Сибирское отделение РАН

Институт почвоведения и агрохимии

ПРИКЛАДНАЯ СТАТИСТИКА НА КОМПЬЮТЕРЕ

2-е издание

Сорокин О.Д.

Новосибирск 2012 г

УДК 311:681.3.06

Сорокин О.Д. Прикладная статистика на компьютере. 2-е изд. Новосибирск, 2012 – 282 с.

Аннотация

В книге представлено доступное изложение некоторых базовых понятий прикладной статистики, информация о наиболее используемых классических методах обработки данных, справочный материал по программам пакета SNEDECOR ®. Книга рекомендуется молодым исследователям-биологам (студентам, стажерам, аспирантам), но, естественно, может быть полезна всем, кто использует пакет в своей работе.

Данные об авторе:

Сорокин Олег Дмитриевич, ведущий инженер Института почвоведения и агрохимии СО РАН, Новосибирск; моб. тел. 8-953-804-94-96.

E-mail: magnum46@ngs.ru

WEB: odssoft.narod.ru

Содержание

Предисловие

Автор, будучи по образованию биохимиком, в 80-х годах столкнулся с проблемой обработки больших массивов данных, получаемых при работе на аминокислотном анализаторе. Примерно в 1984-85 гг. в институт земледелия, где я работал, начали поступать первые отечественные микро-ЭВМ (Электроника-60, ДВК и другие). Начав с изучения различных языков программирования, и решив, таким образом, свои химико-аналитические задачи, я незаметно для себя перешел к обработке экспериментальных данных своих коллег различными статистическими методами.

Выяснилось, что российские биологи в общем плохо владеют математико-статистической наукой, по сравнению с зарубежными школами. В университетах США уже в 50-е годы прошлого столетия имелись штатные должности «Professor of Biostatistics», подобных нет в наших университетах и по сей день.

Используя формулы из различных кних по прикладной статистике и благодаря неоценимой поддержке д.б.н. А.И.Южакова (СибНИИЗХим), удалось запрограммировать и отладить значительную коллекцию алгоритмов в виде библиотеки процедур на языке Pascal. Как впоследствии оказалось, на базе этого языка была создана система программирования Delphi, с помощью которой программирование прикладных задач для операционной системы Windows весьма упрощается.

Первые программы для обработки данных, написанные для операционной среды MS DOS, до сих пор работают в научно-исследовательских организациях СО РАСХН. Различные версии пакета SNEDECOR [37], написанного для системы Windows, установлены в десятках лабораторий НИИ Сибирского региона, в Новосибирском агроуниверситете.

За 20 лет работы над пакетом выяснилось следующее. Исследователи, работающие в различных направлениях биологической науки (селекционеры, агрохимики, медики, инженеры-технологи), довольно часто предпочитают работать в среде пакета SNEDECOR, хотя на рынке есть много программ зарубежного происхождения, иногда даже русифицированных (STATGRAPHICS, SPSS, STATISTICA, NCSS и др.), а также отечественных пакетов (STADIA, ЭВРИСТА, МЕЗОЗАВР). Как оказалось, практически все пакеты ориентированы на специалистов с серьезной подготовкой в области прикладной статистики, интерфейс пакетов, как правило, достаточно сложный, и для освоения пакетов требуются определенные усилия, значительные затраты времени.

Пакет SNEDECOR целенаправленно создавался для биологов, следуя принципам:

– простота и стандартизация программного интерфейса;

– простота структур массивов данных (стандартный текст ASCII);

– работа с небольшими выборками (хотя возможны массивы до 200000 чисел);

– специализация и полная самостоятельность программ, входящих в пакет;

– возможность верификации программ тестовыми данными из руководств по прикладной статистике (массивы данных в комплекте пакета).

Цель данной книги – доступное изложение некоторых базовых понятий прикладной статистики, информация о наиболее используемых классических методах обработки данных, справочный материал по программам пакета SNEDECOR. Книга рекомендуется в первую очередь молодым исследователям (студентам, аспирантам), но, естественно, может быть полезна всем, кто использует пакет в своей работе.

5-я версия пакета отличается от 4-й следующим:

– объединены родственные программы дисперсионного, корреляционного, регрессионного анализа;

– исключены программы фенотипического анализа по Животовскому (они остаются в пакете BIOGEN [64]);

– добавлены программы METRIC, REPLICS, SERIES, PROBIT, GAUSS;

– во всех программах переработан интерфейс ввода-вывода массивов данных;

– сделано значительное количество добавлений и модификаций в различные программы.

Работа по совершенствованию пакета продолжается, планируется создание новых программ, с новыми алгоритмами обработки данных. Автор надеется, что принципы, заложенные в основу пакета, и дальше будут помогать биологам в их исследованиях.

В многолетней работе над пакетом мне помогали многие замечательные личности, которым я приношу свои искренние благодарности.

В первую очередь это Александр Иванович Южаков, которого я почитаю своим учителем, он был крупнейшим специалистом по применению математико-статистических методов в сельскохозяйственных исследованиях. Его алгоритмы

работают во многих программах дисперсионного анализа, а идея многомерного ранжирования объектов является пионерской.

В программах многомерного анализа реализованы алгоритмы Вадима Михайловича Ефимова (ИЦИГ СО РАН), консультации которого неоднократно выручали меня в джунглях математических формул. Его блестящие идеи применения главных компонент к анализу временных рядов реализованы в программе PROGNOZ.

Особую благодарность следует принести Юрию Константиновичу Галактионову за множество критических замечаний, позволивших серьёзно улучшить работу программ, за моральную поддержку на непростом пути разработчика программного обеспечения для биологов.

Одним из поворотных пунктов на стартовых этапах работы над пакетом явилась реализация алгоритма вычисления вероятности для эмпирического значения критерия Фишера-Снедекора – одного из самых важных критериев в прикладной статистике. Григорий Сигизмундович Гросбарт обнаружил ошибку (!!) в формуле, приведённой в известном “Справочнике по специальным функциям” Абрамовиц, Стиган, которую я ни за что не смог бы идентифицировать.

1. Введение в прикладную статистику

Человеческая история – это многие тысячелетия экспериментов. Методом проб и ошибок люди находили оптимальные условия существования, иногда ценой жизни некоторой части сообщества. Это позволило человечеству в целом выжить и достичь нынешней численности к началу XXI столетия.

Чтобы продолжать экспериментирование с наибольшей эффективностью, математическая наука выработала ряд правил, позволяющих как снизить затраты на проведение опытов, так и получать результаты с заданным уровнем достоверности, надежности.

Уровень достоверности (общепринятый термин – значимости) выбирается исследователем, исходя из специфики системы, с которой ведется работа. Например, в исследованиях взрывчатых веществ, испытаниях аэрокосмической техники ошибочное заключение допустимо, если можно так выразиться, в одном случае из тысячи (P=0.001), или даже из десяти тысяч (P=0.0001) подобных исследований. В экспериментировании с лекарственными препаратами, пищевыми продуктами и ряде подобных областей, связанных со здоровьем человека, допустима ошибка в одном случае из ста P=0,01 (или 1%). В инженерных исследованиях, биологических науках принят норматив уровня значимости P=0,05 (5%) – ошибочное заключение в одном опыте из 20. Эти нормативы узаконены государственными стандартами, являются естественными критериями качества научных исследований, принятыми во всем мире.

Исследователь может понизить уровень значимости по своей инициативе, например до 10%, только для предварительных, стартовых экспериментов – с целью минимизировать затраты или из иных соображений, но конечные выводы следует формулировать на базе результатов, полученных в экспериментах со стандартным уровнем значимости.

Общая логика экспериментальной работы такова:

а/ на основе собственного опыта, литературных данных, экспертных оценок и иных соображений исследователь делает предположение о некотором параметре изучаемой системы;

б/ для определения значений этого параметра вырабатывается план эксперимента (число факторов, вариантов, повторений и т.п.) – с учетом предполагаемой изменчивости (вариабельности, варьирования) системы в целом;

в/ фиксируются в рабочем журнале 0-гипотеза об изучаемом параметре системы, уровень значимости для эксперимента, план и прочие условия проведения опыта, метод обработки результатов;

г/ выполняется эксперимент в соответствии с планом;

д/ полученный в эксперименте числовой материал обрабатывается с помощью выбранного метода прикладной статистики, результат обработки данных – некоторое число – эмпирическое значения статистического критерия, однозначно определяемое свойствами массива экспериментальных данных;

е/ эмпирическое значение критерия сравнивается с табличным значением из специальных таблиц по математической статистике, вычисленных на базе фундаментальной теории распределения вероятностей случайных величин – на выбранном уровне значимости;

ж/ в результате сравнения эмпирического и табличного значений критерия 0-гипотеза либо считается подтвержденной, либо отвергается – в пользу Контр-гипотезы (на принятие которой обычно рассчитывает экспериментатор).

Как правило, обработка данных выполняется на персональном компьютере с помощью какого-либо пакета прикладной статистики. В этом случае эмпирическое значение критерия сопровождается одним или двумя целыми числами, называемыми степенями свободы (определяются из размеров массива данных) и значением вероятности, вычисленным программой для статистического критерия и этим степеням свободы. Тогда таблицы по математической статистике не используются, а вывод по принятию/отклонению 0-гипотезы выполняется сравнением вычисленной вероятности с нормативом уровня значимости. Например, в опыте получено значение критерия Фишера-Снедекора, вычислена вероятность:

$F_{эмп}=2,14$ ст. свободы: $n_1=2$, $n_2=31$ $P_{эмп}=0,1347$

Значение вероятности больше стандартного уровня значимости ($0,1347>0,05$), следовательно, 0-гипотеза подтверждена. Еще пример:

$F_{эмп}=6,52$ ст. свободы: $n_1=3$, $n_2=8$ $P_{эмп}=0,0153$

Значение вероятности меньше стандартного уровня значимости ($0,0153<0,05$), поэтому 0-гипотеза отклоняется, принимается Контр-гипотеза.

Вероятность, вычисляемая по эмпирическому значению критерия и степеням свободы, называется «вероятностью ошибки в случае отклонения 0-гипотезы», или вероятностью ошибки 1-го рода. Таким образом, экспериментатор с этой малой вероятностью отклоняет 0-гипотезу, зная, что все-таки в каком-то аналогичном эксперименте он обязательно сделает ошибку – отклонит 0-гипотезу,

когда на самом деле она верна. Из этого следует вывод, что экспериментатору нужно стремиться как можно более тщательно проводить опыт, чтобы вероятность ошибки получилась максимально близкой к нулю.

Вышеизложенное – это некая идеализированная схема экспериментирования, в реальных условиях план эксперимента зависит от многих причин, и в первую очередь определяется трудоемкостью получения той или иной информации о системе. Реальный план действий обычно формируется исследователем, исходя из его искусства, опыта, интуиции, в определенной степени может помочь теория планирования эксперимента [34, 62]. Как правило, одновременно исследуется некоторый комплекс параметров, метод обработки данных определяется после эксперимента, 0-гипотеза, а чаще целый блок 0- и Контр-гипотез, формулируются a posteriori.

1.1. Некоторые термины прикладной статистики

Факторы – внешние переменные, свойства, параметры, при изменении которых изучаемая система реагирует изменением своих внутренних параметров, свойств, характеристик. Часть факторов может быть задана экспериментатором в виде условий, градаций, вариантов – планом эксперимента, прочие факторы считаются неконтролируемыми, случайными, мешающими.

0-гипотеза – некоторое утверждение, высказанное для проверки какого-либо предположения об изучаемой системе, доказательства неочевидного факта, закономерности. Примеры типичных 0-гипотез:

– средние вариантов исследуемого фактора различаются только из-за действия множества случайных (неконтролируемых) факторов, фактор не влияет на изучаемую систему (все средние фактически равны между собой);

– отсутствует линейная связь между переменной (фактором, признаком) «X» и переменной «Y», значение коэффициента парной корреляции R_{xy} неравно нулю только вследствие действия множества случайных факторов;

– отсутствует функциональная линейная связь между независимой переменной «X» и зависимой переменной «Y», значение коэффициента регрессии ненулевое только вследствие действия множества случайных факторов.

Контр-гипотеза – утверждение, противоположное 0-гипотезе, примеры типичных Контр-гипотез:

– средние некоторых вариантов (как минимум одна пара средних) исследуемого фактора различаются достоверно, фактор влияет на изучаемую систему;

– имеется линейная связь между переменной (фактором, признаком) «X» и переменной «Y», значение коэффициента парной корреляции R_{xy} неравно нулю;

– имеется функциональная линейная связь между независимой переменной «X» и зависимой переменной «Y», значение коэффициента линейной регрессии достоверно отличается от нуля.

Выборка – конечное множество чисел, отражающее значение некоторого параметра системы в каком-то стационарном состоянии, или близком к стационарному. В биологических системах обычно действует множество случайных факторов, поэтому значение определяемого параметра постоянно дрейфует относительно своего истинного значения. Таким образом, чтобы оценить это истинное значение, исследователь должен несколько раз измерить каким-то способом значение параметра, вычислить среднее значение и определить разброс, служащий характеристикой вариабельности, изменчивости параметра, а также степени доверия к этому вычисленному среднему значению.

Дисперсия (вариация, изменчивость) – характеристика параметров любых систем, фундаментальное свойство природы. Чтобы отразить изменчивость параметров в стандартизованном виде, принято вычислять числовую характеристику выборки под названием «дисперсия» и величину σ^2:

$\mathbf{D = \frac{1}{n}\sum_{n}^{i=1}(x_i - \bar{x})^2}$; n = численность выборки; $\bar{x}$ = среднее выборки.

$\mathbf{\sigma^2 = D \cdot n/(n-1)}$ средний квадрат отклонений с поправкой на смещение.

Обычно для анализа изменчивости используют квадратный корень из среднего квадрата (в просторечии «сигма», среднеквадратическое отклонение):

$\mathbf{\sigma = \sqrt{D \cdot n/(n-1)}}$, D = дисперсия, n = численность выборки,

и коэффициент вариации, выражаемый в процентах:

$\mathbf{C_v = \sigma \times 100/\bar{x};}$.

Еще одной характеристикой изменчивости параметра является сумма абсолютных отклонений от среднего, а также среднее абсолютное отклонение:

$\mathbf{D_{abs} = \sum_{n}^{i=1} \left| (x_i - \bar{x}) \right|;} \qquad \mathbf{\sigma_{abs} = D_{abs}/n;}$

2. Методы прикладной статистики

Экспериментальный материал в виде массивов числовых данных подвергается математической обработке различными методами прикладной статистики. Множество методов можно сгруппировать примерно в семь разделов.

1. Дескриптивная (описательная) статистика: первичная обработка данных. Выборочный анализ (вычисление стандартных характеристик выборок), проверка на выбросы (артефакты, аномальные значения), анализ принадлежности выборочного распределения к тому или иному теоретическому распределению вероятностей, анализ нормальности (основной предпосылки применимости методов классической прикладной статистики).

2. Дисперсионный анализ: изучение действия одного или нескольких факторов на некоторый параметр системы по анализу средних в вариантах фактора (факторов). Существенность изменений средних доказывается критерием Фишера-Снедекора, известным как дисперсионное отношение.

3. Корреляционный анализ: определение связанности, взаимозависимости между переменными характеристиками систем, выражаемой обычно в виде коэффициентов парной корреляции по Пирсону от –1,0 до +1,0. Достоверность значения коэффициента корреляции доказывается критерием Стьюдента на заданном уровне значимости.

4. Регрессионный анализ: определение зависимостей между переменными в виде математических функций (регрессионных уравнений), доказательство достоверности соответствующих функциональных зависимостей по значимости коэффициентов регрессии с помощью критериев Стьюдента и Фишера-Снедекора.

5. Многомерный анализ – совокупность различных методов математико-статистического анализа данных, важнейшие из них:

метод главных компонент,
дискриминантный анализ,
факторный анализ,
кластерный анализ,
многомерный дисперсионный анализ,
многомерное ранжирование объектов.

В первую очередь, это методы преобразования данных с целью отразить неочевидные, скрытые свойства системы в виде новых переменных, объясняющих наиболее общие закономерности, факторов, реально действующих в изучаемой системе (главные компоненты, факторный анализ).

Эти же методы используются для решения различных задач выявления групп (популяций) однородных объектов, и, соответственно, классификации произвольных объектов по группам, это главная цель дискриминантного анализа, кластерного, а также многомерного ранжирования.

6. Анализ временных рядов. Статистический анализ в этой области в основном используется в акустике, радиофизике, теории информации, климатологии. В биологии программы анализа временных рядов используются для прогнозирования численности вредителей (насекомых, грызунов), заболеваемости населения, ожидаемого урожая – в зависимости от цикличности изменения глобальных погодных условий. Прогноз параметра строится с помощью либо уравнений авторегресии, либо уравнений, состоящих из суммы тригонометрических функций.

7. Специальные методы статистического анализа. К этой группе методов, естественно, попадает все, что не вошло в предыдущие группы: анализ частот, согласованность экспертов, анализ однородности, коэффициенты ассоциации и многое другое. Существуют направления прикладной статистики, например анализ экономической информации, анализ данных для селекционеров-генетиков, анализ систем массового обслуживания, психологов. Для этих областей развиваются весьма специализированные подходы обработки данных, иногда для обработки результатов сугубо конкретного эксперимента.

Большинство методов анализа данных, упомянутых выше, относится к так называемой параметрической статистике, которая основана на совокупности предположений о характере вероятностных распределений в исследуемой системе, в частности о нормальном законе распределения вероятностей (распределение Гаусса). Это означает, что распределение экспериментальных данных должно быть непрерывным (есть распределения дискретных величин – биномиальное, Пуассона), симметричным, унимодальным (одногорбым). Эмпирические распределения, однако, могут быть асимметричными, двугорбыми, островершинными и в виде различных комбинаций этих типов. Если распределение экспериментальных данных значительно отличается от нормального, мощность статистических критериев снижается, реальный уровень значимости меняется, в итоге можно сделать ошибку, например, отклонить 0-гипотезу, когда на самом деле она верна.

Чтобы быть уверенным в применимости стандартных методов статистики, следует перед обработкой данных проверить нормальность распределения с помощью специальной программы, или хотя бы визуально – анализом гистограммы

распределения. Если обнаружена асимметрия, можно произвести преобразование данных с помощью какой-либо математической функции.

При умеренных отклонениях от нормальности классические методы в общем работают, однако всегда следует помнить, что существует спектр методов непараметрической статистики, которые не требуют выполнения предположений о виде распределения, и обычно опираются на аппарат анализа рангов. Это означает, что исходный массив данных заменяется на массив рангов – некоторых чисел, отражающих номер позиции каждого значения в преобразованной выборке, построенной из исходной с упорядочиванием по возрастанию или убыванию. Непараметрические методы несколько грубее классических методов, однако бывают ситуации, когда об истинном типе распределения данных нет никакой информации, и применение непараметрических подходов остается единственной возможностью статистического анализа.

Многолетняя практика статистического анализа данных в биологии говорит о том, что наиболее часто используемыми методами являются дисперсионный, корреляционный и регрессионный методы анализа.

2.1. Дисперсионный анализ

Простейший вид дисперсионного анализа (ДА) – однофакторный, с фиксированными уровнями вариантов фактора. Пример массива данных для ДА из 4-х вариантов фактора «Доза азота» и 3-х повторений:

1-й вариант, контрольный	8,3	7,8	9,1
2-й вариант, 30 кг/га азота	12,1	11,6	10,9
3-й вариант, 60 кг/га азота	15,5	16,9	14,4
4-й вариант, 90 кг/га азота	17,9	18,2	22,5

Фиксированность уровней фактора («Fixed») означает, что вариация уровней воздействия на систему (0, $30^{\pm 0.5}$, $60^{\pm 0.5}$, $90^{\pm 0.5}$ кг/га) на 2-3 порядка меньше вариации изучаемого параметра ($\approx 15^{\pm 10}$ ц/га, урожай зерновых) системы.

Альтернативный вид уровней – случайный, «Random», в этом случае варьирование уровней вариантов фактора превосходит или сравнимо с варьированием изучаемого параметра. Например, фактор «Годы» (многолетний опыт, изменчивость погодных условий), или «Пункты выращивания» (сортоиспытание, изменчивость географических и климатических условий). От типа фактора зависит формула вычисления критерия Фишера-Снедекора и вид 0- и Контр-гипотез:

«Fixed» фактор: ***средние*** как минимум пары вариантов различаются достоверно,

«Random» фактор: ***дисперсии*** как минимум пары вариантов различается достоверно.

При проведении эксперимента обычно руководствуются принципом рандомизации, который заключается в выборе (размещении) экспериментальных единиц (делянок, вегетационных сосудов, растений, животных) случайным образом. Случайный выбор можно осуществлять различным способом, например слепым выбором номеров из списка, из таблиц случайных чисел, компьютерным датчиком случайных чисел. При этом возможны два типа рандомизации при формировании плана эксперимента с последующей обработкой ДА. Первый тип – полная рандомизация, второй – рандомизация вариантов в блоках повторений. Для плана 4х3 это будет примерно такая структура размещения делянок:

5	8	11
2	10	6
12	4	9
3	1	7

полная рандомизация

1 блок	2 блок	3 блок
4в	1в	3в
2в	4в	1в
1в	3в	2в
3в	2в	4в

рандомизация в блоках повторений

Формула для вычисления критерия Фишера-Снедекора корректируется для типа рандомизации в блоках вычленением дисперсии от блоков, это позволяет уточнить результаты эксперимента, исключив, например, возможный градиент плодородия почвы. Когда план эксперимента нельзя однозначно отнести к одному из этих типов, принято либо обрабатывать результаты по типу полной рандомизации, либо использовать специальные алгоритмы дисперсионного анализа.

Двухфакторные эксперименты с повторениями немногим сложнее, например, результаты опыта с 4 вариантами фактора А «Доза азота», 4 вариантами фактора В «Доза фосфора» и 3 повторениями:

1А контроль	**1В 0 кг Р**	**7,4**	**9,2**	**8,6**
	2В 40 кг Р	**8,7**	**8,3**	**9,1**
	3В 80 кг Р	**6,5**	**9,9**	**8,8**
	4В 120 кг Р	**8,3**	**10,8**	**9,8**
2А 30 кг N	**1В 0 кг Р**	**12,3**	**13,4**	**10,0**
	2В 40 кг Р	**11,7**	**12,8**	**14,3**
	3В 80 кг Р	**13,5**	**14,2**	**15,7**
	4В 120 кг Р	**13,1**	**14,6**	**14,9**
3А 60 кг N	**1В 0 кг Р**	**14,6**	**15,5**	**13,4**
	2В 40 кг Р	**15,3**	**16,7**	**15,2**
	3В 80 кг Р	**16,1**	**16,7**	**17,3**
	4В 120 кг Р	**15,5**	**16,9**	**18,4**
4А 90 кг N	**1В 0 кг Р**	**16,4**	**17,8**	**16,0**
	2В 40 кг Р	**17,7**	**18,5**	**20,7**
	3В 80 кг Р	**19,6**	**22,8**	**23,1**
	4В 120 кг Р	**22,9**	**26,2**	**24,5**

Анализ результатов на ПК программой D2MAXI:

Фактор	Степень влияния	Критерий Фишера-Снедекора F	ст.своб.	вероятность	НСР(5%)
A	0,8361	274,964	3, 30	0,00000*	0,854
B	0,0753	25,661	3, 30	0,00000*	0,854
AB	0,0520	5,262	9, 30	0,00025*	

Стандартная Ошибка = 0,5917 (4,04% от общего среднего)

Выявлено действие обоих факторов с высоким уровнем достоверности (звездочки у вероятностей ошибки 1 рода), фактор «Доза азота» значительно сильнее влияет на урожай, чем фактор «Доза фосфора». Достоверен эффект взаимодействия факторов (совместное применение азотных и фосфорных удобрений дает больший прирост, нежели простая сумма эффектов).

Если F-критерием **подтверждено** действие фактора, приступают к анализу различий средних с помощью критерия Стьюдента в форме НСР (Наименьшей Существенной Разницы) на стандартном уровне значимости:

	Фактор-"B"						
	1	2	3	4	Средние	Разница	Значима?
Фактор"A"							
1	8,400	8,700	8,400	9,633	8,783	Контроль	
2	11,90	12,93	14,47	14,20	13,38	4,592	Да!
3	14,50	15,73	16,70	16,93	15,97	7,183	Да!
4	16,73	18,97	21,83	24,53	20,52	11,73	Да!
Средние	12,88	14,08	15,35	16,33	14,660	5,877	Да!
Разница	Контр,	1,20	2,47	3,44	1,777		
Значима?		Да!	Да!	Да!	Да!		

Анализ многофакторных экспериментов (3-х, 4-х) аналогичен. Если доказано действие фактора “Random” типа, ***различие средних не анализируется.***

В случае явного различия вариации в вариантах Fixed типа (нарушение предпосылки однородности дисперсий) либо выполняют преобразование массива с помощью функций ArcSin, Ln или “квадратный корень”, либо используют непараметрические аналоги дисперсионного анализа – по Краскелу-Уоллесу, Фридману, Уилсону, Джонкхиеру.

В руководствах по прикладной статистике обычно приводятся математические модели различных видов дисперсионного анализа в виде суммы генерального среднего изучаемой системы, эффектов вариантов, возможных взаимодействий факторов и случайной составляющей. Например, модель 2-факторного анализа с повторениями:

$$y_{ijk} = \mu + a_i + b_j + ab_{ij} + e_{ijk};$$

μ – генеральное среднее изучаемой системы;

a_i – эффект варианта фактора A типа Fixed;

b_j – эффект варианта фактора B типа Fixed;

ab_{ij} – эффект взаимодействия факторов;

e_{ijk} – ошибка от случайных факторов, распределенная по N(0, σ).

Модель подразумевает оценку эффектов a_i, b_j и ab_{ij}, дающих скорректированное значение изучаемого параметра системы в некоторой точке плана. Подчеркнем, что эти эффекты – вклады в значение параметра ***относительно генерального среднего***, тогда как исследователя обычно интересуют различия факторных средних ***относительно контрольного варианта*** или произвольных пар вариантов.

2.2. Корреляционный анализ

Коэффициент парной корреляции Пирсона между двумя переменными вычисляется по формуле:

$$R_{xy} = \left(\sum_{n}^{i=1} (x_i - \bar{x}) \times (y_i - \bar{y}) \right) \Big/ \sqrt{\sum_{n}^{i=1} (x_i - \bar{x})^2 \times \sum_{n}^{i=1} (y_i - \bar{y})^2};$$

n – число пар X–Y, $\bar{x}$ и $\bar{y}$ – средние.

Достоверность коэффициента корреляции определяется либо по пороговому значению, взятому из статистических таблиц, либо критерием Стьюдента. Помимо корреляции по Пирсону, возможна оценка взаимосвязанности переменных коэффициентами корреляции рангов по Спирмену и Кендаллу. Эти корреляции не требуют выполнения предпосылки нормальности распределения.

При анализе связанности ансамбля переменных в виде массива «Признаки-объекты», взятых из единой системы, вычисляют матрицу парных корреляций, симметричную относительно диагонали, на которой стоят 1,0. Вследствие множественности связей между переменными значения корреляций могут быть завышенными, поэтому, чтобы определить истинные значения парных корреляций, прибегают к вычислению матриц ***частных*** корреляций. При этом с помощью математических преобразований исключается возможная связанность с третьими признаками, значения корреляций предстают в “очищенном” виде. Степень исключения может быть различной – элиминацией одного, двух и более признаков до максимальной степени очищения.

Довольно часто требуется оценить степень связанности какого-то признака с целой группой переменных, ее можно выразить коэффициентом ***множественной*** корреляции. Достоверность коэффициентов частной и множественной корреляций подтверждается критерием Фишера-Снедекора.

2.3. Регрессионный анализ

Простейшим регрессионным уравнением выражается линейная зависимость между двумя переменными в виде полинома 1-й степени:

$\mathbf{Y(x) = B_0 + B_1 \times X}$ X – независимая, Y – зависимая переменная,

B_0 – свободный член, B_1 – коэффициент регрессии.

Коэффициенты регрессии вычисляются методом наименьших квадратов, который является одним из самых мощных и наиболее используемых методов прикладной статистики.

Для описания нелинейных зависимостей чаще всего используют полиномы более высоких степеней, например, квадратичной параболой, кубической:

$\mathbf{Y(x) = B_0 + B_1 \times X + B_2 \times X^2}$; $\mathbf{Y(x) = B_0 + B_1 \times X + B_2 \times X^2 + B_3 \times X^3}$.

Достоверность уравнения регрессии в целом определяется критерием Фишера-Снедекора, затем с помощью критерия Стьюдента выясняется значимость коэффициентов регрессии в соответствующих членах уравнения. Если значимость какого-то коэффициента не доказана, этот член может быть исключен из уравнения, и вычисления повторены. Исключение незначащих членов обычно увеличивает достоверность уравнения регрессии.

Уравнением ***множественной*** регрессии описывается функциональная зависимость между ансамблем независимых переменных и зависимой переменной, обычно называемой «откликом». Как правило, построение такого уравнения – результат целенаправленного многошагового эксперимента на поиск оптимума какого-то параметра системы.

В простых случаях уравнение множественной регрессии – сумма линейных членов:

$$\mathbf{Y = B_0 + B_1 \times X_1 + B_2 \times X_2 + B_3 \times X_3 + ... + B_n \times X_n}$$

например, Y – урожай, X_1 – доза азота, X_2 – доза фосфора, X_3 – доза калия, и так далее. Добавлением дополнительных членов решается проблема нелинейности для некоторых независимых переменных, а также их взаимодействия:

$$Y = B_0 + B_1 \times X_1 + B_2 \times X_2 + B_3 \times X_3 + B_4 \times X_1^2 + B_5 \times X_2^2 + B_6 \times X_1 \times X_2 .$$

Коэффициенты регрессии вычисляются методом наименьших квадратов, определяется их достоверность критерием Стьюдента, и по вероятности ошибки принимается решение – оставить соответствующий член (переменную) в уравнении, или нет.

В случае большого числа членов приходится много раз выполнять как исключение, так и включение тех или иных переменных, в этом случае регрессия называется шаговой, и тогда обычно используют процедуры автоматического определения наилучшего уравнения регрессии.

Помимо регрессионного анализа, основанного на методе наименьших квадратов, существуют итерационные методы поиска наилучших уравнений множественной регрессии на базе метода наименьших модулей отклонений. Часто они дают более качественное решение, определяемое большим значением критерия Фишера-Снедекора.

3. Пакет программ SNEDECOR V5

3.1. Общие сведения о пакете

Пакет SNEDECOR предназначен для обработки экспериментальных данных различного происхождения (биология, медицина, исследования в области сельского хозяйства, инженерный эксперимент) методами прикладной статистики [37]. Пакет функционирует на ПК под управлением операционных систем Windows-95/98/ME/2000/XP/Vista/Seven. Существует версия пакета (V3), функционирующая в среде MS DOS [38].

Пакет назван в честь американского статистика Джорджа У. Снедекора, много сделавшего для развития прикладной статистики в биологических и сельскохозяйственных исследованиях [20].

Пакет состоит из набора программ, с помощью которых экспериментальные данные можно обрабатывать различными методами прикладной статистики. Все программы представляют собой полностью самостоятельные модули, поэтому возможно формирование специализированных комплексов, исходя из специфики экспериментальной работы. Помимо стандартных статистических методов, имеются реализации авторских разработок:

– прогноз временных рядов по В.М.Ефимову [35];

– многомерное ранжирование объектов по А.И.Южакову [63];

– классификация объектов по Байесу (А.Афифи, С.Эйзен, [4]).

В ряде программ пакета применяется принцип анализа данных с помощью ансамбля статистических критериев. Например, существует проблема проверки нормальности небольших выборок (10-40 дат), так как классический метод проверки с помощью критерия Hi^2 работает с большими выборками (>50 дат). В этой ситуации предлагается такой подход: использовать одновременно 5-6 различных критериев проверки нормальности (Hi^2 плюс Колмогоров-Смирнов, Уилк-Шапиро, Мизес-Смирнов, Гири, асимметрия/эксцесс). Решение проблемы нормальности становится простым, если все критерии единодушны в своих выводах относительно какой-то выборки.

Аналогичный подход применяется при множественном сравнении средних, двухвыборочном анализе, для теста однородности дисперсий.

Принципиальное отличие от аналогичных программных продуктов заключается в том, что пакет ориентирован не на профессиональных пользователей (статистиков, математиков), а на исследователей-биологов, имеющих, как правило, минимальные знания в области прикладной статистики, и занимающихся обработкой данных спорадически. Необходимы: знание элементарной статистики, клавиатуры, некоторое понимание функционирования системы Windows – как с помощью манипулятора "мышь" запустить программу, как включить принтер и т.п.

Пакет SNEDECOR V5 состоит из 47 программ. Общий объем, занимаемый пакетом – около 50 мегабайт. Во время первичной установки на ПК выполняется кодирование всех программ пакета (за исключением программы #START) с целью защиты от несанкционированного копирования.

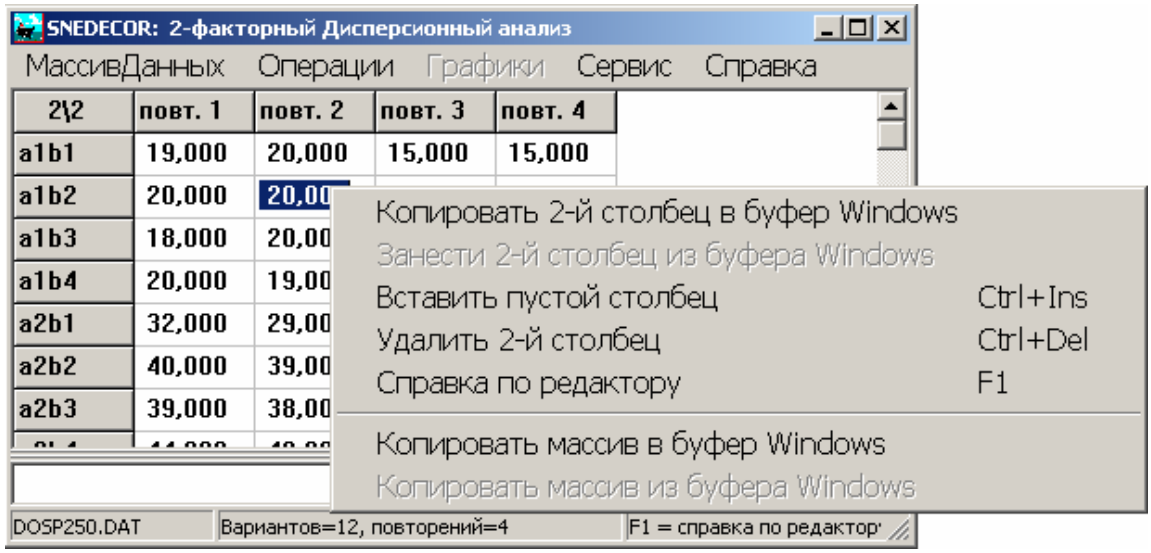
Каждая программа может функционировать независимо от всех прочих программ, во всех модулях есть собственный аппарат ввода/редактирования данных. Начинать работу с пакетом удобно вызовом головной программы – #START, из которой можно загрузить любую программу пакета, а также получить различную справочную информацию.

Все программы пакета имеют практически одинаковый интерфейс – стандартную организацию диалога с пользователем. Как только освоена работа с первой программой, это означает, что изучение прочих программ не составит труда.

После запуска выбранной программы появляется стандартная форма с Меню в стиле Windows; любую операцию можно выбрать щелчком мыши. Ре-

комендации по последовательности операций можно получить нажатием клавиши <F1>.

В большистве ситуаций дополнительные возможности можно обнаружить кликом **правой** клавиши мышки на различных областях программ – в поле Табличного Редактора, в поле текста результатов, на графике. Появляется при этом Меню дополнительных операций, помогающих в дальнейшей работе:



3.2. Установка пакета на компьютер

Первичная установка пакета на ПК выполняется стандартным способом с помощью процедуры инсталляции с компакт-диска. Предполагается, что на компакт-диске хранятся все файлы пакета, и в случае необходимости всегда можно восстановить работоспособность любых программ, входящих в комплект поставки. Если все же возникает необходимость реконфигурации пакета (например, после переустановки операционной системы), рекомендуем действовать следующим образом.

1. Выбрать том винчестера (C:, D: и т.п.) для установки пакета, проверить наличие на этом томе свободной области размером не менее 50 мегабайт.

2. Если пакет устанавливается из дистрибутива, запустить с компакт-диска установку программой SETUP.exe, или же скопировать с архивного компакт-диска копию пакета – каталог (папку) SNEDECOR со всем ее содержимым.

3. Скопировать в подкаталог \SNEDECOR\DATA массивы данных, если они имеются.

4. Создать ярлык на рабочем столе Windows для программы #START.exe с именем "Snedecor" (кликом правой клавиши мышки в поле рабочего стола Windows или с помощью проводника).

Вносить какие-либо коррективы в реестр Windows нет необходимости. Настройка на поддиректории описана в разделе, касающемся файла CONFIG.sdc. Далее следует запустить программу #START и проверить работоспособность программ на тестовых или реальных данных.

3.3. Табличный редактор

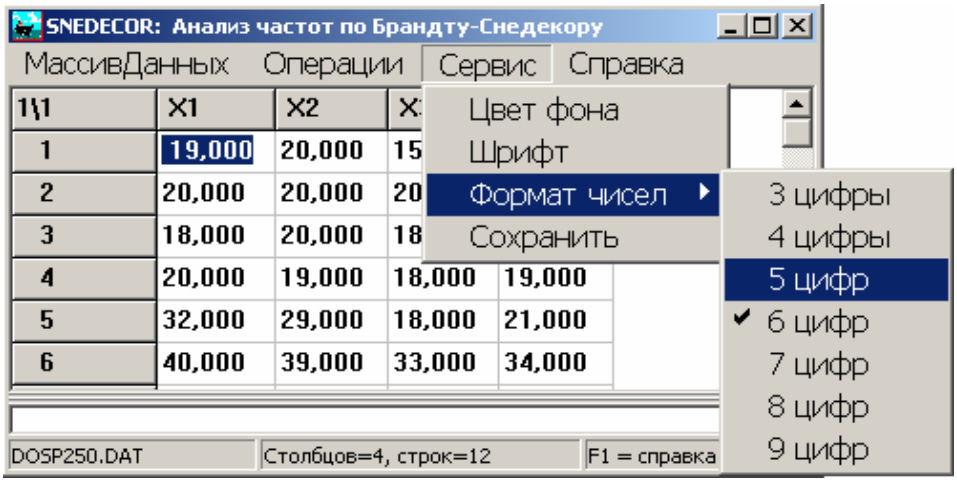
Во всех программах пакета имеется встроенный Табличный Редактор данных, приемы работы с которым во многом схожи с электронными таблицами. Подробное описание методов работы есть в справочном тексте, сопровождающем каждую программу. Для большинства программ редактор имеет модификации, отражающие специфику данных, обрабатываемых конкретной программой, но общие принципы функционирования следующие.

Размеры редактируемого массива задаются пользователем при обращении к пункту Меню "Ввод нового массива данных с клавиатуры" (например, число вариантов/повторностей, признаков/объектов и т.п.). При этом формируется двумерная структура – прямоугольная матрица – все значения которой равны "-999" (признак отсутствия значения); на дисплее это отражается пустыми ячейками. Максимально возможный размер массива данных – 200000 элементов – для программ, работающих с данными типа "признаки/объекты" (MCOM, VARS, MATRIX и т.п.), для некоторых – 16000 и менее. Во всех программах максимальные размеры массива (число признаков, вариантов, групп, повторений и т.д.) уточняются в справочном тексте и контролируются, как правило, при вводе с клавиатуры и загрузке массива из файла.

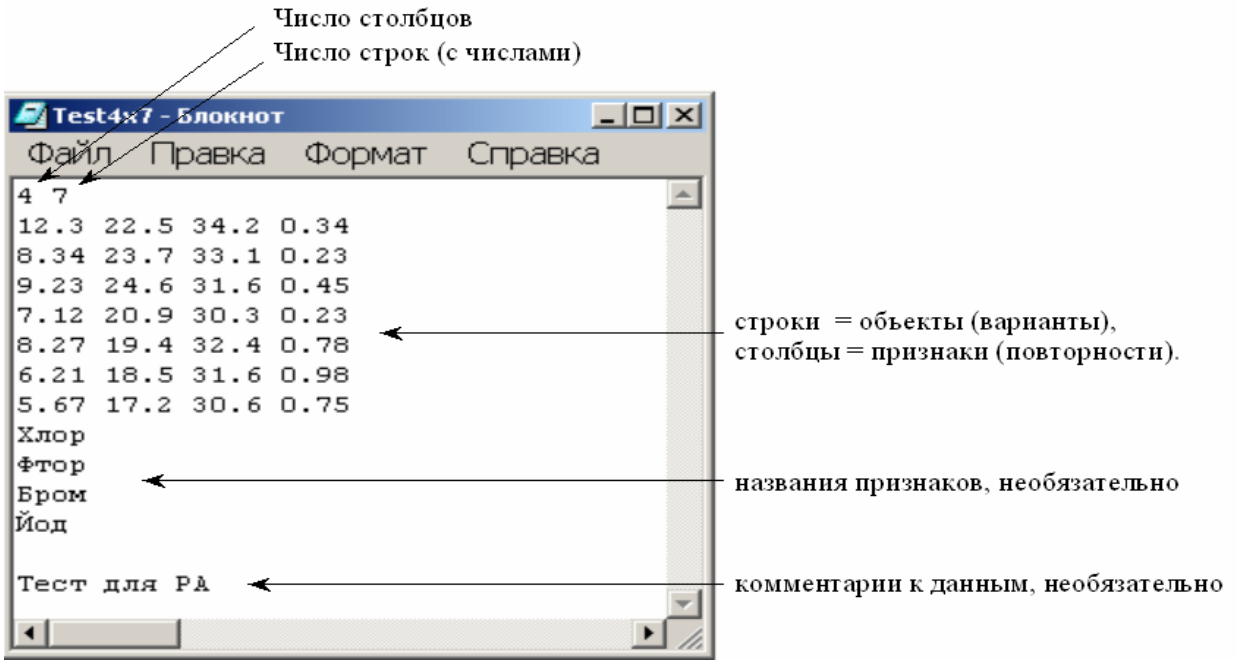
Редактор большинства программ позволяет исключать строки/столбцы массива с помощью комбинации клавиш <Ctrl/Y> или <Ctrl/Delete> – в том месте массива, где находится маркер, а также вставлять строки/столбцы с помощью клавиш <Insert>, <Ctrl/Insert>, соответственно.

Текущая редактируемая ячейка массива выбирается мышкой или клавишами "стрелки" в поле Редактора, вторым кликом мышки или нажатием клавиши <F2> инициируется режим редактирования ячейки. Завершить ввод числа можно клавишами <Пробел>, <Enter>, <Стрелки вверх/вниз>. Формат чисел в ячей-

ках редактора можно менять с помощью файла CONFIG.sdc, а также непосредственно выбром нужного значения из Меню:



Файл данных, записанный на магнитном носителе, представляет из себя обычный ASCII-текст, который может быть обработан с помощью любого редактора текстов типа Блокнота Windows. Его можно посмотреть на экране с помощью клавиши <F3> (лучше Alt/F3) в Коммандере Нортона или в FARe, а с помощью клавиши <F4> подредактировать, например, первую строку файла – это иногда требуется для массивов, которые обрабатываются программами дисперсионного, регрессионного и дискриминантного анализа. Пример массива из 4-х признаков и 7-и объектов в текстовом файле (Блокнот Windows):



Перед вычислениями ячейки массива тестируются на значение «-999,0» (может быть изменено по выбору пользователя); если обнаруживается хотя бы одно число с таким значением, вычисления блокируются с соответствующим сообщением. Это, однако, не препятствует записи частично введенного массива в

виде файла данных на винчестер или дискету, в последующем можно вызвать его и продолжить ввод числового материала.

1-я строка файлов данных иногда содержит более чем два числа (4, 5 и более). Дополнительные числа используются некоторыми программами для структурирования информации в основном массиве. Например, программы многофакторного дисперсионного анализа используют эти значения для определения числа вариантов в каждом факторе. Программы множественного регрессионного анализа определяют по дополнительным числам количество повторностей зависимой переменной, и т.д. Особенности, касающиеся 1-й строки файлов данных, описаны в справочных текстах конкретных программ. Однако следует заметить, что файлы с любой 1-й строкой могут считываться программами общего назначения – IODATA, VARS, NORMAL – и использоваться для обработки, игнорируя дополнительную информацию из этой строки.

Отсюда следует, что в ряде случаев массивы данных, введенные с клавиатуры в рамках какой-либо программы и записанные затем на диск, могут быть корректно использованы только этой же программой, и только как чисто двумерный массив – другими программами. Массивы, подготовленные для многофакторного дисперсионного анализа, могут быть использованы любыми 1-факторными программами, и наоборот.

В пакете SNEDECOR имеются две специализированные программы для ввода и редактирования массивов данных – IODATA и COMPAR. Первая программа – универсальный редактор двумерных массивов, предоставляющая следующие возможности для операций с данными:

– транспонирование массива: X[M x N] -> X[N x M];

– математические операции с признакими;

– сортировка данных в любом признаке;

– генерация псевдослучайных значений;

– объединение массивов данных добавлением "справа" или "снизу";

– "восстановление" отсутствующих значений;

– анализ экстремальных значений и т.д.

Программа COMPAR предназначена для ввода и редактирования массивов, обрабатываемых в дальнейшем программами многофакторного дисперсионного анализа. Помимо стандартных возможностей по вводу данных, в программе возможны различные арифметические преобразования данных с целью стабили-

зации дисперсии в вариантах (однородность дисперсий – одна из предпосылок классического дисперсионного анализа).

3.4. Логика анализа экспериментальных данных

Приступая к анализу своих экспериментальных данных, пользователь должен представлять, что он хочет извлечь из своих данных. Возможна следующая рабочая классификация основных методов статистической обработки с точки зрения экспериментатора:

1. Анализ выборки: оценить стандартные статистические характеристики массива данных (среднее, сигму, коэффициент вариации, асимметрию, эксцесс и т.п.), проверить нормальность распределения дат в выборке, исключить выбросы, определить доверительные интервалы для среднего.

2. Дисперсионный анализ: выяснить, влияет ли исследуемый фактор (или несколько факторов) на исследуемую систему – по достоверному изменению средних в вариантах опыта.

3. Корреляционный анализ: выяснить, есть ли линейная связь между двумя (тремя, и более) переменными (факторами, признаками) – по достоверности различных коэффициентов корреляции.

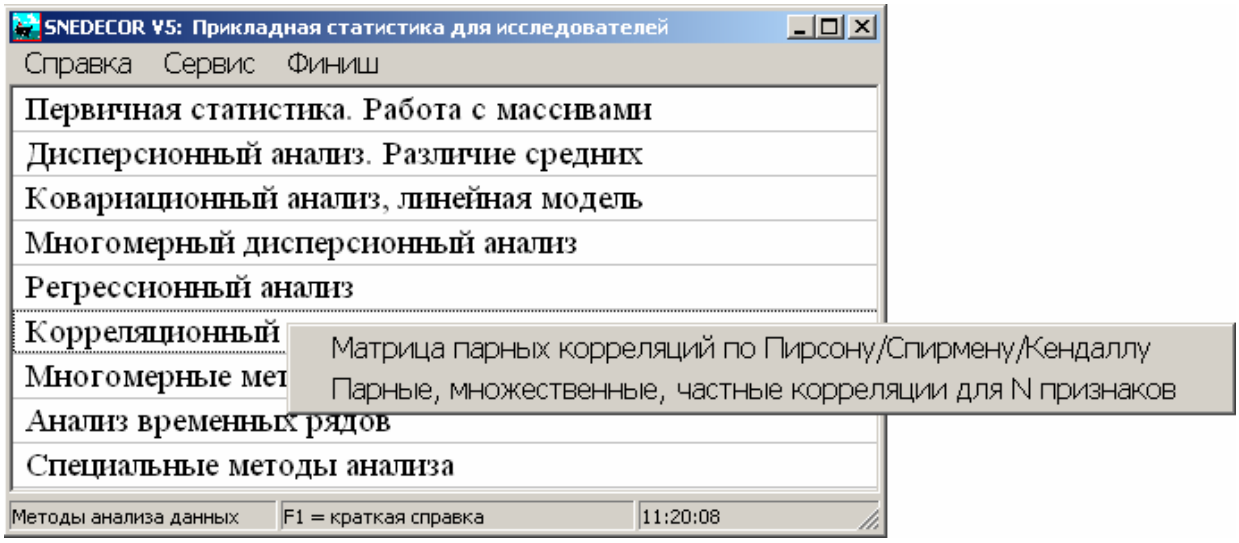
4. Регрессионный анализ: выяснить, есть ли какая-либо связь (не обязательно линейная) между зависимой переменной (Y) и одной или несколькими независимыми переменными (X_1, X_2 и т.д.), и если есть, то выразить эту связь в виде уравнения с достоверными коэффициентами регрессии.

5. Многомерный анализ: статистические методы для профессионалов; рекомендуем работать в этой области после изучения специальной литературы и в контакте со специалистами.

6. Анализ временных рядов. В биологии используется в основном для прогноза какого-либо показателя, процесса.

7. Специальные виды статистического анализа: самые разнообразные методы обработки данных, часто весьма специфичны для конкретных направлений и исследований.

Исходя из этой классификации, пользователь должен выбрать для себя группу методов. Пакет SNEDECOR структурирован в соответствии с вышеперечисленными разделами, которые предлагает программа #START:



Если после просмотра программ в выбранной группе методов отыскивается нужная, ее следует запустить и просмотреть справочную информацию о программе. Если вы убедились, что данные могут быть обработаны этой программой, следует потренироваться на тестовых задачах. Их можно взять из каких-либо руководств по биометрии и статистике, или же ввести с клавиатуры небольшой массив случайных чисел, похожих на ваши экспериментальные данные. Для большинства программ есть тестовые массивы из комплекта поставки пакета SNEDECOR (например, D2MAXI.dat для программы D2MAXI). После запуска программы можно получить справку по работе с Табличным редактором с помощью клавиши <F1>.

Затем следует ввести с клавиатуры массив ваших данных, и после этого настоятельно рекомендуем записать введенный массив на магнитный диск. После завершения обработки данных этот массив можно записать на компакт-диск или флэшку для архивных целей, а также использовать для обработки другой программой.

Если вы затрудняетесь с выбором программы для обработки ваших данных, следует обратиться за консультацией к опытному пользователю или специалисту.

Большинство методов статистического анализа данных базируется на стандартном подходе, состоящем из 3 этапов:

1) фиксируется "0-гипотеза" (в норме должна быть сформулирована перед экспериментом), или набор 0-гипотез; исходя из специфики задачи исследования также фиксируется "Уровень значимости" – 1, 5 или 10 процентов (строгий эксперимент, обычный или предварительный; в ряде руководств по прикладной статистике фигурируют значения 99, 95 и 90%);

2) для полученного массива экспериментальных данных вычисляется некоторый статистический критерий (T-, F-, Hi^2- и т.п.) с соответствующими степенями свободы (или несколько – по числу 0-гипотез);

на основании сравнения эмпирического значения критерия с табличным значением 0-гипотеза либо принимается, либо отклоняется – позволяя принять тем самым "Контр-гипотезу" (зачастую и являющуюся целью эксперимента).

