МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

А.Ю. Гришенцев

ТЕОРИЯ И ПРАКТИКА ТЕХНИЧЕСКОГО И ТЕХНОЛОГИЧЕСКОГО ЭКСПЕРИМЕНТА

Учебное пособие

Санкт-Петербург

2010

УДК 519.2: 519.6

Гришенцев А.Ю. Теория и практика технического и технологического эксперимента /учебное пособие.– СПб: СПбГУ ИТМО, 2010.–102 с.

В пособии рассмотрены основные понятия теории вероятностей и математической статистики применительно к курсу теория и практика технического и технологического эксперимента.

Пособие адресовано студентам высших учебных заведений обучающихся в соответствии с требованиями ОС ВПО по направлению подготовки 211000 «Конструирование и технология электронных средств», специализации магистра 211000.68 – «Технология и инструментальные средства проектирования электронных систем»

СПбГУ ИТМО стал победителем конкурса инновационных образовательных программ вузов России на 2007-2008 годы и успешно реализовал инновационную образовательную программу «Инновационная система подготовки специалистов нового поколения в области информационных и оптических технологий», что позволило выйти на качественно новый уровень подготовки выпускников и удовлетворять возрастающий спрос на специалистов в информационной, оптической и других высокотехнологичных отраслях науки. Реализация этой программы создала основу формирования программы дальнейшего развития вуза до 2015 года, включая внедрение современной модели образования.

ОГЛАВЛЕНИЕ

1. ВВЕДЕНИЕ

Учебное пособие содержит значительный объем теоретической и практической информации. При разработке пособия широко использовалась хорошо зарекомендовавшие себя работы: ставшая уже фундаментальной в области машинной обработки данных Вукулов Э. А. «Основы статистического анализа» [3] и классический труд Большакова В. Д. «Теория ошибок наблюдений» [2].

1.1. Область определений

В начале изучения курса будет полезно ознакомится с рядом определений, которые являются ключевыми для понимания рассматриваемой области знаний.

Эксперимент (*от лат. experimentum* — проба, опыт), метод познания, при помощи которого в *контролируемых* и *управляемых* условиях исследуются явления действительности. Отличаясь от наблюдения активным оперированием изучаемым объектом, эксперимент осуществляется на основе теории, определяющей постановку задач и интерпретацию его результатов. Нередко главной задачей эксперимента служит проверка гипотез и предсказаний теории, имеющих принципиальное значение (так называемый решающий эксперимент). В связи с этим эксперимент, как одна из форм практики, выполняет функцию критерия истинности научного познания в целом [4].

Современная наука использует разнообразные виды эксперимента В сфере фундаментальных исследований простейший тип эксперимента — качественный эксперимент, имеющий целью установить наличие или отсутствие предполагаемого теорией явления. Более сложен измерительный эксперимент, выявляющий количественную определённость какого-либо свойства объекта. Ещё один тип эксперимента, находящий широкое применение в фундаментальных исследованиях,— так называемый мысленный эксперимент. Относясь к области теоретического знания, он представляет собой систему мысленных, практически не осуществимых процедур, проводимых над идеальными объектами. Будучи теоретическими моделями реальных эксперимент, ситуаций, мысленные эксперименты проводятся в целях выяснения согласованности основных принципов теории. В области прикладных исследований применяются все указанные виды эксперимента. Их задача — проверка конкретных теоретических моделей. Для прикладных наук специфичен модельный эксперимент, который ставится на материальных моделях, воспроизводящих существ, черты исследуемой природной ситуации или технического устройства. Он тесно связан с производственным экспериментом. Для обработки результатов

эксперимента применяются методы математической статистики, специальная отрасль которой исследует принципы анализа и планирования эксперимента.

Измерение (*англ. measurement*) в простейшем случае есть процесс сравнения определяемой физической величины с другой однородной ей величиной, значение которой известно. В результате измерения получают численное значение, показывающее, во сколько раз определяемая физическая величина больше или меньше величины с которой её сравнивали, и в конечном итоге больше или меньше величины, принятой за меру измерения.

Наблюдение – регистрация различных факторов естественного или искусственного происхождения. Наблюдение является основой всех научных исследований; признаки наблюдаемого объекта, выявляемые при наблюдении, могут быть как качественными, так и количественными. Количественные признаки выявляются, как правило, путём измерения. Наблюдение можно рассматривать как совокупность измерений, произведённых над наблюдаемым объектом в один момент времени. В большинстве случаев под *наблюдением* и *измерением* понимают один и тот же процесс; под наблюдением следует понимать сложный вид измерений.

Планирование эксперимента, (*experimental design techniques*) — математико-статистическая дисциплина, изучающая методы рациональной организации экспериментальных исследований — от оптимального выбора исследуемых факторов и определения собственно плана эксперимента в соответствии с его целью до методов анализа результатов.

Технический эксперимент, натурный эксперимент, связанный с исследованием качественных и (или) количественных характеристик некоторого технического устройства.

Технологический эксперимент, эксперимент, в большинстве случаев основной целью которого является разработка новой или улучшение имеющейся технологии т.е. технологическая эволюция.

1.2. Неизбежность погрешностей

Результаты всех измерений, как бы тщательно и на каком бы научном уровне они ни выполнялись, подвержены некоторым погрешностям. *Теория ошибок* – наука, занимающаяся изучением и оценкой погрешностей; эти две её функции позволяют исследователю определить, насколько велики погрешности в его измерениях, и помогают уменьшить их, когда это необходимо [1]. Поскольку в основе любой науки и её применений лежат измерения, исключительно важно уметь рассчитывать ошибки и сводить их к минимуму.

В науку слово *ошибка* не имеет значение чего-то неправильного. Ошибка в научном измерении означает неизбежную погрешность, которая сопутствует всем измерениям. Ошибки как таковые нельзя отнести к промахам экспериментатора; ошибок нельзя избежать стараясь быть очень внимательным. Лучшее на, что можно рассчитывать – это свести ошибки к возможному минимуму и надёжно рассчитать их величины.

Чтобы показать неизбежность появления ошибок, мы должны лишь тщательно проанализировать любое измерение.

Рассмотрим хорошо известный пример из работы (Дж. Тейлора [1]).

Чтобы установить дверь плотник, должен измерить высоту дверного проема. Делая прикидку, он мог бы просто взглянуть на дверной проем и оценить его высоту в 210 см. Это грубое «измерение» определенно содержит погрешность. При необходимости плотник мог бы учесть эту погрешность, допуская, что высота может, быть и меньше (205 см), и больше (215 см).

Если бы он захотел произвести более строгое измерение, он мог бы использовать рулетку и определить, что высота равна 211,3 см. Это измерение определенно является более точным, чем его первоначальная прикидка, но и оно, очевидно, содержит некоторую погрешность, поскольку невероятно, чтобы он мог знать, что высота равна точно 211,3000 см, а не, например, 211,3001 см.

Имеется много причин, влияющих на эту остающуюся погрешность. Часть из них мы будем рассматривать. Некоторые из источников ошибок можно было бы устранить, если бы плотник проявил больше внимания к измерению. Например, одним из источников ошибки могло служить плохое освещение, затрудняющее считывание с рулетки. Эту причину ошибки можно было бы устранить, улучшив освещение.

С другой стороны, некоторые из источников ошибки присущи самому процессу измерения и никогда не могут быть полностью устранены. Например, предположим, что рулетка плотника проградуирована полусантиметровыми делениями. Верх дверного проема, по всей вероятности, не совпадает точно ни с одним из полусантиметровых делений. В этом случае плотник должен оценить положение верха проема между двумя делениями. Если же верх проема совпал с одним из делений, то, учитывая, что само деление имеет ширину порядка миллиметра, он должен оценить положение верха в пределах деления. В любом случае плотник должен, в конечном счете, оценить, где лежит верх дверного проема относительно делений на его рулетке, и это приводит к некоторой ошибке в его отсчете.

Купив другую рулетку с чаще расположенными и более тонкими делениями, плотник может уменьшить ошибку, но не может ее полностью

устранить. Если бы он преисполнялся решимости определить высоту проема с наилучшей точностью, допускаемой современным техническим уровнем, он мог бы купить дорогой лазерный интерферометр. Но даже точность интерферометра ограничена величиной порядка длины волны света (около $0{,}5 \cdot 10^{-6}$ м). Хотя теперь плотник был бы в состоянии проводить измерения с фантастической точностью, ему все же не удалось бы точно определить высоту дверного проема.

Более того, стремясь достигнуть все более высокой точности, наш плотник столкнется с важной и принципиальной проблемой. Он определенно обнаружит, что высота в разных местах различна. Даже в одном и том же месте он найдет, что высота изменяется, если меняются температура и влажность или даже если он случайно сотрет тонкий слой пыли. Другими словами, он обнаружит, что нет такой величины, как высота дверного проема. Такого рода проблема называется *проблемой определения* (высота дверного проема не является точно определяемой количественной характеристикой). Она играет важную роль во многих научных измерениях.

Опыты нашего плотника иллюстрируют известную истину. Ни одну физическую величину (длину, время, температуру и т. д.) нельзя измерить с полной определенностью. Ценой особых усилий мы можем свести ошибки до очень малых значений, но исключить их полностью невозможно.

В повседневных измерениях мы обычно не затрудняем себя обсуждением ошибок. Иногда ошибки просто не имеют значения. Если мы говорим, что расстояние между домом и университетом равно 3 км, то (в большинстве случаев) не важно, значит ли это, что оно лежит «между 2,5 и 3,5 км» или «между 2,99 и 3,01 км». Часто ошибки важны, но их нельзя оценить интуитивно без точного анализа.

Теория вероятностей и математическая статистика относятся к фундаментальным областям математики и содержат в себе достаточно объемный кластер человеческих знаний. Размеры данной работы относительно небольшие, поэтому материал изложен достаточно плотно, некоторые первые главы содержат сведенья, концептуальная основа которых изложена в последующих разделах.

1.3. Несколько слов о пакете STATISTICA

STATISTICA – программный пакет для всестороннего статистического анализа, разработанный компанией StatSoft [7]. В пакете STATISTICA (рис. 1.3.1) реализованы процедуры для анализа данных(data analysis), управления данными (data management), добычи данных (data mining), визуализации данных (data visualization).

Пакет STATISTICA имеет модульную структуру. Каждый модуль содержит уникальные процедуры и методы анализа данных:

– **Base** – базовый, включает в себя обширный выбор основных статистик, широкий набор методов для разведочного анализа;

– **Advanced Linear/Non-Linear Models** – развитые линейные и нелинейные модели, предлагает широкий спектр линейных и нелинейных средств моделирования, регрессионный анализ, анализ компонент дисперсий, анализ временных рядов и т. д;

– **Multivariate Exploratory Techniques** – многомерные разведочные технологии, анализа STATISTICA предоставляет широкий выбор разведочных технологий, начиная с кластерного анализа до расширенных методов классификационных деревьев, в сочетании с бесчисленным набором средств интерактивной визуализации для построения связей и шаблонов;

– **QC** (quality control) – контроль качества, предоставляет широкий спектр аналитических методов управления качеством, а также контрольные карты презентационного качества, непревзойденной гибкости и разнообразия;

– **Neural Networks** – нейронные сети, единственный в мире программный продукт для нейросетевых исследований, полностью переведенный на русский язык, реализовано в виде отдельного модуля;

– **Data Miner** – добыча данных, интеллектуальный анализ данных.

STATISTICA – является достаточно популярным исследовательским программным приложением для OS Windows. STATISTICA позволяет существенно расширить свою функциональность благодаря применению технологии позволяющей пользоваелю создавать свои макро программы с применением функцианальности STATISTICA.

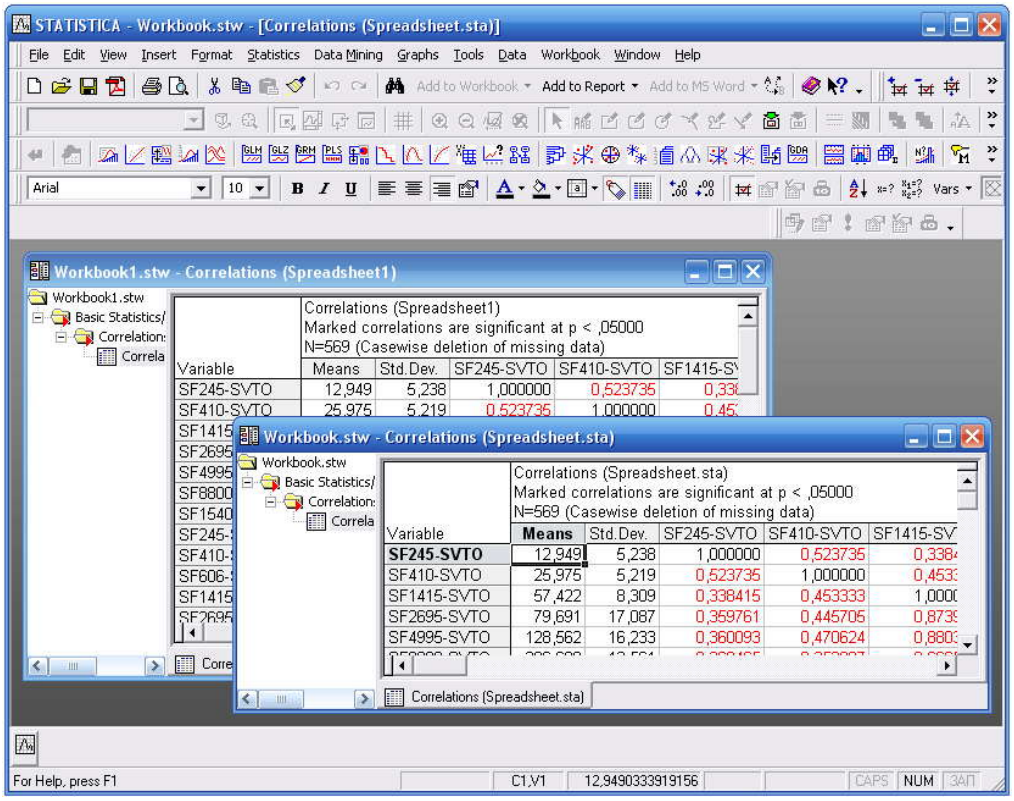

Рисунок 1.3.1. Общий вид программы STATISTICA.

2. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ ЭЛЕМЕНТАРНОЙ ТЕОРИИ ВЕРОЯТНОСТЕЙ

2.1. Виды событий

Осуществление каждого отдельного наблюдения, опыта или измерения при воспроизведении комплекса условий будем называть испытанием. Результат испытания называют *событием* (иногда *исходом*). Понятие события является одним из основных понятий теории вероятностей.

Примеры.

а) При подбрасывании, монеты могут происходить события: «появление герба», «появление цифры».

б) При измерениях трех углов в плоском треугольнике сумма измеренных углов вследствие несовершенства органов чувств наблюдателя, наличия приборных ошибок, влияния внешних условий и других причин будет отличаться от 180° в сторону увеличения либо уменьшения, либо будет случайно совпадать с теоретической суммой. Это соответствует следующим событиям: «появление положительной ошибки (невязки) в сумме измеренных углов», «появление отрицательной ошибки в сумме измеренных углов», «появление в сумме измеренных углов ошибки, равной нулю».

Элементарным (простым) *событием* назовем такой результат испытания, который полностью описывается одним (и только одним) событием. Элементарное событие не может быть разделено на составляющие события.

Сложным событием назовем событие, которое составляется из двух или более элементарных событий [2].

Обозначим через *A, B, C, D, . . . , W* различные события; в последующем будем писать: «событие *A*», «событие *B*» и т. д.

Событие, которое при воспроизведении одного и того же опыта (при осуществлении данного комплекса условий) может наступить, а может и не наступить, называют *случайным событием*.

Соответствует этому явление, которое при многократном воспроизведении одного и того же опыта протекает каждый раз несколько по-иному в зависимости от некоторой изменчивости условий, в которых производится опыт, называют *случайным явлением*.

Например, результатами стрельбы из орудия при одной и той же установке являются случайные события: «попадание в цель», «недолет», «перелет», «отклонение влево», «отклонение вправо» и т. д. Эти результаты обусловлены тем, что, хотя стрельба производится и при одной установке орудия, неучтенные факторы (отклонение массы снаряда и заряда от стандарта, точность наведения, точность введения поправок и т. д.) оказывают свое влияние и опыт протекает каждый раз несколько по-иному.

Как показывает практика, при большом числе испытаний, производимых в одинаковых условиях, обнаруживаются вполне устойчивые закономерности, что является основополагающим при применении методов теории вероятностей и математической статистики к обработке массовых наблюдений. Под одинаковыми условиями здесь и далее будем понимать условия, которые характеризуются примерно равными показателями основных действующих факторов (температуры, давления, влажности, силы и направления ветра и т. д.).

В свою очередь случайные события подразделяют на совместные, несовместные, единственно возможные, равновозможные.

Несколько событий называются *совместными*, если при выполнении комплекса условий при испытании они могут наступить одновременно. Если *A, B, C, ... , W* — совместные события и известно, что они наверняка произойдут, то пишут $A\,B\,C\,\ldots\,W = U$.

События *несовместны*, если при одном испытании они не могут произойти одновременно.

Примеры.

в) В урне имеются белые и черные шары. При одном испытании (взятии) событие A — появление белого шара — и событие B — появление черного шара — несовместные события.

б) Производится один выстрел из орудия. События: «разрыв снаряда» и «неразрыв снаряда» — несовместные события.

Единственно возможными называются события, если в результате испытания появление одного и только одного из них является достоверным событием. Эти события попарно несовместны.

События *равновозможны*, если ни одно из них не является объективно возможным больше, чем любое другое, например выпадение герба или цифры при бросании монеты.

2.2. Полная группа событий

События образуют *полную группу событий*, если при опыте одно из них обязательно должно совершиться.

Два несовместных события составляющих полную группу событий, называются *противоположными*.

Событие, противоположное событию A, обозначается $\overline{A}$ или «не А».

2.3. Относительная частота и вероятность событий

Относительной частотой (частостью) некоторого события называется отношение числа появлений этого события к числу всех произведенных испытаний при выполнении определенного комплекса условий.

Пусть Q — относительная частота, M — число появлений события, N — число всех произведенных испытаний. В соответствии с определением относительной частоты получаем:

$$Q = \frac{M}{N} \qquad (2.3.1)$$

Пример. Произведено 100 измерений одной и той же величины, при этом число отрицательных ошибок оказалось равным 45. Следовательно, M = 45, N = 400, относительная частота появления отрицательной ошибки Q = 0,40.

На основании определения относительной частоты легко установить, что она не может быть отрицательной и что ее значение заключено между нулем и единицей, т. е.:

$$0 < Q < 1 \qquad (2.3.2)$$

При большом числе испытаний, производимых в одинаковых условиях, обнаруживаются вполне устойчивые закономерности. Самым ярким проявлением подобных закономерностей является свойство устойчивости относительной частоты случайных событий, т. е.

уменьшение разброса значений частоты, получаемых в разных сериях испытаний, при увеличении числа испытаний в каждой серии. Поэтому, выполнив достаточно большую серию испытаний, можно с высокой точностью предсказать результат других таких же серий испытаний.

Так, английский ученый *К. Пирсон,* определяя относительную частоту появления герба при подбрасывании монеты 12 000 и 24 000 раз, получил значения этой частоты соответственно 0,5016 и 0,5005. Нетрудно для данного опыта установить, пользуясь обычными представлениями, что относительные частоты появления герба или цифры должны быть близки одна к другой, а их «точное» значение, около которого колеблются опытные данные, равно 0,5.

Итак, при большом числе испытаний N относительная частота Q обнаруживает устойчивость, которая характеризует объективную связь между комплексом условий, в которых производится опыт, и событием. Так как при увеличении числа испытаний N в сериях колебания относительной частоты Q уменьшаются, то есть основания предполагать, что существует некоторое постоянное определенное значение относительной частоты, от которого она отклоняется в ту и другую сторону. Этой постоянной величиной является количественная мера степени объективной возможности появления события при одном опыте, называемая вероятностью события.

Обозначим через $p(A)$ вероятность события A. В дальнейшем для упрощения будем обозначать вероятность просто буквой p.

В теории вероятностей часто рассматривают несовместные, равновозможные события, образующие полную группу. Такие события называют случаями (или шансами). Тогда, как принято говорить, опыт «сводится к схеме случаев» и вероятность может быть вычислена непосредственно как отношение числа благоприятствующих случаев к общему числу всех возможных случаев (классическое определение вероятности), по формуле:

$$p = \frac{m}{n},\ m \le n, \qquad (2.3.3)$$

где: m — число случаев, благоприятствующих некоторому событию; n — число всех возможных случаев; p — вероятность события.

Случай называется благоприятствующим некоторому событию, если появление этого случая влечет за собой появление данного события.

Свойство устойчивости относительной частоты *события* явилось предметом исследования многих ученых. Первым выразил эту закономерность в виде **теоремы *Бернулли***: при числе испытаний неограниченно большом с вероятностью, сколь угодно близкой к единице, относительная частота события сколь

угодно мало отличается от его вероятности в отдельном опыте. (Вероятность p в отдельном опыте постоянна).

Это положение — простейшая формулировка закона больших чисел. Математически теорема Бернулли в принятых обозначениях может быть выражена формулой:

$$\underset{N\to\infty}{p}\left\{\left|\frac{M}{N}-p\right|<\varepsilon\right\}>1-\delta\,, \qquad (2.3.4)$$

где ε и δ — сколь угодно малые положительные числа.

Формула (2.3.4) лежит в основе эмпирического определения вероятности в тех случаях, когда нет возможности применить формулу (2.3.3). В этих случаях вероятность принимают приближенно равной относительной частоте, полученной из достаточно большого числа испытаний. Вероятность, полученную таким путем, называют *статистической*. Заметим, кстати, что на практике это бывает как правило.

Из формулы (2.3.3) следует, что чем ближе вероятность к единице, тем чаще происходит событие, и чем ближе вероятность к нулю, тем событие происходит реже.

Если вероятность события сколь угодно близка к единице, его называют *практически достоверным*, если близка к нулю – *практически невозможным*. Степень приближения p к 1 или 0 оценивается, исходя из практических соображений.

Иными словами, нельзя заранее указать абсолютную величину вероятности, при которой можно было бы событие считать или практически достоверным, или практически невозможным, весь вопрос состоит в том, о каком конкретном событии идет речь.

Пример. Если при стрельбе из артиллерийских орудий из 1001 снарядов 1000 разорвутся при падении и один снаряд не разорвется, вероятность события «неразрыв снаряда», равную ≈0,001, можно уверенно считать величиной пренебрежимо малой, а указанное событие — практически невозможным. Но если 0,001 есть вероятность события «нераскрытие парашюта после прыжка», то вряд ли мы так же легко отнесем это событие к практически невозможным.

Из изложенного следует, что для каждого испытания, исходя из практических соображений, необходимо установить как бы допустимую величину отклонения вероятности от единицы или нуля для того, чтобы можно было считать событие практически достоверным или, наоборот, практически невозможным.

Рассмотрим следствия, которые вытекают из определения вероятности.

Вероятность невозможного события равна нулю, т. е.

$$p(V) = 0\,, \tag{2.3.5}$$

где V – невозможное событие.

Вероятность достоверного события равна единице, т. е.

$$p(U) = 0\,, \tag{2.3.6}$$

где U – достоверное событие.

Вероятность любого события, так же как и его относительная частота, всегда рациональная правильная дробь, т. е.

$$\begin{cases} 0 \le p \le 1 \\ 0 \le m \le n\,. \\ p = \dfrac{m}{m} \end{cases} \tag{2.3.7}$$

Вероятность случайного события, как следует из теоремы Бернулли, всегда отвечает условию

$$0 < p < 1\,. \tag{2.3.8}$$

Сумма вероятностей событий полной группы равна единице, поскольку одно из событий полной группы обязательно должно совершиться при испытании.

Сумма вероятностей противоположных событий всегда равна единице, т. е.

$$p(\overline{A}) + p(A) = 1\,, \tag{2.3.9}$$

это следует из того, что противоположные события составляют полную группу. Кроме того:

$$\begin{cases} p(A) = \dfrac{m}{n} \\ p(\overline{A}) = \dfrac{n-m}{n} \end{cases}. \tag{2.3.10}$$

2.4. Сложение вероятностей

Суммой нескольких несовместных событий называется сложное событие, состоящее в появлении хотя бы одного из этих событий.

Пусть даны несовместные события $A_1, A_2, \ldots A_n$. Если событие B – сложное событие, состоящее в появлении хотя бы одного из этих несовместных событий, то:

$$B = (\vee A_1, \vee A_2, \ldots \vee A_n)\,. \tag{2.4.1}$$

Более удобной записью выражения (2.4.1) могут служить следующие выражения:

$$\begin{cases} B = A_1 + A_2 + ... + A_n \\ \quad\quad ИЛИ \\ B = \sum_{i=1}^{n} A_i \end{cases} . \qquad (2.4.2)$$

Теорема сложения вероятностей. Вероятность суммы нескольких несовместных событий равна сумме вероятностей этих событий, т. е.

$$\begin{cases} p(B) = p(A_1 + A_2 + ... + A_n) = p(A_1) + p(A_2) + ... + p(A_n) \\ \quad\quad ИЛИ \\ p\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} p(A_i) \end{cases} . \qquad (2.4.3)$$

Докажем данную теорему, используя принцип перехода от частного к общему (т.е. принцип индукции).

Доказательство. Пусть в ящике находятся a – белых, b – желтых, c – оранжевых, d – черных и e – коричневых шаров одинакового размера и массы. Определить вероятность того, что шар, извлеченный наудачу при одном взятии, будет иметь светлый тон (т. е. будет белого, желтого или оранжевого цвета).

Обозначим:

В – событие, состоящее в появлении шара светлого тона;

A_1 – событие, состоящее в появлении шара белого цвета;

A_2 – событие, состоящее в появлении шара желтого цвета;

A_3 – событие, состоящее в появлении шара оранжевого цвета;

A_4 – событие, состоящее в появлении шара черного цвета;

A_5 – событие, состоящее в появлении шара коричневого цвета.

По условию можно записать: $B = A_1 + A_2 + A_3$.

Но в то же время, используя формулу (2.3.3) для непосредственного подсчета вероятностей, получим:

$$p(B) = \frac{m}{n} = \frac{a+b+c}{a+b+c+d+e} = \frac{a}{n} + \frac{b}{n} + \frac{c}{n}$$

где

$$\begin{cases} p(A_1) = \dfrac{a}{n} \\ p(A_2) = \dfrac{b}{n}, \\ p(A_3) = \dfrac{b}{n} \end{cases}$$

а следовательно:

$$p(B) = p(A_1) + p(A_2) + p(A_3)$$

что и требовалось доказать.

Распространяя полученный частный вывод на общий случай, можем считать формулу (2.4.3) доказанной.

2.5. Независимые и зависимые события, условная вероятность

Выводы, основанные на положениях теории вероятностей и касающиеся сложных событий, будут существенно различными в зависимости от характера связи между элементарными событиями. Поэтому представляется целесообразным дать определения независимых и зависимых событий и понятие об условной вероятности.

Два события называются *независимыми*, если вероятность появления любого из них не зависит от того, появилось или не появилось другое событие.

Несколько событий называются *попарно независимыми*, если каждые два из них независимы.

Несколько событий называются *независимыми в совокупности*, если каждое из них и любое сложное событие (составленное из всех остальных или части их) – события независимые.

Независимость в совокупности и попарная независимость — не одно и то же.

Примеры.

а) Отсчеты по шкале прибора при одной установке разными наблюдателями.

б) Результаты измерения одной и той же величины в разное время.

в) Результаты стрельбы по одной мишени разными стрелками.

Два или несколько событий называются *зависимыми*, если вероятность появления хотя бы одного из них зависит от того, появляются другие события или не появляются.

Пример.

Если поражение цели достигается двумя попаданиями, то поражение цели при втором выстреле есть событие зависимое, так как оно может совершаться лишь при условии первого попадания в цель.

В связи с тем, что наряду с независимыми событиями при испытаниях приходится иметь дело и с зависимыми событиями, возникает вопрос о так называемой условной вероятности.

Вероятность, вычисленная в предположении, что одно или несколько событий уже произошло, называется *условной вероятностью*.

Примем обозначения:

$p(A/B)$ – условная вероятность события A, вычисленная в предположении, что произошло событие B;

$p(A/B_1, B_2, ...B_n)$ – условная вероятность события Л, вычисленная в предположении, что произошли события $B_1, B_2, ...B_n$.

