МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ УКРАИНЫ

ЗАПОРОЖСКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ

Кафедра медицинской и фармацевтической информатики и НТ

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ РЕЗУЛЬТАТОВ МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Методическое руководство
для студентов фармацевтического факультета
специальностей 7.110201 «Фармация»,
7.110202 «Технология парфюмерно-косметических средств»

Запорожье 2015

УДК: 57.08:004(075.8)
ББК: 5+28]ся73
С 83

Страхова О. П.

Статистические методы обработки результатов медико-биологических исследований: методическое пособие для студентов фармацевтического факультета. – Запорожье : [ЗГМУ], 2015. – 99 с.

Рецензент:

Б.А. Прийменко профессор кафедры органической химии Запорожского государственного медицинского университета, доктор фармацевтических наук

Рекомендовано Центральной методической Радой Запорожского государственного медицинского университета (протокол №_6_ от_20/05/2015.)

В пособии рассмотрены основные понятия и определения важной части доказательной медицины – математической статистики. Они применимы в первичной обработке результатов исследований для количественных данных, представляющих собой выборки с нормальным и ненормальным распределением. Задачи такого типа свойственны медицине и фармации, они служат элементом доказательства наличия либо отсутствия различий между выборками; эффективности новых разработанных методов лечения пациентов; повышения качества производства лекарственных препаратов.

Обеспечено методическое обоснование процессов взаимодействия информации, данных и методов, представлены материалы, которые помогут студентам освоить базовую терминологию теории статистической обработки данных.

Пособие предназначено для студентов и преподавателей фармацевтического факультета

УДК: 57.08:004(075.8)
ББК: 5+28]ся73

Тема 1:Статистические методы обработки информации. Первичная статистическая обработка количественных признаков, оценка значимости их различия с помощью программы STATISTICA.

Цель работы: Получить навыки использования программного обеспечения для первичной обработки медико-биологической информации и проведения регрессионного и корреляционного анализа с использованием MicrosoftExcel и Statistica.

План занятия.

1. Повторить теоретический материал.
2. Запустить демо-версию программы Statistica
3. Загрузить файл “FarmMarket-xx” с индивидуальным заданием.
4. Провести необходимые измерения и вычисления, занести результаты в контрольный файл “Result-xx”.
5. Подготовить отчет о выполненной работе с интерпретацией результатов обработки экспериментальных данных и отправить его по электронной почте на электронный адрес преподавателя strahova@zsmu.zp.ua.
6. Пройти тестирование по итогам занятия.

Студент должен уметь:

- подготовить данные для первичной статистической обработки;
- интерпретировать статистические показатели;
- формулировать гипотезу, которая доказывается с помощью методов статистики;
- провести оценку признаков с помощью статистических критериев.

Студент должен знать:

- основытеориивероятности,
- методысозданиярепрезентативнойвыборки;
- основные параметры оценки статистической выборки,
- основныевидыстатистическогоанализа.

Основные понятия:

- статистическийанализрегрессионный, корреляционный;
- критерийСтьюдента, Фишера;
- нормальноестатистическоераспределение.

Необходимое материальное и программное обеспечение:

- персональныйкомпьютер,
- демо-версияППП Statistica, Microsoft Excel,
- локальнаясеть,
- доступ к электроннойбиблиотеке,
- весы медицинские, тонометр, фонендоскоп, метр.

Медицина - экспериментальная наука, методом исследования в ней является эксперимент (от лат. *experimentum* — проба, опыт). При этом, эксперимент может быть активным, когда исследователь сам каким-либо образом вмешивается в ход процессов, заранее планируя или определяя уже в ходе воздействия на биологический объект направление этого воздействия, его силу, временной интервал, в течение которого проводится эксперимент, и т.д. Примером активного эксперимента может быть подбор оптимальной концентрации действующих веществ в препарате. Так как контролируемые параметры изменяются по заранее известным правилам,

исследователь, получив результаты своего эксперимента, может создать математическую модель процесса. Это даст ему возможность с определенной вероятностью прогнозировать дальнейшие изменения в объекте, испытывающем такое направленное воздействие.

Эксперимент может быть также пассивным, когда исследователь лишь фиксирует наступающие изменения, не оказывая на объект или на происходящие процессы никакого направленного влияния, например, как в случае с применением специфического препарата для лечения определенного заболевания. В такой ситуации исследователь проводит мониторинг наблюдаемого объекта и явлений, протекающих в нем, что также дает возможность построения математической модели и прогнозирования развития дальнейших изменений объекта.

Изучаемые в медицине явления являются сложными системами, функционирующими при воздействии на них множества входных факторов. Часть таких факторов – контролируемые, измеряемые количественно, оцениваемые в баллах. Другие – неконтролируемые, случайные, зачастую неизвестные, не поддаются измерению, но оказывают воздействие на систему, результатом чего является случайность ее состояния и функционирования.

Определения

В силу того, что неконтролируемые и случайные факторы для каждого объекта наблюдений принимают различные случайные значения, выходные параметры, характеризующие состояние и функционирование сложной стохастической системы, являются случайными величинами, для исследования которых следует применять методы теории вероятности и математической статистики.

Статистический анализ сложной системы включает**:**

-статистическое описание переменных;

-оценку гипотез о значимости различия показателей в различных группах объектов;

-определение количественной оценки связи между входными контролируемыми факторами и выходными параметрами;

- моделирование выходных параметров для их прогнозирования при определенных значениях входных факторов;

-применение всего арсенала многомерных исследований систем (регрессионный, дисперсионный, дискриминантный и др. методы анализа).

Множество объектов изучаемого явления называется генеральной совокупностью. Сплошное наблюдение всех объектов генеральной совокупности (ГС) проводится редко, например, при ежедневной регистрации всех больных, обратившихся за медицинской помощью в поликлинику или ведении историй болезни на всех больных, находящихся на стационарном лечении. В научных целях чаще используют выборочный метод наблюдения, в котором используется только часть объектов ГС, по результатам анализа которой делают выводы обо всей ГС. Часть объектов, отобранных из ГС по определенным правилам, называется выборкой, или выборочной совокупностью.

Чтобы выводы, полученные в результате анализа выборки, адекватно отражали свойства ГС, выборка должна быть репрезентативной (представительной). Её можно сформировать при выполнении двух требований:

- случайность отбора объектов однородной ГС в выборку, когда каждый объект ГС должен иметь одинаковую вероятность попадания в выборку;

- выборка должна иметь достаточную численность независимых наблюдений.

Выборкой $x_1, ..., x_n$ объема n из совокупности называется n независимых наблюдений над случайной величиной ξ с функцией распределения $F(x)$.

Вариационным рядом $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ называется выборка, записанная в порядке возрастания ее элементов.

Выборки больших объемов труднообозримы. Разобьем диапазон значений выборки на равные интервалы и подсчитаем для каждого интервала **частоту**- количество наблюдений, попавших в него. Частоты, отнесенные к общему числу наблюдений n, называют **относительными частотами.** Графическое представление распределения частот по интервалам называют **гистограммой.** Н**акопленной частотой** для данного интервала называют сумму частот данного интервала и всех тех, что левее его.

Числовые характеристики эмпирического распределения называются **выборочными характеристиками**:

среднее (математическое ожидание):

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

дисперсия:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2;$$

выборочный **момент** порядка k:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k;$$

выборочные **квантили** ζ_p порядка p - корни уравнения

$$F(\zeta_p)=p,$$

которыми являются члены вариационного ряда

$$\zeta_{(p)}=\xi_{([np]+1)},$$

где *[np]* означает целую часть *np*; частным случаем (p = 0.5) является выборочная **медиана -** центральный член вариационного ряда. Значение выборочных характеристик состоит в том, что при $n \to \infty$ они стремятся к истинным значениям распределения *F(x).*

Число случаев наблюдений в выборке N называется объемом выборки.

Экспериментально установлено, что надежные результаты статистического анализа можно получить, если число случаев наблюдений больше в 3-5 раз числа входных контролируемых факторов и выходных параметров.

Основными задачами статистического описания переменных являются:

-определение числовых характеристик переменных и оценка их точности и надежности;

-определение статистических рядов распределения переменных и оценка их соответствия теоретическим законам распределения;

-оценка значимости различия показателей в зависимых и несвязанных выборках.

По таким числовым характеристикам, как среднее арифметическое, среднее квадратичное отклонение, средняя квадратичная ошибка среднего значения определяют доверительные интервалы, и оценивается значимость различий показателей в различных условиях.

Оценка значимости различия показателей в независимых и связанных выборках – одна из основных задач, решаемых исследователями при сравнении методов профилактики, лечения заболеваний и т.д.

Числовые характеристики переменных подразделяются на три вида:

- характеристики положения;
- характеристики рассеяния:
- характеристики вида распределения.

К характеристикам положения относятся:

- среднее арифметическое значение;
- медиана;
- мода;
- среднее геометрическое значение;
- среднее гармоническое значение.

Среднее арифметическое – характеристика центра дискретного ряда рассчитывается по формуле

$$\overline{x} = \frac{\sum_{j=1}^{N} x_j}{N},$$

медиана соответствует варианте, стоящей в середине ранжированного ряда. её положение в ряду определяется номером

$$N_{Me} = \frac{N+1}{2},$$

где n – число единиц совокупности.

Мода – наиболее часто встречающееся или повторяющееся значение в массиве или интервале данных. как и функция медиана, функция мода является мерой взаимного расположения значений.

её величину определяют по формуле:

$$M_o = x_{Mo} + h\frac{f_{Mo} - f_{Mo-1}}{[f_{Mo} - f_{Mo+1}] + [f_{Mo} - f_{Mo-1}]},$$

где x_{Mo}- нижняя граница модального интервала, f_{Mo} - частота, соответствующая модальному интервалу, f_{Mo-1} - предмодальная частота, f_{Mo+1} - послемодальная частота.

В наборе значений мода — это наиболее часто встречающееся значение; медиана — это значение в середине массива; среднее — это среднее арифметическое значение. ни одна из этих величин не характеризует в полной мере то, в какой степени центрированы данные.

К характеристикам рассеяния значений переменной относятся:

- минимальное и максимальное значение; Хмин и Хмах
- размах вариационного ряда – Р= Хмах-Хмин;
- дисперсия – σ^2
- среднее квадратичное (стандартное) отклонение S;
- 25% (ЛК)и 75% (УК) квартили и межквартильный размах МР= УК-ЛК;
- средняя квадратичная ошибка среднего значения Мх
- 95% доверительный интервал истинного среднего значения.

Доверительный интервал для некоторой величины - это диапазон вокруг значения величины, в котором находится истинное значение этой величины (с определенным уровнем доверия).

Вид распределения характеризуют коэффициенты:

- асимметрии в натуральном и стандартизованном виде As
- эксцесса в натуральном и стандартизованном виде Es.

Асимметрия характеризует степень несимметричности распределения относительно его среднего. Положительная асимметрия указывает на отклонение распределения в сторону положительных значений.

Отрицательная асимметрия указывает на отклонение распределения в сторону отрицательных значений.

Эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение. Отрицательный эксцесс обозначает относительно сглаженное распределение. Иными словами, эксцесс определяет наличие выбросов в значениях.

По числовым характеристикам судят о соответствии эмпирического распределения теоретическому нормальному распределению. Его можно оценить как близкое к нормальному, если:

- среднее арифметическое, геометрическое и гармоническое значения незначительно различаются друг от друга, а также с модой и медианой;
- минимальные и максимальные значения примерно равноудалены от среднего значения;
- стандартизованные коэффициенты асимметрии и эксцесса по абсолютной величине меньше |2|.

Любое исследование должно включать элемент оценки точности и надежности числовых характеристик. Оценкой точности и надежности является 95% доверительный интервал истинного среднего значения. Например, истинное среднее значение ГС находится в доверительном интервале

$$M_{95} = \bar{x} \pm t_{95} * m_{x\bar{x}}$$

где t_{95} - табличное значение t –критерия Стьюдента, отвечающее доверительной вероятности 95% по числу степеней свободы df= n-1, n – количество наблюдений;

$m_{\bar{x}}$ - средняя квадратичная ошибка среднего значения, определяемая по формуле

$$m_x = \frac{S_x}{\sqrt{n}}$$

где Sx – среднее квадратичное отклонение показателя в выборке.

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

Как видим, ошибка уменьшается с увеличением объема выборки. Так, чтобы уменьшить ошибку в два раза, число наблюдений надо увеличить в четыре раза.

В ряде случаев целесообразно определять 95% доверительный интервал для возможных значений показателя.

$$X = \bar{x} \pm t_{95} * S_x$$

Закон нормального распределения случайной переменной адекватно описывает случайные величины, формирующиеся под влиянием большого числа статистически независимых факторов, когда ни один из них не доминирует над остальными.

Предполагать нормальность распределения результатов медико-биологических наблюдений априори нет оснований. Следовательно, нормальность надо проверять. Предварительные выводы о виде распределения переменной можно сделать по статистическому ряду распределения, гистограмме и кумулятивной линии, являющимися аналогами функций плотности распределения и интегральной функции распределения. Для окончательного суждения о соответствии эмпирического распределения определенному теоретическому закону распределения применяют специальные критерии - Пирсона, Стьюдента, Колмогорова-Смирнова и т.д. Рассмотрим критерий Стьюдента.

Критерий Стьюдента был разработан английским химиком У.Госсетом, когда он работал на пивоваренном заводе Гиннеса и по условиям контракта не имел права открытой публикации своих исследований. Поэтому публикации своих статей по t-критерию У.Госсет сделал в 1908г. в журнале "Биометрика" под псевдонимом "Student", что в переводе означает "Студент". Коварная простота вычисления t-критерия Стьюдента, а также его наличие в большинстве статистических пакетов и программ привели к широкому использованию этого критерия даже в тех условиях, когда применять его нельзя.

Наиболее часто t -критерий Стьюдента используется в двух случаях. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух независимых, несвязанных выборок (так называемый двухвыборочный t-критерий). В этом случае есть контрольная группа и опытная группа, состоящая из разных пациентов, количество которых в группах может быть различно. Во втором же случае используется так называемый парный t-критерий, когда одна и та же группа объектов порождает числовой материал для проверки гипотез о средних. Поэтому эти выборки называют зависимыми, связанными.

t-критерий является наиболее часто используемым методом обнаружения различия между средними двух выборок. Например, t-критерий можно использовать для сравнения средних показателей группы пациентов, принимавших определенное лекарство, с контрольной группой, где принималось безвредное лекарство. Теоретически, t-критерий может применяться, даже если размеры выборок очень небольшие (например, 10; некоторые исследователи утверждают, что можно исследовать выборки меньшего размера), и если переменные нормально распределены (внутри групп), а дисперсии наблюдений в группах не слишком различны. Предположение о нормальности можно проверить, исследуя распределение

(например, визуально с помощью гистограммы) или применяя какой-либо критерий нормальности. **Если условия применимости t-критерия не выполнены, следует использовать непараметрические альтернативы t-критерия**

Степень различия между средними в двух группах зависит от внутригрупповой вариации (дисперсии) переменных. В зависимости от того, насколько различны эти значения для каждой группы, "грубая разность" между групповыми средними показывает более сильную или более слабую степень зависимости между независимой (группирующей) и зависимой переменными. Например, если среднее число лейкоцитов равнялось 102 для мужчин и 104 для женщин, то разность внутригрупповых средних только на величину 2 будет чрезвычайно важной, когда все значения мужчин лежат в интервале от 101 до 103, а все значения женщин - в интервале 103 - 105. В этом случае можно довольно хорошо предсказать значение зависимой переменной, исходя из пола субъекта (независимой переменной). Однако если та же разность 2 получена из сильно разбросанных данных (например, изменяющихся в пределах от 0 до 200), то этой разностью вполне можно пренебречь. Таким образом, можно сказать, что уменьшение внутригрупповой вариации увеличивает чувствительность критерия.

t-критерий для зависимых выборок очень полезен в тех довольно часто возникающих на практике ситуациях, когда важный источник внутригрупповой вариации (или ошибки) может быть легко определен и исключен из анализа. Например, это относится к экспериментам, в которых две сравниваемые группы основываются на одной и той же совокупности наблюдений (субъектов), которые тестировались дважды (например, до и после лечения, до и после приема лекарства). В подобных экспериментах значительная часть внутригрупповой изменчивости (вариации) в обеих группах может быть объяснена индивидуальными различиями субъектов.

Однако в случае независимых выборок, вы ничего не сможете поделать с этим, т.к. не сможете определить (или "удалить") часть вариации, связанную с индивидуальными различиями субъектов. Если та же самая выборка тестируется дважды, то можно легко исключить эту часть вариации. Вместо исследования каждой группы отдельно и анализа исходных значений, можно рассматривать просто разности между двумя измерениями (например, "до приема лекарства" и "после приема лекарства") для каждого субъекта. Вычитая первые значения из вторых (для каждого субъекта) и анализируя затем только эти "чистые (парные) разности", вы исключите ту часть вариации, которая является результатом различия в исходных уровнях индивидуумов. Именно так и проводятся вычисления в t-критерии для зависимых выборок. В сравнении с t-критерием для независимых выборок, такой подход дает всегда "лучший" результат (критерий становится более чувствительным).

Чтобы применить t-критерий для независимых выборок, требуется, по крайней мере, одна независимая (группирующая) переменная (например, признак опытной и контрольной групп) и одна зависимая переменная (например, тестовое значение некоторого показателя, кровяное давление, число лейкоцитов и т.д.). Исследователь выдвигает гипотезу H0 (нулевую) о соответствии закона распределения данной зависимой переменной нормальному. Эту гипотезу принимают, если ее вероятность (уровень значимости p) будет больше 0.05, и отвергают, если ее вероятность будет меньше или равна 0.05. В этом случае следует подыскать для описания переменной более подходящий закон распределения.

Теория проверки статистических гипотез является основным инструментом доказательной медицины. Доказа́тельная медици́на (англ. *Evidence-based medicine — медицина, основанная на доказательствах*) — термин описывает такой подход к медицинской практике, при котором

решения о применении профилактических, диагностических и лечебных мероприятий принимаются исходя из полученных доказательств их эффективности и безопасности, и предполагающий поиск, сравнение, обобщение и широкое распространение полученных доказательств для использования в интересах больных

При сравнении показателей, например, в контрольной (прием плацебо) и опытной (прием настоящего препарата) группах выдвигают статистические гипотезы:

H1 – о существенном различии показателя в опытной и контрольной группах;

H0- нулевую гипотезу – о равенстве (отсутствии различий) показателя в опытной и контрольной группах.

Гипотезу H1 принимают, если ее вероятность имеет значение равное или больше 95% и отклоняют, если ее вероятность будет меньше 95%. В этом случае принимают гипотезу H0, а ее вероятность, как альтернативной, будет $p>0.05$.

Вероятность H0 p называют уровнем значимости, а величину 1-p называют доверительной вероятностью гипотезы H1 (альтернативной). Статистическая значимость результата – это мера уверенности в его "истинности". Вопрос, который необходимо при этом задавать: "Насколько можно доверять этому результату?". Представьте, что мы проводили исследование на основе только двух пациентов. Конечно же, в этом случае к результатам нужно относиться с опасением. Если же были обследовано большое количество больных, то сделанным выводам уже можно доверять. Степень доверия и определяется значением p-уровня.

Более высокий р- уровень соответствует более низкому уровню доверия к результатам, полученным при анализе выборки. Например, р-уровень, равный 0.05 (5%) показывает, что сделанный при анализе некоторой группы вывод является лишь случайной особенностью этих объектов с вероятностью только 5%.

Другими словами, с очень большой вероятностью (95%) вывод можно распространить на все объекты.

Во многих исследованиях 5% рассматривается как приемлемое значение р-уровня. Это значит, что если, например, р = 0.01, то результатам доверять можно, а если р=0.06, то нельзя.

Независимыми (несвязанными) называются выборки, в каждой из которых наблюдаются различные объекты, например, первая, контрольная группа, принимающая плацебо, и вторая, опытная, получающая определенный препарат.

По 95% -м доверительным интервалам дается приближенное графическое решение. Если доверительные интервалы не перекрывают друг друга или их перекрытие не превышает 1/3, можно считать, что имеется значимое различие средних значений показателей в двух выборках. Если перекрытие доверительных интервалов превышает 1/3, следует признать, что различие средних значений показателя в этих двух выборках незначимое, недостоверное. Однако, приближенный метод оценки значимости различия по доверительным интервалам может использоваться в качестве экспресс-метода, он хорош для графической демонстрации средних значений признаков и 95% - х доверительных интервалов их истинных значений. Более обоснованное решение получают с применением критерия Стьюдента.