В программах пакета для эмпирического значения статистического критерия вычисляется "Вероятность ошибки в случае отклонения 0-гипотезы", это позволят анализировать результаты без обращения к статистическим таблицам. Как трактовать это значение? Если вероятность

P<=0,01 0-гипотеза отклоняется на строгом уровне экспериментирования – 1%;

0,01<P<=0,05 0-гипотеза отклоняется только на обычном (в биологических исследованиях) уровне – 5%;

0,05<P<=0,10 0-гипотеза отклоняется только для предварительных экспериментов – уровень значимости 10%;

P>0,10 0-гипотеза принимается.

Интервал вероятностей 0,05 .. 0,10 обычно трактуется как неопределенный, и вывод относительно результата эксперимента зависит от позиции исследователя.

3.5. Особенности работы с графикой

Большинство программ пакета позволяют получить графическое изображение какой-либо характеристики обрабатываемых данных в виде диаграммы, графика, гистограммы и т.п. Параметры графического режима подбирались таким образом, чтобы изображение было удовлетворительным на старых дисплеях типа VGA (SVGA) с разрешением 800x600 строк, и не требовало дополнительной подстройки. Желательна 16-битовая или 32-битовая цветовая гамма (65536 или 6 млн. цветовых оттенков), но вполне возможна работа и в режимах 256 и 16 цветов.

На современных мониторах, как правило, можно получить изображение с более высоким разрешением – 1024x768, 1280x1024 строк и более. Выбор разрешения и цветового режима зависит, естественно, от характеристик компьютера и настроек операционной системы.

В некоторых программах многомерного анализа положение объектов на плоскостной проекции отображается по умолчанию овалом, при нажатии клавиши "N" позиции объектов отображаются их номерами.

Распечатку графика можно получить на принтере выбором соответствующего пункта меню. Аналогичным образом можно записать изображение в графический файл с расширением .bmp (стандартный формат Windows "bitmap") или посредством буфера Windows передать в текст MS Word.

3.6. Файл конфигурации пакета CONFIG.sdc

Конфигурационный файл CONFIG.sdc используется программами пакета для корректировки некоторых параметров, связанных с сервисными функциями. Пользователь имеет возможность менять значения этих параметров по своему выбору. Программа в момент запуска считывает содержимое этого файла и модифицирует значения параметров в соответствии с установленными значениями. При выходе из программы файл CONFIG.sdc перезаписывается, с заменой информации в четвертой строке.

Это обычный текстовый файл из 11 строк, который можно редактировать непосредственно в среде программы #START или в любом текстовом редакторе. Содержимое стандартного файла:

```
Config=On
Sound=On
Ck=6
File=C:\WORK\DATA\AB.dat
Long=80
DataDir=C:\SNEDECOR\DATA
TextDir=C:\SNEDECOR\RESULTS
NoValue=-999,0
Color=16754343
Font=Fidexsys
Decimal=,
```

1-я строка: конфигурацию корректировать,
2-я -"- звуковой сигнал включить,
3-я -"- позиций в ячейке редактора = 6,
4-я -"- файл данных для загрузки = AB.dat,
5-я -"- максим.длина строки печати = 80 символов,
6-я -"- каталог для массивов данных,
7-я -"- каталог для текстовых файлов/результатов,
8-я -"- значение "отсутствующей" даты,
9-я -"- цвет фона редактора, окна результатов
10-я -"- шрифт редактора, окна результатов.
11-я -"- разделитель целой и дробной частей чисел

1-я строка служит для указания программе, корректировать или нет все прочие параметры; Config=On означает подтверждение корректировки, Config=Off – игнорировать значения параметров, указанные в строках 2..11, устанавливаются поэтому программой на значения по умолчанию.

2-я строка указывает, включить или нет режим звуковой поддержки клавиатуры, если он ранее не был включен. В любой момент пользователь может включить/выключить этот режим с помощью клавиши ScrollLock.

3-я строка устанавливает формат вывода значений в Табличном Редакторе массивов данных. Это значение можно менять в пределах от 3 до 8. В некоторых программах нельзя переустановить размер ячейки.

4-я строка содержит имя файла данных, который загружался в программу в последний раз перед выходом из нее. При запуске следующей программы этот файл данных будет автоматически считываться для возможной обработки. Если файл с таким именем отсутствует, ничего загружаться не будет. Также не будет загружаться файл, если информация в первой строке файла данных не будет соответствовать специфике программы. Указание имени файла в 4-ой строке можно убрать, оставив в ней "File=".

5-я строка ограничивает максимальное количество символов в строке при выводе результатов на принтер или в текстовый файл; например, 72 или 80 символов – при выводе на "узкие" принтеры, 132 символа – стандартное значение для "широких" принтеров, однако при печати больших массивов мелкими шрифтами логичнее устанавливать значение Long=160, 200 или даже 230 символов в строке. Максимально допустимое значение Long=254 символа. В некоторых случаях значение параметра Long игнорируется.

6-я строка указывает программе поддиректорию "по умолчанию" при обращении к винчестеру для чтения/записи массивов данных.

7-я строка указывает программе поддиректорию, в которую можно записывать текстовые файлы с результатами анализа данных,

8-я строка определяет, какое число будет использоваться в качестве "отсутствующего" значения, по умолчанию это –999,0, в частности, в некоторых тестовых массивах встречается именно это значение.

9-я строка устанавливает цвет фона в Табличном Редакторе, в окне представления результатов анализа в стандарте RGB (Red, Green, Blue); в режимах видеокарты 16 и 256 цветов цвет фона меняться не будет;

10-я строка определяет шрифт в Табличном Редакторе, который может быть выбран пользователем из набора шрифтов Windows; в некоторых случаях система осуществляет замену шрифтов по собственной инициативе;

11-я строка определяет разделитель целой и дробной части числа в Табличном Редакторе, результатах анализа. В России принято в качестве разделителя использовать запятую, в других странах – точку.

Следует соблюдать последовательность строк, как указано в стандартном файле CONFIG.sdc, не должно быть пробелов между именем параметра и значе-

нием. Файл CONFIG.sdc может отсутствовать в директории с пакетом SNEDECOR, все параметры в этом случае устанавливаются на стандартные.

3.7. Работа с принтером

Для распечатки результатов счета может быть использован любой принтер – матричный, струйный или лазерный. В случае, когда результаты обработки данных имеют таблицы шириной более 80 символов, и при печати на стандартном принтере формата A4 получается искажение таких таблиц, следует записать результаты в текстовый файл (с расширением .rtf или .txt), потом распечатать его с помощью программ WordPad (Write.exe) или MS Word, имеющие возможность настраивать размеры страницы перед печатью и подбирать размер шрифта, и затем распечатать обычным способом. Также можно распечатать справочные тексты, сопровождающие каждую программу. Графики также можно выводить на печать на любом принтере, при этом изображение будет автоматически масштабироваться по ширине формата A4.

3.8. Программа #START.exe

Программа #START предназначена для запуска программ прикладной статистики, входящих в состав пакета SNEDECOR, и для вывода различной справочной информации по пакету. Любой справочный раздел может быть выведен на принтер или в текстовый файл. Рекомендуем именно эту программу указывать в качестве ярлыка при создании метки пакета на Рабочем столе Windows.

Меню "Сервис" позволяет посмотреть и подредактировать в случае необходимости файл настроек пакета CONFIG.sdc.

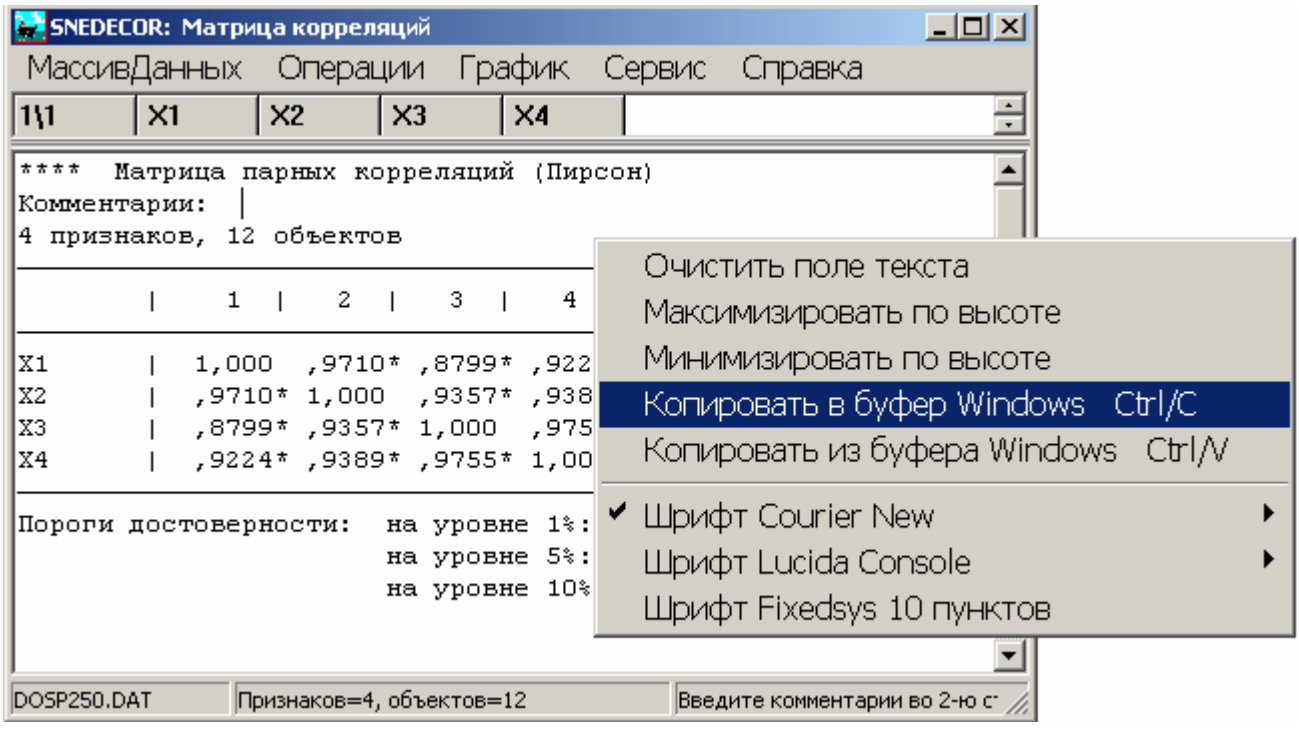
Выбор раздела статистического анализа производится с помощью мышки – кликом по строке с названием раздела. При запуске выбранной программы анализа окно #START минимизируется, располагаясь в нижней части экрана; после завершения работы с программой статистики окно восстанавливается автоматически до нормального размера. Одновременно можно работать с несколькими программами пакета, располагая их в разных местах экрана, корректируя размеры окон мышкой.

3.9. Работа с буфером Windows

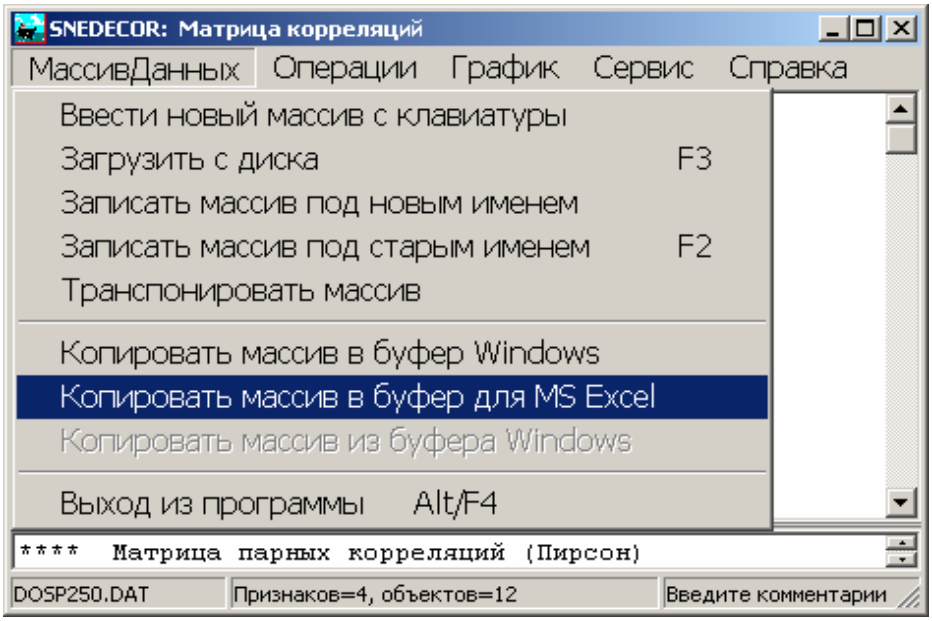
Во всех программах пакета имеется возможность использовать буфер в оперативной памяти (Clipboard) для переноса информации в другие программы – в редакторы текстов или электронные таблицы. Например, текст с результатами обработки данных можно передать в редактор MS Word, дополнить его графиком, занести какие-либо комментарии и затем распечатать.

Для копирования текста в буфер следует кликнуть ПРАВОЙ клавишей мышки в поле текста – появится «выскакивающее» меню операций, среди них следует выбрать «Копировать в буфер Windows».

Если при вставке текста из буфера в поле редактора MS Word произойдет некоторое искажение табличных результатов, надо выделить мышкой блок текста и выбрать для него один из стандартных моноширинных шрифтов – Courier, Courier New или Lucida Console, имеющихся в стандартной комплектации системы Windows. Размер шрифта также иногда требует подбора – в зависимости от типа документа, целей пользователя.



Массивы данных также можно переносить с помощью буфера, например, в среду программы MS Excel, для этого нужно выбрать пункт основного меню программ «Копировать массив в буфер для MS Excel»:



Кликом ПРАВОЙ клавиши мышки в поле Табличного Редактора можно вызвать выскакивающее меню, в котором тоже есть операции работы с буфером. При копировании в буфер переносится весь массив, причем в операции «Копировать в буфер для MS Excel» будет сделана замена разделителя целой и дробной части чисел на «запятую» и символа «разделитель чисел», при операции «Копировать в буфер Windows» такая замена не выполняется.

Аналогичным образом можно передавать массивы данных из MS Excel: выделить мышкой прямоугольный блок данных (только числа!), скопировать его в буфер, затем передать массив данных в Табличный Редактор программы пакета.

Для копирования в буфер графика следует выбрать соответствующий пункт основного меню формы графики, или использовать комбинацию клавиш <Ctrl/C>. Перед копированием желательно подобрать размеры графика, «хватая» края формы с помощью мышки.

Если пункт «Копировать из буфера Windows» неактивен, это значит, что в буфере – нечисловое содержимое. При передаче массива данных из буфера в некоторые программы пакета следует учесть, что дополнительная информация о структуре данных (число факторов в массивах многофакторного анализа, число групп в массивах данных для дискриминантного анализа и т.п.), комментарии – будет утеряна – программе передается только прямоугольная матрица.

3.10. Массивы тестовых данных

Помимо программ в состав пакета SNEDECOR входят более 130 массивов тестовых данных. Эти данные взяты из реальных экспериментов, различных ру-

ководств по прикладной статистике, оригинальных публикаций, или же сгенерированы методом Монте-Карло.

Например, массив FISH_IRI.dat – данные Р.Фишера по 3-м видам ирисов – классический тест для дискриминантного анализа.

Массив SSP6x30.dat – тестовый массив из пакета SSP [26].

Имеются тестовые массивы из руководств Снедекора, Доспехова, Лакина, Рао, Аренса – Лейтера, Хикса и других. Как правило, это можно увидеть из имени файла, иногда дополненного номером страницы, более полная информация обычно есть в комментариях (последняя строка файла).

В многих случаях имя файла данных несет информацию о его структуре, например, SSP6x30 = 6 столбцов (признаков), 30 строк (объектов). Опытный пользователь сразу же сообразит, что это массив – скорее всего для корреляционного или регрессионного анализа. Небольшие массивы (5x3, 4x8 и т.п.) – как правило, для дисперсионного анализа, большие массивы (10x100, 20x300) – для многомерного анализа.

Тестовыми массивами следует пользоваться для изучения методов прикладного статистического анализа, верификации алгоритмов, использованных в программах пакета. Находятся эти массивы обычно в каталоге \SNEDECOR\Tests.

4. Первичная статистика. Работа с массивами

Программы для ввода и редактирования массивов данных, стандартной вариационной статистики. Программа VARS – наиболее используемая из всех программ пакета, распространяется бесплатно. Программа INTER – дань прошлому прикладной статистики, когда из-за отсутствия компьютеров для облегчения обработки данных в таблицы заносились не исходные числа, а частоты попадания значений в заданные интервалы.

4.1. IODATA: Ввод, редактирование массивов данных

Программа IODATA предназначена для ввода и редактирования массивов экспериментальных данных "признаки-объекты", "выборки-наблюдения" в форме "столбцы/строки". Для введенного массива данных можно выполнить стандартную статистическую обработку:

1/ вариационная статистика для выборок в столбцах массива;

2/ вычисление матрицы парных корреляций (также для столбцов);

3/ проверка на выбросы (артефакты) в выборках (4 критерия).

Результаты анализа могут быть отредактированы непосредственно в среде программы при выводе результатов на дисплей (заголовок, комментарии, удаление ненужной информации и т.п.).

Максимальный размер массива данных – не более 200000 элементов, размер выборки (столбца из этого массива) – не более 8000 элементов.

Данные в виде двумерного массива "признаки-объекты" могут быть введены:

– с клавиатуры непосредственно в среде программы;

– переданы через буфер Windows из программы MS Excel, для этого надо выделить прямоугольный блок ячеек, состоящий только из чисел, скопировать его в буфер в среде MS Excel (Ctrl/C) и передать программе IODATA комбинацией клавиш Ctrl/V. Через буфер Windows может быть передан массив данных, подготовленный в каком-либо текстовом редакторе (WordPad, MS Word), разделителями чисел могут быть пробелы и символы табуляции.

– из файла в стандарте пакета SNEDECOR, подготовленного заранее с помощью какой-либо программы пакета или любого редактора текстов типа Блокнота Windows. Этот массив должен представлять из себя обычный текст ASCII, который можно посмотреть на экране с помощью клавиши <F3> в оболочках типа FAR, Norton, DiskCommander.

Пример формирования массива из 6-и выборок и 11-и наблюдений:

```
6 11                                    <- начало файла
12,3 22,5 34,2 0,34 1,45 3,11
8,34 23,7 33,1 0,23 1,66 3,65           массив данных:
9,23 24,6 31,6 0,45 1,89 2,79             строки = объекты,
7,12 20,9 30,3 0,23 1,73 3,09             столбцы = признаки
8,27 19,4 32,4 0,78 1,77 3,35
6,21 18,5 31,6 0,98 1,85 -999           в 6-м признаке – 2 пропуска
5,67 17,2 30,6 0,75 1,57 3,51
8,55 16,3 33,9 0,77 1,33 3,40
7,23 17,6 32,1 0,82 1,21 -999
6,47 15,5 31,7 0,79 1,42 3,74
5,18 16,0 31,2 0,78 1,63 3,71
Калий                                   <- названия признаков
Натрий                                    (необязательно)
Кальций
Железо
Кобальт
Молибден

Данные 1997 г.                          <- необязательный комментарий
```

В качестве примера формирования массива можно посмотреть файл SSP6x30.dat (6 выборок, 30 повторений).

Помимо текстовых массивов данных, программа может считывать числовую информацию из файлов dBase-III/IV, имеющих простую структуру: все поля dBase-файла должны быть только числовые (”N”), формат полей может быть одинаковым (например, все 6:2, 7:3, 8:4, 9:5 и т.п.) и различным, число полей определяет число признаков, число записей в dBase-файле = число объектов, названия полей dBase-таблицы используются в качестве названий признаков.

При желании можно изменить размер ячейки Табличного Редактора для удобства работы с конкретными данными, а также указать формат вывода значений массива на экран (и, соответственно, на принтер или при записи в файл данных на винчестере):

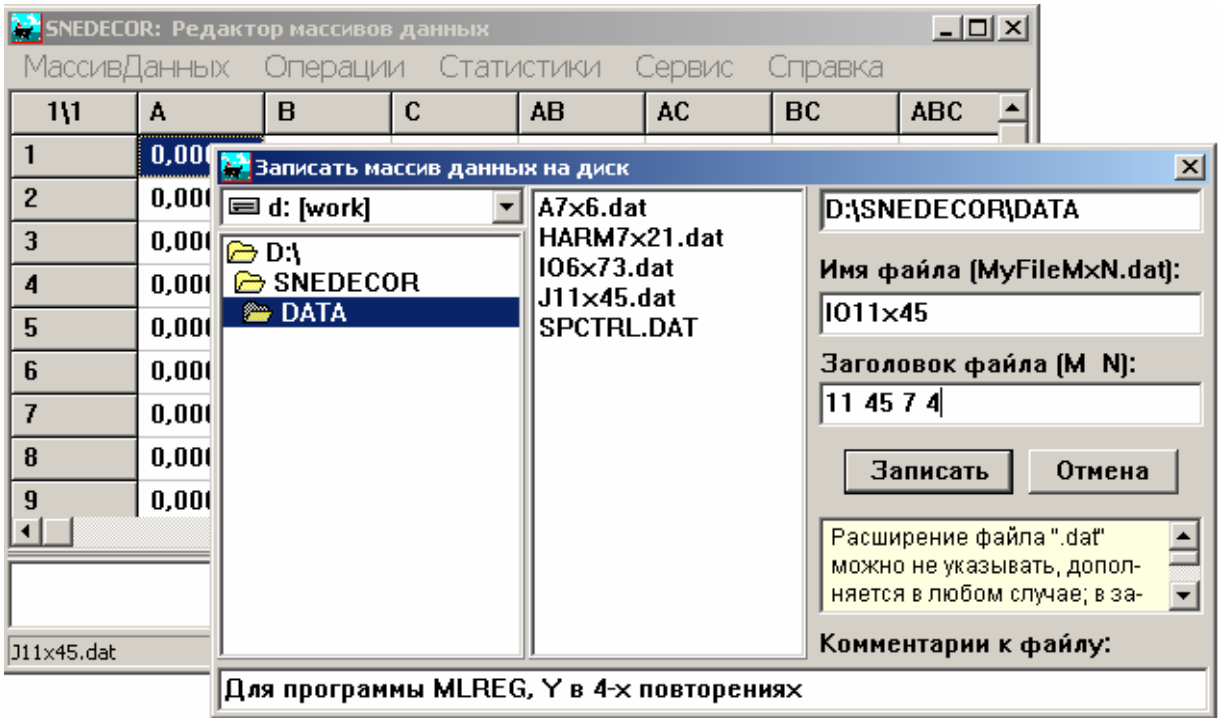
8 позиций – диапазон от –9999999 до 99999999
7 позиций – от –999999 до 9999999
6 позиций – от –99999 до 999999 <= по умолчанию
5 позиций – от –9999 до 99999
4 позиции – от –999 до 9999
3 позиции – от –99 до 999

Если число выходит за текущий диапазон представления данных, программа отражает это звездочками "****", но само значение, естественно, сохраняется в массиве.

Программа предоставляет возможность производить различные операции с данными:

– транспонирование массива: M x N -> N x M;
– математические операции с признаками;
– сортировка объектов по возрастанию/убыванию любого признака;
– генерация псевдослучайных значений;
– объединение массивов данных; и т.д.

Перед записью массива данных на диск программа запрашивает имя файла данных, строку комментариев до 60 символов (необязательно) и предлагает скорректировать первую строку файла:



Если массив готовится для программ дискриминантного анализа, следует указать разбиение объектов на группы; например, M=15 признаков, N=40 объектов, 3 группы 10+15+15, тогда заголовок файла должен выглядеть следующим образом:

15 40 10 15 15

Заголовок файла данных для программ множественного линейного регрессионного анализа также следует скорректировать; например, M=10 переменных, N=50 вариантов, 9 независимых переменных, зависимая переменная в одном повторении (в последнем столбце):

10 50 9 1

Заголовок массивов для одномерного дисперсионного анализа формируется автоматически в программе COMPAR или в соответствующих программах; правила формирования 1-й строки для других программ подробно описываются в справочных текстах.

4.1.1. Вариационная статистика

Для быстрого ознакомления с наиболее востребованными статистическими характеристиками выборок программа предоставляет операцию "Вариационный анализ".

Вычисляется 10 стандартных статистик (среднее, ошибка, сигма и т.д.) для всех признаков массива данных. Для углубленного вариационного анализа выбо-

рок (доверительные интервалы, робастные оценки параметров, графический анализ) следует использовать программу VARS.

4.1.2. Матрица корреляций Пирсона

Для оценки линейной связанности признаков массива данных программа выполняет операцию "Матрица парных корреляций". Это бывает нужно перед проведением множественного регрессионного анализа, дискриминантного анализа, кластерного.

Вычисляется матрица парных корреляций Пирсона, значения, достоверные на уровне значимости 5%, помечаются звездочками. Для углубленного корреляционного анализа массива данных (непараметрические корреляции, частные корреляции) следует использовать программы MATRIX и MCOR.

4.1.3. Проблема пропусков в массиве данных

Специальная операция "восстановления" отсутствующих по какой-либо причине дат может быть выполнена одним из 4-х способов:

– по принципу замены пропущенного числа на значение в этом признаке ближайшего в многомерном пространстве объекта (наиболее корректный способ, однако возможно некоторое изменение дисперсии признака);

– заменой пропущенных значений на среднее в признаке; в этом случае неизбежно уменьшение дисперсии данных, поэтому этот метод следует использовать только при малом числе выпавших дат (1-3 в признаке);

– заменой пропусков на среднее в строке; это допустимо только в том случае, когда все признаки измерены в единой шкале и с примерно одинаковой дисперсией;

– заменой на некоторые числа, генерируемые датчиком случайных чисел, распределенных по нормальному закону со средним и сигмой, оцененным по имеющимся в признаке значениям; этот метод может быть неприемлем для признаков с большой дисперсией из-за появления отрицательных значений.

Для этого отсутствующие значения нужно ввести как –999, при этом они представлены пустыми ячейками в Табличном Редакторе.

4.1.4. Дублирование признаков со сдвигом

Меню операций содержит пункт "Дублирование со сдвигом" – "вправо/вниз" и "вправо/вверх; выполняются следующим образом, например, исходный массив 2 признака, 9 объектов, сдвиг вниз 2 раза:

```
11 24      11 24                                  новый массив: 6 признаков
13 21      13 21 11 24                                        7 объектов
15 22      15 22 13 21 11 24       15 22 13 21 11 24
17 25 -->  17 25 15 22 13 21  --> 17 25 15 22 13 21
19 27      19 27 17 25 15 22       19 27 17 25 15 22
14 28      14 28 19 27 17 25       14 28 19 27 17 25
16 26      16 26 14 28 19 27       16 26 14 28 19 27
12 23      12 23 16 26 14 28       12 23 16 26 14 28
10 29      10 29 12 23 16 26       10 29 12 23 16 26
                 10 29 12 23
                       10 29
```

Эта операция может быть применена для данных типа временных рядов – для исследования автокорреляций, авторегрессий.

4.1.5. Анализ выбросов

Специальная операция "Анализ экстремальных значений" позволяет проверить каждый признак (столбец) или каждую строку (например, варианты массива для дисперсионного анализа) на наличие выбросов – аномально больших или аномально малых значений по нескольким критериям. Следует заметить, что в экспериментальной работе практически всегда неизвестен тип вероятностного распределения данных в выборках, а если и предполагается из некоторых соображений нормальность распределения, то остаются неизвестными параметры распределения – генеральное среднее и дисперсия. Поэтому в программе используются критерии, учитывающие такие свойства экспериментальных данных.

Рекомендации из текста на сайте http://www.himikatus.ru:

Существует несколько правил, соблюдение которых необходимо для получения корректных результатов при выявлении грубых промахов:

– недопустим произвольный отброс подозрительно выделяющихся значений;

– применение тестов для выявления грубых промахов применимы к некоррелированным данным, и их нельзя применять к взаимозависимым результатам измерений;

– к каждой выборке может применяться любой подходящий тест, но только один и только однократно. Выявленные выбросы не учитываются при

статистических расчетах, но не должны забываться: вообще их показывают на гистограммах, сообщают при выдаче результатов измерений и т.д. Кроме того, каждый выброс должен анализироваться с точки зрения причин его появления.

Критерии анализа выбросов:

1. Стандартный критерий “Наибольшего по абсолютной величине нормированного выборочного отклонения” из [15], стр. 60.

$$\zeta_{(\mu,\sigma)} = \mathbf{Max}|x_i - \mu|/\sigma;\ \mu = \frac{1}{n}\sum_{i=1}^{n} x_i;\ \sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \mu)^2};$$

Может быть применён для выборок любого размера, предполагается, что данные в выборках распределены по нормальному закону, но параметры распределений неизвестны и оцениваются по значениям выборок.

2. Критерий Диксона – отношения разностей к размаху в отсортированной по возрастанию значений выборке, где X_1 и X_n – экстремумы, возможные выбросы:

$$\zeta_{min} = (x_2 - x_1)/(x_n - x_1);\ \zeta_{max} = (x_n - x_{n-1})/(x_n - x_1);$$

Метод может быть применён для малых выборок (N<31). Также предполагается нормальность распределения данных в выборках. Описан в [15], стр. 61. В случае равенства пар значений (X_2=X_1 или X_n=X_{n-1}) возникают сомнения в применимости критерия.

3. Критерий на основе MAD (Median Absolute Deviation). Med(x_i) – медиана выборки:

$$\zeta_{extr} = |x_{extr} - \mathbf{Med}(x_i)|/\mathbf{MAD};\ \mathbf{MAD} = \mathbf{Med}(|x_i - \mathbf{Med}(x_i)|);$$

Рекомендуется в случае отсутствия информации о возможном типе вероятностного распределения, и, предположительно, при явном отсутствии нормальности распределения. Может быть применён для анализа выборок любого размера. Описание метода можно найти на сайте www.himikatus.ru. Пороговое значение критерия равно 5 для выборок с произвольным распределением вероятностей и не зависит от размера выборок (?!).

4. Критерий Граббса, описан в ГОСТ Р ИСО 5725-2–2002 [76], немного отличается по формулам от 1-го критерия, табличные значения иные. Может быть применён в случае небольших выборок (N<41), распределённых по нормальному закону.

При обнаружении выброса, можно либо поставить "–999" на этом месте и далее "восстановить" это значение, либо исключить объект из массива в Табличном Редакторе.

Если в выборках обнаружены выбросы (очень большие или очень маленькие величины), то вполне оправданным будет исключение этих значений – выбором соответствующего пункта меню (эта операция допустима только для массивов, не содержащих пропуски). При этом удаляются по два значения из каждой выборки (процедура усечения данных). Следует помнить, что после исключения экстремумов операция транспонирования, вообще говоря, лишена смысла, так как практически всегда экстремумы располагаются в различных местах выборок.

4.1.6. Метод Монте-Карло

В программе имеется возможность сгенерировать выборки, значения в которых будут распределены по нормальному закону (распределение Гаусса), или равномерно распределены. Это бывает нужно для тестирования различных статистических процедур, моделирования систем, процессов.

С помощью датчика псевдослучайных значений программа формирует последовательность чисел, обладающих заданными свойствами (естественно, в статистическом смысле). Качество распределения сильно зависит от размера выборки, Не следует ждать аккуратного "колокола" на выборках 20-40 значений, более или менее приличные выборки должны быть размером 100-200-500 значений.

Для углубленного моделирования массивов данных с различными вероятностными законами распределения следует использовать программу GAUSS.

4.2. NORMAL: Тест нормальности распределения данных

Программа NORMAL предназначена для проверки одной из важнейших предпосылок классического статистического анализа – нормальности распределения данных в выборках – массивах экспериментальных данных. Возможно графическое представление выборочного распределения данных в виде гистограммы или интегральной функции.

Данные могут быть представлены двумерным массивом "Признаки-объекты", из которого программе можно передать для анализа любой признак, а также массивом "Варианты-повторности" (подготовленным для дисперсионного

анализа). В этом случае можно передать для анализа весь массив данных, который используется как единая выборка размером M*N. Следует заметить, что в ряде программ дисперсионного анализа имеется возможность формирования "массива остатков" для более корректной проверки предпосылки нормальности. Ограничения на размер массива данных: не более 100000 элементов.

В программе могут быть использованы несколько методов анализа нормальности распределения данных:

– критерии асимметрии и эксцесса распределений, табулированы для больших выборок (>30 дат);

– классический метод – критерий Hi^2 (по Пирсону или Кульбаку) для больших выборок (не менее 30 дат);

– критерий Колмогорова-Смирнова для выборок любого размера; мощность этого метода (способность отклонить гипотезу о нормальности в случае действительно ненормально распределенных данных) выше, чем в случае применения критерия Hi^2. В качестве пороговых используются табличные значения из [1], стр. 376. – для случая оценки среднего и ср.кв. отклонения по выборке;

– критерий Уилка-Шапиро для малых выборок (3 – 50 дат), являющийся одним из наиболее мощных критериев проверки нормальности [12];

– метод, рекомендованный ГОСТ 8.207-76 – D-критерий "нормированного абсолютного отклонения"; описан в [15], стр. 56, 258. Этот метод позволяет проверять выборки от 10 дат и более;

– метод, также рекомендованный ГОСТ 8.207-76 – критерий Мизеса-Крамера-Смирнова (Омега-квадрат); описан в [1], стр. 372-376. Позволяет проверять нормальность выборок любого размера; для лучшего согласования статистики Омега-квадрат с предельным распределением при малых выборках критерий модифицирован по формуле Стефенса. Пороговые значения критерия выбраны в соответствии с рекомендациями Мартынова – для учета оценки среднего и ср.кв. отклонения по выборке.

Для критерия Hi^2 можно задать уточнения:

1/ метод вычисления критерия Hi^2: по Пирсону (классический) или по Кульбаку (теоретически более обоснован);

2/ число классов для разбиения выборки: пользователь должен задать число классов, руководствуясь своим опытом и следующими ориентирами:

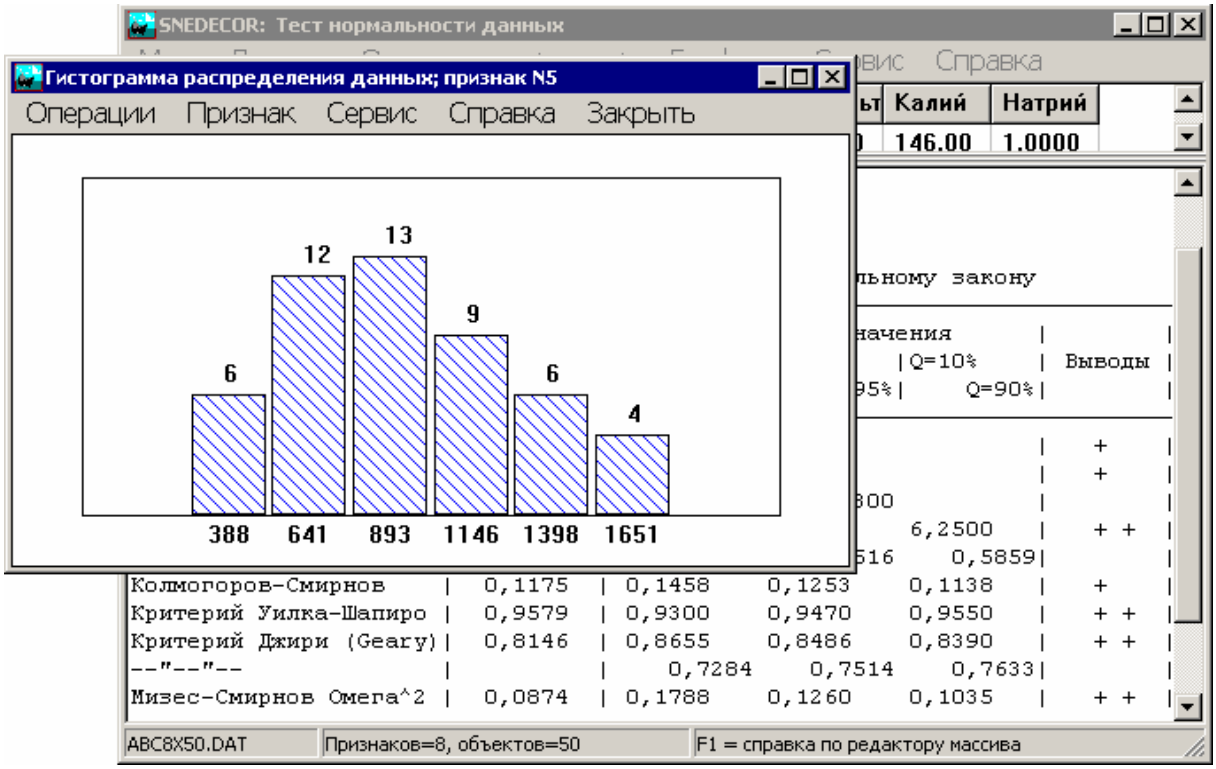
N = 15 – 20 3..5 классов;

N = 21 – 40 4..6 классов;

N = 41 – 60 5..7 классов;
N = 61 – 90 6..8 классов;
N = 91 – 120 6..9 классов;
N > 120 дат 7..10 классов;
N > 200 дат 7..11 классов;

3/ способ разбиения выборки на классы: традиционный способ – разбиение на классы с одинаковыми интервалами; но в некоторых случаях невозможно разбить выборку в классы с одинаковыми интервалами из-за появления пустых классов, тогда необходимо использовать разбиение в интервалы различной длины, но с примерно равным числом дат.

В ряде специфических ситуаций программа может автоматически сократить число классов, при этом на экране появляется соответствующее сообщение.



Результаты анализа трактуются следующим образом: если какой-либо критерий приемлем для данной выборки, программа выставляет в графе "Выводы" символы "+" и "-":

++ 0-гипотеза подтверждена данным критерием,

\+ 0-гипотеза не отвергается на уровне значимости 5%,

\- 0-гипотеза отвергается на уровне значимости 5%,

-- 0-гипотеза отвергается на уровне значимости 1%.

Для некоторых критериев имеются табличные значения для уровней значимости 90, 95 и 99%; 0-гипотеза должна быть отвергнута [1], если эмпирическое

значение критерия говорит о **подозрительно хорошем** согласии с нормальным распределением (попадает в диапазон значений для уровней 95..99%, или даже за уровень 99%).

Для визуального анализа типа вероятностного распределения данных имеется возможность сравнить выборочное распределение с нормальным (наличие асимметрии, двугорбость и т.п.).

Число классов группировки данных можно менять посредством Меню, как и просматривать гистограммы других признаков (столбцов массива данных). Для упрощения анализа выбросов крайние столбцы, содержащие только одну дату, помечаются номером объекта с этим значением.

Результаты анализа могут быть отредактированы непосредственно в среде программы при выводе результатов на дисплей (заголовок, комментарии, удаление ненужной информации и т.п.).

Для более корректных оценок среднего и среднеквадратичного отклонения выборки разработан алгоритм поиска минимума суммы квадратов отклонений эмпирического распределения от функции вероятности нормального распределения. В большинстве случаев удается получить робастные оценки генерального среднего и среднеквадратического отклонения, особенно эффективные для малых выборок, засоренных артефактами.

4.2.1. Нестандартные операции с графикой

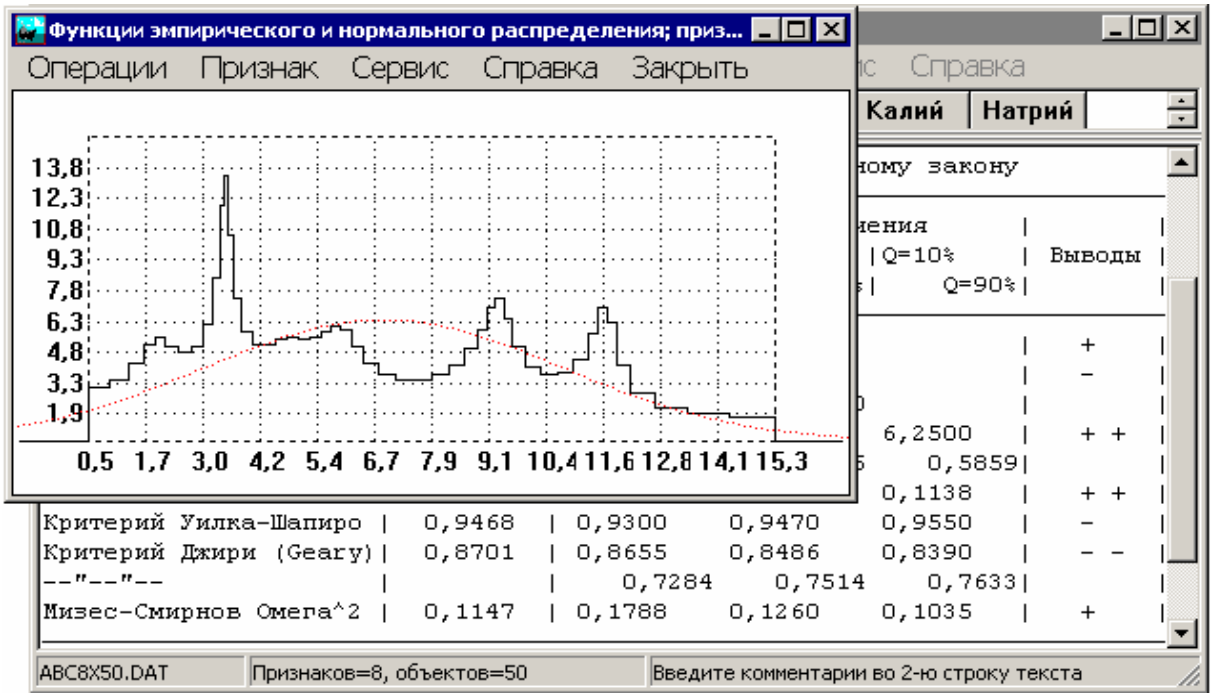
К числу нестандартных возможностей программы относятся:

– манипуляция видом гистограммы – изменением числа классов разбиения выборки – с целью получить наиболее презентабельное отображение типа распределения (Pop-Up меню правым кликом мышки); изменение числа классов влияет на значение критерия Hi^2 (вычисляется для выборок не менее 30 дат);

– график "функция плотности эмпирического распределения" – оригинальная разработка (в литературе не описан) позволяет визуально оценить плотность эмпирического распределения, не зависит от субъективного фактора (числа классов в случае гистограммы); для более эффективного представления характера распределения следует несколько раз сделать сглаживание функции распределения выбором способа сглаживания из Меню дополнительных операций (клик правой клавишей мышки):

– анализ нормальности последовательности выборок может весьма эффективно сопровождаться графиком гистограммы или функции распределения; "пе-

релистывание" мышкой текстового результата совместно с графиком автоматически меняет содержимое на обеих формах; в сомнительных случаях этот прием особенно полезен.



Для гистограммы возможно получение таблицы интервалов с оценкой центральных точек на дисплее или с выводом на принтер, в текстовый файл.

4.3. VARS: Вариационные статистики выборок

Программа VARS предназначена для обработки экспериментальных данных, представленных двумерным массивом "выборки-наблюдения", "признаки-объекты", "варианты-повторения", для которого необходимо получить оценки стандартных вариационных статистик – либо по строкам, либо по столбцам массива.

1. Стандартная обработка – все выборки в единой таблице, для каждой выборки рассчитываются среднее, среднеквадратическое отклонение, ошибка среднего, Min и Max, коэффициент вариации, мода, медиана, асимметрия, эксцесс.

2. Углубленный вариационный анализ выборки – вычисляются различные виды среднего, ср.-кв. отклонение, коэффициент вариации, мода, медиана, асимметрия, эксцесс – и все оценки сопровождаются стандартными ошибками и доверительными интервалами, вычисленными на основе классической теории. Дополнительно выполняется Bootstrap-процедура для оценки вариабельности этих статистик.

3. Вариационный анализ выборки – вычисляются среднее, дисперсия, коэффициент вариации и медиана, их стандартные ошибки, **непараметрические доверительные интервалы**, вычисленные по формулам А.И.Орлова (см. далее).

4. Анализ независимости и стационарности рядов наблюдений – для каждой выборки выполняется **тест серий** на основе медианного **τ**-критерия.

5. Робастное оценивание вариационных статистик выборок методом **выборочных квантилей** по Тьюки.

Максимальный размер массива данных – не более 200000 элементов, размер выборки (строки или столбца из этого массива) – не более 8000 элементов.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла.

Выборки в массиве могут быть неравной величины; программа определяет размер выборки, отбрасывая пропуски (-999) и даты "для восстановления" (-1,0). В качестве примера формирования массива можно посмотреть файл SSP6x30.dat (6 выборок, 30 наблюдений).

После вывода результатов (1. Стандартная обработка) на дисплей возможна манипуляция точностью представления чисел – либо увеличив число цифр (максимально 8 цифр у среднего), либо уменьшив число цифр (минимально 3 цифры). При выводе на принтер или в текстовый файл используется выбранное пользователем представление результатов (массив ABC8x50.dat):

Название выборки	Числ дат	Среднее арифм.	Максимум Минимум		Станд. ошибка	Ср.кв. отклон	Коэффиц. вариации	Медиана Мода		Эксцесс Асимметрия	
Уран	50	6,8580	0,500	15,30	0,544	3,848	56,11%	6,900	6,200	0,280	-0,976
Плутоний	50	15,616	3,600	36,00	1,039	7,348	47,06%	15,60	14,95	0,792	0,354
Радий	50	101,51	6,500	443,3	12,90	91,19	89,84%	102,5	76,70	1,830	3,596
Стронций	50	135,42	21,00	293,0	9,641	68,17	50,34%	132,0	125,0	0,324	-0,793
Цезий	50	930,80	286,0	1753	51,06	361,0	38,79%	915,0	871,0	0,448	-0,526
Кобальт	50	1943,6	694,0	3550	100,4	709,8	36,52%	1944	1923	0,315	-0,583
Калий	50	367,86	78,00	878,0	28,23	199,6	54,26%	362,0	343,0	0,558	-0,240
Натрий	50	4,7800	1,000	8,000	0,308	2,179	45,58%	5,000	5,000	-0,301	-0,960

В случае, когда аналогичные статистические показатели необходимо получить для данных по строкам массива, нужно предварительно сделать транспонирование массива, т.е. поменять местами строки и столбцы – путем выбора подпункта Меню "Транспонирование".

Если в выборках имеются выбросы (очень большие или очень маленькие величины – это можно протестировать с помощью программы IODATA), то вполне оправданным будет исключение этих значений – выбором соответст-

вующего пункта меню (эта операция допустима только для массивов, не содержащих пропуски или даты "для восстановления"). При этом удаляются по два значения из каждой выборки (процедура цензурирования данных). Следует помнить, что после исключения экстремумов операция транспонирования, вообще говоря, лишена смысла, так как практически всегда экстремумы располагаются в различных местах выборок.

При выводе графики следует помнить, что помимо столбцовых диаграмм для средних, сигм, ошибок, коэффициентов вариации возможно получение круговых диаграмм (полигонов Дебеца) этих показателей. Этот способ наиболее интересен для тех массивов данных, в которых все выборки (признаки) измерены в одной и той же единице (в сантиметрах, граммах, баллах и т.д.). С помощью пункта Меню "Сервис" можно изменить цвет фона, замкнутых фигур, сменить шрифт, а при выводе графика "столбчатая диаграмма" сделать некоторые дополнительные операции.

4.3.1. Квантили выборочных распределений

Анализ квантилей выборочных распределений может быть применен, если все значения массива отражают вариабельность одного и того же параметра, а выборки – это либо варианты опыта (некоторым образом упорядоченные), либо подвыборки генеральной совокупности, или же временные точки какого-то процесса. Объекты – либо повторности опыта (допустимы неравное число повторений, отсутствующие данные), либо произвольное множество наблюдений.

Программа позволяет проанализировать структуру изменений в выборках с помощью квантилей эмпирических распределений, которые можно весьма эффективно представить в графической форме. Используется техника обработки данных, изложенная в [39]. Помимо квантилей, в программе вычисляются робастные характеристики выборок: "центральные" средние, размах, коэффициент вариации и т.п.

Выборки в массиве могут быть неравной величины; программа определяет выборку неполного размера по значениям -999. Для каждой выборки выполняются следующие операции.

1. Выбираются значения вероятностей из таблицы:

Объем выборки	Значения вероятностей, p
3 – 4	0,50
5 – 15	0,25 0,50 0,75

16 – 24		0,15	0,25	0,50	0,75	0,85	
25 – 70		0,10	0,25	0,50	0,75	0,90	
71 – 99	0,06	0,10	0,25	0,50	0,75	0,90	0,94
>= 100	0,05	0,10	0,25	0,50	0,75	0,90	0,95

2. Выборка упорядочивается по возрастанию;

3. Вычисляются квантили эмпирических распределений, X(p=0,nn);

4. На базе квантилей вычисляются некоторые "центральные" статистические характеристики выборок, обладающие значительной устойчивостью к различным выбросам, отклонениям от нормальности. Например,

центральное среднее: $\mathbf{X_{ср} = X_{(p=0,75)} + X_{(p=0,25)} - X_{(p=0,5)};}$

центральный размах: $\mathbf{S = 0,674 * [X_{(p=0,75)} - X_{(p=0,25)}];}$

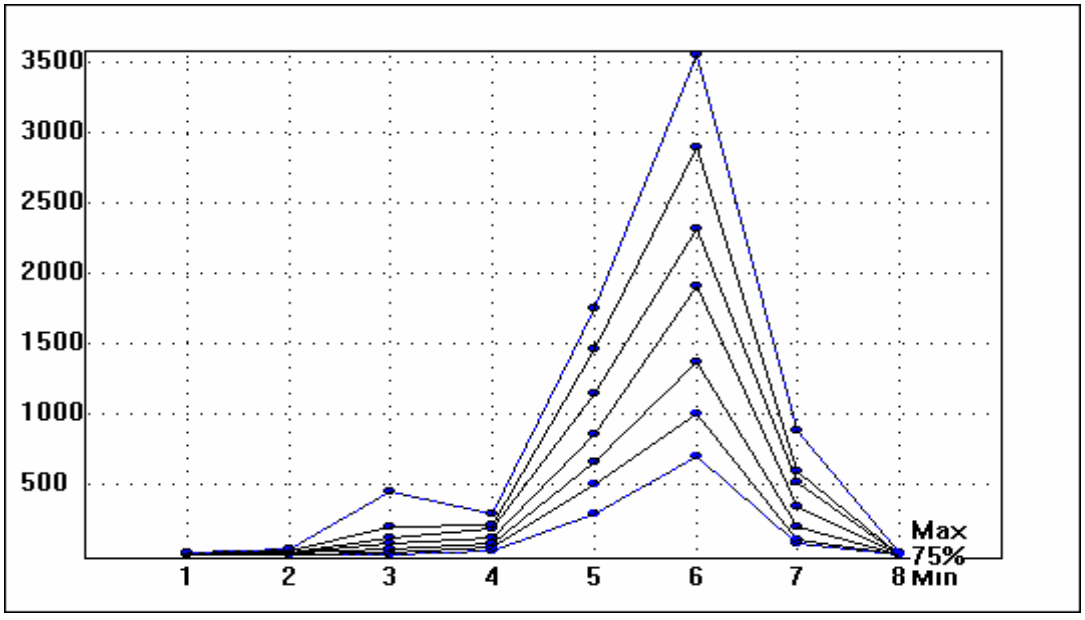
центральный коэффициент вариации: $\mathbf{V = S * 100 / X_{ср}}$;

асимметрия распределения: $\mathbf{As=[X_{(p=0,75)}+X_{(p=0,25)}-2*X_{(p=0,5)}]/[X_{(p=0,75)}-X_{(p=0,25)}]}$

Пример обработки данных по Тьюки (массив ABC8x50.dat):

No выборки	Объем	Центральные среднее	Центральные размах	Центральные вариация	Асимметрия 25%-75%	Асимметрия 10%-90%
1	50	6,85000	4,0777	59,53%	0,1074	0,1340
2	50	14,8000	5,9986	40,53%	0,0337	0,1959
3	50	88,9000	57,492	64,67%	0,1618	0,3685
4	50	146,500	74,477	50,84%	0,2127	0,1018

Для визуального анализа динамики сдвигов в структуре выборок имеется графическое представление квантилей:



4.3.2. Непараметрические доверительные интервалы

Вычисление доверительных интервалов производится по формулам, изложенным в [59]. Основной аргумент для использования непараметрических доверительных интервалов – распределения вероятностей в реальных экспериментальных данных "как правило, отличны от нормальных". Вычисление классических доверительных интервалов основано на асимптотической нормальности выборочных моментов, используются при этом табличные значения критериев Стьюдента и Hi^2. Если выборочное распределение отличается от нормального (например, асимметрично), границы доверительного интервала могут измениться непредсказуемым образом и серьезно отличаться от значений, вычисленных стандартным методом.