События A и B зависимы, если выполняются неравенства:

$$\begin{cases} p(A/B) \neq p(A) \\ p(B/A) \neq p(B) \end{cases}. \qquad (2.5.1)$$

В отличие от условной вероятности вероятность независимого события называют иногда безусловной вероятностью (когда эти понятия используются вместе). События A и B независимы, если выполняются условия:

$$\begin{cases} p(A/B) = p(A) \\ p(B/A) = p(B) \end{cases}. \qquad (2.5.1)$$

В приведенном выше примере условная вероятность поражения цели при втором выстреле равна вероятности попадания в случае, если первый выстрел сопровождался попаданием в цель, и равна нулю в случае промаха при первом выстреле.

Независимыми переменными называются переменные, которые варьируются исследователем, тогда как *зависимые переменные* - это переменные, которые измеряются или регистрируются. Может показаться, что проведение этого различия создает путаницу в терминологии, поскольку как говорят некоторые студенты "все переменные зависят от чего-нибудь". Тем не менее, однажды отчетливо проведя это различие, вы поймете его необходимость. Термины зависимая и независимая переменная применяются в основном в экспериментальном исследовании, где экспериментатор манипулирует некоторыми переменными, и в этом смысле они "независимы" от реакций, свойств, намерений и т.д. присущих объектам исследования. Некоторые другие переменные, как предполагается, должны "зависеть" от действий экспериментатора или от экспериментальных условий. Иными словами, зависимость проявляется в ответной реакции исследуемого объекта на посланное на него воздействие. Отчасти в противоречии с данным разграничением понятий находится использование их в исследованиях, где вы не варьируете независимые переменные, а только приписываете объекты к "экспериментальным

группам", основываясь на некоторых их априорных свойствах. Например, если в эксперименте мужчины сравниваются с женщинами относительно числа лейкоцитов (WCC), содержащихся в крови, то Пол можно назвать независимой переменной, а WCC зависимой переменной.

2.6. Умножение вероятностей

Одним из примеров сложного события является произведение событий.

Произведением двух или нескольких событий называется сложное событие, состоящее в совместном появлении всех этих событий.

Пример.

Произведено три выстрела по цели. Пусть событие B – попадание при первом выстреле, событие C – попадание при втором выстреле, событие D – попадание при третьем выстреле. Тогда сложное событие A – попадание всеми тремя выстрелами есть: $A = B \cdot C \cdot D$.

Таким образом, если В — сложное событие, состоящее в совместном появлении событий $A_1, A_2, \ldots A_n$, то произведение событий будет выражено:

$$\begin{cases} B = (\wedge A_1, \wedge A_2, \ldots \wedge A_n) \\ \quad ИЛИ \\ B = (A_1 \cdot A_2 \cdot \ldots \cdot A_n) \end{cases}. \qquad (2.6.1)$$

Теорема умножения вероятностей. Вероятность произведения двух или нескольких зависимых событий равна произведению вероятности одного из этих событий на условные вероятности других, т. е.

$$\begin{aligned} & p(B) = p(A_1 \cdot A_2 \cdot \ldots \cdot A_n) = \\ & = p(A_1) \cdot p(A_2 / A_1) \cdot p(A_3 / A_1 A_2) \cdot \ldots \cdot p(A_n / A_1 A_2 \ldots A_{n-1}) = \\ & = p\left(\prod_{i=1}^{n} A_i \right) \end{aligned} \qquad (2.6.2)$$

Для вероятности произведения двух зависимых событий A и B формула (2.6.2) принимает вид:

$$p(AB) = p(A) \cdot p(B / A) = p(B) \cdot p(A / B). \qquad (2.6.3)$$

Если события, составляющие произведение, независимы, то теорема умножения вероятностей упрощается, принимая следующую формулировку:

Вероятность произведения двух или нескольких независимых событий равна произведению вероятностей этих событий.

Таким образом, если в формуле (2.6.2):

$$\begin{cases} p(A_2) = p(A_2 / A_1) \\ p(A_3) = p(A_3 / A_1 A_2) \\ \ldots \\ p(A_n) = p\left(A_n / \prod_{i=1}^{n-1} A_i \right) \end{cases}$$

то можно записать:

$$p(A_1 \cdot A_2 \cdot \ldots \cdot A_n) = p(A_1) \cdot p(A_2) \cdot p(A_3) \cdot \ldots \cdot p(A_n) \qquad (2.6.3)$$

Докажем теорему умножения вероятностей для независимых событий, имея при этом в виду, что этот случай для теории ошибок наблюдений является преимущественным, так как в большинстве испытаний удается обеспечить независимость измерений. Доказательство этой теоремы аналогично доказательству теоремы сложения вероятностей проведем по принципу перехода от частного к общему.

Доказательство. Пусть в двух ящиках имеется: в первом – a_1 белых и b_1 – черных шаров, во втором – a_2 белых и b_2 черных шаров. Определить вероятность того, что будут вынуты два белых шара из двух ящиков, если из каждого ящика возьмут по одному шару.

Обозначим: события:

A_1 – появление белого шара из первого ящика;

A_2 – появление белого шара из второго ящика;

$A = A_1 \cdot A_2$ – совместное появление двух белых шаров при взятии по одному из каждого ящика.

Следовательно,

$$p(A) = p(A_1 \cdot A_2) .$$

Вероятность определим непосредственным подсчетом по формуле (2.3.3).

С этой целью сначала определим число благоприятствующих случаев m и число всех возможных случаев n, причем числом случаев при определении m и n будет число пар соответствующих шаров (белых или белых и черных вместе).

Каждый белый шар из первого ящика может выпасть в паре с одним из белых шаров из второго ящика a_2 раз. Следовательно, всего независимых пар белых шаров может быть:

$$m = a_1 \cdot a_2 .$$

Число всех возможных пар, которые можно составить при опыте, равно:

$$n = (a_1 + b_1)(a_2 + b_2) .$$

Вероятность появления двух белых шаров по одному из каждого ящика будет равна:

$$p(A) = \frac{a_1 a_2}{(a_1 + b_1)(a_2 + b_2)} \qquad (2.6.4)$$

Но в выражении (2.6.4):

$$p(A_1) = \frac{a_1}{(a_1 + b_1)}$$

– вероятность появления белого шара из первого ящика, а:

$$p(A_2) = \frac{a_2}{(a_2 + b_2)}$$

– вероятность появления белого шара из второго ящика.

С учетом этого формула (2.6.4) примет вид:

$$p(A) = p(A_1) \cdot p(A_2) \qquad (2.6.5)$$

что и требовалось доказать.

Распространяя полученный результат (2.6.5) на случай, когда сложное событие (произведение) составлено из *n* независимых событий, можно считать доказанным, что

$$p\left(\prod_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \left(p(A_i)\right) \qquad (2.6.6)$$

В частном случае, который, кстати сказать, при рассмотрении вопросов в теории ошибок встречается довольно часто, а именно, когда:

$$p(A_1) = p(A_2) = ... = p(A_n) = p$$

можно записать:

$$\prod_{i=1}^{n} \left(p(A_i)\right) = p^n \qquad (2.6.7)$$

3. МНОГОКРАТНЫЕ ИСПЫТАНИЯ

3.1. Распределение вероятностей при многократных испытаниях, биномиальное распределение

При исследовании новых приборов или при проверке новых методов работ независимые испытания производятся многократно, т. е. при сохранении определенного комплекса условий повторяются большое число раз. Испытателя в этом случае интересует конечный результат опыта, например, сколько раз при n испытаниях появится ожидаемое событие.

Пусть A – некоторое событие, на появление (или непоявление) которого производятся многократные испытания. Вероятность этого события известна и постоянна. Результатом каждого отдельного опыта может быть появление или непоявление события A, т. е. совершение A, или $\overline{A}$. Требуется определить вероятность того, что при n испытаниях событие появится k раз, т. е. необходимо найти $P_n(k)$. Проанализируем

результаты *n* испытаний в «схеме случаев», т. е. последовательно: после первого испытания, после двух испытаний и т. д. и наконец, после n испытаний.

Когда произведено одно испытание, получим A, или $\overline{A}$, но

$$p(\overline{A}+A)=p(\overline{A})+p(A)=1.$$

После двух испытаний возможны следующие комбинации сложных событий:

$$\overline{A}\,\overline{A} \text{ или } \overline{A}A \text{ или } A\overline{A} \text{ или } AA.$$

Из теоремы умножения вероятностей следует:

$$p(\overline{A}\,\overline{A})+p(\overline{A}A)+p(A\overline{A})+p(AA)=\left(p(\overline{A})+p(A)\right)^2=1$$

Рассуждая аналогично, для *n* испытаний получим:

$$\left(p(\overline{A})+p(A)\right)^n=1. \qquad (3.1.1)$$

Выражение (3.1.1) после разложения его в ряд по формуле бинома Ньютона даст вероятность для (*n + 1*) произведений сложных событий при многократных испытаниях без учета последовательности их появления.

В общем случае имеем:

$$P_n(k)=C_n^k q^{n-k}p^k, \qquad (3.1.2)$$

где: $P_n(k)$–вероятность появления *k* раз при *n* испытаниях;

$$C_n^k=\frac{k!}{k!(n-k)!} \qquad (3.1.3)$$

– число сочетаний из *n* по *k*; *p* – вероятность события; *q* – вероятность противоположного события.

При больших значениях *n* и *k* для вычисления факториала, применяют приближённую формулу Стирлинга:

$$n!\approx\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad (3.1.4)$$

Абсолютная ошибка вычислений факториалов по формуле (3.1.4) увеличивается с увеличением *n*. Относительная же точность (характеризуемая отношением абсолютной ошибки к значению факториала и выражаемая обычно в %) повышается с увеличением *n*. По этой причине формула Стирлинга принадлежит к числу асимптотических формул. Сравнивая формулы (3.1.3) и (3.1.4), легко подсчитать, что при *n = 10* относительная ошибка вычислений составит 0,8 %, при *n = 20* она будет равна 0,4 %.

В практических целях представляется полезным уметь вычислять вероятность появления события хотя бы один раз при выполнении данного комплекса условий. Поскольку сумма вероятностей противоположных событий равна 1, т. е. $p(A)+p(\overline{A})=1$, вероятность того, что некоторое событие *A* ни разу не появится при *n* испытаниях, равна $\left[p(\overline{A})\right]^n$.

Следовательно, вероятность того, что событие A произойдет хотя бы один раз, равна:

$$p(A) = 1 - \left[p(\overline{A})\right]^n. \qquad (3.1.5)$$

3.2. Вероятнейшее число появлений события при многократных испытаниях

При испытаниях новых приборов или методов работ, а также при различных теоретических расчетах исследователь, естественно, ставит перед собой вопрос: какое число появлений ожидаемого события (при многократных испытаниях) наиболее возможно, если выполнен определенный комплекс условий, который в обобщающем виде характеризуется постоянством в процессе опыта вероятности появления события при одном испытании. Конечно, каждый раз, вычисляя все члены разложения в формуле (3.1.1) и зная из определения вероятности, что большая вероятность порождает большую уверенность в получении желаемого результата опыта, можно определить число появлений k_0 события, соответствующее вероятности p при числе испытаний n.

$$k_0 \approx np \qquad (3.2.1)$$

где k_0 – вероятнейшее число появления события при многократных испытаниях.

Следует подчеркнуть, что несмотря на свою простоту формула (3.2.1) имеет исключительно важное значение как для последующих теоретических выкладок, так и в особенности при решении задач на появление числа ошибок в заданных пределах.

Основную трудность при решении практических вопросов с использованием формулы (3.2.1) составляет определение вероятности появления события.

4. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ, ИХ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ, МОДЕЛИРОВАНИЕ

4.1. Понятие случайной величины. Прерывные и непрерывные случайные величины.

Случайной величиной называют переменную величину, сопутствующую случайному событию и отражающую многообразие неучтенных колебаний условий, при которых производится данный опыт. Таким образом, при проведении опыта приходится иметь дело со случайной величиной, конкретное значение которой заранее неизвестно.

Различают два типа случайных величин: *прерывные* (дискретные) и *непрерывные*.

Прерывной или дискретной случайной величиной называют такую случайную величину, возможные значения которой можно заранее указать (например, число попаданий при трех выстрелах — 0, 1, 2, 3; число появлений отрицательной ошибки при пяти измерениях— 0, 1, 2, 3, 4, 5; число появлений события при многократных испытаниях).

Непрерывной случайной величиной называют такую, которая может принять любые значения на некотором непрерывном интервале, которые не могут быть перечислены заранее. Примером может послужить значение радиусов разлёта осколков после взрыва.

Разумеется, конкретные значения случайной величины при ограниченном числе испытаний могут заметно отличаться друг от друга и в этом смысле она не является непрерывной, хотя эти значения и нельзя перечислить заранее. Однако речь идет о возможных значениях случайной величины, т. е. о таких, которые она может принимать в процессе опыта. Таким образом, число наблюденных значений случайной величины всегда конечное; однако возможное число значений непрерывной случайной величины – «несчетное множество».

Здесь и далее будем обозначать случайные величины прописными латинскими буквами *X, Y, Z* …, а их возможные значения соответствующими строчными буквами *x, y, z,* … (например, X — случайная величина; $x_1, x_2, x_3, \ldots, x_n$ — возможные значения случайной величины *X* при *n* испытаниях; $p_1, p_2, p_3, \ldots, p_n$ — соответствующие величинам *x* с соответствующими коэффициентами вероятности).

4.2. Способы задания распределения случайных величин

Для полной характеристики случайной величины недостаточно знать ее значения, а необходимо также знать вероятности, соответствующие этим значениям, т. е. $p_1, p_2, p_3, \ldots, p_n$, или ожидаемые относительные частоты (статистические вероятности) этих значений.

Всякое соотношение, при помощи которого устанавливается связь между значениями случайной величины и соответствующими им вероятностями, называют *законом распределения случайной величины.*

Закон распределения прерывной случайной величины может быть задан следующими способами:

– *аналитически* в виде формулы;

– *численно* в виде простой таблицы распределения, в которой приведены возможные значения случайной величины и соответствующие им вероятности, таблицу распределения часто называют также рядом распределения случайной величины *X*;

– *графически* в виде так называемого многоугольника распределения, часто результаты исследований задаются в виде

интервальных статистических рядов, в этих случаях для графического представления полученных результатов используется гистограмма.

4.3. Вычисление вероятностей для дискретных случайных величин

Хорошим примером распределения дискретной случайной величины *X*, является *биномиальное распределение* $B(n, p)$, *x* принимает значения: *0, 1, 2, ..., n*. Распределение случайной величины *X* определяется следующей формулой:

$$P[K = k] = C_n^k q^{n-k} p^k, \tag{4.3.1}$$

где: $k = 0,1, \ldots n$; $q = 1 - p$; $C_n^k = \dfrac{n!}{(n-k)!k!}$ – число сочетаний из *n* по *k*.

Биномиальное распределение случайной величины может быть рассчитано в пакете STATISTICA [7], при помощи функции **binom(x; p; n)** – возвращающей, вероятность того, что случайная величина *X*, имеющая биномиальное распределение с параметрами *n, p* примет значение *x*.

Суммарная, накопленная вероятность $P[X \le x]$, может вычислена по формуле:

$$\sum_{k=0}^{x} P[K = k] = \sum_{k=0}^{x} C_n^k q^{n-k} p^k \tag{4.3.2}$$

Суммарную, накопленная вероятность так же можно вычислить в пакете STATISTICA при помощи функции **ibinom(x; p; n).**

Распределение Пуассона:

$$P[K = k] = \frac{\lambda^k}{k!} e^{-\lambda}, \tag{4.3.3}$$

возможно вычислить в пакете STATISTICA применяя функцию **poisson(k; λ)**, а накопленную суммарную вероятность при помощи **ipoisson(k; λ)**.

Геометрическое распределение:

$$P[K = k] = (1 - p)^{k-1} \tag{4.3.4}$$

возможно вычислить в пакете STATISTICA применяя функцию **geom(k; p)**, а накопленную суммарную вероятность при помощи **igeom(k; p)**.

4.4. Вычисление вероятностей для непрерывных случайных величин

Равномерное распределение на интервале [*a,b*], R(*a,b*), имеет плотность, определяемую формулой:

$$p(x) = \begin{cases} \dfrac{1}{b-a}, x \in [a,b] \\ 0, x \notin [a,b] \end{cases} \tag{4.4.1}$$

Функция распределения $P(x)$ случайной величины, имеющей равномерное распределение равна:

$$P(x) = \int_{-\infty}^{x} p(x)dx = \begin{cases} 0, x \le a; \\ \dfrac{x-a}{b-a}, x \in (a,b) \\ 0, x \ge b. \end{cases} \qquad (4.4.2)$$

Нормальное распределение имеет плотность, определяемую формулой:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-m)^2}{2\sigma^2}}, \qquad (4.4.3)$$
$$-\infty < x < \infty$$

Функция распределения *P(x)* нормального распределения:

$$P(x) = \int_{-\infty}^{x} p(y)dy = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{\frac{-(y-m)^2}{2\sigma^2}} dy \qquad (4.4.4)$$

Параметры m и σ нормального распределения равны соответственно математическому ожиданию (m) и дисперсии (σ^2) случайной величины X:

$$M[X] = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-m)^2}{2\sigma^2}} dx = m, \qquad (4.4.5)$$

$$D[X] = M[X^2] - (M[X])^2 = \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-m)^2}{2\sigma^2}} dx - m^2 = \sigma^2 \qquad (4.4.6)$$

О смысле математического ожидания и дисперсии будет сказано далее.

В пакете STATISTICA нормальную плотность распределения (4.4.3) можно вычислить, используя функцию **normal(x; m; σ)**. Значение функции нормального распределения в точке x: **inormal(x; m; σ)**.

Нормальное распределение с нулевым математическим ожиданием *m=0*, и дисперсией равной единице, называют $\sigma^2=1$, называют *стандартным нормальным распределением*.

Функция плотности стандартного нормального распределения:

$$p(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-(x)^2}{2}}. \qquad (4.4.7)$$
$$-\infty < x < \infty$$

Функция распределения:

$$P(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{\frac{-(t)^2}{2}} dt \qquad (4.4.8)$$

Экспоненциальное распределение, имеет плотность, определяемую следующим выражением:

$$p(x)=\begin{cases}\lambda e^{-\lambda x}, x \geq 0\\ 0, x<0\end{cases} \quad (4.4.9)$$

где λ – параметр экспоненциального распределения.

Функция плотности P(x) случайной величины X, имеющий экспоненциальное распределение:

$$P(x)=\begin{cases}\int\limits_0^x \lambda e^{-\lambda t}dt = 1-e^{-\lambda x}, x>0\\ 0, x \leq 0\end{cases}. \quad (4.4.10)$$

В пакете STATISTICA экспоненциальную плотность распределения (4.4.9) можно вычислить, используя функцию **expon(x; λ)**. Значение функции экспоненциального распределения в точке x: **iexpon(x; λ)**.

В технических приложениях математической статистики экспоненциальное распределение можно встретить в таких параметрах устройств, как, например, «наработка времени на отказ».

Распределение Стьюдента, имеет плотность, определяемую следующим выражением:

$$p(x)=\frac{\int\limits_0^\infty t^{\frac{n+1}{2}-1}e^{-t}dt}{\sqrt{\pi n}\int\limits_0^\infty t^{\frac{n}{2}-1}e^{-t}dt}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (4.4.11)$$

где: n – число степеней свободы.

В пакете STATISTICA плотность распределения Стьюдента можно вычислить, используя функцию **student(x, n)**. Значение функции экспоненциального распределения в точке x: **istudent(x, n)**.

4.5. Двумерное нормальное распределение

Двумерное нормальное распределение – это распределение системы двух случайных величин (X,Y) с плотностью распределения $p(x,y)$, определяемые формулой:

$$p(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot$$
$$\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2}-\frac{2\rho(x-m_1)(y-m_2)}{\sigma_1\sigma_2}+\frac{(y-m_2)^2}{\sigma_2^2}\right]\right\} \quad (4.5.1)$$

где m_1, m_2, σ_1, σ_2, математическое ожидание (m) и дисперсии (σ^2) для величин X, Y соответственно, параметр будет рассмотрен ниже.

Плотность распределения случайной величины X, $p(x)$, можно вычислить:

$$p_X(x)=\int\limits_{-\infty}^{\infty}p(x,y)dy=\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left[-\frac{(x-m_1)^2}{2\sigma_1^2}\right] \quad (4.5.2)$$

Плотность распределения случайной величины Y, $p(y)$, можно вычислить:

$$p_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[-\frac{(x-m_2)^2}{2\sigma_2^2}\right] \qquad (4.5.3)$$

По формулам (4.5.2, 4.5.3) можно заметить, что в случае двумерного нормального распределения с плотностью $p(x,y)$ компоненты X, Y имеют нормальное распределение, причём: $M[X]=m_1$, $D[X]=\sigma_1^2$ и $M[Y]=m_2$, $D[Y]=\sigma_2^2$.

Ковариацию X и Y можно рассчитать как:

$$\operatorname{cov}(X,Y) = M\left[(X-m_1)(Y-m_2)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-m_1)(y-m_2)\,p(x,y)dxdy = \rho\sigma_1\sigma_2 . \qquad (4.5.4)$$

Из выражения (4.5.4) можно заметить, что параметр ρ есть коэффициент корреляции между X и Y.

Если X и Y некорелированны то $\rho=0$, в этом случае функцию (4.5.1) можно записать в виде:

$$p(x,y) = p_X(x)\cdot p_Y(y) . \qquad (4.5.5)$$

В случае двумерного нормального распределения из некоррелированности компонент следует их независимость.

4.6. Моделирование распределений при помощи пакета STATISTICA

Далее рассмотрена работа с версией программы STATISTICA 8.0**.** Для начала моделирования необходимо создать документ, который будет содержать данные. В программе STATISTICA возможно создать несколько видов документов, нас будет интересовать:

– Spreadsheet – (просторный холст англ.), лист данных, далее л и с т может содержать векторы данных;

или

–Workbook – (рабочая книга англ.), может содержать множество Spreadsheet или графических интерпретаций, с возможностью группировки в виде дерева.

Целесообразнее использовать Workbook, т.к. в этом случае можно манипулировать одним проектом при работе с множеством различных данных. Создать Workbook достаточно просто, для этого надо выполнить следующую последовательность действий: **File→New… Workbook** далее включить флаг Insert empty spreadsheet (включить пустой лист) и нажать **OK**.

Размер spreadsheet по умолчанию (10×10), в большинстве случаев такого объёма недостаточно. При вставке данных, например, из Excel размер листа автоматически будет увеличен, можно увеличить размер

листа, выполнив следующие действия: **Insert→Add Cases…**, далее в появившемся окошке выбрать число добавляемых строк (How many), и после какой строки добавить (Insert after cases). Аналогичным образом можно добавить столбцы **Insert→Add Variables…**

Итак, создав новый проект и сохранив (**File→Save As…**) под уникальным именем можно перейти к моделированию.

Для моделирования выберем нормальное распределение (4.4.3, 4.4.4), высоту столбцов для рассмотренных ниже примеров лучше выбрать равную 50-ти.

Вектор **Var1** заполним данными функции **=normal(x; m; σ)**, для этого необходимо сделать два щелчка левой кнопкой мыши, разместив указатель мыши над ячейкой **Var1**, или расположив указатель мыши над **Var1** выбрать в контекстном меню (правая кнопка мыши) элемент **Variable Specs…** Далее в поле **Long name**, занести требуемую формулу **=normal(v0;25;10)** и нажать кнопку **OK**. Символ **v0**, означает перебор номеров (натуральных чисел) 1, 2, 3 …, а символ **v0-1** – перебор номеров 0, 1, 2 …

В поле **Long name** (**Var2**) запишем формулу **=inormal(v0;25;10)**, вектор **Var3** нам понадобится для построения графиков полученных функций, в поле **Long name** (**Var3**) запишем выражение **=v1**.

После заполнения векторов данных значениями введённых функций можно перейти к визуализации, графические возможности STATISTICA очень широки. Выберем в меню пункт **Graphs→2D Graphs→Scatterplots…**, далее вкладку **Advanced** в поле **Graph type→Double-Y** и в поле линии тренда для нашего случая подойдёт сплайн: **Fit→Spline.** Следующим действием выберем переменные нажав кнопку **Variables** для **X**: **Var3**, для **Y Left**: **Var1** и **Y Right**: **Var2**. Далее нажимаем кнопку OK. Результат построения графиков изображен на рисунке 4.6.1. Редактирование цветов, надписей, линий сетки, диапазонов и способов подписи шкал, достаточно интуитивен, меню редактирования можно вызвать двойным щелчком левой кнопки мыши расположив указатель над интересующим элементом графика.

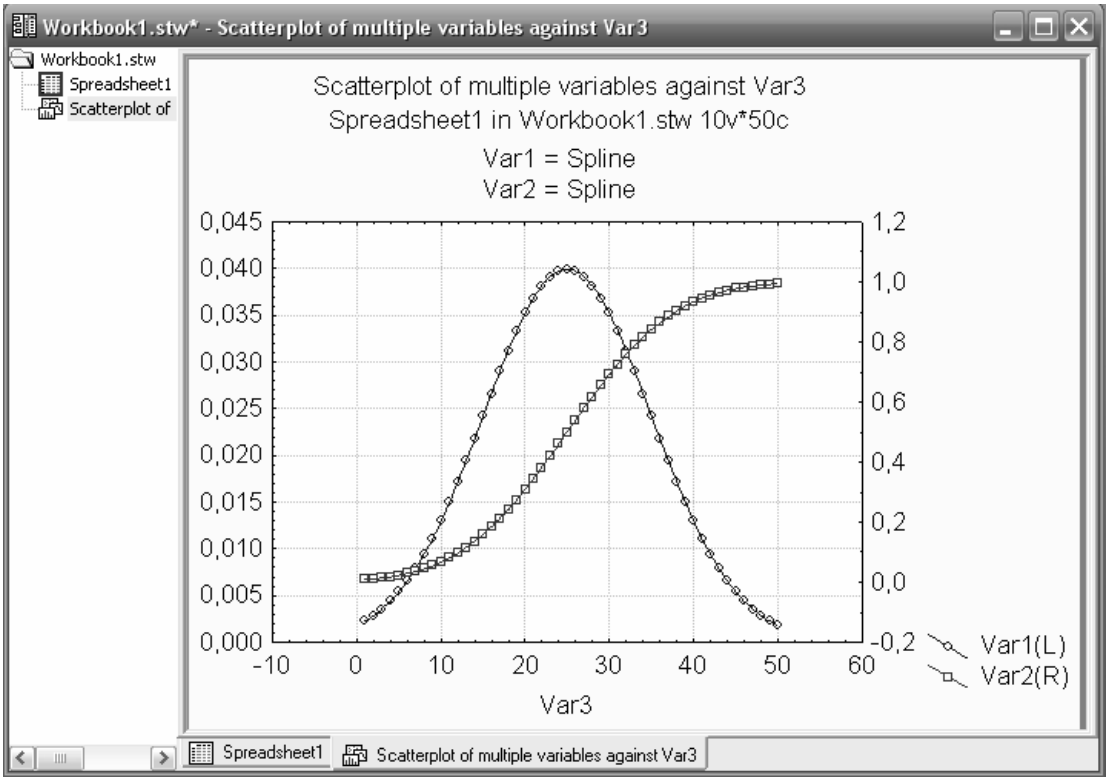

Рисунок 4.6.1. Результаты моделирования нормального распределения.

Часто для различных исследований может понадобится случайная величина с определённым законом распределения. Пакет STATISTICA позволяет производить такое моделирование. Рассмотрим случай генерации данных имеющих нормальное распределение. За генерацию случайной величины в диапазоне (*0,max*) функция **=rnd(max)**, где *max* – максимальное значение. Введя в нашей рабочей книге в поле **Long name** (**Var4**) функцию **=vnormal(rnd(1);25;10)**, мы получим вектор случайных чисел с нормальным законом распределения. Далее можно построить график плотности распределения полученной выборки, воспользуемся разделом главного меню **Statistics→Basic Statistics/Tables** в появившемся окошке **Descriptive statistics**, далее выбрать вектор **Var4**, нажав кнопку **Variables**, затем нажать кнопку **Histograms**, результат моделирования показан на рисунке 4.6.2. Будет полезным при работе с различными разделами главного меню и опциональных меню STATISTICA поинтересоваться их назначением нажав клавишу **F1**.