Надежные значения критерия можно получить по формуле:

$$t = \frac{\left| \bar{X}_1 - \bar{X}_2 \right|}{\sqrt{\frac{(S_1^2 * (n_1 - 1) + S_2^2 * (n_2 - 1)) * (n_1 + n_2)}{(n_1 + n_2 - 2) * n_1 * n_2}}}$$

Связанными (или зависимыми) называют выборки, состоящие из одних и тех же объектов, наблюдающихся в различных условиях, например, до некоторого воздействия и после него, или в период разгара заболевания, на 3, 9 и т.д. дни лечения.

Исходными данными для решения служат числовые характеристики разностей показателя, получаемые по исходной матрице наблюдений. Расчет критерия Стьюдента проводится по формуле:

$$t = \frac{\left| \Delta \bar{x} \right|}{m_{\Delta \bar{x}}}$$

где $\Delta \bar{x}$ - средняя разность показателя в сравниваемых группах;

$m_{\Delta \bar{x}}$ - средняя квадратичная ошибка средней разности показателей;

$m_{\Delta \bar{x}} = \frac{S_{\Delta x}}{\sqrt{n}}$ **,** где $S_{\Delta x}$ - среднее квадратичное отклонение разности показателей.

Полученное расчетное значение критерия Стьюдента следует сравнить с табличным, определенным для данного числа степеней свободы, при заданном уровне значимости, который обычно равен или менее 0.05. Вывод о том, что **различие показателя** в сравниваемых парах связанных выборок

значимо, можно сделать**, если табличное** значение критерия Стьюдента **меньше рассчитанного** исследователем.

Величина t–критерия значимости различия p зависит от разности $\left|\bar{x}_1 - \bar{x}_2\right|$, Δx и числа наблюдений в выборках n_1,n_2.

При небольших объемах выборки увеличиваются средние квадратичные ошибки m_{x_i} и уменьшается величина t-критерия, а следовательно уменьшается вероятность гипотезы H1 о различии, увеличивается вероятность гипотезы H0 о соответствии показателя в сравниваемых выборках. При желании исследователя получить значимое различие показателя следует увеличивать число наблюдений в выборках. Так, при заданном уровне значимости p=0.05 требуемое число наблюдений должно удовлетворять следующим требованиям:

-при сравнении независимых выборок 1 и 2

$$n_1 \text{ и } n_2 \geq \frac{t_{05}^2 * (S_{x_1}^2 + S_{x_2}^2)}{(\bar{x}_1 - \bar{x}_2)^2};$$

- при сравнении связанных выборок

$$n \geq \frac{t_{05}^2 * S_{\Delta x}^2}{(\Delta x)^2}$$

Например, при получении незначимого различия показателя в независимых выборках

1. $\bar{x}_1$ =10, S_1 = 5, n_1=25;
2. $\bar{x}_2$ = 12, S_2 = 6, n_2 = 25,

для получения значимого различия с p< 0.05 (при f = n_1+ n_2 -2 = 48 - t_{05} ~ 2.00; получено из Таблицы значений критерия Стьюдента) следует иметь число наблюдений

$$\mathbf{n_1} \text{ и } \mathbf{n_2} \geq \frac{2^2 * (5^2 + 6^2)}{(10-12)^2} = 6\ 1 \text{ наблюдение.}$$

При количестве наблюдений в каждой выборке не менее 61 возможно принятие гипотезы Н1 о значимом различии показателей

Ход работы.

Для выполнения лабораторной работы, необходимо получить файл с заданием. Порядковый номер файла соответствует номеру варианта.

Пример. Исследовали влияние рекламы на динамику продаж лекарственного препарата, применяемого при симптоматическом лечении сезонных респираторных заболеваний.

Для выявления этого влияния наблюдался показатель - количество проданных упаковок препарата.

В контрольную группу отобрано 15 аптечных пунктов, не проводивших сезонных рекламных кампаний.

В исследовательской группе - 10 аптечных пунктов, разместивших на видных местах у входа баннеры- «раскладушки» с рекламой данного препарата на 1, 3 и 9-й день от установки баннера.

Таблица 1. Количество проданных упаковок препарата симптоматического лечения острых респираторных заболеваний в контрольной и исследовательской группе на 1, 3 и 9 дни от появления баннеров перед входом в аптечные пункты.

№ п/п	X1	Исследовательскаягруппа		
	Контроль	X2на 1-й день	X3на 3-й день	X4 на 9-й день
1	2	28	15	5
2	5	35	13	3
3	3	40	19	8

4	0	25	5	3
5	1	33	18	7
6	5	42	18	8
7	3	19	5	4
8	2	21	10	5
9	8	28	16	2
10	1	31	15	2
11	0			
12	6			
13	4			
14	2			
15	7			

Требуется:

I. Определить выборочные характеристики в каждой группе.

2. Оценить значимость различий характеристик в независимых и связанных выборках.

3. Сформулировать выводы.

Выполним данные действия с помощью пакета MicrosoftExcel.

Для выполнения этих действий следует перенести данные на лист1 книги MicrosoftExcel, озаглавив ее «FarmMarket», выбрать из предлагаемого списка функций «Статистические» и вычислить для данного набора значений:

1. Среднее арифметическое – «СРЗНАЧ»
2. Среднее квадратичное отклонение –«СТАНДОТКЛОН.В»,
3. Моду – «МОДА.ОДН»,
4. Медиану – «МЕДИАНА»,
5. Дисперсию – «ДИСП.В»
6. Среднее отклонение – «СРОТКЛ»,
7. Асимметрию – «СКОС»,
8. Эксцесс – «ЭКСЦЕСС»,
9. Максимальное значение - »МАКС»,
10. Минимальное значение – «МИН»,

11. Доверительный интервал – «ДОВЕРИТ.НОРМ»,
12. Количество значений в каждой выборке.

№ п/п	X1	Исследовательская группа		
	Контроль	X2 на 1-й день	X3 на 3-й день	X4 на 9-й день
1	2	28	15	5
2	5	35	13	3
3	3	40	19	8
4	0	25	5	3
5	1	33	18	7
6	5	42	18	8
7	3	19	5	4
8	2	21	10	5
9	8	28	16	2
10	1	31	15	2
11	0			
12	6			
13	4			
14	2			
15	7			

Таблица 2	Переменные			
Статистические характеристики	X1	X2	X3	X4
Минимум	0			
Максимум	8			
Мода	2			
Медиана	3			
Среднее арифметическое	3,27			
Среднеквадатичное откло	2,49			
Среднее отклонение	2,05			
Дисперсия	6,21			
Доверительный интервал	1,26			
Эксцесс	-0,75			
Асимметрия	0,48			
Количество значений n	15			

Результаты занести в таблицу на этом же листе.

Таблица 3. Несвязанные выборки

	X1	X2	X3	X4
Число наблюдений	15	10	10	10
Число степеней свободы		23		
Стьюдент.Тест (Хвосты=2, тип=2)		5,0616E-12	0,160837645	0,156296

Таблица 4. Связанные выборки

	X1	X2	X3	X4
Число наблюдений	15	10	10	10
Число степеней свободы			18	
Стьюдент.Тест (Хвосты=1, тип=1)			5,23206E-07	8,43E-05

Рис. 1. Лист файла MicrosoftExcel с исходными данными и примерами вычислений.

Затем для несвязанных и связанных выборок следует построить гистограммы и графики, нанести линии тренда и определить расчетные функции для них, пользуясь возможностями MicrosoftExcel. Их разместить на этом же листе файла «FarmMarket».

Для оценки значимости показателей в связанных и несвязанных выборках применим статистическую функцию «СТЬЮДЕНТ.ТЕСТ». Она используется, чтобы определить, насколько вероятно, что две выборки взяты из генеральных совокупностей, которые имеют одно и то же среднее.

Несвязанными выборками будут пары XI и X2, XI и X3, XI и X4; связанными выборками - X2 и X3, X2 и X4, X3 и X4 (объясните почему).

К параметрам вычисляемой функции относятся хвосты и тип теста. Хвосты — число хвостов распределения. **Если хвосты = 1, то число элементов в двух выборках одинаковое,** функция СТЬЮДЕНТ.ТЕСТ использует одностороннее распределение. Если хвосты = 2, то число элементов в выборках разное и функция СТЬЮДЕНТ.ТЕСТ использует двустороннее распределение.

Определим число степеней свободы для каждой выборки, обратив внимание на различие количества значений в контрольной и исследовательской группах.

Итоги размещаем в таблицах отдельно для связанных и несвязанных выборок рядом с исходными данными.

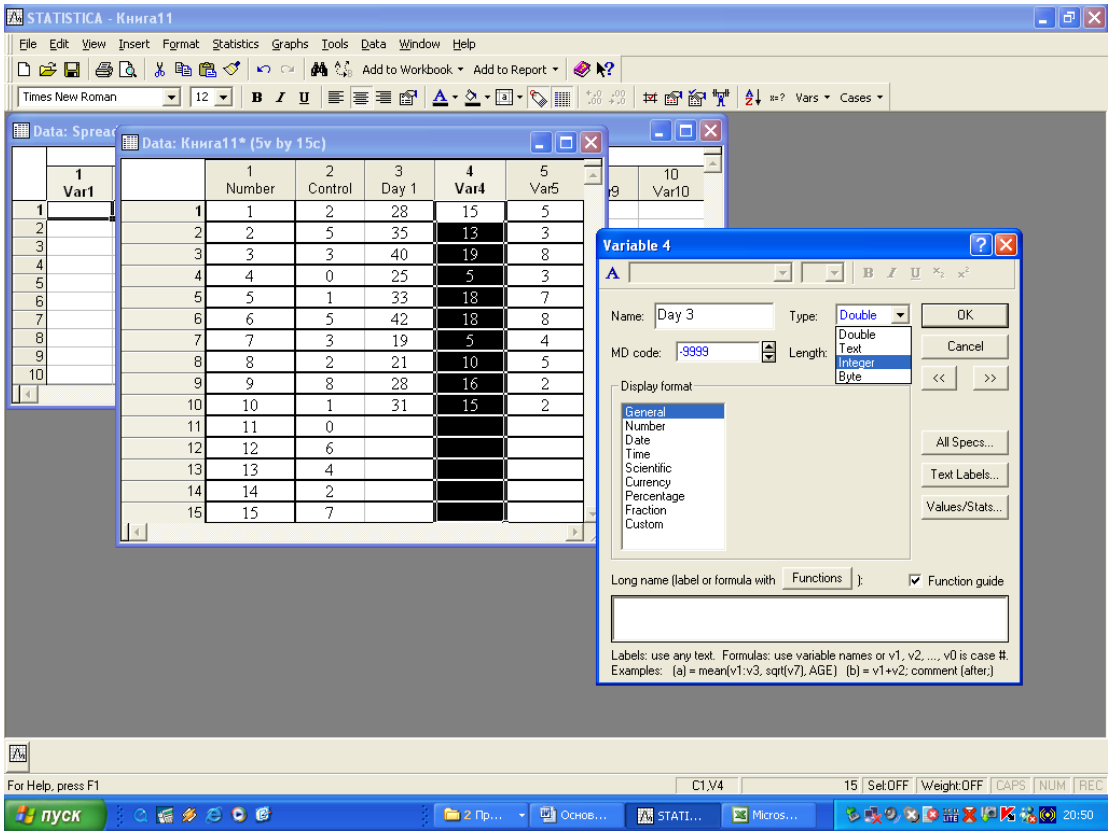
По результатам проведенных вычислений необходимо сделать выводы:

1. о существенности признака, выбранного для проведения исследования,
2. о его значимости для оценки эффективности влияния рекламных акций на изменение объема продаж препаратов,
3. о нормальности распределения данных в выборках,
4. о величинах доверительных интервалов, уровнях значимости признаков,

5. о правильности применения выбранного метода статистической обработки данных,

6. о верности или ложности выдвинутой гипотезы.Выводы следует поместить на втором листе файла «FarmMarket».

Рассмотрим аналогичные вычисления, выполняемые с помощью пакета «Статистика». Для этого данные следует перенести из таблицы, сформированной с помощью MicrosoftExcel, в таблицу пакета Статистика. Это делается открытием файла формата MicrosoftExcel в окне Статистики.



Наименования столбцов и характеристики находящихся в них данных можно изменить, открыв меню двойным щелчком на заголовке соответствующего столбца.

Рис. 2. Вид главного окна пакета Статистика.

Затем необходимо выбрать раздел – базовые статистики. Для этого выбираем на верхней линейке-меню Statistics элемент Basicstatistics/Tables и в

следующем раскрывшемся списке – Descriptivestatistics. Перечисляем переменные для обработки и выбираем вычисляемые значения.

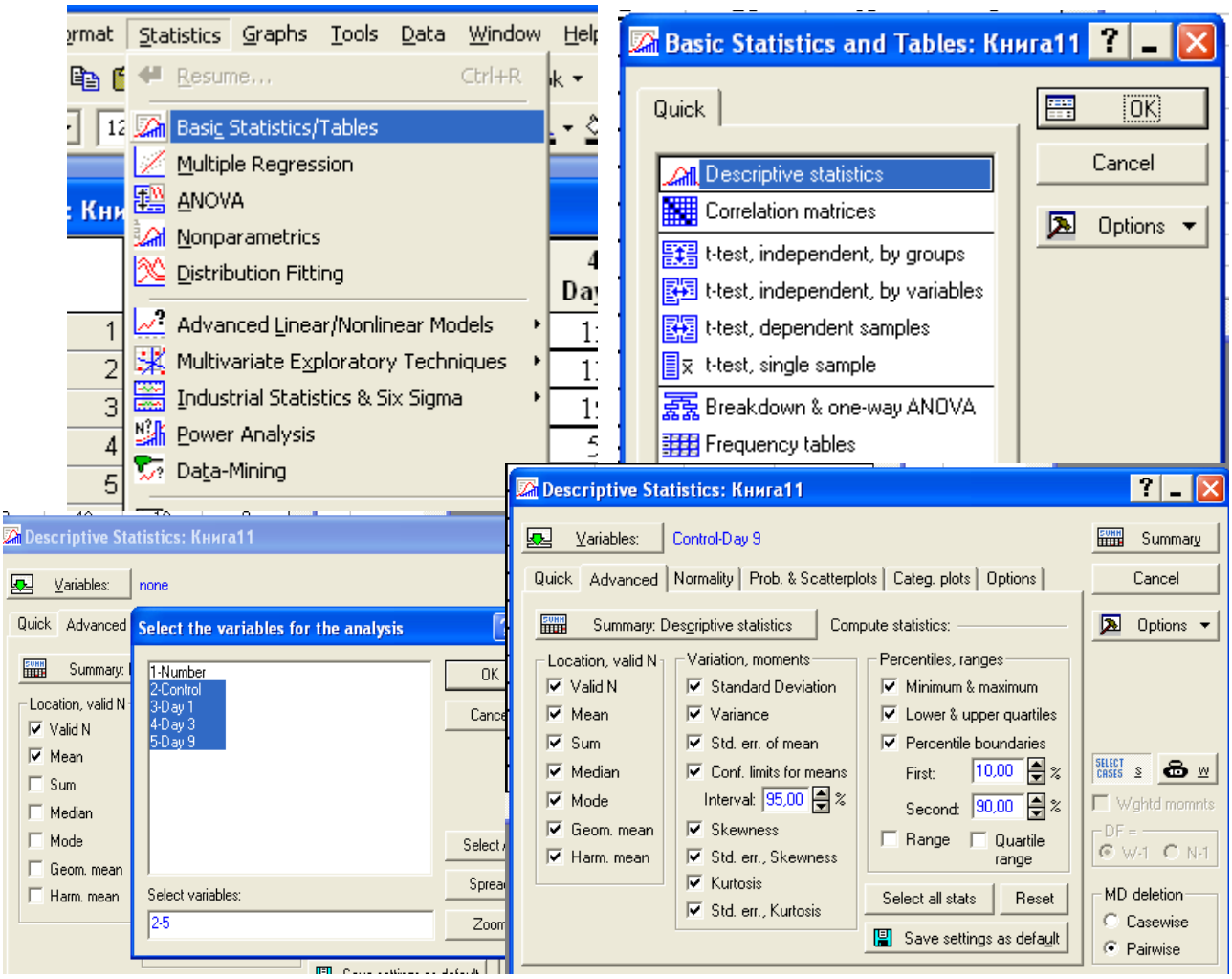


Рис. 3. Последовательно раскрывающиеся окна Статистики.

С помощью Статистики найдите значение Т-теста. Сравните между собой все значения, полученные двумя методами, и охарактеризуйте их. Совпадают ли полученные результаты? Сохраните созданный программой файл Workbook.

Данные можно визуализовать. Для этой цели в блоке базовых статистик существует возможность построения графиков. Выберем

Box&Whiskerplot, а затем Histograms , указав, для каких переменных необходимо строить график

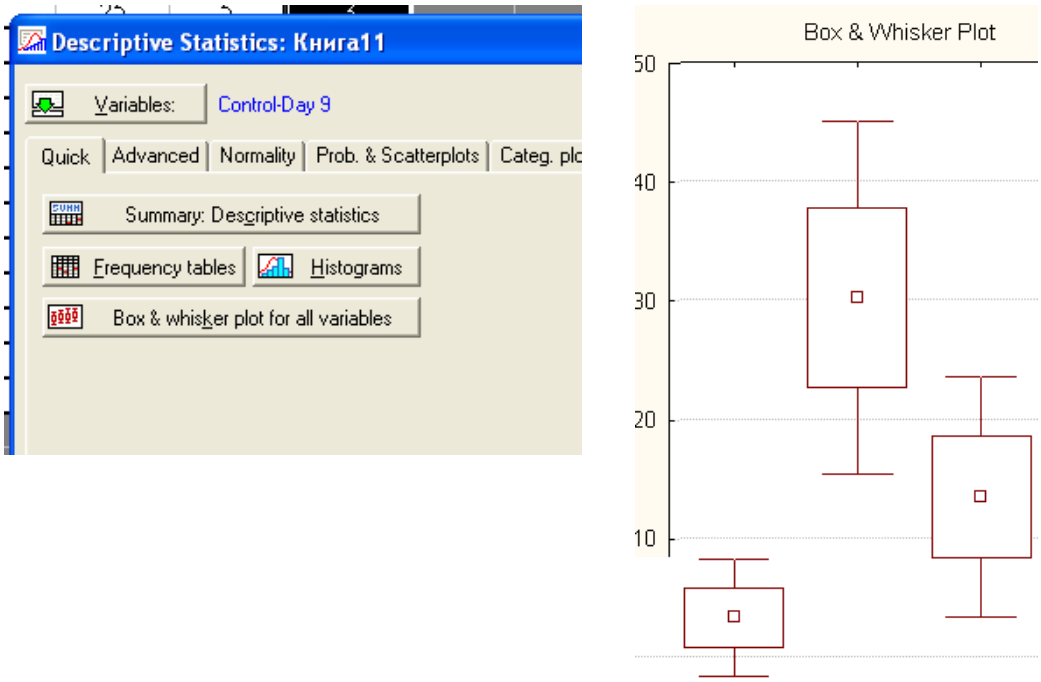


Рис. 4. Создание графика Бокса – Вискера.

На изображении внешние отрезки – минимальные и максимальные значения, прямоугольник задает границы нижнего и верхнего квартилей, маленький квадрат в середине – медиана.

Сохраните файл с рассчитанными значениями, графиком и гистограммой, созданный пакетом Статистика.

Сравните результаты, полученные при использовании MicrosoftExcel и Статистики.

Выводы занесите на второй лист файла Стат1.

Задание. Провести определение веса, артериального давления и частоты сердечных сокращений у всех студентов группы. Измерить размер запястья и талии. Результаты занести в таблицу 2 формата Excel и Statistica.

Таблица 2. Результаты антропометрических измерений у студентов группы.

	A	B	C	D	E	F	G	H	I
1	ФИО	Пол	Рост, см	Вес, кг	Сист.	Диаст.	ЧСС	Запястье	Талия
2									
3									

Сохраните таблицу. Пользуясь описанным выше примером выполнения статистической обработки данных, для всех измеренных величин определите:

- выборочные характеристики,
- вид распределения данных.