Вычисление непараметрических доверительных интервалов среднего базируется на существовании первых двух моментов для любой выборки экспериментальных данных и центральной предельной теореме (ЦПТ) теории вероятностей. Для доверительных интервалов дисперсии должен существовать четвертый момент, коэффициента вариации – третий и четвертый моменты. Непараметрическим доверительным интервалам присуще свойство асимптотичности – они становятся тем точнее, чем больше объем выборки. В отличие от классического метода, непараметрический подход можно применять всегда, "когда случайная величина имеет математическое ожидание и дисперсию, что в силу финитности (ограниченности шкал) бывает практически всегда в реальных ситуациях".

В качестве теста можно использовать массив данных из статьи Орлова ORLOV1x50.dat (гамма-распределение)

4.3.3. Неклассические оценки среднего и ср.-кв. отклонения

Стандартные оценки генерального среднего и среднеквадратичного отклонения по выборкам небольшого размера могут быть искажены единичными артефактами (ошибки измерения, выбросы и т.п.).

Для более корректных оценок среднего и среднеквадратичного отклонения разработан оптимизационный алгоритм поиска минимума суммы квадратов отклонений выборочного распределения от функции вероятности нормального распределения, среднее и ср.-кв. отклонение – параметры оптимизации. Начальным значениям присваиваются стандартные выборочные оценки среднего и сигмы. В большинстве случаев удается получить робастные значения среднего и ср.-кв. отклонения, особенно эффективные для малых выборок.

В основе алгоритма лежит двупараметрический градиентный спуск в область минимума суммы модулей отклонений или суммы квадратов отклонений.

Оценки среднего и ср.-кв. отклонения этим методом рекомендуем использовать для выборок из симметричных распределений. Методом Монте-Карло проверена сходимость таких оценок к генеральным среднему и ср.-кв. отклонению при увеличении объема выборок.

4.3.4. Метод Bootstrap

Метод предложен Б.Эфроном в 1977 году для оценки вариабельности различных статистик, получаемых в результате обработки данных стандартными методами.

Например, для одномерной выборки размером N получены оценки среднего и дисперсии. Если по какой-либо причине исследователя не удовлетворяют значения стандартных ошибок этих оценок и доверительных интервалов, полученных на основе классической теории, можно оценить их вариабельность следующим образом:

1. Задается некоторое большое число (500..1000..10000), определяющее, сколько раз генерировать случайные выборки (того же размера N) из значений имеющегося массива данных.

2. Методом Монте-Карло генерируются эти выборки, и каждый раз вновь вычисляются оценки среднего и дисперсии. Эти оценки накапливаются в соответствующих массивах.

3. По этим массивам определяются средние, среднеквадратические отклонения, экстремумы, квантили, на основании которых можно судить о вариабельности интересующих исследователя статистик. Среднеквадратическое отклонение bootstrap-средних используется для вычисления 95% доверительного интервала среднего по формуле:

$\overline{\mathbf{X}} \pm \mathbf{D}_{95\%} = \overline{\mathbf{X}} \pm \sigma \times \mathbf{T}(\mathbf{n},\mathbf{0.95})$, T(n, 0.95) – критерий Стьюдента.

В случае необходимости можно получить графическое представление распределения bootstrap-оценок в виде гистограмм с переменным числом классов (соответствующий пункт в Меню “Графики”).

4.3.5. Анализ независимости и стационарности рядов

Анализ независимости и стационарности рядов наблюдений выполняется для проверки одной из предпосылок стандартного параметрического анализа (дисперсионный, регрессионный). Для каждой выборки выполняется **тест серий** на основе медианного Tau-критерия. Используется метод, изложенный в [55].

Нуль-гипотеза: последовательность значений ряда случайна – отклонения значений “вверх” и “вниз” от среднего уровня носят случайный характер.

Вначале определяется медиана выборки, это устойчивая к выбросам характеристика среднего значения ряда, затем по значениям выборки формируется последовательность из 1 и –1, означающая, что соответствующее значение либо больше медианы, либо меньше. Далее анализируются серии из единиц – подсчитывается число серий, и определяется размер максимальной серии. Эти значения сравниваются с пороговыми значениями Тау-критерия для числа серий и максимального размера серии на заданном уровне значимости. Если число серий или размер максимальной серии больше пороговых, нуль-гипотеза о стационарности ряда отклоняется, и принимается контр-гипотеза – последовательность значений ряда имеет скрытые зависимости, ряд не стационарен.

4.4. INTER: Анализ группированных данных

INTER – специализированная программа для статистической обработки двумерных массивов данных, представляющих собой частоты значений исследуемого параметра в группах (интервалах значений). Размер массива данных – не более 16000 элементов. Методы обработки данных:

1. Стандартный вариационный анализ выборок в строках массива, сравнение средних по Т-критерию Стьюдента. Предполагается, что весь массив данных – это один параметр (признак), измеренный для нескольких вариантов опыта, хотя возможна обработка и одной выборки.

2. Вычисление коэффициента парной корреляции по частотам в строках/столбцах двумерной таблицы. Массив частот примерно такой же, как для первого метода, по горизонтали – группы/частоты переменной "X", по вертикали – группы/частоты переменной "Y", в крайнем правом столбце – среднегрупповые значения переменной "Y" (этот столбец обычно пустой в массивах для 1-го метода).

Формат двумерного массива:

Число строк = число вариантов + 1 строка значений границ интервалов в группах;

Число столбцов = число интервалов группировки данных + 1 столбец для для правой границы последнего интервала.

В первую строку заносятся значения параметра – левых границ интервалов; например, имееем 7 вариантов опыта и 5 интервалов группировки данных:

2.0-3.999, 4.0-4.999, 5.0-5.999, 6.0-6.999, 7.0-9.0;

эти значения заносятся в первую строку массива:

```
6 8                                          <= начало файла: 6 столбцов, 8 строк
2,00 4,00 5,00 6,00 7,00 9,00                <= левые границы групп
-999  2    6    1   -999 -999                1-й вариант:  9 значений
 2    4    6   -999  2   -999                2-й вариант:  14 значений
 2    5    11   7    2   -999                3-й вариант:  27 значений
-999  1    8    3    1   -999                4-й вариант:  13 значений
 1    2    9    5    3   -999                5-й вариант:  20 значений
 1   -999  6    2   -999 -999                6-й вариант:  9 значений
-999  1   -999  4    2   -999                7-й вариант:  7 значений

Данные по овсу, 1997 г.                      <- необязательный комментарий
```

Средние для интервалов в этом случае будут следующие:

3.0 4.5 5.5 6.5 8.0

Эти значения будут использоваться программой для вычисления статистических характеристик выборок, поэтому границы интервалов должны выбираться соответственно реальным средним. В качестве теста можно использовать массив INTER5x6.dat.

Результаты счета могут быть отредактированы непосредственно в среде программы при выводе результатов на дисплей (заголовок, комментарии, удаление ненужной информации и т.п.). В результате работы программы рассчитываются различные виды статистических характеристик: средние, ср.-квадр. отклонения, коэффициенты вариации, асимметрии, эксцесса, и т.д.

Для анализа достоверности различия средних используются формулы из [10], стр. 99-100. Для эмпирического значения Т-критерия вычисляется вероятность ошибки в случае отклонения нуль-гипотезы: данная пара выборок взята из одной генеральной совокупности, средние различаются только из-за действия случайных факторов. Если вероятность меньше выбранного уровня значимости (обычно 0.05), нуль-гипотеза должна быть отвергнута: выборки по-видимому взя-

ты из различных генеральных со вокупностей. В качестве контроля можно выбрать любой вариант, по умолчанию контрольным считается 1-й вариант.

Удобной возможностью анализа произвольного негруппированного массива данных с большим количеством объектов является передача через буфер Windows любого признака (столбца) из другой программы пакета или из таблиц MS Excel. Для этого следует вначале занести этот признак в буфер, затем кликнуть левой клавишей мышки в поле Редактора массива данных программы, выбрать пункт в Меню "Взять из буфера признак, разнести по классам", при этом автоматически формируются классы (группы), в которые заносятся частоты попадания значений признака после сортировки по возрастанию.

Число классов определяется в соответствии со следующей эмпирической таблицей “число дат в выборке – число классов”:

	Число дат	Число классов
1	5..20	3
2	21..30	4
3	31..40	5
4	41..50	6
5	51..60	7
6	61..70	8
7	71..80	9
8	81..100	10
9	101..300	11
10	301..500	12
11	501..700	13
12	701..900	14
13	901..1200	15
14	1201..1500	16
15	1501..1800	17
16	1801..2100	18
17	более 2100	19

Аналогичная возможность предоставляется при загрузке массива данных "признаки/объекты" посредством Главного Меню программы, но при этом значения всех признаков разносятся по группам, сформированным по первому (крайне левому) признаку массива данных, что не всегда удобно.

4.4.1. Коэффициенты связанности признаков

Вычисление коэффициента парной корреляции Пирсона возможно не только по исходным значениям пар X_i-Y_i, но и по частотам, сгруппированным в двумерной таблице. Используются формулы из[7], стр. 385-386.

Достоверность коэффициента парной корреляции проверяется критерием Фишера-Снедекора, с вычислением вероятности $R_{xy}=0$ (отсутствие линейной зависимости). Если вероятность меньше 0,05, корреляция значима.

Аналогичными показателями степени связанности признаков в двумерных таблицах частот являются коэффициенты взаимной сопряжённости Чупрова и Пирсона ([74], стр. 55). Формулы коэффициентов:

$$K_{ч} = \sqrt{\frac{\varphi^2}{\sqrt{(N-1)(M-1)}}};\ K_{п} = \sqrt{\frac{\varphi^2}{\varphi^2+1}};\ \varphi^2 = \left(\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{x_{ij}^2}{n_i m_j}\right) - 1$$

N и M – число строк/столбцов таблицы частот, n_i и m_j – суммы частот по строкам/столбцам таблицы. Достоверность коэффициентов сопряженности проверяется критерием $\chi 2$, для которого вычисляется вероятность ошибки 1-го рода. Поскольку в формулы не входят классовые значения признаков, эти коэффициенты следует считать непараметрическими показателями связи признаков, поэтому не зависящими от распределения данных в выборках.

В качестве тестов можно использовать массивы ZAX7x6.dat, Glinsky5x6.dat.

4.5. GAUSS: Анализ эмпирических распределений вероятности

GAUSS – специализированная программа для статистической обработки двумерных массивов данных, представляющих собой частоты значений исследуемого параметра в группах (интервалах значений). В программу можно загружать два типа массивов. Первый тип – частоты в группах, экспериментальные данные уже ранее отсортированы по возрастанию и разбиты на группы. Размер такого массива данных – не более 16000 элементов. Второй тип – стандартные массивы "признаки/объекты". Это исходные несортированные данные, подготовленные ранее в какой-либо другой программе пакета, или переданные через буфер Windows, например, из MS Excel. Размер такого массива – до 50 тысяч значений. Предполагается, что любой столбец/признак из этого массива будет передаваться через буфер Windows в среду редактора массива 1-го типа, преобразовываться в массив группы/частоты и далее анализироваться.

Наличие двух типов массивов позволяет обрабатывать данные в виде частот из руководств по прикладной статистике и иных литературных источников.

Методы, реализованные в программе:

1. Анализ типа распределения вероятности по частотам в группах массива данных с помощью критериев $\chi 2$ и Колмогорова.

2. Вероятностный калькулятор: вычисление квантилей и вероятностей для распределения, оценённого программой, или любого другого теоретического распределения вероятности из списка распределений, имеющихся в программе.

3. Моделирование законов распределения вероятностей: генерация массивов псевдослучайных значений с заданным законом распределения.

Формат двумерного массива первого типа:

Число строк = число вариантов + 1 строка значений границ интервалов в группах;

Число столбцов = число интервалов группировки данных + 1 столбец для для правой границы последнего интервала.

В первую строку заносятся значения параметра – левых границ интервалов; например, имееем 7 вариантов опыта и 5 интервалов группировки данных:

2.0-3.999, 4.0-5.999, 6.0-7.999, 8.0-9.999, 10.0-12.0;

эти значения заносятся в первую строку массива:

```
6 8
2,00 4,00 6,00 8,00 10,0 12,0
 1    2    6    8    5   -999
 0    3    7   20   15   -999
 2    5   11    7    2   -999
 1    4    8   12    9   -999
 1    2    9    5    3   -999
 1    3    6    4    2   -999
12    8    6    4    2   -999

Данные по овсу, 1997 г.
```

<= начало файла: 6 столбцов, 8 строк
<= левые границы групп
1-й вариант: 22 значения
2-й вариант: 45 значений
3-й вариант: 27 значений
4-й вариант: 34 значения
5-й вариант: 20 значений
6-й вариант: 16 значений
7-й вариант: 32 значения

<- необязательный комментарий

Средние для интервалов в этом случае будут следующие:

3.0 5.0 7.0 9.0 11.0

Эти значения будут использоваться программой для вычисления параметров распределений, поэтому границы интервалов должны выбираться соответственно реальным средним. В качестве теста можно использовать массив GAUSS7x2.dat.

Результаты счета могут быть отредактированы непосредственно в среде программы при выводе результатов на дисплей (заголовок, комментарии, удаление ненужной информации и т.п.).

Анализ вида распределения вероятностей в экспериментальных данных основан на методах, изложенных в руководствах Г.Н.Зайцева [22], Г.Хана, С.Шапиро [12] и В.А.Бостанджияна [70]. Массив данных в этом случае может состоять из двух строк – в первой строке – нижние границы интервалов группировки, во второй строке – частоты попадания дат в интервалы. Анализ всех типов распределений основан на **равенстве всех интервалов группировки** данных.

Выбор того или иного типа распределения, наиболее соответствующего распределению экспериментальных данных, может быть сделан с помощью критериев χ2 и Колмогорова, а также визуальным анализом гистограммы распределения.

Удобной возможностью анализа произвольного негруппированного массива данных с большим количеством объектов является передача через буфер Windows любого признака (столбца) из массива в поле редактора 2-го типа, из другой программы пакета или из таблиц MS Excel. Для этого следует вначале занести этот признак в буфер, затем кликнуть левой клавишей мышки в поле Редактора массива 1-го типа данных, выбрать пункт в Меню "Взять из буфера признак, разнести по классам", при этом автоматически выполняется сортировка по возрастанию, формируются классы (группы), в которые заносятся частоты попадания значений признака. После этого можно приступать к анализу распределения.

4.5.1. Распределения экспериментальных данных

А.К.Митропольский [27], стр. 163: “Вопрос об установлении вида функции распределения принадлежит к одному из наиболее важных вопросов статистического исчисления”.

В программе GAUSS имеется возможность оценить тип возможного распределения экспериментальных данных методом подбора вида распределения из списка 24 типов различных непрерывных распределений и 6 типов дискретных. Наиболее информативным является визуальный анализ гистограммы распределения с графиком теоретической функции плотности распределения вероятности, параметры которой оценены по экспериментальным данным.

Различия между теоретическими и эмпирическими частотами используются для вычисления стандартного критерия согласия – χ^2 Пирсона. В случае дискрет-

ных распределений теоретические значения частот округляются до ближайшего целого, в таблицах результатов приводятся неокругленные значения. Для значения χ^2 вычисляется вероятность ошибки в случае отклонения 0-гипотезы, чем меньше эта вероятность отличается от единицы, тем ближе распределение данных к выбранному распределению. Минимум значения χ^2 служит дополнительным аргументом для выбора того или иного типа распределения для экспериментальных данных.

Следует заметить, что анализ типа распределения вероятности экспериментальных данных – нетривиальная задача. В литературе можно встретить немало критических замечаний по этому поводу, как по применению критерия χ^2, так и критерия Колмогорова. Автор программы попытался хотя бы немного помочь исследователям в решении этой проблемы. Невозможно охватить все случаи, которые могут быть в экспериментальной практике, следует в первую очередь смотреть гистограмму распределения группированных данных, которая всегда может быть получена при выборе нормального типа распределения, или равномерного.

Для выполнения анализа в программе используются некоторые правила (на уровне постулатов):

1/ равенство всех интервалов группировки данных;

2/ определение числа групп в соответствии с эмпирической таблицей, приведённой выше (стр. 53);

3/ полусуммы пар границ интервалов формируют центр класса (интервала) непрерывных распределений; в массивах данных дискретных распределений значения в ячейках первой строки являются центрами классов.

Эти правила должны свести к минимуму субъективизм при разбиении выборки на классы, позволяют сравнивать между собой результаты анализа разных выборок.

Параметры всех распределений непрерывного типа (за исключением равномерного) рекомендуем оптимизировать быстрой итерационной процедурой, цель которой – поиск *минимума суммы модулей отклонения эмпирических частот от теоретических*. Тем самым достигается наилучшее возможное прохождение кривой функции плотности вероятности по вершинам прямоугольников гистограммы.

Дополнительно, по желанию пользователя, может быть выполнено ещё одно уточнение параметров распределения методом шаговой минимизации суммы модулей отклонений долей (частостей) эмпирического распределения от инте-

гральной функции теоретического распределения вероятности. Тем самым можно получить наилучшее возможное прохождение интегральной кривой (кумуляты) по "ступенькам" эмпирической функции распределения. Эта операция может длиться несколько десятков секунд, и не всегда заканчивается успехом, но в некоторых случаях может существенно улучшить подгонку теоретического распределения под экспериментальные данные.

В случае непрерывных распределений в качестве дополнительного критерия согласия вычисляется критерий Колмогорова и также вероятность ошибки в случае отклонения 0-гипотезы. Как правило, результаты анализа по χ^2 и Колмогорову совпадают, в противном случае решение о принадлежности к тому или иному типу распределения следует принимать на основе графического анализа данных. Критерий Колмогорова вычисляется следующим образом:

а/ эмпирические и теоретические частоты делятся на общее число объектов, получаются значения функции плотности вероятности;

б/ начиная с частот, соответствующих левому хвосту распределения, эмпирические и теоретические частные шаг-за-шагом суммируются, формируя значения интегральной функции вероятности, на каждом шаге вычисляется модуль разницы сумм, его значение заносится в последний столбец таблицы результатов;

в/ максимальное значение разницы – D-критерий Колмогорова, который в форме $D_n = D \cdot \sqrt{N}$, где N – общее число объектов, используется для вычисления вероятности ошибки P в случае отклонения 0-гипотезы;

г/ если $P \leq 0.01$ 0-гипотеза отклоняется на уровне значимости 1%, если $P \leq 0.05$ – отклоняется на на уровне 5%, если $P > 0.10$ – 0-гипотеза подтверждена, распределение эмпирических данных соответствует выбранному.

Для изучения методов анализа распределений экспериментальных данных следует потренироваться на тестовых массивах данных, взятых из руководства Г.Н.Зайцева. Они находятся в подкаталоге \Snedecor\Tests, их название соответствует тому или иному типу распределения: Maxwell12x2.dat, Normal10x2.dat, Exponen10x2.dat, Lognor21x2.dat, Pareto9x2.dat, Puasson17x2.dat, Relay10x2.dat, Gauss7x2.dat, Negativ.dat, BetaOne11x2.dat, Binom11x2.dat, Laplas12x2.dat. Массив Loto37x2.dat – соответствует равномерному распределению вероятностей. Следует учесть, что возможны некоторые несовпадения с результатами в руководстве Г.Н.Зайцева, автор явно делал расчеты на логарифмической линейке или с помощью арифмометра, поэтому иногда случаются неточности и ошибки, а критерий

согласия Колмогорова расчитывался по упрощенной схеме, приводящей к ошибкам в выводах. Критерий Колмогорова неприменим к анализу распределений дискретных данных, однако у Г.Н.Зайцева используется, по-видимому ошибочно (стр.86, 99, 101).

Опыт работы с различными данными позволяет рекомендовать следующее:

1/ Минимально возможный размер массива данных 9-10 значений (тест: массив GAUSS4x2.dat), на меньших массивах практически невозможно достоверно определить тип распределения. ГОСТ 8.207-76 прямо указывает о неэффективности тестирования выборок менее 10 значений. Уверенное определение типа распределения начинается примерно со 100 значений, 8-9 классов. При выборках 40-60 значений также можно делать корректные выводы при "хороших" данных.

2/ Минимальное число классов – три. Для двух классов в принципе нельзя определить тип распределения, по крайней мере, для критерия χ^2 в случае распределений с двумя параметрами будет отрицательное значение числа степеней свободы, в случае распределений с одним параметром – нулевое значение числа степеней свободы. При наличии классов с нулевыми частотами вычисление критерия $\chi 2$ автоматически корректируется объединением таких классов в более значимые частоты, с соответствующим уменьшением числа классов и степеней свободы. Если в результате такого объединения число степеней свободы – нулевое, вероятность ошибки 1 рода для критерия χ^2 не вычисляется.

3/ Желательно, чтобы в среде программы был исходный несортированный массив, признаки которого передаются для анализа, в этом случае процедура дополнительной оптимизации параметров распределения по интегральной функции вероятности будет выполняться по исходной выборке, а не по центральным значениям классов, которые, естественно, несколько снижают информационное содержание массива данных. При наличии исходного массива вычисляется ешё один критерий согласия – Ω-квадрат (Мизеса-Краммера-Смирнова). Вероятность для критерия Ω-квадрат вычсляется интерполяцией по значениям таблицы 6.4а сборника таблиц Л.Н.Большева, Н.В.Смирнова.

4/ Критерий χ^2 плохо работает в случае небольших выборок (30..100 значений), критерий Колмогорова предпочтительнее, но лучше всего решения принимать в результате анализа графика функции плотности веропятости.

Дополнительно вычисляется мера неопределенности в массиве данных – энтропия, на основе методологии, изложенной в книге С.Кульбака [13], 408 с., формула энтропии для группированных данных:

$$H(p_1, p_2 ... p_k) = -\sum_{i=1}^{k} p_i Log_2(p_i)$$

p_i – доли частот от общего размера массива данных, k – число интервалов группировки данных.

Энтропия обычно измеряется в **битах**, это определяется применением логарифмов по основанию 2, но её можно вычислять с помощью натуральных логарифмов, тогда энтропия выражается в **нитах.** Энтропия определяется вероятностями всех элементарных событий в исследуемой системе, называемой полем, и служит мерой его неопределённости. Для всякого поля, связанного с неопределенностью результатов, энтропия всегда положительна, поэтому перед знаком суммы стоит знак "минус".

После того, как сделан выбор типа распределения и получены результаты анализа, можно использовать Вероятностный Калькулятор, вызываемый из под-Меню "Операции". С его помощью можно вычислить вероятность по выборочному значению, а также выборочный квантиль по значению вероятности.

4.5.1.1. Нормальное распределение

Нормальное распределение имеет наибольшее значение для теории и практики экспериментальных исследований. Оценка нормальности данных должна быть первым шагом перед использованием всех прочих методов прикладной статистики. Чем ближе распределение экспериментальных данных к нормальному, тем больше корректность применимости большинства методов обработки данным, тем больше достоверность получаемых результатов.

Теоретические предположения о сути нормального распределения: переменная, значения которой распределены по нормальному закону, формируется в результате действия суммы большого числа произвольно распределённых факторов, вклад каждого из которых относительно невелик.

В программе GAUSS выполняется оценка параметров возможного нормального распределения, формирование на их основе теоретического распределения частот для центральных значений интервалов группировки.

Значения теоретических частот предполагаемого нормального распределения вычисляются по формуле (Г.Н.Зайцев, стр. 73):

$$f_{teor} = N \cdot c \cdot f(x,\mu,\sigma);\ f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} Exp(-0{,}5(x-\mu)^2/\sigma^2);$$

$f(x,\mu,\sigma)$ – функция плотности вероятности, μ – параметр положения центра распределения, оценка средним арифметическим, N – общее количество дат в выборке, c – размер классового интервала, σ – параметр рассеяния, оценивается средним квадратическим отклоненим, вычисляется с поправкой Шеппарда для группированных данных.

В пакете SNEDECOR есть специализированная программа для анализа нормальности выборок любого размера – NORMAL. В ней используются 7 различных критериев согласия эмпирического распределения с нормальным.

4.5.1.2. Распределение Максвелла

Распределение эмпирических частот вариационного ряда положительных значений с умеренной положительной асимметрией может быть аппроксимировано распределением Максвелла. Это распределение вероятностей непрерывного типа, вид уравнения определяется одним параметром, теоретическая частота распределения вычисляется по формуле (Г.Н.Зайцев, стр. 108):

$$f_{teor} = N \cdot c \cdot f(x,\alpha);\ f(x,\alpha) = \sqrt{\frac{2}{\pi}} \cdot \frac{x^2}{\alpha^3} \cdot Exp\left(\frac{-x^2}{2\alpha^2}\right);\ \alpha = \mu \cdot \sqrt{\frac{\pi}{8}}$$

$f(x,\alpha)$ – функция плотности вероятности, N – общее количество дат в выборке, c – размер классового интервала, α – параметр распределения, оцениваемый первым центральным моментом μ, x – центральные значения интервалов группированных данных. Распределение Максвела имеет фиксированное значение коэффициента вариации, равное 0,42203 или в процентах – 42,2%. Мода и медиана являются функциями параметра распределения: Mo=α*1,41421; Me=α*1,538.

Распределение Максвелла имеет фундаментальное значение в физике (анализ распределения скоростей молекул газов и жидкостей), часто используется в анализе качества технологических процессов.

4.5.1.3. Распределение Рэлея

Распределение вероятностей по закону Рэлея пригодно для аппроксимации вариационных рядов положительных значений с умеренной асимметрией. Это

распределение данных непрерывного типа, характер распределения определяется одним параметром. Оценкой применимости этого распределения для некоторого массива группированных данных служит выражение:

$\sigma = -1 + \bar{x} \cdot \sqrt{4/\pi}$; $\bar{x}$ – среднее арифметическое

Это примерно 0,523 среднего арифметического. Теоретические частоты распределения Рэлея вычисляются по формуле (Г.Н.Зайцев, стр. 111):

$$f_{teor} = N \cdot c \cdot f(x,\alpha);\ f(x,\alpha) = \frac{x}{\alpha^2}\,Exp(-0{,}5x^2/\alpha^2);\ \alpha^2 = 2 \cdot \mu^2/\pi$$

$f(x,\alpha)$ – функция плотности вероятности, N – общее количество дат в выборке, c – размер классового интервала, α – параметр распределения, x – центральные значения интервалов группированных данных. Мода, дисперсия и медиана распределения данных, подчиняющихся закону Рэлея – простые функции от параметра распределения α:

$$Mo = \alpha;\ \sigma^2 = \alpha^2(2 - \pi/2);\ Me = \sqrt{2\alpha^2 \cdot Ln(2)};$$

4.5.1.4. Показательное распределение

Массив положительных значений с резко выраженной левосторонней асимметрией может быть аппроксимирован показательным, или экспоненциальным, распределением. Характерной особенностью этого распределения непрерывного типа является равенство значений среднего и среднеквадратического отклонения. Примерное равенство выборочных среднего и сигмы служит критерием применимости теста выборки на возможность показательного распределения. Поведение показательного распределения определяется одним параметром, теоретические частоты вычисляются по формуле (Г.Н.Зайцев, стр. 114):

$$f_{teor} = N \cdot c \cdot f(x,\lambda);\ f(x,\lambda) = \lambda \cdot Exp(-\lambda x);\ \lambda = 1/\mu;\ 0 < x < \infty;$$

N – общее количество дат в выборке, λ – параметр распределения, c – размер классового интервала. Дисперсия и медиана данных, подчиняющихся показательному распределению, определяются через параметр λ: $\sigma^2 = 1/\lambda^2$, $Me = Ln(2)/\lambda$. Интегральная функция вероятности показательного распределения определяется формулой (Г.Хан, С.Шапиро, стр. 111):

$$F(x,\lambda) = \int_0^x \lambda \cdot Exp(-\lambda t) \cdot dt = 1 - Exp(-\lambda x);$$

Показательному распределению соответствует время до момента появления одного события, если события появляются независимо друг от друга с постоянной средней интенсивностью, или же это распределение интервала времени между моментами появления независимых случайных событий с постоянной средней интенсивностью. Используется в теории массового обслуживания, теории надежности систем и оборудования.

4.5.1.5. Распределение Парето

Двупараметрическое распределение значений выборок, в которых большая часть данных непрерывного типа очень тесно сгруппирована в области малых положительных значений, а частота больших значений далее бысто падает, может быть аппроксимировано распределением Парето. Основной параметр распределения β определяется нижней границей размаха варьирования выборочных значений. Функция плотности вероятности распределения Парето и интегральная функция вероятности определяются формулами ([1], стр. 186):

$$f(x,\alpha,\beta)=\frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1};\ F(x,\alpha,\beta)=1-\left(\frac{\beta}{x}\right)^{\alpha};\ \beta\le x<\infty$$

Параметры распределения оцениваются методом моментов:

$$\alpha=1+\sqrt{1+(\hat{\mu}/\hat{\sigma})^2};\ \beta=(1-1/\alpha)\cdot\hat{\mu};$$

Медиана и дисперсия данных, подчиняющихся распределению Парето, определяются формулами:

$$Me=\beta\cdot Exp(Ln(2)/\alpha);\ \sigma^2=\beta^2/(\alpha-2)/(\alpha-1)^2$$

но дисперсия может быть вычислена только при $\alpha>2$. Цитата из [1] по поводу распределения Парето:

Характер случайного варьирования исследуемого признака в генеральных совокупностях, из которых заранее изъяты все элементы со значением данного признака, не превосходящим заданного порогового уровня β. Таким образом, это распределение сильно усечённых слева выборок, используемое в экономике, социологии (анализ доходов).

4.5.1.6. Равномерное распределение

Распределение дискретных и непрерывных величин с равной вероятностью попадания в любой интервал группировки данных, является равномерным рас-

пределением. Функция распределения частот этого распределения – прямая линия, паралельная оси абсцисс, ограниченная слева и справа границами соответствующих классов. Вероятность попадания значения в какой-либо класс определяется только числом классов, средняя (теоретическая) частота для любого класса:

$$P = 1/K;\ \bar{f} = \frac{1}{K}\sum_{i=1}^{k} f_i;\ K - \text{число классов}$$

f_i – эмпирические частоты попадания в классы, полученные в эксперименте.

Г.Н.Зайцев (стр.123-124) приводит следующий пример применения теста на равномерность распределения в селекции сортов культурных растений: правомерно ли считать сортовым признаком, например, высоту растений у совокупности сортов какой-либо культуры? Если распределение частот достоверно отличается от равномерного, только тогда высота растений у данной совокупности сортов может служить сортовым признаком.

На базе псевдослучайных значений равномерного распределения в интервале [0..1] формируются все остальные непрерывные распределения с помощью различных формул, алгоритмов.

4.5.1.7. Логарифмически нормальное распределение

Распределение случайной величины, логарифм которой распределён по нормальному закону, называется логарифмически нормальным. Считается, что случайная величина, подчиняющаяся логнормальному закону, формируется как произведение множества малых случайных факторов.

Логнормальное распределение весьма удобно для приближения самых различных экспериментальных выборок с умеренной положительной асимметрией. Характер этого распределения определяется двумя параметрами, оцениваемые следующими формулами ([1], стр. 215):

$$\hat{\mu} = x_{Мод};\ \ \hat{\sigma} = \sqrt{\frac{1}{(n-1)} \cdot \sum_{i=1}^{n} (Ln(x_i) - Ln(\hat{\mu}))^2};$$

μ – аналог среднего для данного типа распределения (модальное значение выборки, разбитой на классы), σ – аналог среднеквадратического отклонения. n – общее количество дат в выборке. После предварительной оценки параметров выполняется оптимизирующая процедура поиска значений параметров – с минимизацией суммы модулей отклонений эмпирических частот в интервалах от теоре-

тических частот, вычисленных по функции плотности вероятности логнормального распределения:

$$\mathbf{f(x,\mu,\sigma) = \frac{1}{\sigma x\sqrt{2\pi}} \cdot Exp\left(-\frac{(Ln(x) - Ln(\mu))^2}{2 \cdot \sigma^2} \right);\ 0 < x < \infty;}$$

Логарифмически нормальное распределение применяется в самых различных областях – для описания процессов, в которых наблюдаемое значение составляет случайную долю от предыдущего значения. Широко используется в экономической статистике, в биологии. Г.Хан, С.Шапиро, стр. 120:

"Примером может служить распределение размеров организма, развитие которого происходит под влиянием большого числа незначительных воздействий, эффект каждого из которых пропорционален мгновенному значению размера организма".

4.5.1.8. Гамма-распределение

Двупараметрическое асимметричное распределение положительных значений [52], стр. 46. В некоторых справочниках приведена однопараметрическая форма гамма-распределения ([51], стр. 185). Используется в теории массового обслуживания, теории надежности систем, математической статистике. Формула функции плотности вероятности:

$$\mathbf{f(x,\eta,\lambda) = \frac{\lambda^{\eta}}{\Gamma(\eta)} x^{\eta-1} Exp(-\lambda x);\ \hat{\lambda} = \bar{x}/s^2;\ \hat{\eta} = \hat{\lambda}\bar{x};\ \Gamma(\eta) = (\eta-1)!}$$

η, λ – параметры гамма-распределения оцениваемые методом моментов, $\Gamma(\eta)$ – гамма-функция Эйлера, факториал произвольного положительного числа. Коэффициенты асимметрии и эксцесса – функции параметров:

$$\mathbf{C_{ass} = 2/\sqrt{\eta};\ C_{exc} = 6/\lambda;}$$

Массив данных, передаваемый для анализа на гамма-распределение, не должен содержать отрицательных значений. Критерием возможности гамма-распределения данных в выборке является примерное равенство выборочных оценок среднего арифметического и дисперсии.

Частными видами гамма-распределения являются распределения χ^2, Эрланга и показательное (экспоненциальное).

4.5.1.9. Бета-распределение 1 рода

В тех случаях, когда диапазон изменения исследуемой переменной ограничен слева и справа некоторыми положительными значениями, для моделирования распределения вероятности используется бета-распределение 1-го рода. Формула функции плотности бета-распределения, определенного для интервала [0..1] ([12], стр. 112):

$$\mathbf{f(x,\gamma,\eta) = \frac{\Gamma(\gamma+\eta)}{\Gamma(\gamma)\cdot\Gamma(\eta)} x^{\gamma-1}(1-x)^{\eta-1};\ 0 \le x \le 1;\ \gamma > 0;\ \eta > 0}$$

Γ(a) – гамма-функция Эйлера, факториал произвольного положительного числа, γ и η – параметры распределения, определяющие его форму, которая может быть одновершинной (с различной степенью асимметрии), U-образной, убывающей и J-образной. Возможно моделирование данных для произвольного интервала [a..b], для этого используется обобщённая форма бета-распределения. Параметры обычно оцениваются методом моментов ([12], стр. 117):

$$\mathbf{\hat{\eta} = \frac{1-\bar{x}}{s^2}[\bar{x}(1-\bar{x}) - s^2];\ \hat{\gamma} = \bar{x}\cdot\hat{\eta}/(1-\bar{x});}$$

Бета-распределение используется для моделирования процессов при статистическом контроле качества, в теории надёжности. Частными видами бета-распределения являются равномерное, треугольное и параболическое распределения.

4.5.1.10. Бета-распределение 2 рода

В отличие от Бета-распределения 1 рода, диапазон изменения исследуемой переменной ограничен только слева (нулевым значением), справа возможны произвольно большие значения. Бета-распределение 2-го рода – также двупараметрическое, оба параметра влияют на форму функции плотности вероятности. Формула функции плотности бета-распределения 2 рода, определенного для интервала [0..∞] ([27], стр. 258):

$$\mathbf{f(x,\gamma,\eta) = \frac{\Gamma(\gamma+\eta)}{\Gamma(\gamma)\cdot\Gamma(\eta)} \cdot \frac{x^{\eta-1}}{(1+x)^{\gamma+\eta}};\ 0 \le x < \infty;\ \gamma > 0;\ \eta > 0;}$$

Γ(x) – Гамма-функция Эйлера. Параметры распределения оцениваются методом моментов:

$$\hat{\gamma} = \frac{2 \cdot m_2 - m_1^2 + m_1}{m_2 - m_1^2};\ \hat{\eta} = m_1 \cdot (\hat{\gamma} - 1);$$

m_1 и m_2 – первый и второй нулевые моменты распределения, вычисленные по экспериментальным данным. Бета-распределение 2 рода используется для моделирования процессов при статистическом контроле качества, в теории надёжности.

4.5.1.11. Распределение Коши

Симметричное двупараметрическое распределение вероятностей непрерывного типа, характеризующееся длинными хвостами (наличие значительного количества экстремальных значений в обе стороны от центра распределения). Формула функции плотности вероятности ([12], стр. 123):

$$f(x,\mu,\sigma) = \frac{1}{\sigma \cdot \pi \cdot (1 + (x-\mu)^2 / \sigma^2)};\ -\infty < x < \infty;\ -\infty < \mu < \infty;$$

μ – параметр положения распределения, σ – параметр, аналогичный среднеквадратическому отклонению относительно центра распределения. Своеобразие распределения Коши заключается в том, что конечные моменты (математическое ожидание, дисперсия и прочие) не существуют. Вместо них используются: в качестве оценки центра – медиана или мода, в качестве дисперсии – сумма квадратов отклонений от оценки центра распределения.

Как статистическая модель, распределение Коши используется для представления случайных величин, у которых в норме могут быть значения, отстоящие очень далеко от центра μ в любом направлении, но не являющиеся выбросами в смысле обычных распределений вероятности.

4.5.1.12. Распределение Вейбулла – Гнеденко

Распределение Вейбулла-Гнеденко используется как статистическая модель распределения данных об отказах в работе оборудования, в анализе смертности/выживания живых организмов. При этом, в отличие от моделей, описываемых показательным (экспоненциальным) распределением, интенсивность отказов (поломок, летальных исходов) может быть переменной, зависящей от времени, и обычно подчиняющейся следующей зависимости:

$$\mathbf{In(t)} = \frac{\eta}{\sigma}\left(\frac{t}{\sigma}\right)^{\eta-1}$$

Отсюда плотность двупараметрического распределения Вейбулла-Гнеденко имеет вид:

$$\mathbf{f(x,\eta,\sigma)} = \frac{\eta}{\sigma}\left(\frac{x}{\sigma}\right)^{\eta-1} \cdot \mathbf{Exp}\left[-\left(\frac{x}{\sigma}\right)^{\eta}\right];\ x \geq 0,\ \sigma > 0,\ \eta > 0$$

σ – параметр масштаба, η – параметр формы. Плотность распределения положительных значений может принимать самые разнообразные формы – колоколообразные ($\eta>1$) с различной степенью асимметрии, убывающие ($\eta\leq1$). Частными случаями распределения Вейбулла являются экспоненциальное распределение ($\eta=1$), распределение Релея ($\eta=2$).

Интегральная функция вероятности распределения Вейбулла-Гнеденко выражается формулой:

$$\mathbf{F(x,\eta,\sigma)} = 1 - \mathbf{Exp}\left[-\left(\frac{x}{\sigma}\right)^{\eta}\right];\ x \geq 0$$

Параметры распределения оцениваются методом преобразования интегральной функции к линейной зависимости ([12], стр. 327) с последующим применением метода наименьших квадратов.

Предполагается, что первый интервал массива данных начинается с нулевого значения, в этом случае не требуется вводить третий параметр, характеризующий центр распределения и обычно определяющий длительность начального периода, в течение которого отказы не происхододят. При наличии такого параметра следует просто вычесть его величину из выборочных значений.

4.5.1.13. Логистическое распределение

Двупараметрическое распределение данных непрерывного типа с симметричной функцией плотности вероятности, близкой к плотности нормального распределения. Используется в качестве модели распределения численности особей в анализе динамики популяций, анализе смертности/выживания объектов (пробит-анализ, логит-анализ), в теоретической биологии.

Формула функции плотности вероятности ([68], стр. 44):

$$\mathbf{f(x,\mu,\sigma) = \frac{\pi \cdot Exp(-\pi(x-\mu)/\sigma/\sqrt{3})}{\sigma \cdot \sqrt{3} \cdot \left[1 + Exp(-\pi(x-\mu)/\sigma/\sqrt{3})\right]^2};}$$

μ и σ – первый и второй центральные моменты распределения, оцениваются средним и средне-квадратическим отклонением по экспериментальным данным, аналогично нормальному распределению. Логистическое распределение может быть представлено в упрощённой форме, без нормирующих коэффициентов (функция плотности вероятности, интегральная функция распределения вероятности):

$$\mathbf{f(x) = Exp(x)/\left[1 + Exp(x)\right]^2;\ F(x) = 1/\left[1 + Exp(-x)\right];\ \mu = 0;\ \sigma = 1;}$$

4.5.1.14. Распределение Лапласа

Однопараметрическое распределение данных непрерывного типа с симметричной функцией плотности вероятности, характеризуемое острой вершиной в центральной точке функции плотности распределения. Другое название этого распределения – двустороннее экспоненциальное, так как оно как бы "склеено" из двух зеркальных функций показательного распределения.

Формула функции плотности вероятности ([1], стр. 185):

$$\mathbf{f(x,\lambda) = \frac{\lambda}{2} \cdot Exp(-\lambda \cdot |x - \bar{x}|);\ -\infty < x < \infty}$$

Распределение используется в прикладной статистике при анализе остаточных отклонений от регрессии (случайных ошибок), может быть применено как модель данных для выборок, значения которых в основном группируются около центра (тестовый массив – Laplas12x2.dat).

Параметр распределения оценивается методом моментов:

$$\mathbf{\lambda = \frac{N}{\sum_{i=1}^{N} |x_i - \bar{x}|};\ N - размер\ выборки,\ \bar{x} - среднее;}$$

Интегральная функция вероятности распределения Лапласа состоит из двух формул (Википедия):

$$\mathbf{F(x,\lambda) = Exp(\lambda \cdot (x - \bar{x}))/2 \ \ при\ x \le \bar{x};}$$

$$\mathbf{F(x,\lambda) = 1 - Exp(-\lambda \cdot (x - \bar{x}))/2 \ \ при\ x > \bar{x};}$$

4.5.1.15. Распределение Стьюдента

Симметричное колоколообразное однопараметрическое распределение данных непрерывного типа, интенсивно используемое в прикладной статистике (доверительные интервалы, различие средних, достоверность коэффициентов регрессии и т.д.). Может служить модельным распределением в тех же ситуациях, что и нормальное распределение или логистическое. Функция плотности вероятности стандартизованной случайной величины **X** (с центром в точке **X**=0):

$$\mathbf{f(x,n) = \frac{\Gamma((n+1)/2)}{\sqrt{n \cdot \pi} \cdot \Gamma(n/2)} \left(1 + x^2/n\right)^{-(n+1)/2};\ -\infty < x < +\infty;}$$

n – параметр масштаба, обычно целое число, называемое числом степеней свободы. В принципе n может быть вещественным числом, поэтому в программе этот факт используется. С ростом n распределение Стьюдента приближается к стандартному нормальному распределению. Распределение Коши является частным случаем распределения Стьюдента – для n=1**.**

Параметр n оценивается методом моментов: $n=2\sigma^2/(\sigma^2-1)$.

4.5.1.16. Распределение Hi-квадрат

Распределение χ^2 является частным случаем Гамма-распределения, один из параметров которого равен 1/2, а удвоенный второй параметр, обычно целое число (=n), называется "числом степеней свободы". Это связано с тем, что случайная величина, распределённая по χ^2, представляет собой сумму квадратов n случайных величин, распределённых нормально со средними=0.0 и единичными сигмами. Функция плотности распределения χ^2 может иметь асимметричную колоколообразную форму, приближающуюся к нормальному распределению с ростом n, а также L-образную форму, характерную для экспоненциального (показательного) распределения.

Формула функции плотности вероятности (А.К.Митропольский, стр. 325):

$$\mathbf{f(x,n) = \frac{x^{(n-2)/2}}{2^{n/2} \cdot \Gamma(n/2)} Exp(-x/2);\ 0 < x < \infty}$$

Г(a) – Гамма-функция Эйлера. Целочисленный параметр n, вообще говоря, может быть вещественным числом (дробное число степеней свободы). В при-

кладном статистическом анализе этот факт используется в случаях, связанных с нарушениями предпосылок классического анализа данных.

Параметр n вначале оценивается по экспериментальным данным как $n=X_{сред}–X_{миним}$, затем выполняется итерационный поиск оптимального значения n, с минимизацией суммы модулей отклонений эмпирических частот от теоретических частот, вычисленных по функции плотности вероятности.

В качестве модельного распределения χ^2 может использоваться в тех же ситуациях, что и Гамма-распределение, с учетом единственности параметра формы.

Распределение χ^2 широко используется в теоретической и прикладной статистике в качестве критерия согласия, в анализе таблиц сопряженности и т.п.

4.5.1.17. Распределение Фишера–Снедекора

Отношение единичных дисперсий двух независимых случайных величин, распределённых по нормальному закону, называется F-статистикой Фишера. Распределение вероятности того, что это отношение, **x**, больше некоторого порогового значения, определяемого из размеров этих выборок, подчиняется закону Фишера – Снедекора. Формула функции плотности вероятности этого распределения:

$$\mathbf{f(x,m,n) = \frac{m^{m/2} \cdot n^{n/2} \cdot x^{(m-2)/2}}{B(m/2,n/2) \cdot (n + x \cdot m)^{(m+n)/2}}; 0 \le x < \infty;}$$

B(a,b) – beta-функция Эйлера; параметры **m** и **n**, обычно целые числа, называемые числами степеней свободы, определяются, исходя из размеров выборок, но, согласно теории, могут быть и вещественными, нецелыми. В программе параметры **m** и **n** оцениваются процедурой быстрой минимизации суммы модулей отклонений частот, вычисленных по формуле функции плотности вероятности, от экспериментальных частот.

Распределение Фишера – Снедекора широко используется в прикладной статистике (дисперсионный анализ, корреляционный, регрессионный). В качестве модельного может использоваться в случае унимодальных распределений со значительной асимметрией, ограниченных слева.

4.5.1.18. Распределение Пуассона

Распределение вероятностей по закону Пуассона используется для аппроксимации данных дискретного характера, то есть целочисленных значений, отражающих частоты некоторых редких событий, возникающих с малой вероятностью за определенный период. Эмпирические частоты при распределении Пуассона являются числом одинаковых проб, имеющих ту или иную долю наблюдаемого признака. Характерным признаком данных, распределенных по Пуассону, является примерное равенство средней частоты признака и дисперсии: $\mu=\sigma^2$.

Это однопараметрическое распределение, теоретические частоты которого вычисляются по формуле:

$$\mathbf{f_{teor} = N_n \cdot f(x,\lambda);\ f(x,\lambda) = \frac{\lambda^x}{x!} \cdot Exp(-\lambda);\ \lambda = \mu;}$$

λ – параметр распределения, N_n – общее число проб, x – частота признака в некотором классе. Функция распределения вероятности пуассоновской случайной величины ([68], стр. 27):

$$\mathbf{F(x,\lambda) = \sum_{i=0}^{x} \frac{\lambda^i}{i!} \cdot Exp(-\lambda);}$$

Коэффициенты асимметрии, эксцесса, вариации вычисляются по формулам:

$$\mathbf{C_{Ass} = \lambda^{-1/2};\ C_{Exc} = \lambda^{-1};\ \sigma = \lambda^{1/2};\ C_v = 100/\sigma;}$$

Википедия: "Распределение Пуассона моделирует случайную величину, представляющую собой число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга. Распределение Пуассона играет ключевую роль в теории массового обслуживания."

При вводе данных следует учесть, что в качестве значения частоты признака в некотором классе используется левая граница интервала, а не центр интервала, как в случае непрерывных распределений, но при вводе данных следует создавать на один класс больше – для согласования с общей структурой данных программы.

4.5.1.19. Гипергеометрическое распределение

Если в эксперименте имеется некоторое число групп (=G) с равным числом объектов (=m), и у части объектов имеется, а у других отсутствует некоторый признак, то частоты **f** появления групп с определенным числом [0..m] объектов,

имеющих этот признак, могут быть распределены по биномиальному или гипергеометрическому распределению.

Аппроксимация распределения таких дискретных данных гипергеометрическим распределением используется в тех случаях, когда нет возможности каждый раз возвращать обследованный объект обратно в генеральную совокупность и перемешивать все объекты надлежащим образом, чтобы вероятность отбора новой информации по учитываемому признаку оставалась неизменной от пробы к пробе. Гипергеометрическое распределение, в отличие от биномиального, предоставляет возможность учесть общий объем совокупности объектов, а также не возвращать объекты обратно в совокупность и не перемешивать её, что достаточно удобно в большинстве случаев.

Теоретические частости (фактически плотности вероятности) для появления групп с определеным, равным X, числом объектов, имеющих некоторый признак, вычисляются по формуле (Г.Н.Зайцев, стр. 89):

$$\mathbf{f(x,m,N)=\frac{C_{n_p}^{x}\,C_{n_q}^{m-x}}{C_N^m};\ C_n^m=\frac{n!}{m!\cdot(n-m)!};\ \ m\le n}$$

m – размер групп объектов, N – общее число объектов, n_p – общее число объектов, имеющих некоторый признак, n_q – число объектов, не имеющих требуемый признак; $N=n_p+n_q$.