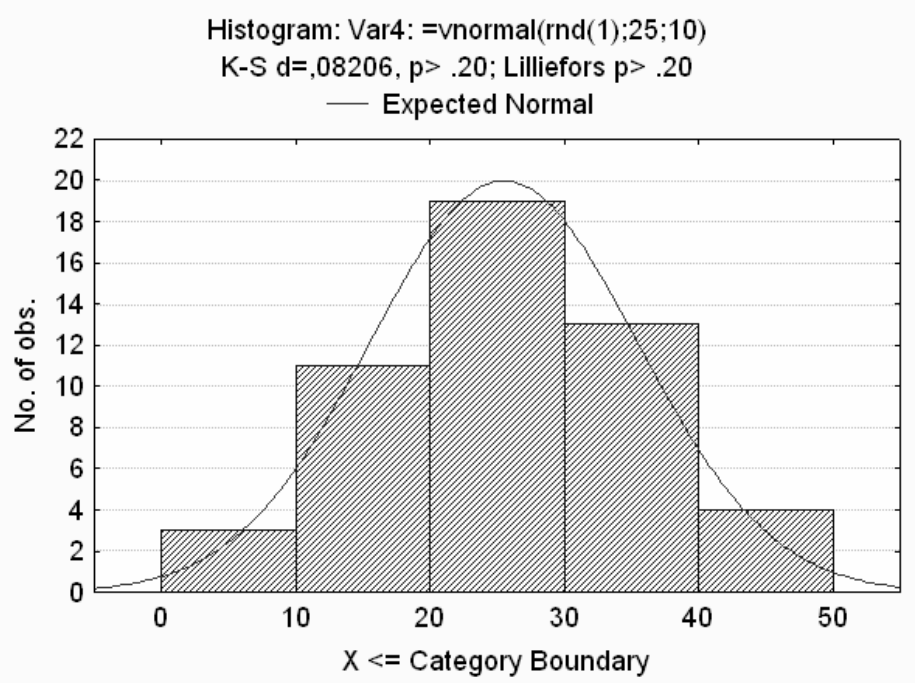

Рисунок 4.6.2. Моделирование случайной величины с распределением по нормальному закону.

5. ТИПЫ СТАТИСТИЧЕСКИХ ДАННЫХ

5.1. Типы статистических данных

Статистические данные представляют собой наблюдаемые значения одного или нескольких признаков обследуемой совокупности объектов [3]. Различают *количественные* и *качественные* признаки. Значения количественных признаков могут быть *непрерывными* или *дискретными*. Примерами качественных признаков являются, например, пол, семейное положение, цвет кожи, качество товара. В свою очередь качественные признаки в зависимости от вида данных делятся на *номинальные* (классификационные) и *ординальные* (порядковые). Говорят также, что соответствующие качественные признаки измеряются в номинальной или порядковой шкале. Разница между этими шкалами состоит в следующем.

Признак, измеряемый в номинальной шкале, принимает одно значение из конечного числа заведомо установленных градаций. Примерами признаков, измеряемых в номинальной шкале, являются тип дома (каменный, деревянный); семейства в биологии (покрытосемянные, голосемянные); тип покрытия дороги (грунтовая, брусчатка, асфальт) и т. п. Статистические данные, измеряемые в номинальных шкалах, могут быть представлены в виде таблиц, в которых приводятся частоты появления той или иной градации признака. Часто номинальные данные появляются при

обработке социологических опросов, в медицине. Таблицы таких данных называются таблицами.

Значения качественных признаков, измеряемых в ординальной шкале, могут быть упорядочены, например, по возрастанию. Примерами таких признаков являются тестовые баллы и школьные оценки, качество условий жизни. Для признаков, измеряемых в ординальных шкалах, операции сложения и вычитания не имеют смысла. Так, нельзя сказать, что студент, получивший на экзамене «пять» по статистике знает предмет на одну единицу лучше, чем студент, получивший по этому предмету «четыре», поскольку для знаний не существует единицы измерения. Однако можно сказать, что первый студент знает статистику лучше, чем второй.

Для представления значений ординальных признаков в числовой форме используется следующий способ. Все значения признака записываются в порядке возрастания в виде ряда. Каждому значению поставим в соответствие натуральное число, равное его номеру в ряду. Это число называется *рангом*. Например, качество условий жизни (плохое, удовлетворительное, хорошее, очень хорошее) будет представлено рангами 1, 2, 3, 4. Для ординальных признаков, представленных в виде рангов, разработаны специальные статистические методы, позволяющие измерять степень близости признаков (например, *ранговая корреляция*), проверять гипотезы о виде распределения, проводить дисперсионный, кластерный анализ.

Для данных, представленных в номинальной шкале, также не определены операции сложения и вычитания. Эти данные (в отличие от ординальных признаков) не могут быть упорядочены и, следовательно, оцифрованы с помощью рангов. Применяя специальные статистические методы для номинальных признаков, можно проверить гипотезы о независимости признаков и о принадлежности двух или нескольких *выборок* к одному виду. Для оцифровки номинальных признаков используются числовые метки. Выбор меток в зависимости от цели статистического анализа может проводиться по различным критериям [5].

Введём понятие – *шкалы измерений*. Переменные различаются тем "насколько хорошо" они могут быть измерены или, другими словами, как много измеряемой информации обеспечивает шкала их измерений. Очевидно, в каждом измерении присутствует некоторая ошибка, определяющая границы "количества информации", которое можно получить в данном измерении. Другим фактором, определяющим количество информации, содержащейся в переменной, является тип шкалы, в которой проведено измерение. Различают следующие типы шкал:(a) *номинальная*, (b) порядковая (*ординальная*), (c) *интервальная* (d)

относительная (шкала отношения). Соответственно, имеем четыре типа переменных: (a) *номинальная*, (b) порядковая (*ординальная*), (c) *интервальная* и (d) *относительная*.

a.) Номинальные переменные используются только для качественной классификации. Это означает, что данные переменные могут быть измерены только в терминах принадлежности к некоторым, существенно различным классам; при этом вы не сможете определить количество или упорядочить эти классы. Например, вы сможете сказать, что 2 индивидуума различимы в терминах переменной A (например, индивидуумы принадлежат к разным национальностям). Типичные примеры номинальных переменных - пол, национальность, цвет, город и т.д. Часто номинальные переменные называют категориальными.

b.) Порядковые переменные позволяют ранжировать (упорядочить) объекты, указав какие из них в большей или меньшей степени обладают качеством, выраженным данной переменной. Однако они не позволяют сказать "на сколько больше" или "на сколько меньше". Порядковые переменные иногда также называют ординальными. Типичный пример порядковой переменной - социоэкономический статус семьи. Мы понимаем, что верхний средний уровень выше среднего уровня, однако сказать, что разница между ними равна, скажем, 18% мы не сможем. Само расположение шкал в следующем порядке: номинальная, порядковая, интервальная является хорошим примером порядковой шкалы.

c.) Интервальные переменные позволяют не только упорядочивать объекты измерения, но и численно выразить и сравнить различия между ними. Например, температура, измеренная в градусах Фаренгейта или Цельсия, образует интервальную шкалу. Вы можете не только сказать, что температура 40 градусов выше, чем температура 30 градусов, но и что увеличение температуры с 20 до 40 градусов вдвое больше увеличения температуры от 30 до 40 градусов.

d.) Относительные переменные очень похожи на интервальные переменные. В дополнение ко всем свойствам переменных, измеренных в интервальной шкале, их характерной чертой является наличие определенной точки абсолютного нуля, таким образом, для этих переменных являются обоснованными предложения типа: x в два раза больше, чем y. Типичными примерами шкал отношений являются измерения времени или пространства. Например, температура по Кельвину образует шкалу отношения, и вы можете не только утверждать, что температура 200 градусов выше, чем 100 градусов, но и что она вдвое выше. Интервальные шкалы (например, шкала Цельсия) не обладают данным свойством шкалы отношения. Заметим, что в большинстве

статистических процедур не делается различия между свойствами интервальных шкал и шкал отношения.

Множество всех обследуемых объектов называется *генеральной совокупностью*. Если это множество содержит небольшое число элементов, то возможно полное обследование всех его элементов. Однако в большинстве случаев в силу того, что генеральная совокупность имеет очень много элементов либо ее элементы труднодоступны, либо по другим причинам обследуется некоторая часть генеральной совокупности – *выборка*. В этом случае основные характеристики генеральной совокупности (их называют статистиками: *среднее*, *дисперсия* и т. д.) оцениваются (т. е. определяются приближенно) по выборке. Соответствующие статистики называются «*выборочное среднее*», «*выборочная дисперсия*» и т. д. Очевидно, что не всякая выборка правильно отражает свойства генеральной совокупности. Например, нельзя судить о среднем душевом доходе населения по выборке, составленной из доходов служащих финансовых компаний. Выборка должна давать правильное, неискаженное представление о генеральной совокупности, или, как говорят, быть репрезентативной. Если свойства генеральной совокупности заранее неизвестны, то, за неимением лучшего, следует использовать простой случайный выбор. Это означает, что все элементы генеральной совокупности должны иметь равные шансы попасть в выборку.

Например, при выяснении мнения всех студентов университета по какому-либо вопросу, выборка, составленная из студентов первого курса, не будет репрезентативной. Процедуру случайного выбора можно организовать, например, так. Запишем фамилии всех студентов на отдельные карточки, которые затем тщательно перетасуем, и из всего множества карточек отберем нужное количество. Ответы выбранных таким способом студентов составят репрезентативную выборку. Если требуется, чтобы в выборке были представлены элементы различных групп, составляющих генеральную совокупность, используется процедура типического отбора. Так, если студенты первого курса составляют 15 % всех студентов университета, то и в выборке они должны составлять 15 %. В некоторых случаях необходимо учитывать не только курс, но и специализацию студентов, если это может повлиять на результаты опроса.

Как правило, статистические данные в силу ошибок измерений, влияния внешней среды, присущей индивидуумам случайной изменчивости, имеют разброс. Рассмотрим, например, результаты выборочного контроля партии расфасованной продукции. С большой вероятностью можно сказать, что в выборке не найдется ни одной пары пакетов, имеющих один и тот же вес. Основная задача статистики состоит

в получении осмысленных заключений именно из такого типа данных, т. е. данных, подверженных случайной изменчивости.

Математическая модель статистических данных содержит *детерминированную* и *случайную* составляющие. В простейшей модели случайная компонента – это случайная величина. Например, математическая модель данных, представляющих вес расфасованной продукции, есть сумма двух величин: номинального веса m пакета (детерминированная компонента) и отклонения Δm истинного веса пакета от номинального. Это отклонение (имеющее случайный характер) можно рассматривать как сумму очень большого числа случайных факторов всегда имеющих место в производственных условиях (износ оборудования, влажность и температура продукта и др.).

Выборка – это множество случаев выбранных из генеральной совокупности, с помощью определённой процедуры. Выборка характеризуется качественной и количественной характеристиками. Качественная характеристика выборки – критерий выбора из генеральной совокупности – *способы построения выборки*. Количественная характеристика выборки – *объём выборки*.

Случайная выборка – выборка, из генеральной совокупности формируется случайным образом. Объём выборки m считается произвольной, но фиксированной, неслучайной величиной. Формально это означает, что с генеральной совокупностью X связывается вероятностное пространство $\langle X_m, \Sigma_m, P_m \rangle$, где X_m – множество всех выборок длины m, Σ_m – заданная на этом множестве комплекс условий, P_m – вероятностная мера, как правило, неизвестная. Случайная выборка $X_m = (x_1, x_2 ..., x_m)$ – это последовательность из m прецедентов, выбранная из множества X согласно вероятностной мере P_m.

Однородная выборка – это такая выборка из генеральной совокупности, при которой любые два претендента X_k, X_l, где k, l – номера произвольных выборок, имеют одинаковые распределения.

Зависимая выборка – выборка X_l из генеральной совокупности X называется зависимой, если для случая $x^l \in X_l$, в выборке X_k из X найдётся гомоморфный образ $x^k \in X_k$. В противном случае выборка называется *независимой*. Примером зависимой выборки может быть множество экспериментальных измерений произведённых до и после некоторого воздействия.

Простая выборка – это случайная, однородная, независимая выборка.

Применительно к *нейронным сетям* различают *обучающие* и *тестовые* выборки.

5.2. Математическое ожидание, дисперсия

Числовые характеристики случайной величины определяют основные свойства распределения: *среднее значение*, *разброс значений* случайной величины относительно среднего значения и другие. Важнейшей характеристикой случайной величины является ее *математическое ожидание* или *среднее значение*.

Определение. Пусть X – дискретная случайная величина, принимающая значения $x_1, x_2, \ldots$ с вероятностями $p_1, p_2, \ldots$ Математическое ожидание $M[X]$ случайной величины X определяется формулой:

$$M[X] = \sum_i x_i p_i \qquad (5.2.1)$$

в предположении, что ряд (5.2.1) абсолютно сходится. Если ряд (5.2.1) расходится, то говорят, что случайная величина X не имеет конечной величины математического ожидания.

Для абсолютно непрерывного распределения:

$$M[X] = \int_{-\infty}^{\infty} x \cdot p(x)dx \qquad (5.2.2)$$

где $p(x)$ – непрерывная плотность распределения величины X.

Рассмотрим основные свойства математического ожидания.

1. $M[c] = c$, где c – константа.

2. $M[cX] = cM[X]$ и $M[X+ c] = M[X] + c$, где c – константа.

3. Пусть случайная величина Z является заданной функцией случайной величины X: $Z = h(X)$. Например, $Z = tg(X)$. Случайная величина Z имеет конечное математическое ожидание $M[Z]$, вычисляемое по формуле:

$$M[Z] = \sum_i h(x_i) p_i \qquad (5.2.2)$$

при условии, что ряд (5.2.2) абсолютно сходится.

4. Пусть $X_1, X_2, \ldots, X_n$ – случайные величины с конечными математическими ожиданиями. Математическое ожидание их суммы равно сумме их математических ожиданий:

$$M[X_1 + X_2 + \ldots + X_n] = M[X_1] + M[X_2] + \ldots + M[X_3]. \qquad (5.2.3)$$

5. Пусть X и Y – взаимно независимые случайные величины с конечными математическими ожиданиями. Математическое ожидание произведения $X \cdot Y$ равно произведению их математических ожиданий:

$$M[X \cdot Y] = M[X] \cdot M[Y]. \qquad (5.2.4)$$

Это правило распространяется на любое конечное число взаимно *независимых* случайных величин.

Заметим, что равенство (5.2.4) для *зависимых* случайных величин, вообще говоря, не выполняется.

Другой важной характеристикой случайной величины X является ее *дисперсия* $D[X]$. Дисперсия характеризует разброс значений случайной величины относительно ее математического ожидания и вычисляется по формуле:

$$D[X] = M[(X - M[X])^2] = \sum_{i=1} (x_i - M[X])^2 \, p_i \,, \qquad (5.2.5)$$

предположении, что ряд (5.2.5) сходится.

Величина $\sigma = \sqrt{D[X]}$ называется средним *квадратическим* или *стандартным отклонением* случайной величины X.

Основные свойства дисперсии.

1. $D[X] \geq 0$.

2. Если случайная величина X постоянна, то ее дисперсия равна нулю: $D[X] = 0$.

3. $D[cX] = c^2 D[X]$, где c=*const*.

4. $D[X + c] = D[X]$, где c=*const*.

5. Дисперсия суммы независимых случайных величин X и Y равна сумме их дисперсий:

$$D[X + Y] = D[X] + D[Y]$$

На практике, в большинстве случаев бывает, что плотность распределения выборки X из генеральной совокупности неизвестна и имеет ограниченный объём N. Поэтому используют величину называемую *среднее значение* или *наилучшая оценка* или *среднее арифметическое* величины X:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (5.2.6)$$

Дисперсию в этом случае можно рассчитать как:

$$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \qquad (5.2.7)$$

При извлечении корня из величины σ_x^2 получим *стандартное отклонение* или *среднее квадратическое отклонение*:

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2} \,. \qquad (5.2.8)$$

Есть другое «улучшенное» определение для *стандартного отклонения* σ_x:

$$\sigma_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} \,. \qquad (5.2.9)$$

Очевидно, что определение (5.2.9) даёт большее значение, чем (5.2.8) (особенно при малых значениях N), что помогает компенсировать оценку погрешности в результатах измерений X.

5.3. Дополнительные интервальные оценки, оценка характеристик рассеяния

Пусть $x_1, x_2, ...x_n$ – положительные числа, тогда *среднее геометрическое* вычисляется по формуле:

$$x_g = \sqrt{x_1 \cdot x_2 \cdot ... \cdot x_n} \qquad (5.3.1)$$

Логарифм среднего геометрического:

$$\lg(x_g) = \frac{\sum_{i=1}^{n} \lg(x_i)}{n} \qquad (5.3.2)$$

Среднее гармоническое x_H вычисляется по формуле:

$$x_H = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1} \qquad (5.3.3)$$

Среднее гармоническое используется для характеристики данных, размерность которых выражена отношением различных физических величин: [км/ч], [кг/с] и т.д.

Между средним арифметическим, средним гармоническим и средним геометрическим существует следующая связь:

$$\bar{x} \geq x_g \geq x_H . \qquad (5.3.4)$$

Наиболее распространенными мерами рассеяния величины X являются: размах, средний межквартильный размах, персентильный размах, дисперсия и среднее квадратическое отклонение.

Размах определён как разность между максимальным и минимальным значением их выборки *X*.

Квантиль – такое число, что заданная случайная величина не превышает его с фиксированной вероятностью.

При разделении вариационного ряда X тремя квартилями (Q_1, Q_2, Q_3) на четыре равные части (рис. 5.3.1), средний межквартильный размах равен половине разностей верхнего и нижнего квартилей или половине межквартильного размаха:

$$R_Q = \frac{Q_3 - Q_1}{2} . \qquad (5.3.5)$$

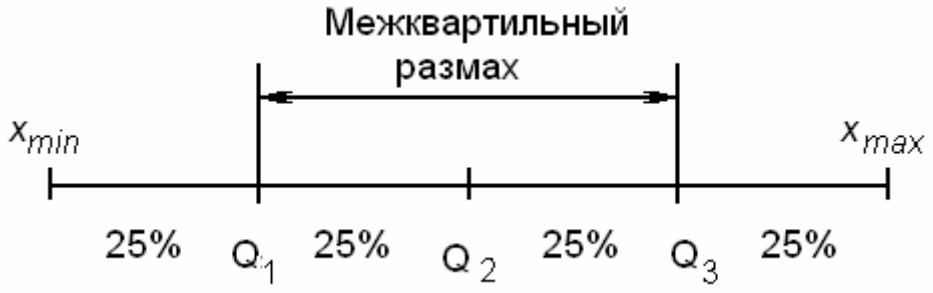

Рисунок 5.3.1. Межквартильный размах.

Персентильный размах равен разности 90- и 10-го персентилей:

$$P_{90} - P_{10} = x_{0.9} - x_{0.1} \qquad (5.3.6)$$

В качестве меры относительного разброса данных используют *коэффициент вариации*:

$$V = \frac{\sigma_x}{\bar{x}}, \qquad (5.3.7)$$

или выраженный в процентах:

$$C_V = \frac{\sigma_x}{\bar{x}} \cdot 100\% . \qquad (5.3.8)$$

Доверительный интервал для x называется интервал (Θ_1, Θ_2), накрывающий истинное значение с заданной вероятностью $p = 1 - \alpha$:

$$P[\Theta_1 < x < \Theta_2] = 1 - \alpha . \qquad (5.3.9)$$

Число $1 - \alpha$ называют *доверительной вероятностью*, а значение α *уровнем значимости*. Обычно используют значения $1 - \alpha$, равные 0,90; 0,95; 0,99. Для отыскания доверительного интервала необходимо знать закон распределения величины X.

В последнее время стала широко использоваться понятие *информационной энтропии*, особенно при медицинских исследованиях, энтропия является мерой беспорядка, применительно к данным мерой хаотичности распределения величины X. Энтропию можно рассчитать по формуле:

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i . \qquad (5.3.10)$$

для непрерывного распределения:

$$H(x) = -\int_{-\infty}^{\infty} p(x) \log_2 p(x)\, dx \qquad (5.3.11)$$

В случае использования выражения (5.3.9) значение энтропии будет зависеть от числа n – элементов дискретного распределения, этот факт необходимо учитывать т.к. изменение n, например, для различных выборок будет влиять на значение энтропии.

В общем случае выбор характеристики рассеяния зависит от особенностей (объема данных, критичности и т.д.) решаемой задачи.

5.4. Применение пакета STATISTICA

Нормальное распределение важно по многим причинам. В большинстве случаев оно является хорошим приближением функций распределения случайной величины. Распределение многих статистик является нормальным или может быть получено из нормальных с помощью некоторых преобразований. Рассуждая философски, можно сказать, что нормальное распределение представляет собой одну из

эмпирически проверенных истин относительно общей природы действительности и его положение может рассматриваться как один из фундаментальных законов природы. Точная форма нормального распределения (характерная "колоколообразная кривая") определяется только двумя параметрами: средним и стандартным отклонением.

Характерное свойство нормального распределения состоит в том, что 68% всех его наблюдений лежат в диапазоне ±1 стандартное отклонение от среднего, а диапазон ±2 стандартных отклонения содержит 95% значений. Другими словами, при нормальном распределении, стандартизованные наблюдения, меньшие -2 или большие +2, имеют относительную частоту менее 5% (Стандартизованное наблюдение означает, что из исходного значения вычтено среднее и результат поделен на стандартное отклонение (корень из дисперсии)). Если у вас имеется доступ к пакету STATISTICA, Вы можете вычислить точные значения вероятностей, связанных с различными значениями нормального распределения, используя Вероятностный калькулятор; например, если задать *z*-значение (т.е. значение случайной величины, имеющей стандартное нормальное распределение) равным 4, соответствующий вероятностный уровень, вычисленный STATISTICA будет меньше .0001, поскольку при нормальном распределении практически все наблюдения (т.е. более 99.99%) попадут в диапазон ±4 стандартных отклонения.

Выбрав в главном меню раздел **Statistics→Basic Statistics/Tables** откроется окно в котором можно выбрать **Descriptive statistics** нужно можно выбрать вектор для которого будет произведён расчёт и нажать кнопку **Summary Statistics**. В результате (по умолчанию) будет произведён расчёт числа валидных элементов вектора (**Valid N**), среднего арифметического значения (**Mean**), стандартного отклонения (**Std.Dev.**), произведён выбор минимального (**Minimum**) и максимального (**Maximum**) значений. Выбрав в окне **Descriptive statistics** опции можно рассчитать дополнительные параметры.

Достаточно часто бывает полезным вероятностный калькулятор (**Probability Calculator**). Вероятностный калькулятор (рис. 5.4.1.) позволяет производить расчёт ряда связанных параметров. Выбрать его можно в главном меню **Statistics→Probability Calculator→Distributions** в появившемся окошке можно выбрать тип распределения, и исследовать взаимосвязь (в рамках функциональной связи выбранного распределения), следующих величин: среднего значения (**mean**), стандартного отклонения (**st.dev.**), вероятности (**p**) и значения (**x:** $x \in X$). Параллельно с выводом значений вероятностный калькулятор производит построение функции плотности (**Density Function**) и функции распределения (**Distributtion Function**).

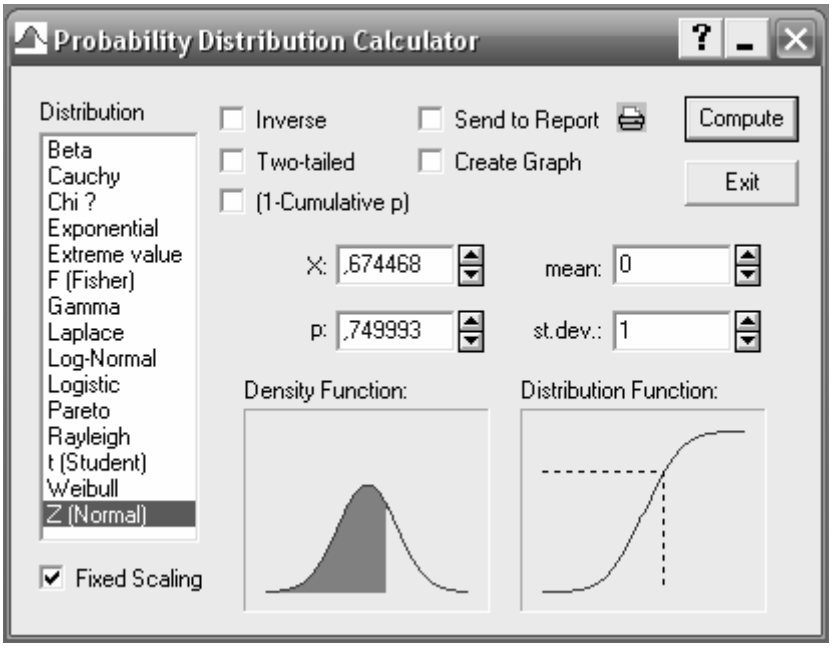

Рисунок 5.4.1. Вероятностный калькулятор.

6. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

6.1. Методы математической статистики проверки гипотез

Во многих случаях результаты наблюдений используются для проверки предположений (гипотез) относительно тех или иных свойств распределения генеральной совокупности. В частности, такого рода задачи возникают при сравнении различных технологических процессов или других исследованиях по определенным измеряемым признакам.

Пусть X – наблюдаемая дискретная или непрерывная случайная величина. *Статистической гипотезой* H называется предположение относительно параметров или вида распределения случайной величины X. Статистическая гипотеза H называется простой, если она однозначно определяет распределение случайной величины X; в противном случае, гипотеза H называется сложной. Например, *простой гипотезой* является предположение о том, что случайная величина X распределена по нормальному закону $N(0, 1)$; если же высказывается предположение, что случайная величина имеет нормальное распределение $N(m, 1)$, где $a \le m \le b$, то это *сложная гипотеза*. Другим примером сложной гипотезы является предположение о том, что непрерывная случайная величина X с вероятностью $1/3$ принимает значение из интервала (1, 5); в этом случае распределение случайной величины X может быть любым из класса непрерывных распределений [3].

Часто распределение случайной величины X известно и по выборке наблюдений необходимо проверить предположения о значении параметров этого распределения. Такие гипотезы называются *параметрическими*.

Проверяемая гипотеза называется нулевой гипотезой и обозначается H_0. Наряду с гипотезой H_0 рассматривают одну из альтернативных (конкурирующих) гипотез H_1,. Например, если проверяется гипотеза H_0 о параметре Θ, пишут $H_0 : \Theta = \Theta_0$, где Θ_0 – известное значение, в качестве альтернативной гипотезы можно рассмотреть одну из следующих гипотез:

$$H_1^{(1)} : \Theta > \Theta_0; H_1^{(2)} : \Theta < \Theta_0; H_1^{(3)} : \Theta \neq \Theta_0; H_1^{(4)} : \Theta = \Theta_1;$$

где Θ_0, Θ_1, – известное значение. Выбор альтернативной гипотезы определяется конкретной формулировкой задачи.

Правило, по которому принимается решение принять или отклонить гипотезу H_0, называется критерием. Так как решение принимается на основе выборки наблюдений случайной величины X, необходимо выбрать подходящую статистику Z, называемую в этом случае *статистикой критерия*. При проверке простой параметрической гипотезы $H_0 : \Theta = \Theta_0$ в качестве статистики критерия обычно выбирают ту же статистику, что и для оценки параметра Θ.

Проверка статистической гипотезы основывается на принципе, в соответствии с которым маловероятные события считаются невозможными, а события, имеющие большую вероятность, – достоверными. Этот принцип можно реализовать следующим образом. Перед анализом выборки назначается некоторая малая вероятность α, называемая *уровнем значимости*. Пусть V – множество значений статистики Z, а V_k — такое подмножество, что, при условии истинности гипотезы H_0, вероятность попадания статистики критерия в V_k равна α :

$$P[Z \in V_k / H_0] = \alpha . \qquad (6.1.1)$$

Пусть z_B – выборочное значение статистики Z, вычисленное по выборке наблюдений. Критерий формулируется следующим образом: отклонить гипотезу H_0, если $z_B \in V_k$; принять гипотезу H_0, если $z_B \in V \setminus V_k$. Критерий, основанный на использовании заранее заданного уровня значимости α, называют *критерием значимости*. Множество V_k всех значений статистики критерия Z, при которых принимается решение отклонить гипотезу H_0, называется *критической областью*; область $V \setminus V_k$ называется областью *принятия гипотезы* H_0.