Постройте графики распределения данных, создайте модели распределения, определите виды полученных уравнений регрессии.

Установите, есть ли соответствие между ростом и весом человека, ростом и давлением, давлением и частотой сердечных сокращений, между полом студента и его весом, полом и окружностью талии, окружностью талии и величиной запястья.

Получите таблицы данных других групп. Сохраните их. Определите выборочные характеристики для всех групп. Результатысохраните в файледлядальнейшейстатистическойобработки.

Вопросы.

1. Цель применения статистической обработки данных.
2. Чтотакоегенеральнаясовокупность.
3. Дайте определение выборки.
4. Каковы требования к репрезентативной выборке?
5. Дайте определение вариационного ряда.

6. Что означают термины частота, накопленная и относительная частота?

7. Что такое гистограмма и кумулятивная линия?

8. Перечислите выборочные характеристики.

9. Как определить объем выборки, необходимый для получения достоверных статистических выводов?

10. На основании каких характеристик выборки делается вывод о ее соответствии нормальному закону распределения?

11. Какой вид распределения имеет выборка, представленная в лабораторной работе?

Литература.

1. Пінчук Н.С., Галузинський Г.П. Орленко Н.С. Інформаційні системи і технології в маркетингу: Навч. посібник. - КНЕУ.1999. -328с.;

2. Гужва В.М. Інформаційні системи і технології на підприємствах: Навч. посібник.-К.: КНЕУ, 2001.-400с.

3. Юнкеров В.И., Григорьев С.Г. Математико-статистическаяобработкаданныхмедицинскихисследований.-СПб.: ВМедА, 2002.- 206с.

4. Герасимов А.Н. Медицинская статистика. – М.: МИА, 2007.-480 с

5. Клюшин Д.А., Петунин Ю.И. Доказательная медицина. Применение статистических методов. - К.: Диалектика, 2007.-320 с

Приложение

Критические значения коэффициента Стьюдента (t-критерия) для различной доверительной вероятности p*и числа степеней свободы* f:

f	***p***							
	0.80	**0.90**	**0.95**	**0.98**	**0.99**	**0.995**	**0.998**	**0.999**
1	3.0770	6.3130	12.7060	31.820	63.656	127.656	318.306	636.619

2	1.8850	2.9200	4.3020	6.964	9.924	14.089	22.327	31.599
3	1.6377	2.35340	3.182	4.540	5.840	7.458	10.214	12.924
4	1.5332	2.13180	2.776	3.746	4.604	5.597	7.173	8.610
5	1.4759	2.01500	2.570	3.649	4.0321	4.773	5.893	6.863
6	1.4390	1.943	2.4460	3.1420	3.7070	4.316	5.2070	5.958
7	1.4149	1.8946	2.3646	2.998	3.4995	4.2293	4.785	5.4079
8	1.3968	1.8596	2.3060	2.8965	3.3554	3.832	4.5008	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.780
10	1.3720	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.363	1.795	2.201	2.718	3.105	3.496	4.024	4.437
12	1.3562	1.7823	2.1788	2.6810	3.0845	3.4284	3.929	4.178
13	1.3502	1.7709	2.1604	2.6503	3.1123	3.3725	3.852	4.220
14	1.3450	1.7613	2.1448	2.6245	2.976	3.3257	3.787	4.140
15	1.3406	1.7530	2.1314	2.6025	2.9467	3.2860	3.732	4.072
16	1.3360	1.7450	2.1190	2.5830	2.9200	3.2520	3.6860	4.0150
17	1.3334	1.7396	2.1098	2.5668	2.8982	3.2224	3.6458	3.965
18	1.3304	1.7341	2.1009	2.5514	2.8784	3.1966	3.6105	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.3253	1.7247	2.08600	2.5280	2.8453	3.1534	3.5518	3.8495
21	1.3230	1.7200	2.2.0790	2.5170	2.8310	3.1350	3.5270	3.8190
22	1.3212	1.7117	2.0739	2.5083	2.8188	3.1188	3.5050	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.1040	3.4850	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.0905	3.4668	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.4502	3.7251
26	1.315	1.705	2.059	2.478	2.778	3.0660	3.4360	3.7060
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.4210	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.0469	3.4082	3.6739
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.0360	3.3962	3.8494
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.0298	3.3852	3.6460
32	1.3080	1.6930	2.0360	2.4480	2.7380	3.0140	3.3650	3.6210
34	1.3070	1.6909	2.0322	2.4411	2.7284	3.9520	3.3479	3.6007
36	1.3050	1.6883	2.0281	2.4345	2.7195	9.490	3.3326	3.5821

38	1.3042	1.6860	2.0244	2.4286	2.7116	3.9808	3.3190	3.5657
40	1.303	1.6839	2.0211	2.4233	2.7045	3.9712	3.3069	3.5510
42	1.320	1.682	2.018	2.418	2.6980	2.6930	3.2960	3.5370
44	1.301	1.6802	2.0154	2.4141	2.6923	3.9555	3.2861	3.5258
46	1.300	1.6767	2.0129	2.4102	2.6870	3.9488	3.2771	3.5150
48	1.299	1.6772	2.0106	2.4056	2.6822	3.9426	3.2689	3.5051
50	1.298	1.6759	2.0086	2.4033	2.6778	3.9370	3.2614	3.4060
55	1.2997	1.673	2.0040	2.3960	2.6680	2.9240	3.2560	3.4760
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.9146	3.2317	3.4602
65	1.2947	1.6686	1.997	2.3851	2.6536	3.9060	3.2204	3.4466
70	1.2938	1.6689	1.9944	2.3808	2.6479	3.8987	3.2108	3.4350
80	1.2820	1.6640	1.9900	2.3730	2.6380	2.8870	3.1950	3.4160
90	1.2910	1.6620	1.9867	2.3885	2.6316	2.8779	3.1833	3.4019
100	1.2901	1.6602	1.9840	2.3642	2.6259	2.8707	3.1737	3.3905
120	1.2888	1.6577	1.9719	2.3578	2.6174	2.8598	3.1595	3.3735
150	1.2872	1.6551	1.9759	2.3515	2.6090	2.8482	3.1455	3.3566
200	1.2858	1.6525	1.9719	2.3451	2.6006	2.8385	3.1315	3.3398
250	1.2849	1.6510	1.9695	2.3414	2.5966	2.8222	3.1232	3.3299
300	1.2844	1.6499	1.9679	2.3388	2.5923	2.8279	3.1176	3.3233
400	1.2837	1.6487	1.9659	2.3357	2.5882	2.8227	3.1107	3.3150
500	1.2830	1.6470	1.9640	2.3330	2.7850	2.8190	3.1060	3.3100

Тема 2: Планирование эксперимента. Модули Дисперсионного, канонического и кластерного анализа в программе СТАТИСТИКА

Цель работы: Ознакомиться со способами планирования эксперимента. Провести различные виды анализа исходных данных, сравнить полученные результаты.

План занятия.

1. Повторить теоретический материал.

2. Запустить демо-версию программы Statistica
3. Загрузить файл “Stat3-xx” с индивидуальным заданием.
4. Провести необходимые измерения и вычисления, занести результаты в контрольный файл “Result3-xx”.
5. Подготовить отчет о выполненной работе с интерпретацией результатов обработки экспериментальных данных и отправить его по электронной почте на электронный адрес преподавателя strahova@zsmu.zp.ua.
6. Пройти тестирование по итогам занятия.

Студент должен уметь:

- подготовитьплан экспериментас учетом последующей статистической обработки его результатов;
- ориентироваться в способах статистической обработки данных по виду шкал;
- проводить дисперсионный, канонический и кластерный анализ данных;
- формулировать заключение о качестве проведенного статистического исследования данных.

Студент должен знать:

- условия организации факторного эксперимента;
- виды планов эксперимента – полнофакторный, дробнофакторный;
- виды измерительных шкал и их соответствие применяемым статистическим методам;
- виды кластерного анализа**.**

Основные понятия:

- план эксперимента насыщенный, ненасыщенный, сверхнасыщенный;
- статистический анализ кластерный, канонический, дисперсионный;
- непараметрическиекритерии,их соответствие параметрическим;
- нормальноестатистическоераспределение.

Необходимое материальное и программное обеспечение:

- персональный компьютер,
- демо-версияППП Statistica, Microsoft Excel,
- локальная сеть,
- доступ к электронной библиотеке.

Тема: Планирование эксперимента. Модули Дисперсионного, канонического и кластерного анализа в программе СТАТИСТИКА

Цель работы: Ознакомиться со способами планирования эксперимента. Провести различные виды анализа исходных данных, сравнить полученные результаты.

План занятия.

7. Повторить теоретический материал.
8. Запустить демо-версию программы Statistica
9. Загрузить файл “Stat3-xx” с индивидуальным заданием.
10. Провести необходимые измерения и вычисления, занести результаты в контрольный файл “Result3-xx”.

11. Подготовить отчет о выполненной работе с интерпретацией результатов обработки экспериментальных данных и отправить его по электронной почте на электронный адрес преподавателя strahova@zsmu.zp.ua.
12. Пройти тестирование по итогам занятия.

Необходимое материальное и программное обеспечение:

- персональный компьютер,
- демо-версия ППП Statistica, Microsoft Excel,
- локальная сеть,
- доступ к электронной библиотеке.

Планирование эксперимента

Планирование эксперимента (активный эксперимент) в фармации - раздел математической статистики, изучающий методы организации совокупности опытов с различными условиями для получения наиболее достоверной информации о свойствах исследуемого объекта при наличии неконтролируемых случайных возмущений. Величины, определяющие условия данного опыта, обычно называются факторами (например, температура, концентрация), их совокупность - факторным пространством. Набор значений факторов характеризует некоторую точку факторного пространства, а совокупность всех опытов составляет так называемый факторный эксперимент. Расположение точек в факторном пространстве определяет план эксперимента, который задает число и условия проведения опытов с регистрацией их результатов.

Начало планированию экспериментов положили труды Р. Фишера (1935). Он показал, что рациональное планирование экспериментов дает не

менее существенный выигрыш в точности оценок, чем оптимизированная обработка результатов измерений.

Планирование экспериментов используют для изучения и математического описания процессов и явлений путем построения мат. моделей в форме уравнений регрессии - соотношений, связывающих с помощью ряда параметров значения факторов и результаты эксперимента, называемых откликами. Основное требование, предъявляемое к планам факторного эксперимента, в отличие от пассивного эксперимента, - минимизация числа опытов, при которой получают достоверные оценки вычисляемых параметров при соблюдении приемлемой точности мат. моделей в заданной области факторного пространства. В этом случае задача обработки результатов факторного эксперимента заключается в определении численных значений указанных параметров.

Одним из способов повышения точности обработки результатов планирования эксперимента служит замена переменных, при которой от исходных (физических, или натуральных) значений переменных, выраженных в соответствующих единицах измерений, переходят к безразмерным значениям, определяемым формулой:

$$x_j = (z_j - z_j^0)/\Delta z_j,\ j = 1,2,\ \ m$$

где m-число факторов;

x_j- безразмерное значение переменной;

z_j - значение физической переменной;

$z_j^0 = (z_j^{max} + z_j^{min})/2$ - среднее значение физической переменной,

$\Delta z_j = (z_j^{max} - z_j^{min})/2$ - интервал ее варьирования;

z_j^{max} и z_j^{min}-максимальные и минимальные значения физической переменной, которые могут быть заданы в опытах. При таком преобразовании значения всех x_j или уровни факторов, изменяются в

одинаковых пределах: от -1 до +1. Точка факторного пространства, отвечающая нулевым значениям факторов, называется центром плана.

Область применения планирования экспериментов распространяется на процессы и явления, зависящие от так называемых управляемых факторов, т. е. факторов, которые можно изменять и поддерживать на заданных уровнях.

Основные направления использования планирования экспериментов в фарм. технологии:

1) выделение значимых факторов, существенно влияющих на изучаемый процесс;

2) получение математических моделей объектов исследования (аппроксимационные задачи);

3) поиск оптимальных условий протекания процессов, т. е. совокупности значений факторов, при которой заданный критерий оценки эффективности процесса имеет наилучшее значение (экстремальные задачи);

4) построение диаграмм состав-свойство;

5) изучение кинетики и механизма процессов.

Выделение значимых факторов осуществляется в ходе отсеивающего эксперимента. Число опытов в нем может быть больше, равно или меньше числа проверяемых факторов. Планы, отвечающие таким экспериментам, называются соответственно ненасыщенными, насыщенными или сверхнасыщенными.

Ненасыщенные планы используют, если предварительно исследованию подлежат сравнительно небольшое число факторов ($n < 6 - 7$) и их возможные взаимодействия. Эффект взаимодействия двух или нескольких факторов проявляется при одновременном их варьировании, когда влияние каждого фактора на отклик зависит от уровней, на которых находятся другие факторы. Ненасыщенные планы обычно включают

значительное число опытов и поэтому достаточно трудоемки. В качестве таких планов часто применяют планы так называемого полного факторного эксперимента (ПФЭ), в котором каждый фактор изменяется одинаковое число раз q (где $q \geqslant 2$ - число выбранных уровней); при этом реализуются все возможные опыты, различающиеся значением хотя бы одного фактора. Число опытов в ПФЭ

$n = q^m$:

например, для $m = 2$ и $q = 2$ число $n = 2^2 = 4$ опыта.

Условия проведения опытов могут быть представлены в графической (см. рис. на стр. 50) или табличной (см. табл.4) форме. В последнем случае первый столбец (i-номер опыта) и совокупность значений факторов (второй и третий столбцы) образуют матрицу плана ПФЭ, к которой предъявляют такие требования:

1) сумма элементов столбца каждого фактора равна нулю:

$$\sum_{u=1}^{n} x_{ju} = 0$$

(u-текущий номер опыта);

2) сумма квадратов элементов столбца каждого фактора равна числу опытов:

$$\sum_{u=1}^{n} x_{ju}^2 = n$$

3) сумма почленных произведений любых столбцов двух любых факторов равна нулю:

$$\sum_{u=1}^{n} x_{i\,u} x_{j\,u} = 0 (i \neq j; i, j = 0,1,2, \quad m)$$

Представление условий проведения опытов в табличной форме.

Таблица 4. Представление условий проведения опытов в табличной форме.

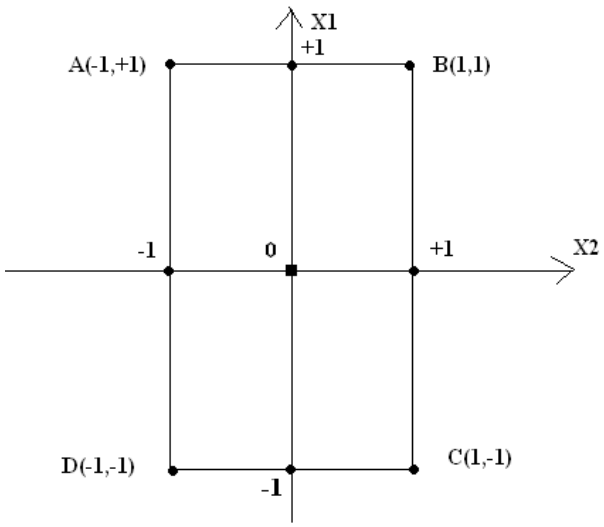
i	Кодированные переменные		Отклик у
	x_1	x_2	
1	-1	+ 1	y_1
2	- 1	- 1	y_2
3	+ 1	+ 1	y_3
4	+ 1	-1	y_4

Значения физических переменных, соответствующие матрице, выбранной для реализации опытов, рассчитывают по формуле:

$$z_j = z_j^0 + \Delta z_j x_j,\ j = 1,2,\ m.$$

При числе опытов в ПФЭ, значительно превышающем число определяемых параметров модели, применяют так называемые дробные реплики (или дробный факторный эксперимент -ДФЭ), которые представляют собой часть плана ПФЭ. ДФЭ может содержать половину, четверть и т.д. опытов от ПФЭ. Соответственно различают полуреплики (q^{m-1}), четверть реплики (q^{m-2}) и т. п. В общем случае ДФЭ может быть обозначен как q^{m-l}, где l-дробность реплики. К матрице ДФЭ предъявляют те же требования, что и к матрице ПФЭ. Планы, полученные с использованием ПФЭ или его дробных реплик, в которых переменные варьируются на двух уровнях, называются линейными либо планами 1-го порядка, т.к. при их применении можно построить уравнение модели, включающее исследуемые факторы лишь в 1-й степени.

Расположение точек в факторном пространстве в случае ПФЭ 2^2. Цифры около точек A,B,C,D характеризуют в кодированных переменных условия проведения опытов



Насыщенные планы используют, если математическая модель предполагается в виде полинома (уравнения регрессии) 1-го порядка, общий вид которого может быть представлен выражением:

$$y = b_0 + \sum_{j=1}^{m} b_j x_j$$

где y-отклик, b_0 и b_j-параметры модели. В качестве насыщенных планов наиболее часто применяют планы ДФЭ.

Алгоритм выделения значимых факторов в этом случае включает следующие этапы:

1) по формуле определяют параметры математической модели.

$$b_i = \sum_{j=0}^{m} x_j y_i / n, i = 1,2, \quad n$$

2) По результатам параллельных опытов вычисляют дисперсию воспроизводимости, характеризующую разброс значений отклика. Например, при проведении r параллельных опытов в одной точке факторного пространства:

$$S_b^2 = \sum_{i=1}^{r} (y_i - \bar{y})^2 / 9r - 1)$$

где

$$\bar{y} = \sum_{i=1}^{r} y_i / r$$

3)По формуле определяют дисперсию каждого параметра.

$$S_{b\ j}^2 = S_b^2 / n, j = 0,1, \ .m$$

4) Для оценки точности найденных значений параметров, а также полученной мат. модели используют статистические критерии соответствия Стьюдента (t-критерий) и Фишера (F-критерий). При этом количественными мерами служат доверительная вероятность β или уровень значимости p = 1 — β и число степеней свободы f, т. е. число экспериментов за вычетом числа констант, рассчитываемых по результатам этих опытов. Число констант определяется видом выбранной дисперсии; например, в случае дисперсии воспроизводимости по результатам параллельных опытов находят величину $\bar{y}$, поэтому f_b = r — 1. При заданных требованиях на точность результатов измерений доверительная вероятность (уровень значимости) определяет надежность полученной оценки. Значения указанных критериев табулированы и приводятся в специальной литературе.

5) Значимость каждого фактора проверяют оценкой значимости соответствующего параметра, т.к. вклады факторов в значение отклика пропорциональны значениям параметров. Для оценки их значимости рассчитывают соответствующее значение t-критерия по формуле:

$$t_{b\ i} = |b_i| / S_b \ , i = 0,1, \ .m$$

Полученное значение сравнивают с табличным t^T, найденным на предыдущем этапе. При выбранной доверительной вероятности параметр считается значимым, если $t_{bi}. > t^T$. В противном случае параметр незначим и соответствующий фактор можно исключить из построенной математической модели.

Сверхнасыщенные планы используют, если на процесс может влиять большое число факторов и их взаимодействий. Наиболее часто с целью уменьшения их числа применяют метод случайного баланса,

позволяющий вместо ПФЭ и ДФЭ применять эксперименты, в которых значения факторов распределены по уровням случайным образом (рандомизированы). Метод имеет высокую разрешающую способность (возможность выделять сильно влияющие факторы), но малую чувствительность (т. е. способность выделять значимые параметры модели, характеризующие факторы, которые имеют относительно слабое влияние). Используют также метод последовательного отсеивания: все изучаемые факторы на основе априорной информации подразделяют на группы, каждую из которых в дальнейшем рассматривают как отдельный комплексный фактор. В зависимости от полученной при этом информации остальные факторы снова разбивают на группы и выполняют новый цикл расчетов.

Аппроксимационные задачи. Для учета нелинейностей объекта исследований его математическое описание часто получают в виде полинома i-го порядка, который в общем виде выражается формулой:

$y = b_i x^i + b_{i-1} x^{i-1} + \ldots + b_0$

Например, полином 2-го порядка для двух факторов записывается следующим образом:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2.$$

Для нахождения параметров таких моделей недостаточно варьирования значений факторов на двух уровнях, поскольку нелинейность не может быть определена двумя точками. Поэтому для указанных моделей обычно применяют так называемые композиционные планы, включающие изменения факторов более чем на двух уровнях, что позволяет использовать их для построения моделей порядка выше первого. Общий алгоритм решения аппроксимационной задачи включает этапы.