Массив данных, предназначенный для теста на гипергеометрическое распределение, должен быть дополнен значением размера группы объектов, которое следует ввести в крайнюю правую (пустую у массивов данных для теста всех прочих распределений) ячейку строки. Это значение будет сохраняться при записи массива на диске или ином носителе информации.

Среднее, дисперсия и коэффициент вариации массива данных, распределённых гипергеометрически:

$$\mathbf{\bar{x}=m\cdot n_p/N;\ \sigma^2=\bar{x}\frac{n_q(N-m)}{N(N-1)};\ C_v=100\cdot\sqrt{\sigma^2/\bar{x}};}$$

4.5.1.20. Отрицательное биномиальное распределение

Отрицательное, или негативное биномиальное распределение данных относится к виду дискретных распределений с двумя параметрами. Цитата из [1], стр.208:

Число v(k) независимых экспериментов, которое пришлось провести в ожидании заданного числа k – появления интересующего нас события, когда вероятность появления этого события в одном испытании остается постоянной и равна P. Формула функции вероятности отрицательного биномиального распределения:

$$\mathbf{P\left\{v_p(k)=x\right\}=C_{x-1}^{k-1}p^k(1-p)^{x-k};\ p=\frac{k}{n}\sum_{i=1}^{n}\frac{1}{x_i};\ \bar{x}=k/p;\sigma^2=\frac{k(1-p)}{p^2};}$$

Характерное свойство данного распределения – дисперсия должна быть больше средней арифметической. Параметры k и p распределения оцениваются методом моментов (Г.Н.Зайцев, стр. 96):

$$\mathbf{k=\bar{x}/(\sigma^2/\bar{x}-1);\ p=1/(1+\bar{x}/k);}$$

4.5.1.21. Биномиальное распределение

Если в эксперименте имеется некоторое число групп (=G) с равным числом объектов (=m), и у части объектов имеется, а у других отсутствует некоторый признак, то частоты **f** появления групп с определенным числом [0..m] объектов, имеющих этот признак, могут быть распределены по биномиальному или гипергеометрическому распределению.

Аппроксимация распределения таких дискретных данных биномиальным распределением используется в тех случаях, когда обследованные объекты возвращаются обратно в генеральную совокупность и все объекты перемешиваются надлежащим образом, чтобы вероятность отбора новой информации по учитываемому признаку оставалась неизменной от пробы к пробе.

Теоретические частости (фактически плотности вероятности) для появления групп с определеным, равным x, числом объектов, имеющих некоторый признак, вычисляются по формуле ([1], стр. 208):

$$\mathbf{f(x,m,p)=C_m^x p^x(1-p)^{m-x};\ p=n_p/N;}$$

C_m^x – биномиальные коэффициенты, m – размер групп объектов, N – общее число объектов, n_p – число объектов, имеющих некоторый признак, x=0,1,...m. Если ξ – случайная величина из [0..m], где m – фиксированное число наблюдений, то распределение вероятности случайной величины ξ имеет следующий вид ([68], стр. 23):

$$P(\xi \le k) = F(k,m,p) = \begin{cases} 0, & k < 0; \\ \sum_{x=0}^{k} C_m^x p^x (1-p)^{m-x}; \\ 1, & k \ge m; \end{cases}$$

Массив данных, предназначенный для теста на биномиальное распределение, должен быть дополнен значением размера группы объектов, которое следует ввести в крайнюю правую (пустую у массивов данных для теста всех прочих распределений) ячейку строки. Это значение будет сохраняться при записи массива на диске или ином носителе информации.

Среднее, дисперсия и коэффициент вариации массива данных, распределённых биномиально:

$$\bar{x} = m \cdot p;\ \sigma^2 = m \cdot p \cdot (1-p);\ C_v = 100\sqrt{(1-p)/(m \cdot p)}$$

4.5.1.22. Геометрическое распределение

Если в испытаниях по схеме Бернулли необходимо найти вероятность такого события, что потребуется последовательно провести **x**-1 независимых экспериментов, каждый с постоянной вероятностью успеха P, чтобы добиться успеха в эксперименте **x**, то распределение вероятности такой случайной величины **x** называется геометрическим распределением. Функция вероятности этого распределения ([12], стр. 181):

$$f(x,P) = P \cdot (1-P)^{x-1};\ x = 1,2,3,...;\ 0 \le P \le 1;$$

Можно рассматривать и обратную схему – найти вероятность появления первого неудачного исхода после серии **x**-1 успешных экспериментов. Тогда формула функции вероятности выглядит следующим образом:

$$f(x,1-P) = P^{x-1} \cdot (1-P);\ x = 1,2,3,...;\ 0 \le P \le 1;$$

Параметр P оценивается с помощью итерационного процесса – минимизируется сумма модулей отклонений экспериментальных частот от теоретических, вычисленных по первой формуле. Тестовый массив – GEOM8x2.dat.

4.5.1.23. Распределение Паскаля

Если в испытаниях по схеме Бернулли необходимо оценивать вероятность такого события, что потребуется провести x+k независимых экспериментов, каждый с постоянной вероятностью успеха P, чтобы получить ровно k успешных исходов, то распределение вероятности такой случайной величины x называется

распределением Паскаля. Функция вероятности этого распределения ([12], стр. 183):

$$f(x,k,P) = C_x^{x+k-1} \cdot P^k \cdot (1-P)^x;\ x = 0,1,2,...;\ k = 1,2,...;\ 0 \le P \le 1;$$

Геометрическое распределение является частным случаем распределения Паскаля – для k=1. Параметр P оценивается с помощью итерационного процесса – минимизируется сумма модулей отклонений экспериментальных частот от теоретических, вычисленных по данной формуле. Целочисленный параметр k (число успешных исходов) должен задаваться пользователем перед началом обработки данных.

Обобщением распределения Паскаля является отрицательное биномиальное распределение. В нём параметр k – вещественное число, а множитель C (число сочетаний) может заменяться гамма-функциями. Для теста можно использовать массив PASCAL13x2.dat

4.5.2. Распределения Пирсона

Английский математик Карл Пирсон, исходя из анализа множества гистограмм, предположил, что практически все эмпирические одномерные распределения случайных величин могут быть аппроксимированы функциями специального вида, которые выводятся из единственного дифференциального уравнения:

$$\frac{df(x)}{dx} = \frac{(x-a)\cdot f(x)}{b_0 + x\cdot b_1 + x^2 \cdot b_2}$$

Интегрирование этого уравнения приводит к семейству 13 функций плотности вероятности, из которых 9 считаются основными и ещё 4 – частными случаями основных при определённых значениях параметров – число которых от 3 до 4, позволяет весьма гибко моделировать эмпирические распределения.

Частными случаями распределений Пирсона являются классические распределения – нормальное (VIII тип), экспоненциальное (IX тип), Парето, гамма- и бета-распределения, и, соответственно, все частные виды этих классических распределений.

Функции плотности вероятности распределений Пирсона позволяют моделировать распределения с самыми различными видами гистограмм – колоколообразными с разной степенью ассимметрии, J-образными, U-образными. Выбор типа распределения может быть сделан с помощью значения некоторого параметра κ, вычисляемого перед анализом данных. Если

κ <0 – область I типа;

κ =0 – II тип при отрицательном эксцессе, или VII тип при положительном,

0< κ <1 – область IV типа,

κ =1 – обычно соответствует V типу,

κ >1 – область VI типа,

при κ –>+∞ или κ –>–∞ – III тип.

Значение κ визуально представляется на графике в виде желтой окружности на оси возможных значений. Рекомендация выбора типа распределения с помощью κ – не догма, довольно часто эмпирическое распределение может быть смоделировано несколькими видами распределений.

Дополнительной возможностью подгонки некоторых типов распределений под экспериментальные данные является указание левой (и/или правой) границы выборочных значений. Она может быть равна значению крайнего левого класса или немного меньше его (и, соответственно, крайнего правого класса, или немного больше его).

Следует заметить, что если на экране появляется сообщение типа "Дискриминант<0.0" или "Минус под радикалом", это означает непригодность выбранного типа распределения для введённых экспериментальных данных, следует выбрать другой тип.

Анализ данных по типам распределений Пирсона основывается на формулах, изложенных в книге В.А.Бостанджияна [70].

4.5.2.1. Распределение Пирсона I типа

4-параметрическое распределение, ограниченное слева и справа, может быть применено для моделирования различных эмпирических распределений с ограниченным размахом – колоколообразных, J-образных, U-образных. Формула функции плотности вероятности ([70], стр. 27):

$$\mathbf{f(x,a_1,a_2,m_1,m_2)=\frac{(x-a_1)^{m_1}(a_2-x)^{m_2}}{(a_2-a_1)^{(m_1+m_2+1)}B(m_1+1,m_2+1)}}$$

B – Beta-функция Эйлера, a_1 и a_2 – левая и правая границы функции распределения, m_1 и m_2 – параметры формы распределения. Распределение I типа может считаться обобщением Beta-распределения, так как после замены $Y=(x-a_1)/(a_2-a_1)$ и при границах [0..1] оно становится классическим двупараметрическим Beta-

распределением 1 рода. Параметры распределения оцениваются методом моментов ([70], стр. 30-32).

4.5.2.2. Распределение Пирсона II типа

3-параметрическое симметричное распределение, частный случай распределения I типа при $m_1=m_2$, так же ограниченное слева и справа. Формула функции плотности вероятности ([70], стр. 44):

$$\mathbf{f(x,a_1,a_2,m)=\frac{(x-a_1)\cdot(a_2-x)^m}{(a_2-a_1)^{(2m+1)}B(m+1,m+1)}}$$

B — Beta-функция Эйлера, a_1 и a_2 — левая и правая границы функции распределения, m —параметр формы распределения. Параметры оцениваются методом моментов. Может быть применено для моделирования симметричных распределений с различным эксцессом, ограниченных слева и справа вследствие некоторых свойств объектов, ограничений изучаемой системы.

4.5.2.3. Распределение Пирсона III типа

Распределение односторонне ограниченных случайных величин, являющееся трёхпараметрическим Gamma-распределением, может быть использовано для моделирования колоколообразных и J-образных эмпирических распределений. Функция плотности вероятности ([70], стр. 56):

$$\mathbf{f(x,m,\gamma,a)=\frac{\gamma^m}{\Gamma(m)}(x-a)^{m-1}Exp(-\gamma(x-a));\ a\le x<\infty;\ m,\gamma>0}$$

Γ(x) – Gamma-функция Эйлера, **a** – левая граница выборки, **m** – параметр формы **γ** – параметр масштаба. Параметры оцениваются методом моментов ([70], стр. 63).

Таким образом, распределение Пирсона III типа можно эффективно использовать для приближения самых различных экспериментальных данных со значительной асимметрией распределения.

4.5.2.4. Распределение Пирсона IV типа

Четырёхпараметрическое распределение IV типа – наиболее сложное по формуле плотности вероятности и более трудное в реализации вычислительных процедур. В отличие от предыдущих типов, границы распределения отсутствуют. Формула функции плотности ([70], стр. 72):

$$f(x,a,\sigma,m,v)=\frac{(1+y^2)^m}{\sigma\cdot F(2m-2,v)}\mathrm{Exp}(-v\cdot\mathrm{Arctg}(y));\ y=\frac{x-a}{\sigma};\ -\infty<x,a<\infty;$$

a – параметр сдвига, **σ** – масштаба, **m,v** – параметры формы, **F(b,c)** – нормирующая функция довольно сложного вида, вычисляется приближенно (бесконечное произведение дробей).

Может быть применимо для моделирования самых различных эмпирических распределений. Интегральная функция вероятности не выражается через элементарные или специальные функции и не табулирована. Вследствие численного интегрирования функции вероятности возникают небольшие временные задержки, точность интегрирования может быть недостаточной в некоторых случаях.

Параметры распределения оцениваются методом моментов.

4.5.2.5. Распределение Пирсона V типа

Трёхпараметрическое распределение V типа может иметь только колоколообразную форму с различной степенью асимметрии, при правосторонней асимметрии выборка должна иметь границу слева, при левосторонней асимметрии – границу справа. Формула функции плотности вероятности ([70], стр. 86):

$$\mathbf{f(x,a,\sigma,m)=\frac{\sigma^m}{\Gamma(m)}\cdot y^{-m-1}Exp(-\sigma/y);\ y=Sgn\cdot(x-a);\ a\le x<\infty\ (-\infty<x\le a)}$$

a – параметр сдвига, σ – масштаба, m – параметр формы, Γ(x) – гамма-функция Эйлера, Sgn=1 при левосторонней асимметрии, Sgn=–1 при правосторонней асимметрии. Параметры распределения оцениваются методом моментов. Интегральная функция вероятности выражается посредством неполной гамма-функции:

$$\mathbf{F(x,a,\sigma,m)=1-I(\sigma/y,m)\ \ или\ \ F(x,a,\sigma,m)=I(\sigma/y,m);}$$

при правосторонней или левосторонней асимметрии выборочного распределения.

4.5.2.6. Распределение Пирсона VI типа

Четырёхпараметрическое распределение VI типа может быть применено для моделирования ограниченных слева эмпирических распределений самой разнообразной формы. Его частным случаем является Beta-распределение 2-го рода,

которое ещё называют стандартным распределением VI типа. Формула функции плотности вероятности ([70], стр. 96):

$$\mathbf{f(x,a_1,a_2,m_1,m_2)=\frac{(a_1-a_2)^{m_2-m_1-1}\cdot(x-a_1)^{m_1}}{B(m_2-m_1-1,m_1+1)\cdot(x-a_2)^{m_2}};a_1\le x<\infty;m_2>m_1+1}$$

a1 – левая граница, a2 – параметр сдвига, m1 и m2 – параметры формы, B(x,y) – Beta-функция Эйлера. Параметры распределения оцениваются методом моментов. Интегральная функция вероятности распределения VI типа может быть получена преобразованием неполной Beta-функции. В качестве тестового примера можно использовать массив Pears_VI_9x2.dat.

4.5.2.7. Распределение Пирсона VII типа

3-параметрическое симметричное распределение, являющееся обобщением хорошо известного в прикладной статистике однопараметрического распределения Стьюдента. В отличие от симметричного распределения Пирсона II типа, не имеет границ. Формула функции плотности вероятности ([70], стр. 108):

$$\mathbf{f(x,a,\sigma,m)=\frac{\left(1+(x-a)^2/\sigma^2\right)^{-m}}{\sigma\cdot B(m-0.5,0.5)};-\infty<x<\infty;\ \sigma>0;\ m>0.5;}$$

a – параметр сдвига, σ – параметр масштаба, m – параметр формы, B(x,y) – Beta-функция Эйлера. Параметры оцениваются методом моментов ([70], стр. 111). Может быть использовано для моделирования симметричных распределений с различной степенью эксцесса. В качестве теста можно использовать массив Mitro18x2.dat.

4.5.3. Оценка вероятности с помощью гистограммы распределения

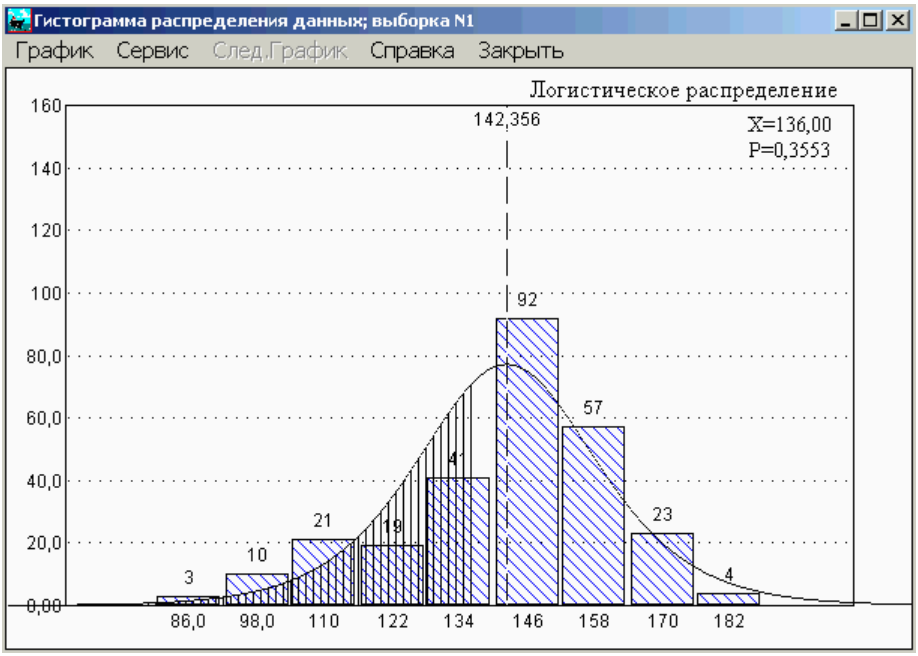
Для практического использования результатов моделирования эмпирического распределения можно использовать следующий приём:

– выбрать желаемый тип модели, получить результаты анализа критериями согласия;

– вывести на экран график “Гистограмма распределения”, с помощью мышки сделать размер графика максимально удобным для анализа,

– кликами мышки в поле под функцией плотности вероятности будет выполняться штриховка области под функцией, соответствующей интегралу функции вероятности, значение исследуемой переменной (квантиль) и соответствующей ему вероятности будет выведено в верхнем левом углу графика.

Для дискретных распределений вместо интеграла вычисляется сумма соответствующих прямоугольных областей графика, поэтому значения вероятности будут ступенчатыми, соответственно сумме интервалов группировки.



Аналогичная возможность имеется на графике интегральной фунции вероятности – клик мышки приводит к оценке квантиля и соответствующей ему вероятности, приводимых в нижнем правом углу графика.

4.5.4. Вероятностный калькулятор

После того, как сделан выбор типа распределения и получены результаты анализа, можно использовать Вероятностный Калькулятор, вызываемый из под-Меню "Операции". С его помощью можно вычислить вероятность для какого-либо выборочного значения (“клавиша” [P = ?]), а также выборочный квантиль по значению вероятности (“клавиша” [X = ?]). Перед этим следует ввести с клавиатуры нужное значение в соответствующее место формы. При сравнении с аналогичными значениями, которые можно получить кликом мышки на графиках "Гистограмма" и "Интегральная функция", могут быть небольшие расхождения, связанные с численным интегрированием сложных функций.

Значения параметров выбранного распределения можно редактировать, поэтому можно вычислить вероятность и, соответственно, квантиль для произвольного распределения вероятностей, а не только того, что было получено оценкой экспериментальных данных. Точность вычислений – не более 5 значащих цифр, чаще всего 4.

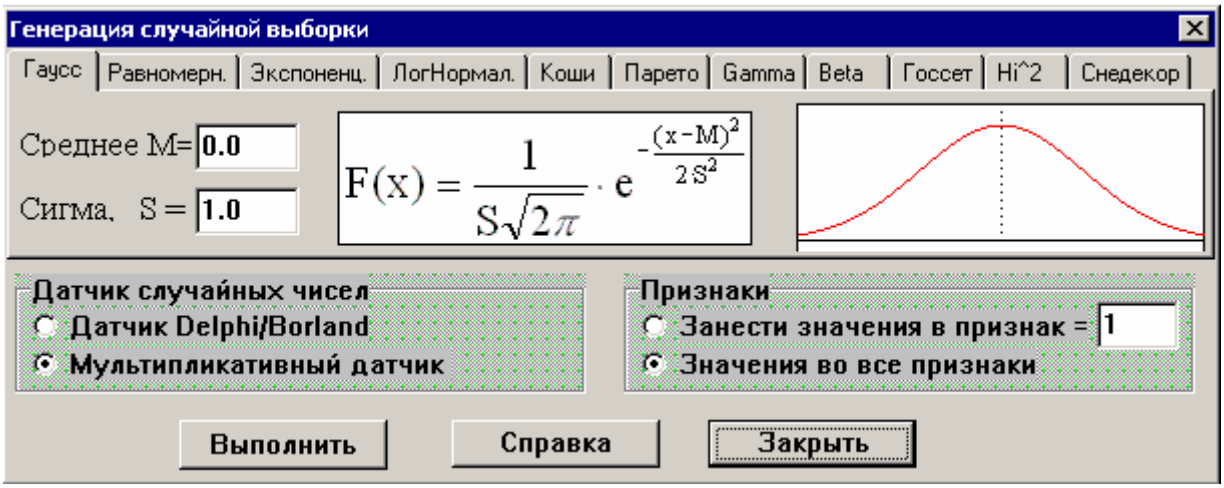
Следует заметить, что для некоторых типов распределний приводятся дополнительные параметры, помимо тех параметров, которые входят в формулу плотности вероятности. Это связано с тем, что в теории интервал значений многих распределений ограничен слева нулевым значением, но в реальных данных левая граница может быть иной, следовательно необходимо указывать это значение. Для Beta-распределения 1-го типа указываются обе границы – слева и справа, так как в теории это распределение ограничено интервалом [0..1]. Для реальных данных с распределением Стьюдента необходимо указывать среднее и среднеквадратическое отклонение, так как в теории у распределения Стьюдента нулевое среднее и единичная сигма.

4.5.5. Генерация выборок с заданным законом распределения

В ряде случаев бывает необходимо сравнить выборочное распределение с каким-либо известным теоретическим распределением вероятностей, протестировать те или иные методы обработки данных с помощью модельных массивов данных с известным законом распределения.

В меню «Операции» имеется пункт «Генерация случайных выборок», с помощью которого можно сформировать массивы псевдослучайных данных, законы распределения которых соответствуют следующим теоретическим непрерывным распределениям вероятностей:

– нормальному распределению (Гаусса),
– равномерному распределению,
– показательному распределению (экспоненциальному),
– логарифмически нормальному распределению,
– распределению Коши,
– распределению Парето,
– gamma-распределению,
– beta- распределению 1 рода,
– распределению Стьюдента (Госсета),
– распределению Hi-квадрат,
– распределению Фишера-Снедекора.



Для создания выборок с этими законами распределения используются два типа алгоритмов генерации псевдослучайных чисел:

1) процедура Random, встроенная в математическую библиотеку системы программирования Delphi,

2) мультипликативный процесс порождения случайного числа из предыдущего с очень большим периодом [50].

Общие рекомендации: размер генерируемой выборки должен быть не менее 50 значений, на меньших выборках характерные свойства распределения будут слабо выражены. Мультипликативный датчик обычно дает более "качественные" распределения, чем датчик Delphi.

Для практического знакомства с вероятностными распределениями следует сформировать массив, например из 10 признаков х 1000 значений, затем генерировать различные типы распределений по всем десяти признакам, далее анализировать характер распределения данных на графиках "Гистограмма" и "Интегральная функция вероятности ".

Последовательность действий при работе по генерации выборок:

1/ Выбрать кликом мышки редактор массива данных 2-го типа;

2/ Создать новый (пустой) массив 2-го типа (признаки/объекты), выбрав пункт главного Меню "Ввести новый массив с клавиатуры", задать требуемые размеры массива, например, 5 признаков, 700 объектов;

3/ Выбрать пункт главного Меню "Генерация случайных массивов", на открывшейся форме выбрать тип распределения, ввести желаемые параметры распределения, номер признака – куда должен быть занесён массив сгенерированных значений, или же во все признаки массива, тип датчика случайных чисел..

4/ После выполнения операции либо передать какой-нибудь признак в среду редактора массивов 1-го типа для разбиения на группы и последующего анализа,

либо записать массив "признаки/объекты" для обработки массива другими программами.

5. Дисперсионный анализ

Классический (параметрический) дисперсионный анализ Р.Фишера предполагает выполнение нескольких предпосылок [8, стр. 376] для обрабатываемого массива данных:

1. Нормальность распределения ошибок измерения.

2. Равенство дисперсий ошибок измерения.

3. Статистическая независимость ошибок в последовательности измерений данных в эксперименте.

Равенство дисперсий ошибок обычно трактуется как равенство выборочных дисперсий вариантов опыта, так как иным способом оценить вариабельность ошибок измерения не представляется возможным.

Статистическая независимость ошибок может быть проверена различными методами, например, тестом серий, тестом достоверности автокорреляций; для этого надо двумерный массив “варианты-повторения” преобразовать в одномерный одностолбцовый массив, например:

```
12,3 12,5 14,2          12,3
8,34 11,0 13,1     ->   12,5
8,48 10,3 11,7          14,2
                        8,34
                        11,0
                        13,1
                        8,48
                        10,3
                        11,7
```

затем вычислить матрицу автокорреляций (программа MATRIX), либо выполнить тест серий (SERIES или VARS).

В программах пакета обычно есть возможность записать массив остатков для теста нормальности программой NORMAL, выполнить тест однородности дисперсий в вариантах массива данных (программа COMPAR).

Проверка нормальности малых выборок 6-10 дат обычно весьма проблематична. Согласно ГОСТ проверка нормальности таких выборок вообще не делается, так как практически всегда результат будет подтверждать гипотезу нормаль-

ности. Есть, однако, критерии (Уилк-Шапиро, Колмогоров-Смирнов), которые могут быть использованы для теста нормальности малых выборок.

В руководствах по прикладной статистике обычно имеются фразы типа “при умеренных отклонениях от нормальности допустимо использование классических методов”. Следует по возможности добиваться **равного** числа повторений, избегать использования дисперсионного анализа для данных из явно дискретных распределений (целочисленные значения – баллы, экспертные оценки, численности, частоты, данные в виде 0 и 1, и т.п.). Для таких данных нужно использовать непараметрические аналоги дисперсионного анализа.

Основной результат работы программы параметрического дисперсионного анализа – F-критерий Фишера-Снедекора. 0-гипотеза чаще всего формулируется следующим образом: отсутствует действие изучаемого фактора типа Fixed, средние имеют разные значения вследствие действия случайных факторов. Для F-критерия вычисляется "вероятность ошибки в случае отклонения 0-гипотезы". Если

P <= 0,01 действие фактора подтверждено на уровне значимости 1%,

P <= 0,05 действие фактора подтверждено на уровне значимости 5%,

P > 0,10 действие фактора не подтверждено.

В многофакторном анализе обычно допускается возможность взаимодействия факторов, для проверки этого вычисляются F-критерии для каждого вида взаимодействия (2-, 3-факторные и т.д.).

0-гипотеза для проверки взаимодействия: факторы влияют на изучаемую систему как простая сумма воздействий, отсутствует эффект взаимоусиления (синергизм) или взаимоподавления (антагонизм) факторов. F-критерий для взаимодействия и соответствующая ему вероятность трактуются аналогично:

P <= 0,01 взаимодействие факторов подтверждено на уровне 1%,

P <= 0,05 взаимодействие факторов подтверждено на уровне 5%,

P > 0,10 взаимодействие факторов не подтверждено.

После доказательства действия фактора типа “Fixed” выполняется анализ достоверности различия факторных средних по Т-критерию Стьюдента в форме НСР (Наименьшей Существенной Разницы, в иностранных публикациях LSD, Least Significant Difference) на заданном уровне значимости (1, 5 или 10%). Следует помнить, что Т-критерий полностью корректен **только при 2-х вариантах** фактора, и может привести к ошибочным выводам при большем числе вариантов. Для строгого анализа достоверности различия факторных средних необходимо

использовать программу COMPAR. В этой программе множественное сравнение средних выполняется критериями Шеффе, Тьюки и другими.

Существует проблема относительно трактовки факторов типа “Random”. Определение Хикса [34, стр. 204-205]**:** уровни вариантов “Random” типа выбираются случайным образом из бесконечной совокупности возможных уровней. Насколько это применимо к практике экспериментирования? Например, принято считать, что фактор “Годы” (многолетние полевые опыты) – Random типа. В этом случае действие фактора означает лишь доказательство различия дисперсий по некоторым годам, относительно различия среднемноголетних теория ничего говорит, тогда как экспериментаторам как раз нужна оценка достоверности различия средних по годам.

В этой ситуации можно рекомендовать следующий подход: вначале обработать данные по модели “Mixed”; если для “Random” фактора 0-гипотеза **не отклоняется** (нет различий дисперсий по годам), выполнить стандартный анализ по модели “Fixed”. Если в этом случае выявится действие фактора “Годы”, приступить к анализу различия среднемноголетних. Отсутствие различия дисперсий по годам будет всего лишь выполнение одной из предпосылок параметрического дисперсионного анализа – однородности дисперсий. В случае, когда обнаружено различие дисперсий, следует анализировать среднемноголетние данные непараметрическими методами.

5.1. D1MAXI: 1-факторный дисперсионный анализ

Программа D1MAXI предназначена для обработки экспериментальных данных методом однофакторного дисперсионного анализа, с возможностью анализа различий средних по величине НСР (Наименьшей Существенной Разницы). Массив данных может иметь выпавшие значения, в этом случае используется итерационный алгоритм "восстановления" данных по Снедекору [20], стр. 294-295. В программе можно создать новый массив, заполнить его числами, использовать для редактирования ранее введенный массив, сформированный в стандарте пакета SNEDECOR.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами –

через буфер Windows, из текстового файла. Пример формирования массива из 3-х повторностей и 6-и вариантов в текстовом файле:

```
3 6                  <- начало файла
12,3 12,5 14,2
8,34 -1,0 13,1       <- 2-я повторность – для восстановления
8,48 10,3 11,7          массив данных:
9,23 14,6 11,0            строки = варианты,
7,12 10,9 10,3            столбцы = повторности
8,27 10,4 12,4
```

В качестве примеров формирования массива для программ, обрабатывающих массивы "варианты-повторности" можно посмотреть файлы D1MAXI.dat, PEREG233.dat, LAKIN222.dat. Массивы данных, подготовленные для обработки программами дисперсионного анализа большей факторности, могут быть переданы программе D1MAXI и обработаны как 1-факторные данные. Существует возможность обрабатывать массивы данных с неравным числом повторений. В этом случае число повторений определяется по значению "-999" при анализе строки массива слева направо:

```
4 6                  <- начало файла
12,3 12,5 14,2 13,4  4 повт.
8,34 9,37 8,25 -999  3 повт.  массив данных:
9,23 10,6 -999 -999  2 повт.    строки = варианты,
7,12 6,39 8,27 7,55  4 повт.    столбцы = повторности.
8,27 19,4 32,4 -999  3 повт.   "-999" – признак завершения
6,21 8,15 9,66 7,21  4 повт.    повторностей варианта.

Вес листьев          <- необязательный комментарий
```

В качестве примера формирования массива с неравным числом повторений можно использовать файлы DAN1.dat, DAN2.dat. Программа автоматически распознает тип массива, обрабатывая такие данные по типу "полной рандомизации".

F-критерий Фишера-Снедекора вычисляется для двух методов организации эксперимента: полной рандомизации вариантов/повторностей и рандомизации вариантов в блоках повторностей. Дисперсионный анализ подразумевает математическую модель данных (полная рандомизация):

$\mathbf{y_{ij} = \mu + a_i + e_{ij}}$; $\boldsymbol{\mu}$ – генеральное среднее изучаемой системы;

a_i – эффект варианта фактора типа Fixed;

e_{ij} – ошибка от случайных факторов, распределена по $N(0, \sigma)$.

0-гипотеза: все $a_i=0$, контр-гипотеза: некоторые $a_i=/=0$. Математическая модель в случае рандомизации в блоках:

$y_{ij} = \mu + a_i + r_j + e_{ij}$; r_j – возможный эффект блока повторений.

Для проверки одной из предпосылок классического дисперсионного анализа – нормальности распределения остатков – программа может сформировать 1-мерный массив остатков в соответствии с моделью данных (полная рандомизация/рандомизация в блоках), который далее записывается в виде файла для обработки программой NORMAL.

Пример обработки 1-факторных данных:

Таблица разложения дисперсии ANOVA. Рандомизация в блоках.

Дисперсия	Сумма квадратов	Доля вариации	Степени свободы	Средний квадрат	F-критерий
Общая	44,498	1.0000	20	2,225	
Фактор	27,885	0,6267	6	4,647	4,172
Повторения	3,247	0,0730	2	1,623	
Сл.Факторы	13,367	0,3004	12	1,114	

F-критерий = 4,1723, ст.св.=6, 12, P=0,0170
Степень влияния по Снедекору = 0,5140
Станд.Ошибка = 0,6093 (5,20% от общего среднего)
НСР(1%)= 2,6322 НСР(5%)= 1,8776 НСР(10%)= 1,5359

Вероятность, вычисленная для значения F-критерия, меньше 0,05; таким образом, принимается контр-гипотеза: фактор действует, некоторые средние достоверно различаются. Значение (5,2% от общего среднего) в некоторых руководствах трактуется как показатель “точности опыта”, в других книгах такая характеристика игнорируется. Важный показатель “Влияние фактора” – можно использовать либо “степень влияния по Снедекору”, либо “Долю вариации” из таблицы ANOVA.

5.1.1. Анализ различия средних

После выявления действия фактора приступают к анализу различия средних. В качестве контрольного варианта автоматически предлагается первый вариант; если же в действительности в опыте контрольным был другой вариант, его номер нужно ввести в окне установок параметров анализа. Достоверность различия средних в сравнении с контрольным вариантом может быть проверена любым из четырёх критериев – Шеффе, Тьюки, T(n), Стьюдента. Число средних (вариантов) фактора не должно превышать 20 [ограничение, связанное с размерами таблиц критериев Тьюки и T(n)]; при выборе критериев Шеффе и Стьюдента число вариантов может быть произвольным.

————Повторности————————————————————

Варианты	1	2	3	Средние	Разница	Значима?
1	12,30	12,50	14,20	13,00	Контроль	
2	10,00	13,00	12,10	11,70	-1,300	Нет
3	9,000	10,00	11,00	10,00	-3,000	Да!
4	13,00	14,00	11,50	12,83	-0,167	Нет
5	11,00	10,90	10,30	10,73	-2,267	Да!
6	10,00	10,40	12,00	10,80	-2,200	Да!
7	13,00	14,00	12,00	13,00	0,000	Нет
Средние	11,19	12,11	11,87	11,724	-1,276	Нет

Критерий Стьюдента предлагается программой по умолчанию, это связано с общепринятой практикой Российских естествоиспытателей, однако следует помнить, что это оправдано только при малом числе вариантов (2..4) и при больших значениях критерия Фишера-Снедекора, с вероятностью ошибки 1-го рода порядка 0,005 и менее.

Средние каждого фактора вычисляются на основе фактически имеюшихся значений (исключаются пропуски, даты для восстановления), затем анализируются сравнением с величиной Наименьшей Существенной Разницы (НСР, в англоязычной литературе LSD – Least Significant Differens) на выбранном уровне значимости.

Все критерии сравнения средних, используемые в программе, полностью эквивалентны для случая эксперимента из 2-х вариантов; как только число вариантов увеличивается, достоверность выявления ДЕЙСТВИТЕЛЬНО различающихся пар средних может падать.

Самым жестким критерием достоверности различия средних является критерий Шеффе; менее категоричными, но достаточно строгими в теоретическом плане являются критерии Тьюки и T(n); критерий Стьюдента может привести к завышенному числу "достоверно" различающихся пар средних при большем числе вариантов. Известный критерий Дункана основан на эмпирическом правиле выявления достоверной разницы, и не может считаться строгим критерием для определения достоверных различий (однако, имеется в программе COMPAR). Формулы для анализа средних взяты из [21], стр. 30-34.

Рекомендуем пользоваться для анализа средних, формируя НСР на базе критерия Тьюки.

5.1.2. Анализ данных с возможными отклонениями от нормального распределения

В классическом дисперсионном анализе Фишера предполагается, что данные во всех вариантах распределены по нормальному закону. В большинстве руководств по прикладной статистике упомянуто, что при умеренных отклонениях распределения данных от нормального дисперсионный анализ Фишера применим, но с различными оговорками:

– должна быть равночисленность всех вариантов,

– число повторений в вариантах должно быть достаточно большим, отсутствие пропусков,

– дисперсии вариантов должны быть равными (однородными),

– данные не должны быть целыми числами – баллами, частотами (подозрение на дискретное распределение).

Если распределение дат в вариантах действительно иное, критерий Фишера – Снедекора неприменим, так как может быть серьёзно снижена его мощность – способность правильно различать ситуации справедливости 0- и конт-гипотез на заданном уровне достоверности.

В реальных ситуациях весьма сложно проверить нормальность малых выборок, типичное значение числа повторений – 3..8 дат, большинство критериев нормальности с такими выборками не работают (кроме критерия Уилка – Шапиро).

Если возникает подозрение, что данные в каких-то вариантах распределены не по нормальному закону, следует либо использовать непараметрические аналоги дисперсионного анализа (по Краскелу-Уоллесу, Фридману, Уилсону и т.п.), либо сделать критерий Фишера – Снедекора устойчивым к отклонениям от нормальности и в то же время сохранить достаточную мощность, чтобы быть чувствительным к действию фактора. В программе используется метод Бокса – Андерсена ([42]. стр. 28-33). В этом случае степени свободы F-критерия корректируются в сторону уменьшения, в общем случае делая их нецелыми (вещественными). Используется лишь предположение о независимости и одинаковой распределённости дат в вариантах.

Метод Бокса – Андерсена может быть применён как к случаю полной рендомизации вариантов/повторений, так и к плану с рендомизацией в блоках повторений. Метод применим к неравночисленному массиву данных, который обрабатывается по типу полной рендомизации.

Для использования коррекции степеней свободы критерия Фишера – Снедекора следует выбрать страницу "Метод вычислений" на форме "Параметры анали-

за данных" и пометить пункт "Устойчивый к ненормальности данных". Так же, как и для классического F-критерия, вычисляется вероятность ошибки в случае отклонения 0-гипотезы – интеграл Бета-распределения 1-го рода, параметры которого формируются из значений F-критерия и его степеней свободы.

5.2. D2MAXI: 2-факторный дисперсионный анализ

Программа D2MAXI предназначена для обработки экспериментальных данных различными методами параметрического 2-факторного дисперсионного анализа, с возможностью анализа различий средних сравнением с Наименьшей Существенной Разницей, вычисляемой на базе критериев Шеффе, Тьюки, T(n), Стьюдента. Массив данных может иметь выпавшие значения (помечаемые как -1.0).

Данные в виде двумерного массива "варианты-повторения " могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла.

Ограничения на размер массива данных: общее число вариантов – не более 8000, максимальное количество вариантов в любом факторе – 100, максимальное количество повторностей – 100, общий размер массива – не более 16000 элементов. Пример формирования массива из 4-х повторностей, 2-х вариантов фактора "A" и 3-х вариантов фактора "B" в текстовом файле:

```
4 6 2 3                  <- начало файла
12,3 12,5 14,2 13,1      1a1b вар. Всего 2*3=6 вариантов;
8,34 13,7 -1,0 11,9      1a2b вар. <- в 3-й повторности дата для восстан.
13,3 14,6 11,0 15,3      1a3b вар.
7,12 9,08 10,3 11,5      2a1b вар.  Массив данных:
8,27 9,56 7,33 10,3      2a2b вар.  Строки  = варианты,
11,5 12,1 11,7 13,5      2a3b вар.  Столбцы = повторности.

Данные за 1998 г         <- необязательный комментарий
```

В качестве примера формирования массива для программы D2MAXI можно посмотреть файлы D2MAXI.dat, DOSP250.dat. Массивы данных, подготовленных для обработки методами дисперсионного анализа большей факторности (3 и выше), не могут быть переданы программе D2MAXI и обработаны как 2-факторные данные.

Существует возможность обрабатывать массивы данных с неравным числом повторений. В этом случае число повторений определяется по значению "-999" при анализе строки массива слева направо:

```
4 6 2 3
12,3 12,5 14,2 13,1
8,34 13,7 8,25 -999
13,3 14,6 -999 -999
7,12 9,08 10,3 11,5
8,27 9,56 7,33 -999
11,5 12,1 11,7 13,5

Вес листьев
```

<- начало файла
4 повт.
3 повт. Массив данных:
2 повт. Строки = варианты,
4 повт. Столбцы = повторности.
3 повт. "-999" – признак завершения
4 повт. Повторностей варианта.

<- необязательный комментарий

В качестве примера формирования массива с неравным числом повторений можно использовать файл DAN2.dat. Программа автоматически распознает тип массива, обрабатывая такие данные по типу "полной рандомизации".

F-критерий Фишера-Снедекора вычисляется для двух методов организации эксперимента: полной рандомизации вариантов/повторностей и рандомизации вариантов в блоках повторностей. Дисперсионный анализ подразумевает математическую модель данных (полная рандомизация):

$\mathbf{y_{ijk} = \mu + a_i + b_j + ab_{ij} + e_{ijk}}$; μ – генеральное среднее изучаемой системы;

a_i – эффект варианта фактора A типа Fixed;

b_j – эффект варианта фактора B типа Fixed;

ab_{ij} – эффект взаимодействия факторов;

e_{ijk} – ошибка от случайных факторов, распределена по $N(0, \sigma)$.

0-гипотезы: все $a_i=0$, все $b_j=0$, все $ab_{ij}=0$;

контр-гипотезы: некоторые $a_i=/=0$, $b_j=/=0$, $v_{ij}=/=0$.

Математическая модель в случае рандомизации в блоках:

$\mathbf{y_{ijk} = \mu + a_i + b_j + ab_{ij} + r_k + e_{ijk}}$; r_k – возможный эффект блока повторений.

Для проверки одной из предпосылок классического дисперсионного анализа – нормальности распределения остатков, программа может сформировать 1-мерный массив остатков в соответствии с моделью данных (полная рандомизация/рандомизация в блоках), который далее записывается в виде файла для обработки программой NORMAL.

Пример обработки 2-факторных данных (файл LAKIN224.dat):

Таблица ANOVA. Полная рендомизация.

Дисперсия	Сумма квадратов	Доля вариации	Степени свободы	Средний квадрат	F-критерий

Общая	15,779	1.0000	35	0,451	
Случайные Факторы	6,227	0,3946	24	0,259	
Варианты	9,552	0,6054	11	0,868	3,3471
=> Фактор "A"	0,511	0,0324	2	0,255	0,9839
=> Фактор "B"	7,939	0,5031	3	2,646	10,200
=>Взаимодействие	1,103	0,0699	6	0,184	0,7084

Анализ действия факторов

Фактор	Степень влияния	Критерий Фишера-Снедекора F	ст.своб.	вероятность	Наим.Существ.Разность 1%	5%	10%
A	0,0000	0,984	2, 24	0,38842	0,582	0,429	0,356
B	0,5055	10,200	3, 24	0,00016*	0,672	0,496	0,411
AB	0,0000	0,708	6, 24	0,64603	1,163	0,858	0,712
Частные средние	3,347	11, 24	0,00645*		1,163	0,858	0,712

Стандартная Ошибка = 0,2941 (9,21% от общего среднего)

Нет оснований для отклонения 0-гипотезы относительно действия фактора A (F-критерий меньше единицы, следовательно все средние вариантов фактора A не различаются. F-критерий для фактора B позволяет отклонить 0-гипотезу на очень высоком уровне значимости (P<0,001), имеются достоверные различия средних фактора B. F-критерий для частных средних также говорит о наличии достоверно различающихся средних в ячейках плана:

Варианты / Фактор "A"	Фактор "B" 1	2	3	4	Средние	Разница	Значима?
1	2,500	3,767	2,833	3,100	3,050	Контроль	
2	2,800	4,167	3,433	2,967	3,342	0,292	Нет
3	2,433	3,700	3,600	3,033	3,192	0,142	Нет
Средние	2,578	3,878	3,289	3,033	3,1944	0,144	Нет
Разница	Контр.	1,30	0,71	0,46	0,617		
Значима?		Да!	Да!	Нет	Да!		

Для анализа различий средних в качестве контрольного варианта автоматически предлагается первый вариант; если же в действительности в опыте контрольным был другой вариант, его номер нужно ввести в окне установок параметров анализа (перед выводом на дисплей). В случае, если в массиве имеются выпавшие значения, для анализа различия средних НСР вычисляется для конкретной пары средних – с учетом действительного числа дат в вариантах.

Здесь же можно указать на необходимость объединения дисперсии взаимодействия с дисперсией от случайных факторов, если a priori известно, что взаимодействие факторов отсутствует. Если о возможности взаимодействия нет данных, дисперсия взаимодействия может быть объединена с дисперсией от случайных факторов в случае подтверждения 0-гипотезы (отсутствие взаимодействия). Тогда несколько изменятся значения F-критериев для главных эффектов. Это рекомендуется делать при значении вероятности ошибки более 0.5.

Наличие взаимодействия факторов можно анализировать графически: линии на графике должны быть расходящимися в случае взаимодействия, или более-менее параллельными при отсутствии эффекта взаимодействия.

5.2.1. Методы 2-факторного анализа

В программе D2MAXI реализованы несколько методов параметрического дисперсионного анализа экспериментальных данных из опытов с повторениями:

1/ с равным числом повторений (полная рандомизация, рандомизация в блоках);

2/ с неравным числом повторений (вплоть до одного повторения в некоторых точках плана);

3/ анализ экспериментов типа "расщепленные делянки" (равное число повторений, рандомизация в блоках);

4/ анализ иерархических планов;

5/ анализ многолетних однофакторных экспериментов по Томилову.

Для каждого из этих методов предназначены свои блоки вычислений и формы выдачи, отражающие специфику метода. Например, для теста однородности дисперсий используется критерий Кокрена для данных с равным числом повторений, и критерий Бартлета в случае неравночисленных данных. Также различны методы восстановления "выпавших данных" – в соответствии с моделью анализа данных.

При выборе метода анализа в соответствующем диалоговом окне (с помощью мышки) программа может автоматически менять тип рандомизации.

Анализ стандартных перекрестных планов с равным числом повторений. Предполагается, что уровни факторов в эксперименте фиксированы (модель дисперсионного анализа типа "Fixed"), или случайны (модели дисперсионного анализа типа "Random" или "Mixed"). Для восстановления выпавших данных используется итерационный алгоритм по Снедекору [20, стр. 294-295]. Для проверки программы использовались тестовые примеры из руководств Снедекора, Хикса, Доспехова и других книг. Анализ различия средних вариантов фактора фиксированного типа в случае модели "Mixed" выполняется на базе остаточного среднего квадрата – как и в случае двух факторов типа "Fixed", хотя это, по-видимому, не совсем корректно.

Анализ опытов с расщеплением: более крупные экспериментальные объекты – повторения фактора "A" состоят из более мелких экспериментальных объ-

ектов – полного набора вариантов фактора "B". Эта ситуация обычно имеет место в экспериментах с удобрениями, сортами растений и т.п., когда большие делянки (повторности фактора "A") разбиваются на несколько мелких делянок – по числу вариантов фактора "B". Предполагается, что опыт организован методом случайных блоков (равное число повторностей), а уровни факторов в эксперименте фиксированы, то есть используется модель дисперсионного анализа типа "Fixed". Корректность работы программы по этому методу проверялась по [20, стр. 344] и [11, стр. 256].

Для "восстановления" отсутствующих данных используется метод Йетса [41, стр. 33], с учетом модели данных (рандомизация в блоках).

Анализ 2-факторных иерархических экспериментов. Фактор "A" должен быть фиксированного типа, а фактор "B" – любого (Fixed или Random), с равным числом повторностей (не менее 2-х) в каждой точке плана. В отличие от стандартной схемы перекрестного факторного эксперимента, в иерархических планах (или планах с группировкой) каждый вариант фактора "A" содержит собственный набор вариантов фактора "B". Следствием этого является невозможность выявления взаимодействия факторов.

Анализ 1-факторных многолетних экспериментальных данных методом Томилова. В основе этого метода – вычисление критерия Фишера-Снедекора для проверки действия фактора на протяжении нескольких лет по формулам, изложенным в [36]. Ежегодные 1-факторные опыты должны быть с равным числом повторностей; массив данных должен быть введен следующим образом: фактор “A” – годы, “B” – исследуемый фактор (сорта, удобрения, методы обработки и т.п.). В программе выполняется анализ различий среднемноголетних сравнением с Наименьшей Существенной Разницей (на базе критерия Стьюдента), вычисляемого также по формулам Томилова. Предполагается, что уровни факторов в эксперименте фиксированы, то есть используется модель дисперсионного анализа типа "Fixed", по типу полной рандомизации.

0-гипотеза для факторов фиксированного типа формулируется следующим образом: отсутствует действие изучаемого фактора, средние имеют разные значения вследствие действия случайных факторов.

0-гипотеза для фактора случайного типа: дисперсии вариантов фактора одинаковы, фактор не влияет на изменчивость (вариабельность) изучаемой системы. Контр-гипотеза: по меньшей мере одна пара вариантов фактора имеет различные дисперсии. Следует серьезно отнестись к заданию типа фактора B, так

как от этого выбора зависит способ вычисления критерия Фишера-Снедекора для фактора А.

Частные средние можно анализировать с помощью критерия НСР, который вычисляется с учетом модели данных, и, таким образом, отличается от НСР стандартных 2-факторных анализов.

Программа тестировалась по [33], стр.49, тестовый массив в файле JOHNS49.dat, [4], стр. 264, файл AFIFI264.dat, [34], стр. 230, файл HICKS230.dat. Для восстановления выпавших данных используется метод Йетса (по Снедекору [20], стр. 294-295).

5.2.2. Анализ различия средних

После доказательства действия фактора (типа Fixed) приступают к анализу различия средних. Достоверность различия средних в сравнении с контрольным вариантом может быть проверена любым из четырёх критериев – Шеффе, Тьюки, T(n), Стьюдента. Число средних (вариантов) фактора не должно превышать 20 [ограничение, связанное с размерами таблиц критериев Тьюки и T(n)]; при выборе критериев Шеффе и Стьюдента число вариантов может быть больше.

Критерий Тьюки предлагается программой по умолчанию, этот критерий корректно работает во всех практических ситуациях, и является серьёзной альтернативой критерия Стьюдента. Применение НСР на базе критерия Стьюдента – общепринятая практика Российских естествоиспытателей, однако следует помнить, что это оправдано только при малом числе вариантов и при больших значениях критерия Фишера-Снедекора, с вероятностью ошибки 1-го рода порядка 0,001 и менее.

Средние каждого фактора вычисляются на основе фактически имеюшихся значений (исключаются пропуски, даты для восстановления), затем анализируются сравнением с величиной Наименьшей Существенной Разницы (НСР, или НЗР, Наименьшего Значимого Различия; в англоязычной литературе LSD – Least Significant Differens) на выбранном уровне значимости. Если разница средних (по абсолютной величине) превысила НСР на уровне 1%, программа печатает "**", на уровне 5% – программа печатает "*".

Все критерии сравнения средних, используемые в программе, полностью эквивалентны для случая эксперимента из 2-х вариантов; как только число вариантов увеличивается, достоверность выявления ДЕЙСТВИТЕЛЬНО различающихся пар средних может падать.

Самым жестким критерием достоверности различия средних является критерий Шеффе; менее категоричными, но достаточно строгими в теоретическом плане являются критерии Тьюки и T(n); критерий Стьюдента может привести к завышенному числу "достоверно" различающихся пар средних при большем числе вариантов. Известный критерий Дункана основан на эмпирическом правиле выявления достоверной разницы, и не может считаться строгим критерием для определения достоверных различий (однако, имеется в программе COMPAR). Формулы для анализа средних взяты из [21], стр. 30-34.