Уровень значимости α определяет «размер» критической области V_k. Положение критической области на множестве значений статистики Z зависит от формулировки альтернативной гипотезы H_1. Например, если проверяется гипотеза $H_0 : \Theta = \Theta_0$, а альтернативная гипотеза H_1, формулируется как $H_1 : \Theta > \Theta_0 \, (\Theta < \Theta_0)$, то критическая область размещается на правом (левом) «хвосте» плотности распределения статистики Z, т. е. имеет вид неравенства:

$$Z > z_{1-\alpha} \text{ или } Z < z_\alpha , \qquad (6.1.2)$$

где $z_{1-\alpha}$ и z_{α} – квантили распределения статистики Z соответственно порядка $1-\alpha$ и α вычисленные при условии, что верна гипотеза H_0. В этом случае критерий называется *односторонним*, соответственно право- и левосторонним. Если альтернативная гипотеза формулируется как $H_1 : \Theta \neq \Theta_0$, критическая область размещается на обоих «хвостах» плотности распределения Z, т. е. определяется совокупностью неравенств:

$$\begin{cases} Z < z_{1-\frac{\alpha}{2}} \\ Z < z_{\frac{\alpha}{2}} \end{cases}. \qquad (6.1.3)$$

В этом случае критерий называют *двусторонним*.

В общем случае проверку статистической гипотезы при помощи критерия значимости может быть разбита на следующие этапы:

1. Сформулировать проверяемую H_0 и альтернативную H_1, гипотезы.

2. Назначить уровень значимости α.

3. Выбрать статистику Z критерия для проверки гипотезы H_0.

4. Определить выборочное распределение статистики Z критерия при условии, что верна гипотеза H_0.

5. Определить критическую область V_k в зависимости от формулировки альтернативной гипотезы одним из неравенств: $Z > z_{1-\alpha}$ или $Z < z_{\alpha}$, или совокупностью неравенств:

$$\begin{cases} Z < z_{1-\frac{\alpha}{2}} \\ Z < z_{\frac{\alpha}{2}} \end{cases}$$

6. Получить выборку наблюдений и вычислить выборочное значение статистики z_B критерия.

7. Принять статистическое решение:

– если $z_B \in V_k$, отклонить гипотезу H_0 как не согласующуюся с результатами наблюдений;

если, $z_B \in V \setminus V_k$ принять гипотезу H_0, т. е. считать, что гипотеза H_0 не противоречит результатам наблюдений.

Если для проверки статистической гипотезы H_0 применяется критерий значимости то в соответствии с принципом проверки гипотез, гипотеза H_0 отклоняется при попадании статистики критерия в критическую область. Если, тем не менее, гипотеза H_0 верна, то принимаемое решение неверно. Ошибка, совершаемая при отклонении правильной гипотезы H_0, называется *ошибкой первого рода*. Очевидно, вероятность ошибки первого рода равна вероятности попадания

статистики в критическую область при условии, что верна гипотеза H_0, т. е. равна уровню значимости α (6.1.2).

Ошибка второго рода происходит в том случае, если гипотеза H_0 принимается, но в действительности верна альтернативная гипотеза H_1. Вероятность ошибки второго рода β можно вычислить (при простой альтернативной гипотезе H_1) по формуле [3]:

$$\beta = P[Z \in V \setminus V_k / H_1]. \qquad (6.1.4)$$

Хорошим примером является проверка непараметрической гипотезы о виде распределения по критерию χ^2.

Критерий χ^2 использует тот факт, что случайные величины

$$\frac{(n_k - np_k)}{\sqrt{np_k}},\ k = 1,2,...,r,$$

имеют распределения, близкие к нормальному *N(0,1)*. Чтобы это утверждение было достаточно точным, необходимо, чтобы для всех интервалов выполнялось условие $np_k \geq 5$. Если для некоторых интервалов это условие не выполняется, то их следует объединить с соседними.

Пусть $x_1, x_2, ...x_n -$ выборка наблюдений случайной величины *X*. ГТроверяется гипотеза H_0, утверждающая, что *X* имеет функцию распределения *F*(*x*).

Проверка гипотезы H_0 при помощи критерия χ^2 осуществляется по следующей схеме. По выборке наблюдений находят оценки неизвестных параметров предполагаемого закона распределения случайной величины *X*. Далее, область возможных значений случайной величины *X* разбивается на Δ множеств $\Delta_1, \Delta_2, ...\Delta r$, например, r – интервалов в случае, когда *X* – непрерывная случайная величина, или *г* групп, состоящих из отдельных значений, дискретной случайной величины *X*.

Пусть n_k – число элементов выборки, принадлежащих множеству Δ_k, $k = 1,2,..,r$. Очевидно, что $\sum_{k=1}^{r} n_k = n$. Используя предполагаемый закон распределения случайной величины *X*, находят вероятности p_k того, что значение *X* принадлежит множеству Δ_k, т. е. $p_k = P[X \in \Delta_k], k = 1,2,...,r$. Очевидно, что $\sum_{k=1}^{r} p_k = 1$.

Выборочное значение статистики критерия вычисляется по формуле:

$$\chi^2 = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k} \qquad (6.1.5)$$

Гипотеза H_0 согласуется с результатами наблюдений на уровне значимости α, если:

$$\chi^2 < \chi^2_{1-\alpha}(r-l-1), \qquad (6.1.5)$$

где $\chi^2_{1-\alpha}(r-l-1)$ – квантиль порядка $1-\alpha$ распределения χ^2 с $(r-l-1)$ степенями свободы, а l – число неизвестных параметров распределения, оцениваемых по выборке; если же:

$$\chi^2 \geq \chi^2_{1-\alpha}(r-l-1), \qquad (6.1.6)$$

то гипотеза H_0 отклоняется.

6.2. Непараметрические методы математической статистики

Основные методы математической статистики: оценка параметров распределения, проверка статистических гипотез, дисперсионный анализ применяются в предположении, что распределение генеральной совокупности известно. В частности *t-критерий* для сравнения средних двух генеральных совокупностей и однофакторный дисперсионный анализ для сравнения средних нескольких совокупностей пригодны только в случае нормального распределения последних. Однако часто встречаются данные для которых эти предположения не выполняются. Например, результаты социологических опросов обычно имеют форму ответов вида «да» или «нет» и представляются в виде таблиц, содержащих частоты положительных и отрицательных ответов. Традиционные методы математической статистики не могут быть использованы для обработки таких данных. В этих случаях используются непараметрические методы, т. е. методы независящие от распределения генеральной совокупности.

Непараметрические методы применяются для качественных данных, представленных в номинальной шкале и для данных, измеряемых в порядковой шкале (т. е. представленных в виде рангов), а также для количественных данных в том случае, когда распределение генеральной совокупности неизвестно.

При решении конкретной задачи необходимо выбрать тот или иной метод. Помощь в таком выборе может оказать следующая классификация непараметрических методов, используемых для проверки гипотезы о том, что анализируемые данные – это выборки из однородных генеральных совокупностей. Заметим, что понятие однородности генеральных совокупностей понимается достаточно широко: это могут быть генеральные совокупности, имеющие одну и ту же функцию распределения, либо совокупности, у которых совпадают характеристики положения (средние, медианы) и/или характеристики разброса (дисперсии).

Первым критерием для выбора метода является, очевидно, вид шкалы, в которой представлены исходные данные.

Вторым критерием является вид выборок (независимые или связанные) и их количество.

Поясним понятие связанной выборки. Если над каждым из n объектов или индивидуумов проводятся два наблюдения: одно до, а другое после некоторого воздействия (приема лекарства, обучения, рекламной компании, обработки тем или иным способом и т. д.), то результаты наблюдений представляют две связанные (зависимые) выборки объема n. В случае если каждый из n объектов подвергается k воздействиям, то результаты наблюдений представляют k связанных выборок объема n. Например, множество оценок, проставленных k судьями каждому из n спортсменов – это k связанных выборок объема n, измеренных в порядковой шкале. Таким образом, рассматриваемые ниже методы можно классифицировать следующим образом.

1. Исходные данные: две независимые выборки объемов n_1 и n_2. Проверяемая гипотеза H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

1) критерий серий Вальда–Вольфовица;

2) критерий Манна–Уитни;

3) двухвыборочный критерий Колмогорова–Смирнова.

2. Исходные данные: пары наблюдений $(x_i, y_i), i = 1, 2, ..., n$ двух признаков X и Y, измеренных в порядковых или количественных шкалах.

Проверяемая гипотеза H_0: признаки X и Y некоррелированны. Меры статистической зависимости: *ранговый коэффициент корреляции Спирмена*, *коэффициент корреляции τ Кендалла*.

3. Исходные данные: к независимых выборок объемов $n_1, n_2, ..., n_k$. Проверяемая гипотеза H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

1) однофакторный дисперсионный анализ Краскела–Уоллиса;

2) медианный критерий.

4. Исходные данные: две связанные выборки объемов n. Проверяемая гипотеза H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

1) критерий знаков;

2) критерий Вилкоксона.

5. Исходные данные: k связанных выборок объемов n. Проверяемая гипотеза H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

1) двухфакторный анализ Фридмана;

2) меры связи – коэффициент конкордации Кендалла.

6. Связанные выборки, измеряемые в номинальной шкале.

1) Исходные данные: две связанные выборки объемов *n* переменных *X* и *Y*, каждая из которых принимает два значения («0», «1» или «+», «–» или «да», «нет» и т. д.).

Проверяемая гипотеза H_0: эффект воздействия отсутствует.

Метод: критерий Макнимара.

2) Исходные данные: две связанные выборки объемов *n* переменных $X_1, X_2, ..., X_k$, каждая из которых принимает два значения.

Проверяемая гипотеза H_0: эффект воздействия отсутствует.

Метод: критерий Кокрена.

7. Выборки, измеряемые в номинальной шкале.

1) Исходные данные: выборки двух случайных переменных *X* и *Y*, каждая из которых принимает два значения.

Проверяемая гипотеза H_0: *X* и *Y* независимы.

Метод: анализ таблицы сопряженности 2×2 (*точный критерий Фишера*, *критерий* χ^2).

2) Исходные данные: выборки двух переменных *X* и *Y*, представленных в номинальных шкалах. *X* принимает *k* значений, *Y* – *r* значений.

Проверяемая гипотеза H_0: *X* и *Y* – независимы.

Метод: анализ таблицы сопряженности $k \times r$ (критерий χ^2).

Рассмотрение особенностей каждого отдельного метода выходит за рамки данного пособия, для получения более подробного описания можно обратится к литературным источникам, например [3–5].

6.3. Проверка гипотез в пакете STATISTICA

Статистическая значимость (*p*-уровень) результата представляет собой оцененную меру уверенности в его "истинности" (в смысле "репрезентативности выборки"). Выражаясь более технически, *p*-уровень (этот термин был впервые использован в работе Brownlee, 1960) это показатель, находящийся в убывающей зависимости от надежности результата. Более высокий *p*-уровень соответствует более низкому уровню доверия к найденной в выборке зависимости между переменными. Именно, *p*-уровень представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю популяцию. Например, *p*-уровень = .05 (т.е. 1/20) показывает, что имеется 5% вероятность, что найденная в выборке связь между переменными является лишь случайной особенностью данной выборки. Иными словами, если данная зависимость в популяции отсутствует, а вы многократно проводили

бы подобные эксперименты, то примерно в одном из двадцати повторений эксперимента можно было бы ожидать такой же или более сильной зависимости между переменными. (Отметим, что это не то же самое, что утверждать о заведомом наличии зависимости между переменными, которая в среднем может быть воспроизведена в 5% или 95% случаев; когда между переменными популяции существует зависимость, вероятность повторения результатов исследования, показывающих наличие этой зависимости называется статистической мощностью плана). Во многих исследованиях *p*-уровень .05 рассматривается как "приемлемая граница" уровня ошибки.

Не существует никакого способа избежать произвола при принятии решения о том, какой уровень значимости следует действительно считать "значимым". Выбор определенного уровня значимости, выше которого результаты отвергаются как ложные, является достаточно произвольным. На практике окончательное решение обычно зависит от того, был ли результат предсказан априори (т.е. до проведения опыта) или обнаружен апостериорно в результате многих анализов и сравнений, выполненных с множеством данных, а также на традиции, имеющейся в данной области исследований. Обычно во многих областях результат *p=0.05* является приемлемой границей статистической значимости, однако следует помнить, что этот уровень все еще включает довольно большую вероятность ошибки (5%). Результаты, значимые на уровне *p =0.01* обычно рассматриваются как статистически значимые, а результаты с уровнем *p=0.005* или *p=0.001* как высоко значимые. Однако следует понимать, что данная классификация уровней значимости достаточно произвольна и является всего лишь неформальным соглашением, принятым на основе практического опыта в той или иной области исследования.

Понятно, что чем больше число анализов вы проведете с совокупностью собранных данных, тем большее число значимых (на выбранном уровне) результатов будет обнаружено чисто случайно. Например, если вы вычисляете корреляции между 10 переменными (имеете 45 различных коэффициентов корреляции), то можно ожидать, что примерно два коэффициента корреляции (один на каждые 20) чисто случайно окажутся значимыми на уровне р .05, даже если переменные совершенно случайны и некоррелированы в популяции. Некоторые статистические методы, включающие много сравнений, и, таким образом, имеющие хороший шанс повторить такого рода ошибки, производят специальную корректировку или поправку на общее число сравнений. Тем не менее, многие статистические методы (особенно простые методы разведочного анализа данных) не предлагают какого-либо способа

решения данной проблемы. Поэтому исследователь должен с осторожностью оценивать надежность неожиданных результатов.

Рассмотрим следующий пример, заимствованный из Nisbett, et al., 1987. Имеются 2 больницы. Предположим, что в первой из них ежедневно рождается 120 детей, во второй только 12. В среднем отношение числа мальчиков, рождающихся в каждой больнице, к числу девочек 50/50. Однажды девочек родилось вдвое больше, чем мальчиков. Спрашивается, для какой больницы данное событие более вероятно? Ответ очевиден для статистика, однако, он не столь очевиден неискушенному. Конечно, такое событие гораздо более вероятно для маленькой больницы. Объяснение этого факта состоит в том, что вероятность случайного отклонения (от среднего) возрастает с уменьшением объема выборки.

Рассмотренный пример показывает, что если связь между переменными "объективно" слабая (т.е. свойства выборки близки к свойствам популяции), то не существует иного способа проверить такую зависимость кроме как исследовать выборку достаточно большого объема. Даже если выборка, находящаяся в вашем распоряжении, совершенно репрезентативна, эффект не будет статистически значимым, если выборка мала. Аналогично, если зависимость "объективно" (в популяции) очень сильная, тогда она может быть обнаружена с высокой степенью значимости даже на очень маленькой выборке. Рассмотрим пример. Представьте, что вы бросаете монету. Если монета слегка несимметрична, и при подбрасывании орел выпадает чаще решки (например, в 60% подбрасываний выпадает орел, а в 40% решка), то 10 подбрасываний монеты было бы не достаточно, чтобы убедить кого бы то ни было, что монета асимметрична, даже если был бы получен, казалось, совершенно репрезентативный результат: 6 орлов и 4 решки. Не следует ли отсюда, что 10 подбрасываний вообще не могут доказать что-либо? Нет, не следует, потому что если эффект, в принципе, очень сильный, то 10 подбрасываний может оказаться вполне достаточно для его доказательства. Представьте, что монета настолько несимметрична, что всякий раз, когда вы ее бросаете, выпадает орел. Если вы бросаете такую монету 10 раз, и всякий раз выпадает орел, большинство людей сочтут это убедительным доказательством того, что с монетой что-то не то. Другими словами, это послужило бы убедительным доказательством того, что в популяции, состоящей из бесконечного числа подбрасываний этой монеты орел будет встречаться чаще, чем решка. В итоге этих рассуждений мы приходим к выводу: если зависимость сильная, она может быть обнаружена с высоким уровнем значимости даже на малой выборке.

В качестве примера рассмотрим проверку гипотезы о виде распределении по критерию χ^2 (6.1.5). Для этого в пакете STATISTICA

необходимо выполнить следующие действия. В главном меню выбрать **Statistics→Distribution Fitting**, далее выбрать вид распределения (рис. 6.3.1), выбираем Normal. В появившемся окошке необходимо выбрать переменные (вектор), во вкладке **Parametrs** можно задать дополнительные параметры. Результатом вычислений будет таблица в заголовке которой будет содержаться значение статистики критерия (**Chi-Square**), число степеней свободы (**df**) и вычисленный уровень значимости (**p**). На основании величины (**p**) принимается или отклоняется гипотеза о нормальном распределении выбранных данных.

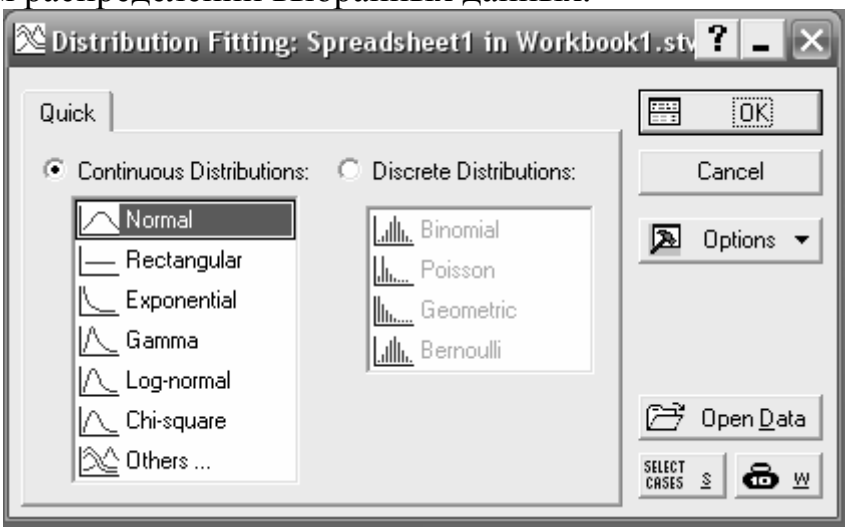

Рисунок 6.3.1. Окно подбор распределений.

В пакете STATTSTICA непараметрические процедуры выполняются в модуле **Statistics→Nonpametrics** (рис.6.3.2).

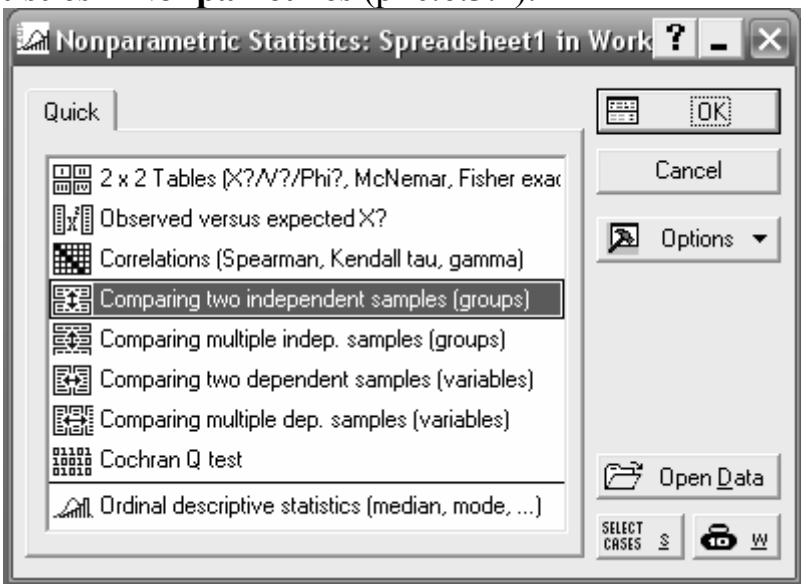

Рисунок 6.3.2. Непараметрические статистики.

В модуле **Nonpametrics** содержится значительное количество процедур. Рассмотрим сравнение двух независимых выборок при помощи критерия Манна-Уитни. После выбора соответствующего раздела (рис. 6.3.2), необходимо в появившемся окне (рис. 6.3.2), выбрать сравниваемые выборки и затем нажать кнопку соответствующую критерию (рис. 6.3.3), в нашем случае (**Mann-Whitney U-test**). В результаты расчётов будут представлены в виде таблицы на основании уровня значимости (**p-level**) принимается решение о принятии гипотезы H_0 о том, что две выборки получены из однородных генеральных совокупностей, и имеют равные средние и медианы.

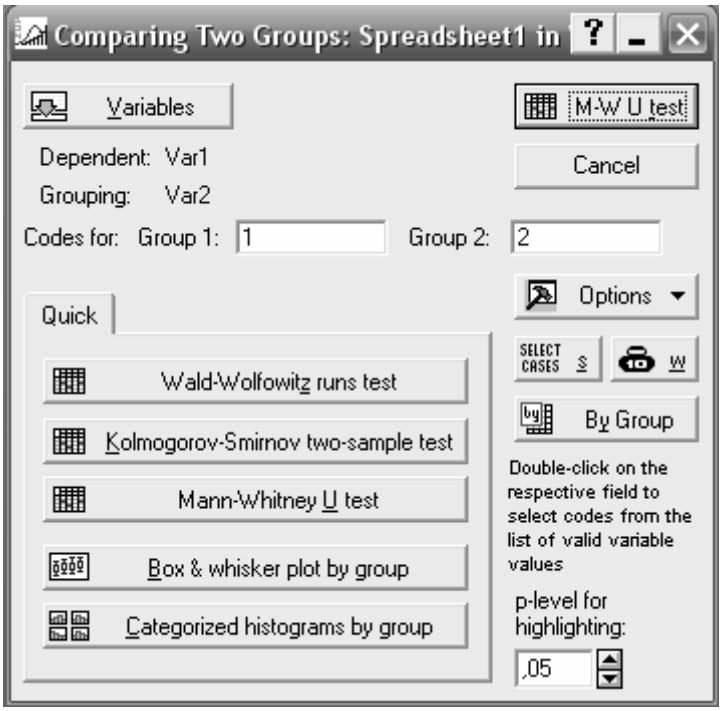

Рисунок 6.3.3. Сравнение двух выборок.

7. РЕГРЕССИОННЫЙ АНАЛИЗ

7.1. Понятие регрессии

Во многих случаях исследуются объекты, характеризующиеся несколькими признаками. Например, у каждого человека можно измерить рост, вес, частоту пульса и ряд других физиологических показателей; работу торгового предприятия можно оценивать по объему товарооборота и величине прибыли. Совокупность данных такого типа представляет выборку из многомерной генеральной совокупности. Для таких данных интерес представляет не только определение характеристик распределения каждого признака, но и то, насколько тесно эти признаки связаны между собой, можно ли по значению одного признака сделать какие-либо выводы о предполагаемом значении другого признака и т. д.

Большинство эмпирических исследований данных можно отнести к одному из названных выше типов. В исследовании корреляций (зависимостей, связей...) вы не влияете (или, по крайней мере, пытаетесь не влиять) на переменные, а только измеряете их и хотите найти зависимости (корреляции) между некоторыми измеренными переменными, например, между температурой эксплуатации электронной платы и временем наработки на отказ. В экспериментальных исследованиях, напротив, вы варьируете некоторые переменные и измеряете воздействия этих изменений на другие переменные. Например, исследователь может искусственно увеличивать температуру, а затем на определенных уровнях и достаточном для статистического анализа числе случаев измерять время наработки на отказ. Анализ данных в экспериментальном исследовании также приходит к вычислению "корреляций" (зависимостей) между переменными, а именно, между переменными, на которые воздействуют, и переменными, на которые влияет это воздействие. Тем не менее, экспериментальные данные потенциально снабжают нас более качественной информацией. Только экспериментально можно убедительно доказать причинную связь между переменными. Например, если обнаружено, что всякий раз, когда изменяется переменная А, изменяется и переменная В, то можно сделать вывод - "переменная А оказывает влияние на переменную В", т.е. между переменными А и В имеется причинная зависимость. Результаты корреляционного исследования могут быть проинтерпретированы в каузальных (причинных) терминах на основе некоторой теории, но сами по себе не могут отчетливо доказать причинность.

Зависимые и независимые переменные. Независимыми переменными называются переменные, которые варьируются исследователем, тогда как зависимые переменные - это переменные, которые измеряются или регистрируются. Может показаться, что проведение этого различия создает путаницу в терминологии, поскольку как говорят некоторые студенты "все переменные зависят от чего-нибудь". Тем не менее, однажды отчетливо проведя это различие, вы поймете его необходимость. Термины зависимая и независимая переменная применяются в основном в экспериментальном исследовании, где экспериментатор манипулирует некоторыми переменными, и в этом смысле они "независимы" от реакций, свойств, намерений и т.д. присущих объектам исследования. Некоторые другие переменные, как предполагается, должны "зависеть" от действий экспериментатора или от экспериментальных условий. Иными словами, зависимость проявляется в ответной реакции исследуемого объекта на посланное на него воздействие. Отчасти в противоречии с данным разграничением понятий находится использование их в исследованиях, где

вы не варьируете независимые переменные, а только приписываете объекты к "экспериментальным группам", основываясь на некоторых их априорных свойствах. Например, если в эксперименте мужчины сравниваются с женщинами относительно числа лейкоцитов (WCC), содержащихся в крови, то Пол можно назвать независимой переменной, а WCC зависимой переменной.

Вообще говоря, конечная цель всякого исследования или научного анализа состоит в нахождение связей (зависимостей) между переменными. Философия науки учит, что не существует иного способа представления знания, кроме как в терминах зависимостей между количествами или качествами, выраженными какими-либо переменными. Таким образом, развитие науки всегда заключается в нахождении новых связей между переменными. Исследование корреляций по существу состоит в измерении таких зависимостей непосредственным образом. Тем не менее, экспериментальное исследование не является в этом смысле чем-то отличным. Например, отмеченное выше экспериментальное сравнение WCC у мужчин и женщин может быть описано как поиск связи между переменными: Пол и WCC. Назначение статистики состоит в том, чтобы помочь объективно оценить зависимости между переменными.

Регрессионный анализ – это один из наиболее известных статистических методов, применяемых для решения задач такого рода. Основная цель регрессионного анализа состоит в определении связи между некоторой характеристикой *Y* наблюдаемого явления или объекта и величинами $x_1, x_2, ...x_n$, которые обусловливают, объясняют изменения *Y*. Переменная *Y* называется зависимой переменной (откликом), объясняющие переменные $x_1, x_2, ...x_n$ называются предикторами, регрессорами или факторами.

Например, при исследовании электронной сети нас может интересовать, как число ошибочных пакетов зависит от параметров сети. Для ответа на этот вопрос необходимо собрать данные о числе ошибочных пакетов при различных значениях параметров сети.

В данном случае нужно выяснить, как число ошибочных пакетов связано параметрами сети, какой фактор является наиболее важным при прогнозе числа ошибочных пакетов, имеется ли в исходных данных режимы работы сети, обладающие какими-либо специфическими свойствами (выбросы).

Если в рассматриваемом примере в качестве объясняющих факторов использовать только *m* определенных факторов $x_1, x_2, ...x_m$, то регрессионная модель может быть записана в виде:

$$Y = f(x_1, x_2, ...x_n) + \varepsilon \,, \qquad (7.1.1)$$

где: $f(x_1, x_2, ...x_n)$ – детерминированная составляющая отклика *Y*, зависящая от $f(x_1, x_2, ...x_n)$, а ε – случайная составляющая. Случайная составляющая может быть обусловлена множеством неучтённых факторов. И соответственно, чем больше факторов будет учтено тем меньше будет доля в сумме (7.1.1) случайной составляющей. При выборе влияющих факторов важно понимать «механику» происходящих процессов, что позволит выделить среди множества факторов те, что оказывают влияние, а среди них в свою очередь те, что оказывают существенное (значимое) влияние. Выбор таких факторов, можно производить, с применением регрессионного анализа в виде последовательных итераций, последовательно исключая факторы влияние которых не выявлено. При таком подходе необходимо соблюдать осторожность т.к. при некоторых обстоятельствах может возникнуть потеря существенных влияющих факторов, которые могут проявляться при возникновении определённых условий.

Часто (например, в пакете STATISTICA) объясняющие переменные $x_1, x_2, ...x_n$ называют независимыми переменными. Такое название во многих случаях не соответствует реальной ситуации: «независимые» переменные могут быть зависимы и влиять одна на другую. Часто понятие «независимые переменные» используется в другом контексте: это переменные, значения которых в процессе определения отклика, могут устанавливаться произвольно, независимо.