1) Выбирают число существенных факторов, их средние значения и интервалы варьирования - эта информация может быть получена после

проведения отсеивающего эксперимента или на основании знаний и интуиции исследователя.

2) Строят матрицу плана; на начальном этапе исследования в зависимости от числа факторов выбирают, как правило, планы 1-го порядка (ПФЭ или ДФЭ).

3) Рандомизируют опыты - для уменьшения влияния систематических ошибок опыты проводят в условиях, соответствующих строкам матрицы плана, выбираемым в случайном порядке (целесообразность такого приема подтверждена на практике).

4) Обрабатывают полученные результаты - рассчитывают параметры и составляют уравнение регрессии, оценивают значимость параметров и проверяют адекватность полученной мат. модели имеющимся экспериментальным данным. Для проверки адекватности модели анализируют разность между опытными значениями и значениями отклика, предсказанными по полученной мат. модели в разных точках факторного пространства. В качестве последних могут быть взяты как точки плана (при ненасыщенных планах), так и дополнительные точки. Их обычно выбирают в области, представляющей наибольший интерес, либо располагают таким образом, чтобы полученные результаты можно было использовать для построения более точной модели высокого порядка.

5) Принимают решение о дальнейших действиях: если на этапе 4 получено адекватное уравнение регрессии, вывод аппроксимационной зависимости на этом заканчивают; в противном случае выясняют причину неадекватности и проводят новую серию экспериментов с использованием планов 1-го порядка (уменьшают интервалы варьирования факторов, включают в мат. модель новый фактор и т.д.) или более высоких порядков (выбор определяется целями исследователя).

В результате проверки адекватности модель может оказаться неадекватной вследствие того, что:

а) в нее включены не все факторы, существенно влияющие на процесс. В этом случае выбирают более полную модель и для определения ее параметров строят, реализуют и обрабатывают новую матрицу планирования;

б) не учтены эффекты взаимодействия разных факторов. Для их учета предполагаемые взаимодействия включают в модель и, если позволяет исходный план (число опытов не менее числа определяемых параметров новой модели), повторно обрабатывают результаты эксперимента. Если начальный план не дает возможности провести такую обработку ($n < т$), выполняют дополнительные опыты с расширенным планом (например, от полуреплики переходят к ПФЭ и т.п.), причем реализуются только те опыты, которые не входили в исходный план;

в) принятый порядок модели ниже требуемого. Для проверки необходимо расширить используемый композиционный план, включив опыты, обеспечивающие получение модели более высокого порядка. Если модель высшего порядка будет адекватной, то это предположение подтверждается.

При проведении эксперимента исследователь может предъявлять к мат. модели различные требования: получение определенных оценок ее параметров; обеспечение желаемых предсказательных свойств и т. п. Это приводит к необходимости выбора специальных планов, подчиненных поставленным требованиям (критериям). Среди критериев, удовлетворяющих первому требованию, наиболее общим является D-критерий, соответствующий обобщенной дисперсии всех оценок параметров математической модели. Кроме него применяют А-критерий, отвечающий средней дисперсии оценок параметров; Е-критерий, соответствующий длине максимальной оси эллипсоида рассеяния оценок параметров; критерий ортогональности, обеспечивающий независимость определения параметров модели, и т.д. Среди критериев, удовлетворяющих

второму требованию, особенно часто используют G-критерий, отвечающий макс. дисперсии предсказанных значений функции отклика; Q-критерий, соответствующий среднему значению дисперсий предсказанных значений; критерий ротатабельности, отвечающий дисперсии оценки предсказанных значений отклика во всех точках, равноудаленных от центра плана, и др.

Планы, минимизирующие приведенные выше критерии, называют соотв. D-оптимальными, A-оптимальными и т.д. Как правило, не удается построить план, одновременно удовлетворяющий нескольким критериям. Исключение составляют линейные планы: например, планы ПФЭ и ДФЭ не только ортогональны и ротатабельны, но еще и D-, G-, A- и E-оптимальны. Поэтому, если цель исследования - построение некоторой описательной математической модели, аппроксимирующей опытные данные, рекомендуют использовать планы, отвечающие D-критерию; если модель должна обладать наилучшими предсказательными свойствами, используют планы, соответствующие G- или Q-критерию. Если, наконец, цель эксперимента - поиск оптимальных условий функционирования объекта, часто применяют ротатабельные планы.

Экстремальные задачи имеют целью определить наилучшее значение целевой функции, в качестве которой принимают значение интересующей исследователя характеристики процесса. Такие задачи могут быть решены по крайней мере двумя способами: с построением и без построения математической модели.

Планирование экспериментов с построением математической модели процесса. На основе выбранного плана строят модель, отвечающую рассматриваемому отклику, и, используя ее, с помощью известных методов поиска экстремума находят значения факторов, при которых целевая функция, определенная по модели, будет экстремальной. Если найденные значения факторов, соответствующие экстремальной точке, лежат на границе примененного плана, область планирования либо смещается, либо

расширяется и строится новая модель, после чего поиск экстремума повторяется. Задача считается решенной, если вычисленные координаты точки экстремума находятся внутри области, характеризуемой использованным планом.

На практике такой подход часто реализуют методом «крутого восхождения» (метод Бокса-Уилсона). Выбирают начальную точку, в окрестности которой проводят ПФЭ или ДФЭ (в зависимости от числа факторов); по его результатам рассчитывают параметры математической модели 1-го порядка. Если модель адекватна, с ее помощью определяют направление изменения факторов, соответствующее движению к экстремальному значению целевой функции в направлении градиента или антиградиента (соответственно при поиске максимума или минимума). Движение в выбранном направлении осуществляют с помощью последовательно выполняемых опытов и производят до тех пор, пока отклик изменяется желаемым образом. В найденной наилучшей (для выбранного направления) точке снова выполняют ПФЭ или ДФЭ и т.д. Изложенную процедуру повторяют до построения адекватной модели на каждом этапе. Неадекватность модели, полученной на очередном этапе, свидетельствует о том, что, возможно, достигнута область экстремума, в которой линейную модель уже нельзя использовать. Для уточнения положения экстремума в этой области можно применять модель 2-го порядка, построенную посредством соответствующих планов.

Непосредственно эксперимент на объекте (без построения модели). Стратегия проведения опытов определяется выбранным методом оптимизации. При этом значение целевой функции вычисляют не по модели, а находят непосредственно из опыта, выполненного в соответствующих условиях. Наиболее часто для поиска наилучшего значения целевой функции используют последовательный симплексный метод, метод Гаусса-Зейделя и т.п.

Построение диаграмм состав-свойство. Построение таких диаграмм - важная часть физико-химических исследований различных смесей. Для смесей, содержащих k компонентов, характерно наличие следующего ограничения:

$$\sum_{i=1}^{k} x_i = c\ o\ n$$

Сумма концентраций компонентов смеси обычно нормируется, поэтому соотношение имеет вид:

$$\sum_{i=1}^{k} x_i = 1$$

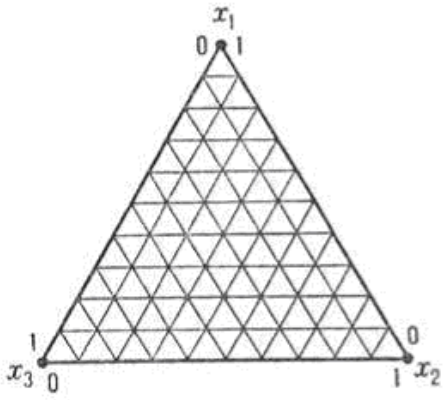
где x_i-относительная концентрация i-го компонента смеси. При обработке результатов активного эксперимента это выражение определяет в n-мерном пространстве переменных x_i область их допустимых изменений, называемую симплексом. Например, в случае трех переменных симплекс представляет собой равносторонний треугольник (рис. 2). Вершинам симплекса соответствуют чистые компоненты. Точки на границах симплекса (ребрах) отвечают бинарным смесям соответствующих пар компонентов. Любая точка внутри симплекса отвечает составу смеси, в которой присутствуют все три компонента (указанные точки отмечены на рис. штриховкой). Для четырехкомпонентной смеси симплексом служит тетраэдр, грани которого- симплексы, соответствующие трехкомпонентным смесям, и т.д.

Согласно условию $\sum_{i=1}^{k} x_i = 1$, упомянутые выше факторные эксперименты непригодны для построения диаграмм состав-свойство из-за невозможности независимого варьирования каждого фактора. На практике для построения таких моделей иногда применяют симплекс-решетчатые планы (планы Шеффе), представляющие собой набор точек, равномерно распределенных на границе и внутри симплекса. Эти планы обычно насыщены и могут быть композиционными; например, точки плана 1-го порядка входят во все послед. композиции. Предложены также насыщенные

симплекс - центроидные планы, которые состоят из точек, расположенных в вершинах симплекса, серединах ребер, центрах граней различной размерности и в центре симплекса.

Адекватность моделей, построенных на основе симплекс - решетчатых и симплекс -центроидных планов, вследствие их насыщенности проверяют по результатам дополнительных опытов в контрольных точках. Их координаты целесообразно выбирать так, чтобы они могли быть использованы, если возникнет необходимость получения уточненной модели более высокого порядка.

Изучение объектов, характеризуемых наличием неоднородностей. В общем случае источники неоднородностей могут быть непрерывного или дискретного типа. Источники непрерывного типа характеризуются изменением свойств объекта (его дрейфом) во времени или по какой-либо другой переменной (например, неравномерное старение катализатора по длине аппарата). В случае невысоких (по сравнению с продолжительностью проведения всех опытов эксперимента) скоростей дрейфа можно использовать обычные методы планирования экспериментов. При высоких скоростях дрейфа применяют специальные планы, построенные, например, на основе «ортогональных полиномов Чебышева» и т. п.



Общий вид простейшего симплекса.

A	B			
	B_1	B_2	B_3	B_4
A_1	C_1	C_2	C_3	C_4
A_2	C_2	C_3	C_4	C_1
A_3	C_3	C_4	C_1	C_2
A_4	C_4	C_1	C_2	C_3

Пример латинского квадрата 4 × 4.

Рис. 7. Симплекс, латинский квадрат

Источники дискретного типа: различие в сырье, технологических аппаратах, способах проведения процессов, исполнителях и т. д. В данном случае задача планирования экспериментов заключается в сокращении числа оцениваемых возможных сочетаний изучаемых факторов, т.е. относится к классу комбинаторных задач. Их решают с помощью планов, основанных на специальных правилах размещения факторов по уровням в каждом опыте. Существует множество способов организации таких планов, из которых наиболее распространены планы, использующие свойства «латинских» и «греко-латинских квадратов», кубов и др. Например, латинский квадрат представляет собой таблицу, состоящую из n строк и n столбцов и заполненную n элементами (числами или буквами) так, что каждый элемент повторяется в каждой строке и каждом столбце только один раз.

Изучение кинетики и механизмов процессов связано, как правило, с разработкой детерминированных моделей, отражающих физико-химическую сущность исследуемых явлений и содержащих описания механизмов (кинетики) протекающих в них элементарных процессов. Среди задач, решаемых методами планирования экспериментов, можно выделить:

1) определение (уточнение) параметров моделей;

2) дискриминацию, т.е. отбрасывание проверяемых механизмов элементарных процессов.

Для уточнения параметров детерминированных моделей необходимо выбрать такой план эксперимента, который обеспечит наилучшие оценки определяемых величин. Наиболее часто для этих целей используют, как указано выше, D-оптимальные планы. При уточнении параметров планирования экспериментов сталкиваются с рядом трудностей. К основным из них можно отнести:

1) необходимость иметь отдельный план для каждого класса моделей, т. е. в каждой конкретной ситуации исследователь должен

вычислить оптимальное расположение точек в факторном пространстве для постановки уточняющих экспериментов;

2) необходимость расчета параметров детерминированных моделей с использованием методов оптимизации; это обусловлено обычно нелинейностью данных моделей относительно определяемых параметров.

Задача дискриминации заключается в выборе такой модели среди нескольких конкурирующих, которая наиболее правильно отражает механизм процесса и обладает наилучшей предсказательной способностью. Эта задача реализуется сопоставлением результатов оценки соответствия модели опытным данным при использовании различных описаний одного и того же процесса или явления. Самый простой метод дискриминации состоит в вычислении параметров каждой предложенной модели по экспериментальным данным и последующем сравнении остаточных дисперсий. В качестве выбранной модели принимают модель с минимальной остаточной дисперсией. Если не удается выбрать механизм, не противоречащий опытным данным, то либо расширяют исследуемую область, либо смещают расположение точек в факторном пространстве и операцию повторяют. Достоинство такого подхода заключается в том, что исследователь одновременно решает обе задачи - вычисление параметров и дискриминацию моделей. К недостаткам можно отнести то, что при этом часто требуются большие затраты времени на эксперименты и расчет параметров моделей.

Данные, полученные в ходе нашего эксперимента, необходимо обработать. Как мы видели, они подчиняются нормальному закону распределения, и, следовательно, к ним можно применить определенные виды статистического анализа.

Основными элементами математических моделей являются признаки, которыми описываются объекты наблюдения. Их подразделяют на факторы-причины, воздействующие на объекты, и

показатели – отклики, характеризующие состояние изучаемой системы. Вероятность прогнозируемого значения показателя-отклика – степень возможности проявления какого – либо определённого события в тех или иных условиях.

В тех случаях, когда факторы-причины и показатели-отклики измерялись в количественных шкалах и между ними установлена сильная и значимая корреляционная связь, моделирование выполняется методами **РЕГРЕССИОННО - КОРРЕЛЯЦИОННОГО АНАЛИЗА.** Регрессионный анализ (линейный) — статистический метод исследования зависимости между зависимой переменной Y и одной или несколькими независимыми переменными X1,X2,...,Xp. Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные — критериальными. Целью регрессионного анализа является определение наличия и характера (математического уравнения, описывающего зависимость) связи между переменными, определение степени детерминированности вариации критериальной переменной предикторами, предсказание значения зависимой переменной с помощью независимой и определение вклада независимых переменных в вариацию зависимой (модули **Множественная регрессия / Multiply Regression, Непараметрические данные/ Nonparametrics в программе СТАТИСТИКА).**

Когда в исследовании имеются факторы только неколичественного характера (порядковые или номинальные) и они задаются в эксперименте на некоторых качественных уровнях, то для моделирования значений показателей-откликов на воздействия таких факторов и решения задач исследования применяется **ДИСПЕРСИОННЫЙ АНАЛИЗ**. Он выявляет структуру связи между показателем-откликом и факторами-причинами, позволяет оценить степень влияния каждого из изучаемых качественных факторов, а также

их взаимодействий на дисперсию показателя-отклика (модуль **Анализ вариантов /ANOVA**).

Иногда результаты эксперимента включают как количественные, так и качественные факторы, воздействующие на объекты наблюдения. В этих условиях для моделирования показателя-отклика не только в зависимости от основных качественных факторов, но и с учетом влияния сопутствующих количественных эффективен **КОВАРИАЦИОННЫЙ АНАЛИЗ** (модуль **Дополнительные линейные/нелинейные модели/Структурное моделирование уравнения / Advanced Linear/Nonlinear Models/Structural Equation Modeling**).

Для решения задач классификации (распознавания образов) и отнесения объекта с определенным набором признаков к одному из известных классов используется **ДИСКРИМИНАНТНЫЙ АНАЛИЗ**. В медицине такой вид анализа применяется для решения диагностических, прогностических, экспертных задач, выбора методов и схем лечения. Для классификации определяется линейная комбинация (линейная дискриминантная функция), которая максимизирует различия между классами, но минимизирует дисперсию внутри классов. В итоге определяются линейные классификационные функции для каждого класса (модуль **Многомерные исследовательские методы/Дискриминантный анализ / Multivariate Exploratory techniques/Discriminant Analysis).**

Результаты исследования, в котором все переменные являются только качественными, традиционно сводятся в таблицы сопряженности. Моделировать по таким таблицам лучше всего посредством процедур **ЛОГЛИНЕЙНОГО АНАЛИЗА**. Такой анализ обеспечивает установление силы и значимости связей между признаками с учетом их взаимодействия, определение степени влияния исходных факторов на выходные результирующие признаки-отклики, прогнозирование

ожидаемых частот наблюдений при определенных сочетаниях уровней факторов (модуль **Основная статистика/Таблицы/Блок таблиц / Basic Statistics/Tables**)

Анализ результатов эксперимента, содержащих качественные факторы и количественный признак-отклик, оценивающий продолжительность жизни (продолжительность ремиссии хронического заболевания, многолетней выживаемости онкологических больных после оперативного лечения и т.д.) и построение модели функции продолжительности жизни проводится методом **АНАЛИЗА ВРЕМЕНИ ВЫЖИВАНИЯ**. (модули **Дополнительные линейные/нелинейные модели/Анализ выживания /Advanced Linear/Nonlinear Models/Survival Analysis** и **Дополнительные линейные/нелинейные модели/Прогноз-Серия времени /Advanced Linear/Nonlinear Models/Time Series Forecasting**).

В последнее время получил распространение метод моделирования с помощью логистической регрессии. Показаниями к применению этого метода являются:

признак-отклик является дихотомическим (измеряется на двух уровнях и является альтернативным);

факторы-причины преимущественно качественные.

Логистическая регрессионная модель позволяет получить вероятность наступления благоприятного или неблагоприятного исхода изучаемого явления в зависимости от степени выраженности конкретного набора признаков-причин и степени влияния одного или группы показателей-причин, в процентах, на вероятность наступления прогнозируемого события.

Непараметрические статистики и подгонка распределения.

Термин «Непараметрические статистики» был впервые введен Wolfowitz, 1942. Для многих изучаемых переменных невозможно сказать с уверенностью, что рассматриваемая переменная имеет нормальное распределение. Случаи редких болезней не являются нормально распределенными в популяции, число автомобильных аварий также не является нормально распределенным, как и многие переменные, интересующие исследователя. Другим фактором, часто ограничивающим применимость критериев, основанных на предположении нормальности, является объем или размер выборки, доступной для анализа. До тех пор пока выборка достаточно большая (например, 100 или больше наблюдений), можно считать, что выборочное распределение нормально, даже если вы не уверены, что распределение переменной в популяции, действительно, является нормальным. Тем не менее, если выборка очень мала, то критерии, основанные на нормальности, следует использовать только при наличии уверенности, что переменная действительно имеет нормальное распределение. Однако нет способа проверить это предположение на малой выборке. Использование критериев, основанных на предположении нормальности, кроме того, ограничено точностью измерений. Например, рассмотрим исследование, в котором средний балл успеваемости является основной переменной. Можно ли сказать, что средняя успеваемость студента A в два раза выше, чем успеваемость студента C? Является ли различие между средним баллом студентов B и A сравнимым с различием между студентами D и C? Индекс среднего балла успеваемости является грубой мерой, позволяющей только ранжировать студентов в порядке "хороший" - "плохой". После этого введения становится ясной необходимость наличия статистических процедур, позволяющих обрабатывать данные "низкого качества" из выборок малого объема с

переменными, про распределение которых мало что или вообще ничего не известно. **НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ** как раз и разработаны для тех ситуаций, достаточно часто возникающих на практике, когда исследователь ничего не знает о параметрах исследуемой популяции (отсюда и название методов - непараметрические). Иными словами, непараметрические методы не основываются на оценке параметров, таких как среднее или стандартное отклонение, при описании выборочного распределения интересующей величины. Поэтому эти методы иногда также называются свободными от параметров или свободно распределенными. По существу, для каждого параметрического критерия имеется, по крайней мере, один непараметрический аналог. Эти критерии можно отнести к одной из следующих групп:

- критерии различия между группами (независимые выборки);
- критерии различия между группами (зависимые выборки);
- критерии зависимости между переменными.