Рекомендуем пользоваться для анализа средних, формируя НСР на базе критерия Тьюки.

5.3. DISLAT: Дисперсионный анализ “Латинских квадратов”

Программа DISLAT предназначена для дисперсионного анализа экспериментальных данных, полученных из специально организованного опыта по схеме "Латинский квадрат", который может быть одно-, двух- или трехфакторным. В программе выполняется также анализ различий средних по критерию НСР (по Т-критерию Стьюдента). Предполагается, что основной изучаемый фактор "шифруется" латинской буквой (1-й вариант – A, 2-й вариант – B, и т.д.); варианты второго фактора в двухфакторном эксперименте располагаются в строках квадрата, и, соответственно, варианты третьего фактора в трехфакторном плане размещаются в столбцах квадрата.

Модель дисперсионного анализа латинских квадратов:

$\mathbf{y_{ijk} = \mu + a_i + b_j + c_k + e_{ijk}}$; μ – генеральное среднее изучаемой системы;

a_i – эффект варианта фактора A типа Fixed;

b_j – эффект строк или вариантов фактора B типа Fixed;

c_k – эффект столбцов или вариантов фактора C типа Fixed;

e_{ijk} – ошибка от случайных факторов, распределена по $N(0, \sigma)$.

0-гипотезы: все $a_i=0$, все $b_j=0$, все $c_k=0$;

контр-гипотезы: некоторые $a_i=/=0$, $b_j=/=0$, $c_k=/=0$.

Ограничения на размер массива данных: не более 20 вариантов (20x20). Пример формирования массива 5x5 в текстовом файле:

```
5 5
35,3 31,1 32,6 33,4 33,8
40,8 33,7 39,3 37,7 37,3
35,8 27,7 37,2 31,8 35,8
34,2 35,3 36,9 40,0 33,9
32,2 33,7 26,4 33,7 31,2
DCABE
BAECD
EBDAC
ADCEB
CEBDA

Сорт 235
```

<- начало файла

массив данных:
строки = объекты,
столбцы = признаки

Схема размещения вариантов основного фактора

<- необязательный комментарий

В качестве примера можно посмотреть файл DISLAT.dat. Массивы данных, подготовленных для обработки методом 1-факторного дисперсионного анализа, в которых число повторностей равно числу вариантов, могут быть переданы программе DISLAT, но в этом случае необходимо дополнительно ввести массив "латинских букв".

Для анализа различий средних в качестве контрольного варианта автоматически предлагается первый вариант; если же в действительности в опыте контрольным был другой вариант, его номер нужно ввести в окне установок параметров анализа перед выводом на дисплей (массив DISLAT.dat):

Действие факторов, влияние по Снедекору.

Фактор	Степень влияния	Критерий Фишера	Степени свободы	Вероятность ошибки	Наим.Существ.Различие 1%	5%	10%
Латынь	0,3169	3,320	4, 12	0,0475*	4,305	3,071	2,512
Строки	0,5106	6,217	4, 12	0,0060*	4,305	3,071	2,512
Столбцы	0,1264	1,723	4, 12	0,2094	4,305	3,071	2,512

Средние фактора "латинская буква"

Вариант	Средние	Разница	Значима?
A	32,70	Контроль	
B	32,44	-0,260	нет
C	34,74	2,040	нет
D	35,76	3,060	нет
E	36,52	3,820	да
Общее	34,43	1,732	нет

Стандартная ошибка = 0,9965 (2,89% от общего среднего)

Принимаются контр-гипотезы: факторы “Латинская буква” и “Строки” действуют, нет оснований для отклонения 0-гипотезы для фактора “Столбцы”. Среднее варианта “E” достоверно отличается от среднего варианта “A” фактора “Латинская буква” по НСР (по критерию Стьюдента).

Для проверки одной из предпосылок классического дисперсионного анализа – нормальности распределения остатков, программа может сформировать 1-мерный массив остатков в соответствии с моделью данных, который записывается в виде файла для обработки программой NORMAL.

Программа позволяет проверить еще одно предположение классического дисперсионного анализа – однородность дисперсий, для этого вычисляется критерий Кокрена. Статистика Кокрена проверяет 0-гипотезу: все дисперсии равны, контр-гипотеза: выборка с максимальным значением дисперсии – из другой генеральной совокупности.

5.4. DSPAN: Дисперсионный анализ опытов без повторений

Программа DSPAN предназначена для обработки экспериментальных данных, полученных в многофакторных опытах без повторностей, методом параметрического дисперсионного анализа. Предполагается, что уровни всех факторов в вариантах эксперимента фиксированы, то есть используется модель дисперсионного анализа типа I, например, математическая модель 3-факторных данных:

$$y_{ijk} = \mu + a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk} + e_{ijk};$$

μ – генеральное среднее изучаемой системы;

a_i – эффект варианта фактора A типа Fixed;

b_j – эффект вариантов фактора B типа Fixed;

c_k – эффект вариантов фактора C типа Fixed;

ab_{ij}, ac_{ik}, bc_{jk} – эффекты взаимодействия факторов;

e_{ijk} – ошибка от случайных факторов, распределена по $N(0, \sigma)$.

0-гипотезы: все $a_i=0$, все $b_j=0$, все $c_k=0$, все $ab_{ij}=0$, все $ac_{ik}=0$, все $bc_{jk}=0$;

контр-гипотезы: некоторые $a_i=/=0$, $b_j=/=0$, $c_k=/=0$, $ab_{ij}=/=0$, $ac_{ik}=/=0$, $bc_{jk}=/=0$.

Помимо собственно дисперсионного анализа, выполняется анализ достоверности различия факторных средних по Т-критерию Стьюдента в форме НСР (Наименьшей Существенной Разницы).

Помимо обычных предположений классического дисперсионного анализа, для данного метода анализа постулируется ***отсутствие взаимодействия высшего порядка*** (всех факторов вместе взятых). В противном случае возможны ошибочные выводы вследствие смещения значений критерия Фишера-Снедекора.

Данные в виде двумерного массива " варианты-повторения " могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла.

Ограничения на размер массива данных: число строк – не более 8000, максимальное количество вариантов в любом факторе – 100, максимальное количество факторов – 8; общий размер массива – не более 16000 элементов. Пример формирования массива из 3-х факторов: 2 варианта фактора "A", 3 варианта фактора "B" и 4 варианта фактора "C" в текстовом файле:

```
4 6 2 3                                <= начало файла 4 столбца, 6 строк;
12,3 12,5 14,2 13,1    1a1b      2 = вариантов фактора "A",
8,34 13,7 11,6 11,9    1a2b      3 = вариантов фактора "B";
13,3 14,6 11,0 15,3    1a3b
7,12 9,08 10,3 11,5    2a1b    строки = варианты "A" x "B";
8,27 9,56 7,33 10,3    2a2b    столбцы = варианты фактора "C";
6,21 8,44 7,25 5,29    2a3b

Данные Петрова         <= необязательный комментарий
  c1   c2   c3   c4    <= варианты фактора "C"
```

В качестве примера формирования массива для программы DSPAN можно посмотреть файлы DOSP250.dat, DIS3_F.dat. Массивы данных, подготовленных для обработки методами дисперсионного анализа с повторениями, могут быть переданы программе DSPAN и обработаны как данные без повторений, считая повторения как дополнительный фактор.

Отсутствующие даты должны кодироваться как "-1". Допустимо не более 5% выпавших значений, в противном случае возможо получение некорректных результатов. Используется подстановка среднего по вариантам всех факторов для ячейки с отсутствующим значением, с соответствующим уменьшением числа степеней свободы для остаточного среднего квадрата.

Наличие взаимодействия можно анализировать графически: линии на графике должны быть расходящимися в случае взаимодействия, или более-менее параллельными при отсутствии эффекта взаимодействия.

Для анализа различий средних в качестве контрольного варианта автоматически предлагается первый вариант; если же в действительности в опыте контрольным был другой вариант, его номер нужно ввести в окне установок параметров анализа (перед выводом на дисплей). В случае, если в массиве имеются выпавшие значения, для анализа различия средних НСР вычисляется для конкретной пары средних – с учетом действительного числа дат в вариантах (файл DOSP303.dat):

1. Дисперсионный анализ. Таблица ANOVA

Источник вариации	Сумма квадратов	Доля вариации	Степени свободы	Средний квадрат	F-критерий
Общая	10354,200	1.0000	19	544,9579	
Фактор "A"	5324,700	0,5143	4	1331,1750	9,1150
Фактор "B"	3277,000	0,3165	3	1092,3333	7,4796
Ошибка(+AB)	1752,500	0,1693	12	146,0417	

2. Анализ различия факторных средних.

Фактор<A>	Фактор-<B> 1	2	3	4	Средние	Различия
1	91,00	102,0	94,00	126,0	103,3	Контроль
2	112,0	104,0	148,0	142,0	126,5	23,25 *
3	134,0	115,0	158,0	167,0	143,5	40,25 *
4	122,0	148,0	144,0	166,0	145,0	41,75 *
5	145,0	134,0	148,0	154,0	145,3	42,00 *
Средние	120,8	120,6	138,4	151,0	132,7	
Различия	Контр.	-0,20	17,6*	30,2*		

Стандартная ошибка опыта = 12,085, (9,11% от общего среднего)

3. Анализ действия факторов, влияние по Снедекору.

Фактор	Степень влияния	Критерий Фишера	Степени свободы	Вероятность ошибки	НСР(1%)	НСР(5%)	НСР(10%)
A	0,4691	9,115	4, 12	0,0013*	26,101	18,618	15,230
B	0,2997	7,480	3, 12	0,0044*	23,346	16,653	13,622

Доказано действие обоих факторов с высоким уровнем значимости (P<0,01), средние вариантов отличаются от средних контрольных вариантов.

Программа тестировалась числовыми данными из [22], массив ZAJC300.dat.

5.5. DIS8: Многофакторный дисперсионный анализ опытов с повторениями

Программа DIS8 предназначена для обработки экспериментальных данных, полученных в многофакторных опытах, различными видами параметрического дисперсионного анализа:

1/ Стандартный метод анализа многофакторных планов с **полной рандомизацией** вариантов/повторений с фиксированными уровнями вариантов факторов (тип "Fixed"), или с **рандомизацией вариантов в блоках повторений** (метод "случайных блоков");

2/ Анализ многофакторных планов с **расщеплением вариантов факторов**: более крупные экспериментальные объекты (начиная с вариантов фактора "A")

состоят из более мелких экспериментальных объектов – полного набора вариантов следующего фактора ("B"), и так далее до последнего фактора. Эта ситуация обычно имеет место в экспериментах с удобрениями, сортами растений и т.п., когда большие делянки разбиваются на несколько мелких делянок – по числу вариантов субфактора. Предполагается, что опыт организован методом случайных блоков, а уровни факторов в эксперименте фиксированы, то есть используется модель дисперсионного анализа типа I. Вычислительная часть программы основана на алгоритме А.И.Южакова (Сибирский НИИ земледелия и химизации).

3/ Анализ многофакторных планов **смешанного типа ("Mixed")**, в которых варианты первого фактора выбраны случайным образом (типа Random), варианты остальных факторов – фиксированного типа. Этому типу данных соответствуют многолетние факторные эксперименты в с.-х. исследованиях. Для более глубокого анализа данных дополнительно выполняется серия стандартных дисперсионных анализов для каждого варианта фактора типа Random. Вычислительная часть программы также базируется на алгоритме А.И.Южакова. В результатах анализа первый фактор помечен буквой "T", остальные – "B", "C" и так далее.

4/ Анализ многофакторных **иерархических “Nested”** планов. Специфика таких планов (nest = гнездо) заключается в том, что каждый уровень старшего фактора содержит собственный набор вариантов следующего фактора, каждый уровень которого, в свою очередь, также содержит собственный набор вариантов следующего фактора, и так далее. Следствием такой структуры данных является невозможность выявления взаимодействия факторов.

5/ Анализ экспериментальных данных, полученных в многофакторных опытах специального типа "**Repeated Measures**" ("повторные измерения"). Предполагается, что уровни всех факторов, кроме последнего, выбраны неслучайным образом ("Fixed"), а последний фактор – типа "Random". Таким образом, используется модель дисперсионного анализа смешанного типа ("Mixed"), причем в каждой точке плана имеется только одно значение (эксперимент без повторений в классическом смысле).

Помимо собственно дисперсионного анализа (вычисление критериев Фишера-Снедекора), выполняется анализ достоверности различия факторных средних по T-критерию Стьюдента в форме НСР (Наименьшей Существенной

Разницы, в иностранных публикациях LSD, Least Significant Difference) на заданном уровне значимости (1, 5 или 10%). Следует помнить, что Т-критерий полностью корректен только при 2-х вариантах фактора, и может привести к ошибочным выводам при большем числе вариантов. Для строгого анализа достоверности различия факторных средних необходимо использовать программу COMPAR.

Данные в виде двумерного массива "варианты-повторения" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла.

Ограничения на размер массива данных: число строк – не более 8000, максимальное количество вариантов в любом факторе – 50, максимальное количество факторов – 8; общий размер массива – не более 32000 элементов. Пример формирования массива из 4-х повторностей, 2-х вариантов фактора "A" (или "T" для 3-го типа анализа), 3-х вариантов фактора "B" и 2-х вариантов фактора "C" в текстовом файле:

```
4 12 2 3 2                    <- начало файла
12,3 12,5 14,2 13,1           1a1b1c  факторы 2x3x2 = 12 вариантов
8,34 13,7 13,1 11,9           1a1b2c
13,3 14,6 11,0 15,3           1a2b1c  массив данных:
7,12 9,08 10,3 11,5           1a2b2c  строки = варианты
8,27 9,56 7,33 10,1           1a3b1c  столбцы = повторения.
6,21 8,44 7,25 5,29           1a3b2c
12,2 12,6 14,1 13,2           2a1b1c
8,35 13,8 13,2 11,8           2a1b2c
13,4 14,5 11,4 -1,0           2a2b1c  <- в 4-м повторении дата
7,13 9,03 10,2 11,4           2a2b2c      для "восстановления"
8,28 9,57 7,32 10,2           2a3b1c
6,22 8,45 7,24 5,24           2a3b2c

Данные Петрова, 1998         <- необязательный комментарий
```

В качестве примера формирования массива для программы DIS8 можно посмотреть файлы DOSP250.dat, DIS3_F.dat. Массивы данных, подготовленных для обработки методами 1-факторного дисперсионного анализа с повторениями, могут быть переданы программе DIS8 и обработаны 1-м методом. 2-й и 3-й методы предполагают число факторов два и более.

Отсутствующие даты должны кодироваться как "-1". Допустимо не более 5% выпавших значений, в противном случае возможо получение некорректных результатов. Используется подстановка среднего по вариантам всех факторов для ячейки с отсутствующим значением, или итерационный алгоритм вычисления этих значений, не изменяющий общую дисперсию массива данных, с соответст-

вующим уменьшением числа степеней свободы для остаточного среднего квадрата. Для 2-го метода анализа предполагается, что выпавшей является субделянка самого нижнего уровня расщепления.

0-гипотеза для факторов типа Fixed формулируется следующим образом: отсутствует действие изучаемого фактора, **средние вариантов** имеют разные значения вследствие действия неконтролируемых факторов.

0-гипотеза для фактора типа Random: отсутствует действие изучаемого фактора, **дисперсии вариантов** имеют разные значения только вследствие действия неконтролируемых факторов.

Наличие взаимодействия можно анализировать графически: линии на графике должны быть расходящимися в случае взаимодействия, или более-менее параллельными при отсутствии эффекта взаимодействия.

Анализ различия средних какого-либо фактора выполняется после доказательства его действия по F-критерию. НСР вычисляется по формуле:

$$\mathbf{HCP = T \times \sqrt{Er \times 2 / n}},$$ где

Er – остаточный средний квадрат из таблицы ANOVA;

T – табличное значение критерия Стьюдента на уровне значимости 1%, 5% или 10%, с числом степеней свободы остаточного среднего квадрата;

n – число повторений в вариантах анализируемого фактора;

Для анализа различий средних в качестве контрольного варианта автоматически предлагается первый вариант; если же в действительности в опыте контрольным был другой вариант, его номер нужно ввести в окне установок параметров анализа. В случае, если в массиве имеются выпавшие значения, для анализа различия средних НСР вычисляется для конкретной пары средних – с учетом действительного числа дат в вариантах.

5.5.1. Исключение некоторых незначимых эффектов

При обработке многофакторных данных тройные, четверные (иногда и парные) взаимодействия факторов обычно недостоверны, тем более, если взаимодействие принципиально невозможно по сути действия факторов. В этом случае их вклад можно считать случайным, объединить с дисперсией от случайных факторов, и скорректировать число степеней свободы ошибки, тем самым уточнив «Остаточный средний квадрат», на основе которого строится критерий Фишера-Снедекора и критерии НСР для анализа средних.

Эта операция может быть выполнена после стандартного дисперсионного анализа (полная рандомизация/рандомизация в блоках). С помощью мышки выбирается строка на вспомогательной таблице разложения дисперсий с недостоверным взаимодействием, и затем исключается также щелчком мышки. При этом в нижней части формы пересчитываются значения случайной дисперсии (SSE), числа степеней ошибки (DF) и остаточного среднего квадрата (SSE/DF), затем пересчитываются F-критерии всех основных факторов и оставшихся взаимодействий.

После исключения всех нежелаемых эффектов можно получить скорректированные результаты в обычном виде.

5.5.2. Стандартный перекрестный план

Стандартный метод анализа многофакторных планов, уровни всех факторов фиксированного типа "Fixed", равное число повторений:

– **полная рандомизация** вариантов/повторений, это типично для лабораторных экспериментов (вегетационные опыты), в которых несложно создавать одинаковые условия для всех объектов эксперимента;

– **рандомизация вариантов в блоках повторений** (метод "случайных блоков"), это обычно полевые деляночные эксперименты; формируя блоки повторений, стараются учесть возможный градиент какого-либо неконтролируемого фактора (близость к лесополосе, водоему, наличие склона и т.п.);

Если действительный план размещения объектов не подпадает под эти два способа, следует обрабатывать данные по типу полной рандомизации. Например, математическая модель 3-факторных данных этого типа:

$$y_{ijkl} = \mu + a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk} + abc_{ijk} + e_{ijkl};$$

μ – генеральное среднее изучаемой системы;

a_i – эффект варианта фактора A типа Fixed;

b_j – эффект вариантов фактора B типа Fixed;

c_k – эффект вариантов фактора C типа Fixed;

ab_{ij}, ac_{ik}, bc_{jk}, abc_{ijk} – эффекты взаимодействия факторов;

e_{ijkl} – ошибка от случайных факторов, распределена по $N(0, \sigma)$.

0-гипотезы: все $a_i=0$, все $b_j=0$, все $c_k=0$, все $ab_{ij}=0$, все $ac_{ik}=0$, все $bc_{jk}=0$, все $abc_{ijk}=0$

контр-гипотезы: некоторые $a_i \neq 0$, $b_j \neq 0$, $c_k \neq 0$, $ab_{ij} \neq 0$, $ac_{ik} \neq 0$, $bc_{jk} \neq 0$, $abc_{ijk} \neq 0$

Для восстановления выпавших данных используется итерационный алгоритм по Снедекору [20, стр. 294-295]. Для проверки программы использовались тестовые примеры из руководств Снедекора, Хикса, Доспехова и других книг (файл LAKIN234.dat, обработка по типу полной рандомизации):

2. Действие факторов.

Фактор	Сумма квадратов	Доля вариации	Средний квадрат	Критерий Фишера	Степ. своб.	Вероятность ошибки	НСР(5%)
B	7,939	0,50313	2,6463	10,200	3	0,00016	0,4956
A	0,511	0,03236	0,2553	0,984	2	0,38842	0,4292
AB	1,103	0,06989	0,1838	0,708	6	0,64603	0,8584

Степеней свободы знаменателя F-критерия = 24

3. Анализ факторных средних по НСР(5%)

Вариант	Число дат	Среднее	Разница	Достоверна?
Фактор A				
1	12	3,050	контроль	-
2	12	3,342	0,292	Нет
3	12	3,192	0,142	Нет
Фактор B				
1	9	2,578	контроль	-
2	9	3,878	1,300	Да
3	9	3,289	0,711	Да
4	9	3,033	0,456	Нет

Нет оснований для отклонения 0-гипотезы для фактора “A” (F-критерий меньше единицы), для фактора “B” принимается контр-гипотеза ($P<0,001$): фактор действует, средние 2-го и 3-го вариантов достоверно отличаются от контроля (1-го варианта) по НСР (критерию Стьюдента). Эффект взаимодействия отсутствует.

5.5.3. Многофакторный план с расщеплением вариантов

Анализ многофакторных планов с **расщеплением вариантов факторов**: более крупные экспериментальные объекты (начиная с вариантов фактора "A") состоят из более мелких экспериментальных объектов – полного набора вариантов следующего фактора ("B"), и так далее до последнего фактора. Эта ситуация обычно имеет место в экспериментах с удобрениями, сортами растений и т.п., когда большие делянки разбиваются на несколько мелких делянок – по числу вариантов субфактора. Предполагается, что все варианты имеют равное число повторений, опыт организован методом **случайных блоков**, а уровни факторов в эксперименте **фиксированы**, то есть используется модель дисперсионного ана-

лиза типа Fixed. Вычислительная часть программы основана на алгоритме А.И.Южакова (Сибирский НИИ земледелия и химизации).

Корректность работы программы по этому методу проверялась по [20, стр. 344] и [11, стр. 256]. Для восстановления отсутствующих данных используется метод Йетса [41, стр. 33], с учетом модели данных (рандомизация в блоках).

5.5.4. Многофакторные планы смешанного типа, "Mixed"

Анализ многофакторных планов **смешанного типа ("Mixed")**, в которых варианты первого фактора выбраны случайным образом (типа Random), варианты остальных факторов – фиксированного типа, с равным числом повторений. Этому типу данных соответствуют многолетние факторные эксперименты в с.-х. исследованиях. Для более глубокого анализа данных дополнительно выполняется серия стандартных дисперсионных анализов для каждого варианта фактора типа Random. Вычислительная часть программы также базируется на алгоритме А.И.Южакова. В результатах анализа первый фактор помечен буквой "T", остальные – "B", "C" и так далее.

Примеры факторов типа Random:

– годы эксперимента;

– пункты эксперимента (хозяйства, опытные станции);

– типы почв.

В соответствии с теорией дисперсионного анализа, действие Random-фактора проявляется в различии дисперсий вариантов, а не в различии средних. Пример обработки 3-факторного массива (J2_DIS8.dat), фактор T – Random типа, B, C – Fixed типа:

Источник вариации	Сумма квадратов	Средний квадрат	Критерий Фишера	Степени свободы		Вероятность P	Доля вариации	НСР(5%)
C	58,313	14,578	1,312	4	8	0,3437	0,0356	1,469
B	110,037	55,018	0,851	2	4	0,4920	0,0672	1,138
BC	49,488	6,186	1,514	8	16	0,2282	0,0302	2,544
T	565,998	282,999	85,518	2	133	0,0000*	0,3458	
TC	88,897	11,112	3,358	8	133	0,0015*	0,0543	
TB	258,519	64,630	19,530	4	133	0,0000*	0,1579	
TBC	65,375	4,086	1,235	16	133	0,2501	0,0399	
Остаток	440,129	3,309		133			0,2689	
Сумма	1636,757	9,144		179			1,0000	

Нет оснований для отклонения 0-гипотез относительно факторов B и C, средние вариантов этих факторов не различаются. 0-гипотеза для фактора T отклоняется на очень высоком уровне значимости (P<0,0001), дисперсии вариантов

фактора Т достоверно различаются. Достоверность взаимодействий ТВ и ТС должна быть отвергнута, поскольку не доказаны главные эффекты факторов (большие значения F-критериев для взаимодействий требуют дополнительного экспертного анализа – случайно это или есть некая закономерность).

5.5.5. Многофакторные иерархические планы, "Nested"

Анализ многофакторных **иерархических “Nested”** планов. Специфика таких планов (nest = гнездо) заключается в том, что каждый уровень старшего фактора содержит собственный набор вариантов следующего фактора, каждый уровень которого, в свою очередь, также содержит собственный набор вариантов следующего фактора, и так далее. Следствием такой структуры данных является невозможность выявления взаимодействия факторов. Такие данные типичны в животноводстве, например, анализируются несколько поколений быка-производителя: каждое поколение – фактор со своим собственным набором вариантов – потомством.

Алгоритм обработки данных, реализованный в программе, предполагает следующее:

– число факторов от 2 до 8;

– в каждом варианте текущего уровня содержится равное число вариантов следующего уровня;

– все факторы – типа "Fixed";

– число повторений в ячейках плана – не менее 2-х, допустимо **неравное** число повторений, в некоторых ячейках может быть только одно значение, но для такой точки плана не будет выполнен анализ различий средних.

Пример обработки 3-факторного массива (J2_DIS8.dat):

Вариация	Сумма квадратов	Доля вариации	Степени свободы	Средний квадрат	Критерий Фишера	вероятность
Общая	1817,076	1.00000	179	10,151		
Фактор A	524,210	0,28849	2	262,105	62,81	0,0000
Фактор B(A)	341,540	0,18796	6	56,923	13,64	0,0000
Фактор C(AB)	387,971	0,21351	36	10,777	2,583	0,0000
Случайная	563,355	0,31003	135	4,173		

Отклоняются 0-гипотезы для всех факторов на высоком уровне значимости, средние вариантов фактора А различаются достоверно, частные средние фактора В (в каждом варианте фактора А – свои средние вариантов В) различаются достоверно, частные средние фактора С (в каждом частном варианте фактора В – свои

средние вариантов С) различаются достоверно. После доказательства действия факторов приступают к анализу средних по НСР – какие конкретно пары средних различаются по критерию Стьюдента, или по другим, более жестким критериям.

5.5.6. Многофакторные планы "Repeated Measures"

Анализ экспериментальных данных, полученных в многофакторных опытах специального типа "**Repeated Measures**" ("повторные измерения"). Предполагается, что уровни всех факторов, кроме последнего, выбраны неслучайным образом ("Fixed"), а последний фактор – типа "Random". Таким образом, используется модель дисперсионного анализа смешанного типа ("Mixed"), причем в каждой точке плана имеется только одно значение (эксперимент без повторений в классическом смысле).

Для каждого изучаемого эффекта (действие факторов, взаимодействий) используется собственное значение ошибки (знаменателя F-отношения), определяемое дисперсией "взаимодействия" эффекта с фактором типа Random, и соответствующее ей число степеней свободы. Помимо классического дисперсионного анализа с вычислением критериев Фишера-Снедекора, выполняется анализ факторов типа "Fixed" – с вычислением поправок по Гринхаузу-Гейсеру (см. ниже).

Если варианты последнего фактора ("Random") представляют собой эксперименты, организованные последовательно во времени на одних и тех же объектах (делянках, растениях, животных и т.п.), или же это одномоментный эксперимент с вариантами фактора на подобных экспериментальных объектах, в этом случае может иметь место значительная коррелированность между вариантами этого фактора. При этом возможно получение ложных выводов относительно как действия факторов, так и их взаимодействия.

Если к тому же имеет место и неоднородность дисперсий в вариантах фактора, это формирует так называемую "несферичность" структуры данных, при которой стандартный F-критерий некорректен, так как дисперсионное отношение "фактор/ошибка" имеет другое распределение, отличное от распределения Фишера-Снедекора.

Тест сферичности каждого фактора выполняется с помощью критерия Hi^2; 0-гипотеза: дисперсии вариантов фактора однородны, парные корреляции между вариантами равны (или все нулевые).

Помимо этого теста, для каждого фактора вычисляется коэффициент Гринхауза-Гейсера [41]. Это коэффициент к степеням свободы числителя и зна-

менателя F-отношения, который уменьшает их значения и таким образом позволяет использовать стандартный F-критерий. Поправки к степеням свободы для анализа взаимодействий вычисляются как произведения коэффициентов Гринхауза-Гейсера. Степени свободы числителя обычно получаются дробными, поэтому используется полиномиальная аппроксимация по целым степеням свободы для вычисления вероятности для значения F-критерия. Если в массиве имеется небольшое число отсутствующих дат (не более 3-4), можно рекомендовать подставить вместо них средние по столбцу или по строке. При большем числе отсутствующих дат следует подставить вместо них "восстановленные" с помощью программы COMPAR значения.

5.6. WILSON: 1-,2-,3-факторный анализ данных по Уилсону

Программа WILSON предназначена для обработки экспериментальных данных, полученных в одно-, двух- или трехфакторных опытах **с равным или различным числом повторностей**, методом непараметрического анализа по Уилсону, являющимся аналогом дисперсионного анализа, но не требующим выполнения стандартных предпосылок классического анализа (нормальности распределения ошибок измерения и однородности дисперсий). В основе метода лежит анализ частот распределения данных в вариантах относительно медианы.

Данные в виде двумерного массива "варианты-повторения" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла.

Пример формирования массива из 4-х повторностей, 2-х вариантов фактора "A", 3-х вариантов фактора "B" и 2-х вариантов фактора "C" в текстовом файле:

```
4 12 2 3 2                 <- начало файла
12,3 12,5 14,2 13,1        1a1b1c  всего 2x3x2 = 12 вариантов
8,34 13,7 13,1 11,9        1a1b2c
13,3 14,6 11,0 15,3        1a2b1c  массив данных:
7,12 9,08 10,3 11,5        1a2b2c  строки = варианты
8,27 9,56 7,33 10,1        1a3b1c  столбцы = повторности.
6,21 8,44 7,25 5,29        1a3b2c
12,2 12,6 14,1 13,2        2a1b1c
8,35 13,8 -999 -999        2a1b2c   <= 2 повторности
13,4 14,5 11,4 12,0        2a2b1c
7,13 9,03 10,2 -999        2a2b2c   <= 3 повторности
8,28 9,57 7,32 10,2        2a3b1c
6,22 8,45 7,24 5,24        2a3b2c

Данные Петрова, 1992      <- необязательный комментарий
```

В качестве примера формирования массива для программы WILSON можно посмотреть файлы DOSP250.dat, DIS3_F.dat, DAN2.dat. Отсутствующие даты кодируются обычно значением -999 или иным, указанным в файле конфигурации CONFIG.sdc.

Основной результат работы программы – критерии Hi^2. 0-гипотеза для факторов формулируется следующим образом: отсутствует действие изучаемого фактора, средние имеют разные значения вследствие действия неконтролируемых случайных факторов.

Для Hi^2-критерия вычисляется вероятность ошибки в случае отклонения 0-гипотезы. Если

P<=0,01 действие фактора подтверждено на уровне 1%,

P<=0,05 действие фактора подтверждено на уровне 5%,

P>0,10 действие фактора не подтверждено.

0-гипотеза для проверки взаимодействия: факторы влияют на изучаемую систему как простая сумма воздействий, отсутствует эффект взаимоусиления (синергизм) или взаимоподавления (антагонизм) факторов. Hi^2-критерий для взаимодействия и вероятность трактуются аналогичным образом:

P<=0,01 взаимодействие факторов подтверждено на уровне 1%,

P<=0,05 взаимодействие факторов подтверждено на уровне 5%,

P > 0,10 взаимодействие факторов не подтверждено.

5.7. FRIDMAN: 1-факторный непараметрический анализ

Программа FRIDMAN предназначена для обработки экспериментальных данных непараметрическими аналогами дисперсионного анализа:

– методом Краскела – Уоллеса,

– методом Фридмана,

– методом анализа медиан,

– методом Уилсона,

– методом Джонкхиера – Терпстра.

В случае ввода массивов без пропусков имеется возможность сделать непараметрический анализ множественного различия вариантов по Уилкоксону – Уилкоксу.

Ограничения на размер массива данных: общее число вариантов – не более 1000, общий размер массива не более 16000 элементов.

Данные в виде двумерного массива "варианты-повторения" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х повторностей и 6-и вариантов в текстовом файле:

```
4 6                          <= начало файла
12,3 12,5 14,2 13,4
8,34 9,37 8,25 8,89          массив данных:
9,23 10,6 11,2 10,5              строки = варианты,
7,12 6,39 8,27 7,55              столбцы = повторности.
8,27 19,4 32,4 25,8
6,21 8,15 9,66 7,21

Вес листьев                  <= необязательный комментарий
```

В качестве примеров формирования массива для программ, обрабатывающих массивы "варианты-повторности" можно посмотреть файлы D1MAXI.dat (3 варианта, 4 повторности), PEREG233.dat, LAKIN222.dat Массивы данных, подготовленных для обработки программами стандартного дисперсионного анализа большей факторности, могут быть переданы программе FRIDMAN и обработаны как однофакторные данные.

Существует возможность обрабатывать массивы данных с неравным числом повторений (методом Краскела-Уоллеса или методом медиан). В этом случае число повторений определяется по значениям "-999" при анализе строки:

```
4 6                          <= начало файла
12,3 12,5 14,2 13,4          4 повт.
8,34 9,37 8,25 -999          3 повт.   Массив данных:
9,23 -999 -999 10,5          2 повт.       Строки = варианты,
7,12 6,39 8,27 7,55          4 повт.       Столбцы = повторности.
-999 19,4 32,4 25,8          3 повт.
6,21 8,15 9,66 7,21          4 повт.

Вес листьев                  <= необязательный комментарий
```

В качестве примера формирования массива с неравным числом повторений можно использовать файл DAN1.dat. Программа автоматически распознает тип массива, обрабатывая такие данные по Краскелу-Уоллесу.

Для всех методов анализа 0-гипотеза формулируется следующим образом: отсутствует действие изучаемого фактора, варианты различаются только вследствие действия множества случайных факторов. Для критерия Hi^2 вычисляется вероятность ошибки в случае отклонения 0-гипотезы. Если

P<=0,01 действие фактора подтверждено на уровне 1%,

P<=0,05 действие фактора подтверждено на уровне 5%,

P > 0,10 действие фактора не подтверждено.

5.7.1. Анализ данных по Краскелу – Уоллесу

Метод Краскела-Уоллеса – аналог однофакторного дисперсионного анализа (с равным или неравным числом повторностей), не требует выполнения предпосылок классического анализа (нормальности распределения отклонений от средних, однородности дисперсий и др.) [5], стр. 131. Действие фактора проверяется на основе H-критерия с последующим сравнением с табличными значениями, либо по аппроксимации критерием Hi-квадрат.

Результат работы программы – H-критерий Краскела – Уоллеса.

В качестве дополнительного критерия можно использовать эффективную аппроксимацию порядковой статистики Краскела – Уоллеса – критерий Имана-Давенпорта [5], стр. 11; его применение аналогично анализу с помощью критерия Hi-квадрат.

5.7.2. Анализ данных по Фридману

Метод Фридмана – аналог двухфакторного дисперсионного анализа без повторностей, или однофакторного с рандомизацией в блоках повторений, также не требующий выполнения предпосылок классического дисперсионного анализа [5], стр. 154. Действие факторов проверяется на основе статистики Фридмана с последующим сравнением с табличными значениями, либо по аппроксимации критерием Hi-квадрат.

Результат работы программы – критерий Фридмана или его аппроксимация критерием Hi-квадрат, который приемлем при числе вариантов больше шести. При меньших значениях необходимо пользоваться таблицами статистики Фридмана, имеющиеся в справочниках.

5.7.3. Анализ различия медиан

Метод анализа медиан аналогичен однофакторному дисперсионному анализу с равным или неравным числом повторностей, также не требует выполнения предпосылок стандартного дисперсионного анализа. Действие фактора доказывается не по различию средних, а по различию медиан в вариантах опыта.

5.7.4. Анализ данных по Уилсону

В основе непараметрического анализа данных по Уилсону лежит различие частот распределения данных в вариантах относительно медианы. Массив данных может быть с равным или различным числом повторностей в вариантах опыта.

Результат работы программы – критерий Hi-квадрат.

В составе пакета имеется программа Wilson для непараметрического анализа 2- и 3-факторных данных с повторениями.

5.7.5. Анализ данных по Джонкхиеру

Анализ по Джонкхиеру эффективен в тех случаях, когда предполагается монотонное возрастание действия фактора, варианты упорядочены в соответствии с этим возрастанием. Например, варианты с возрастающими дозами удобрения, пестицида, лекарства [5], стр. 136–140.

5.7.6. Множественный анализ различия вариантов по Уилкоксону – Уилкоксу

Проблема множественного анализа различия вариантов в случае подозрения на ненормальность распределения в экспериментальных данных может быть решена при отсутствии пропусков в массиве данных "варианты–повторения". Программа выполняет непараметрический анализ рангов по Уилкоксону – Уилкоксу [7], стр. 504-507. Наиболее близка структура массива данных методу анализа по Фридману (однофакторный анализ с рендомизацией в блоках повторений, или двухфакторный опыт без повторений).

Столбцы массива заменяются столбцами рангов, при наличии в столбце равных значений (троек или двоек) ранги в соответствующих позициях усредняются. Для каждого варианта вычисляется сумма рангов, и модуль разницы сумм пары вариантов сравнивается с табличным значением критерия на уровне значимости 5%.

5.8. TWOSAMP: Анализ различия 2-х выборок

Программа TwoSamp предназначена для углубленного анализа возможных различий двух выборок с помощью различных статистических методов; даты в этих двух выборках могут быть попарно связанными (2-й метод анализа). В общем случае (1-й метод анализа) допускается неравный размер выборок (до 2000

дат), отсутствующие значения могут быть в произвольных местах, при чтении массива данных из файла пропуски определяются по числу -999.0 (значение по умолчанию, может быть изменено в файле CONFIG.sdc).

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Для анализа используется любая пара столбцов загруженного массива; при необходимости массив данных может быть транспонирован для использования строк в качестве выборок. Пример данных:

```
3 10                    <- начало файла
12,3 23,45 4,567
8,34 22,89 5,233          массив данных:
9,23 21,22  -999          строки  =10 наблюдений ,
7,12 54,56 2,991          столбцы =3 выборки.
8,27 34,77 11,23
6,21 33,91 25,67          3-я выборка – из 8-и значений;
6,55 30,04  -999
5,90 31,25 23,36
7,34 29,81 22,13
8,93 28,17 21,95

Данные за 1998 г        <- необязательный комментарий
```

Результаты счета могут быть отредактированы непосредственно в среде программы при выводе результатов на дисплей (заголовок, комментарии, удаление ненужной информации и т.п.).

5.8.1. Анализ различия произвольных выборок.

Выборки могут считаться идентичными (взятыми из одной генеральной совокупности), если у них равны (в статистическом смысле) средние, дисперсии, асимметрии распределений. Но если хотя бы по одному из этих трех параметров равенство отсутствует, выборки должны считаться принадлежащими к различным генеральным совокупностям.

Равенство асимметрий распределений проверяется визуально анализом гистограмм выборок или графиков эмпирических распределений, а также анализом стандартных ошибок коэффициентов асимметрии.

Статистические тесты **не доказывают** равенство или неравенство параметров выборок, отклоняя или подтверждая 0-гипотезу на заданном уровне значимости – они лишь утверждают, что в большинстве аналогичных ситуаций ошибочный вывод будет сделан с некоторой малой вероятностью.

Исследователь обязан четко сформулировать, по каким параметрам он считает результаты экспериментов (выборки) идентичными или различающимися, а также зафиксировать (до эксперимента) уровень значимости для последующего анализа данных. Для большинства реальных экспериментальных данных обычно проверяется равенство (или неравенство) выборочных средних, однако стандартный критерий сравнения средних по Стьюденту базируется на предпосылках нормальности распределения ошибок измерения и равенства выборочных дисперсий. Если это не так, могут быть сделаны неправильные выводы.

В программе вычисляются различные статистические критерии для выборок произвольного размера, приводятся табличные значения критериев, взятых из различных источников, а для классических критериев типа Стьюдента, Фишера, Hi^2 вычисляется вероятность ошибки в случае отклонения 0-гипотезы:

если P> 0,05 0-гипотеза подтверждена данным критерием;

если P<=0,05 0-гипотеза отвергается, принимается контр-гипотеза.

Уровень значимости 0,05 (один ошибочный вывод на 20 подобных экспериментов) выбирается программой по умолчанию, его можно изменить перед выводом результатов.

Нормальность выборочных распределений может быть проверена для выборок значительного размера (программа NORMAL). В соответствии с ГОСТ, нормальность выборок менее 11 дат вообще не проверяется никакими критериями (есть однако, критерий Уилка-Шапиро).

1. Критерий Стьюдента. 0-гипотеза – средние выборок равны; предполагается непрерывность и нормальность распределения данных, равенство выборочных дисперсий. Может быть применен при умеренных отклонениях от этих предпосылок. Если дисперсии выборок apriori известны, вместо критерия Стьюдента используется критерий Гаусса (квантиль нормального распределения).

Если дисперсии выборок неизвестны и заведомо неравны (проблема Беренса-Фишера), используется критерий Пагуровой [53].

2. Критерий Фишера. 0-гипотеза – дисперсии выборок равны. Классический метод сравнения вариабельности данных в двух выборках с помощью дисперсионного отношения Фишера-Снедекора, для которого предполагается нормальность распределений выборок, средние которых известны. В качестве критерия используется отношение большей дисперсии к меньшей; критерий Фишера чувствителен к отклонениям от нормальности выборочных распределений.

3. Критерий Фишера в модификации Мардиа-Земроч [35, стр. 33-34]. Степени свободы дисперсионного отношения корректируются специальным образом, делая его более устойчивым к отклонениям от нормальности выборочных распределений. Степени свободы в этом случае – нецелые числа, для вычисления табличных значений используются аппроксимации по целым степеням свободы.

4. Робастный критерий Бокса-Андерсена. 0-гипотеза – средние выборок равны. Предполагается только одинаковая распределенность наблюдений (примерное равенство дисперсий и асимметрий) и независимость значений; критерий устойчив к отклонениям от нормальности распределений (по [35], стр. 28-30). Анализ сводится к вычислению стандартного критерия Фишера-Снедекора, для которого корректируются специальным образом степени свободы, обычно в виде дробных чисел. Табличное значение вычисляется интерполяцией по целым степеням свободы числителя F-отношения.

5. Критерий Уэлча-Беренса-Фишера. 0-гипотеза – средние выборок равны. Этот критерий используется в случае явного неравенства выборочных дисперсий, то есть сравниваются выборки из заведомо различных генеральных совокупностей. Характеризуется в общем случае нецелым числом степеней свободы; эмпирическое значение сравнивается с табличным для выборок менее 20 значений, или аппроксимируется критерием Стьюдента с помощью линейной интерполяции по целым степеням свободы.

6. Критерий Манна-Уитни. 0-гипотеза – медианы выборок равны, распределения дат в выборках идентичны. Если по некоторым соображениям невозможно считать нормальными выборочные распределения (явная асимметрия, двугорбость и т.п.), следует использовать для сравнения выборок непараметрическую статистику ранговых сумм Манна-Уитни, предполагающую только непрерывность распределений и независимость значений между выборками. Из таблиц ([36], стр. 159) извлекаются два значения: если вычисленное значение критерия попадает в интервал, образуемый этими табличными значениями, 0-гипотеза подтверждается на заданном уровне значимости, и отвергается в ином случае. Для выборок размером более 20 дат для анализа используется эффективная аппроксимация критерия Уилкоксона по Иману [23], стр. 11.

7. Критерий Уилкоксона. С точки зрения прикладной статистики эквивалентен критерию Манна-Уитни, используется только для тех пользователей, которые не знают этого факта.

8. Критерий Ван дер Вардена. 0-гипотеза – средние выборок равны. Этот критерий является непараметрическим, используется "...в случае, когда функции распределения исследуемых совокупностей могут отличаться лишь параметром сдвига. Особенно полезен этот критерий, если обе совокупности нормальны или близки к нормальным" [15], стр. 95. Табличное значение критерия и вероятность ошибки в случае отклонения 0-гипотезы вычисляются на основе формул в [15], стр. 96.

9. Критерий серий Вальда-Вольфовица. 0-гипотеза – распределения дат в выборках идентичны. Несколько менее эффективный непараметрический метод двухвыборочного сравнения, также не предполагает нормальности данных и равенства дисперсий. При размерах выборок менее 21 дат используются табличные значения критерия, для больших выборок применяется преобразование критерия серий к нормальному распределению. "Критерий Вальда-Вольфовица является чувствительным по отношению к целому ряду различий, включая различия в медианах, мерах изменчивости и асимметрии" [36], стр. 82.

10. Медианный критерий. 0-гипотеза – медианы выборок равны. Предполагается непрерывность распределения данных. При (N1+N2)>19 ([23], т.2, стр. 127) в качестве статистики используется приближенный критерий Hi^2, при меньших размерах выборок – точный критерий Фишера из таблиц [36]; стр. 134-141. Медиана служит хорошим аналогом среднего для больших и средних выборок, особенно для симметричных распределений. Отклонение 0-гипотезы медианным критерием предоставляет дополнительные аргументы для доказательства различия выборок.

11. Критерий Смирнова. 0-гипотеза – распределения дат в выборках идентичны; предполагается взаимная независимость значений, непрерывность распределений. Используется третий критерий Смирнова, D(m,n) (максимум модулей разниц между эмпирическими распределениями). Критерий обнаруживает любые различия в выборках (смещение средних, неравенство дисперсий, асимметрий). Используются табличные значения ([15], стр. 350) при размерах выборок менее 21 значения, и приближение критерием Колмогорова при больших выборках.

12. Критерий Ансари-Брэдли. 0-гипотеза – дисперсии выборок равны. Свободный от распределения ранговый критерий для проверки различия в мерах рассеяния двух выборок произвольного размера. Обе выборки должны иметь извест-

ные и равные медианы; если различие медиан превышает 10%, выборки корректируются вычитанием выборочных медиан, но в этом случае критерий Ансари-Брэдли становится *асимптотически* свободным от распределения. Используются методы вычислений, изложенные в [5], стр. 101-108. Тестовый массив из этой книги – HOLL2x20.dat, еще один тестовый массив – TWO2x12.dat из книги И.Гайдышева [72]. В качестве критерия для теста равенства дисперсий используется приближение для больших выборок нормальным распределением.

13. Критерий Сигела-Тьюки-Уилкоксона. 0-гипотеза – дисперсии выборок равны. Свободный от распределения ранговый критерий для проверки различия в мерах рассеяния двух выборок произвольного размера [54], [55].

Если в отношении каких-либо данных критерии не проявляют единодушия, следует посмотреть график "Гистограммы выборок" и принять субъективное решение, исходя из характера данных.

Немаловажным этапом анализа является выбор контр-гипотезы, которая может быть односторонней и двусторонней. Если предполагается **неравенство** средних (дисперсий, ранговых сумм, медиан) – это двусторонняя контр-гипотеза, для нее используются более "строгие" табличные значения с половинным уровнем значимости (например, 0,05/2=0,025); если же исследователь из каких-либо соображений предполагает, что среднее (дисперсия, ранговая сумма, медиана) одной выборки **больше** среднего (дисперсии, ранговой суммы, медианы) другой выборки – это односторонняя контр-гипотеза, используются табличные значения с заданным уровнем значимости, 0,05.

Во время графического анализа выборок можно менять в некоторых пределах число интервалов разбиения (столбиков гистограммы) с помощью соответствующего пункта меню; график "эмпирические распределения" особенно информативен для анализа сходства/различия выборок.

Для тестирования программы использовались данные из различных руководств по статистике: AZOT.dat, ANDER299.dat, ARENS169.dat, LLOYD125.dat, DAI2x15.dat, AFIFI264.dat, RUNION76.dat, RUNION84.dat, RUNION85.dat, HOLL236.dat, их можно найти в директории \SNEDECOR\TEST. Рекомендуем использовать эти массивы для ознакомления с программой, выяснения чувствительности различных критериев.

5.8.2. Анализ выборок с попарно связанными данными

2-й метод анализа предназначен для корректной проверки гипотезы об отсутствии различий между двумя выборками одинакового размера, данные в которых попарно связаны между собой какой-либо внутренней характеристикой. Например, для группы однородных делянок собираются два ряда показателей (биомасса, или количество насекомых, агрохимические параметры и т.д.) – до обработки, и через некоторый промежуток времени после обработки. Необходимо выяснить, достоверен ли эффект воздействия на систему, или нет.

Стандартным методом анализа в этой ситуации является проверка различия средних по парному Т-критерию Стьюдента. В программе используется тест Стьюдента в качестве основного критерия; однако, так как в ряде случаев возможны нарушения некоторых предпосылок применения этого метода (дискретный характер данных, эксцесс или асимметрия выборочного распределения), дополнительно вычисляется W-критерий Уилкоксона, эффективность которого мало зависит от типа распределения. Например, имеются две связанные выборки:

1-явыборка	2-я выборка	Разница	Критерий Стьюдента
1,0	1,1	-0,1	
2,0	2,2	-0,2	T(выч,)=1,626
3,0	3,3	-0,3	
4,0	4,5	-0,5	T(таб,)=1,833
5,0	5,8	-0,8	(односторонний)
6,0	7,3	-1,3	
7,0	8,0	-1,0	T(таб,)=2,262
8,0	10,0	-2,0	(двусторонний)
9,0	20,0	-11,0	
10,0	9,9	+0,1	
5,50	7,21	-1,71	Средние

Налицо эффект обработки – только в одном случае из 10 значение меньше исходного, однако проверка по Т-критерию не дает значимой разницы (для односторонней и двусторонней контр-гипотез). Более того, даже если последнее значение во второй выборке было бы равно 11.0 – и в этом случае эффект обработки не доказывался бы парным Т-критерием на уровне значимости 5%. Такой результат – следствие анормальности распределения значений в третьем столбце, что можно доказать критерием Уилка-Шапиро.

В этой ситуации применение непараметрического W-критерия Уилкоксона для связанных выборок дает корректные результаты, которые в обычных случаях совпадают с выводами, полученными с помощью Т-критерия. Метод вычисления W-критерия в программе – на основе формул, изложенных в [21]. Ис-

пользуется точное значение статистики до N=100 с соответствующим значением вероятности ошибки в случае отклонения 0-гипотезы, и аппроксимация Имана-Давенпорта – J-критерий для больших выборок.

Ограничения на размер массива данных: число дат не должно превышать 2000. Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 2-х выборок по 7 значений в текстовом файле:

```
2 7
12,3 22,5
8,34 23,7
9,23 24,6
7,12 20,9
8,27 19,4
6,21 18,5
5,67 17,2

Опыт с калием
```

<- начало файла

массив данных:
строки = пары значений,
столбцы = выборки.

<-- необязательный комментарий

Если имеется файл с массивом данных типа "переменные-варианты" с числом переменных больше двух, программе можно указать любую пару переменных из этого массива для выполнения анализа. В качестве примера формирования массива можно посмотреть файл SSP6x30.dat (6 переменных, 30 вариантов).

0-гипотеза формулируется следующим образом: выборки взяты из одной генеральной совокупности, эффект изменений "до воздействия – после воздействия" отсутствует. Пользователь, исходя из своего представления об исследуемой системе, должен указать программе тип проверяемой контр-гипотезы. По умолчанию программа выполняет анализ по двустороннему тесту (DELTA = разница средних):

Двусторонний тест: DELTA=/=0

Односторонний тест: DELTA > 0 (или DELTA<0)

Если из общих соображений известно, как влияет обработка на измеряемый показатель, используют односторонний тест; если же эффект обработки неизвестен (может в сторону увеличения показателя, а может и в сторону снижения), то применяют двусторонний тест.