Существуют различные регрессионные модели, определяемые выбором:

простая линейная регрессия:

$$Y = \beta_0 + \beta_1 x + \varepsilon; \qquad (7.1.2)$$

множественная регрессия

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1} + \varepsilon; \qquad (7.1.3)$$

3. полиномиальная регрессия

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_{k-1} x^{k-1} + \varepsilon; \qquad (7.1.4)$$

4. регрессионная модель общего вида:

$$Y = \beta_0 + \beta_1 \varphi_1(x_1, x_2, ...x_n) + \beta_2 \varphi_2(x_1, x_2, ...x_n) + ... + \beta_{k-1} \varphi_{k-1}(x_1, x_2, ...x_n) + \varepsilon; \qquad (7.1.5)$$

где: $\varphi_i(x_1, x_2, ...x_n)$ – заданные функции факторов.

Коэффициенты $\beta_0, \beta_1, \beta_2, ...\beta_{k-1}$ называются *параметрами регрессии*. Модели (1-4) называют *линейными* (по параметрам) моделями, а математические методы анализа этих моделей – *линейным регрессионным анализом*.

Существуют *нелинейные регрессионные модели*, которые в некоторых случаях можно свести к линейным.

Примером нелинейной модели является логистическая функция:

$$y = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad (7.1.6)$$

которая после замены $y' = \ln\left(\frac{y}{1-y}\right)$ примет вид:

$$y' = \beta_0 + \beta_1 x\,.$$

После выбора вида регрессионной модели, используя результаты наблюдений зависимой переменной и факторов, нужно вычислить оценки (приближенные значения) параметров регрессии, а затем проверить значимость и адекватность модели результатам наблюдений.

Две основные черты всякой зависимости между переменными. Можно отметить два самых простых свойства зависимости между переменными: (a) величина зависимости и (b) надежность зависимости.

a.) *Величину* зависимости легче понять и измерить, чем надежность. Например, если любой мужчина в вашей выборке имел значение WCC выше чем любая женщина, то вы можете сказать, что зависимость между двумя переменными (Пол и WCC) очень высокая. Другими словами, вы могли бы предсказать значения одной переменной по значениям другой.

b.) *Надежность* взаимозависимости - менее наглядное понятие, чем величина зависимости, однако чрезвычайно важное. Надежность зависимости непосредственно связана с репрезентативностью определенной выборки, на основе которой строятся выводы. Другими словами, надежность говорит нам о том, насколько вероятно, что зависимость, подобная найденной вами, будет вновь обнаружена (иными словами, подтвердится) на данных другой выборки, извлеченной из той же самой популяции. Следует помнить, что конечной целью почти никогда не является изучение данной конкретной выборки; выборка представляет интерес лишь постольку, поскольку она дает информацию обо всей популяции. Если ваше исследование удовлетворяет некоторым специальным критериям, то надежность найденных зависимостей между переменными вашей выборки можно количественно оценить и представить с помощью стандартной статистической меры называемой *p*-уровень или статистический уровень значимости.

7.2. Простая линейная регрессия

Простая линейная регрессия – регрессионная модель, описывающая зависимость переменной Y одного фактора x.

Пусть $(x_i, y_i),\ i = 1,2,...n$ – выборка наблюдений из двумерной генеральной совокупности. Предварительное представление о зависимости между случайными величинами X и Y можно получить, отображая элементы выборки как точки на плоскости. Такое представление выборки называется *диаграммой рассеяния*.

При построении диаграммы рассеяния рекомендуется масштабы по осям *x* и *y* выбирать так, чтобы значения обоих признаков укладывались на отрезках приблизительно равной длины.

Возможны различные варианты расположения «облака» точек, по которым можно судить о виде и степени взаимосвязи между признаками *X* и *Y*.

Количественной характеристикой степени линейной зависимости между случайными величинами *X* и *Y* является коэффициент корреляции ρ.

$$\rho(X,Y) = \frac{Q_{xy}}{\sqrt{Q_x Q_y}}, \qquad (7.2.1)$$

где:

$$\begin{cases} Q_x = \sum \left(x_i - \bar{x}\right)^2 \\ Q_y = \sum \left(y_i - \bar{y}\right)^2 \\ Q_{xy} = \sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right). \\ \bar{x} = \frac{1}{n}\sum x_i \\ \bar{y} = \frac{1}{n}\sum y_i \end{cases}$$

Для коэффициента корреляции справедливо:

– $(-1 \le \rho \le 1)$;

– если $|\rho| = 1$, то между *X* и *Y* имеет место функциональная линейная зависимость;

– если $\rho = 0$, то говорят, что *X* и *Y* некоррелированны, т.е. между ними нет линейной зависимости;

– если *X* и *Y* имеют двумерное нормальное распределение, то из равенства $\rho = 0$ следует, что они статически независимы.

7.3. Множественная регрессия

Если переменная Y зависит от нескольких факторов $x_1, x_2, \ldots x_{k-1}$, то регрессионная модель определяется уравнением множественной регрессии:

$$\tilde{y}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon \ . \qquad (7.3.1)$$

Исходными данными для регрессионного анализа представляют собой результаты наблюдений зависимой переменной *Y* и факторов $x_1, x_2, \ldots x_{k-1}$.

При множественной регрессии, регрессионный анализ простой линейной регрессии обобщается. Для нахождения оценок параметров уравнения (7.3.1) применяют метод наименьших квадратов (МНК), т.е. ищут минимум отклонений зависимой переменной $\tilde{y}$ от фактического значения при выбранных параметрах уравнения (7.3.1), ищется минимум функции:

$$Q(\beta_0, \beta_1, \beta_2, ..., \beta_{k-1}) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)$$
$$\frac{\partial Q}{\partial \beta_i} = 0 \qquad (7.3.2)$$

Для проверки адекватности полученной модели результатам наблюдений надо найти остатки, т. е, разности между наблюдаемыми и предсказанными моделью значениями переменной *Y*. Вектор остатков равен:

$$e_i = y_i - \tilde{y}_i . \qquad (7.3.2)$$

Далее вычисляются статистики, на основе которых можно проверить выполнение основных предположений регрессионного анализа и адекватность полученной модели.

Для адекватной модели, кроме некоррелированности остатков, их нормального распределения, должно выполняться условие гомоскедаксичности, т. е. постоянства дисперсии ошибок наблюдений для всех наблюдений. Оценка выполнимости этого условия проводится по графику остатков в зависимости от номера наблюдений: если все остатки укладываются в симметричную относительно нулевой линии полосу, то, можно считать, что дисперсия ошибок наблюдений постоянна.

Более тщательная проверка адекватности регрессионной модели может быть проведена, если для зависимой переменной *Y* проведены повторные наблюдения. В этом случае для проверки адекватности модели используется следующая процедура дисперсионного анализа.

Пусть при *i*-м наборе независимых переменных проведено n_i, повторных наблюдений переменной Y, $i = 1, 2, ...m$. Объем всей выборки $n = \sum_i n_i$. Обозначим y_{ij}, $j = 1,2,...,n$, результаты повторных наблюдений Y при *j*-м наборе независимых переменных. Если модель адекватна данным, то среднее $\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n_i} y_{ij}$, должны быть близки к значениям $\tilde{y}_i$, предсказанным регрессионной моделью (7.3.1).

Мерой неадекватности модели будет сумма квадратов

$$Q_n = \sum n_i (\bar{y}_i - \tilde{y}_i) . \qquad (7.3.3)$$

Чем меньше будет Q_n, тем лучше результаты наблюдений согласуются с моделью.

Возведя обе части тождества $y_{ij} - \tilde{y}_i = (\bar{y}_i - \tilde{y}_i) + (y_{ij} - \bar{y}_i)$ в квадрат и просуммировав их по *i* и *j*, получим, что остаточная сумма квадрата может быть разбита на две суммы:

$$Q_e = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - \tilde{y}_i\right)^2 = \sum_{i=1}^{m} n_i \left(\bar{y}_i - \tilde{y}_i\right)^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 \qquad (7.3.4)$$

или

$$Q_e = Q_n + Q_p , \qquad (7.3.5)$$

где второе слагаемое называют суммой квадратов чистой ошибки. Если модель адекватна результатам наблюдений, то статистики $\frac{Q_n}{\sigma^2}$ и $\frac{Q_p}{\sigma^2}$ независимы и имеют распределение χ^2 соответственно с (*m*-*k*) и (*n*-*m*) степенями свободы, где *k* – число параметров в уравнении множественной регрессии.

В этом случае статистика:

$$F = \frac{Q_n /(m-k)}{Q_p /(n-m)}$$

имеет распределение Фишера с (*m*-*k*) и (*n*-*m*) степенями свободы.

Вычисленное значение статистики сравнивается с квантилью распределения Фишера $F_{1-\alpha}(m-k, n-m)$. Если $F < F_{1-\alpha}(m-k, n-m)$, то гипотеза об адекватности модели принимается на уровне значимости α.

Однозначный прогноз и частная корреляция. Регрессионные коэффициенты представляют независимые вклады каждой независимой переменной в предсказание зависимой переменной. Другими словами, переменная X_1, к примеру, коррелирует с переменной *Y* после учета влияния всех других независимых переменных. Этот тип корреляции упоминается также под названием частной корреляции (этот термин был впервые использован в работе Yule, 1907). Вероятно, следующий пример пояснит это понятие. Кто-то мог бы, вероятно, обнаружить значимую отрицательную корреляцию в популяции между длиной волос и ростом (невысокие люди обладают более длинными волосами). На первый взгляд это может показаться странным; однако, если добавить переменную Пол в уравнение множественной регрессии, эта корреляция, скорее всего, исчезнет. Это произойдет из-за того, что женщины, в среднем, имеют более длинные волосы, чем мужчины; при этом они также в среднем ниже мужчин. Таким образом, после удаления разницы по полу посредством ввода предиктора Пол в уравнение, связь между длиной волос и ростом

исчезает, поскольку длина волос не дает какого-либо самостоятельного вклада в предсказание роста помимо того, который она разделяет с переменной Пол. Другими словами, после учета переменной Пол частная корреляция между длиной волос и ростом нулевая. Иными словами, если одна величина коррелирована с другой, то это может быть отражением того факта, что они обе коррелированы с третьей величиной или с совокупностью величин.

Предсказанные значения и остатки. Линия регрессии выражает наилучшее предсказание зависимой переменной (*Y*) по независимым переменным (*X*). Однако, природа редко (если вообще когда-нибудь) бывает полностью предсказуемой и обычно имеется существенный разброс наблюдаемых точек относительно подогнанной прямой. Отклонение отдельной точки от линии регрессии (от предсказанного значения) называется остатком.

Остаточная дисперсия и коэффициент детерминации *R*-квадрат. Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем, очевидно, лучше прогноз. Например, если связь между переменными *X* и *Y* отсутствует, то отношение остаточной изменчивости переменной *Y* к исходной дисперсии равно 1.0. Если *X* и *Y* жестко связаны, то остаточная изменчивость отсутствует, и отношение дисперсий будет равно 0.0. В большинстве случаев отношение будет лежать где-то между этими экстремальными значениями, т.е. между 0.0 и $\pm$1.0 минус это отношение называется *R*-квадратом или коэффициентом детерминации. Это значение непосредственно интерпретируется следующим образом. Если имеется *R*-квадрат равный 0.4, то изменчивость значений переменной *Y* около линии регрессии составляет 1–0.4 от исходной дисперсии; другими словами, 40% от исходной изменчивости могут быть объяснены, а 60% остаточной изменчивости остаются необъясненными. В идеале желательно иметь объяснение если не для всей, то хотя бы для большей части исходной изменчивости. Значение *R*-квадрата является индикатором степени подгонки модели к данным (значение *R*-квадрата близкое к 1.0 показывает, что модель объясняет почти всю изменчивость соответствующих переменных).

Интерпретация коэффициента множественной корреляции *R*. Обычно, степень зависимости двух или более предикторов (независимых переменных или переменных *X*) с зависимой переменной (*Y*) выражается с помощью коэффициента множественной корреляции *R*. По определению он равен корню квадратному из коэффициента детерминации. Это неотрицательная величина, принимающая значения между 0 и 1. Для интерпретации направления связи между переменными смотрят на знаки (плюс или минус) регрессионных коэффициентов. Если регрессионный

коэффициент положителен, то связь этой переменной с зависимой переменной положительна (например, чем больше IQ, тем выше средний показатель успеваемости оценки); если отрицателен, то и связь носит отрицательный характер (например, чем меньше число учащихся в классе, тем выше средние оценки по тестам). Если регрессионный коэффициент равен нулю, связь между переменными отсутствует [8]. Хотя большинство предположений множественной регрессии нельзя в точности проверить, исследователь может обнаружить отклонения от этих предположений. В частности, выбросы (т.е. экстремальные наблюдения) могут вызвать серьезное смещение оценок, "сдвигая" линию регрессии в определенном направлении и тем самым, вызывая смещение регрессионных коэффициентов. Часто исключение всего одного экстремального наблюдения приводит к совершенно другому результату.

7.4. Регрессионный анализ в пакете STATISTICA

Пакет STATISTICA позволяет производить регрессионный анализ. Простую однофакторную регрессию можно вычислить используя раздел главного меню **Statistics→Basic Statistics/Tables** в появившемся окне (рис. 7.4.1) необходимо выбрать раздел **Correlation matrices**, далее выбрать векторы между которыми будет произведён расчёт корреляции. Результат вычислений можно посмотреть в виде корреляционной матрицы или построить корреляционные диаграммы, в том числе трёхмерные (для трёх векторов).

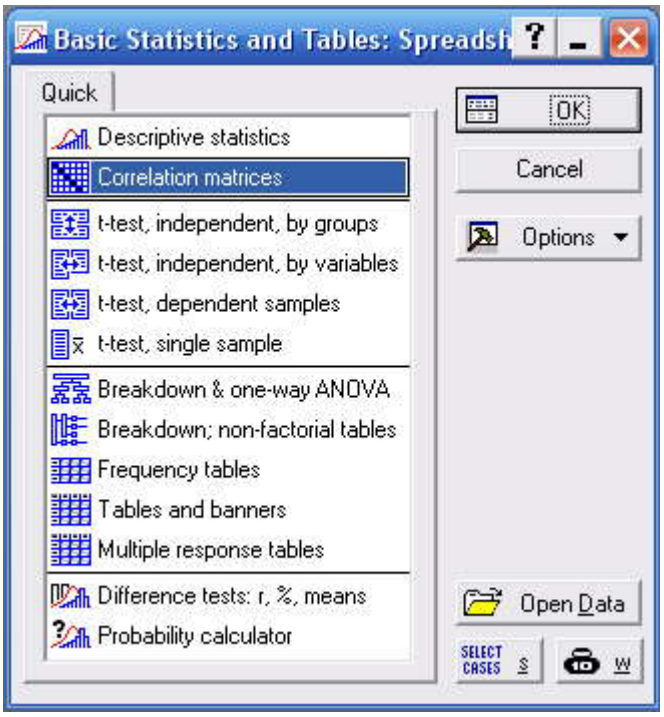

Рисунок 7.4.1. Корреляционный анализ.

Анализ множественной регрессии можно произвести при помощи раздела **Statistics→Multiple Regression.** В появившемся окне (7.4.2) необходимо выбрать зависимые и независимые переменные, в случае необходимости указать дополнительные опции, произвести расчёт. Результатом расчёта буде матрица содержащая значения регрессионных коэффициентов и уровней значимости для каждого фактора.

Рисунок 7.4.2. Множественная регрессия.

Так же сформированном в окне после вычислений (рис. 7.4.3) во вкладке **Residuals/assumptions/prediction** возможно сделать ряд полезных операций, например, произвести прогнозирование значений на основе построенного уравнения регрессии.

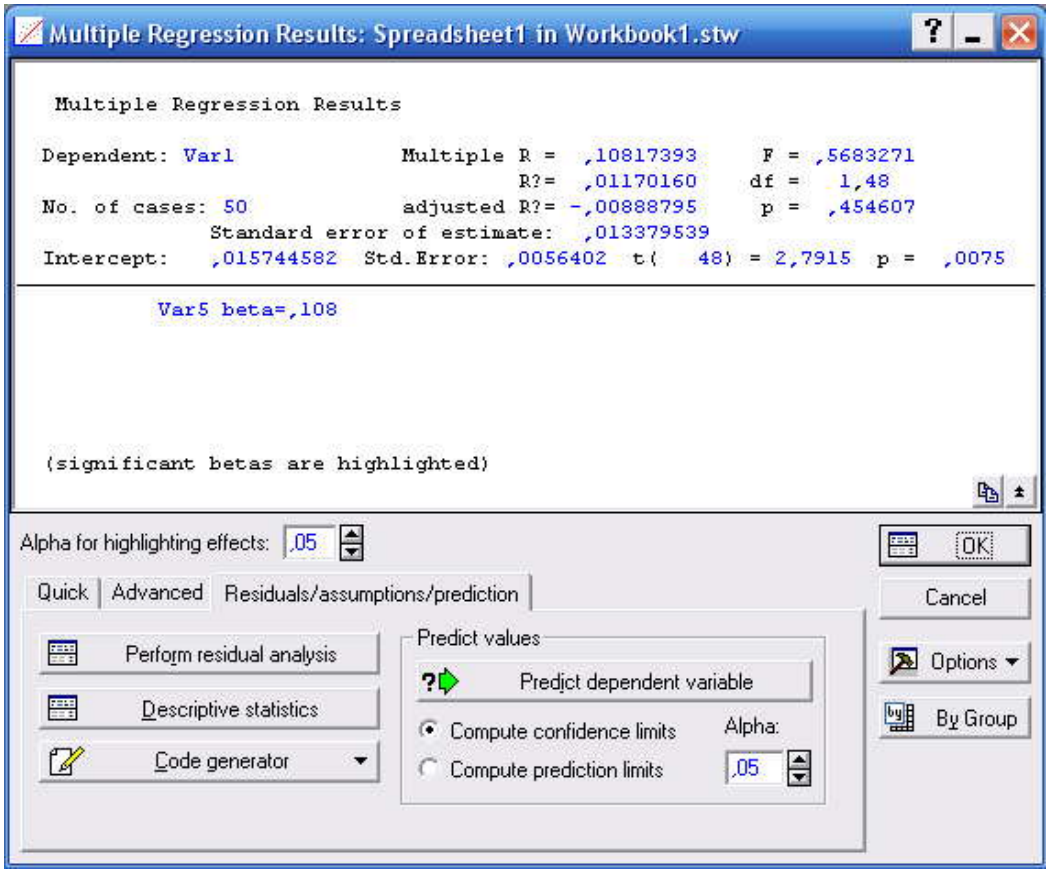

Рисунок 7.4.3. Результирующее окно расчёта множественной регрессии.

8. КЛАСТЕРНЫЙ АНАЛИЗ

8.1. Кластерный анализ, основные понятия

Термин *кластерный анализ* (впервые ввел Tryon, 1939) в действительности включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры, т.е. развернуть таксономии. Например, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. Заметьте, что в этой классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. Человек имеет больше сходства с другими приматами (т.е. с обезьянами), чем с "отдаленными" членами семейства млекопитающих (например, собаками) и т.д.

Заметим, что предыдущие рассуждения ссылаются на алгоритмы кластеризации, но ничего не упоминают о проверке статистической значимости. Фактически, кластерный анализ является не столько обычным статистическим методом, сколько "набором" различных алгоритмов

"распределения объектов по кластерам". Существует точка зрения, что в отличие от многих других статистических процедур, методы кластерного анализа используются в большинстве случаев тогда, когда вы не имеете каких-либо априорных гипотез относительно классов, но все еще находитесь в описательной стадии исследования. Следует понимать, что кластерный анализ определяет "наиболее возможно значимое решение". Поэтому проверка статистической значимости в действительности здесь неприменима, даже в случаях, когда известны *p*-уровни (как, например, в методе K средних).

Техника кластеризации применяется в самых разнообразных областях. Хартиган (Hartigan, 1975) дал прекрасный обзор многих опубликованных исследований, содержащих результаты, полученные методами кластерного анализа. Например, в области медицины кластеризация заболеваний, лечения заболеваний или симптомов заболеваний приводит к широко используемым таксономиям. В археологии с помощью кластерного анализа исследователи пытаются установить таксономии каменных орудий, предметов быта и т.д. Широко применяется кластерный анализ при климатических исследованиях. Известны широкие применения кластерного анализа в маркетинговых исследованиях. В общем, всякий раз, когда необходимо классифицировать "горы" информации к пригодным для дальнейшей обработки группам, кластерный анализ оказывается весьма полезным и эффективным.

Пусть $X_1, X_2, ...X_n$ – исходная совокупность объектов, каждый из которых задан набором p признаков. Например, объектами могут быть пациенты клиники, а признаками – физические данные (вес, давление и т. д.) и результаты амбулаторного обследования каждого пациента (содержание сахара в крови, уровень гемоглобина и т. д.)

Задача кластерного анализа состоит в разбиении исходной совокупности объектов на группы схожих, близких между собой объектов. Эти группы называют кластерами или таксонами.

Другими словами, *кластерный анализ* это один из способов классификации объектов по их признакам. Желательно, чтобы результаты классификации имели содержательную интерпретацию.

Результаты, полученные методами кластерного анализа применяются в самых разнообразных областях. Например, в области медицины кластеризация заболеваний и симптомов заболеваний приводит к классификациям используемым для выбора методов лечения. Кластерный анализ широко применяется в маркетинговых исследованиях. В общем, всякий раз, когда необходимо классифицировать большое количество информации такого рода и представить ее в виде пригодным

для дальнейшей обработки кластерный анализ оказывается весьма полезным и эффективным.

Фактически, кластерный анализ является «набором» различных алгоритмов «распределения объектов по кластерам».

В настоящее время известно огромное число алгоритмов кластеризации. Их разнообразие объясняется не только разными вычислительными методами, но и различными концепциями, лежащими в основе кластеризации.

Одна из концепций состоит в построении разбиения исходного множества объектов доставляющего оптимальное значение определенной целевой функции. Большая группа методов кластеризации использует в качестве целевой функции внутригрупповую сумму квадратов: разбиение каждого множества должно быть таково, чтобы оно минимизировало внутригрупповые суммы квадратов. Эти методы используют евклидову метрику и называются методами минимальной дисперсии.

Большинство алгоритмов кластеризации основано на использовании эвристических методов. Дать рекомендации для выбора того или иного метода кластеризации можно только в общих чертах, а основной критерий выбора – практическая полезность результата.

Пусть $X_1, X_2, ...X_n$ – объекты, каждый из которых задан набором *p* признаков. Распределения объектов по кластерам на однородные в некотором смысле группы должно удовлетворять критерию оптимальности, который выражается в терминах расстояния $\rho(X_i, X_j)$ между любой парой объектов рассматриваемой совокупности.

В качестве расстояния (метрики) может быть взята любая неотрицательная действительная функция $\rho(X_i, X_j)$, определенная на множестве $X_1, X_2, ...X_n$ и удовлетворяющая следующим условиям:

а) $\rho(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$;

б) $\rho(X_i, X_j) = \rho(X_j, X_i)$;

в) $\rho(X_i, X_j) \le \rho(X_i, X_k) + \rho(X_k, X_j)$.

Выбор расстояния между объектами неоднозначен и в этом состоит основная сложность.

Наиболее популярной метрикой является евклидова. Эта метрика отвечает интуитивным представлениям близости. При этом на расстояние между объектами могут сильно влиять изменения масштабов (единиц измерения) по осям. Например, если один из признаков измерен в метрах, а затем его значение переведены в сантиметры (т. е. умножены на 100), то евклидово расстояние между объектами сильно изменится и это приведет к тому, что результаты кластерного анализа могут значительно отличаться от предыдущих.

Если признаки измерены в разных единицах измерения, то требуется их предварительная нормировка – такое преобразование исходных данных, которое переводит их в безразмерные величины.

Наиболее известные способы нормировки следующие:

$$z_1 = \frac{x - \bar{x}}{\sigma},\ z_2 = \frac{x}{\bar{x}},\ z_3 = \frac{x}{x'},\ z_4 = \frac{x}{x_{\max}},\ z_5 = \frac{x - \bar{x}}{x_{\max} - x_{\min}}. \qquad (8.1.1)$$

где z_i; i= *1, 2, ..., 5* — нормированное значение; x – исходное значение; $\bar{x}$ и σ – соответственно среднее и среднее квадратическое отклонение; x' – эталонное (нормативное) значение; $x_{\min}, x_{\max}$ – наименьшее и наибольшее значение x.

Рассмотрим некоторые наиболее употребительные метрики (в скобках указано английское обозначение некоторых метрик, используемых в пакете STATISTICA в опции **Distance measure**).

1. *Евклидова метрика* (**Euclidean distance**):

$$\rho_E(X_i, X_j) = \left[\sum_{k=1}^{p}\left(X_{ki} - X_{kj}\right)\right]^{\frac{1}{2}}, \qquad (8.1.2)$$

где: X_{ki} – значение k-го признака i-го объекта.

2. «Взвешенная» евклидова метрика:

$$\rho_{WE}(X_i, X_j) = \left[\sum_{k=1}^{p} W_k\left(X_{ki} - X_{kj}\right)^2\right]^{\frac{1}{2}}, \qquad (8.1.3)$$

где: W_k – «вес» $к$-то признака. Применяется в тех случаях, когда каждому признаку можно приписать «вес», пропорциональный степени важности данного признака в задаче классификации. Цель «взвешивания» признака состоит в том, чтобы обеспечить максимальную дискриминирующую способность признака для разделения на кластеры.

3. l_m – *нормы*:

$$\rho_m(X_i, X_j) = \left[\sum_{k=1}^{p}\left|X_{ki} - X_{kj}\right|^m\right]^{\frac{1}{m}}. \qquad (8.1.4)$$

В частности, при m = *1* получаем меру l_1 – манхэттоновское расстояние или расстояние городских кварталов (**City-block (Manhatten) distance**).

Это расстояние является просто средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для евклидова расстояния. Однако отметим, что для этой метрики влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

4. Супремум – норма (*расстояние Чебышева* (**Chebychev distance metric**)):

$$\rho_{ch}(X_i, X_j) = \sup_{k=1,2,...p} \left\{ \left| X_{ki} - X_{kj} \right| \right\}. \qquad (8.1.5)$$

Это расстояние может оказаться полезным, если желают определить два объекта как «различные», если они различаются по какому-либо одному признаку.

5. Степенное расстояние (**Power**):

$$\rho_p(X_i, X_j) = \left[\sum_{k=1}^{p} \left(X_{ki} - X_{kj} \right)^m \right]^{\frac{1}{r}}, \qquad (8.1.2)$$

где: m, r – параметры задаваемые пользователем.

Процедуры классификации на основе кластерного анализа часто применяют для кластеризации классов (множества объектов).

8.2. Кластерный анализ в пакете STATSTICA

Пакет STATISTICA предоставляет значительные возможности для кластерного анализа. В модуле **Cluster Analysis** реализуются следующие методы кластеризации:

- соединения (древовидная кластеризация), **Joining** (**tree clustering**);
- метод К-средних (**K-means clustering**);
- двухвходовое объединение (**Two-way joining**).

Первая опция (**Joining**) представляет группу так называемых иерархических алгоритмов кластеризации. В основе этих алгоритмов лежит идея последовательной кластеризации. Пусть исходное множество содержит *n* объектов $X_1, X_2, ...X_n$

В качестве расстояния между объектами X_i, X_j выбирается некоторая метрика $\rho(X_i, X_j)$. Выбор метрики необходимо сделать в опции **distance measure** панели **Joining**.

На начальном шаге каждый объект рассматривается как отдельный кластер. На следующем шаге некоторые из ближайших друг к другу кластеров будут объединяться в один новый кластер. В зависимости от выбора меры, по которой определяется расстояние между кластерами, реализуются следующие методы объединения объектов в кластеры (выбор осуществляется в зависимости от меры расстояния между кластерами в опции: **Amalgamation** (**linkage**) **rule**).

1. *Метод одиночной связи (**Single Linkage**).* Кластеры объединяются исходя из расстояния, измеряемого по методу «ближайшего соседа». Группы, между которыми расстояния самые маленькие, объединяются. Каждое объединение уменьшает число групп на единицу. Расстояние между группами определяется как расстояние между ближайшими членами групп. Метод приводит к «цепным» кластерам.