Обычно, когда имеются две выборки (например, мужчины и женщины), которые вы хотите сравнить относительно среднего значения некоторой изучаемой переменной, вы используете t **-КРИТЕРИЙ ДЛЯ НЕЗАВИСИМЫХ ВЫБОРОК** (в модуле **Основные статистики и таблицы / Basic statistics / Tables**). Непараметрическими альтернативами этому критерию являются**: критерий серий Вальда-Вольфовица, U критерий Манна-Уитни и двухвыборочный критерий Колмогорова-Смирнова**. Если вы имеете несколько групп, то можете использовать **ДИСПЕРСИОННЫЙ АНАЛИЗ.** Его непараметрическими аналогами являются**: ранговый дисперсионный анализ Краскела-Уоллиса и медианный тест.**

Если вы хотите сравнить две переменные, относящиеся к одной и той же выборке (например, математические успехи студентов в начале и в конце семестра), то обычно используется ***t*-КРИТЕРИЙ ДЛЯ ЗАВИСИМЫХ ВЫБОРОК** (в модуле **Основные статистики и таблицы / Basic statistics / Tables**). Альтернативными непараметрическими тестами являются: **критерий знаков и критерий Вилкоксона парных сравнений.** Если рассматриваемые переменные по природе своей категориальны или являются категоризованными (т.е. представлены в виде частот попавших в определенные категории), то подходящим будет критерий **хи-квадрат Макнемара.** Если рассматривается более двух переменных, относящихся к одной и той же выборке, то обычно используется **дисперсионный анализ (ANOVA)** с **повторными измерениями**. Альтернативным непараметрическим методом является **ранговый дисперсионный анализ Фридмана или Q критерий Кохрена** (он применяется, например, если переменная измерена в номинальной шкале). *Q* критерий Кохрена используется также для оценки изменений частот (долей).

Для того, чтобы оценить зависимость (связь) между двумя переменными, обычно вычисляют коэффициент корреляции. **Непараметрическими аналогами стандартного коэффициента корреляции Пирсона являются статистики Спирмена R, тау Кендалла и коэффициент Гамма**. Если две рассматриваемые переменные по природе своей категориальны, подходящими непараметрическими критериями для тестирования зависимости будут: **Хи-квадрат, Фи коэффициент, точный критерий Фишера. Дополнительно** доступен **критерий зависимости между несколькими переменными** так называемый **коэффициент конкордации Кендалла**. Этот тест часто используется для оценки согласованности мнений независимых экспертов (судей), в частности, баллов, выставленных одному и тому же субъекту.

Если данные не являются нормально распределенными, а измерения, в лучшем случае, содержат ранжированную информацию, то вычисление обычных описательных статистик (например, среднего, стандартного отклонения) не слишком информативно. Модуль **Непараметрическая статистика/ Nonparametrics** вычисляет разнообразный набор характеристик положения (среднее, медиану, моду и т.д.) и рассеяния (дисперсию, гармоническое среднее, квартильный размах и т.д.), позволяющий представить более "полную картину" данных.

Каждая непараметрическая процедура в модуле имеет свои достоинства и свои недостатки. Например, **двухвыборочный критерий Колмогорова-Смирнова** чувствителен не только к различию в положении двух распределений, например, к различиям средних, но также чувствителен и к форме распределения. **Критерий Вилкоксона парных сравнений** предполагает, что можно ранжировать различия между сравниваемыми наблюдениями. Если это не так, лучше использовать критерий знаков. В общем, если результат исследования является важным (например, оказывает ли людям помощь определенная очень дорогостоящая и болезненная терапия?), то всегда целесообразно применить различные непараметрические тесты. Возможно, результаты проверки (разными тестами) будут различны. В таком случае следует попытаться понять, почему разные тесты дали разные результаты. С другой стороны, непараметрические тесты имеют меньшую статистическую мощность (менее чувствительны), чем их параметрические конкуренты, и если важно обнаружить даже слабые отклонения (например, является ли данная пищевая добавка опасной для людей), следует особенно внимательно выбирать статистику критерия.

Непараметрические методы наиболее приемлемы, когда объем выборок мал. Если данных много (например, $n > 100$), то не имеет смысла

использовать непараметрические статистики. Когда выборки становятся очень большими, то выборочные средние подчиняются нормальному закону, даже если исходная переменная не является нормальной или измерена с погрешностью (теорема Гливенко, т.н. Центральная предельная теорема, см. стр. 22). Таким образом, **параметрические методы**, являющиеся более чувствительными (имеют большую статистическую мощность), **всегда подходят для больших выборок.**

Если наблюдаемые значения переменной являются **результатом очень редких событий**, то переменная будет иметь **распределение Пуассона** (которое иногда называется распределением редких событий). Например, несчастные случаи на производстве можно рассматривать как результат пересечения ряда неудачных событий (на житейском языке стечением маловероятных обстоятельств), поэтому их частота приближенно описывается распределением Пуассона.

Ход работы.

Сначала повторим вычисление значения Т-теста, как раннее это было выполнено с помощью Microsoft Excel. В Basic statistics выберем вычисление значения Т-теста сначала для независимых выборок, затем для зависимых.

T-test for Independent Samples (Книга11)
Note: Variables were treated as independent samples

Group 1 vs. Group 2	**Mean Group 1**	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
Control vs. Day 1	3,266667	30,20000	-12,9110	23	0,000000	15	10	2,491892	7,554248	9,190184	0,000337
Control vs. Day 3	3,266667	13,40000	-6,5998	23	0,000001	15	10	2,491892	5,146736	4,265849	0,015651
Control vs. Day 9	3,266667	4,70000	-1,4490	23	0,160838	15	10	2,491892	2,311805	1,161865	0,843709

T-test for Dependent Samples (Книга11) Fill Color
Marked differences are significant at p < ,05000

Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
Day 1	30,20000	7,554248						
Day 3	13,40000	5,146736	10	16,80000	4,589844	11,57474	9	0,000001
Day 1	30,20000	7,554248						
Day 9	4,70000	2,311805	10	25,50000	6,570134	12,27343	9	0,000001

Рис. 8. Вычисление Т – теста в программе Статистика.

Выводы о результате вычислений занесем в файл «Стат1» .

В пакете Статистика красным цветом выделяются данные, полученные для выборок, имеющих достоверное расхождение.

Можно вычислить группу коэффициентов, объединяемых названием Корреляционная матрица. Для этого на линейке – меню выберем Statistics --- Correlation matrices.

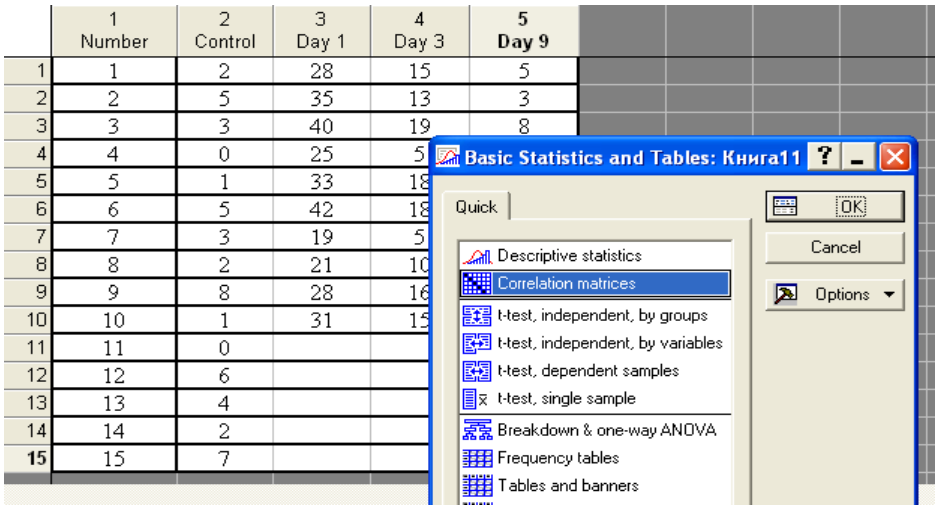


Рис. 9. Выбор «корреляционных матриц».

Укажем переменные, подлежащие обработке, и на вкладке Options выберем Display detailed table of results. Получим развернутую матрицу корреляции. Чтобы сделать выводы по результатам проведенных вычислений, надо понимать, что собой представляет корреляция.

Корреляция.

Корреляция — статистическая взаимосвязь двух или нескольких случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). Изменения одной или нескольких из этих величин приводят к систематическому изменению другой или других величин. Математической мерой корреляции двух случайных величин служит коэффициент корреляции.

Корреляционный анализ — метод обработки статистических данных, заключающийся в изучении коэффициентов (корреляции) между переменными. При этом сравниваются коэффициенты корреляции между одной парой или множеством пар признаков, для установления между ними статистических взаимосвязей.

Коэффициент корреляции или парный коэффициент корреляции — показатель характера изменения двух случайных величин.

Корреляция может быть положительной и отрицательной (возможна также ситуация отсутствия статистической взаимосвязи — например, для независимых случайных величин).

Отрицательная корреляция — корреляция, при которой увеличение одной переменной связано с уменьшением другой переменной, при этом коэффициент корреляции отрицателен.

Положительная корреляция — корреляция, при которой увеличение одной переменной связано с увеличением другой переменной, при этом коэффициент корреляции положителен.

Автокорреляция — статистическая взаимосвязь между случайными величинами из одного ряда, но взятых со сдвигом, например, для случайного процесса — со сдвигом по времени.

Корреляция отражает лишь линейную зависимость величин, но не отражает их функциональной связности.

Есть некоторые ограничения, налагаемые на принципиальное применение методов корреляционного анализа. Применение возможно в случае наличия достаточного количества случаев для изучения: для конкретного вида коэффициента корреляции составляет от 25 до 100 пар наблюдений. Второе ограничение вытекает из гипотезы корреляционного анализа, в которую заложена линейная зависимость переменных. Во многих случаях, когда достоверно известно, что зависимость существует, корреляционный анализ может не дать результатов просто ввиду того, что зависимость не линейна (выражена, например, в виде параболы). Сам по себе факт корреляционной зависимости не даёт основания утверждать, какая из переменных предшествует или является причиной изменений, или что переменные вообще причинно связаны между собой, например, ввиду действия третьего фактора.

Кроме того, корреляция может быть ложной. Часто заманчивая простота корреляционного исследования подталкивает исследователя делать ложные интуитивные выводы о наличии причинно-следственной связи между парами признаков, в то время как коэффициенты корреляции устанавливают лишь статистические взаимосвязи. Иллюстрацией этому служит хорошо известный анекдот: если выйти на улицу и измерить у 1000 случайных прохожих размер обуви и IQ, между ними будет обнаружена статистически значимая корреляция. Однако это не значит, что размер ноги влияет на интеллект, так как на наличие этой взаимосвязи, скорее всего, повлияли такие факторы, как пол и возраст участников исследования.

Marked correlations are significant at p < ,05000
(Casewise deletion of missing data)

Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	r?	t	p	N	Constant dep: Y	Slope dep: Y	Constant dep: X	Slope dep: X
Control	3,00000	2,403701									
Control	3,00000	2,403701	1,000000	1,000000			10	0,00000	1,000000	0,00000	1,000000
Control	3,00000	2,403701									
Day 1	30,20000	7,554248	0,257001	0,066050	0,752174	0,473494	10	27,77692	0,807692	0,53037	0,081776
Control	3,00000	2,403701									
Day 3	13,40000	5,146736	0,323331	0,104543	0,966430	0,362132	10	11,32308	0,692308	0,97651	0,151007
Control	3,00000	2,403701									
Day 9	4,70000	2,311805	-0,099976	0,009995	-0,284199	0,783475	10	4,98846	-0,096154	3,48857	-0,103950
Day 1	30,20000	7,554248									
Control	3,00000	2,403701	0,257001	0,066050	0,752174	0,473494	10	0,53037	0,081776	27,77692	0,807692
Day 1	30,20000	7,554248									
Day 1	30,20000	7,554248	1,000000	1,000000			10	0,00000	1,000000	0,00000	1,000000
Day 1	30,20000	7,554248									
Day 3	13,40000	5,146736	0,803618	0,645802	3,819193	0,005096	10	-3,13474	0,547508	14,39430	1,179530
Day 1	30,20000	7,554248									
Day 9	4,70000	2,311805	0,550976	0,303575	1,867414	0,098799	10	-0,39213	0,168614	21,73805	1,800416
Day 3	13,40000	5,146736									
Control	3,00000	2,403701	0,323331	0,104543	0,966430	0,362132	10	0,97651	0,151007	11,32308	0,692308
Day 3	13,40000	5,146736									
Day 1	30,20000	7,554248	0,803618	0,645802	3,819193	0,005096	10	14,39430	1,179530	-3,13474	0,547508
Day 3	13,40000	5,146736									
Day 3	13,40000	5,146736	1,000000	1,000000			10	0,00000	1,000000	0,00000	1,000000
[redacted]											
			0,496805	0,246815	1,619124	0,144081	10	1,70973	0,223154	8,20166	1,106029

Рис.10. Пример вычислений.

Коэффициент корреляции r равен ±1 тогда и только тогда, когда две выборки линейно зависимы. Коэффициент r является случайной величиной, поскольку вычисляется из случайных величин. Если модуль коэффициента (без учета знака) более 0.95, говорят о наличии практически линейная зависимость. При $0.75<|r|<0.95$ говорят о сильной степени линейной зависимости между параметрами. При $0.45<|r|<0.75$ - о существовании линейной связи между параметрами. при $|r|<0.45$ говорят, что линейную связь выявить не удалось.

В этом случае применяют так называемые непараметрические методы оценивания. Такими методами можно оценивать данные, которые не подчиняются нормальному закону распределения. В Статистике непараметрические методы реализованы в модуле Nonparametrics/Distrib.

Выводы о сути корреляционной зависимости между параметрами заносим в файл «Стат3».

Оценка значимости различия частот наблюдений в независимых выборках по χ^2-критсрию Пирсона.

Для оценки значимости различия частот наблюдения изучаемого признака в нескольких независимых группах без расчета относительных величин частоты и оценки их точности и надежности рекомендуется непараметрический критерий Пирсона «хи-квадрат».

Критерий хи-квадрат — это наиболее простой критерий проверки значимости связи между двумя категорированными переменными. Точное название - "хи-квадрат Пирсона". Критерий Пирсона основывается на том, что в двухвходовой таблице ожидаемые частоты при гипотезе "между переменными нет зависимости" можно вычислить непосредственно. Например, что 20 мужчин и 20 женщин опрошены относительно выбора газированной воды марки А или марки В. Если между предпочтением и полом нет связи, то естественно ожидать равного выбора марки А и марки В для каждого пола.

Критерий представляется уравнением:

$$\chi^2 = \Sigma(n_{1i}-n_{2i})^2/ n_{2i},$$

где n_{1i} - наблюдавшееся число случаев признака в i-ой ячейке частотной таблицы;

n_{2i} - теоретическое (рассчитанное, как среднеожидаемое) число случаев признана в i-й ячейке частотной таблицы.

При точном совпадении n_{1i} и n_{2i} во всех ячейках таблицы $\chi^2 = 0$, что свидетельствует о полном соответствии числа наблюдений в группе по данному признаку.

При увеличении разности $| n_{1i} - n_{2i} |$ величина χ^2 возрастает, увеличивается вероятность различия, и когда она становится равна или больше 95% считают, что различие групп по данному критерию значимо.

Решение получают, сравнивая рассчитанное значение χ^2 с критическими значениями χ^2_{05}, χ^2_{01}, χ^2_{001} , которые берут из соответствующей таблицы по уровням значимости p=0,05; 0,01; 0,001 и числу степеней свободы

$$n` = (m-1)*(s-1),$$

где m - число сравниваемых групп,

s- число уровней изучаемого признака.

При $\chi^2 < \chi^2_{05}$ различие групп по данным признакам незначимо (p >0,05)

При $\chi^2 \geq \chi^2_{05}$ или χ^2_{01}, или χ^2_{001} - различие значимо с уровнем значимости соответственно p≤0,05; p≤0,01; p≤0,001.

Исходной для решения задачи служит частотная таблица, содержащая m строк и s столбцов по числу уровней изучаемого признака. Корректное решение может быть получено, если число наблюдений будет ≥ 5. При меньшем числе наблюдений можно получить лишь приблизительное решение.

Вычислим критерий Пирсона для всех имеющихся выборок. Значение критерия, уровень значимости и число степеней свободы можно увидеть в строке сверху рассчитанной таблицы

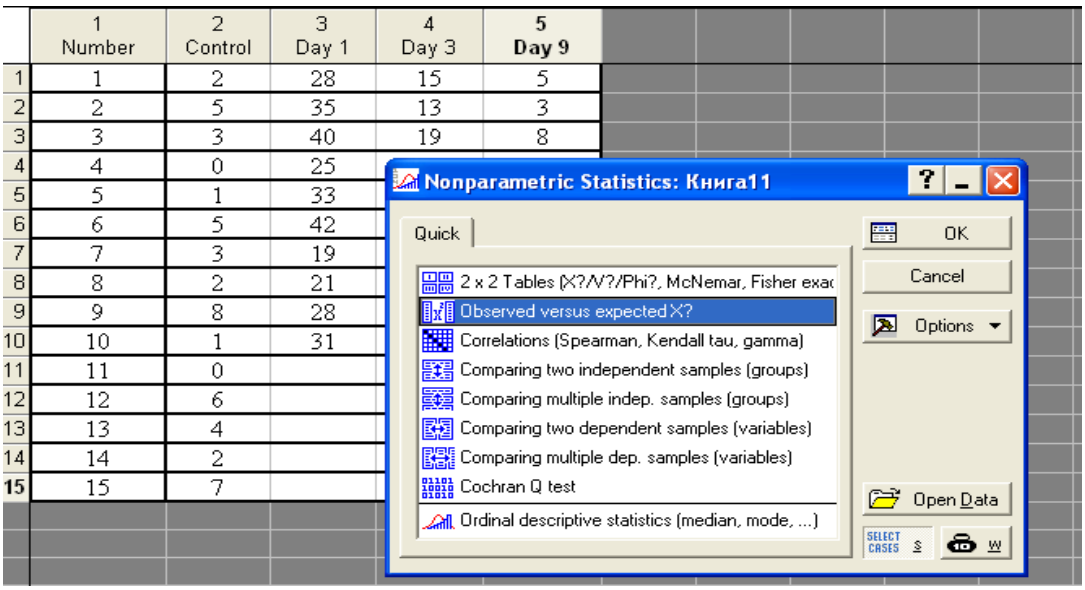


Рис. 11. Выбор расчета критерия «Хи-квадрат».

Observed vs. Expected Frequencies (Книга11)
Chi-Square = 36,07619 df = 9 p < ,000038
NOTE: Unequal sums of obs. & exp. frequencies

Case	observed Control	expected Day 9	O - E	(O-E)**2 /E
	2,00000	5,00000	-3,0000	1,80000
	5,00000	3,00000	2,0000	1,33333
	3,00000	8,00000	-5,0000	3,12500
	0,00000	3,00000	-3,0000	3,00000
	1,00000	7,00000	-6,0000	5,14286
	5,00000	8,00000	-3,0000	1,12500
	3,00000	4,00000	-1,0000	0,25000
	2,00000	5,00000	-3,0000	1,80000
	8,00000	2,00000	6,0000	18,00000
	1,00000	2,00000	-1,0000	0,50000
Sum	30,00000	47,00000	-17,0000	36,07619

Рис. 12. Расчетное значение «хи-квадрат».

Сравним полученные значения с табличными. Вывод поместим в файл «Стат3».

Применяемый в непараметрических методах парный критерий Колмогорова-Смирнова позволяет сравнить между собой две связанные выборки и оценить их подчиненность одному закону распределения. Как и

критерий хи-квадрат, он вызывается из Nonparametrics/Distrib., Comparing two independent samples (groups).