Так как критерий Уилкоксона не табулирован для выборок размером 2-5 дат, для таких выборок программа вычисляет только Т-критерий Стьюдента.

5.9. COMPAR: Анализ средних, преобразование данных

Программа COMPAR предназначена для:

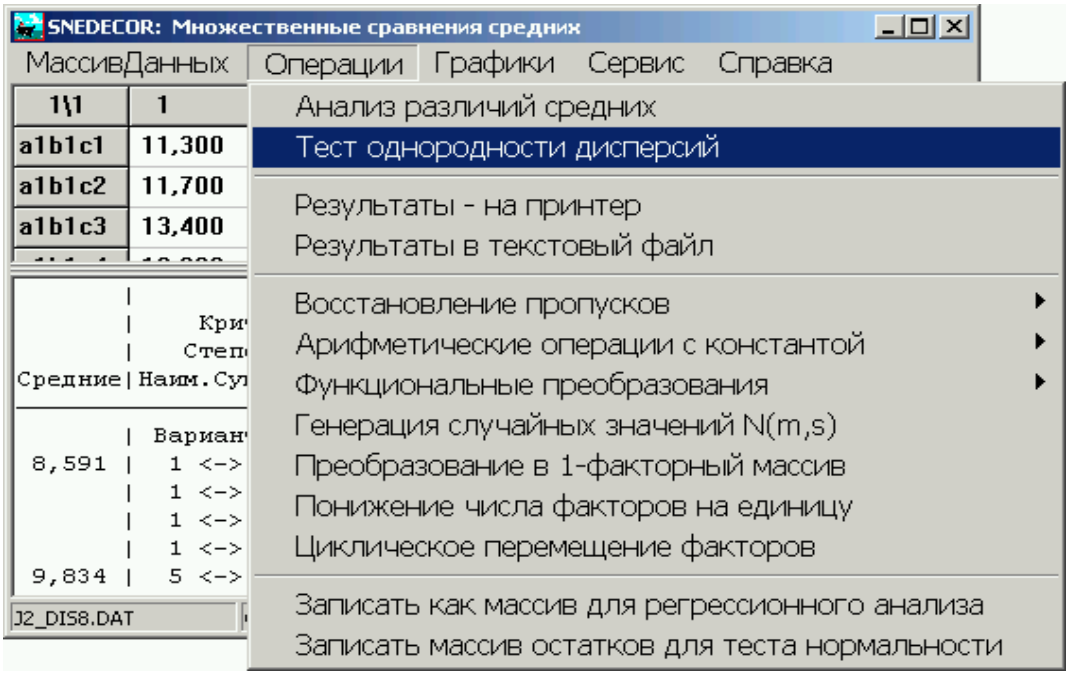
1/ Редактирования и преобразования массивов данных для последующей обработки программами одномерного многофакторного дисперсионного анализа (DIS8, DSPAN, D2MAXI, WILSON), другими программами пакета;

2/ Проверки одной из предпосылок классического дисперсионного анализа – **однородности дисперсий** – различными тестами;

3/ Обработки данных, полученных из факторных экспериментов, методом **множественного сравнения средних**; достоверность различия любой пары средних проверяется на основе нескольких критериев – Стьюдента, Шеффе, Тьюки, Дункана, T(n). Число средних (вариантов) фактора не должно превышать 20 (ограничение, связанное с размерами таблиц критериев Тьюки, Дункана и T(n)). Анализ различия средних основан на предположениях нормальности распределения отклонений от средних и однородности дисперсий.

4/ Множественного сравнения средних на основе **непараметрических критериев** – Даннета, Ньюмена-Кейлса, Данна (последний – наиболее строгий).

5/ Формирования массива данных, предназначенного для **множественного линейного регрессионного анализа** – на основе массива «варианты-повторения» обычного формата для дисперсионного анализа:



Программе можно передавать для обработки массивы "варианты-повторности", подготовленные для программ дисперсионного анализа различной факторности (до 8 факторов) **с равным числом повторений**, а также массивы

данных из факторных экспериментов **без повторений**. При этом программа вычисляет средние для каждого фактора и остаточный средний квадрат; для массива "факторы+повторения" пользователем может быть указан тип рандомизации повторностей (по умолчанию в программе устанавливается тип "полной рандомизации"). Остаточный средний квадрат (оценка ошибки эксперимента) вычисляется, исходя из модели данных "все факторы типа Fixed". В массиве могут быть "выпавшие" данные, введенные как "-1", при вычислении среднего квадрата программа подставит вместо них некоторые "восстановленные" значения в соответствии с типом эксперимента.

Ограничения на размер массива данных: число строк – не более 8000, максимальное количество вариантов в любом факторе – 50, максимальное количество факторов – 8; общий размер массива – не более 16000 элементов.

Данные в виде двумерного массива "варианты-повторения" могут быть загружены из текстового файла, либо переданы через буфер Windows. Пример формирования массива из 5-и повторностей, 2-х вариантов фактора "A" (или "T" для 3-го типа анализа), 3-х вариантов фактора "B" и 2-х вариантов фактора "C" в текстовом файле:

```
5 12 2 3 2                          <- начало файла
12,3 12,5 14,2 13,1 12,9            1a1b1c   всего 2x3x2 = 12 вариантов
8,34 13,7 13,1 11,9 12,4            1a1b2c
13,3 14,6 11,0 15,3 14,2            1a2b1c   массив данных:
7,12 9,08 10,3 11,5 10,7            1a2b2c      строки  = варианты,
8,27 9,56 7,33 10,1 8,26            1a3b1c      столбцы  = повторности
6,21 8,44 7,25 5,29 5,22            1a3b2c
12,2 12,6 14,1 13,2 12,8            2a1b1c
8,35 13,8 13,2 11,8 12,6            2a1b2c
13,4 14,5 11,1  -1  14,3            2a2b1c    <- в 4-й повторности дату
7,13 9,03 10,2 11,4 10,8            2a2b2c    необходимо "восстановить"
8,28 9,57 7,32 10,2 8,28            2a3b1c
6,22 8,45 7,24 5,24 5,23            2a3b2c

Данные Петрова, 1992 г              <- необязательный комментарий
```

Отсутствующие даты должны кодироваться как "-1". Допустимо не более 5% выпавших значений, в противном случае возможо получение некорректных результатов. Массив данных можно распечатать на принтере, при этом вычисляются частные средние вариантов.

5.9.1. Множественное сравнение средних

Средние каждого фактора вычисляются на основе фактически имеюшихся значений (исключаются пропуски, даты для восстановления), затем

упорядочиваются по возрастанию. Далее, по мере увеличения последовательных разниц средних, на основе сравнения с Наименьшим Существенным Различием (НСР, в англоязычной литературе LSD – Least Significant Differens) на уровне значимости 5% для каждого критерия, программа печатает "*" при превышении этого значения. Если разница средних превысила НСР на уровне 1%, программа печатает "**".

Все критерии сравнения средних, используемые в программе, полностью эквивалентны для случая эксперимента из 2-х вариантов; как только число вариантов увеличивается, достоверность выявления ДЕЙСТВИТЕЛЬНО различающихся пар средних может падать.

Самым жестким критерием достоверности различия средних является критерий Шеффе (в форме НСР):

$$\mathbf{HCP}_{0.05} = \sqrt{(\mathbf{v}-1)\times \mathbf{F}(\mathbf{v}-1, \mathbf{de}, 0.05)\times \mathbf{Er}\times 2/\mathbf{r}}$$

v – число вариантов фактора,

F(v-1, de, 0.05) – табличное значение критерия Фишера-Снедекора,

de – степней свободы ошибки из таблицы ANOVA,

Er – остаточный средний квадрат из таблицы ANOVA,

r – число повторений в каждом из сравниваемых вариантов.

Менее категоричными, но достаточно строгими в теоретическом плане являются критерии Тьюки и T(n):

$$\mathbf{HCP}_{0.05} = \mathbf{Tu}(\mathbf{v}, \mathbf{de}, 0.05)\times \sqrt{\mathbf{Er}/\mathbf{r}}$$

$$\mathbf{HCP}_{0.05} = \mathbf{Tn}(\mathbf{v}-1, \mathbf{de}, 0.05)\times \sqrt{\mathbf{Er}\times 2/\mathbf{r}}$$

Tu(..) и Tn(..) – табличные значения критериев Тьюки и T(n).

Критерий Стьюдента может привести к завышенному числу "достоверно" различающихся пар средних при большем числе вариантов:

$$\mathbf{HCP}_{0.05} = \mathbf{S}(\mathbf{v}-1, 0.05)\times \sqrt{\mathbf{Er}\times 2/\mathbf{r}}$$

S(..) – табличное значение критерия Стьюдента.

Критерий Дункана основан на эмпирическом правиле выявления достоверной разницы, и не может считаться строгим критерием для определения достоверных различий. Формулы для анализа средних взяты из [21, стр. 30-34]. Пример обработки данных (файл COMPAR.dat, 11 вариантов, часть распечатки):

Метод	Стьюдент	Шеффе	Тьюки	T(n)	Дункан
Критерий Q=5%	1,997	1,977	4,648	2,891	2,802 - 3,338
Степени свободы	66	10 66	11 66	20 66	2..11 66

Средние	Наим.Сущест.Разница	54,06	120,4	88,99	78,28	53,65 - 63,91

	Варианты	Delta					
190,0	9 <-> 5	60,00	*	-	-	-	*
	9 <->11	62,86	*	-	-	-	*
	9 <-> 1	65,71	*	-	-	-	*
	9 <-> 3	74,29	**	-	-	-	*
	9 <-> 6	87,14	**	-	-	*	**
	9 <-> 4	101,4	**	-	*	**	**
	9 <-> 7	124,3	**	*	**	**	**
	9 <-> 2	157,1	**	**	**	**	**
	9 <-> 8	157,1	**	**	**	**	**
	9 <->10	205,7	**	**	**	**	**

Для упрощенного анализа различий средних приводятся таблицы, в которых одинаковыми латинскими буквами отмечаются варианты, не различающиеся на уровне 5% внутри таких групп, и, соответственно, вариант из такой группы с наибольшим средним достоверно отличается от любого меньшего среднего, не помеченного данной буквой. Например, НСР(5%)=0,43:

1	5.55	
2	6.01	DEF
3	6.34	DE
4	6.43	D
5	7.45	ABC
6	7.66	AB
7	7.86	A

Средние вариантов 5, 6 и 7 не различаются достоверно, но варианты 1..4 достоверно отличаются от 7-го, 6-го или 5-го вариантов; варианты 2, 3 и 4 не различаются достоверно, но любой из этих вариантов достоверно отличается от 1-го варианта.

Рекомендуем пользоваться для анализа средних этим методом, формируя НСР на базе критерия Тьюки, так как стандартный НСР на базе критерия Стьюдента будет давать завышенное число достоверно различающихся пар вариантов, которые на самом деле статистически неразличимы.

На диаграммах средних формируются доверительные интервалы в виде: среднее ±НСР5%.

5.9.2. Редактирование и преобразование массивов данных

Специальная операция "восстановления" отсутствующих по какой-либо причине дат может быть выполнена одним из 3-х способов:

– заменой пропущенных значений на среднее в столбце; в этом случае неизбежно уменьшение дисперсии данных, поэтому этот метод следует использовать только при малом числе выпавших дат (1-3 в столбце);

– заменой пропусков на среднее в строке; этот способ используется чаще всего, однако он также неизбежно приводит к уменьшению общей и факторной дисперсии;

– заменой на некоторые числа, генерируемые датчиком случайных чисел, распределенных по нормальному закону со средним и сигмой, оцененным по имеющимся в массиве значениям; этот метод может быть неприемлем для данных с большой дисперсией из-за появления отрицательных значений.

Для этого отсутствующие значения нужно ввести как -1, либо они должны быть представлены пустыми ячейками в Табличном Редакторе (обычно это значения -999).

Обычно эту операцию нет необходимости выполнять, так как в большинстве программ дисперсионного анализа имеется автоматическое "восстановление" выпавших данных, специализированное для конкретного типа анализа.

Пункты Меню "Арифметические операции" и "Функциональные преобразования" предназначены для модифицирования всех значений массива с целью уменьшения дисперсий по вариантам факторов. Как известно, однородность дисперсий – одна из предпосылок классического дисперсионного анализа. В ряде программ пакета однородность дисперсий проверяется по критериям Кокрена или Бартлета, в программе COMPAR есть полный тест однородности дисперсий, включая анализ по Хартли. Если для обрабатываемого массива данных была обнаружена неоднородность дисперсий, нужно выполнить какое-либо стабилизирующее преобразование, и вновь сделать проверку. Ячейки со значением "-1" (данные для "восстановления") не преобразовываются, также как и отсутствующие значения (обычно -999).

В пункте Меню "Понижение факторности массива данных" возможно осуществление следующих операций:

1. Преобразование всего массива в однофакторный массив для какого-либо фактора (число повторностей умножается на число вариантов всех остальных факторов) – с формированием одного массива. Например, 3-факторный массив 3A x 4B x 5C x 2 повт. может быть преобразован в однофакторные массивы:

для фактора "A": 3 варианта, 40 повторностей;

для фактора "B": 4 варианта, 30 повторностей;

для фактора "C": 5 вариантов, 24 повторности.

Однофакторные массивы (после транспонирования) могут быть использованы другими программами, например:

VARS – для вычисления различных вариационных статистик,

NORMAL – для проверки нормальности,

IODATA – для проверки на аномальные значения, и т.д.

2. Вычленение нескольких массивов – по числу вариантов какого-либо фактора – с записью нескольких файлов. Массивы будут записываться в текущий каталог с добавлением 2-х цифр к имени исходного файла данных. Например, 3-факторный массив 3A x 4B x 5C x 2 повт. с именем Meta.dat может быть преобразован следующим образом:

для фактора "A": 3 массива 4B x 5C x 2 повт. –
Meta_01.dat, Meta_02.dat, Meta_03.dat;
для фактора "B": 4 массива 3A x 5B x 2 повт. –
Meta_01.dat, ..., Meta_04.dat;
для фактора "C": 5 массивов 3A x 4B x 2 повт. –
Meta_01.dat, ..., Meta_05.dat.

Такое преобразование позволяет сделать более тщательный анализ данных, например, выявить различия частных средних, которые не вычисляются в программах DIS8 и DSPAN.

Операция "циклический сдвиг факторов" выполняет преобразование массива данных, которое может быть полезным в некоторых ситуациях, например для того, чтобы фактор типа Random (остальные типа Fixed) стал первым (A) – это требуется для программы DIS8 при обработке многолетних факторных экспериментов. Пример перемещения факторов в трехфакторном массиве:

B -> C; A -> B; C -> A

3A x 4B x 5C x 2 повт. -> 5A x 3B x 4C x 2 повт., далее

5A x 3B x 4C x 2 повт. -> 4A x 5B x 3C x 2 повт., далее

4A x 5B x 3C x 2 повт. -> 3A x 4B x 5C x 2 повт. (исх.форма).

Массив данных можно распечатать на принтере, при этом вычисляются частные средние вариантов.

5.9.3. Тест однородности дисперсий

Программа позволяет проверить одну из основных предпосылок классического дисперсионного анализа – однородность дисперсий. Массив данных должен быть типа "варианты-повторности" (с равным числом повторений), допускается наличие некоторого числа выпавших дат. Однородность дисперсий проверяется на основе стандартных критериев – Кокрена, Хартли, Бартлетта – сравнением с табличными значениями на уровнях значимости 1% и 5%.

Число вариантов любого фактора не должно превышать 15 (ограничение, связанное с имеющимися таблицами критериев однородности). Программа вы-

числяет дисперсии всех факторов, выполняя "восстановление" выпавших данных, если таковые имеются.

Статистика Кокрена проверяет 0-гипотезу: все дисперсии равны, контргипотеза: выборка с максимальным значением дисперсии – из другой генеральной совокупности.

Статистика Бартлетта несколько мощнее критерия Кокрена, но если выборки взяты из совокупностей, распределение которых отличается от нормального, это может привести к отклонению 0-гипотезы, когда на самом деле она верна ([15], стр.46). В программе используется аппроксимация критерия Бартлетта статистикой Hi^2.

Результаты анализа формируются для каждого критерия однородности следующим образом:

– – однородность дисперсий отклоняется на уровне 1%;

– однородность дисперсий отклоняется на уровне 5%;

\+ однородность дисперсий не отклоняется на уровне 5%;

\+ + однородность дисперсий подтверждена.

Программа тестировалась по [15], стр.49, [20], стр.271, [21], стр.26.

5.9.4. Формирование массива для множественной регрессии

Операция «Записать как массив для регрессионного анализа» может быть использована в тех случаях, когда уровни всех факторов имеют количественный, равномерно возрастающий характер. Например, фактор А: 4 варианта дозы азотных удобрений (0 кг/га, 30 кг/га, 60 кг/га, 90 кг/га), фактор В: 3 варианта дозы фосфорных удобрений (0 кг/га, 40 кг/га, 80 кг/га). Для упрощения структуры массива уровни факторов кодируются числами натурального ряда.

Первые три столбца – “независимые” переменные (фактор А, фактор В, взаимодействие факторов АхВ), четыре оставшихся столбца – “зависимая” переменная Y в четырех повторениях:

исходный массив	массив для регрессионного анализа
4 12 4 3	7 12 3 4
12,3 12,5 14,2 13,1	0 0 0 12,3 12,5 14,2 13,1
8,34 13,7 13,1 11,9	0 1 0 8,34 13,7 13,1 11,9
13,3 14,6 11,0 15,3	0 2 0 13,3 14,6 11,0 15,3
7,12 9,08 10,3 11,5	1 0 0 7,12 9,08 10,3 11,5
8,27 9,56 7,33 10,1	1 1 1 8,27 9,56 7,33 10,1
6,21 8,44 7,25 5,29	1 2 2 6,21 8,44 7,25 5,29

```
12,2 12,6 14,1 13,2          2 0 0 12,2 12,6 14,1 13,2
8,35 13,8 13,2 11,8          2 1 2 8,35 13,8 13,2 11,8
13,4 14,5 11,4 11,0          2 2 4 13,4 14,5 11,4 11,0
7,13 9,03 10,2 11,4          3 0 0 7,13 9,03 10,2 11,4
8,28 9,57 7,32 10,2          3 1 3 8,28 9,57 7,32 10,2
6,22 8,45 7,24 5,24          3 2 6 6,22 8,45 7,24 5,24
```

Загрузив такой массив в программу множественного линейного регрессионного анализа MLREG, можно проверить существование возможных нелинейных эффектов действия удобрений (квадратичного, кубического), вычислить значение отклика в произвольной точке (например, при N=45 кг/га, P=60 кг/га: X1=1,5, X2=1,5, X3=2,25).

Для факторов неколичественного характера (например, «Сорта», «Способы вспашки», «Тип гербицида») в принципе тоже можно сформировать массивы для множественного регрессионного анализа, но коэффициенты полученного таким образом уравнения регрессии будут носить абстрактный характер, интерпретировать их значения каким-либо образом вряд ли возможно.

5.10. REPLICS: оценка оптимального числа повторений

Типичная проблема, стоящая перед многими исследователями – определение минимального числа повторений в факторном эксперименте. Минимизация числа повторений позволяет снизить трудоемкость эксперимента, затраты на его проведение. Как правило, эксперимент выполняется для **доказательства действия** фактора, выявления вариантов с достоверным различием средних.

Распространено мнение, что число повторений должно быть не менее 4-х, однако практически никогда это не анализируется с точки зрения вариабельности предполагаемых экспериментальных данных и структуры факторного плана. Это довольно часто приводит плачевным результатам – случайные факторы, накладываясь на действие исследуемых факторов, приводят к низким значениям критерия Фишера-Снедекора, подтверждая тем самым 0-гипотезу, отклоняя действие факторов на стандартном уровне значимости, хотя они должны действовать, по мнению экспериментатора. Если бы эксперимент был проведен не с 4-мя повторениями, а с 5-ю или 6-ю, действие факторов возможно было бы доказано.

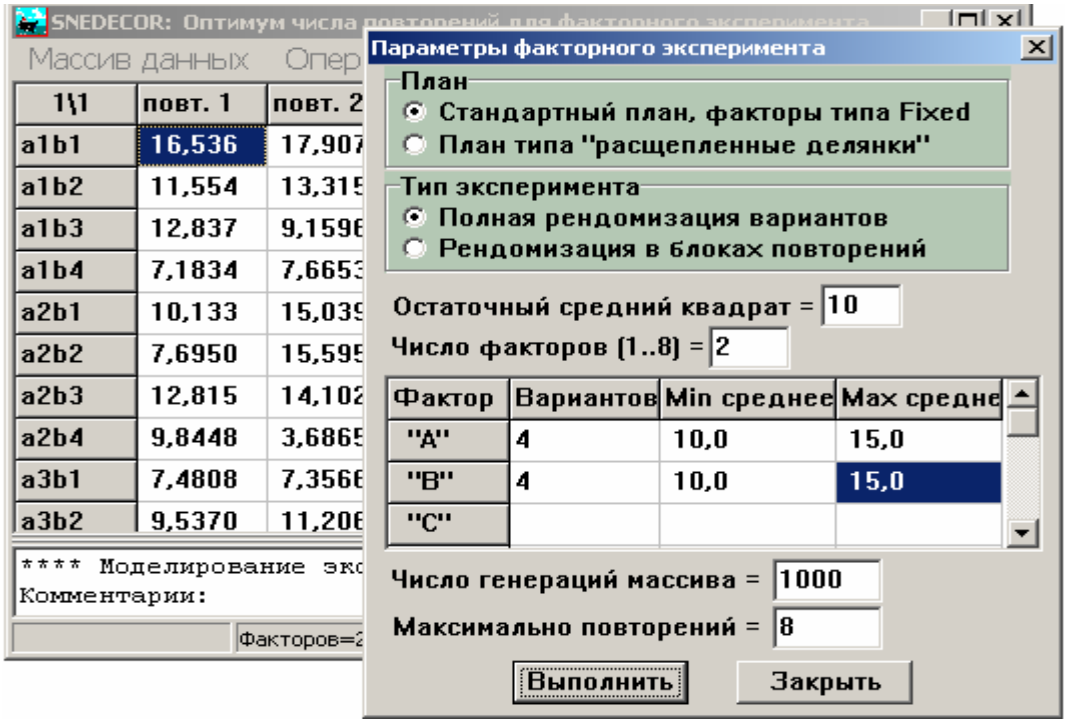
Программа REPLICS предназначена для определения оптимального количества повторений в полнофакторном эксперименте методом Монте-Карло. Для этого у экспериментатора должны быть:

– оценка предполагаемой ошибки опыта в виде остаточного среднего квадрата (из таблицы дисперсионного анализа),

– оценки минимума и максимума средних каждого фактора.

Эти значения обычно могут быть определены на основе ранее проведенных аналогичных экспериментов или предложены из экспертных оценок.

После ввода с клавиатуры плана опыта (число факторов, вариантов в каждом факторе, типа рандомизации) и вышеуказанных оценок, программа многократно (10..5000 раз) формирует случайные массивы в соответствии с моделью предполагаемого факторного эксперимента и обрабатывает их методом стандартного дисперсионного анализа.



При этом накапливаются средние значения решающих характеристик, отражающих возможные результаты будущего факторного эксперимента:

– критерий Фишера-Снедекора (доказательство действия фактора);

– вероятность ошибки при отклонении 0-гипотезы (об отсутствии действия фактора);

– Наименьшая Существенная Разница (доказательство различия хотя бы одной пары факторных средних критерием Стьюдента).

Анализируя результаты моделирования эксперимента в циклах с различным числом повторений, можно сделать продуктивные выводы об оптимальном числе повторений, при котором достигаются цели экспериментатора. Естественно, меняя значения остаточного среднего квадрата и размах средних, можно многократ-

но повторять моделирование, исследуя поведение результирующих характеристик.

Такие циклы выполняются последовательно, начиная с 2-х повторений и до 8-и или больше, исходя из целей экспериментатора:

Естественные предположения, на основе которых формируются случайные массивы:

1/ Общее среднее массива = полусумма минимального и максимального средних любого фактора;

2/ Значения факторных средних равномерно возрастают в соответствии с номером варианта;

3/ Любые взаимодействия факторов – только аддитивные, без эффектов синергизма или антагонизма;

4/ Сумма случайных факторов (ошибка) добавляется к каждому значению массива как $\mathbf{N_{rand} * \sqrt{Er}}$, где N_{rand} – случайное число с нормальным законом распределения (среднее=0, сигма=1), Er – остаточный средний квадрат из таблицы дисперсионного анализа.

Пример теста: 2-факторный план: по 4 варианта в каждом факторе, полная рандомизация, остаточный средний квадрат = 20:

Повторений =2, циклов =1000
Общая дисперсия =919,78290
Остаточный средний квадрат =19,7663
Общее среднее =11,972

Параметр	A	B
Вариантов	4	4
Дисперсия	131,748	136,044
F-критерий	2,5177	2,6051
Вероятность	0,095	0,088
НСР(5%)	4,6388	4,6388

Повторений =3, циклов =1000
Общая дисперсия =1384,3814
Остаточный средний квадрат =19,9301
Общее среднее =12,038

Параметр	A	B
Вариантов	4	4
Дисперсия	161,594	171,757
F-критерий	2,9128	3,0667
Вероятность	0,049	0,042
НСР(5%)	3,6839	3,6839

```
Повторений =4,  циклов =1000
Общая дисперсия =1855,6203
Остаточный средний квадрат =19,8752
Общее среднее =12,016
```

Параметр	A	B
Вариантов	4	4
Дисперсия	206,071	195,424
F-критерий	3,5914	3,4069
Вероятность	0,020	0,025
НСР(5%)	3,1539	3,1539

Двух повторений недостаточно (P_a=0,095 P_b=0,088), трех повторений, по-видимому, будет достаточно для доказательства действия обоих факторов (P_a=0,049 P_b=0,042), 4-х повторений наверняка хватит для доказательства (P_a=0,02 P_b=0,025).

План эксперимента может быть:

1/ Стандартным многофакторным планом с полной рандомизацией вариантов/повторений с фиксированными уровнями вариантов факторов (тип “Fixed”), или с рандомизацией вариантов в блоках повторений (метод "случайных блоков");

2/ Многофакторным планом с расщеплением вариантов факторов: более крупные экспериментальные объекты (начиная с вариантов фактора "A") состоят из более мелких экспериментальных объектов – полного набора вариантов следующего фактора ("B"), и так далее до последнего фактора. Эта ситуация обычно имеет место в экспериментах с удобрениями, сортами растений и т.п., когда большие делянки разбиваются на несколько мелких делянок – по числу вариантов субфактора. Предполагается, что опыт организован методом случайных блоков, а уровни факторов в эксперименте фиксированы, то есть используется модель дисперсионного анализа типа “Fixed”.

Ограничения на размер массива данных: число строк – не более 8000, максимальное количество вариантов в любом факторе – 50, максимальное количество факторов – 8; общий размер массива – не более 32000 элементов.

Оптимальное значение количества повторений определяется значениями вероятностей ошибки в случае отклонения 0-гипотезы (ошибки 1-го рода). Все вероятности должны быть меньше 0,05 для стандартных экспериментов, и меньше 0,01 – для строгих экспериментов. Значения Наименьших Существенных Разниц (НСР 5%) должны быть не менее размаха факторных средних, задаваемого экспериментатором.

6. Ковариационный анализ. Линейная модель

Ковариационный анализ – эффективная модификация параметрического дисперсионного анализа, позволяющего уточнить результаты факторного эксперимента, сделать более надежными выводы о действии фактора, различии факторных средних.

Как и в стандартном дисперсионном анализе, для массива данных должны выполняться предпосылки нормальности распределения ошибок измерения, однородности дисперсии и независимости в последовательности значений.

Данные для ковариационного анализа состоят из двух массивов одинаковой структуры: 1-й массив (Y, основной параметр изучаемой системы) – идентичен массиву для обычного дисперсионного анализа, 2-й массив (X, дополнительный параметр) используется для корректировки основного параметра в случае существования функциональной связи между этими параметрами, обычно предполагается наличие линейной зависимости Y=A+B*X.

6.1. COVAR1: 1-факторный ковариационный анализ

Программа COVAR1 предназначена для обработки экспериментальных данных методом 1-факторного ковариационного анализа, с возможностью анализа различий средних по критерию НСР. Предполагается, что уровни фактора в эксперименте фиксированы, то есть используется модель данных типа "Fixed". Подразумевается следующая математическая модель данных (полная рандомизация):

$\mathbf{y_{ij} = \mu + a_i + q*(x_{ij} - \bar{x}) + e_{ij}}$; $\boldsymbol{\mu}$ – генеральное среднее изучаемой системы;

a_i – эффект i-го варианта фактора типа Fixed;

q – коэффициент линейной регресии, q=/=0;

x_{ij} – независимая переменная;

e_{ij} – ошибка от случайных факторов, распределена по N(0, σ).

0-гипотеза: все a_i=0, контр-гипотеза: некоторые a_i=/=0.

Математическая модель в случае рандомизации в блоках:

$\mathbf{y_{ij} = \mu + a_i + q*(x_{ij} - \bar{x}) + r_j + e_{ij}}$;

r_j – возможный эффект j-го блока повторений.

Варианты исследуемого фактора могут иметь неравное или равное число повторностей, в последнем случае возможно указание типа организации опыта (полная рандомизация или рандомизация в блоках). В случае данных с неравным числом повторностей программа автоматически переключается на тип "полной рандомизации".

Данные в виде двумерного массива "варианты-повторения" могут быть введены с клавиатуры непосредственно в среде программы, либо из текстового файла в стандарте SNEDECOR/COVAR1. Ограничения на размер массива данных: число вариантов – не более 1000, максимальное количество повторностей – 100, общий размер массива – не более 32000 элементов. Пример формирования массива из 4-х повторностей, 5-и вариантов в текстовом файле:

```
4  5                        <- начало файла
12,3 12,5 14,2 13,1         Y 1-й вар.
8,34 13,7 13,1 11,9         Y 2-й вар.
13,3 14,6 11,0 15,3         Y 3-й вар.  Массив данных:
7,12 9,08 10,3 11,5         Y 4-й вар.  Строки = варианты,
8,27 9,56 7,33 10,3         Y 5-й вар.  Столбцы = повторности.
1,21 1,44 1,25 1,29         X 1-й вар.
1,17 1,87 1,69 1,53         X 2-й вар.
1,25 1,23 1,27 1,26         X 3-й вар.
1,34 1,30 1,37 1,34         X 4-й вар.
1,41 1,45 1,43 1,49         X 5-й вар.

Данные 1997 г               <- необязательный комментарий
```

В качестве примера формирования массива для программы COVAR1 можно посмотреть файл DOSP303.dat. Массивы данных, подготовленных для обработки методами дисперсионного анализа различной факторности, не могут быть переданы программе COVAR1, но обратная передача массива (зависимой переменной Y) в программы D1MAXI или DIS8 возможна.

Основной результат работы программы – критерии Фишера-Снедекора для двух способов обработки данных: стандартного дисперсионного анализа, и анализа с учетом связи изучаемого параметра (Y) с некоторой независимой переменной (X). Предполагается, что эта связь имеет линейный характер, и в программе вычисляются коэффициенты линейной регрессии, с помощью которой корректируются средние по вариантам (массив COV1.dat):

Таблица разложения дисперсий ANACOVA

Дисперсия	X	XY	Y	Доля вариации	Степени свободы	Средний квадрат
Общая	10468,550	6738,100	10354,200	1.0000	19	544,96

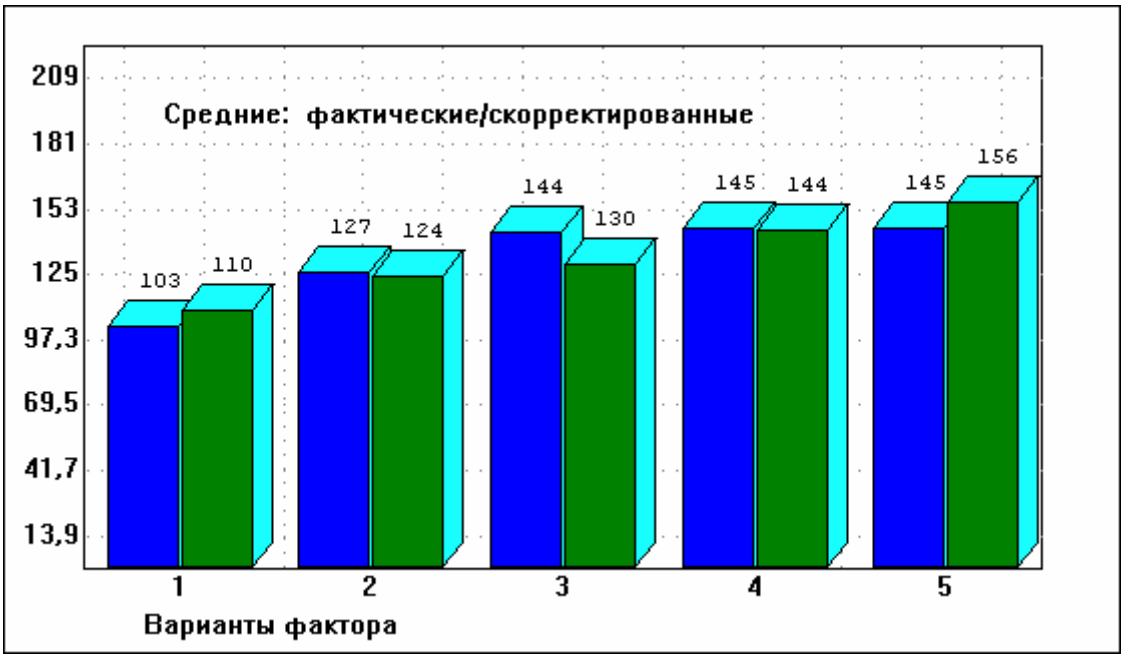
```
Вариантов          |  2698,300 | 1157,100 | 5324,700 | 0,5143 |    4   | 1331,2 |
Случайных факторов|  7770,250 | 5581,000 | 5029,500 | 0,4857 |   15   | 335,30 |
 |=> от регрессии |                        4008,566 | 0,3871 |    1   | 4008,6 |
 |=> сл.факторы - регрессия                1020,934 | 0,0986 |   14   | 72,924 |
```

```
Анализ с учетом ковариации от независимой переменной:
Достоверность регрессии:   Y = 67,806 + 0,7183 * X;
F-критерий= 54,969,  ст.св.= 1, 14;  P=0,00000
Действие фактора:
F-критерий (по Снедекору) = 17,128,  ст.св.=4, 14;  P=0,00003
Сила влияния фактора по Снедекору = 0,8118
F-критерий (по Доспехову) = 18,254,  ст.св.=4, 14;  P=0,00002
Стандартная ошибка = 4,2698
НСР(1%)= 17,975  НСР(5%)= 12,951  НСР(10%)= 10,635
```

Вначале проверяется достоверность регрессии по критерию Фишера-Снедекора. Если регрессия доказана, тогда эту связь можно использовать для коррекции средних, и затем проверить 0-гипотезу для фактора. 0-гипотеза в данном случае формулируется следующим образом: отсутствует действие изучаемого фактора, скорректированные средние имеют разные значения вследствие действия случайных факторов:

Программа выводит два значения F-критерия для действия фактора – по Снедекору и Доспехову; как правило, их величины не противоречат друг другу, в противном случае рекомендуем использовать результат анализа по Снедекору [20], стр. 368-384.

Средние могут быть представлены в графическом виде:



6.2. COVAR2: 2-факторный ковариационный анализ

Программа COVAR2 предназначена для обработки экспериментальных данных методом 2-факторного ковариационного анализа, с возможностью ана-

лиза различий средних по критерию НСР. Предполагается, что уровни факторов в эксперименте фиксированы, то есть используется модель данных типа "Fixed". Варианты исследуемых факторов должны иметь равное число повторностей; возможно указание типа организации опыта (полная рандомизация или рандомизация в блоках). Подразумевается следующая математическая модель данных, полная рандомизация:

$$y_{ijk} = \mu + a_i + b_j + ab_{ij} + q * (x_{ijk} - \bar{x}) + e_{ijk};$$

μ – генеральное среднее изучаемой системы;
a_i – эффект i-го варианта фактора A типа Fixed;
b_j – эффект j-го варианта фактора B типа Fixed;
ab_{ij} – эффект взаимодействия факторов;
q – коэффициент линейной регресии, q=/=0;
x_{ijk} – независимая переменная;
e_{ijk} – ошибка от случайных факторов, распределена по N(0, σ).

0-гипотезы: все a_i=0, все b_j=0, все ab_{ij}=0;

контр-гипотезы: некоторые a_i=/=0, b_j=/=0, ab_{ij}=/=0.

Математическая модель в случае рандомизации в блоках:

$$y_{ijk} = \mu + a_i + b_j + ab_{ij} + q * (x_{ijk} - \bar{x}) + r_k + e_{ijk};$$

r_k – возможный эффект k-го блока повторений.

Данные в виде двумерного массива "варианты-повторения" могут быть введены с клавиатуры непосредственно в среде программы, либо загружены из файла в стандарте SNEDECOR/COVAR2, подготовленного заранее. Массивы данных, подготовленных для обработки методами дисперсионного анализа различной факторности, не могут быть переданы программе COVAR2, но обратная передача массива (зависимой переменной Y) в программы D2MAXI или DIS8 возможна.

Пример формирования массива из 4-х повторностей, 2-х вариантов фактора "A" и 3-х вариантов фактора "B" в текстовом файле:

```
4 6 2 3                    <- начало файла
12,3 12,5 14,2 13,1        Y 1a1b
8,34 13,7 13,1 11,9        Y 1a2b   Всего 2*3=6 вариантов;
13,3 14,6 11,0 15,3        Y 1a3b   Массив данных:
7,12 9,08 10,3 11,5        Y 2a1b   Строки  = варианты,
8,27 9,56 7,33 10,3        Y 2a2b   Столбцы = повторности.
11,5 12,1 11,7 13,5        Y 2a3b
1,21 1,44 1,25 1,29        X 1a1b
1,17 1,87 1,69 1,53        X 1a2b
1,25 1,23 1,27 1,26        X 1a3b
1,34 1,30 1,37 1,34        X 2a1b
1,41 1,45 1,43 1,49        X 2a2b
1,46 1,52 1,55 1,53        X 2a3b

Данные за 1998 г           <- необязательный комментарий
```

В качестве примера формирования массива для программы COVAR2 можно посмотреть файл RAO261.dat.

Основной результат работы программы – критерии Фишера-Снедекора для двух способов обработки данных: стандартного дисперсионного анализа, и анализа с учетом связи изучаемого параметра (Y) с некоторой независимой переменной (X). Предполагается, что эта связь имеет линейный характер, и в программе вычисляются коэффициенты линейной регрессии, с помощью которой корректируются средние по вариантам. Вначале проверяется достоверность регрессии по критерию Фишера-Снедекора. Если регрессия доказана, тогда эту связь можно использовать для коррекции средних, и затем проверить 0-гипотезу для фактора. 0-гипотеза формулируется следующим образом: отсутствует действие изучаемого фактора, скорректированные средние имеют разные значения вследствие действия случайных факторов.

0-гипотеза для проверки аддитивности: действие факторов на изучаемую систему есть простая сумма эффектов, взаимодействие (синергизм или антагонизм) факторов отсутствует; вероятность для F-критерия трактуется обычным образом.

По умолчанию программа вычисляет значения F-критерия для факторов классическим методом (по С.Р.Рао); при желании можно указать программе вычислять значения F-критерия по формулам, рекомендованным А.Б.Доспеховым. Как правило, эти методы не противоречат друг другу, в противном случае рекомендуем использовать результат анализа по С.Р.Рао [32, стр. 258-263]. Данные из этой книги (RAO261.dat) использовались в качестве теста:

```
Анализ с учетом ковариации от независимой переменной (по С.Р.Рао).
Достоверность регрессии:   Y =  6,2535 + 0,076148 * X
```

F-критерий = 19,206, Ст.Св.=1, 23; P=0,00022

Фактор	Степень влияния	Критерий Фишера	Степени свободы	Вероятность ошибки	НСР(1%)	НСР(5%)	НСР(10%)
A	0,1922	3,760	2, 23	0,03865*	0,706	0,521	0,431
B	0,0208	3,740	1, 23	0,06554	0,577	0,425	0,352
AB	0,0671	0,191	2, 23	0,82775	0,999	0,736	0,610

Стандартная ошибка = 0,2517 (0,63% от общего среднего)
Остаточный средний квадрат = 0,3167

Регрессия достоверна на высоком уровне значимости (P<0,001), поэтому используется для уточнения дисперсионного анализа; доказано действие фактора А на уровне значимости 5%, действие фактора В возможно на уровне 10%, подтверждена гипотеза аддитивности факторов (отсутствие взаимодействия).

Для анализа различий средних в качестве контроля предлагаются первые варианты факторов; если же в действительности в опыте контрольными были другие варианты, их номера следует ввести перед выполнением расчетов.

7. Многомерный дисперсионный анализ

Если в факторном эксперименте накапливаются данные по нескольким параметрам, существует возможность обработать их совместно. Например, в полевом опыте с каждой делянки получены данные по нескольким признакам:

– урожай зерновых,

– содержание протеина,

– надземная биомасса,

– средняя высота растений.

Совместная обработка такого 4-мерного массива может показать действие исследуемого фактора на заданном уровне значимости, тогда как одномерные дисперсионные анализы возможно не покажут достоверного действия. Вследствие очевидных множественных линейных связей этих признаков многомерный анализ позволит выявить действие фактора.

Анализ различия факторных средних в многомерном случае преобразуется в анализ расстояний между векторами средних в Эвклидовом пространстве, с помощью F-критерия проверятся 0-гипотеза: расстояние между векторами равно нулю.

7.1. MANOVA1. 1-факторный N-мерный дисперсионный анализ

Программа MANOVA1 предназначена для обработки данных методом многомерного однофакторного дисперсионного анализа с равным или неравным числом повторений в вариантах. Предполагается, что для всех однофакторных массивов используется модель данных типа Fixed.

Ограничения на размер массива: число признаков (размерность данных, M) может быть не более 100, число объектов (сумма повторений в вариантах, N) – не более 4000, но при соблюдении условия M x N <= 100000; число вариантов (групп) – не более 50.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х признаков, 3-х вариантов, 12 объектов в текстовом файле:

```
                повторений в каждом из 3 вариантов
                    |   |   |
4 12 3 4 5                          <- начало файла: 3+4+5=12 объектов
12,3 22,5 34,2 51,4                 \
10,4 20,9 33,2 49,5                  1-й вариант: 3 повторения
9,23 24,6 31,6 59,0                 /
7,12 20,9 30,3 56,3                 \
6,21 18,5 31,6 51,5                  \
5,67 17,2 30,6 55,6                   2-й вариант: 4 повторения
7,23 15,2 32,1 57,3                  /
6,78 16,2 31,9 53,7                 \
9,44 10,2 11,7 59,1                  \
8,34 23,7 33,1 57,6                   3-й вариант: 5 повторений
11,4 23,7 33,5 57,1                  /
10,8 25,3 32,9 55,2                 /

Данные 1998 г, Н-ск                 <- необязательный комментарий
1----2----3----4------признаки
```

В качестве примера формирования массива можно посмотреть файл ARENS131.dat (10 признаков, 31 объект в 3-х группах/вариантах).

Программа выполняет для каждого признака стандартный 1-мерный дисперсионный анализ с вычислением критерия Фишера-Снедекора, с выводом таблицы средних. Затем вычисляется многомерный λ-критерий Уилкса, аппроксимация этого критерия с помощью распределения Hi^2, а также F-критерий для проверки 0-гипотезы: различия векторов средних между какой-либо парой вариантов опыта отсутствуют, фактор не влияет на исследуемую систему. Используется аппроксимация многомерного критерия по С.Р.Рао [31] и по Аренсу-Лейтеру [32] во второй (более точной) форме. Если вероятность

P<=0,01 0-гипотеза отвергается с уровнем значимости 1%, по крайней мере два варианта (два вектора средних) достоверно отличаются;

P<=0,05 0-гипотеза отвергается с уровнем значимости 5%;

P > 0,10 0-гипотеза остается в силе: векторы средних отличаются только из-за действия случайных факторов.

Далее программа выполняет анализ различий векторов средних, вычисляя парный F-критерий для всех вариантов в сравнении с контрольным вариантом. Используется методология, изложенная в [32, стр. 108]. Номер контрольного варианта можно выбирать перед началом анализа. Если вероятность для соответствующего F-критерия меньше пороговой (обычно 0,05), данный вектор средних достоверно отличается от вектора средних контроля (ANDERSON.dat):

Признак N	Суммы квадратов/Средние квадраты: Общее	Вариантов	Остаток	F-критерий: 1-мерный	вероятность
X1	24077,867	18011,078	6066,789		
	830,271	3602,216	252,783	14,250	0,0000
X2	17224,167	10344,589	6879,578		
	593,937	2068,918	286,649	7,218	0,0003
Ст.своб.	29	5	24		

Лямбда-критерий Уилкса = 0,10045 ст.св.=2, 5, 24
Аппроксимация Лямбда-критерия Hi2 = 71,240 ст.св.=10 P=0,0000
N-мерный F-критерий (по C.P.Pao)= 9,91365 ст.св.= 20,0, 46,0 P=0,0000
N-мерный F-критерий (по Лейтеру)= 10,4424 ст.св.= 12,2, 23,0 P=0,0000

Таблица средних. Различия векторов средних.

Признак	Варианты 1	2	3	4	5	6	Общие средние	НСР(5%) 1-мерн.
X1	102,8	155,6	91,60	126,2	90,00	88,20	109,07	20,754
X2	82,80	116,8	119,0	91,80	71,00	77,60	93,167	22,100
Повторений	5	5	5	5	5	5	30	
F-критерий	Контр.	13,34	10,87	2,683	0,904	1,057	0,7863	
Вероятность	-	0,000	0,000	0,090	0,419	0,364	0,4674	

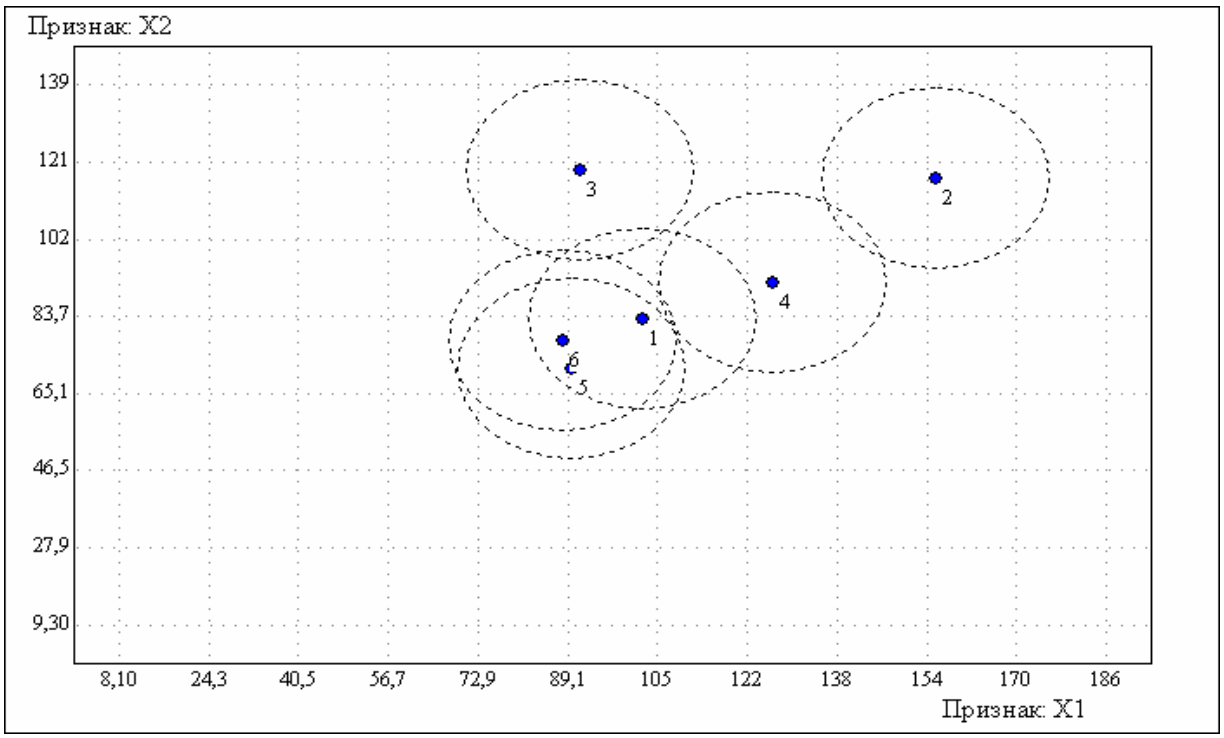
Ст.свободы для N-мерного парного F-критерия =2, 23

Критерием Hi^2, F-критериями по Рао и Лейтеру принимается контргипотеза: фактор действует (P<0,0001), векторы средних 2-го и 3-го вариантов достоверно различаются от вектора средних контроля (1-го варианта).

В случае, если все варианты имеют равное число повторений, программа по умолчанию использует метод обработки данных по типу "полной рандомиза-

ции"; возможно указание типа "рандомизации в блоках повторений" перед анализом.

Для облегчения анализа средних рекомендуем использовать графики "2D средние" и "3D средние". Программа определяет два (или три) признака с максимальными значениями 1-мерного F-критерия и строит в осях X-Y(-Z) координаты средних, сопровождая их эллипсами, формируемыми значениями наименьших существенных разниц (НСР):



7.2. MANOVA2. 2-факторный N-мерный дисперсионный анализ экспериментов с повторениями

Программа MANOVA2.exe предназначена для обработки данных, полученных в 2-факторных опытах с равным числом повторений, методом многомерного дисперсионного анализа. Предполагается, что для всех 2-факторных массивов используется модель данных типа I (фиксированные уровни вариантов обоих факторов).

Ограничения на размер массива: число признаков (размерность данных, M) может быть не более 50, число объектов (N = произведение числа вариантов фактора "A" на число вариантов фактора "B" и на число повторений) – не более 4000, но при соблюдении условия M x N <= 32000.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х признаков, 2-х вариантов фактора "A", 3-х вариантов фактора "B" и 2-х повторений в текстовом файле:

	вариантов фактора «A» | вариантов фактора «B» | | повторений
4 12 2 3 2	<- начало файла: 2*3*2=12 объектов
12,3 22,5 34,2 51,4	1a 1b 1п \
10,4 20,9 33,2 49,5	1a 1b 2п \
9,23 24,6 31,6 59,0	1a 2b 1п \
7,12 20,9 30,3 56,3	1a 2b 2п 1-й вариант фактора A
6,21 18,5 31,6 51,5	1a 3b 1п /
5,67 17,2 30,6 55,6	1a 3b 2п /
7,23 15,2 32,1 57,3	2a 1b 1п \
6,78 16,2 31,9 53,7	2a 1b 2п \
9,44 10,2 11,7 59,1	2a 2b 1п \
8,34 23,7 33,1 57,6	2a 2b 2п 2-й вариант фактора A
11,4 23,7 33,5 57,1	2a 3b 1п /
10,8 25,3 32,9 55,2	2a 3b 2п /
Данные 1998 г, Н-ск	<- необязательный комментарий

1----2----3----4------признаки

В качестве примера формирования массива можно посмотреть файл ARENS169.dat (2 признака, 3 варианта "A", 3 варианта "B", 4 повторения). Массивы данных, подготовленные в других программах дисперсионного анализа, не могут быть переданы для обработки программе MANOVA2, и наоборот. Другие программы (VARS, NORMAL, MCOR, MCOM и т.п.) могут использовать массивы от MANOVA2.