2. *Метод полной связи (**Complete Linkage).*** Расстояние между группами определяется как расстояние, измеряемое по принципу «дальнего соседа». Расстояние между объединяемыми кластерами равно диаметру наименьшей сферы, содержащей оба кластера. Метод создает компактные кластеры в виде гиперсфер, которые плохо объединяются с другими кластерами. Если кластеры имеют удлиненную форму, то метод не работает.

3. *Метод невзвешенного попарного среднего (**Unweightedpair-group average).*** Расстояние между кластерами определяется по принципу «средней связи».

4. *Метод взвешенного попарного среднего (**Weighted pair-group average).*** Расстояние между кластерами определяется по принципу «средней связи», но с учетом в качестве весов числа объектов, содержащихся в кластерах.

5. *Невзвешенный центроидный метод (**Unweighted pair-group centroid).*** Расстояния между кластерами определяется как расстояние между их «центрами тяжести».

6. *Взвешенный центроидный метод (**Weighted pair-group centroid**).* Расстояние между классами определяется как расстояние между их «центрами тяжести», но с учетом весов, определяемых по количеству объектов в каждом кластере (т. е. с учетом размеров кластеров).

7. *Метод Уорда* **(Ward's metod). В** этом методе в качестве целевой функции используется сумма квадратов расстояний между каждым элементом и «центром тяжести» класса, содержащего этот элемент. Кластеризация представляет последовательную процедуру, на каждом шаге которой объединяются два таких класса, при объединении которых происходит минимизация статистического расстояния между классами.

Рассмотрим горизонтальную древовидную диаграмму (рис. 8.2.1). Диаграмма начинается с каждого объекта в классе (в левой части диаграммы). Теперь представим себе, что постепенно (очень малыми шагами) вы "ослабляете" ваш критерий о том, какие объекты являются уникальными, а какие нет. Другими словами, вы понижаете порог, относящийся к решению об объединении двух или более объектов в один кластер.

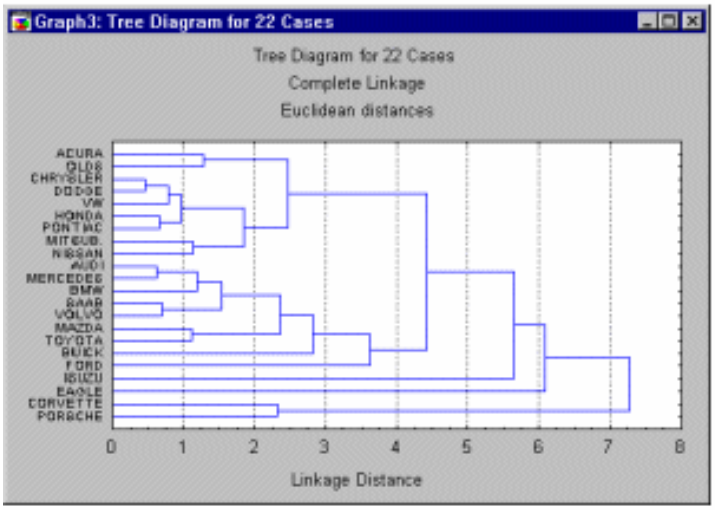

Рисунок 8.2.1. Древовидная диаграмма.

В результате, вы связываете вместе всё большее и большее число объектов и агрегируете (объединяете) все больше и больше кластеров, состоящих из все сильнее различающихся элементов. Окончательно, на последнем шаге все объекты объединяются вместе. На этих диаграммах горизонтальные оси представляют расстояние объединения (в вертикальных древовидных диаграммах вертикальные оси представляют расстояние объединения). Так, для каждого узла в графе (там, где формируется новый кластер) вы можете видеть величину расстояния, для которого соответствующие элементы связываются в новый единственный кластер. Когда данные имеют ясную "структуру" в терминах кластеров объектов, сходных между собой, тогда эта структура, скорее всего, должна быть отражена в иерархическом дереве различными ветвями. В результате успешного анализа методом объединения появляется возможность обнаружить кластеры (ветви) и интерпретировать их.

9. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

9.1. Временные ряды, основные понятия

Временным рядом называется последовательность наблюдений, упорядоченная по времени: $x_1, x_2, ...x_n$, где x_i — числа, представляющие наблюдения некоторой переменной в n равностоящих моментов времени $t= 1, 2, ..., n$. Примерами данных, которые необходимо изучать во времени являются: цены на товар, деловая активность, национальный валовой продукт. Особенностью, выделяющей анализ временных рядов, является зависимость данных, причем характер этой зависимости может определяться положением наблюдений в последовательности.

Основные задачи анализа:

1) прогнозирование, на основе знания прошлого;

2) сжатое описание характерных особенностей ряда;

3) управление процессом, порождающим ряд.

В теории временных рядов разработаны различные методы исследования и анализа: корреляционный и спектральный анализ, методы сглаживания и фильтрации, модели авторегрессии и скользящего среднего.

В анализе временных рядов, как и в большинстве статистических методов, предполагается, что исходные данные содержат детерминированную и случайную составляющие. В общем случае детерминированная составляющая может быть представлена в виде комбинации следующих компонент:

а) тренда определяющего главную тенденцию временного ряда;

б) более или менее регулярных колебаний относительного тренда – циклов;

в) периодических колебаний; такие колебания называются сезонной составляющей.

Временной ряд может быть представлен различными математическими моделями.

Пусть u_i – тренд, W_i, S_i, ε_i – соответственно циклическая, сезонная и случайная остаточная составляющие.

Аддитивная модель записывается в виде:

$$x_i = u_i + W_i + S_i + \varepsilon_i . \qquad (9.1.1)$$

Мультипликативная модель имеет вид:

$$x_i = u_i \cdot W_i \cdot S_i \cdot \varepsilon_i . \qquad (9.1.2)$$

Мультипликативная модель при переходе к логарифмам сводится к аддитивной модели.

Если предположить, что сезонная составляющая *S*, пропорциональна сумме тренда и циклической составляющей:

$$(u_i = W_i),\ S_i = (u_i + W_i)C_i$$

то временной ряд будет представлен в виде *смешанной модели*:

$$x_i = (u_i + W_i)(1 + C_i) + \varepsilon_i . \qquad (9.1.3)$$

Выбор модели зависит от конкретной совокупности явлений, определяющих данный временной ряд и их взаимосвязей.

Представление временного ряда в виде той или иной композиции его компонент естественно приводит к идее последовательного выделения этих компонент и прогнозирования на основе полученной модели.

Хорошим примером временных рядов является плотность потока солнечного радиоизлучения (рис. 9.1.1).

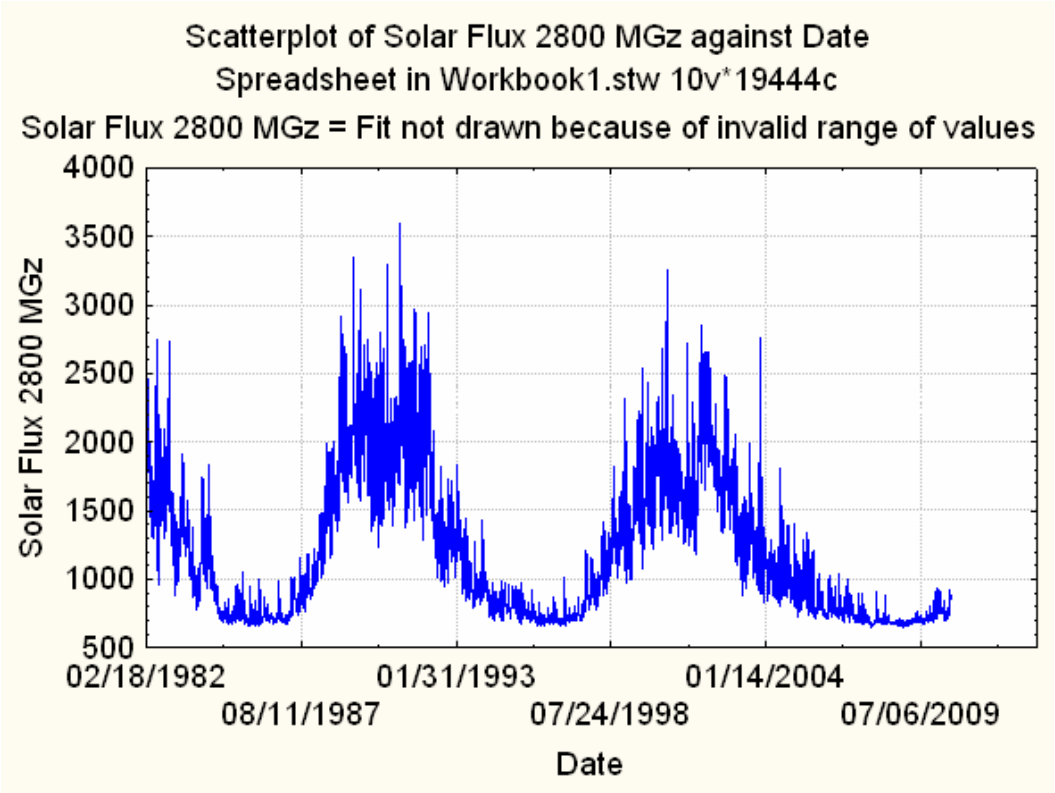

Рисунок 9.1.1. Временной ряд плотности потока солнечного радиоизлучения на частоте 2800 МГц.

9.2. Анализ временных рядов в пакете STATISTICA

Пакет STATISTICA позволяет производить анализ временных рядов в модуле **Statistics→Advanced Linear/Nonlinear Modes→Time Series Analysis (**рис. 9.1.1**)**.

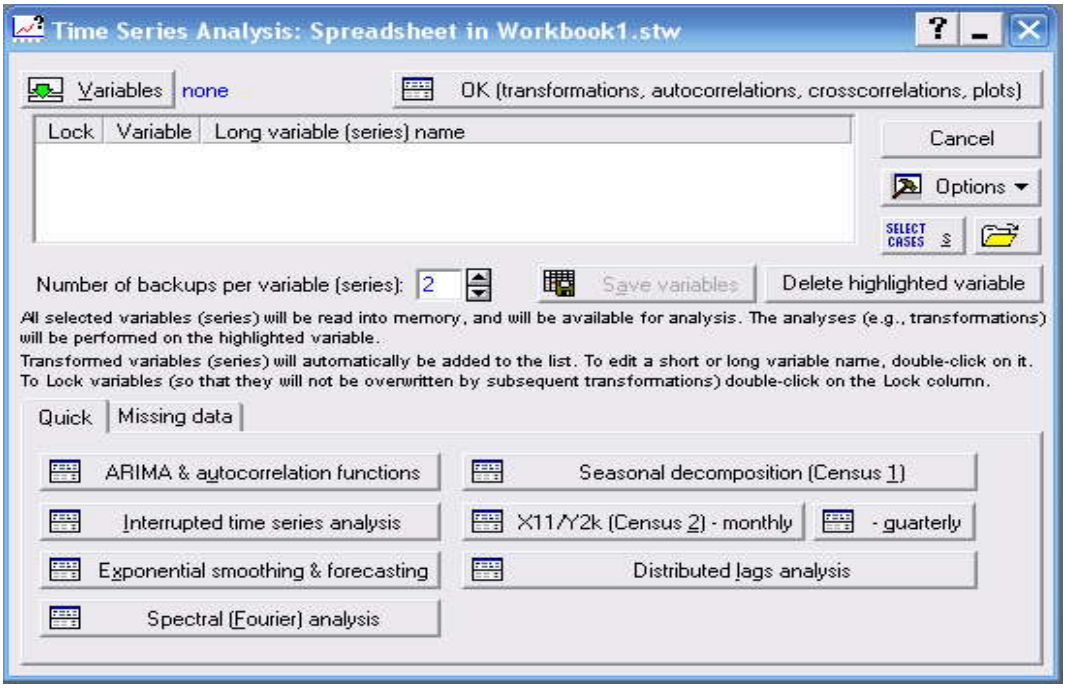

Рисунок 9.2.1. Окно модуля анализ временных рядов.

Данный модуль позволяет производить анализ временных рядов и разделение их на составляющие в соответствии с выражениями (9.1.1–9.1.3). Производить прогнозирование, фильтрацию (детрендирование, выделение сезонной составляющей и случайной компоненты) преобразование временных рядов.

Большинство регулярных составляющих временных рядов принадлежит к двум классам: они являются либо трендом, либо сезонной составляющей. Тренд представляет собой общую систематическую линейную или нелинейную компоненту, которая может изменяться во времени. Сезонная составляющая - это периодически повторяющаяся компонента. Оба эти вида регулярных компонент часто присутствуют в ряде одновременно (рис. 9.2.2). Например, продажи компании могут возрастать из года в год, но они также содержат сезонную составляющую (как правило, 25% годовых продаж приходится на декабрь и только 4% на август).

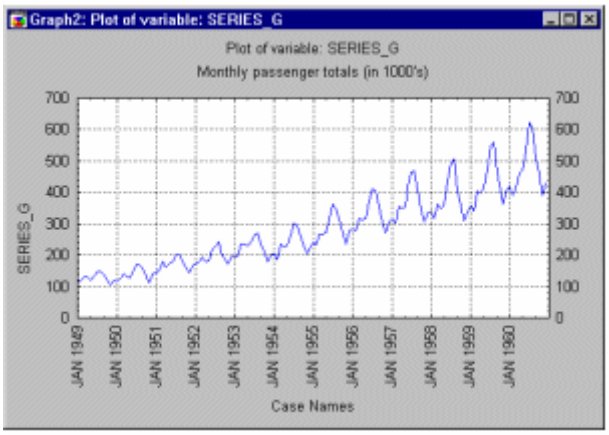

Рисунок 9.2.2. График временного ряда.

Эту общую модель можно понять на "классическом" ряде - Ряд G (Бокс и Дженкинс, 1976, стр. 531), представляющем месячные международные авиаперевозки (в тысячах) в течение 12 лет с 1949 по 1960 (см. файл Series_g.sta). График месячных перевозок ясно показывает почти линейный тренд, т.е. имеется устойчивый рост перевозок из года в год (примерно в 4 раза больше пассажиров перевезено в 1960 году, чем в 1949). В то же время характер месячных перевозок повторяется, они имеют почти один и тот же характер в каждом годовом периоде (например, перевозок больше в отпускные периоды, чем в другие месяцы). Этот пример показывает довольно определенный тип модели временного ряда, в которой амплитуда сезонных изменений увеличивается вместе с трендом. Такого рода модели называются моделями с мультипликативной сезонностью.

Не существует "автоматического" способа обнаружения тренда в временном ряде. Однако если тренд является монотонным (устойчиво возрастает или устойчиво убывает), то анализировать такой ряд обычно нетрудно. Если временные ряды содержат значительную ошибку, то первым шагом выделения тренда является *сглаживание*.

Сглаживание всегда включает некоторый способ локального усреднения данных, при котором несистематические компоненты взаимно погашают друг друга. Самый общий метод сглаживания - скользящее среднее, в котором каждый член ряда заменяется простым или взвешенным средним n соседних членов, где n - ширина "окна" (см. Бокс и Дженкинс, 1976; Velleman and Hoaglin, 1981). Вместо среднего можно использовать медиану значений, попавших в окно. Основное преимущество медианного сглаживания, в сравнении со сглаживанием скользящим средним, состоит в том, что результаты становятся более устойчивыми к выбросам (имеющимся внутри окна). Таким образом, если в данных имеются выбросы (связанные, например, с ошибками

измерений), то сглаживание медианой обычно приводит к более гладким или, по крайней мере, более "надежным" кривым, по сравнению со скользящим средним с тем же самым окном. Основной недостаток медианного сглаживания в том, что при отсутствии явных выбросов, он приводит к более "зубчатым" кривым (чем сглаживание скользящим средним) и не позволяет использовать веса.

Относительно реже, когда ошибка измерения очень большая, используется метод сглаживания методом наименьших квадратов, взвешенных относительно расстояния или метод отрицательного экспоненциально взвешенного сглаживания. Все эти методы отфильтровывают шум и преобразуют данные в относительно гладкую кривую (см. соответствующие разделы, где каждый из этих методов описан более подробно). Ряды с относительно небольшим количеством наблюдений и систематическим расположением точек могут быть сглажены с помощью бикубических сплайнов.

Многие монотонные временные ряды можно хорошо приблизить линейной функцией. Если же имеется явная монотонная нелинейная компонента, то данные вначале следует преобразовать, чтобы устранить нелинейность. Обычно для этого используют логарифмическое, экспоненциальное или (менее часто) полиномиальное преобразование данных.

Периодическая и сезонная зависимость (сезонность) представляет собой другой общий тип компонент временного ряда. Это понятие было проиллюстрировано ранее на примере авиаперевозок пассажиров. Можно легко видеть, что каждое наблюдение очень похоже на соседнее; дополнительно, имеется повторяющаяся сезонная составляющая, это означает, что каждое наблюдение также похоже на наблюдение, имевшееся в том же самом месяце год назад. В общем, периодическая зависимость может быть формально определена как корреляционная зависимость порядка k между каждым i-м элементом ряда и (i-k)-м элементом (Kendall, 1976). Ее можно измерить с помощью автокорреляции (т.е. корреляции между самими членами ряда); k обычно называют лагом (иногда используют эквивалентные термины: сдвиг, запаздывание). Если ошибка измерения не слишком большая, то сезонность можно определить визуально, рассматривая поведение членов ряда через каждые k временных единиц.

Сезонные составляющие временного ряда могут быть найдены с помощью *автокоррелограммы*. Коррелограмма (автокоррелограмма) показывает численно и графически автокорреляционную функцию (АКФ), иными словами коэффициенты автокорреляции (и их стандартные ошибки) для последовательности лагов из определенного диапазона

(например, от 1 до 30). На коррелограмме обычно отмечается диапазон в размере двух стандартных ошибок на каждом лаге, однако обычно величина автокорреляции более интересна, чем ее надежность, потому что интерес в основном представляют очень сильные (а, следовательно, высоко значимые) автокорреляции.

Исследование коррелограмм. При изучении коррелограмм следует помнить, что автокорреляции последовательных лагов формально зависимы между собой. Рассмотрим следующий пример. Если первый член ряда тесно связан со вторым, а второй с третьим, то первый элемент должен также каким-то образом зависеть от третьего и т.д. Это приводит к тому, что периодическая зависимость может существенно измениться после удаления автокорреляций первого порядка, т.е. после взятия разности с лагом (рис. 9.2.3).

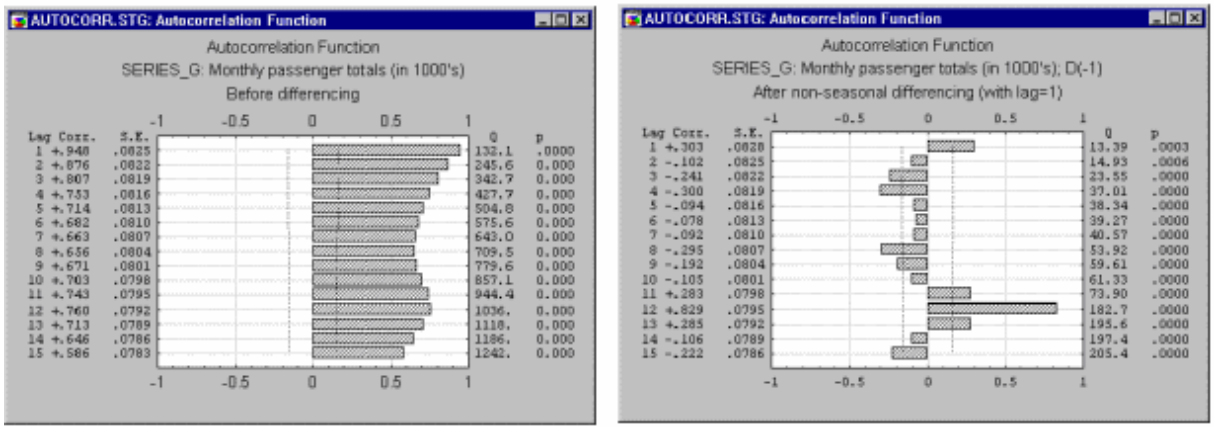

Рисунок 9.2.3. Автокоррелограмма до и после взятия разности ряда.

Другой полезный метод исследования периодичности состоит в исследовании частной автокорреляционной функции (ЧАКФ), представляющей собой углубление понятия обычной автокорреляционной функции. В ЧАКФ устраняется зависимость между промежуточными наблюдениями (наблюдениями внутри лага). Другими словами, частная автокорреляция на данном лаге аналогична обычной автокорреляции, за исключением того, что при вычислении из нее удаляется влияние автокорреляций с меньшими лагами (см. Бокс и Дженкинс, 1976; см. также McDowall, McCleary, Meidinger, and Hay, 1980). Частная автокорреляция дает более "чистую" картину периодических зависимостей.

Удаление периодической зависимости. Как отмечалось выше, периодическая составляющая для данного лага k может быть удалена взятием разности соответствующего порядка. Это означает, что из каждого

i-го элемента ряда вычитается (*i*-*k*)-й элемент. Имеются два довода в пользу таких преобразований.

Во-первых, таким образом можно определить скрытые периодические составляющие ряда. Напомним, что автокорреляции на последовательных лагах зависимы. Поэтому удаление некоторых автокорреляций изменит другие автокорреляции, которые, возможно, подавляли их, и сделает некоторые другие сезонные составляющие более заметными.

Во-вторых, удаление сезонных составляющих делает ряд стационарным, что необходимо для применения авторегрессии и скользящего среднего и других методов, например, спектрального анализа.

10. НЕЙРОННЫЕ СЕТИ

10.1. Принципы построения нейронных сетей

В последние десятилетия в мире бурно развивается новая прикладная область математики, специализирующаяся на *искусственных нейронных сетях* (НС) или *artificial neural network*. Актуальность исследований в этом направлении подтверждается массой различных применений НС. Это автоматизация процессов распознавания образов, адаптивное управление, аппроксимация функционалов, прогнозирование, создание экспертных систем, организация ассоциативной памяти и многие другие приложения. С помощью НС можно, например, предсказывать показатели биржевого рынка, выполнять распознавание оптических или звуковых сигналов, создавать самообучающиеся системы, способные управлять автомашиной при парковке или синтезировать речь по тексту.

Широкий круг задач, решаемый НС, не позволяет в настоящее время создавать универсальные, мощные сети, вынуждая разрабатывать специализированные НС, функционирующие по различным алгоритмам.

Модели НС могут быть программного и аппаратного исполнения.

Несмотря на существенные различия, отдельные типы НС обладают несколькими общими чертами.

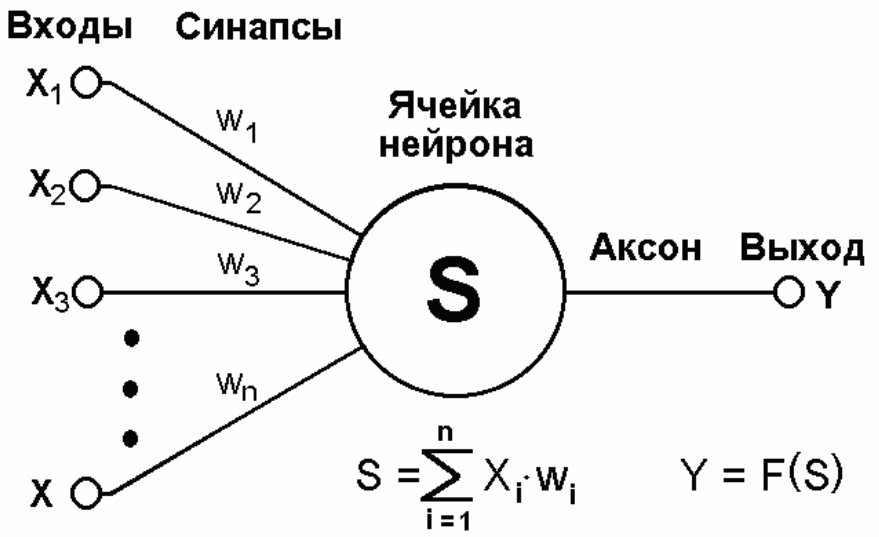

Рисунок 10.1.1. Структурная схема искусственного нейрона.

Во-первых, основу каждой НС составляют относительно простые, в большинстве случаев – однотипные, элементы (ячейки), имитирующие работу нейронов мозга. Далее под нейроном будет подразумеваться искусственный нейрон, то есть ячейка НС. Каждый нейрон характеризуется своим текущим состоянием по аналогии с нервными клетками головного мозга, которые могут быть возбуждены или заторможены. Он обладает группой синапсов – однонаправленных входных связей, соединенных с выходами других нейронов, а также имеет аксон – выходную связь данного нейрона, с которой сигнал (возбуждения или торможения) поступает на синапсы следующих нейронов. Общий вид нейрона приведен на рисунке 10.1.1. Каждый синапс характеризуется величиной синаптической связи или ее весом w_i, который по физическому смыслу эквивалентен электрической проводимости, и определяет чувствительность каждого канала и его вес в общей сумме (10.1.1).

Текущее состояние нейрона определяется, как взвешенная сумма его входов:

$$s = \sum_{i=1}^{n} x_i \cdot w_i \,, \qquad (10.1.1)$$

где: x_i – сигнал раздражителя, подаваемый на соответствующий синапс.

Сигнал на выходе нейрона формируется при помощи функции состояния:

$$y = f(s) \,. \qquad (10.1.2)$$

Нелинейная функция f называется активационной и может иметь вид, как показано на рисунке 10.1.2. Одной из наиболее распространенных является нелинейная функция с насыщением, так называемая логистическая функция или сигмоид (т.е. функция S-образного вида):

$$f(x) = \frac{1}{1 + e^{-a x}}, \qquad (10.1.3)$$

где a – параметрический коэффициент.

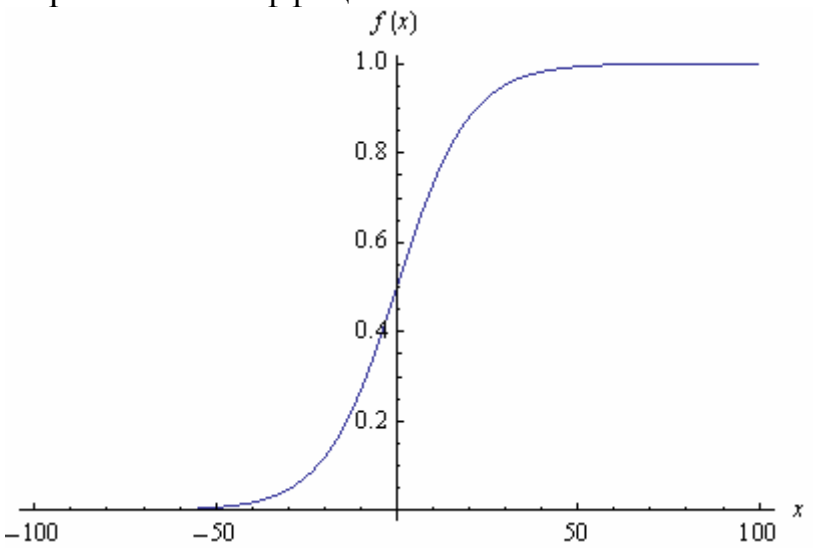

Рисунок 10.1.2. Функция сигмоид при *a=0.1*.

Подбор коэффициента a позволяет получить необходимую крутизну функции, а соответственно чувствительность нейрона к суммарному сигналу раздражителей. Трёхмерная диаграмма функции (10.1.3) отображена на рисунке 10.1.3.

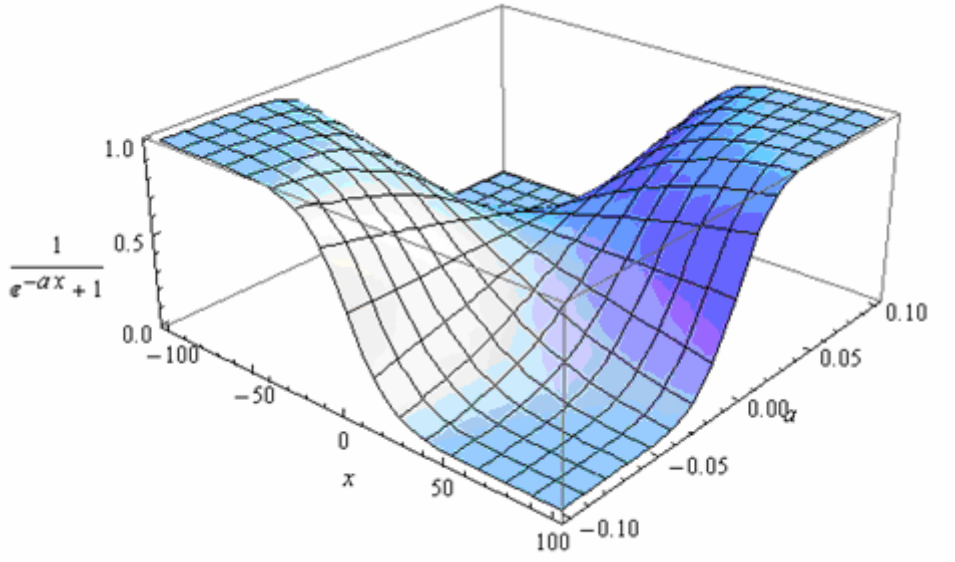

Рисунок 10.1.3. Сигмоид как *f(x,a)*.