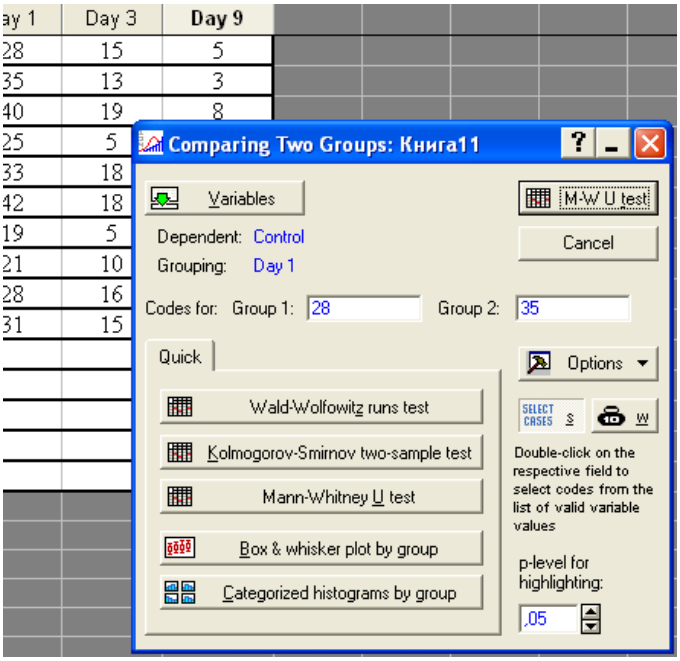


Рис. 13. Вкладка предназначена для выбора тестов.

Аналогично может быть выбран “U-тест Манна- Уитни“. При выполнении всех перечисленных вычислений наиболее существенным является полученное значение p-уровня. Если p<0.05, различие между переменными считается статистически доказанным.

Wilcoxon Matched Pairs Test (Книга11)
Marked tests are significant at p <,05000

Pair of Variables	Valid N	T	Z	p-level
Day 1 & Control	10	0,00	2,803060	0,005062

Рис. 14. Вычисленное значение теста Вилкоксона с рассчитанной величиной p-уровня.

Сравнив полученные различными методами уровни значимости различий между выборками, сделайте вывод о применимости этих методов к имеющимся выборками и поместите его в файл «Стат3»

Дисперсионный анализ

Основной целью дисперсионного анализа является исследование значимости различия между средними. Если вы просто сравниваете средние в двух выборках, дисперсионный анализ даст тот же результат, что и обычный t-критерий для независимых выборок (если сравниваются две

независимые группы объектов или наблюдений) или t-критерий для зависимых выборок (если сравниваются две переменные на одном и том же множестве объектов или наблюдений). Откуда произошло название Дисперсионный анализ? Может показаться странным, что процедура сравнения средних называется дисперсионным анализом. В действительности, это связано с тем, что при исследовании статистической значимости различия между средними двух (или нескольких) групп, мы на самом деле сравниваем (т.е. анализируем) выборочные дисперсии. Фундаментальная концепция дисперсионного анализа предложена Фишером в 1920 году. Возможно, более естественным был бы термин анализ суммы квадратов или анализ вариации, но в силу традиции употребляется термин дисперсионный анализ.

Дисперсионный анализ проводится за некоторыми общими схемами, которые мы рассмотрим на примере однофакторного дисперсионного анализа.

Первичные данные, которые подлежат дисперсионному анализу, группируют в виде корреляционной таблицы, в которой градации организованного (регулируемого) фактора *A* обычно располагают по горизонтали в верхней части таблицы, а числовые значения признака *X,* то есть варианты *Xi,* размещают соответственно по градациям фактора *A*.

Сгруппировав выборочный материал в таблицу, находят средние величины: среднюю арифметическую всего комплекса, так наз. *общую* среднюю *Хобщ* и *групповые средние* x , - по градациям фактора *A:*

$$\bar{x}_{общ} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{x}_j = \frac{\sum_{i=1}^{n_j} x_i}{n_j}$$

где *n*- общее количество опытов, n_j – количество опытов по каждой из градаций. Если количество испытаний в градациях одинаково, то $n_j = L$.

Определяют общую сумму квадратов отклонений *Cy,* равную сумме квадратов отклонений всех значений случайной величины от общей средней, то есть:

$$C_y = \sum_{i=1}^{n} (x_i - \bar{x}_{о\ б})^2_{щ}$$

Определяют сумму квадратов отклонений по организованным факторам, с учетом статистического веса групповых средних (характеризует рассеивание между группами). При одинаковом числе вариант в градациях комплекса эта сумма

$$C_x = \sum_{j=1}^{r} (\bar{x}_j - \bar{x}_{о\ б})^2_{щ} L ,$$

если L постоянно для всех опытов. При разных количествах вариант

–

$$C_x = \sum_{j=1}^{r} (\bar{x}_j - \bar{x}_{о\ б})^2_{щ} n_j$$

Вычисляют сумму квадратов отклонений по случайным факторам (характеризует рассеивание в группах):

$$C_z = \sum_{i=1}^{n} \sum_{j=1}^{r} (x_i - \bar{x}_j)^2$$

Практически остаточную сумму находят по формуле:

$$C_z = C_y - C_x$$

Определив сумму квадратов отклонений, устанавливают число степеней свободы, которые могут быть равны:

для общей дисперсии $k_y = n - 1$

для факторной $k_x = r - 1$

для случайной (остаточной) $k_z = n - r$

Отношением сумм квадратов отклонений к соответствующему числу степеней свободы определяются дисперсии:

общая: $\sigma_y^2 = \frac{C_y}{k_y}$

факторная: $\sigma_x^2 = \frac{C_x}{k_x}$

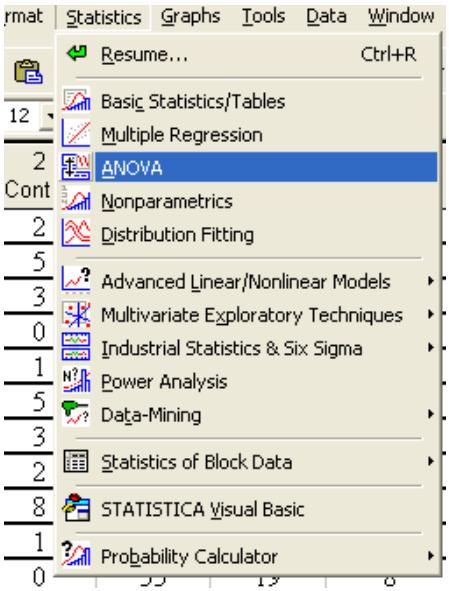
случайная (остаточная) $\sigma_z^2 = \frac{C_z}{k_z}$

Если $\sigma_x^2 > \sigma_z^2$, то делаем предварительный вывод о том, что фактор имеет влияние; если $\sigma_x^2 < \sigma_z^2$,то фактор не оказывает существенного влияния.

Для установления значимости вывода о влиянии регулируемого фактора на результативный признак используют критерий Фишера - Снедекора, в виде отношения факторной дисперсии к дисперсии случайной (остаточной):

$$F_{эксп} = \frac{\sigma_ч^2}{\sigma_я^2}$$

Заключительным этапом дисперсионного анализа является сравнение фактической величины критерия $F_{эксп}$ с его стандартным значением $F\ (a\ ;k_z;\ k_x)$ найденным по Таблице 1 Приложений для уровня значимости a и значениям степеней свободы k_z и k_x.



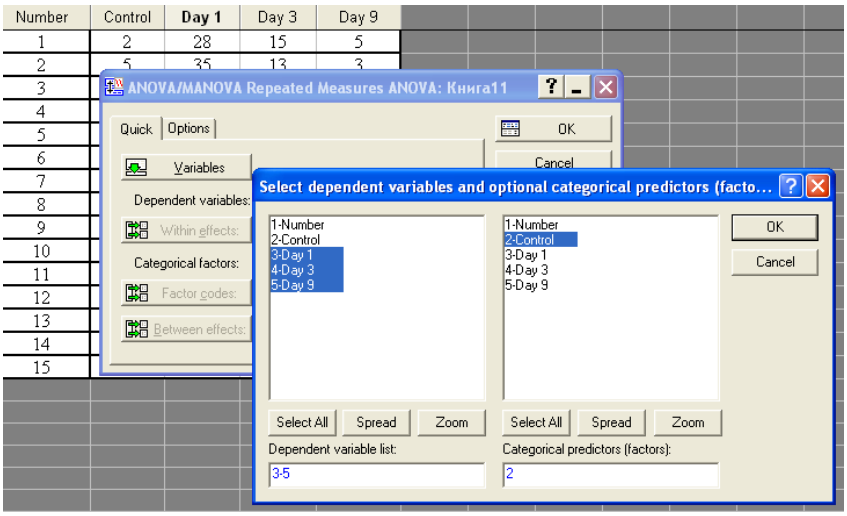
Если Fэ*кспи* $>F$, - делают вывод о статистической значимости влияния на результат исследования организованных факторов, если Fэ*кспи* $<F$, то такой вывод делать нельзя.

В пакете Статистика для проведения дисперсионного анализа используется модуль **ANOVA/MANOVA** (**Analysis of variance** в случае однофакторного анализа и **Multiply analysis of variance** в случае многофакторного).

На линейке меню выберем **ANOVA/MANOVA. - Repeated measures ANOVA**

Зададим переменные: независимые – факторы **Categorical predictors** и зависимые – исследуемые **Dependent variables list.**

Как упоминалось выше, при дисперсионном анализе желательно, чтобы переменные имели значения в виде натуральных чисел. В появившемся затем окне следует выбрать



На следующей вкладке выберем – **All effects**. Результат вычислений:

Multivariate Tests of Significance (Книга11)
Sigma-restricted parameterization
Effective hypothesis decomposition

GENERAL Effect	Test	Value	F	Effect df	Error df	p
Intercept	Wilks	0,018439	70,97595	3	4,00000	0,000634
Control	Wilks	0,021028	1,41696	24	12,20243	0,266909

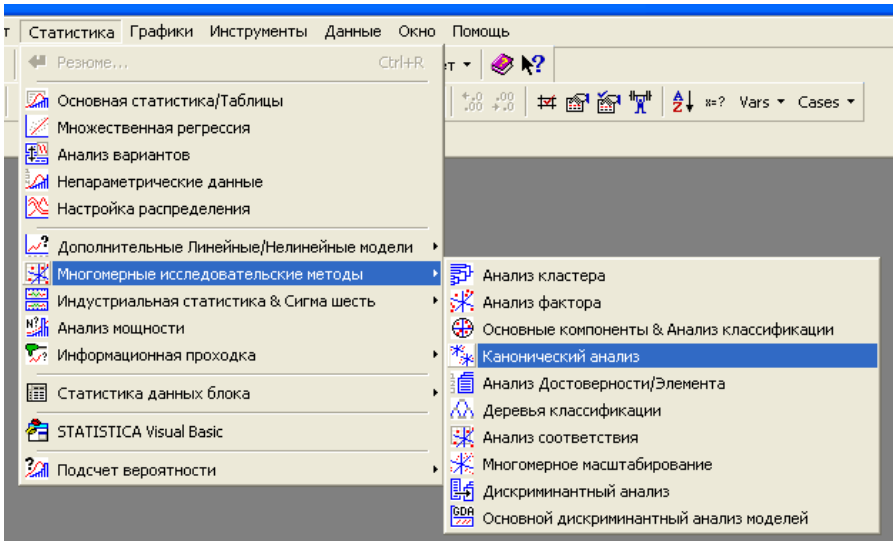
Рис. 15. Результаты вычислений.

Здесь F- значение критерия Фишера. Сравнив его значение с табличным при определенном уровне значимости p, можно сделать вывод о статистической значимости влияния исследованных признаков на результат

исследования. Этот вывод также следует поместить в файл «Стат3», сохранить и отправить по электронной почте на адрес strahova@zsmu.zp.ua.

Канонический анализ.

Модуль **Канонический анализ** предназначен для анализа зависимостей между списками переменных, тем самым развивая возможности других модулей программы. В медицине и фармации он может служить для изучения, например, зависимости между влиянием различных неблагоприятиных для здоровья факторов и появлением определенной группы симптомов, либо для исследования зависимости результатов лечения пациентов какой-либо нозологической группы с помощью определенной медикаментозной схемы и полученными результатами лечения. Рассмотрим последовательность шагов при работе с модулем **Канонический анализ.** В верхнем меню Статистика выберите **Многомерные исследовательские методы**, затем **Канонический анализ.** Откроется стартовая панель модуля. Рассмотрим функциональное назначение основных кнопок.



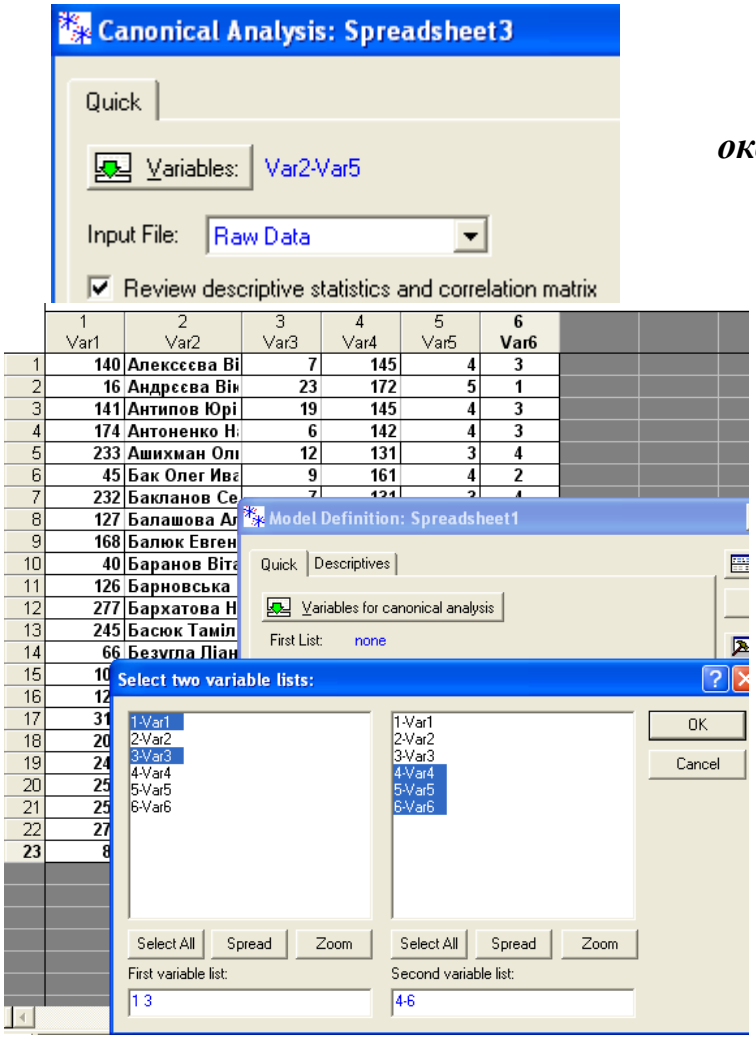


Рис. 16 – 19. Внешний вид окон канонического анализа.

После нажатия кнопки **Variables** открывается стандартное окно выбора переменных, в котором можно выбрать переменные для анализа. Только переменные, выбранные в этом окне, впоследствии доступны для канонического анализа. Матрица корреляции будет вычислена для всех переменных, выбранных в этом окне. Мы возьмем все, кроме первой (объясните почему). Поле списка **Input File** имеет два возможных значения: необработанные исходные данные *Raw Data* и матрица корреляции *Correlation Matrix.* Если выбрана первая опция, программа ожидает на входе файл с необработанными исходными данными. Если выбрана вторая опция, в качестве файла данных необходимо указать файл, содержащий соответствующую матрицу корреляции в стандартном формате матричного файла *STATISTICA.* Файлы корреляционных матриц могут быть созданы в различных модулях системы (например, в модулях **Basic Statistics/Tables, Factor Analysis, Multiply Regression).** Выберите *Raw Data.*

Установите флажок на *Review descriptive statistics and correlation matrix* (отображать описательные статистики и корреляционные матрицы), чтобы

после выхода из стартовой панели открыть окно **Review Descriptive statistics** (просмотр описательных статистик). Это окно позволяет просмотреть описательные статистики для выбранных переменных. (этого можно не делать). Нажмите ОК.

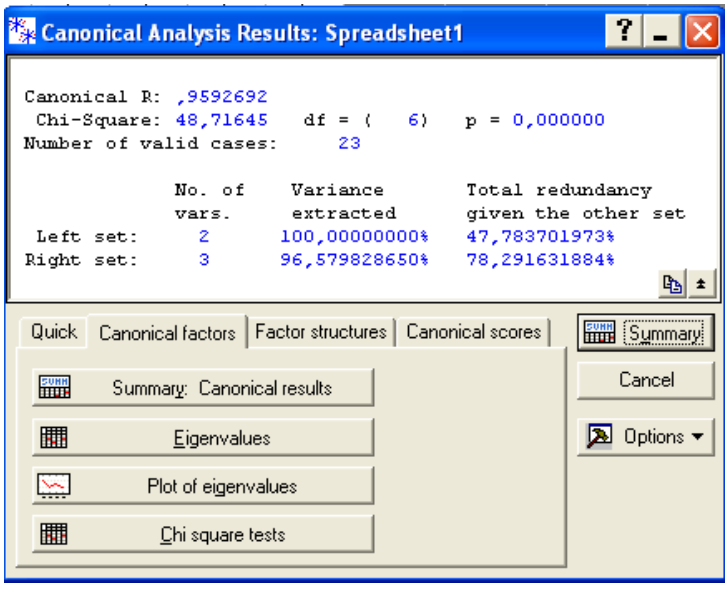
В открывшемся окне **Model Definition** (определение модели) нажмите кнопку **Variables** и выберите в первом списке переменных *Var1,Var3*, во втором – *Var4-Var6 (можете выбрать другие, на свое усмотрение)*.

Если установить флажок на *Bath processing/reporting* (пакетная обработка/печать) и выбрать вывод на принтер, все выходные результаты будут распечатаны без вмешательства пользователя.

Если выбрать вкладку **Descriptives,** откроется окно просмотра описательных статистик.

Нажмите ОК, появится окно **Canonical Analysis Results** (результаты анализа канонической корреляции) Наиболее существенные результаты анализа приведены в верхней информационной части окна.

Рассмотрим внимательно полученные значения. Каноническая корреляция *R*, приведенная в верхней строке окна, соответствует корреляции между первыми каноническими переменными (взвешенными суммами). Она равна максимальному извлеченному каноническому корню. Ее значение свидетельствует о наличии или отсутствии зависимости между группами переменных. Значения *Chi-Square* и уровень значимости p = 0,00 показывают значимость *R*.



Number of valid cases - число наблюдений. В нижней информационной части окна приведены *No. Of vars* (число переменных в левом и правом множествах); *Variance extracted* (процентное число извлеченных дисперсий из левого и правого множеств переменных); *Total redundancy given the other set* (общая избыточность при заданном втором множестве).

Нажмите последовательно все кнопки, имеющиеся на показанной вкладке и рассмотрите полученные результаты вычислений. Затем перейдите на другую вкладку рассматриваемого модуля и также просмотрите все возможные варианты статистической обработки данных в рамках модуля Канонический анализ.

Summary. Canonical results (итоговые результаты) - на экран будет выведена таблица результатов с итоговыми значениями статистик для текущего анализа.

С нажатием кнопки **Eigenvalues п**оявится таблица результатов, содержащая собственные значения, соответствующие каноническим корням.

.

Кнопка **Plot of Eigenvalues** предоставит изображение кусочно-линейного графика убывающих собственных значений.

Кнопка **Chi-Square tests дает** таблицу результатов, содержащую для каждого канонического корня значения R, χ^2, число степеней свободы, *p - уровень*.

Результаты этой таблицы показывают, какие канонические корни следует считать статистически значимыми, чтобы использовать их для дальнейшего рассмотрения (т.е. для интерпретации).

Выделите вкладку **Canonical scores** (канонические значения, крайняя справа вкладка) и нажмите кнопку **Left & right set canonical weights** (канонические веса для левого и правого множества). На экран будут выведены таблицы результатов с каноническими весами для каждого множества

переменных. Чем больше **абсолютное** значение веса, тем больше вклад соответствующей переменной в значение канонической переменной.

Вспомните, с помощью каких модулей мы проводили ранее вычисление значений коэффициента корреляции, критерия Пирсона, каким образом определяли уровень значимости. Сравните значения этих величин, вычисленных для одного и того же набора исходных данных, полученные с помощью различных модулей программы СТАТИСТИКА. Результаты вычислений сохраните в файле формата .xls или .sta (файлы создаются с помощью прохрамм Excel и СТАТИСТИКА). Интерпретируйте полученные значения R, p и χ^2. Какие выводы относительно результатов статистической обработки данных и сравнимости результатов, полученных с помощью различных модулей программы СТАТИСТИКА, вы можете сделать, используя весь пройденный теоретический материал?

Кластерный анализ.