Программа выполняет для каждого признака стандартный 1-мерный дисперсионный анализ с вычислением критериев Фишера-Снедекора, с выводом таблиц средних. Затем вычисляются многомерные критерии для проверки 0-гипотез для факторов: различия векторов средних между какой-либо парой вариантов опыта отсутствуют, фактор не влияет на исследуемую систему. Используется аппроксимация многомерного критерия к F-распределению по Аренсу-Лейтеру [31], с вычислением вероятности ошибки в случае отклонения 0-гипотезы. Если

P<=0,01 0-гипотеза отвергается с уровнем значимости 1%, по крайней мере два варианта (два вектора средних) достоверно отличаются;

P<=0,05 0-гипотеза отвергается с уровнем значимости 5%;

P > 0,10 0-гипотеза остается в силе: векторы средних отличаются только из-за действия случайных факторов.

0-гипотеза для проверки возможного взаимодействия факторов: отсутствует эффект синергизма или антагонизма факторов при любом сочетании вариантов факторов, которые действуют на исследуемую систему как простая сумма возможных эффектов. Значение вычисленной вероятности трактуется аналогичным образом:

3. Многомерные критерии действия факторов по Аренсу-Лейтеру.

Фактор	F-критерий	Ст.свободы	Вероятность
A	1,71296	4,17 26	0,1756
B	13,5067	4,17 26	0,0000
A x B	0,71141	9,09 26	0,6947

4. Таблица средних фактора A, Различия векторов средних.

Признак	Варианты 1	2	3	Общие средние
X1	7,917	7,833	9,500	8,4167
X2	7,250	7,917	9,000	8,0556
Выборка	12	12	12	36
F-критерий	Контр.	0,644	2,124	0,5276
Вероятность	-	0,533	0,140	0,5962

Степеней свободы для парного F-критерия =2, 26

5. Таблица средних фактора B, Различия векторов средних.

Признак	Варианты 1	2	3	Общие средние
X1	6,833	7,250	11,17	8,4167
X2	7,167	8,083	8,917	8,0556
Выборка	12	12	12	36
F-критерий	Контр.	0,488	19,87	3,4252
Вероятность	-	0,619	0,000	0,0478

Степеней свободы для парного F-критерия =2, 26

Для фактора A: нет оснований отвергать 0-гипотезу (P>0,1), векторы средних различаются только из-за множества случайных факторов. Для фактора B: 0-гипотеза отклоняется, принимается контр-гипотеза – фактор действует, вектор

средних 3-го варианта достоверно (P<0,001) отличается от вектора средних 1-го варианта. Взаимодействие факторов отсутствует.

Следует заметить, что число степеней свободы числителя F-отношения может быть дробным числом, в этом случае программа интерполирует к значению вероятности по целым значениям степеней свободы. Существуют "Таблицы F-распределения" [42], в которых можно найти значения F-критерия для дробных степеней свободы.

Далее программа выполняет анализ различий векторов средних по обоим факторам, вычисляя парные F-критерии для всех вариантов в сравнении с контрольным вариантом. Используется методология, изложенная в [31, стр. 108]. Номера контрольных вариантов можно выбирать в пункте Меню "Установ..." (по умолчанию программа считает 1-е варианты факторов контрольными). Если вероятность для соответствующего F-критерия меньше пороговой (обычно 0,05), данный вектор средних достоверно отличается от вектора средних контроля. В этом же пункте Меню возможно указание метода организации эксперимента (полная рандомизация/случайные блоки повторений), по умолчанию программа обрабатывает данные по типу полной рандомизации.

Программа тестировалась по [31, стр. 169].

Для облегчения анализа средних рекомендуем использовать графики "2D средние" и "3D средние". Программа определяет два (или три) признака с максимальными значениями 1-мерного F-критерия и строит в осях X-Y(-Z) координаты средних, сопровождая их эллипсами (или "усами"), формируемыми значениями наименьших существенных разниц (НСР).

7.3. MNV2: 2-факторный N-мерный дисперсионный анализ экспериментов без повторений

Программа MNV2 предназначена для обработки данных, полученных в опытах без повторений, методом многомерного 2-факторного дисперсионного анализа. Предполагается, что для всех 2-факторных массивов используется модель данных типа Fixed.

Ограничения на размер массива: число признаков (размерность данных, M) может быть не более 100, число объектов (произведение числа вариантов фактора "A" на число вариантов фактора "B") – не более 10000.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х признаков, 3-х вариантов фактора "A", 5-и вариантов фактора "B" в текстовом файле:

```
        вариантов фактора «A»
        |  вариантов фактора «B»
4 15 3 5                          <- начало файла: 3*5=15 объектов
12,3 22,5 34,2 51,4               1a 1b  \
10,4 20,9 33,2 49,5               1a 2b  \
9,23 24,6 31,6 59,0               1a 3b   1-й вариант фактора A
7,12 20,9 30,3 56,3               1a 4b  /
6,21 18,5 31,6 51,5               1a 5b /
5,67 17,2 30,6 55,6               2a 1b  \
7,23 15,2 32,1 57,3               2a 2b  \
6,78 16,2 31,9 53,7               2a 3b   2-й вариант фактора A
9,44 10,2 11,7 59,1               2a 4b  /
8,34 23,7 33,1 57,6               2a 5b /
11,4 23,7 33,5 57,1               3a 1b  \
10,8 25,3 32,9 55,2               3a 2b  \
9,44 10,2 11,7 59,1               3a 3b   3-й вариант фактора A
8,34 23,7 33,1 57,6               3a 4b  /
9,23 24,6 31,6 54,4               3a 5b /

Данные 1998 г, H-ск               <- необязательный комментарий
  1----2----3----4                признаки
```

В качестве примера формирования массива можно посмотреть файлы DOSP257.dat (4 признака, 2 варианта "A", 5 вариантов "B"), GIV4x15.dat.

Программа выполняет для каждого признака стандартный 1-мерный дисперсионный анализ с вычислением критериев Фишера-Снедекора, с выводом таблиц средних. Затем вычисляется многомерный λ-критерий Уилкса, аппроксимация этого критерия с помощью распределения Hi^2, а также F-критерий для проверки 0-гипотезы: различия векторов средних между какой-либо парой вариантов опыта отсутствуют, фактор не влияет на исследуемую систему. Используется аппроксимация многомерного критерия по C.P.Pao [32] и по Аренсу-Лейтеру [31]. Если вероятность

P<=0,01 0-гипотеза отвергается с уровнем значимости 1%, по крайней мере два варианта (два вектора средних) достоверно отличаются;

P<=0,05 0-гипотеза отвергается с уровнем значимости 5%;

P >0,10 0-гипотеза остается в силе: векторы средних отличаются только из-за действия случайных факторов.

Далее программа выполняет анализ различий векторов средних по обоим факторам, вычисляя парный F-критерий для всех вариантов в сравнении с контрольным вариантом. Используется методология, изложенная в [31, стр. 108]. Номера контрольных вариантов можно выбирать перед анализом (по умолчанию программа считает 1-е варианты факторов контрольными). Если вероятность для соответствующего F-критерия меньше пороговой (обычно 0,05), данный вектор средних достоверно отличается от вектора средних контроля.

Для облегчения анализа средних рекомендуем использовать графики "2D средние" и "3D средние". Программа определяет два (или три) признака с максимальными значениями 1-мерного F-критерия и строит в осях X-Y(-Z) координаты средних, сопровождая их эллипсами, формируемыми значениями наименьших существенных разниц (НСР).

7.4. MANOVA8. Многофакторный N-мерный дисперсионный анализ экспериментов с повторениями

Программа MANOVA8 предназначена для обработки данных, полученных в многофакторных опытах с равным числом повторений, методом многомерного дисперсионного анализа. Предполагается, что используется модель опыта типа I (фиксированные уровни вариантов всех факторов). Возможен учет типа рандомизации эксперимента (полная рандомизация / рандомизация в блоках повторений).

Ограничения на размер массива: число факторов – не более 8, число признаков (размерность данных, M) – не более 50, число объектов (N = произведение числа вариантов на число повторений) – произвольно, но при соблюдении условия M x N <= 32000.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла.

В массиве данных не должно быть "отсустствующих" значений; если все же они есть, следует подставить вместо них средние из имеющихся повторностей, или же "восстановить" с помощью программы IODATA (с последующей корректировкой 1-й строки файла данных).

Пример формирования массива из 4-х признаков, 2-х вариантов фактора "A", 3-х вариантов фактора "B" и 2-х повторений в текстовом файле:

```
        вариантов фактора «A»
        |  вариантов фактора «B»
        |  |  повторений
4 12 2 3 2                    <- начало файла: 2*3*2=12 объектов
12,3 22,5 34,2 51,4           1a 1b 1п \
10,4 20,9 33,2 49,5           1a 1b 2п  \
9,23 24,6 31,6 59,0           1a 2b 1п  \
7,12 20,9 30,3 56,3           1a 2b 2п   1-й вариант фактора A
6,21 18,5 31,6 51,5           1a 3b 1п  /
5,67 17,2 30,6 55,6           1a 3b 2п /
7,23 15,2 32,1 57,3           2a 1b 1п \
6,78 16,2 31,9 53,7           2a 1b 2п  \
9,44 10,2 11,7 59,1           2a 2b 1п  \
8,34 23,7 33,1 57,6           2a 2b 2п   2-й вариант фактора A
11,4 23,7 33,5 57,1           2a 3b 1п  /
10,8 25,3 32,9 55,2           2a 3b 2п /

Данные 1998 г, Н-ск           <- необязательный комментарий
1----2----3----4------признаки
```

В качестве примера формирования массива можно посмотреть файл ARENS169.dat (2 признака, 3 варианта "A", 3 варианта "B", 4 повторения). Массивы данных, подготовленные в программах одномерного дисперсионного анализа, могут быть переданы для обработки программе MANOVA8 (см. ниже; кроме 1-факторных данных). Другие программы (VARS, NORMAL, MCOR, MCOM и т.п.) могут использовать массивы от MANOVA8. Массивы, созданные в программах MANOVA2, MANODISC, могут быть использованы программой MANOVA8.

Для каждого фактора вычисляется многомерный λ-критерий Уилкса для проверки 0-гипотезы: различия векторов средних между любой парой вариантов отсутствуют, фактор не влияет на исследуемую систему. Так как таблицы квантилей распределения этого критерия труднодоступны, используется аппроксимация критерия Уилкса к F-распределению по Рао [32], с вычислением вероятности ошибки в случае отклонения 0-гипотезы. Если

P<=0,01 0-гипотеза отвергается с уровнем значимости 1%, по крайней мере два варианта (два вектора средних) достоверно отличаются;

P<=0,05 0-гипотеза отвергается с уровнем значимости 5%;

P >0,10 0-гипотеза остается в силе: векторы средних отличаются только из-за действия случайных факторов.

0-гипотеза для проверки возможного взаимодействия факторов: отсутствует эффект синергизма или антагонизма факторов при любом сочетании вариантов факторов, которые действуют на исследуемую систему как простая сумма возможных эффектов. Значение вычисленной вероятности трактуется аналогичным образом.

Следует заметить, что число степеней свободы числителя F-отношения может быть дробным числом, в этом случае программа интерполирует к значению вероятности по целым значениям степеней свободы. Существуют "Таблицы F-распределения" [43], в которых можно найти значения F-критерия для дробных степеней свободы.

С помощью программы MANOVA8 можно обрабатывать обычные 1-мерные массивы данных, для которых имеет место "несферичность" структуры данных. Под этим понимается коррелированность повторений (repeated mesures) совместно с "неоднородностью" дисперсий блоков повторений. Критерий сферичности можно получить с помощью программы MATRIX, которая может использовать массив данных без каких-либо преобразований. Программа MANOVA8 обрабатывает 1-мерные массивы, воспринимая их как многомерные, в которых повторения играют роль признаков. 1-мерные данные должны быть как минимум 2-факторные, при этом варианты последнего фактора (для 2-факторных – "B", для 3-факторных – "C", и т.д.) воспринимаются программой как повторения. Многомерный дисперсионный анализ не требует выполнения предпосылки сферичности данных, поэтому выводы относительно главных эффектов являются корректными. Выводы об эффекте последнего фактора можно получить после преобразования массива данных – изменив порядок следования факторов. Анализ различий факторных средних в этом случае выполняется на основе сравнения векторов средних; достоверность различий выявляется многомерными F-критериями, вычисляемыми для каждой пары векторов (сравнение с контрольным вариантом) с помощью билинейных форм от матрицы остаточных ковариаций [31], стр. 108. Различие векторов средних для задачи множественного сравнения считается доказанным (на заданном уровне значимости, обычно 5%), если значения F-критериев для соответствующих пар вариантов не меньше порогового значения, вычисленного методом, аналогичным S-методу Шеффе в 1-

мерном дисперсионном анализе [31], стр. 111 (наиболее строгий метод анализа множественных сравнений).

Программа тестировалась по [31], стр. 169, [43], стр. 299.

7.5. MANODISC: Многомерный дисперсионный + шаговый дискриминантный анализ

Программа MANODISC предназначена для обработки данных, полученных в многофакторных опытах с равным числом повторений, методом многомерного дисперсионного анализа с повторениями, с возможностью дискриминантного анализа по каждому фактору. Предполагается, что используется модель опыта типа I (фиксированные уровни вариантов всех факторов). Возможен учет типа рандомизации эксперимента (полная рандомизация / рандомизация в блоках повторений).

Ограничения на размер массива: число факторов – не более 8, число признаков (размерность данных, M) – не более 50, число объектов (N = произведение числа вариантов на число повторений) – произвольно, но при соблюдении условия M x N <= 32000.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х признаков, 2-х вариантов фактора "A", 3-х вариантов фактора "B" и 2-х повторений в текстовом файле:

	вариантов фактора "A", | вариантов фактора "B", | | повторений
4 12 2 3 2	<- начало файла
12,3 22,5 34,2 51,4	1a1b1п \ всего 2x3x2 = 12 вариантов
10,4 20,9 33,2 49,5	1a1b2п \
9,23 24,6 31,6 59,0	1a2b1п \
7,12 20,9 30,3 56,3	1a2b2п 1-й вариант фактора A
6,21 18,5 31,6 51,5	1a3b1п /
5,67 17,2 30,6 55,6	1a3b2п /
7,23 15,2 32,1 57,3	2a1b1п \
6,78 16,2 31,9 53,7	2a1b2п \
9,44 10,2 11,7 59,1	2a2b1п \
8,34 23,7 33,1 57,6	2a2b2п 2-й вариант фактора A
11,4 23,7 33,5 57,1	2a3b1п /
10,8 25,3 32,9 55,2	2a3b2п /
Данные 1998 г, Н-ск	<- необязательный комментарий

```
1----2----3----4------признаки
```

В качестве примера формирования массива можно посмотреть файл ARENS169.dat (2 признака, 3 варианта "A", 3 варианта "B", 4 повторения). Массивы данных, подготовленные в программах одномерного дисперсионного анализа, могут быть переданы для обработки программе MANOVA8 (см. ниже; кроме 1-факторных данных). Другие программы (VARS, NORMAL, MCOR, MCOM и т.п.) могут использовать массивы от MANOVA8. Массивы, созданные в программах MANOVA2, MANOVA8, могут быть использованы программой MANODISC.

В массиве данных не должно быть "отсутствующих" значений; если все же они есть, следует подставить вместо них средние из имеющихся повторностей, или же "восстановить" с помощью программы IODATA (с последующей корректировкой 1-й строки файла данных!).

Для каждого фактора вычисляется многомерный критерий для проверки 0-гипотезы: различия векторов средних между любой парой вариантов отсутствуют, фактор не влияет на исследуемую систему. Используется аппроксимация многомерного критерия к F-распределению по Рао [32], с вычислением вероятности ошибки в случае отклонения 0-гипотезы. Если

P<=0,01 0-гипотеза отвергается с уровнем значимости 1%, по крайней мере два варианта (два вектора средних) достоверно отличаются;

P<=0,05 0-гипотеза отвергается с уровнем значимости 5%;

P >0,10 0-гипотеза остается в силе: векторы средних отличаются только из-за действия случайных факторов.

0-гипотеза для проверки возможного взаимодействия факторов: отсутствует эффект синергизма или антагонизма факторов при любом сочетании вариантов факторов, которые действуют на исследуемую систему как простая сумма возможных эффектов. Значение вычисленной вероятности трактуется аналогичным образом.

Следует заметить, что число степеней свободы числителя F-отношения может быть дробным числом, в этом случае программа интерполирует к значению P-вероятности по целым значениям степеней свободы. Существуют "Таблицы F-распределения" [43], в которых можно найти значения F-критерия для дробных степеней свободы.

Перед выполнением анализа возможно указание метода организации эксперимента (полная рандомизация/случайные блоки повторений). Здесь же можно

задать выполнение дополнительных методов анализа: шагового информационного анализа вклада признаков, дискриминантного анализа для каждого эффекта, включая взаимодействие факторов, вычисление векторов средних для вариантов всех факторов.

Шаговый информационный анализ позволяет выяснить значение признаков с точки зрения вклада в различие вариантов фактора. В некотором смысле это "доля влияния признака" на различимость вариантов, причем достоверность вклада каждого признака оценивается критерием Hi^2. Влияние признаков приводится в долях единицы.

Для количественной оценки различимости векторов средних вариантов каждого фактора вычисляются дискриминантные функции вида

$D_1 = K_1*Y_1 + K_2*Y_2 + ... + K_m*Y_m$;

где Y_1, Y_2, ... Y_m – переменные-признаки,

K_1, K_2, ... K_m – коэффициенты дискриминантной функции.

Дискриминантных функций может быть несколько, они вычисляются и выводятся последовательно – по убыванию "разрешающей способности" относительно различимости вариантов фактора. Подставив значения средних для варианта, можно вычислить "дистант" – некоторое число, обобщенная характеристика варианта. Все варианты фактора образуют ряд дистантов, позволяющий отнести любой обьект к тому или иному варианту, вычислив дистант объекта. В этом случае возможна "двухфакторная" дискриминация обьекта – по любой паре факторов – в отличие от классической "однофакторной" дискриминации; в программе можно получить графическое представление одно- и двухфакторной дискриминации для множества всех обьектов массива данных. В первом случае ось "X" отражает значения первой дискриминантной функции, ось "Y" – второй, если число вариантов фактора больше 2-х. Если число вариантов равно двум, используется одномерная дискриминация обьектов, так как может быть вычислена только одна дискриминантная функция.

Во втором случае (двухфакторная дискриминация) для всех факторов используются первые дискриминантные функции.

Программа тестировалась по [31], стр. 169, [44], стр. 299. В основе программы использованы разработки А.И.Южакова (СибНИИЗХим).

8. Регрессионный анализ

Регрессионный анализ – вычисление коэффициентов различных функций – моделей процессов, зависимостей между параметрами систем, доказательство достоверности этих коэффициентов, функций (уравнений регресии).

Предпосылки применимости регрессионного анализа:

– независимая переменная X измерена с точностью, во много раз превосходящую точность измерения зависимой переменной Y

– нормальность распределения ошибок измерения зависимой переменной Y требуется **только для оценки достоверности** коэффициентов и уравнения;

– ошибки измерения зависимой переменной в последовательности измерений – статистически независимы (гомоскедастичность дисперсии, отсутствие автокорреляций в остатках).

Достоверность коэффициентов определяется по вероятности того, что соответствующий коэффициент равен нулю, вычисляемой на основе T-критерия Стьюдента:

P <= 0,01 коэффициент достоверен на уровне 1%;

P <= 0,05 коэффициент достоверен на уровне 5%;

P > 0,10 достоверность коэффициента не подтверждена.

Общая достоверность уравнения регрессии определяется F-критерием Фишера-Снедекора. 0-гипотеза формулируется следующим образом: отсутствует регрессионая связь между переменными, значения коэффициентов регрессии отличаются от нуля вследствие действия случайных факторов.

Для F-критерия вычисляется "вероятность ошибки в случае отклонения 0-гипотезы". Если

P <= 0,01 уравнение регрессии значимо на уровне 1%,

P <= 0,05 уравнение регрессии значимо на уровне 5%,

P > 0,10 регрессионная связь не доказана.

8.1. PRAN: Полиномиальный регрессионный анализ

Программа PRAN предназначена для обработки экспериментальных данных, представленных двумя связанными переменными, методом полиномиального регрессионного анализа. Используются:

1/ стандартная техника (метод наименьших квадратов) с инвертированием матрицы Грамма, сравнительно экономная по памяти;

2/ метод ортогональных полиномов Чебышева, позволяющий получать практически любую степень регрессионной зависимости, но за счет дополнительных ресурсов оперативной памяти; независимая переменная X может быть любого типа (неранжирована), с произвольными расстояниями между значениями;

3/ непараметрическая линейная регрессия, не требующая выполнения предпосылок классического регрессионного анализа.

Для 1-го и 2-го методов возможен подбор полинома оптимальной степени с визуальной оценкой графика регрессии по полю рассеяния экспериментальных данных. Для полинома любой степени возможно получение доверительных интервалов для Y, также с визуальной оценкой. Переменная Y может быть в нескольких повторениях, в том числе и с неравным числом повторений.

Ограничения на размер массива данных: число пар экспериментальной зависимости X - Y не должно превышать 5000. Максимальная степень полинома – 50, однако из-за алгоритмических проблем для конкретных данных большие степени обычно не достигаются.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 2-х переменных, 10-и пар значений в текстовом файле:

```
2 10
12,3 22,5
8,34 23,7
9,23 24,6
7,12 20,9
8,27 19,4
5,67 17,2
5,33 16,8
4,87 16,1
4,33 15,6
3,89 15,1

Опыт N123
```

<- начало файла

массив данных:
строки = объекты,
столбцы = признаки

<- необязательный комментарий

Если имеется файл с массивом данных типа "признаки-объекты" с числом признаков больше двух, программе PRAN можно указать любую пару признаков из этого массива для выполнения регрессионнного анализа. В качестве примера формирования массива для программ регрессионного анализа можно посмотреть файлы PRAN.dat (2 переменных, 15 вариантов), K8_MLRG.dat (8 незав. переменных, зависимая переменная в 4-х повторениях). Пример ввода массива данных с

повторениями зависимой переменной в 4-м и 5-м столбцах, независимые переменные в первых трех столбцах:

```
5 10 3 2                                  <- начало файла
7,31 22,5 35,2 57,4 55,8                  3 независимых переменных,
8,34 23,7 36,3 56,2 56,3                  зависимая переменная в двух
8,73 24,6 37,5 57,7 57,9                  повторностях;  3+2=5
9,12 20,9 38,5 57,8 58,1
9,27 19,4 39,2 57,9 58,8                  массив данных:
9,51 18,5 40,3 58,2 58,4                    столбцы = переменные,
9,67 17,2 41,4 58,3 59,2                    строки  = значения
9,84 16,8 42,5 58,5 59,3
9,88 15,4 43,6 58,9 58,6
9,93 15,1 44,8 58,7 59,3

Данные 1998 г                             <= необязательный комментарий
  X1    X2    X3 |      Y      |
```

Математическая модель данных в полиномиальном регрессионном анализе:

$$Y_i = b_0 + b_1 * x_i + b_2 * x_i^2 + ... + b_m * x_i^m + e_i$$

m – степень полинома, целое число >0;

$b_0, b_1, \ldots b_m$ – коэффициенты полинома, подлежащие оценке;

e_i – ошибка измерения Y, распределена по $N(0, \sigma)$, $cov(e_i, e_j)=0$.

Программа начинает обработку данных с полинома 2-й степени, далее, после вывода результатов на дисплей, можно менять степень полинома кликами по <+> и <->. Качество подгонки регрессионного уравнения к экспериментальным данным можно оценить визуально, активизируя график регресии. Доверительный "коридор" в интервале изменения независимой переменной формируется в зависимости от выбранного уровня значимости – 5% (по умолчанию) или 1%.

Для проверки достоверности полученного уравнения регрессии вычисляется критерий Фишера-Снедекора. В результате работы программы вычисляются:

1. Коэффициенты регрессии: характеристика связи между данной степенью входной переменной и Y; их достоверность определяется Т-критерием Стьюдента.

2. Коэффициент множественной детерминации – доля общей дисперсии зависимой переменной, объясняемая уравнением регрессии.

3. Если задан полином первой степени (простая линейная регрессия), выводится коэффициент парной корреляции и его стандартная ошибка.

4. Если зависимая переменная имеет повторности, это позволяет оценить ”адекватность” полученного уравнения – степень соответствия вычисленным по уравнению значениям Y – полученным в эксперименте. Оценка адекватно-

сти производится по критерию Фишера с соответствующим значением вероятности того, что уравнение адекватно. Чем ближе вероятность к 1, тем лучше уравнение регресии отражает действующие в эксперименте зависимости.

5. Стандартная таблица дисперсионного анализа.

6. Таблица доверительных интервалов для пяти точек: среднего независимой переменной, минимума и максимума, плюс/минус 1/6 размаха от среднего.

Пример обработки данных, полином 1-й степени:

1. Коэффициенты регрессии, анализ достоверности.

	Коэффициенты регресии	Стандартные ошибки	Критерий Стьюдента	Вероятность ошибки 1 рода
B0	9,3423600374	3,8462864	2,4289	0,01768*
B1	0,7621945001	0,0692772	11,002	0,00000*

Степеней свободы для критерия Стьюдента = 71

2. Общие критерии достоверности регрессии.

Y = B0 + B1*X

Критерий Фишера: F= 121,046 степени свободы: 1, 71
Вероятность нулевых значений коэффициентов: P=0,0000
Относительная ошибка аппроксимации: Er = 85,166%
Коэффициент детерминации: BY = 0,63030
Коэффициент парной корреляции: R = 0,79391
Стандартная ошибка коэффициента: S = 0,07216

Уравнение линейной регрессии достоверно на высоком уровне значимости (P<0,0001), как и коэффициенты регрессии (B0 и B1). Если увеличить степень полинома, значение критерия Фишера-Снедекора уменьшится, значение коэффициента B2 при квадратичном члене недстоверно (P>0,1):

1. Коэффициенты регрессии, анализ достоверности.

	Коэффициенты регресии	Стандартные ошибки	Критерий Стьюдента	Вероятность ошибки 1 рода
B0	7,0729746351	5,0624454	1,3971	0,16678
B1	0,8970301454	0,2066318	4,3412	0,00005*
B2	-0,0012296722	0,0017745	0,6930	0,49063

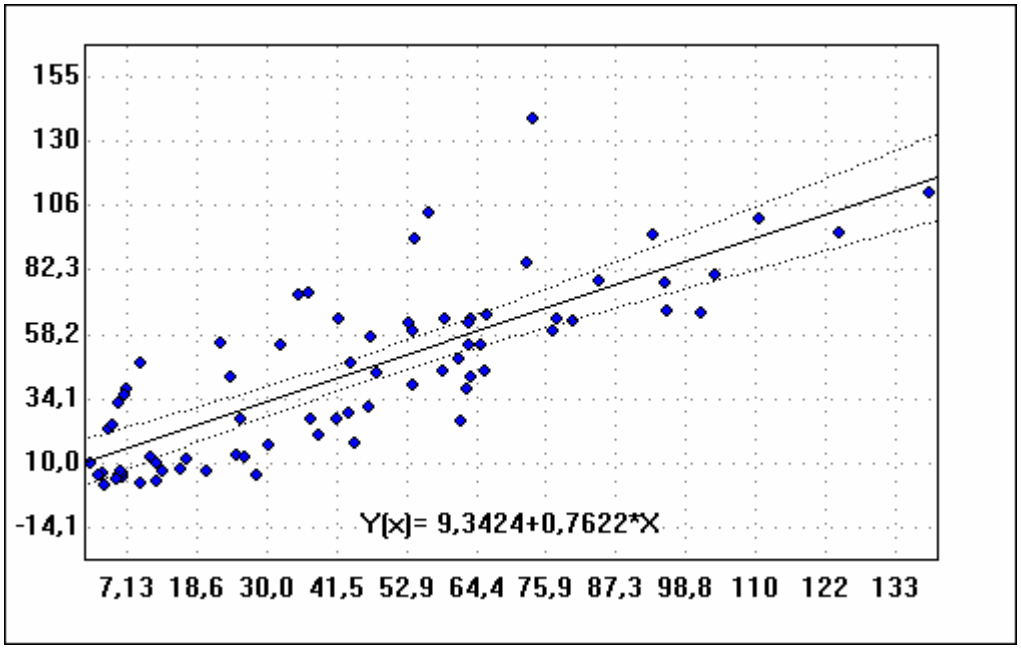
Степеней свободы для критерия Стьюдента = 70

2. Общие критерии достоверности регрессии.

Y = B0 + B1*X + B2*X^2

Критерий Фишера: F= 60,3201 степени свободы: 2, 70
Вероятность нулевых значений коэффициентов: P=0,0000
Относительная ошибка аппроксимации: Er = 80,626%
Коэффициент детерминации: BY = 0,63282

Хотя F-критерий для полинома второй степени значителен, значение коэффициента B2 фактически нулевое, поэтому следует принять в качестве уравнения связи переменных – полином 1-й степени – линейную регрессию:



8.1.1. Ортогональные полиномы

Техника ортогональных полиномов базируется на следующих формулах (по Линнику, [44], стр. 278-279):

$F_0(x)=1.0;$ $\quad Y = S_0 \cdot F_0;$

$F_1(x)=x-A_0;$ $\quad Y = S_0 \cdot F_0+S_1 \cdot F_1;$

$F_2(x)=x^2-B_1 \cdot F_1-B_0;$ $\quad Y = S_0 \cdot F_0+S_1 \cdot F_1+S_2 \cdot F_2;$

$F_3(x)=x^3-C_2 \cdot F_2-C_1 \cdot F_1-C_0;$ $\quad Y = S_0 \cdot F_0+S_1 \cdot F_1+S_2 \cdot F_2+S_3 \cdot F_3;$

$F_4(x)=x^4-D_3 \cdot F_3-D_2 \cdot F_2-D_1 \cdot F_1-D_0;$ $\quad Y = S_0 \cdot F_0+S_1 \cdot F_1+S_2 \cdot F_2+S_3 \cdot F_3+S_4 \cdot F_4;$

. . . . . .

$F_n(x)=x^{m+1}-N_m \cdot F_m- \ldots -N_0;$ $\quad Y = S_0 \cdot F_0+S_1 \cdot F_1+ \ldots +S_n \cdot F_n;$

$F_0..F_n$ – полиномы возрастающих степеней, коэффициенты которых подобраны некоторым специальным образом. Исходная регрессионная модель в виде полинома некоторой степени от X преобразуется в сумму полиномов F_1, F_2, ..., F_n, где n – желаемая степень полинома. В программе вычисляются коэффициенты $S_0..S_n$ при ортогональных полиномах, критерии достоверности этих коэффициентов, затем рассчитываются собственно коэффициенты регрессии при возрастающих степенях независимой переменной:

$$Y = B_0 + B_1*X + B_2*X^2 + \ldots + B_n*X^n$$

при этом различные статистические тесты относительно этих коэффициентов отсутствуют в силу специфики метода. При необходимости эти значения (стандартные ошибки, Т-критерии и т.п.) могут быть получены в большинстве случаев с помощью стандартной полиномиальной регрессии.

Следует заметить, что для некоторых типов данных корректные оценки регрессионных коэффициентов удается получить только с помощью полиномов Чебышева; стандартный метод выдает значения коэффициентов с большой погрешностью из-за вырожденности матрицы Грамма.

Программа начинает обработку данных с полинома второй степени, после вывода результатов на дисплей можно менять степень полинома комбинацией клавиш <Ctrl/+> и <Ctrl/->. Качество подгонки регрессионного уравнения к экспериментальным данным можно оценить визуально. Доверительный "коридор" в интервале изменения независимой переменной формируется в зависимости от выбранного уровня значимости – 5% (по умолчанию) или 1%.

Программа тестировалась по [45], стр. 288-290, тестовый массив LINN2x16.dat, [23], стр. 400-402, массив RAIN2x12.dat. Доверительные интервалы для Y рассчитываются по [2], стр. 351.

8.1.2. Непараметрический регрессионный анализ

Использован метод оценки коэффициента линейной регрессии по Тейлу-Сену (H.Theil, 1950, P.Sen, 1968), изложенный в [5], стр. 219-220.

Непараметрический регрессионный анализ следует использовать, когда у исследователя имеются серьезные сомнения в справедливости предпосылок классического регрессионного анализа:

– независимость наблюдений переменной отклика;

– одинаковая распределенность ошибок измерения переменной отклика в экспериментальных точках;

– нормальность распределения ошибок измерения переменной отклика.

Эти предпосылки могут нарушаться в экспериментах, связанных с измерением параметра, изменяющегося во времени, дрейфом неконтролируемых факторов эксперимента. Оценка коэффициента B1 линейной регрессии

Y = B0 + B1*X

непараметрическим методом позволяет сделать надежное определение значения коэффициента, устойчивое к выбросам и артефактам в экспериментальных данных. Свободный член оценивается следующим образом:

$$B0 = \frac{1}{N}\sum_{i=1}^{n}(Y_i - B1 \times X_i)$$

где X_i и Y_i – экспериментальные данные, N – число пар измеренных точек.

Для коэффициента регрессии определяется доверительный интервал (также непараметрическим методом) на заданном уровне значимости, доказательство достоверности регрессионной зависимости базируется на анализе достоверности коэффициента корреляции рангов по Кендаллу с помощью Z-критерия (приближение нормальным распределением удовлетворительно работает для выборок > 10 пар X и Y).

Дополнительно выполняется классический анализ достоверности регрессии с помощью критерия Фишера-Снедекора, а для данных с повторениями хотя бы в некоторых точках эксперимента – анализ адекватности уравнения регрессии также критерием Фишера-Снедекора.

Корректность работы программы проверялась по тестовым массивам из [60], TURIN6x7.dat, из [5], Hollender2x5.dat.

8.1.3. Bootstrap-процедура для полиномиальной регрессии

Для оценки вариабельности коэффициентов регрессии можно использовать модификацию метода "Bootstrap", предложенного Б.Эфроном в 1977 г.

Например, план в виде массива из независимой переменной X: в N точках измерены значения переменной отклика Y, вычислены значения коэффициентов полиномиальной линейной регрессии. Если по какой-то причине исследователя не удовлетворяют значения стандартных ошибок, полученных на основе классической теории, можно оценить их вариабельность следующим образом.

1. Задается некоторое большое число (500..1000..10000), определяющее, сколько раз генерировать случайные двумерные выборки (того же размера N) из значений имеющегося массива данных.

2. Методом Монте-Карло генерируются эти выборки, и каждый раз вновь вычисляются коэффициенты регрессии. Эти оценки накапливаются в массивах.

3. Помимо коэффициентов регресии, каждый раз вычисляется значение прогноза по заданному значению X, эти значения также накапливаются в массиве.

4. Для этих массивов вычисляются средние, доверительные интервалы, экстремумы, квантили, медиана, мода, на основании которых можно судить о вариабельности коэффициентов регрессии и значений прогноза.. Доверительные интер-

валы для средних вычисляются по формуле (**n** – число bootstrap-генераций, **T** – критерий Стьюдента):

$$\mathbf{B} = \bar{\mathbf{b}} \pm \mathbf{T}_{(1-\alpha,n-1)} \cdot \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (b_i - \bar{b})^2}$$

Дополнительно может быть сделан анализ вариабельности коэффициентов регрессии и результатов вычисления отклика по уравнениям регрессии методом "складного ножа" (М.Кенуй). При этом вычисления коэффициентов регрессии выполняются N раз – с последовательным исключением по одной точке плана X_i-Y_i значений. Для исключенного значения X_i вычисляется значение отклика по получаемым уравнениям, и сравнивается со значением отклика, полученным в эксперименте.

Таким образом, можно сделать некоторые выводы о "качестве" уравнения регрессии.

Также может быть вычислен коэффициент парной корреляции между значениями отклика, полученным в эксперименте и вычисленным по уравнениям регрессии на каждой стадии метода "складного ножа", и оценена его достоверность критерием Стьюдента. Высокое и достоверное значение коэффициента корреляции предоставляет дополнительный аргумент для использования полученного уравнения регрессии для оценивания отклика в любой промежуточной точке плана, а также в некоторой окрестности за интервалами изменения независимых переменных, в частности для вычисления прогноза временных зависимостей.

Для дополнительного анализа результатов бутстреп-процедуры возможен визуальный контроль распределения значений бутстреп-выборки прогнозов отклика по значению X, задаваемому пользователем.

Значения прогноза, вычисленные по трем видам регрессий (МНК, бутстреп, Jack-Nife) сопровождаются доверительными интервалами с вероятностью попадания среднего значения Y 95% и стандартными ошибками. При этом используется формула вычисления доверительных интервалов ([4], стр. 172):

$$\mathbf{Y} = \mathbf{y}_{(x,p)} \pm \mathbf{T}_{(1-\alpha/2,n-p-1)} \cdot \mathbf{S}\sqrt{1/n + \mathbf{d}'\mathbf{A}^{-1}\mathbf{d}};\ \mathbf{d} = (x - \bar{x}_{n,1}, x^2 - \bar{x}_{n,2}, ..., x^p - \bar{x}_{n,p})$$

p – степень полинома, **n** – число пар X–Y, **S** – остаточный средний квадрат из таблицы дисперсионного анализа, **d** – вектор отклонений X_i от средних значений, **A** – матрица Грамма, **T** – критерий Стьюдента с (n-p-1) степенями свободы. Стандартная ошибка прогноза вычисляется по формуле:

$$\mathbf{Y = y_{x,p} \pm S \cdot \sqrt{n}}$$

Для бутстреп-процедуры степень полинома не может превышать 5, степень полинома меняется выбором соответствующих пунктов из Меню дополнительных операций, открываемым **правой** клавишей мышки в поле результатов анализа.

8.2. HARMON: Гармонический регрессионный анализ

Программа HARMON предназначена для обработки временных рядов методом гармонического регрессионного анализа. Используется эффективная модификация метода наименьших квадратов ([33], стр. 85) для ряда с известным периодом, если измерения сделаны через равные промежутки времени. Возможен подбор тригонометрического полинома оптимальной степени с визуальной оценкой графика регрессии по полю рассеяния экспериментальных данных. Анализ произвольных парных зависимостей типа X-Y также может быть выполнен, но при условии, что независимая переменная (X) представляет из себя равномерно возрастающий ряд чисел типа

54, 58, 62, 66, ... или 1.22, 1.25, 1.28, 1.31, ...

Эффективность метода гармонической регрессии заключается в устойчивом равномерном снижении остаточной дисперсии при добавлении следующих гармоник, в отличие от стандартного метода – степенной полиномиальной регрессии.

Если ставится задача **только** аппроксимации произвольного ряда для вычисления значения переменной в любой точке интервала изменения независимой переменной (включая начальные и конечные участки), может быть использовано эффективное сглаживание ряда двойным преобразованием Фурье (прямым и обратным) с переменным числом точек сглаживания. Графический анализ позволяет выбрать наиболее приемлемый вид аппроксимации ряда, значение ряда в любой точке затем определяется с помощью кубических сплайнов.

Качество аппроксимации оценивается коэффициентом детерминации, ошибкой, среднеквадратическим и средним абсолютным отклонениями.

Массив данных может содержать не более 100000 элементов: до 100 временных рядов, до 5000 значений в ряде. Максимальная степень полинома (h) – не более 30 пар коэффициентов ряда Фурье, но при условии N>(2*h). Столбец, передаваемый программе для обработки, должен содержать значения временно-

го ряда за целое число периодов (последнее значение ряда примерно равно первому). В принципе можно обрабатывать любые ряды, но при этом на концах ряда возможны значительные отклонения вычисленных по регрессии значений от экспериментальных точек.

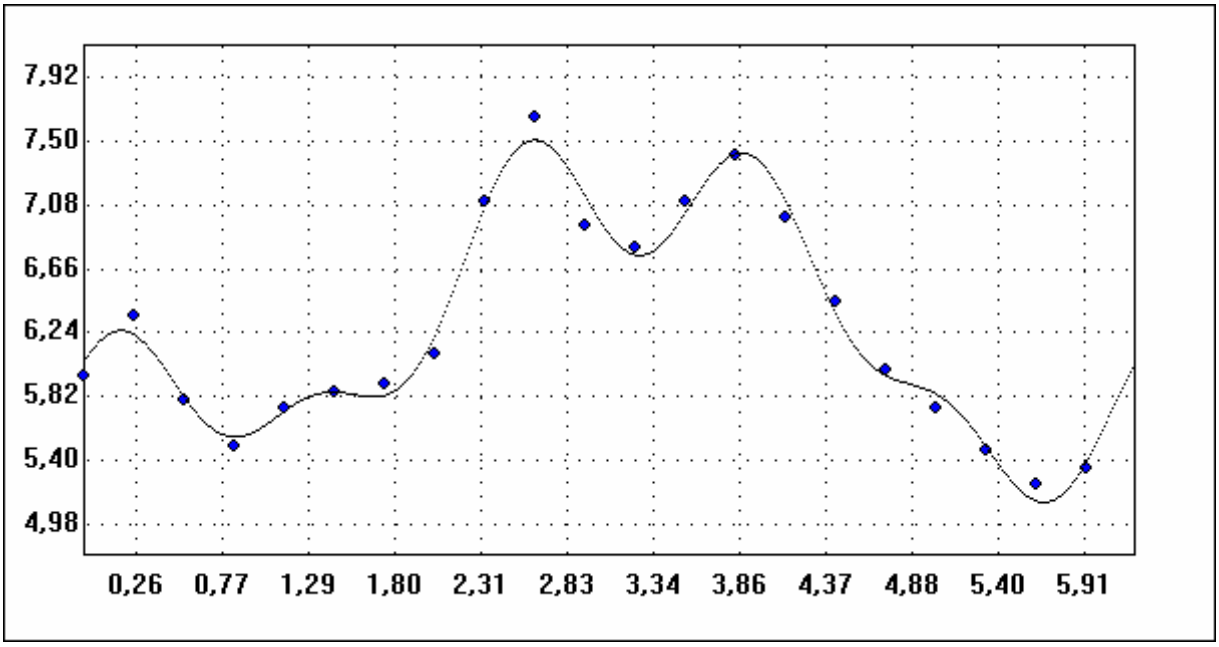
Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Если имеется файл с массивом данных типа "признаки-объекты", программе HARMON можно указать любой признак (столбец) из этого массива для выполнения гармонического анализа. В качестве примера формирования массива можно посмотреть файлы SUN2x78.dat (динамика чисел Вольфа за 78 лет, около 7 периодов), HARMON.dat, [33], стр. 90, период 0.32 сек.:

```
3 21                    <- 3 столбца, 21 строка
1  0,01524  5,95
2  0,03048  6,35
3  0,04571  5,80        n – номер точки измерения,
4  0,06095  5,50        X – отсчеты времени,
5  0,07619  5,75        Y – значения измеряемого параметра
6  0,09143  5,85
7  0,10667  5,90
8  0,12190  6,10
9  0,13714  7,10
10 0,15238  7,65
11 0,16762  6,95
12 0,18286  6,80
13 0,19810  7,10
14 0,21333  7,40
15 0,22857  7,00
16 0,24381  6,45
17 0,25905  6,00
18 0,27429  5,75
19 0,28952  5,47
20 0,30476  5,25
21 0,32000  5,35
```

n X Y

Подразумевается, что все значения ряда отражают изменения некоторого параметра за главный период 2*π, измеренного через равные промежутки времени, причем начальное значение (в массив можно не вводить) должно быть равно конечному. Перед выполнением анализа можно указать программе числовое значение периода (в секундах, днях, годах и т.п.), тогда в распечатке будет выводиться значение переменной X. Программа начинает обработку данных с полинома третьей степени, далее, после вывода результатов на дисплей, можно менять степень полинома клавишами «+» и «-». Качество подгонки регрессионного урав-

нения к экспериментальным данным можно оценить визуально, выбрав в Меню "График регрессии":



Гармоническая регрессия для любого момента времени имеет следующий вид:

$$Y(t) = A_0 + \sum_{i=1}^{h}\left[A_i \times Cos(2\times\pi\times i\times t/T) + B_i \times Sin(2\times\pi\times i\times t/T)\right]$$

h – степень тригонометрического полинома (число гармоник, обычно >2);

i – номер гармоники;

t – произвольный момент времени;

T – период процесса;

$\pi = 3.14159265$;

A_0, A_1, B_1,... A_h, B_h – коэффициенты регрессии (ряда Фурье), вычисляемые программой:

Номер гармоники	Коэффициент A(i)	Критерий Стьюдента	Коэффициент B(i)	Критерий Стьюдента	Амплитуда R(i)	Фаза P(i)
0	6,26048		0,00000		6,26048	0,0000
1	-0,82155	21,23*	-0,07431	1,928	0,82491	-5,1683
2	0,26737	6,935*	0,18335	4,757*	0,32420	-34,440
3	0,19757	5,126*	0,21658	5,619*	0,29316	-47,627
4	-0,06531	1,695	-0,04279	1,110	0,07808	-33,229
5	0,20911	5,425*	0,20496	5,317*	0,29281	-44,427

```
Степеней свободы T-критерия = 10
Коэффициент детерминации = 0,9848
Относительная ошибка аппроксимации = 1,138%
```

Для проверки достоверности полученного уравнения регрессии вычисляется критерий Фишера-Снедекора. 0-гипотеза формулируется следующим образом: отсутствует функциональная зависимость исследуемого параметра (в виде суммы гармоник) от времени, значения коэффициентов регрессии отличаются от нуля вследствие действия случайных факторов.

В результате работы программы вычисляются:

1. Коэффициенты регрессии: характеристика связи между данной гармоникой и Y; их достоверность определяется по Т-критерию Стьюдента; символом '*' отмечаются коэффициенты, значимые на уровне 5%.

2. Коэффициент детерминации – доля общей дисперсии зависимой переменной, объясняемая уравнением регрессии.

3. Стандартная таблица дисперсионного анализа ANOVA.

Дисперсия	СуммаКвадратов	Ст.Своб.	СреднийКвадрат	F-критерий	Вероятность
Общая	10,2710957	20	0,5136		
Регресия	10,1151735	10	1,0115	64,8732	0,00000
Ошибка	0,15592226	10	0,0156		

Программа тестировалась по [33], стр. 93.

8.3. NLREG: Нелинейный регрессионный анализ

Программа NLREG предназначена для обработки экспериментальных данных методами нелинейного регрессионного анализа. Возможен подбор оптимального типа регрессии по различным критериям качества: минимуму остаточной дисперсии, максимуму критерия Фишера-Снедекора или коэффициента детерминации, с визуальной оценкой модели по графику регрессии с коридором доверительных интервалов на поле рассеяния экспериментальных данных.

Для определения коэффициентов регрессии можно выбрать:

1/ стандартный метод линеаризации функции с использованием метода наименьших квадратов;

2/ метод Ньютона – последовательное приближение коэффициентов с минимизацией суммы квадратов отклонений, с обращением матрицы частных производных (матрицы Якоби);

3/ симплекс-метод минимизации суммы квадратов отклонений Нелдера-Мида; в этом методе не требуется вычисления частных производных, но число итераций значительно больше;

4/ ручная аппроксимация экспериментальной зависимости выбранной функцией с помощью постепенного подбора коэффициентов с визуальным анализом качества приближения.

В простых случаях все методы дают практически одинаковые оценки коэффициентов регрессии, для более сложных функций итерационные методы могут дать более эффективные оценки коэффициентов, большую величину критерия Фишера-Снедекора, коэффициента детерминации. Однако, в некоторых случаях методы минимизации не могут быть применены из-за сложности алгоритмов, возникают проблемы машинной арифметики.

Некоторые преимущества метода линеаризации:

– минимум проблем вычислительного характера,

– для каждого коэффициента регрессии вычисляется стандартная ошибка и критерий Стьюдента, определяющий достоверность отличия коэффициента от нуля,

– можно вычислить доверительные интервалы для Y, построить коридор доверия значений отклика на графике регрессии.

Ограничения на размер массива данных: число пар экспериментальной зависимости X-Y не должно превышать 25000, число переменных – не более 200, весь массив – не более 100000 значений. Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 2-х переменных, 10-и пар значений в текстовом файле:

```
2 10
7,31  22,5
8,34  23,7
8,73  24,6
9,12  20,9
9,27  19,4
9,51  18,5
9,67  17,2
9,84  16,8
9,88  15,4
9,93  15,1

Данные 1998 г
```

<- первая строка файла;

массив данных:
столбцы = 2 переменные,
строки = 10 пар значений X(i)-Y(i).

<= необязательный комментарий

Если имеется файл с массивом данных типа "признаки-объекты" с числом признаков больше двух, программе NLREG можно указать любую пару признаков из этого массива для выполнения регрессионнного анализа. В качестве

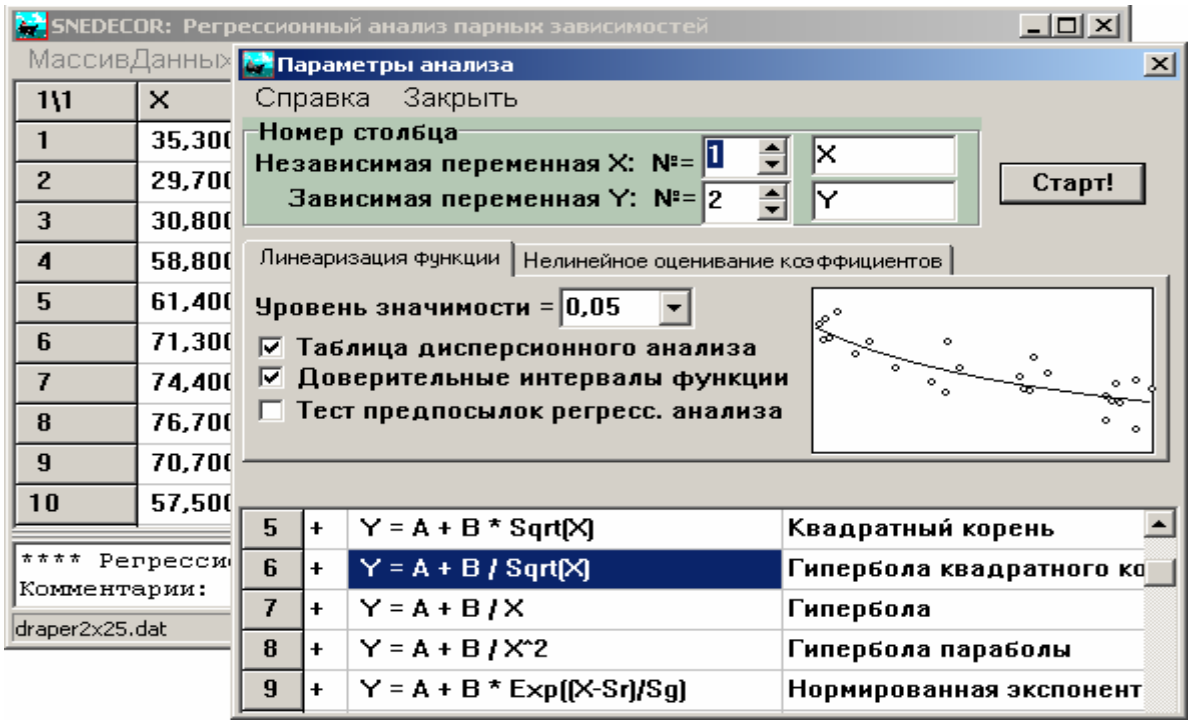
примера формирования массива для программ регрессионного анализа можно посмотреть файлы PRAN.dat (2 переменных, 15 вариантов), ABC8x50.dat.