Из выражения для сигмоида очевидно, что выходное значение нейрона лежит в диапазоне [0,1]. Одно из ценных свойств сигмоидной функции – простое выражение для ее производной:

$$f'(x) = \alpha \cdot f(x) \cdot (1 - f(x)). \qquad (9.1.4)$$

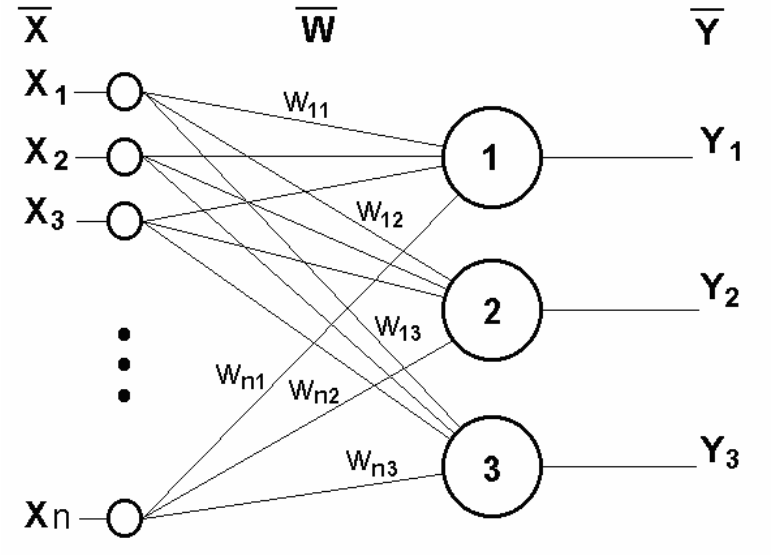

Рисунок 10.1.4. Однослойная нейронная сеть.

В качестве примера простейшей НС рассмотрим трехнейронный персептрон (рис. 10.1.4), то есть такую сеть, нейроны которой имеют активационную функцию $f(x)$ в виде единичного скачка. На n входов поступают некие сигналы, проходящие по синапсам на 3 нейрона, образующие единственный слой этой НС и выдающие три выходных сигнала:

$$y_j = f\left[\sum_{i=1}^{n} x_i \cdot w_{ij}\right], \quad j = 1,2,3 \tag{10.1.5}$$

Очевидно, что все весовые коэффициенты синапсов одного слоя нейронов можно свести в матрицу **W**, в которой каждый элемент w_{ij} задает величину i-ой синаптической связи j-ого нейрона. Таким образом, процесс, происходящий в НС, может быть записан в матричной форме:

$$Y=f(\mathbf{XW}) \tag{10.1.6}$$

где **X** и **Y** – соответственно входной и выходной сигнальные векторы, f(**XW**)– активационная функция, применяемая поэлементно к компонентам вектора **XW**.

Выбор структуры НС осуществляется в соответствии с особенностями и сложностью задачи. Из описания некоторых принципов построения НС следует, что возможные конфигурации и особенности построения НС очень многообразны. Для большинства стандартных задач имеются уже готовые программные средства или библиотеки программ реализующие НС требуемой конфигурации. Для решения некоторых нестандартных задач требуется разрабатывать НС с особенными свойствами.

Очевидно, что процесс функционирования НС, то есть сущность действий, которые она способна выполнять, зависит от величин синаптических связей, поэтому, задавшись определенной структурой НС, отвечающей какой-либо задаче, разработчик сети должен найти оптимальные значения всех переменных весовых коэффициентов (некоторые синаптические связи могут быть постоянными).

Этап подбора параметров НС называется обучением, и от того, насколько качественно он будет выполнен, зависит способность сети решать поставленные перед ней проблемы во время эксплуатации. На этапе обучения кроме параметра качества подбора весов важную роль играет время обучения. Как правило, эти два параметра связаны обратной зависимостью и их приходится выбирать на основе компромисса.

Обучение НС может вестись с учителем или без него. В первом случае сети предъявляются значения как входных, так и желательных выходных сигналов, и она по некоторому внутреннему алгоритму подстраивает веса своих синаптических связей. Во втором случае выходы НС формируются самостоятельно, а веса изменяются по алгоритму, учитывающему только входные и производные от них сигналы.

Существует великое множество различных алгоритмов обучения, которые делятся на два больших класса: детерминистские и стохастические. В первом из них подстройка весов представляет собой жесткую последовательность действий, во втором – она производится на основе действий, подчиняющихся некоторому случайному процессу.

Развивая дальше вопрос о возможной классификации НС, важно отметить существование бинарных и аналоговых сетей. Первые из них оперируют с двоичными сигналами, и выход каждого нейрона может принимать только два значения: логический ноль ("заторможенное" состояние) и логическая единица ("возбужденное" состояние). В аналоговых сетях выходные значения нейронов способны принимать непрерывные значения, что могло бы иметь место после замены активационной функции нейронов персептрона на сигмоид.

Еще одна классификация делит НС на синхронные и асинхронные. В первом случае в каждый момент времени свое состояние меняет лишь один нейрон. Во втором – состояние меняется сразу у целой группы нейронов, как правило, у всего слоя. Алгоритмически ход времени в НС задается итерационным выполнением однотипных действий над нейронами. Сети также можно классифицировать по числу слоев.

До сих пор, говоря о построении и конструировании сети, мы предполагали, что входной и выходной слои заданы, то есть, что мы уже знаем, какие переменные будут подаваться на вход сети, и что будет ее выходом. То, какие переменные будут выходными, известно всегда (по

крайней мере в случае управляемого обучения). Что же касается входных переменных, их правильный выбор порой представляет большие трудности (Bishop, 1995). Часто мы не знаем заранее, какие из входных переменных действительно полезны для решения задачи, и выбор хорошего множества входов бывает затруднен целым рядом обстоятельств.

a.) *Проклятие размерности*. Каждый дополнительный входной элемент сети - это новая размерность в пространстве данных. С этой точки зрения становится понятно следующее: чтобы достаточно плотно "заселить" *N*-мерное пространство и "увидеть" структуру данных, нужно иметь довольно много точек. Необходимое число точек быстро возрастает с ростом размерности пространства (грубо говоря, как 2^N для большинства методов). Большинство типов нейронных сетей (в частности, многослойный персептрон **MLP**) в меньшей степени страдают от проклятия размерности, чем другие методы, потому что сеть умеет следить за проекциями участков многомерного пространства в пространства малой размерности (например, если все веса, выходящие из некоторого входного элемента, равны нулю, то MLP-сеть полностью игнорирует эту входную переменную). Тем не менее, проклятие размерности остается серьезной проблемой, и качество работы сети можно значительно улучшить, исключив ненужные входные переменные. На самом деле, чтобы уменьшить эффект проклятия размерности иногда бывает целесообразно исключить даже те входные переменные, которые несут в себе некоторою (небольшую) информацию.

b.) *Внутренние зависимости между переменными*. Было бы очень хорошо, если бы каждую переменную - кандидата на то, чтобы служить входом сети, можно было бы независимо оценить на "полезность", а затем отобрать самые полезные переменные. К сожалению, как правило, это бывает невозможно сделать, и две или более взаимосвязанных переменных могут вместе нести существенную информацию, которая не содержится ни в каком их подмножестве. Классическим примером может служить задача с двумя спиралями, в которой точки данных двух классов расположены вдоль двух переплетающихся двумерных спиралей. Ни одна из переменных в отдельности не несет никакой полезной информации (классы будут выглядеть совершенно перемешанными), но глядя на обе переменные вместе, классы легко разделить. Таким образом, в общем случае переменные нельзя отбирать независимо.

c.) *Избыточность переменных*. Часто бывает так, что одна и та же информация в большей или меньшей степени повторяется в разных переменных. Например, данные о росте и весе человека, как правило, несут в себе сходную информацию, поскольку они сильно коррелированы. Может оказаться так, что в качестве входов достаточно взять лишь часть

из нескольких коррелированных переменных, и этот выбор может быть произвольным. В таких ситуациях вместо всего множества переменных лучше взять их часть - этим мы избегаем проклятия размерности.

Итак, выбор входных переменных - это исключительно важный этап при построении нейронной сети. Перед тем, как непосредственно начинать работать с НС, имеет смысл произвести предварительный отбор переменных, используя при этом свои знания в предметной области и стандартные статистические критерии. Затем, уже средствами пакета STATISTICA можно будет попробовать различные комбинации входных переменных. Пакет ST Neural Networks фирмы StatSoft, Inc., ориентирован специально на НС, и позволяет производить значительные вычисления с применением НС. В пакете ST Neural Networks имеется возможность "игнорировать" некоторые переменные, так что полученная сеть не будет использовать их в качестве входов. Можно поочередно экспериментировать с различными комбинациями входов, строя всякий раз новые варианты сетей.

При экспериментировании очень полезными оказываются вероятностные и обобщенно-регрессионные сети. Несмотря на то, что они работают медленнее более компактных **MLP** и **RBF** сетей, они обучаются почти мгновенно, и это важно, поскольку при переборе большого числа комбинаций входных переменный приходится каждый раз строить новые сети. Кроме того, **PNN** и **GRNN** (как и **RBF**) - это радиальные сети (в первом слое они имеют радиальные элементы, и аппроксимирующие функция строятся в виде комбинаций гауссовых функций). При отборе входных переменных это является преимуществом, поскольку радиальные сети в меньшей степени страдают от проклятия размерности, чем сети, построенные на линейных элементах.

Чтобы понять причину этого, рассмотрим, что произойдет, если мы добавим в сеть новую, возможно совершенно несущественную входную переменную. Сеть на линейных элементах, например **MLP**, может научиться присваивать весам, идущим от этой переменной, нулевые значения, что означает игнорирование переменной (реально это происходит так: изначально малые веса этой переменной так и остаются малыми, а веса содержательных входных переменных меняются нужным образом). Радиальная сеть типа **PNN** или **GRNN** не может позволить себе такую роскошь: кластеры, образующиеся в пространстве небольшого числа существенных переменных, будут "размазаны" по направлениям несущественных размерностей - для учета разброса по несущественным направлениям требуется большее число элементов. Сеть, в большей степени страдающая от наличия плохих входных данных, имеет преимущество, когда мы стремимся избавиться то этих плохих данных.

Поскольку описанный процесс экспериментирования занимает много времени, в пакете ST Neural Networks имеется инструмент, который может сделать это за Вас. Для выбора подходящей комбинации входных переменных здесь используется так называемый генетический алгоритм (Goldberg, 1989). Генетические алгоритмы хорошо приспособлены для задач такого типа, поскольку они позволяют производить поиск среди большого числа комбинаций при наличии внутренних зависимостей в переменных.

Существует и другой подход к проблеме размерности, который может использоваться как альтернатива или как дополнение к методам отбора переменных: это понижение размерности. Суть его состоит в том, что исходная совокупность переменных преобразуется в новую совокупность, состоящую из меньшего числа переменных, но при этом (как мы надеемся) содержащую по возможности всю информацию, заложенную в исходных данных. В качестве примера рассмотрим данные, все точки которых расположены на некоторой плоскости в трехмерном пространстве. Истинная размерность данных равна двум (поскольку вся информация на самом деле содержится в двумерном подпространстве). Если мы сумеем обнаружить эту плоскость, то на вход нейронной сети можно будет подавать входные данные меньшей размерности, и будет больше шансов на то, что такая сеть будет работать правильно.

Самый распространенный метод понижения размерности - это анализ главных компонент (Bishop, 1995; см. также Факторный анализ). Метод состоит в следующем: к данным применяется линейное преобразование, при котором направлениям новых координатных осей соответствуют направления наибольшего разброса исходных данных. Как правило, уже первая компонента отражает большую часть информации, содержащейся в данных. Поскольку анализ главных компонент (АГК) представляет собой линейный метод, его можно реализовать с помощью линейной сети, и в пакете ST Neural Networks предусмотрена возможность обучать линейную сеть для выполнения АГК. Очень часто метод АГК выделяет из многомерных исходных данных совсем небольшое число компонент, сохраняя при этом структуру информации.

Один из недостатков метода главных компонент (АГК) состоит в том, что это чисто линейный метод, и из-за этого он может не учитывать некоторые важные характеристики структуры данных. В пакете ST Neural Networks реализован также вариант "нелинейного АГК", основанный на использовании так называемой автоассоциативной сети (Bishop, 1995; Fausett, 1994; Bouland and Kamp, 1988). Это такая нейронная сеть, которую обучают выдавать в качестве выходов свои собственные входные данные, но при этом в ее промежуточном слое содержится меньше нейронов, чем

во входном и выходном слоях. Поэтому, чтобы восстановить свои входные данные, сеть должна научиться представлять их в более низкой размерности. Сеть "впихивает" наблюдения в формат промежуточного слоя и только потом выдает их на выходе. После обучения автоассоциативной сети ее внешний интерфейс может быть сохранен и использован для понижения размерности. Как правило, в качестве автоассоциативной сети берется многослойный персептрон с тремя промежуточными слоями. При этом средний слой отвечает за представление данных в малой размерности, а два других скрытых слоя служат соответственно для нелинейного преобразования входных данных в средний слой и выходов среднего слоя в выходной слой. Автоассоциативная сеть с единственным промежуточным слоем может выполнять только линейное понижение размерности, и фактически осуществляет АГК в стандартном варианте.

10.2. Применение нейронных сетей для анализа данных

В пакет STATISTICA интегрирован модуль нейросетевого анализа данных (рис. 10.2.1) (**automated neural networks**), позволяющий создавать нейронную сеть, обучать её и производить: регрессионный анализ (**Regression**, **Time series**); классификацию (**Classification**, **Time series**); кластеризацию (**Cluster analysis**).

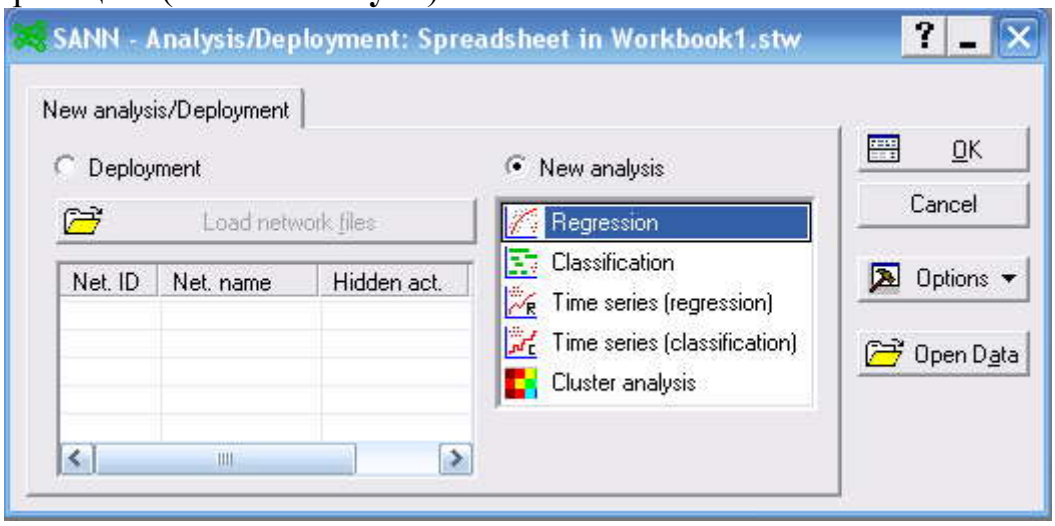

Рисунок 10.2.1. Окно модуля нейросетевого анализа данных.

Обучение нейронной сети производится на поле заранее подготовленных данных, обученную сеть можно сохранить для использования в дальнейшем.

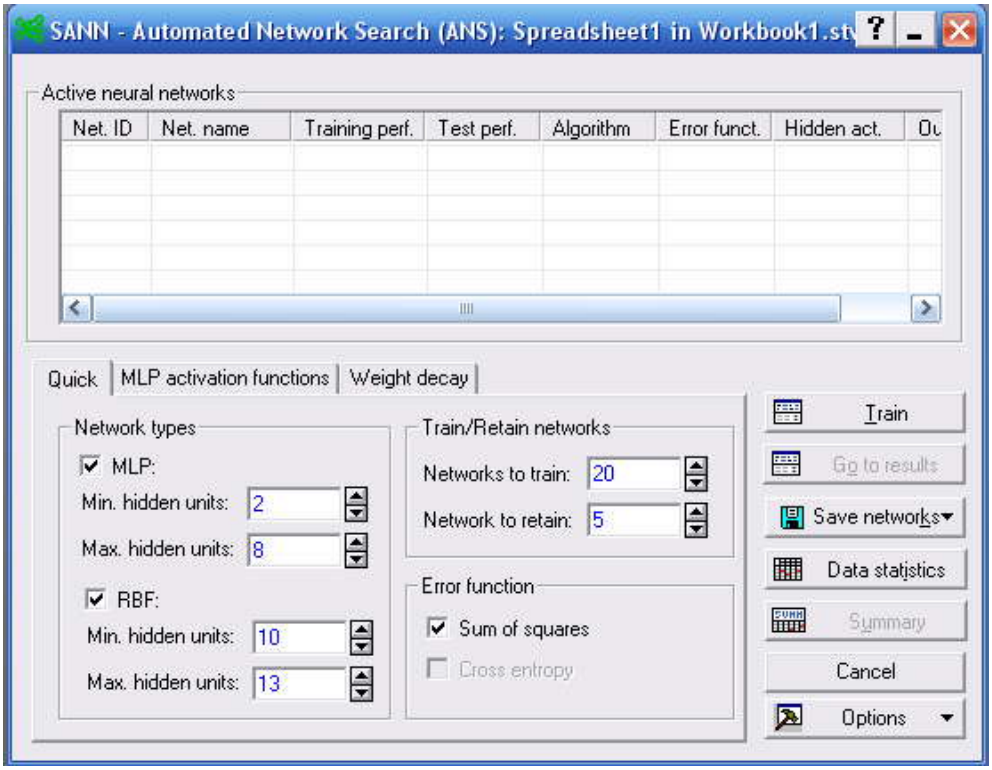

Рисунок 10.2.2. Окно параметров обучения нейронной сети.

При обучении сети можно задать множество параметров, в том числе число скрытых слоёв сети (**hidden units**), тип сети (**MLP, RBF**) и функции активации нейрона (**MLP activation function**). Чем больше число скрытых узлов в модели нейронной сети тем сильнее модель, то есть более способная сеть для моделирования сложных отношений между входными данными и целевыми значениями.

MLP – соответствует использованию многослойной сети на базе персептрона. Многослойный персептрон является наиболее распространенной формой сети. Она требует итерационного обучения, что может быть довольно медленным для большого числа скрытых узлов и наборов данных. Такие сети достаточно компактные, быстрые при работе после обучения, и в большинстве случаев дают лучшие результаты, чем другие типы сетей.

RBF – включает радиальную базисную функцию. Радиальная базисная функция сети, как правило, медленнее и объёмнее (по занимаемой памяти), чем многослойный персептрон, и зачастую уступают в производительности многослойному персептрону, но они очень быстро обучаются. Кроме того, они обычно менее эффективны, чем многослойных персептрон, в случае если имеется большое количество входных переменных (они более чувствительны к включению излишних входов).

11. КОНТРОЛЬ КАЧЕСТВА

11.1. Стандартные карты контроля качества

При организации любого производственного процесса возникает задача установки пределов характеристик изделия, в рамках которых произведенная продукция удовлетворяет своему предназначению. Вообще говоря, существует два "врага" качества продукции: (1) отклонения от плановых спецификаций и (2) слишком большой разброс реальных характеристик изделий (относительно плановых спецификаций). На ранних стадиях отладки производственного процесса для оптимизации этих двух показателей качества часто используются методы *планирования эксперимента*. Методы контроля качества, предназначены для построения процедур контроля качества продукции в процессе ее производства, т.е. текущего контроля качества. За детальным описанием принципов построения контрольных карт и примерам обратитесь к работам Buffa (1972), Duncan (1974), Grant and Leavenworth (1980), Juran (1962), Juran and Gryna (1970), Montgomery (1985, 1991), Shirland (1993) или Vaughn (1974). В качестве превосходных вводных курсов, построенных на основе подхода "как - чтобы", можно указать монографии Hart and Hart (1989) и Pyzdek (1989), а также изданные на немецком языке курсы Rinne and Mittag (1995) и Mittag (1993).

Общий подход к текущему контролю качества достаточно прост. В процессе производства проводятся выборки изделий заданного объема. После этого на специально разлинованной бумаге строятся диаграммы изменчивости выборочных значений плановых спецификаций в этих выборках и рассматривается степень их близости к заданным значениям. Если диаграммы обнаруживают наличие тренда выборочных значений или оказывается, что выборочные значения находятся вне заданных пределов, то считается, что процесс вышел из-под контроля, и предпринимаются необходимые действия для того, чтобы найти причину его разладки. Иногда такие специально разлинованные бумаги называют контрольными картами Шуэрта (в честь W. A. Shewhart, который общепризнанно считается первым, применившим на практике описываемые здесь методы анализа; см. Shewhart, 1931).

Интерпретация контрольных карт. В компьютерном варианте контрольных карт наиболее часто встречается ситуация, когда на экране находятся две карты (и две гистограммы), одна из них называется **X**-картой, а другая - **R**-картой.

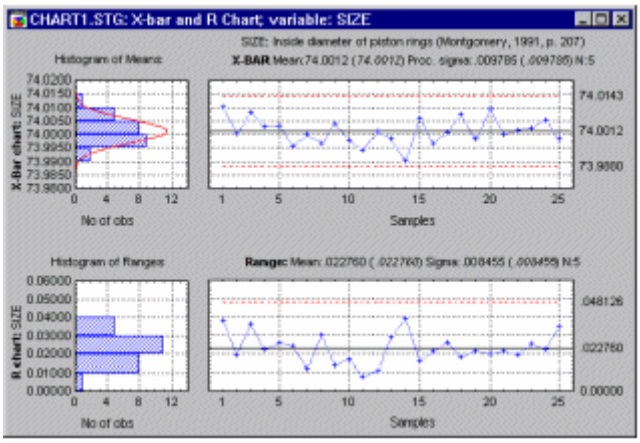

Рисунок 11.1.1. ***X***-карта и ***R***-карта.

В обеих контрольных картах по горизонтальной оси откладываются номера соответствующих выборок; по вертикальной оси в случае *X*-карты отложены выборочные средние исследуемых характеристик, а в случае *R*-карты – размахи соответствующих выборок. Пусть, например, производятся контрольные измерения диаметра поршневых колец, изготавливаемых на вашем предприятии. Тогда центральная линия на *X* - карте будет соответствовать размеру, используемому в качестве стандарта (например, установленному диаметру кольца в миллиметрах), в то время как центральная линия *R*-карты будет соответствовать приемлемому (т.е. находящемуся в пределах плановой спецификации) размаху диаметра поршневого кольца в выборках; таким образом, последняя контрольная карта представляет собой карту изменчивости процесса (чем больше изменчивость, тем больше диапазон отклонения от стандарта). Кроме центральной линии, на карте обычно присутствуют две дополнительные горизонтальные прямые, обозначающие верхний и нижний контрольные пределы (ВКП и НКП соответственно). Принципы определения этих линий обсуждаются ниже. Обычно нанесенные на карты отдельные точки соответствуют выборочным значениям и соединяются прямыми линиями. Если результирующая кривая на графике выходит за верхний или нижний контрольный предел или ее конфигурация выражает определенную тенденцию поведения для следующих друг за другом выборок, то это рассматривается как указание на существование проблем с качеством.

Несмотря на то, что можно достаточно произвольно определить момент разладки производственного процесса (например, при выходе соответствующих значений за границы верхних и нижних контрольных пределов), обычной практикой является применение статистических методов для определения этого момента. В разделе Элементарные понятия статистики обсуждаются свойства выборочного распределения, а также дается сводка характеристик нормального распределения. Метод

установления верхнего и нижнего контрольных пределов представляет собой прямое следствие применения описанных в этом разделе принципов.

Пример. Предположим, вы контролируете среднее значение некоторой величины, например, диаметра поршневых колец. Пусть среднее значение диаметров и дисперсия в процессе производства не меняются. Тогда выборочные средние, полученные для последовательных выборок, будут распределены нормально относительно истинного среднего. Более того, не вдаваясь в тонкости, связанные с выводом формул, можно заключить (согласно центральной предельной теореме и сделанному предположению о нормальности выборочных средних размеров колец; см, например, работу Hoyer and Ellis, 1996), что стандартное отклонение распределения выборочных средних будет равно сигме (стандартному отклонению отдельных наблюдений или измерений диаметра отдельных колец), деленному на квадратный корень из n (n - размер выборки). Следовательно, примерно 95% значений выборочных средних попадут в интервал ±1.96. На практике обычно заменяют 1.96 на 3 (при этом в интервал попадают приблизительно 99% выборочных средних) и определяют верхний и нижний контрольные пределы как плюс-минус 3 сигма соответственно.

Описанный выше частный принцип установления контрольных пределов применяется во всех типах контрольных карт. После выбора контролируемой характеристики (например, стандартного отклонения) оценивается ее ожидаемая изменчивость в выборках того размера, который будет использоваться в контролируемой процедуре. Затем с помощью полученных оценок изменчивости устанавливают контрольные пределы карты.

Классификация типов контрольных карт часто осуществляется согласно типам величин, которые выбраны для отслеживания характеристик качества. Так, различают контрольные карты для непрерывных переменных и контрольные карты по альтернативному признаку. В частности, для контроля по непрерывному признаку обычно строятся следующие контрольные карты:

***X**-карта*. На эту контрольную карту наносятся значения выборочных средних для того, чтобы контролировать отклонение от среднего значения непрерывной переменной (например, диаметров поршневых колец, прочности материала и т.д.);

***R**-карта*. Для контроля за степенью изменчивости непрерывной величины в контрольной карте этого типа строятся значения размахов выборок;

S*-*карта. Для контроля за степенью изменчивости непрерывной переменной в контрольной карте данного типа рассматриваются значения выборочных стандартных отклонений;

S^2*-*карта. В контрольной карте данного типа для контроля изменчивости строится график выборочных дисперсий.

Для контроля качества продукции по альтернативному признаку обычно применяются следующие типы контрольных карт:

C*-*карта. В таких контрольных картах строится график числа дефектов (в партии, в день, на один станок, в расчете на 100 футов трубы и т.п.). При использовании карты этого типа делается предположение, что дефекты контролируемой характеристики продукции встречаются сравнительно редко, при этом контрольные пределы для данного типа карт рассчитываются на основе свойств распределения Пуассона (распределения редких событий);

U*-*карта. В карте данного типа строится график относительной частоты дефектов, то есть отношения числа обнаруженных дефектов к n - числу проверенных единиц продукции (здесь n обозначает, например, число футов длины трубы, объем партии изделий). В отличие от С-карты, для построения карты данного типа не требуется постоянство числа единиц проверяемых изделий, поэтому ее можно использовать при анализе партий различного объема;

Np-карта. В контрольных картах этого типа строится график для числа дефектов (в партии, в день, на станок), как и в случае **C**-карты. Однако, контрольные пределы этой карты рассчитываются на основе биномиального распределения, а не распределения редких событий Пуассона. Поэтому данный тип карт должен использоваться в том случае, когда обнаружение дефекта не является редким событием (например, когда обнаружение дефекта происходит более чем у 5% проверенных единиц продукции). Этой картой можно воспользоваться, например, при контроле числа единиц продукции, имеющих небольшой брак;

P-карта. В картах данного типа строится график процента обнаруженных дефектных изделий (в расчете на партию, в день, на станок и т.д.). График строится так же, как и в случае **U**-карты. Однако контрольные пределы для данной карты находятся на основе биномиального распределения (для долей), а не распределения редких событий. Поэтому **P**-карта наиболее часто используется, когда появление дефекта нельзя считать редким событием (если, например, ожидается, что дефекты будут присутствовать в более чем 5% общего числа произведенных единиц продукции).