Термин кластерный анализ, впервые введенный Трионом (Tryon) в 1939 году, включает в себя более 100 различных алгоритмов.

В отличие от задач классификации, кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальным данным, частотам, бинарным данным). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

Кластерный анализ позволяет сокращать размерность данных, делать ее наглядной.

Кластерный анализ может применяться к совокупностям временных рядов, здесь могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой.

Задачи кластерного анализа можно объединить в следующие группы:

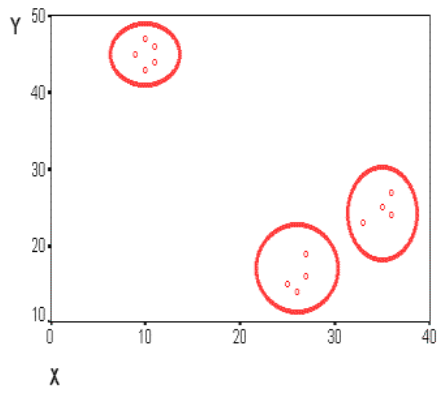
1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Как правило, при практическом использовании кластерного анализа одновременно решается несколько из указанных задач.

Рассмотрим пример процедуры кластерного анализа.

Допустим, мы имеем набор данных A, состоящий из 14-ти примеров, у которых имеется по два признака X и Y. Данные по ним приведены в таблице.

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X	27	11	25	36	35	10	11	36	26	26	9	33	27	10
Y	19	46	15	27	25	43	44	24	14	14	45	23	16	47



Данные в табличной форме не носят информативный характер. Представим переменные X и Y в виде диаграммы рассеивания, изображенной на рис.20.

Рис. 20 Диаграмма рассеивания X и Y

На рисунке мы видим несколько групп "похожих" примеров. Примеры (объекты), которые по значениям X и Y "похожи" друг на друга, принадлежат к одной группе (кластеру); объекты из разных кластеров не похожи друг на друга.

Критерием для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике. Способов определения ***меры расстояния*** между кластерами, называемой еще мерой близости, существует несколько. Наиболее распространенный способ - вычисление ***евклидова расстояния*** между двумя точками i и j на плоскости, когда известны их координаты X и Y:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \qquad (1)$$

Чтобы узнать расстояние между двумя точками, надо взять разницу их координат по каждой оси, возвести ее в квадрат, сложить полученные значения для всех осей и извлечь квадратный корень из суммы.

Когда осей больше, чем две, расстояние рассчитывается таким образом: сумма квадратов разницы координат состоит из стольких слагаемых, сколько осей (измерений) присутствует в нашем пространстве. Например, если нам нужно найти расстояние между двумя точками в пространстве трех измерений (такая ситуация представлена на рис. 21), формула (1) приобретает вид:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \qquad (2)$$

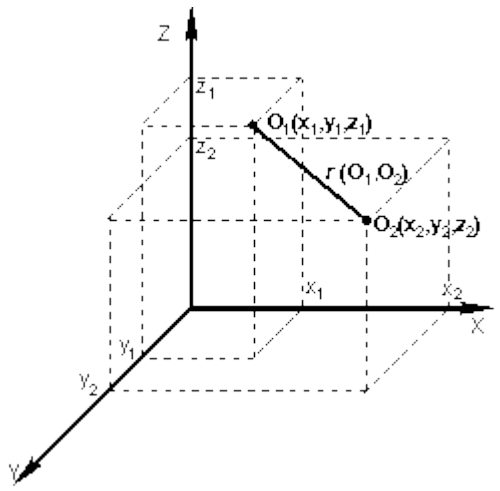


Рис. 21. Расстояние между двумя точками в пространстве трех измерений

Кластер имеет следующие **математические характеристики***:* центр, радиус, среднеквадратическое отклонение, размер кластера.

Центр кластера - это среднее геометрическое место точек в пространстве переменных.

Радиус кластера - максимальное расстояние точек от центра кластера.

Кластеры могут быть перекрывающимися. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют спорными.

Спорный объект - это объект, который по мере сходства может быть отнесен к нескольким кластерам.

Размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше

радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

Неоднозначность данной задачи может быть устранена экспертом или аналитиком.

Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта допускают желательное разбиение совокупности объектов на кластеры. Второе предположение - правильность выбора масштаба или единиц измерения признаков.

Выбор масштаба в кластерном анализе имеет большое значение. Рассмотрим пример. Представим себе, что данные признака x в наборе данных A на два порядка больше данных признака y: значения переменной x находятся в диапазоне от 100 до 700, а значения переменной y - в диапазоне от 0 до 1.

Тогда, при расчете величины расстояния между точками, отражающими положение объектов в пространстве их свойств, переменная, имеющая большие значения, т.е. переменная x, будет практически полностью доминировать над переменной с малыми значениями, т.е. переменной y. Таким образом из-за неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками.

Эта проблема решается при помощи предварительной стандартизации переменных. ***Стандартизация*** (standardization) или нормирование (normalization) приводит значения всех преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака. Существуют различные способы нормирования исходных данных.

Два наиболее распространенных способа:

- деление исходных данных на среднеквадратичное отклонение соответствующих переменных;
- вычисление Z-вклада или стандартизованного вклада.

Наряду со стандартизацией переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов - специалистов предметной области. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных.

В ходе экспериментов возможно сравнение результатов, полученных с учетом экспертных оценок и без них, и выбор лучшего из них.

Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы:

- **иерархические;**
- **неиерархические.**

Каждая из групп включает множество подходов и алгоритмов.

Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

Рассмотрим иерархические и неиерархические методы подробно.

Иерархические методы кластерного анализа

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES)

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA)

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде дендрограммы показан на рис.22.

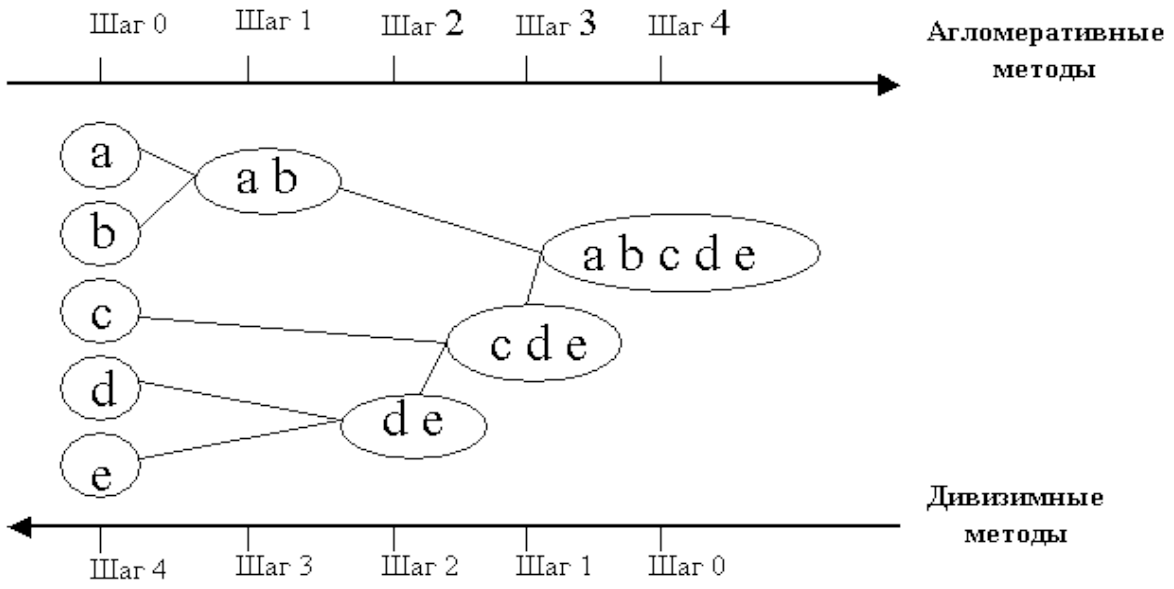


Рис. 22. Дендрограмма агломеративных и дивизимных методов

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).

Иерархические методы кластерного анализа используются при небольших объемах наборов данных.

Преимуществом иерархических методов кластеризации является их наглядность.

Иерархические алгоритмы связаны с построением дендрограмм *(*от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа. **Дендрограмма** описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Дендрограмма (dendrogram) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров.

Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

Существует много способов построения дендограмм. В дендограмме объекты могут располагаться вертикально или горизонтально. Пример вертикальной дендрограммы приведен на рис. 23.

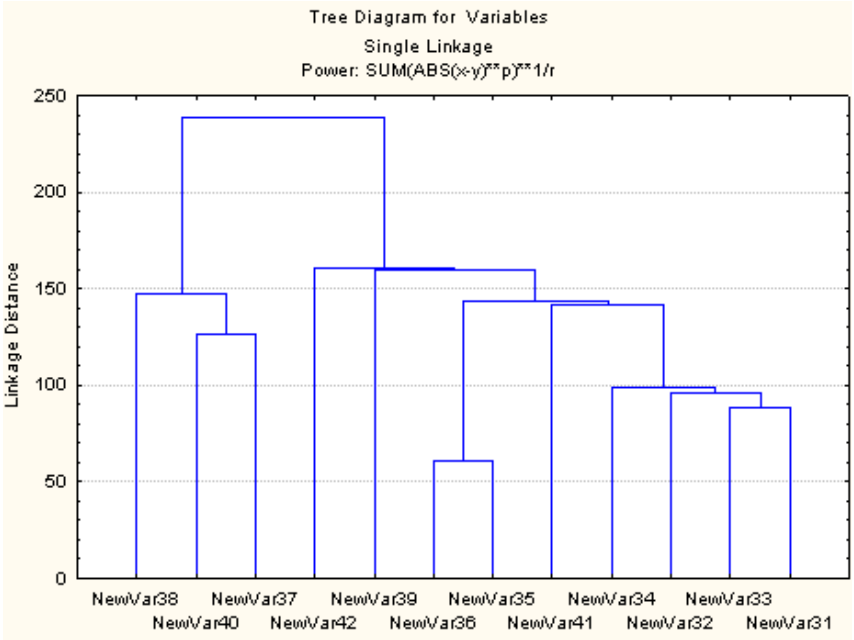


Рис. 23. Пример дендрограммы

Мы видим, что на первом шаге каждое наблюдение представляет один кластер, на втором шаге наблюдаем объединение наблюдений. Данный процесс продолжается до тех пор, пока все наблюдения не объединятся в один кластер.

Меры сходства

Для вычисления расстояния между объектами используются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний. Евклидово расстояние наиболее популярная мера сходства. Для придания больших весов более отдаленным друг от друга объектам можем воспользоваться квадратом евклидова расстояния путем возведения в квадрат стандартного евклидова расстояния.

Манхэттенское расстояние (расстояние городских кварталов), также называемое "хэмминговым" или "сити-блок" расстоянием.

Это расстояние рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния евклида. Однако, для этой меры влияние отдельных выбросов меньше, чем при использовании евклидова расстояния, поскольку здесь координаты не возводятся в квадрат.

Расстояние Чебышева. Это расстояние стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.

Процент несогласия. Это расстояние вычисляется, если данные являются категориальными.

Методы объединения или связи

Когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Возникает следующий вопрос - как определить расстояния между кластерами? Существуют различные правила, называемые методами объединения или связи для двух кластеров.

Метод ближнего соседа или **одиночная связь**. Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее

близкими объектами (ближайшими соседями) в различных кластерах. Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов. В результате работы этого метода кластеры представляются длинными "цепочками" или "волокнистыми" кластерами, "сцепленными вместе" только отдельными элементами, которые случайно оказались ближе остальных друг к другу.

Метод наиболее удаленных соседей или полная связь. Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Метод хорошо использовать, когда объекты действительно происходят из различных "рощ". Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочечным", то этот метод не следует использовать.

Неиерархические методы.

Алгоритм k-средних (k-means)

Наиболее распространен среди неиерархических методов алгоритм k-средних, также называемый быстрым кластерным анализом. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров. Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может

базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Описание алгоритма

Первоначальное распределение объектов по кластерам.

Выбирается число k, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр.

Выбор начальных центроидов может осуществляться следующим образом:

- выбор k-наблюдений для максимизации начального расстояния;
- случайный выбор k-наблюдений;
- выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

Итеративный процесс.

Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

• кластерные центры стабилизировались, т.е. все наблюдения принадлежат ккластеру, которому принадлежали до текущей итерации;

• число итераций равно максимальному числу итераций.

На рис. 24 приведен пример работы алгоритма k-средних для k, равного двум.

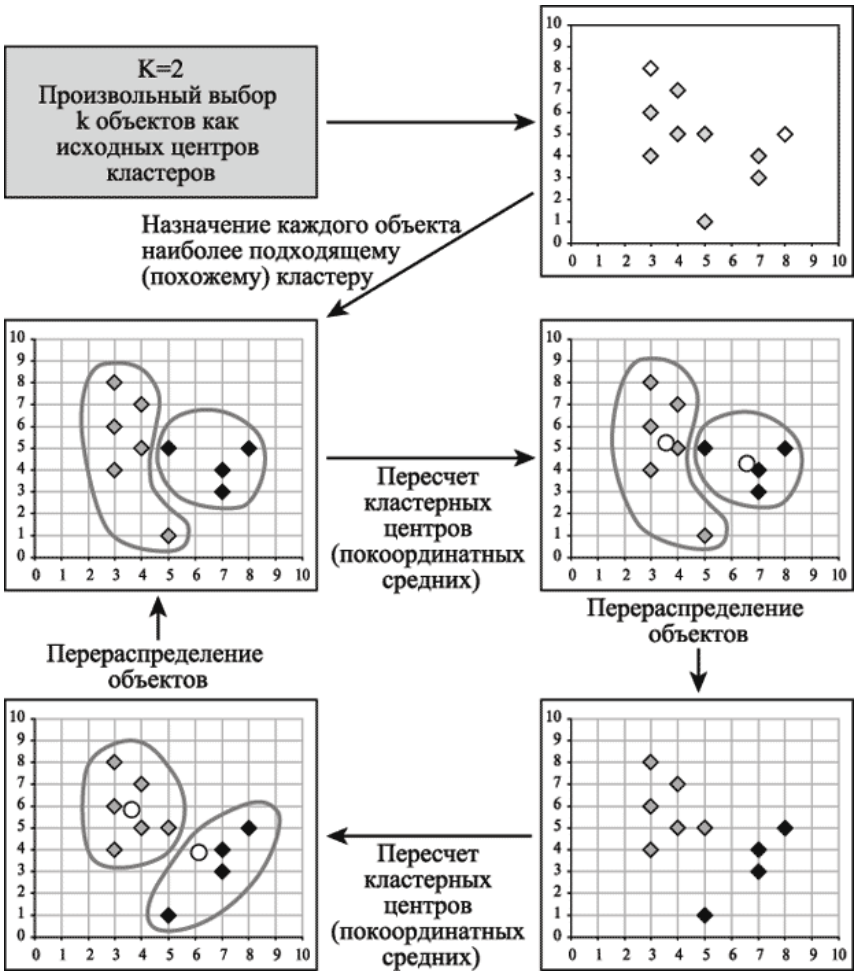


Рис. 24. Пример работы алгоритма k-средних (k=2)

Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т.д., сравнивая полученные результаты.

Проверка качества кластеризации

После получений результатов кластерного анализа методом k-средних следует проверить правильность кластеризации (т.е. оценить, насколько кластеры отличаются друг от друга). Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

Достоинства алгоритма k-средних:

- простота использования;
- быстрота использования;
- понятность и прозрачность алгоритма.

Недостатки алгоритма k-средних:

- алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма - алгоритм k-медианы;
- алгоритм может медленно работать на больших базах данных. Возможным решением данной проблемы является использование выборки данных.

Анализ результатов кластеризации. Этот этап подразумевает решение таких вопросов: не является ли полученное разбиение на кластеры случайным; является ли разбиение надежным и стабильным на подвыборках данных; существует ли взаимосвязь между результатами кластеризации и переменными, которые не участвовали в процессе кластеризации; можно ли интерпретировать полученные результаты кластеризации.

Проверка результатов кластеризации. Результаты кластеризации также должны быть проверены формальными и неформальными методами. Формальные методы зависят от того метода, который использовался для кластеризации. Неформальные включают следующие процедуры проверки качества кластеризации:

- анализ результатов кластеризации, полученных на определенных выборках набора данных;
- кросс-проверка;
- проведение кластеризации при изменении порядка наблюдений в наборе данных;
- проведение кластеризации при удалении некоторых наблюдений;
- проведение кластеризации на небольших выборках.

Один из вариантов проверки качества кластеризации - использование нескольких методов и сравнение полученных результатов. Отсутствие подобия не будет означать некорректность результатов, но присутствие похожих групп считается признаком качественной кластеризации.

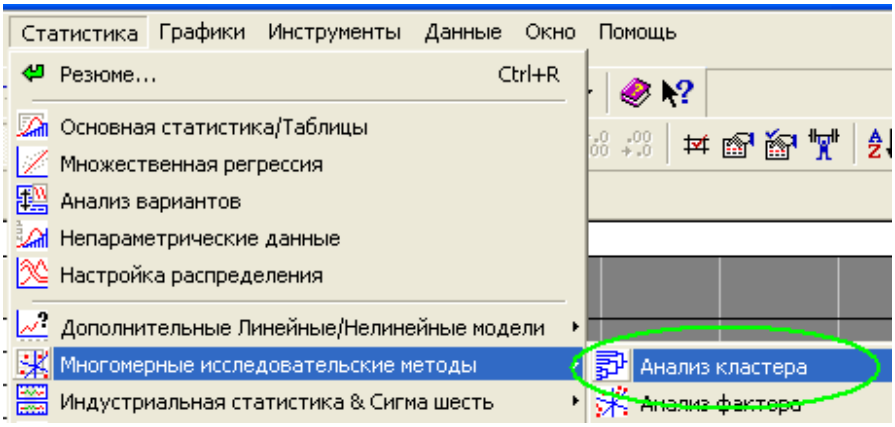
Если нет предположений относительно числа кластеров, рекомендуют использовать иерархические алгоритмы. Однако если объем выборки не позволяет это сделать, возможный путь - проведение ряда экспериментов с различным количеством кластеров, например, начать разбиение совокупности данных с двух групп и, постепенно увеличивая их количество, сравнивать результаты. За счет такого "варьирования" результатов достигается достаточно большая гибкость кластеризации.

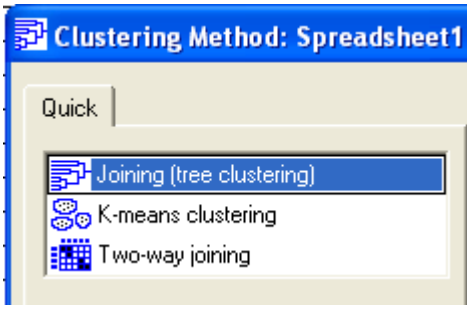
При использовании иерархических методов существует возможность достаточно легко идентифицировать выбросы в наборе данных и, в результате, повысить качество данных. Эта процедура лежит в основе

двухшагового алгоритма кластеризации. Такой набор данных в дальнейшем может быть использован для проведения неиерархической кластеризации.

Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные в некотором смысле группы, или **кластеры**. В области медицины кластеризация заболеваний, лечения заболеваний приводит к широко используемым таксономиям. Иными словани, всякий раз, когда необходимо классифицировать большие информационные массивы на пригодные для дальнейшей обработки группы, применяется кластерный анализ. Большое достоинство кластерного анализа в том, что он позволяет проводить разбиение генеральной совокупности не по одному, а по ряду параметров одновременно. Кроме этого, он не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет исследовать множество исходныхданных практически произвольной природы.

Для запуска этого модуля из меню **Statistics** надо выбрать **Многомерные исследовательские методы** и перейти к разделу **Анализ кластера**. Откроется стартовая панель модуля. На вкладке **Quick** находится список методов кластерного анализа, реализованных в программе *STATISTICA*.





