько зависимых переменных (многомерный анализ). В обычном случае независимые переменные принимают только дискретные значения (и относятся к номинальной или порядковой шкале); в этой ситуации также говорят о факторном анализе. Если же независимые переменные принадлежат к интервальной шкале или к шкале отношений, то их называют ковариациями, а соответствующий анализ – ковариационным.

Факторный анализ

Идея метода состоит в сжатии матрицы признаков в матрицу с меньшим числом переменных, сохраняющую почти ту же самую информацию, что и исходная матрица. В основе моделей факторного анализа лежит гипотеза, что наблюдаемые переменные являются косвенными проявлениями небольшого числа скрытых (латентных) факторов. Хотя такую идею можно приписать многим методам анализа данных, обычно под моделью факторного анализа понимают представление исходных переменных в виде линейной комбинации факторов.

Кластерный анализ

Если процедура факторного анализа сжимает в малое число количественных переменных данные, описанные количественными переменными, то кластерный анализ сжимает данные в классификацию объектов.

Если данные понимать как точки в признаковом пространстве, то задача кластерного анализа формулируется как выделение «сгущений точек», разбиение совокупности на однородные подмножества объектов.

Кластерный анализ является описательной процедурой, он не позволяет сделать никаких статистических выводов, но дает возможность провести своеобразную разведку – изучить «структуру совокупности».

Многомерное шкалирование

Задача многомерного шкалирования состоит в построении переменных на основе имеющихся расстояний между объектами. В частности, если нам даны расстояния между городами, программа многомерного шкалирования должна восстановить систему координат (с точностью до поворота и единицы длины) и приписать координаты каждому городу так, чтобы зрительно карта и изображение городов в этой системе координат совпали.

Таблицы сопряженности

В SPSS имеется большое количество разнообразных процедур, при помощи которых можно произвести анализ связи между двумя переменными. Связь между неметрическими переменными, то есть переменными, относящимися к номинальной шкалу или к порядковой шкале с не очень большим количеством категорий, лучше всего представить в форме таблиц сопряженности. Для этой цели в SPSS проверяется, есть ли значимое различие между наблюдаемыми и ожидаемыми частотами. Кроме того, существует возможность расчета различных мер связанности.

Более тщательно исследовать существование зависимости позволяет вычисление значений ожидаемых частот. Еще одну возможность выявления существования зависимости между переменными дает вычисление остатков. Эти остатки являются показателем того, насколько сильно наблюдаемые и ожидаемые частоты отклоняются друг от друга.

SPSS Statistics является самой распространённой программой для обработки статистической информации. Освоение навыков использования аналитических процедур, предлагаемых в SPSS, тем эффективнее, чем четче и адекватнее понимание пользователем существа и специфики исследовательского процесса в той или иной области социальных наук. Такой подход отражает состояние знаний данной предметной области и позволяет лучше понять перспективы ее развития. В то же время навыки работы в SPSS с успехом могут использоваться в самых различных областях знания и народного хозяйства – от маркетинговых исследований до мониторинга здоровья населения.

Мы рассмотрели приемы сбора, обработки и анализа статистических данных, которые являются основой статистического исследования. Во многих случаях аналитик имеет дело с материалом, полученным из баз данных, бюллетеней информационных агентств, статистических сборников и других источников. Следовательно, работа должна

начинаться с проверки полноты и качества данных, их группировки, а при отсутствии необходимости в этих этапах - с расчета индивидуальных и обобщающих показателей. Рассмотренные приемы и методы с успехом могут использоваться не только в практике статистического анализа. Статистическая методология исследования в настоящее время заняла прочные позиции во многих областях знания. Статистические формулы находят применение в макро- и микроэкономике, оценке бизнеса и недвижимости, финансовом анализе, техническом анализе товарных и финансовых рынков. В большинстве случаев, описанные приемы и показатели будут работоспособны и эффективны при обобщении и анализе технической, биологической, медицинской, демографической и социологической информации.

При работе с большими массивами статистической информации необходимо использовать прикладное программное обеспечение, существенно ускоряющее и упрощающее все расчеты. В данной работе мы рассмотрели практическое применение статистических методов на примере программного пакета SPSS Statistics. Среди наиболее распространенных современных программных продуктов также следует отметить пакеты Мезозавр, ОЛИМП, САНИ, Эвриста, STATISTICA, STATGRAPHICS.

Список литературы

1. Ефимова М.Р., Петрова Е.В., Румянцев В.Н. Общая теория статистики: учебник. М.: ИНФРА-М, 1996. 416с.
2. Агалаков С.А. Курс лекций по статистическим методам анализа данных. Режим доступа: http://www.omsu.ru/file.php?id=4948 (дата обращения 20.10.2014).
3. Наследов А.Д. SPSS: Компьютерный анализ данных в психологии и социальных науках. СПб.: Питер, 2005. 247 с.