Все перечисленные выше типы карт допускают возможность построения кратких карт для производственных серий (краткие

контрольные карты) и контрольных карт для нескольких процессов (многопоточные групповые карты).

Краткая контрольная карта (контрольная карта для кратких производственных серий) представляет собой график наблюдаемых значений характеристик качества (значений непрерывной переменной или альтернативного признака) для нескольких частей процесса, причем все значения контролируемой характеристики наносятся на одну и ту же карту. Разработка кратких контрольных карт стала следствием необходимости адаптации контрольных карт к тем ситуациям, когда требуется выполнить несколько десятков измерений контролируемой характеристики процесса, прежде чем вычислить контрольные пределы. Часто данное требование выполняется с трудом на тех стадиях производственного процесса, в ходе которых изготавливается ограниченное (малое) число деталей, которые необходимо подвергнуть измерениям.

Так, например, на целлюлозно-бумажном комбинате процесс может быть организован следующим образом: выпускается только три-четыре больших рулона бумаги определенного сорта (часть процесса), а затем переходят к выпуску бумаги другого сорта. Однако, если измерения переменных (таких, например, как толщина бумаги или альтернативных признаков, таких, как наличие/отсутствие пятен) производятся для нескольких десятков рулонов, скажем, десяти различных сортов, то контрольные пределы для переменной "толщина бумаги" и признака "наличие/отсутствие пятен" могут быть вычислены на основе преобразованных значений (в рамках краткой производственной серии). Более точно, эти преобразования заключаются в таком изменении масштаба контролируемых переменных, при котором амплитуды их изменения в различных производственных сериях (различных частях процесса) будут сравнимыми. Контрольные пределы, рассчитанные по этим преобразованным значениям, могут применяться в дальнейшем при контроле толщины бумаги и наличия/отсутствия пятен, вне зависимости от сорта выпускаемой бумаги. Для того чтобы определить, произошла разладка процесса или нет, могут быть использованы статистические процедуры контроля процесса. Этими процедурами можно воспользоваться также для постоянного контроля производства и разработки способов постоянного улучшения качества.

Номинальная карта, карта плановых спецификаций. Существует несколько типов кратких контрольных карт. Наиболее часто используются следующие карты: номинальная карта и карта плановых спецификаций. При построении данных карт преобразование наблюдаемых значений контролируемой характеристики в различных частях процесса

производится путем вычитания определенной постоянной из измерений (для наблюдений каждой части используется своя постоянная). В качестве таких постоянных могут выступать как значения номинала для соответствующих частей процесса (результатом такого подхода будет номинальная краткая карта), так и плановые спецификации, рассчитанные по "историческим" средним контролируемой характеристики для каждой части (краткая **X**-карта плановых спецификаций и краткая *R*-карта плановых спецификаций). Так, например, сравнение внутренних диаметров поршневых колец для различных блоков мотора, находящихся в производстве, только тогда может быть обоснованно, когда до проведения сравнения из измерений диаметров будут вычтены средние разности между внутренними диаметрами поршневых колец для моторов различного размера (для определения непротиворечивости значений диаметров). Такое сравнение становится возможным при построении краткой номинальной карты или краткой карты плановых спецификаций. Заметим, что при построении номинальной карты и карты плановых спецификаций делается предположение о равенстве дисперсий различных частей процесса, чтобы применение рассчитанных по общей оценке сигма процесса контрольных пределов можно было считать корректным.

Стандартизованная краткая карта. Если изменчивость различных частей процесса нельзя считать одинаковой, то прежде чем нанести на одну карту данные, относящиеся к разным частям процесса, необходимо провести еще одно преобразование. При построении карты данного типа это преобразование заключается в следующем: вычисляются отклонения выборочных средних контролируемой характеристики от средних для соответствующих частей процесса (т.е. от номинальных значений или плановых спецификаций для частей), далее для каждой части процесса эти отклонения делятся на постоянные, пропорциональные изменчивости соответствующих частей. Так, в случае кратких *X*-карты и *R*-карты, для построения точек графика **X**-карты вначале из каждого выборочного среднего вычитается определенная постоянная, соответствующая рассматриваемой части процесса (т.е. среднее этой части процесса или значение номинала для данной части), затем эта разность делится на другую постоянную – например, на средний размах соответствующей части процесса. В результате таких преобразований масштабы выборочных средних различных частей процесса станут сравнимыми.

Краткие карты по альтернативному признаку. В случае контрольных карт по альтернативному признаку (**C**-, **U**-, **Np**- или **P**-карт) оценка изменчивости процесса (доля, частота и т.д.) зависит от среднего значения процесса (средней доли, средней относительной частоты и т.д.). Следовательно, для альтернативных признаков могут быть построены

только стандартизованные краткие карты. К примеру, точки краткой *P*-карты находятся вычитанием из соответствующих выборочных значений долей *p* средних *p* для части процесса, с последующим делением результата на стандартное отклонение средних *p*.

Групповая контрольная карта дает возможность нанести данные для нескольких потоков наблюдаемых значений непрерывной переменной или альтернативного признака (характеристик качества) на одну и ту же карту. Это упрощает интерпретацию карты при одновременном управлении большим числом процессов или их характеристик. Здесь термином "потоки процесса" могут обозначаться данные, полученные для различных станков, сборочных линий, операторов и так далее. Все эти данные могут быть нанесены на одну контрольную карту.

При построении групповой **X**-карты для каждой из выборок с измерениями контролируемой характеристики на карту наносится две точки, в результате чего на графике образуются две линии. Верхняя из них представляет собой график наиболее высоких средних значений каждой выборки для всех нанесенных на карту потоков переменных или альтернативных признаков, а нижняя – подобный график наименьших средних значений каждой выборки. Для каждой выборки верхняя и нижняя точка представляют собой максимальное и минимальное средние всех нанесенных на карту потоков переменных или альтернативных признаков. Если эти экстремальные значения не выходят за рамки заданных контрольных пределов, очевидно, что все остальные средние также будут находиться в области, ограниченной контрольными пределами. Следовательно, с помощью групповой **X**-карты, можно быстро определить, не началась ли разладка процесса в одном или нескольких потоках процесса или контролируемых характеристиках, не переходя к проверке всех измерений подряд.

Групповая карта для одной части процесса называется стандартной групповой картой или, обычно, просто групповой картой. Групповые карты для нескольких частей процесса называются групповыми краткими картами. Для построения групповых кратких карт используется та же процедура, что и для стандартных групповых карт; единственное их отличие от стандартных состоит в том, что точки на график наносятся только после того, как будут выполнены все преобразования данных в пределах отдельных частей процесса.

При построении на контрольной карте графика для выборок неодинакового объема контрольные пределы, находящиеся по обе стороны от центральной линии (плановой спецификации), не могут быть изображены прямыми линиями. Существует три способа, позволяющих справиться с такой ситуацией.

1.) Средние объемы выборок. В том случае, когда желательно оставить контрольные пределы в виде прямых линий (например, чтобы облегчить чтение карты и ее использование в презентациях), можно найти среднее значение объема выборки n по всем рассматриваемым выборкам и установить контрольные пределы на основе полученного среднего объема выборки. Эту процедуру нельзя назвать "точной". И все же, пока объемы выборок несильно отличаются друг от друга, применение данного метода можно считать вполне адекватным.

2.) Переменные контрольные пределы. С другой стороны, для каждой выборки можно отдельно определить контрольные пределы на основе ее объема. При таком подходе будут получены переменные контрольные пределы. На графике такие пределы будут изображены ступенчатой линией. Этот метод позволяет получить точные контрольные пределы для каждой из использующихся выборок. Однако при этом теряется простота и наглядность контрольных пределов, отмечаемых на карте прямой линией.

3.) Стабилизированная (нормализованная) карта. Наилучший вариант - изображающиеся прямыми линиями контрольные пределы, которые при этом точны - может быть получен путем стандартизации контролируемой численной характеристики (среднего значения, доли и т.д.) согласно единицам сигмы. При этом контрольные пределы изображаются прямыми линиями, но расположение точек выборочных значений на графике определяется не только значениями контролируемой характеристики, но и объемом n соответствующих выборок. Недостаток данного метода заключается в следующем: по вертикальной оси контрольной карты (оси *Y*) величины выражаются в единицах сигма, а не в первоначальных единицах измерения контролируемой характеристики, поэтому их нельзя считывать по выводимому на графике значению. Так, например, выборочная величина со значением 3 отстоит на 3 сигма от плановой спецификации. Для перевода данного значения в первоначальные единицы измерения придется выполнить некоторый объем вычислений.

Иногда инженеру, занимающемуся контролем качества, приходится выбирать между применением контрольной карты для непрерывных переменных и контрольной карты по альтернативному признаку.

Преимущество контрольных карт по альтернативному признаку состоит в возможности быстро получить общее представление о различных аспектах качества анализируемого изделия; то есть, на основании различных критериев качества инженер может сразу принять или забраковать продукцию. Далее, контрольные карты по альтернативному признаку иногда позволяют обойтись без применения дорогих точных приборов и требующих значительных затрат времени

измерительных процедур. Кроме того, этот тип контрольных карт более понятен менеджерам, которые не разбираются в тонкостях методов контроля качества. Таким образом, с помощью таких карт можно более убедительно продемонстрировать руководству наличие проблем с качеством изделий.

Контрольные карты для непрерывных переменных обладают большей чувствительностью, чем контрольные карты по альтернативному признаку (см. Montgomery, 1985, стр. 203). Благодаря этому, контрольные карты для непрерывных переменных могут указать на существование проблемы ухудшения качества, прежде чем в потоке продукции появятся настоящие бракованные изделия, выделяемые с помощью контрольной карты по альтернативному признаку. В работе Montgomery (1985) автор называет контрольные карты для непрерывных переменных основными индикаторами ухудшения качества, которые предупреждают об этих проблемах задолго до того, как в процессе производства резко возрастет доля бракованных изделий.

Кроме выборок, состоящих из нескольких наблюдений, контрольные карты для переменных могут быть построены также для отдельных наблюдений, полученных в ходе производственного процесса. Иногда такой подход необходим в силу дороговизны, неудобства или невозможности анализа выборок, состоящих из ряда наблюдений. Примером может служить ситуация, когда число претензий потребителей или случаев возврата изделий может быть получено только по итогам месяца, тем не менее, существует необходимость в проведении текущего анализа этих данных для выявления ухудшения качества продукции. Другим широко встречающимся примером применения карт данного типа является проверка автоматическим тестирующим прибором каждой единицы произведенной продукции. В этом случае обычно стремятся обнаружить небольшие отклонения качества выпускаемой продукции (например, постепенное ухудшение качества, обусловленное износом оборудования). При этом наилучшее применения находят контрольные карты типа **CUSUM**, **MA**, и **EWMA** (контрольные карты для накопленных сумм и взвешенных средних).

Как уже было отмечено ранее в вводной части, когда точка на контрольной карте, соответствующая выборочному значению контролируемой характеристики (например, среднему значению в X-карте) оказывается вне ограниченной контрольными переделами области, это дает основания предполагать, что производственный процесс разладился. Далее, при этом необходимо отслеживать появление систематической тенденции в расположении точек (например, выборочных средних) на контрольной карте, так как наличие такой тенденции может служить

свидетельством тренда среднего значения контролируемого процесса. Эти критерии иногда называют критериями серий типа AT&T (см. AT&T, 1959) или критериями против альтернатив специального вида (см. Nelson, 1984, 1985; Grant and Leavenworth, 1980; Shirland, 1993). Термин специальные альтернативы, как альтернатива случайным или общим причинам, был использован в работе Шуэрта (Shewhart) для того, чтобы сделать разграничение между нормальным производственным процессом, вариации в котором появляются только в силу действия случайных причин, и вышедшим из-под контроля процессом, в котором вариации характеристик обусловлены некоторыми неслучайными, то есть специальными факторами (см. Montgomery, 1991, стр. 102).

Как и обсуждавшиеся ранее контрольные пределы, выраженные в единицах сигмы, критерии серий имеют в своей основе "статистическое" обоснование. Так, например, вероятность того, что любое выборочное среднее значение для **X**-карты окажется выше центральной линии, равна 0.5 при следующих условиях: (1) производственный процесс находится в нормальном состоянии (т.е. центральная линия проведена через значение, равное среднему контролируемой характеристики генеральной совокупности изделий), (2) средние значения следующих друг за другом выборок независимы (т.е. отсутствует автокорреляция) и (3) выборочные средние значения контролируемой характеристики распределены по нормальному закону. Проще говоря, при таких условиях для выборочного среднего значения шансы попасть выше или ниже центральной линии составляют 50 на 50. Поэтому вероятность того, что два следующих друг за другом выборочных средних окажутся выше центральной линии, будет равна 0.5, умноженному на 0.5 , т.е. 0.25.

Соответственно, вероятность того, что выборочные средние девяти последующих выборок (или серия из 9 точек контрольной карты) окажется с одной стороны от центральной линии, составит 0.59 = .00195. Заметим, что это значение приблизительно равно вероятности того, что отдельное выборочное среднее значение не попадет в интервал, ограниченный контрольными пределами в 3 сигма (при условии нормального распределения выборочных средних и нормальности производственного процесса). Поэтому, в качестве еще одного индикатора разладки производственного процесса можно рассматривать ситуацию, когда девять последовательных выборочных средних находятся с одной стороны от центральной линии. Со статистической интерпретацией других, более сложных критериев можно ознакомиться в работе Duncan (1974).

11.2. Специализированные типы контрольных карт

Далее рассматривается ряд других наиболее широко используемых методов и соответствующих им типов контрольных карт - "рабочих лошадок" контроля качества. Однако, с приходом недорогих персональных компьютеров, все большую популярность приобретают процедуры, требующие проведения большего объема вычислений.

***X**-карты для данных с негауссовским распределением.* Контрольные пределы для стандартных *X*-карт вычисляются, исходя из предположения о приблизительно нормальном распределении выборочных средних. Следовательно, для отдельных наблюдений в выборках нормальность распределения не обязательна, так как. по мере увеличения объема выборок распределение выборочных средних будет приближаться к нормальному (см. обсуждение центральной предельной теоремы в разделе Элементарные понятия статистики. Однако необходимо отметить, что при построении **R**-карты, **S**-карты и $\mathbf{S^2}$-карты предполагается, что отдельные наблюдения обладают нормальным распределением). В монографии Шуарта (Shewhart, 1931) автор экспериментирует с различными негауссовскими распределениями отдельных наблюдений и оценивает полученные в результате распределения средних для выборок объема 4. В результате было обнаружено, что, на самом деле, до тех пор, пока распределение отдельных наблюдений в выборках является приблизительно нормальным, можно применять вычисленные на основе нормального распределения стандартные контрольные пределы. Введение в данный вопрос и обсуждение предположений о распределении данных при контроле качества путем построения контрольных карт можно найти в работе Hoyer and Ellis, 1996.

Однако, как отмечено в работе Ryan (1989), при малых объемах выборок и сильной асимметрии распределения наблюдений, построенные по таким данным стандартные контрольные пределы приводят как к получению большого числа ложных сигналов тревоги (т.е. росту вероятности альфа-ошибки), так и увеличению числа случаев, когда при фактически произошедшей разладке процесс продолжает считаться контролируемым (росту вероятности бета-ошибки). В программе STATISTICA существует возможность расчета контрольных пределов для **X**-карт (а также индексов пригодности процесса) на основе так называемых кривых Джонсона (Johnson, 1949), с помощью которых аппроксимируется асимметрия и эксцесс большой группы негауссовских распределений (см. также раздел Подгонка распределений в модуле Анализ процессов). Негауссовские **X**-карты рекомендуется применять в том случае, когда распределение выборочных средних обладает явной асимметрией или является негауссовским.

Контрольная карта $\mathbf{T^2}$ *Хотеллинга*. Когда исследуется несколько взаимосвязанных характеристик качества (заданных в виде нескольких переменных), для всех средних значений можно построить общий график, воспользовавшись для этого многомерной статистикой Хотеллинга $\mathbf{T^2}$ (впервые предложена в работе Hotelling, 1947).

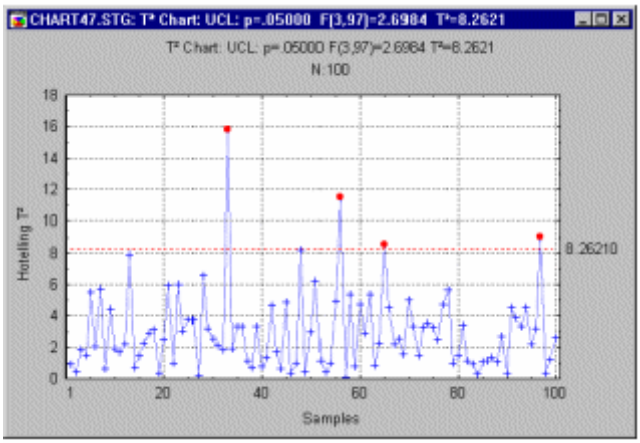

Рисунок 11.2.1. Карта Хотеллинга.

Контрольная карта накопленных сумм (***CUSUM-карта***). Контрольная карта типа **CUSUM** была впервые предложена в работе Page (1954). Обсуждение использующихся при ее построении математических принципов можно найти в работах Ewan (1963), Johnson (1961), а также Johnson and Leone (1962).

Если строить график накопленной суммы отклонений от плановых спецификаций для следующих друг за другом выборочных средних, то даже малые постоянные сдвиги среднего значения процесса постепенно приведут к накоплению ощутимой суммы отклонений. Поэтому данный тип контрольных карт особенно хорошо подходит для обнаружения малых постоянных сдвигов процесса, которые могут оказаться незамеченными при применении **X**-карты. Например, когда из-за износа оборудования процесс медленно "выскальзывает" из-под контроля, в результате чего размеры изделий превышают плановые спецификации (или становятся ниже их), при применении контрольной карты данного типа будет получен монотонно растущий (или снижающийся) график накопленной суммы отклонений от плановых спецификаций.

Для установления контрольных пределов в **CUSUM**-картах в работе Barnhard (1959) было предложено использовать так называемую **V**-*маску*, которая наносится на график после построения точки для последней выборки (самой правой точки на графике). Можно считать, что **V**-маска представляет собой верхний и нижний контрольный пределы для накопленных сумм. Однако, вместо того, чтобы быть параллельными центральной линии, эти прямые сходятся под определенным углом вправо,

образуя в результате фигуру, похожую на лежащую букву **V**. Если график накопленной суммы пересекает любую из линий маски, то процесс считается вышедшим из-под контроля.

Контрольная карта скользящего среднего (**MA**-карта). Возвращаясь к примеру с размером поршневых колец, предположим, что наибольший интерес для инженера по контролю качества представляет обнаружение малых трендов последовательных выборочных средних. Например, необходимо обнаружить износ оборудования, который приводит к медленному, но постоянному ухудшению качества (т.е. отклонению размеров изделий от требований плановой спецификации. Одним из способов отслеживания таких трендов и обнаружения незначительных постоянных сдвигов среднего значения процесса является построение описанной выше **CUSUM**-карты. Другой способ состоит в использовании одной из схем установления весов данных, согласно которой осуществляется суммирование нескольких средних. При движении такого взвешенного среднего вдоль выборочных точек получается контрольная карта скользящего среднего, приведення на следующем рисунке.

Контрольная карта экспоненциально взвешенного скользящего среднего (**EWMA**-карта). Идея построения скользящих средних для последовательных (соседних) выборочных значений может быть обобщена. В принципе, чтобы обнаружить тренд, необходимо присвоить веса следующим друг за другом выборочным значениям, получив таким образом скользящее среднее. Однако, вместо простого арифметического скользящего среднего, можно найти геометрическое скользящее среднее (соответствующая контрольная карта показана на следующем рисунке и называется картой геометрического скользящего среднего, см. работу Montgomery,1985, 1991).

Регрессионные контрольные карты. Иногда может понадобиться обнаружить взаимосвязь между двумя различными параметрами производственного процесса. Например, руководство почтовой организации может захотеть узнать, сколько человеко-часов тратится на обработку некоторого объема корреспонденции. Эти две анализируемые переменные должны быть приблизительно линейно связаны друг с другом. Тогда эту взаимосвязь можно описать с помощью широко известного коэффициента корреляции Пирсона r. Описание свойств этой статистки можно найти в разделе Основные статистики. На регрессионной контрольной карте строится линия регрессии, которая выражает линейную взаимосвязь между двумя рассматриваемыми переменными. На карту также наносятся точки данных для всех наблюдений. Вокруг линии регрессии строится доверительный интервал, в который должна попадать

определенная доля выборки (например, 95%). Присутствие выбросов на этом графике будет свидетельствовать о том, что для некоторых выборок не соблюдается общая тенденция взаимосвязи, которая характерна для рассматриваемых переменных.

Для регрессионных контрольных карт существует множество областей применения. Так, например, профессиональные аудиторы могут с помощью карт данного типа обнаружить, у каких розничных торговцев число наличных трансакций превышает ожидаемое для данного уровня общего объема продаж или выделить те бакалейные магазины, в которых для существующего уровня продаж число погашенных купонов, дающих покупателю право на премию из ассортимента магазина при накоплении определенного числа купонов, превышает ожидаемое. В обоих случаях выбросы на регрессионных контрольных картах (т.е. слишком большое число наличных платежей, слишком большой объем погашенных купонов) могут привлечь к себе внимание и служить основанием для более тщательной проверки.

Контрольные карты Парето. На практике оказывается, что равномерное распределение нарушения качества на различных стадиях производственного процесса или на различных предприятиях, выпускающих продукт, встречается довольно редко. Скорее, причиной большинства проблем является наличие лишь нескольких "паршивых овец в стаде". Данный принцип стал широко известен под названием принципа Парето и утверждает, что потери качества столь "плохо" распределены, что малое число возможных причин его ухудшения отвечает за большинство возникающих проблем. К примеру, вполне возможно, что в основном загрязнение воздуха возникает из-за относительно небольшого числа "грязных" автомобилей. Или, в большинстве компаний основное число убытков является следствием неудачи с одним или двумя выпускаемыми продуктами. Для выявления "паршивых овец в стаде" строят контрольные карты Парето.

Они представляют собой гистограммы, на которых показано распределение потерь от ухудшения качества (например, в долларах) по некоторым категориям. Обычно категории - причины потери качества - приводятся в нисходящем порядке значимости (по частоте возникновения, стоимости в долларах и т.д.). Очень часто карта Парето помогает определить, на что направить усилия по улучшению качества продукта.

12. РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

12.1. Типографские издания

1. Тейлор Дж. Введение в теорию ошибок. Пер. с англ. – М.: Мир, 1985. –272 с.: ил.
2. Большаков В. Д. Теория ошибок наблюдений: Учебник для ВУЗов. – 2-е изд., перераб. и доп. М., Недра, 1983. 223 с.
3. Вукулов Э. А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA, EXCEL: учебное пособие. – М.: ФОРУМ: ИНФРА-М, 2004. 464 с.
4. Колмогоров А. Н. Основные понятия теории вероятностей. Серия «Теория вероятностей и математическая статистика» М.: – 1974. 120 с.: ил.
5. Айвозян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983. – 420 с.: ил.
6. Сергиенко А. Е. Цифровая обработка сигналов: Учебник для вузов. 2-е СПб.: Питер. 2006. – 752 с.: ил.

12.2. Электронные ресурсы

7. Большая Советская Энциклопедия. Российская государственная библиотека [Электронный ресурс] — Электрон. дан. — М. : Рос. гос. б-ка, 2010— Режим доступа: http://slovari.yandex.ru/~книги/БЭС/, свободный. — Загл. с экрана. — Яз. рус., англ.
8. StatSoft Режим доступа: http://www.statsoft.ru/ свободный. StatSoft Russia 2010.
9. Wolfram Research, Inc. (http://www.wolfram.com/). Официальный сайт компании «Wolfram», производителя «Mathematica».
10. Быстрое преобразование Фурье. [Электронный ресурс] Режим доступа: свободный. http://psi-logic.narod.ru/fft/fft.htm.
11. БПФ комплексной функции [Электронный ресурс] Режим доступа: свободный. http://alglib.sources.ru/fft/fft.php.
12. Короткий С. Нейронные сети: основные положения. [Электронный ресурс] Режим доступа: свободный. http://www.gotai.net/documents/doc-nn-002.aspx

СПбГУ ИТМО стал победителем конкурса инновационных образовательных программ вузов России на 2007–2008 годы и успешно реализовал инновационную образовательную программу «Инновационная система подготовки специалистов нового поколения в области информационных и оптических технологий», что позволило выйти на качественно новый уровень подготовки выпускников и удовлетворять возрастающий спрос на специалистов в информационной, оптической и других высокотехнологичных отраслях науки. Реализация этой программы создала основу формирования программы дальнейшего развития вуза до 2015 года, включая внедрение современной модели образования.

13. КАФЕДРА ПРОЕКТИРОВАНИЯ КОМПЬЮТЕРНЫХ СИСТЕМ

1945-1966 РЛПУ (кафедра радиолокационных приборов и устройств). Решением Советского правительства в августе 1945 г. в ЛИТМО был открыт факультет электроприборостроения. Приказом по институту от 17 сентября 1945 г. на этом факультете была организована кафедра радиолокационных приборов и устройств, которая стала готовить инженеров, специализирующихся в новых направлениях радиоэлектронной техники, таких как радиолокация, радиоуправление, теленаведение и др. Организатором и первым заведующим кафедрой был д.т.н., профессор С.И. Зилитинкевич (до 1951 г.). Выпускникам кафедры присваивалась квалификация инженер-радиомеханик, а с 1956 г. – радиоинженер (специальность 0705).

В разные годы кафедрой заведовали доцент Б.С. Мишин, доцент И.П. Захаров, доцент А.Н. Иванов.

1966–1970 КиПРЭА (кафедра конструирования и производства радиоэлектронной аппаратуры). Каждый учебный план специальности 0705 коренным образом отличался от предыдущих планов радиотехнической специальности своей четко выраженной конструкторско-технологической направленностью. Оканчивающим

институт по этой специальности присваивалась квалификация инженер-конструктор-технолог РЭА. Заведовал кафедрой доцент А.Н. Иванов.

1970–1988 КиПЭВА (кафедра конструирования и производства электронной вычислительной аппаратуры). Бурное развитие электронной вычислительной техники и внедрение ее во все отрасли народного хозяйства потребовали от отечественной радиоэлектронной промышленности решения новых ответственных задач. Кафедра стала готовить инженеров по специальности 0648. Подготовка проводилась по двум направлениям – автоматизация конструирования ЭВА и технология микроэлектронных устройств ЭВА. Заведовали кафедрой: д.т.н., проф. В.В. Новиков (до 1976 г.), затем проф. Г.А. Петухов.

1988–1997 МАИ (кафедра микроэлектроники и автоматизации проектирования). Кафедра выпускала инженеров-конструкторов-технологов по микроэлектронике и автоматизации проектирования вычислительных средств (специальность 2205). Выпускники этой кафедры имеют хорошую технологическую подготовку и успешно работают как в производстве полупроводниковых интегральных микросхем, так и при их проектировании, используя современные методы автоматизации проектирования. Инженеры специальности 2205 требуются микроэлектронной промышленности и предприятиям-разработчикам вычислительных систем. Кафедрой с 1988 г. по 1992 г. руководил проф. С.А. Арустамов, затем снова проф. Г.А. Петухов.

С **1997 ПКС** (кафедра проектирования компьютерных систем). Кафедра выпускает инженеров по специальности 210202 «Проектирование и технология электронно-вычислительных средств». Область профессиональной деятельности выпускников включает в себя проектирование, конструирование и технологию электронных средств, отвечающих целям их функционирования, требованиям надежности, проекта и условиям эксплуатации. Кроме того, кафедра готовит специалистов по защите информации, специальность 090104 «Комплексная защита объектов информатизации». Объектами профессиональной деятельности специалиста по защите информации являются методы, средства и системы обеспечения защиты информации на объектах информатизации.

С 1996 г. кафедрой заведует д.т.н., профессор Ю.А. Гатчин.

За время своего существования кафедра выпустила 4264 инженеров. На кафедре защищено 65 кандидатских и 7 докторских диссертаций.