Это:

1. **Joining tree clustering** (древовидная кластеризация),

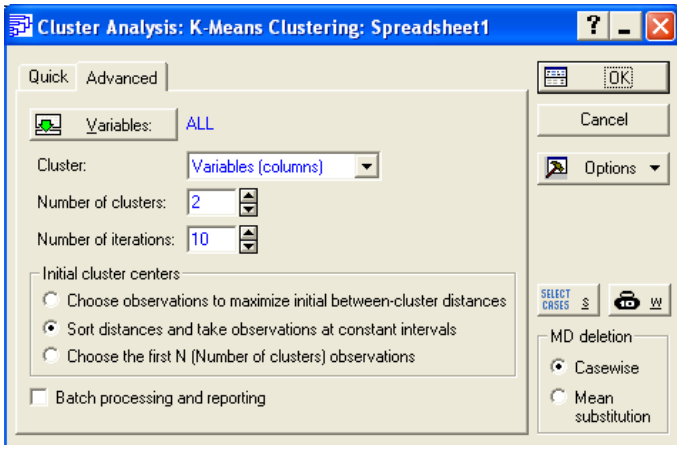
2. **k-means clustering** (метод k-средних)

3. **Two-way joining** (двухвходовая кластеризация)

Рис. 25-33. Внешний вид окон кластерного анализа

1. k-means clustering .

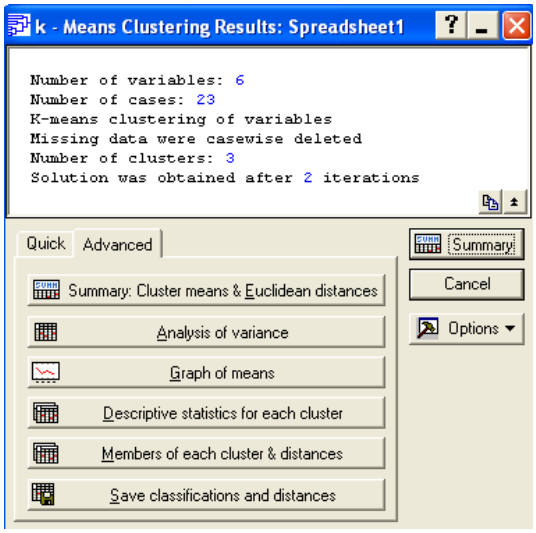
Например, необходимо разбить группу студентов на несколько однородных групп, в которых студенты мало отличаются друг от друга (существенно меньше, чем в совокупности) по результатам сдачи экзамена. Сложность задачи в том, что надо сравнивать участников не по какому-то одному параметру (признаку), а по нескольким параметрам одновременно. В главной части стартовой панели высветите **k-means clustering** и нажмите ОК, на экране появится диалоговое окно **k-means clustering.** Перейдите на вкладку **Advanced** и выберите переменные для анализа.



Для этого нажмите кнопку **Variables** в левом верхнем углу экрана и откройте диалоговое окно **Select variables for analysis.** Для выбора всех параметров нажмите кнопку **Select All,** а затем — ОК. Программа вернется в стартовое окно модуля.

На поле **Claster** надо выбрать объекты для кластеризации. Так как цель исследования — кластеризация студентов, которые являются в файле данных наблюдениями, выберите *Cases (rows)* (наблюдения (строки)).

В поле **Number of clasters** (число кластеров) нужно определить число групп, на которые мы хотим разбить автомобили. Запишите в это поле число 3.



В поле **Number of iterations** (число итераций) задается максимальное число итераций, используемых при построении классов. Задайте, например, число 10. Если необходимо провести кластеризацию не по всем объектам (в данном случае студентам), воспользуйтесь кнопкой **Select cases.**

Группа опций *Initial claster centers* позволяет задать начальные центры кластеров: *Choose observations to maximize initial between-cluster distances* (выбрать наблюдения, максимизирующие начальные расстояния между кластерами); *Sort distances and take observations at constant intervals* (сортировать расстояния и выбрать наблюдения на постоянных интервалах); *Choose the first N (Namber of clasters) observations* (выбрать первые *N* (число кластеров) наблюдений). Выберите, например, вторую опцию и нажмите ОК. Откроется окно результатов **k-means Clustering Results**.

В верхней информационной части окна представлены следующие данные:

• *Number of variables* (количество переменных);

• *Number of cases* (число наблюдений);

• *k-means clustering of cases* (метод *k-средних);*

• *Missing data were casewise deleted* (обработка пропущенных значений опущена);

• *Number of clusters* (число кластеров);

• *Solution was obtained after * iterations* (решение было найдено после* итераций).

Откройте вкладку **Advanced,** так как она содержит более подробную информацию о результатах анализа. Функциональное назначение кнопок открывшегося окна следующее.

Summary:Cluster means & Euclidean distances предназначена для вывода таблиц, в первой из которых указаны средние для каждого кластера (усреднение производится внутри кластера), во второй — евклидовы расстояния и квадраты евклидовых расстояний между кластерами.

Analysis of variance выводит таблицу дисперсионного анализа.

В таблице приведены значения межгрупповых *(Between SS)* и внутригрупповых *(Within SS)* дисперсий признаков. **Чем меньше значение внутригрупповой дисперсии и больше значение межгрупповой дисперсии, тем лучше признак характеризует принадлежность объектов к кластеру и тем «качественнее» кластеризация**. Параметры *F* и *p* также характеризуют вклад признака в разделение объектов на группы. Лучшей кластеризации соответствуют большие значения первого и меньшие значения второго параметра. Признаки с большими значениями *p* (например, больше 0,05) можно из процедуры кластеризации исключить.

Graph of means позволяет просмотреть средние значения для каждого кластера на линейном графике.

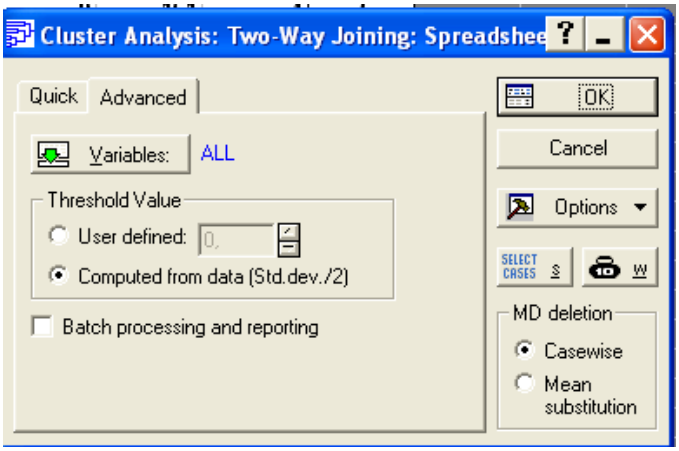
Descriptive statistics for each cluster выводит электронную таблицу с описательными статистиками для каждого кластера (среднее, дисперсия и т.д.).

Members of each clusters & **distances** предназначена для просмотра распределения объектов по кластерам. В таблице также будет указано расстояние от объекта до центра кластера.

Save classifications and distances сохраняет результаты классификации в файле *STATISTICA* для дальнейшего исследования. Чем лучше результаты кластеризации, тем сильнее различаются средние в различных группах. Можно менять количество кластеров и набор переменных, принятых для рассмотрения, чтобы получить наиболее качественное разбиение набора исходных наблюдений на группы. Переменные или наблюдения, имеющие малые значения *F* и большие значения *p* можно исключать из рассмотрения**.**

Все описанные процедуры повторяются до получения приемлемого результата. Если он все-таки не был достигнут, можно сделать вывод о заметной однородности исследуемых данных, и невозможности разделения их на существенно различающиеся по характеристикам группы.

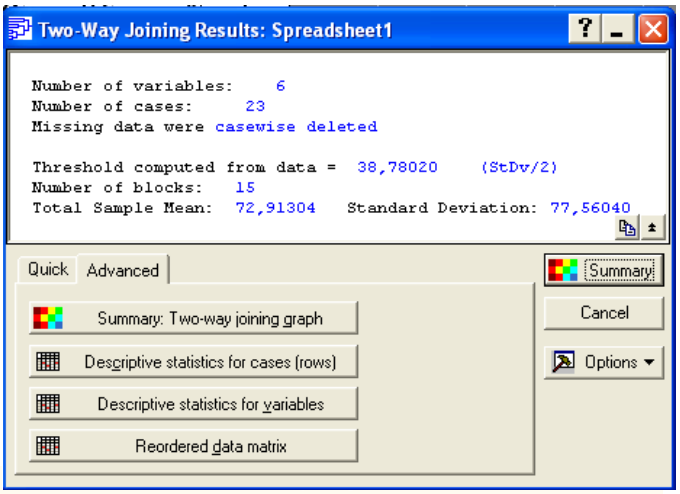
2. Two-way joining.

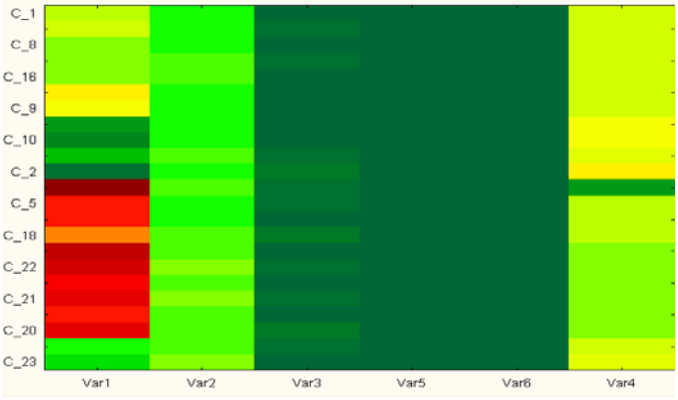


Рассмотрим процедуру одновременной кластеризации по переменным (столбцам) и по случаям (строкам). В стартовой панели модуля выберите **Two-way joining.** Нажмите ОК. Откроется диалоговое окно.

Нажмите кнопку **Variables** и укажите переменные для анализа, например, *Select All.* На вкладке **Advanced** имеется возможность выбрать пороговый параметр *Threshold Value* (значение порога). Пороговый параметр определяет, когда алгоритм рассматривает в матрице данных два числа как равные, а затем приписывает их к одному кластеру. Если эта величина слишком велика (по отношению к числам в матрице данных), то будет сформирован только один кластер; если она очень мала, то кластером будет являться каждая точка данных. Параметр может назначить пользователь — *User defined.* Но для большинства случаев рекомендуется величина по умолчанию — *Computed from data* (общее стандартное отклонение, деленное на 2). Опция *Batch processing and reporting* доступна при определенных установках в *Output Manager* (менеджере вывода). Нажмите ОК. На экране появится окно результатов.

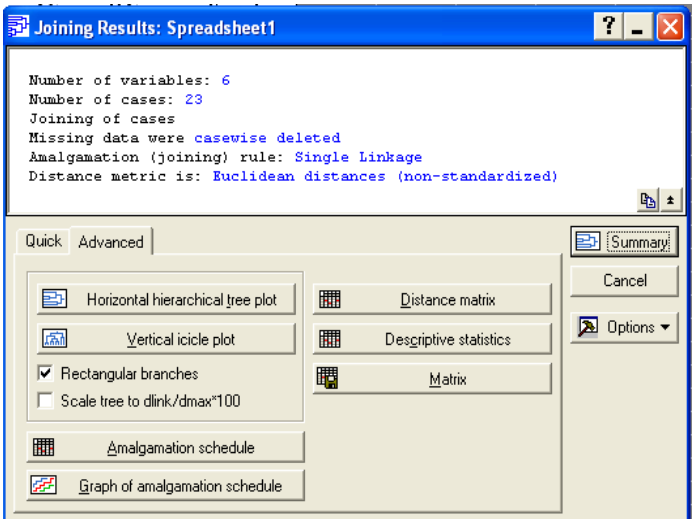
В верхней информационной части окна указано число переменных; число наблюдений; пороговое значение; число полученных блоков разбиения; стандартное отклонение.





Следующие кнопки позволяют провести анализ результатов.

- **Summary: Two-way joining graph** (просмотреть графическое представление результатов).
- **Descriptive statistics for cases (row)** (описательные статистики для наблюдений).
- **Descriptive statistics for variables** (описательные статистики для переменных).
- **Reordered data matrix** (неупорядоченная матрица значений).



Нажмите кнопку **Summary: Two-way joining graph.** Появится цветной график результата кластеризации. Одним цветом обозначены объекты, попавшие в один кластер.

3. **Joining tree clustering**

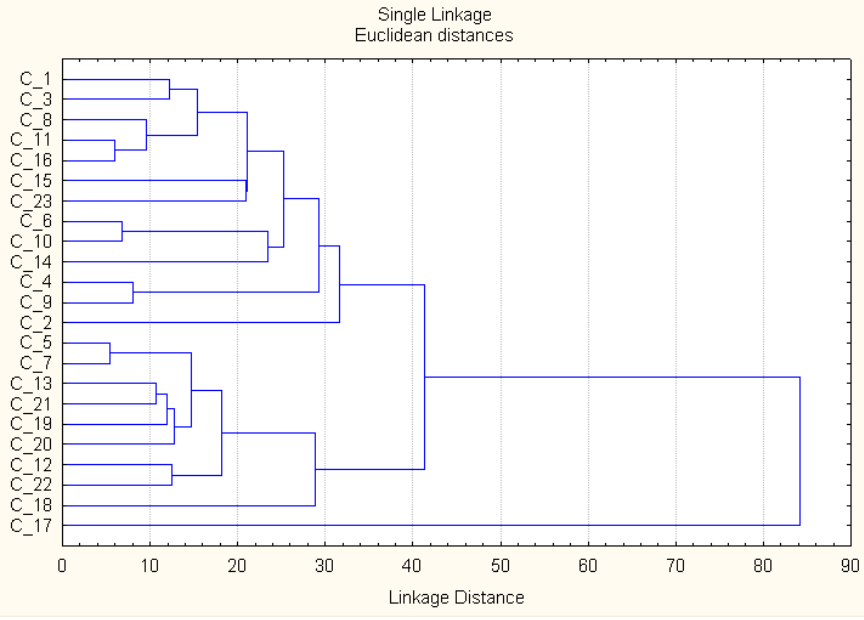
Для кластеризации агломеративным методом древовидной кластеризации надо в стартовой панели высветить **Joining tree clustering** и нажать на ОК. Появится диалоговое окно. Для переменных сделайте установку **Select All.** Так

как кластеризуется список студентов с результатами экзамена, в поле **cluster** выберите пункт *Cases (row).* В качестве исходных данных используется матрица значений (а не матрица расстояний), поэтому в поле **Input file** надо выбрать *Raw data.* В поле **Amalgamation (linkage) rule** (правило иерархического объединения) выберите правило объединения в кластеры, например *Single Linkage* (метод одиночной связи), и нажмите ОК. Появится окно результатов.

В верхней части окна записана информация: число переменных, число наблюдений, метод кластеризации, правило иерархического объединения, выбранная метрика (расстояние между объектами).

Кнопки в нижней части окна на вкладке **Advanced** предназначены для анализа результатов кластеризации:

- **Horizontal hierarchical tree plot** (горизонтальная древовидная диаграмма).
- **Vertical icicle plot** (вертикальная древовидная диаграмма).
- **Amalgamation schedule** (правило объединения в кластеры).
- **Graph of amalgamation schedule** (график порядка объединения).
- **Distance matrix** (матрица расстояний).
- **Descriptive statistics** (описательные статистики).



Нажмите **Horizontal hierarchical tree plot.** Диаграмма начинается с каждого объекта в классе (в левой части диаграммы).

Теперь представьте, что постепенно (очень малыми шагами) «ослабевает» критерий, показывающий, какие объекты

являются уникальными, а какие нет. Другими словами, понижается порог, относящийся к решению об объединении двух или более объектов в один кластер. В результате связывается все большее и большее число объектов и агрегируются (объединяются) все больше кластеров, состоящих из все сильнее различающихся элементов. На последнем шаге все объекты окончательно объединяются. 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) можно определить 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 может быть отражена в иерархическом дереве различными ветвями. В результате успешного анализа методом объединения появляется возможность обнаружить кластеры (ветви) и интерпретировать их. В рассматриваемом случае, можно сделать вывод, насколько разнятся полученные студентами оценки, и каким образом можно сгруппировать наблюдения, учитывая полученные студентами баллы, экзаменационные оценки и номера студенческих групп.

Задание для самостоятельной работы.

Используя антропометрические данные студентов группы, полученные в ходе выполнения предыдущей лабораторной работы (таблица значений роста, веса, окружности запястья и талии, давления), провести кластерный анализ информационного массива. Построить горизонтальную древовидную диаграмму с применением вкладки **Joining tree clustering**, а также, используя вкладку **k-means clustering,** создать рациональное, с вашей точки зрения,

число кластеров. Результаты сохранить в виде файлов в формате STATISTICA.

Для обобщенного информационного массива, полученного от других студенческих групп, произвести вычисление статистических характеристик; оценить характер распределения данных в генеральной совокупности. В случае, если распределение однородных данных (рост, вес, давление и т.д, каждый определенный параметр - отдельно) близко к нормальному или нормальное, провести дисперсионный анализ данных. Если же распределение полученных данных далеко от нормального, вам следует провести оценку полученных данных с помощью непараметрических критериев. Пользуясь описанием непараметричеких критериев, выберите подходящий, с вашей точки зрения. Результаты поместите в файл формата STATISTICA.

Для всего полученного массива проведите кластерный анализ с применением методов построения древовидных диаграмм и k-средних**.**

Приложение.

Таблица 1. Критические точки распределения Фишера - Снедекора при а = 0,05

k; k2	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,3	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,90	6,39	6,60	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4Д	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	309	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,8	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48

Таблица 2. Значение хи-квадрат в зависимости от различных уровней значимости **а** и числа степеней свободы **k**

k	a		k	a	
	0,05	0,01		0,05	0,01
1	3,84	6,64	16	26,3	32,0
2	5,99	9,21	17	27,6	33.4
3	7,82	11,3	18	28,9	34,8
4	9,49	13,2	19	30,1	36,2
5	11,0	15,0	20	31,4	37,6
6	12,5	16,8	21	32,7	38,9
7	14,0	18,4	22	33,9	40,3
8	15,5	20,1	23	35,2	41,6
9	16,9	21,7	24	36,4	43,0
10	18,3	23,2	25	37,7	44,3
11	19,6	24,7	26	38,9	45,6
12	21,0	26,2	27	40,1	47,0
13	22,4	27,7	28	41,3	48,3
14	23,7	29,1	29	42,6	49,6
15	25,0	30,6	30	43,8	50,9

Вопросы.

1. Цели планирования эксперимента.
2. Основные направления использования планирования экспериментов в фарм. технологии
3. Ненасыщенные планы. Полнофакторный эксперимент
4. Алгоритм выделения значимых факторов.
5. Понятие корреляционного анализа
6. Виды корреляции.
7. Коэффициент корреляции. Интерпретация значений коэффициента корреляции.
8. Хи-квадрат критерий Пирсона. Применение критерия для оценки значимости связи.
9. Дисперсионный анализ. Основная цель дисперсионного анализа.
10. F – критерий Фишера.
11. Назначение кластерного анализа.
12. Виды кластерного анализа, реализованные в программе СТАТИСТИКА.
13. Отличие результатов канонического анализа от результатов вычисления описательных статистик.

Литература.

1. Налимов В. В., Чернова Н. А., Статистические методы планирования экстремальных экспериментов, М., 1965;

2. Хикс Ч. Р., Основные принципы планирования эксперимента, пер. с англ., М., 1967;

3. Маркова Е. В., Лисенков А. Н., Планирование эксперимента в условиях неоднородностей, М., 1973;

4. Зедгинидзе И. Г., Планирование эксперимента для исследования многокомпонентных систем, М., 1976;

5. Адлер Ю. П., Маркова Е. Б., Грановский Ю.В., Планирование эксперимента при поиске оптимальных условий, 2 изд., М., 1976;

6. Рузинов Л. П., Слободчикова Р. И., Планирование эксперимента в химии и химической технологии, М., 1980;

7. Новик Ф. С., Арсов Я. Б., Оптимизация процессов технологии металлов методами планирования экспериментов, М.-София, 1980;

8. Ахназарова С. Л., Кафаров В. В., Методы оптимизации эксперимента в химической технологии, 2 изд., М., 1985.

9. Дёрфель К., Статистика в аналитической химии, пер.с нем., под ред. Адлера Ю.П., М., «Мир», 1994.

10. Богуцкая Г.А., Тетерятник О.В. Вища математика і математична статистика, Запоріжжя, 2007.

11. Халафян А.А. STATISTICA 6. Статистический анализ данных. М., Издательство БИНОМ, 2007