Программа NLREG предлагает набор из 33 модельных функций, из которого пользователь должен выбрать наиболее отвечающую экспериментальным данным. Используется метод линеаризации – преобразования нелинейной зависимости к линейным уравнениям типа

F(z) = A + B*z или **F(t,z) = A + B*t + C*z;**

где z и t – новые переменные (различные элементарные функции от исходной переменной X). Методом наименьших квадратов вычисляются коэффициенты линеаризованного уравнения регрессии, достоверность которых определяется по стандартной ошибке и T-критерию Стьюдента, затем они в соответствии с типом модельной функции преобразуются в коэффициенты нелинейного уравнения.

Перед анализом программа проверяет все значения переменных X и Y на равенство нулю и 1.0, отрицательность; для некоторых типов нелинейной регрессии эти значения могут быть недопустимыми (деление на ноль, корень из отрицательного числа и т.п.). В этом случае программа блокирует вычисления по соответствующим типам нелинейной регрессии, в списке модельных уравнений их можно распознать по знаку '–' в левом столбце таблицы функций:



Качество подгонки регрессионного уравнения к экспериментальным данным можно оценить визуально, активизируя график регресии. Доверительный

"коридор" в интервале изменения независимой переменной формируется в зависимости от выбранного уровня значимости – 5% (по умолчанию) или 1%.

Для проверки достоверности полученного уравнения регрессии вычисляется F-критерий Фишера-Снедекора. В некоторых случаях F-критерий для нелинейной регрессии не может быть вычислен из-за превышения дисперсии от регрессии над общей дисперсией зависимой переменной, в этом случае следует использовать другое модельное уравнение.

В результате работы программы также вычисляются:

1. Коэффициенты нелинейной регрессии: характеристика связи между данной функцией входной переменной и Y; если значения коэффициентов нелинейной регрессии и линеаризованного уравнения совпадают, все оценки достоверности справедливы и для коэффициентов нелинейной регрессии, в противном случае Т-критерий может быть только аргументом в пользу той или иной модели.

2. Коэффициенты линеаризованного уравнения регрессии; их достоверность определяется по стандартной ошибке (должна быть значительно меньше коэффициента), а также по Т-критерию Стьюдента.

3. Коэффициент множественной детерминации: доля дисперсии зависимой переменной, объясняемая уравнением нелинейной регрессии (массив DRAPER2x25.dat):

	Коэффициенты нелин.регрессии	Коэффициенты линеариз.уравн.	Стандартные ошибки	Критерий Стьюдента	Вероятность coef=0.0
A	1,7827765	1,7827765	0,9594	1,8582	0,0760
B	53,004917	53,004917	6,5501	8,0922	0,0000*

Степеней свободы для критерия Стьюдента = 23

2. Общие критерии достоверности регрессии.
Исходная функция: **Линеаризованная функция:**

Y = A + B / Sqrt(X) **F(z) = Y = A + B * Z**

Критерий Фишера-Снедекора:
Для нелинейной регрессии: F= 65,48, ст.св.=1, 23 P=0,0000
Относительная ошибка аппроксимации Er= 7,6035%
Коэффициент детерминации: BY= 0,74007

Критерием Фишера-Снедекора подтверждена достоверность выбранного уравнения регрессии на высоком уровне значимости ($P<0.0001$), достоверность коэффициента B подтверждена критерием Стьюдента.

4. Стандартная таблица дисперсионного анализа; оценка остаточной дисперсии. Минимальное значение дисперсии из набора регрессионных зависимостей также может служить основанием для выбора наилучшего типа регрессии.

5. Пять доверительных интервалов для значений зависимой переменной, соответствующих 5 значениям независимой переменной: минимуму, максимуму, среднему, (среднее – Delta), (среднее + Delta), на заданном уровне значимости (по умолчанию используется уровень 5%). Так как доверительные интервалы для нелинейной регрессии вычисляются на базе интервалов линеаризованной регрессии, в некоторых случаях возможны некорректные значения из-за точек разрыва нелинейной функции (Y -> бесконечность).

8.3.1. Выбор стартовых параметров минимизации

Вследствие циклического характера методов шаговой минимизации необходимо указать программе предельное число итераций, после которого программа прекращает вычисления и выводит результаты.

На простых тестовых задачах минимум суммы квадратов отклонений достигается за 5-6 шагов. В сложных случаях – медленной сходимости процесса – минимум может быть достигнут после нескольких сотен или тысяч итераций. В этом случае не следует помечать параметр “Минимизация шаг-за-шагом” из-за возможного исчерпания емкости буфера текста.

Дополнительно можно уточнить значение минимума суммы квадратов отклонений, при достижении которого также прекращаются итерации. Это значение сугубо специфично для конкретных данных, и может быть выбрано путем последовательного перебора значений какого-либо ряда. При явном зацикливании работы алгоритма следует прекращать его работу комбинацией клавиш “Ctrl/Alt/Del” (однократно!) с последующим удалением программы NLREG из памяти компьютера.

Весьма важно указать разумные стартовые значения коэффициентов функции A, B и, возможно, C. Эти значения можно ввести по результатам стандартного регрессионного анализа (методом линеаризации), если он может быть выполнен.

Для метода Нелдера-Мида дополнительно предлагается уточнить значения параметров “Отражения/Сжатия/Растяжения” симплекса. Обычно они не требуют корректировки – значения установлены на основе рекомендаций программистов МехМата МГУ, но в некоторых случаях возможно ускорение сходимости алгоритма

.

8.4. SPLINE: Аппроксимация парных зависимостей сплайном

Программа SPLINE предназначена для аппроксимации значений различных функций типа Y=F(X), заданных таблично, последующего вычисления функции для произвольных значений аргументов в интервале изменения X. Исходные данные могут быть результатами эксперимента в виде парных значений типа $X_i - Y_i$. Такая аппроксимация может быть применена в тех случаях, когда стандартная регрессия методом наименьших квадратов (полиномиальная, гармоническая и т.п.) неприменима или неэффективна.

Парная зависимость приближается совокупностью полиномов 3-й степени, гладко проходящих через выбранные специальным образом точки – узлы сплайна.

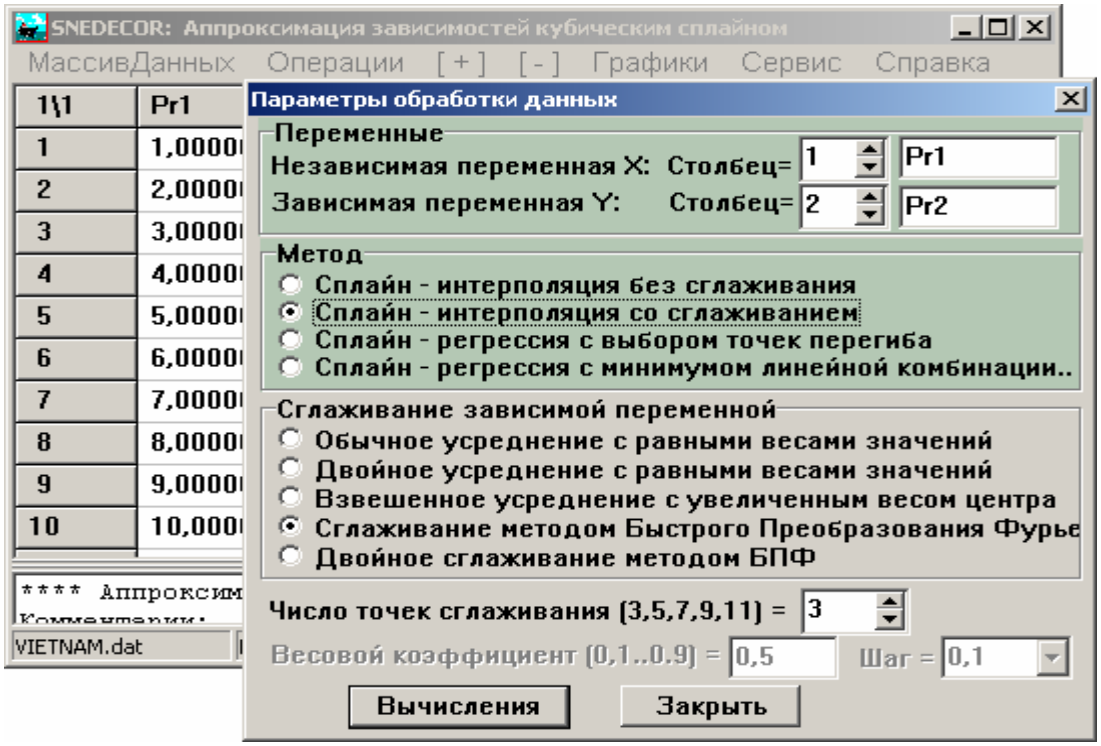
Для вычисления коэффициентов сплайн-функции в интервале изменения независимой переменной могут быть выбраны несколько методов:

1) **сплайн-интерполяция** без сглаживания значений Y;

2) **сплайн-интерполяция** со сглаживанием Y 5-ю способами;

3) **сплайн-регрессия** с минимизацией суммы квадратов отклонений [56];

4) **сплайн-регрессия** с минимизацией суммы квадратов отклонений и интеграла квадратов вторых производных в узлах сплайна [57].

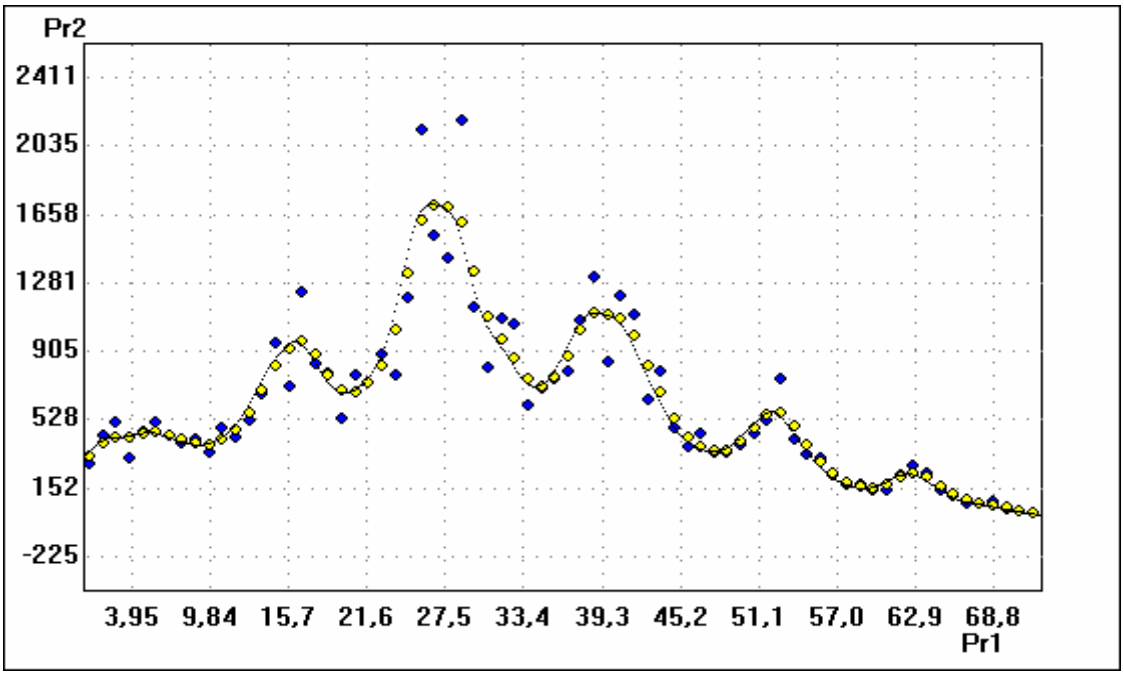


Помимо объективного критерия качества аппроксимации тем или иным методом (минимум остаточной дисперсии), возможен визуальный анализ качества сплайна на графике.

В каждом методе имеется возможность ручной подгонки сплайна для экспериментальной зависимости, исходя из различных экспертных или интуитивных соображений.

Следует отметить, что программа **не предназначена для экстраполяции** значений зависимой переменной за пределы диапазона изменений независимой переменной (для цели прогноза, например), так как на поведение концевого кубического полинома практически никак не влияют точки (и, соответственно, полиномы) предшествующих интервалов (кроме предпоследнего). Для этой задачи следует использовать программу PROGNOZ или специализированные пакеты программ.

Последние два метода (сплайн-регрессия) весьма эффективны для любых типов данных, позволяют построить сплайн, оптимальным образом отражающий специфику экспериментальной зависимости с элиминацией возможных погрешностей измерения. Для определения наилучшего сплайна нужно менять разбиение независимой переменной на интервалы с помощью пунктов Меню [+] и [–]. Первый увеличивает число узлов сплайна, второй уменьшает вплоть до полного использования всех точек. Оптимум определяется по графику – визуальным анализом качества сплайна, а также значением средней ошибки аппроксимации, среднего абсолютного отклонения, остаточной дисперсии. На графике: положение исходных экспериментальных точек отражается синими окружностями, узлов сплайна – желтыми (файл данных Vietnam.dat):



После определения коэффициентов сплайн-функции можно вычислить значение Y для произвольного аргумента X, выбрав соответствующий пункт в Меню “Операции”.

Программа может использовать массив типа "признаки-объекты", из которого любую пару признаков можно использовать как X и Y. Перед обработкой пары значений сортируются по возрастанию значений X.

Ограничения на размер массива данных: число точек временной зависимости не должно превышать 10000. Общее число элементов массива – не более 100000. В массиве "X" не должно быть одинаковых значений, в противном случае вычисления могут блокироваться.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры, переданы через буфер Windows из программы MS Excel, загружены из файла в стандарте пакета SNEDECOR. Пример формирования массива из 2-х переменных, 15 пар значений в текстовом файле (первым признаком здесь введен вектор отсчетов времени):

```
2 15
1961 22,5
1962 23,7
1963 24,6
1964 20,9
1965 19,4
1966 18,5
1967 17,2
1968 17,1
1969 17,9
1970 18,3
1971 18,5
1972 18,7
1973 18,6
1974 18,8
1975 20,2

Урожай зерновых
```

<- начало файла

массив данных:
строки = последов. отсчеты времени,
столбцы = временные ряды, признаки

<- необязательный комментарий

В качестве примера формирования массива можно посмотреть файлы VIETNAM.dat, SPLIN2x20.dat, SPLIN2x27.dat. Эти массивы следует использовать для изучения специфики работы со сплайн-функциями.

8.4.1. Методы сплайн-функций

Парная зависимость аппроксимируется последовательностью полиномов 3-й степени, гладко проходящих через выбранные специальным образом точки. Перед построением сплайна данные сортируются по возрастанию переменной X.

1. Сплайн-интерполяция без сглаживания значений Y.

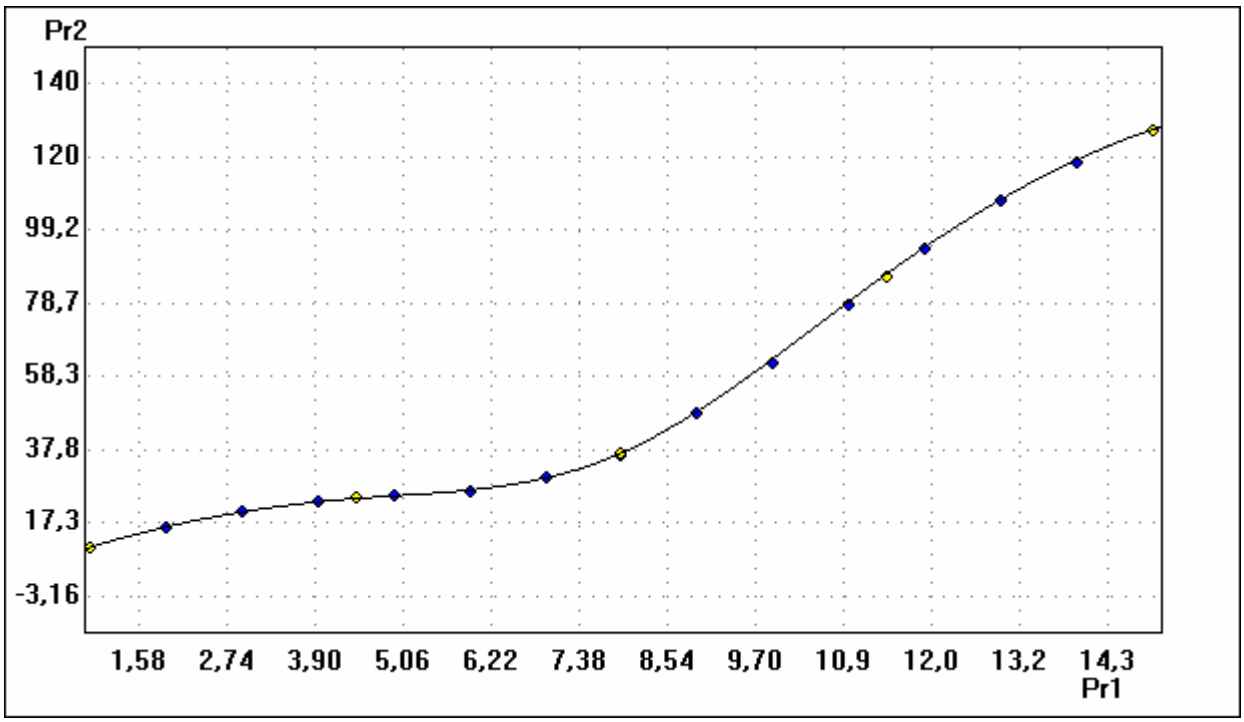
Узлы сплайна выбираются непосредственно по значениям Y – через одну, две, три и т.д. точки экспериментальной зависимости. Выбор числа узлов для вычисления коэффициентов сплайна является ответственной операцией. В [2, стр. 331] рекомендуется минимизировать число узлов, но в интервалах между узлами иметь не менее 4..5 значений. Следует заметить, что первая и последняя точки всегда используются программой в качестве узлов, поэтому коррекцией значений Y, соответствующим крайним значениям независимой переменной, можно в определенной степени управлять наклоном концов сплайна. Этот метод может быть рекомендован для сравнительно “гладких” зависимостей, не засоренных выбросами и ошибками измерения (массив PRAN.dat, сплайн по 5 точкам):

Отклонение от исходных данных.

i	X(i)	Y экспер.	Y вычисл.	Абс.Отклон.	Отн.Отклон.
1	1,0000	10,000	10,039	-0,0394	-0,394%
2	2,0000	16,000	15,862	0,1381	0,863%
3	3,0000	20,000	20,137	-0,1367	-0,684%
4	4,0000	23,000	23,040	-0,0400	-0,174%
5	5,0000	25,000	24,782	0,2178	0,871%
6	6,0000	26,000	26,367	-0,3670	-1,412%
7	7,0000	30,000	29,592	0,4084	1,361%
8	8,0000	36,000	36,287	-0,2873	-0,798%
9	9,0000	48,000	47,647	0,3534	0,736%
10	10,000	62,000	62,306	-0,3056	-0,493%
11	11,000	78,000	78,261	-0,2612	-0,335%
12	12,000	94,000	93,545	0,4547	0,484%
13	13,000	107,00	106,99	0,0105	0,010%
14	14,000	118,00	118,23	-0,2256	-0,191%
15	15,000	127,00	126,92	0,0801	0,063%

Среднее абсолютное отклонение = 0,2217
Средняя ошибка аппроксимации = 0,591%

График сплайна по 5 точкам (массив PRAN.dat):



2. **Сплайн-интерполяция со сглаживанием** Y.

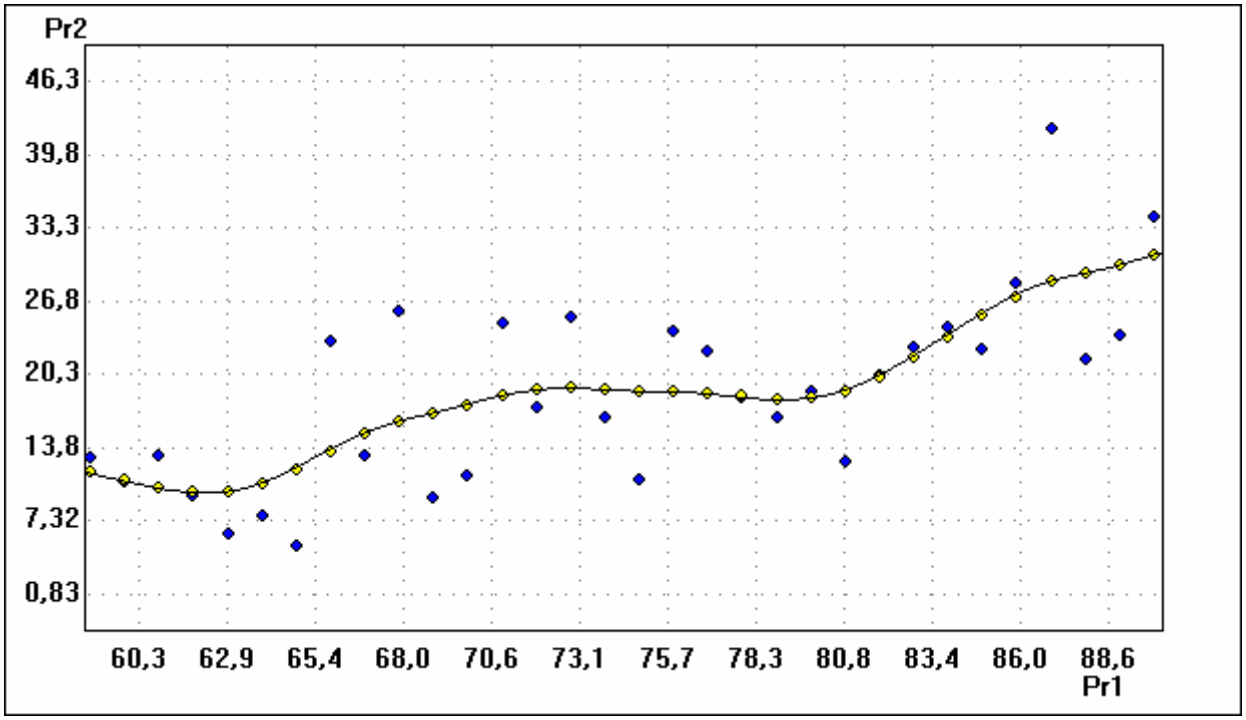
Для лучшей аппроксимации зашумленной экспериментальной зависимости следует использовать предварительное сглаживание значений зависимой переменной Y одним из 5-и способов. Предлагается сглаживание методами простой, двойной и взвешенной скользящей средней (для увеличения влияния центральных членов), в последнем случае коэффициенты вычисляются по формуле Миллера [45]; стр. 39. Узлы сплайна выбираются таким же способом, как в методе без сглаживания.

Весьма эффективным является метод сглаживания с помощью быстрого преобразования Фурье. В этом методе узлы сплайна проходят по всем экспериментальным точкам X, но степенью сглаживания можно манипулировать в широком диапазоне дискретных уровней.

3. **Сплайн-регрессия с минимизацией суммы квадратов** отклонений методом НК. Узлы сплайна располагаются на равных расстояниях друг от друга по оси X и, соответственно, не совпадают с точками X исходных данных. Подгонка сплайн-функции заключается в выборе оптимального числа узлов; эффективность этого метода определяется сглаживанием переменной Y минимизацией суммы квадратов отклонений (МНК по Гауссу).

4. **Сплайн-регрессия с минимизацией суммы квадратов отклонений и интеграла квадратов вторых производных** в узлах сплайна. Узлы проходят по всем экспериментальным точкам X, но степенью сглаживания можно манипулировать в широком диапазоне плавного изменения коэффициента сглаживания пе-

ременной Y – от 0,0000 (максимальное сглаживание в прямую линию) до 0,9999 (сплайн проходит практически по исходным экспериментальным точкам). Шаг изменения коэффициента задается перед вычислением сплайн-функции (0,1, 0,01 или 0,001) и далее автоматически уменьшается при приближении к границам 0,0 .. 1,0 (массив M2x32.dat, коэффициент=0,7):



Программа тестировалась по данным из [46], [57]. Техника сплайнов также описана Поллардом [45].

8.5. MLREG: Множественная линейная регрессия

Программа MLREG предназначена для обработки экспериментальных данных методом множественного линейного регрессионного анализа. Данные могут быть получены как в ходе активного эксперимента, проведенного, например, методом центрального композиционного планирования, так и из различных пассивных экспериментов.

В программе используются следующие методы вычисления коэффициентов регрессии:

1. Стандартный **Метод Наименьших Квадратов (МНК)**, связанный с формированием матрицы ковариаций (матрицы Грамма) и последующим вычислением обратной матрицы.

2. Итерационный метод – минимизация суммы абсолютных отклонений (**Минимизация Суммы Модулей, МСМ**). Оценки коэффициентов регрессии,

полученные этим методом, обычно близки к оценкам МНК. Преимущества метода МСМ:

– аномально большие значения отклика (выбросы, артефакты) значительно меньше влияют на решение, чем в методе НК (линейное влияние отклонений на алгоритм минимизации, а не квадратичное);

– достоверность уравнения регресии, определяемая критерием Фишера-Снедекора, обычно выше, чем в методе НК

– для данных со значительной корреляцией независимых переменных решение МНК вообще может быть не получено из-за вырожденности матрицы Грамма, тогда как метод МСМ позволяет получить решение, в большинстве случаев с оценками достоверности коэффициентов регрессии с помощью Т-критерия Стьюдента.

3. **Метод “Минимакс”**. Итерационный поиск решения, основе которого лежит известный алгоритм “линейного программирования”. Также, как и при методе МСМ, малочувствителен к выбросам; в некоторых случаях позволяет получить уравнение регрессии с максимальным значением F-критерия общей достоверности регрессии.

4. Множественная **регрессия на главных компонентах** (см. соответствующий раздел). Используется в тех случаях, когда матрица Грамма близка к состоянию вырожденности, но выбрать какие-либо переменные для исключения сложно вследствие сильной связи с переменной отклика.

Возможно пошаговое исключение и включение (в ручном или автоматическом режиме) входных переменных для получения уравнения регресии, состоящего только из достоверно влияющих независимых переменных. Имеется возможность добавлять к имеющимся независимым переменным "новые" переменные, полученные либо произведением имеющихся переменных, либо возведением в квадрат – для проверки возможных эффектов взаимодействия, или нелинейных эффектов. Для абсолютного сравнения входных переменных по силе влияния на зависимую переменную дополнительно вычисляются коэффициенты регрессии после центрирования/нормирования входных переменных (без стандартизации зависимой переменной).

Ограничения на размер массива: общее число столбцов должно быть не более 100, число строк (вариантов) – не более 10000, но максимальный размер массива – 100 тысяч элементов. Число вариантов в любом случае должно быть больше числа независимых переменных не менее чем на 1. Зависимая переменная

Y может иметь до 30 повторностей, причем допускается неравное число повторений.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 3-х независимых переменных, 10-и вариантов опыта, зависимой переменной в 2-х повторностях в текстовом файле:

```
5 10 3 2                                <- начало файла: 5 столбцов, 8 строк
12,3 22,5 34,2 0,34 -999                1 повторение
8,34 23,7 33,1 0,23 0,25                2 повторения
9,23 24,6 31,6 0,45 0,43                2 повторения     массив данных:
7,12 20,9 30,3 0,23 -999                1 повторение     строки = объекты,
8,27 19,4 32,4 0,78 0,77                2 повторения     столбцы = признаки
6,21 18,5 31,6 0,98 0,99                2 повторения
5,67 17,2 30,6 0,75 0,71                2 повторения
5,33 16,8 28,9 0,76 0,82                2 повторения
5,38 16,9 28,5 0,76 0,82                2 повторения
5,41 17,1 29,7 0,77 0,81                2 повторения

Опыт N123                               <- необязательный комментарий
  X1    X2    X3 |     Y
```

Можно опустить в заголовке файла указание числа независимых переменных и числа повторностей, тогда перед началом анализа необходимо указать количество независимых переменных. В качестве примера формирования массива для программ регрессионного анализа можно посмотреть файл SSP6X30.dat (6 переменных, 30 вариантов).

Данные могут быть введены с клавиатуры непосредственно в программе MLREG с помощью Табличного редактора. После указания количества переменных и вариантов (не менее чем на 2 больше числа переменных, иначе невозможно вычислить коэффициенты регрессии), следует ввести данные и записать затем в виде файла на магнитном диске.

Для общей проверки достоверности полученного уравнения регресии вычисляется F-критерий Фишера-Снедекора. В результате работы программы также вычисляются:

1. Коэффициенты регрессии: характеристика линейной связи между входной и зависимой переменными; их достоверность определяется T-критерием Стьюдента.

На основе анализа вероятностей можно осуществлять пошаговое исключение незначимых входных переменных, начиная с наименее достоверных.

2. Коэффициенты множественной корреляции и детерминации: наличие множественной линейной связи между зависимой переменной и совокупностью входных переменных; чем ближе они к 1.0, тем выше степень линейной связи (массив SSP6x30.dat):

Коэффициенты регрессии обычные	Стандартные ошибки	Коэффициенты после центр/норм.	Стьюдент T-критерий	Стьюдент вероятн. 1го рода	MAX парной корреляции Xi		Корреляция между Y и Xi парная	Корреляция между Y и Xi частн
B0 -5,5353159	1,76132	2,2666667	3,1427	0,0042*	-	-		
B2 0,0074355	0,00172	0,8508407	4,3176	0,0002*	-0,1838	5	.4219	.6462
B3 0,0149740	0,00551	0,5455148	2,7169	0,0116*	-0,2632	5	.1190	.4702
B5 0,0536255	0,01258	0,8567044	4,2626	0,0002*	-0,2632	3	.3941	.6414

```
Степеней свободы для критерия Стьюдента = 26

 2. Общие критерии достоверности регрессии.
 Y = B0 + B2*X2 + B3*X3 + B5*X5

Критерий Фишера:  F= 10,137  степени свободы: 3, 26
Вероятность нулевых значений коэффициентов:  P=0,00013
Относительная ошибка аппроксимации   Er = 30,880%
Коэффициент множественной корреляции: R = 0,734236
Коэффициент детерминации:            BY = 0,539103
Статистика Маллоуза (>4 при оптим.)   C = 2,11679
```

После исключения переменных X1 и X4 из общего множества 5-и независимых переменных достигнуто максимальное значение критерия Фишера-Снедекора, уравнение регрессии высокозначимо (P<0,001), все коэффициенты уравнения также значимы по критерию Стьюдента. Наибольшее влияние на зависимую переменную оказывают переменные X2 и X5 (по значениям коэффициентов регрессии после центрирования/нормирования, коэффициентов корреляции).

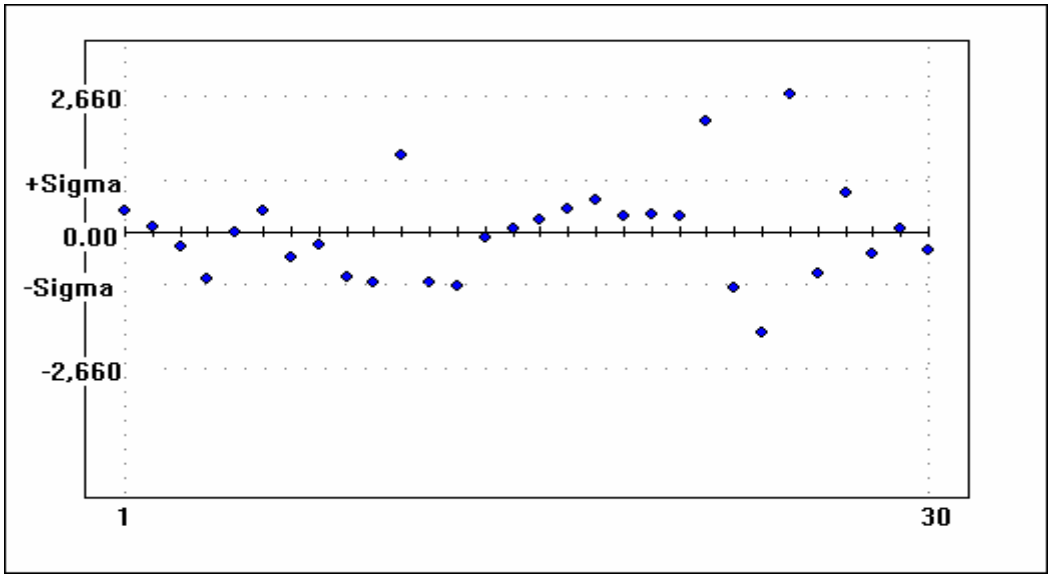
3. Если зависимая переменная имеет повторности, это позволяет оценить “адекватность” полученного уравнения – степень соответствия вычисленным по уравнению значениям Y – полученным в эксперименте. Оценка адекватности производится по критерию Фишера с соответствующим значением вероятности того, что уравнение адекватно. Чем ближе вероятность к 1, тем лучше уравнение регрессии отражает действующие в эксперименте зависимости.

Анализ данных в случае опытов с равным числом повторений может существенно зависеть от типа организации опыта – полной рандомизации или рандомизации в блоках повторностей (случайные блоки). Значимость коэффициентов регрессии определяется "остаточным средним квадратом", вычисляемым с учетом типа рандомизации. По умолчанию программа обрабатывает данные по типу полной рандомизации.

При выполнении пошагового исключения независимых переменных в первую очередь исключаются те переменные, для которых критерий Стьюдента

меньше 1.0 (или вероятность того, что B(i)=0 больше 0.5). Дополнительные соображения для исключения той или иной переменной может дать значение максимального коэффициента парной корреляции этой переменной с какой-либо другой. Большое (по модулю) значение корреляции (0.8-0.9) говорит о дублировании информации, и служит поэтому аргументом для исключения.

В пунктах Меню "Анализ остаточного варьирования, График остатков" можно визуально оценить качество подгонки регрессии для экспериментальных данных по каждому варианту, получить таблицу отклонений с проверкой некоторых предпосылок регрессионного анализа – гомоскедастичности дисперсии, отсутствие автокорреляций в остатках, нормальности распределения по критическим значениям асимметрии и эксцесса (3 станд. ошибки для асимметрии и 5 станд. ошибок для эксцесса):



Программа тестировалась по данным из [4, 18, 20, 26] (массивы SSP6x30.dat, DRAPER59.dat, SAS_MLRG.dat, AFIFI144.dat). Практически все вычисления выполняются с двойной точностью.

Для вычисления поверхности отклика необходимо сформировать двумерный массив значений независимых переменных с желаемым числом комбинаций их уровней. Это можно сделать непосредственно в среде программы выбором соответствующего пункта Меню. После вызова двумерного массива из файла, подготовленного заранее, программа проверяет соответствие числа переменных в этом массиве текущему числу независимых переменных, и при несовпадении выдает предупреждающее сообщение. В качестве такого массива можно использовать исходный массив независимых переменных, который автоматически за-

носится программой в дополнительный редактор данных при загрузке основного массива. Пример массива значений независимых переменных:

```
3 9
10,0  1,0  10,0
10,0  2,0  20,0
10,0  3,0  30,0
15,0  1,0  15,0
15,0  2,0  30,0
15,0  3,0  45,0
20,0  1,0  20,0
20,0  2,0  40,0
20,0  3,0  60,0
```

3 независимых переменных,
9 вариантов для вычисления
отклика по уравнению регрессии
на основе главных компонент

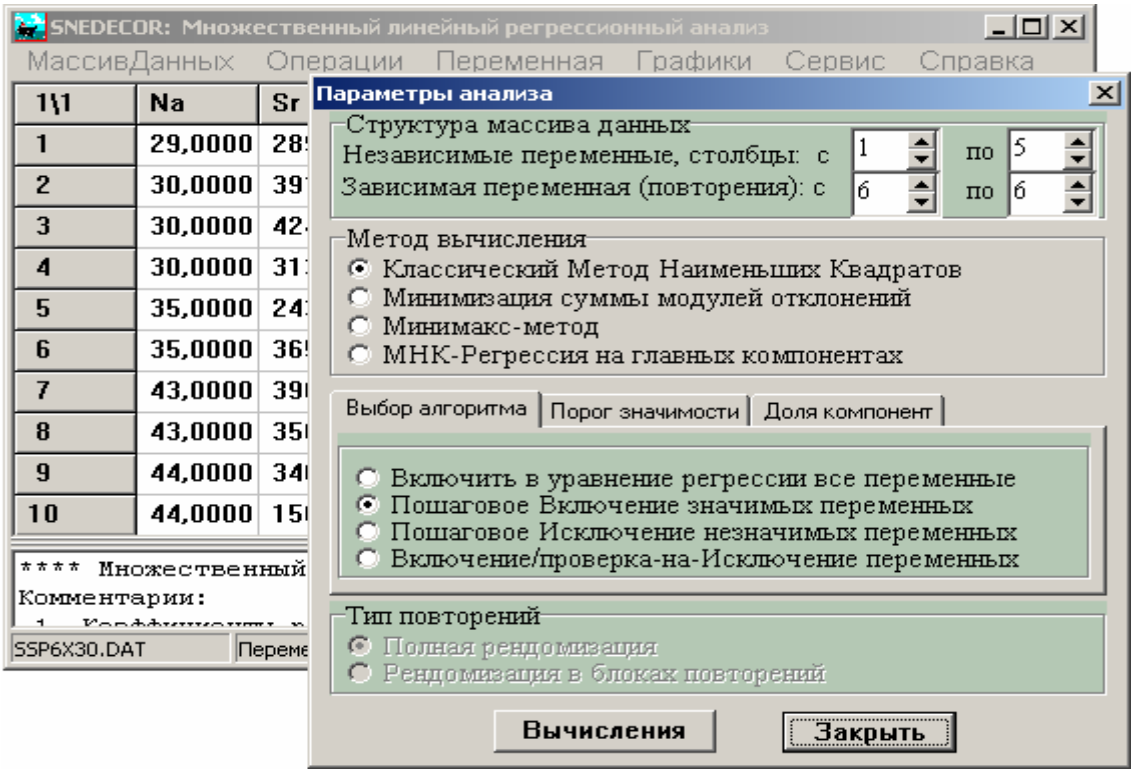
8.5.1. Методы шаговой регрессии

Методы автоматической шаговой регрессии, которые можно выбрать в форме “Параметры анализа”:

1. Пошаговое ВКЛЮЧЕНИЕ переменных.

Вначале вычисляются коэффициенты частной корреляции между зависимой переменной (Y) и всеми независимыми переменными (X1..Xm). Переменная с максимальным (по модулю) значением корреляции включается в качестве первой переменной для построения уравнения регрессии. Вычисляются коэффициенты регрессии, критерии достоверности коэффициентов.

Если вероятность ошибки в случае отклонения 0-гипотезы (Bi=0.0) больше заданного пользователем порогового значения, процесс прекращается, новые переменные не добавляются. Пользователь может задать порог значимости в диапазоне 0,01 – 0.2, обычно рекомендуют значения 0.1 или 0,15.



Если вероятность ошибки меньше порога, добавляется еще одна независимая переменная со следующим максимальным значением коэффициента частной корреляции. Снова вычисляются коэффициенты регрессии, вероятность ошибки при отклонении 0-гипотезы (Bi=0,0), сравнивается с порогом. Процесс добавления переменных прекращается при исчерпании списка входных переменных, либо при вероятности ошибки больше пороговой, при этом последняя добавленная переменная исключается из уравнения регрессии.

2. Пошаговое ИСКЛЮЧЕНИЕ переменных. Вначале в уравнение регрессии включаются все входные переменные. Вычисляются коэффициенты регрессии, Т-критерии их достоверности, вероятности ошибки при отклонении 0-гипотезы (Bi=0,0). Максимальное значение вероятности ошибки сравнивается с порогом (рекомендуется 0,15 – 0,25). Если вероятность больше пороговой, соответствующая переменная исключается из уравнения регрессии, вычисление коэффициентов регрессии повторяется, и т.д. Процесс прекращается, если осталась одна переменная, или максимальная вероятность ошибки меньше пороговой.

3. Пошаговое ВКЛЮЧЕНИЕ/ИСКЛЮЧЕНИЕ переменных. Как и в первом методе, вначале вычисляются коэффициенты частной корреляции между зависимой переменной (Y) и всеми независимыми переменными ($X_1..X_m$). Переменная с максимальным (по модулю) значением корреляции включается в качестве первой переменной для построения уравнения регрессии. Вычисляется коэффициент регрессии, критерий достоверности коэффициента. Если вероятность

ошибки в случае отклонения 0-гипотезы больше порога значимости исключения, дальнейшие операции прекращаются, и выводится решение с этой переменной. В ином случае для каждой оставшейся переменной вычисляются коэффициенты частной корреляции этих переменных с Y, и для коэффициента с максимальным значением корреляции вычисляется вероятность ошибки при отклонении 0-гипотезы [$R(X_i..Y)=0$].

Если вероятность меньше порога включения, эта переменная включается в уравнение регрессии, вычисляются коэффициенты регрессии, и проверяется их достоверность по Т-критерию Стьюдента. Если для какой-либо переменной вероятность ошибки больше порога исключения, она удаляется из уравнения, снова вычисляются коэффициенты регрессии.

Если вероятность меньше порога исключения, снова вычисляются коэффициенты частной корреляции каждой оставшейся переменной с Y, и для переменной с максимальным значением коэффициента определяется достоверность. Если вероятность меньше порога включения, переменная включается в регрессию, и т.д.

Пороги включения/исключения должны быть разными, иначе произойдет зацикливание программы. Например:

Вероятность включения = 0,10, Вероятность исключения = 0,25.

8.5.2. Регрессия на главных компонентах

Метод множественного линейного регрессионного анализа, входными переменными для которого являются "Главные компоненты" – вычисленные некоторым образом новые переменные, являющиеся линейными комбинациями исходных входных переменных, причем корреляции между этими новыми переменными отсутствуют (100% ортогональность).

Важно четко представлять себе структуру массива данных, передаваемую для обработки программе; рассмотрим пример:

Варианты	Независимые переменные X1 X2 X3 X4 X5 X6	Зависимая переменная, Y
1	X11 X21 X31 X41 X51 X61	Y1
2	X12 X22 X32 X42 X52 X62	Y2
3	X13 X23 X33 X43 X53 X63	Y3
4	X14 X24 X34 X44 X54 X64	Y4
5	X15 X25 X35 X45 X55 X65	Y5
6	X16 X26 X36 X46 X56 X66	Y6
7	X17 X27 X37 X47 X57 X67	Y7
8	X18 X28 X38 X48 X58 X68	Y8

В данном случае имеем шесть независимых переменных – X1, X2, ..., X6 (M=6), и восемь вариантов опыта (V=8), в которых для каждой из восьми комбинаций уровней независимых переменных измеряются значения отклика – зависимой переменной, которая здесь в одной повторности (P=1). На основании массива входных переменных [X1...X6] вычисляется матрица корреляций, для которой методом Якоби находятся собственные значения и собственные вектора.

Далее, на основе заданной заранее величины "Доля дисперсии главных компонент", по сумме возрастающих собственных значений определяется число главных компонент для вычисления уравнения регрессии. Обычно устанавливается доля дисперсии в диапазоне 50-90%, что позволяет получить зависимость с числом предикторов меньшим, чем число исходных входных переменных. Например, выбрано 70% дисперсии. После вычисления собственных значений получилось, например, что 3 главных компоненты определяют 76% дисперсии. Из центрированного/нормированного массива входных переменных вычисляются главные компоненты:

Входные переменные							Главн. компоненты			отклик
X1	X2	X3	X4	X5	X6		Z1	Z2	Z3	Y
X11	X21	X31	X41	X51	X61		Z11	Z21	Z31	Y1
X12	X22	X32	X42	X52	X62		Z12	Z22	Z32	Y2
X13	X23	X33	X43	X53	X63	-->	Z13	Z23	Z33	Y3
X14	X24	X34	X44	X54	X64		Z14	Z24	Z34	Y4
X15	X25	X35	X45	X55	X65		Z15	Z25	Z35	Y5
X16	X26	X36	X46	X56	X66		Z16	Z26	Z36	Y6
X17	X27	X37	X47	X57	X67		Z17	Z27	Z37	Y7
X18	X28	X38	X48	X58	X68		Z18	Z28	Z38	Y8

Затем на основании массива главных компонент [Z1...Z3] и переменной отклика [Y] методом стандартной множественной линейной регрессии вычисляются коэффициенты B0, B1, B2, B3 уравнения регрессии:

Y = B0 + B1*Z1 + B2*Z2 + B3*Z3.

Данные могут быть получены как в ходе активного эксперимента, проведенного, например, методом центрального композиционного планирования, так и из различных пассивных экспериментов.

Для облегчения анализа дополнительно вычисляются коэффициенты корреляции между главными компонентами и зависимой переменной.

Доля дисперсии главных компонент определяет в последующем число новых признаков (компонент), являющихся "входными" переменными для множе-

ственной регрессии. Обычно задают 70-80% дисперсии, но возможно и меньше – до 40-50%. Для уточнения доли нужно проанализировать собственные значения матрицы корреляций исходных входных переменных.

9. Корреляционный анализ

Универсальная мера связанности пары переменных – коэффициент парной корреляции. Простота вычисления и интерпретации коэффициента корреляции по Пирсону привела к широкому использованию этого показателя. Предпосылки применимости классического корреляционного анализа:

– нормальность распределения значений в признаках;

– линейность (или монотонность) характера связи между признаками;

– отсутствие автокорреляций в последовательности данных.

Достоверность отличия корреляции от нуля проверяется Т-критерием Стьюдента или пороговыми значениями из таблиц.

9.1. MATRIX: Матрица парных корреляций

Программа MATRIX предназначена для вычисления матрицы корреляций (ковариаций) для массивов экспериментальных данных, которые не могут быть обработаны с помощью программы MCOR:

а/ в случае, когда число объектов меньше числа признаков;

б/ в случае массивов с пропусками;

в/ если не выполняются предпосылки стандартного корреляционного анализа; в этом случае необходимо вычислять непараметрические показатели связи между признаками – корреляции рангов по Спирмену или Кендаллу;

г/ для вычисления матрицы автокорреляций.

д/ для углубленного анализа связи двух признаков вычисляются коэффициенты корреляции Пирсона (классический), корреляции рангов Спирмена и Кендалла; для трех признаков вычисляются коэффициенты частной и множественной корреляции; все коэффициенты сопровождаются критериями достоверности (Стьюдента, Фишера, Z-критерием);

е/ для вычисления бисериальной и точечно-бисериальной корреляции.

Аналогично анализу связанности признаков, программа выполняет анализ сходства/различия объектов. Можно выбрать любую из 9 мер сходства, коэффициент Гауэра допускает одновременное наличие дихотомических признаков (зна-

чения 0 или 1), признаков дискретного характера (ранги, частоты) и обычных признаков из непрерывных распределений.

На размеры массива данных из M признаков и N объектов имеются ограничения: M может быть не более 200, N – не более 5000, но при соблюдении условия M x N <= 100000. Если какой-либо признак имеет нулевую дисперсию, корреляции с этим признаком не могут быть рассчитаны; следует перейти к вычислению матрицы ковариаций.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. В качестве примера формирования массива для программ, обрабатывающих массивы "признаки-объекты" можно посмотреть файл SSP6x30.dat (6 признаков, 30 объектов). В случае массивов с пропусками коэффициенты корреляции и их достоверность вычисляются по имеющимся парам значений.

Пороговые значения коэффициентов корреляции (на уровнях значимости 1%, 5% и 10%) вычисляются на основе критерия Стьюдента или Z-критерия (для корреляций по Кендаллу) и приводятся под таблицей. Для непараметрических корреляций эти значения являются корректными при числе объектов не менее 10, при малых значениях N следует использовать таблицы статистик Спирмена и Кендалла.

Значения матрицы ковариаций вычисляются и выводятся в соответствии с формулой:

$$\mathbf{C_{ij}} = \left[\sum_{\mathbf{n}}^{\mathbf{k=1}} (\mathbf{x_{ki}} - \overline{\mathbf{x_i}}) \times (\mathbf{x_{kj}} - \overline{\mathbf{x_j}}) \right] / \mathbf{N}$$

C_{ij} – элемент матрицы; N – число объектов

это позволяет несколько уменьшить значения ковариаций для больших выборок. Если все же вместо некоторых значений программа выдала "*******", следует увеличить формат выдачи до 7..8 позиций.

Критерий взаимной независимости признаков вычисляется по [3], стр. 357, 0-гипотеза о независимости (некоррелированности, или же о диагональном виде матрицы ковариаций) признаков может быть отвергнута критерием Hi^2 при достаточно больших N. Критерием сферичности, помимо взаимной независимости признаков, проверяется и равенство дисперсий всех признаков [3], стр. 358; 0-гипотеза о сферичности проверяется аналогичным образом. Оба критерия могут

быть рассчитаны только для массивов данных без пропусков; в некоторых случаях из-за сильной коррелированности признаков критерии не могут быть вычислены.

Имеется возможность устанавливать точность представления коэффициентов корреляции при выводе на дисплей или принтер. Для этого нужно зайти в форму "Формат", открывающуюся перед анализом, и выбрать нужную точность.

С помощью программы MATRIX можно вычислить автокорреляции для признака (временного ряда), для этого нужно выбрать желаемый признак и задать "лаг" (число желаемых автокорреляций, от 1 до N-4, обычно 3-6). При этом исходный признак "размножается" следующим образом (например, признак из 9 элементов, требуется вычислить автокорреляции, лаг=3):

```
x1      x1
x2      x2 x1              обрабатываемый массив = 4 x 6
x3      x3 x2 x1
x4      x4 x3 x2 x1        x4 x3 x2 x1
x5  –>  x5 x4 x3 x2  –>    x5 x4 x3 x2   –>  матрица автокорреляций  4 x 4
x5      x6 x5 x4 x3        x6 x5 x4 x3
x7      x7 x6 x5 x4        x7 x6 x5 x4
x8      x8 x7 x6 x5        x8 x7 x6 x5
x9      x9 x8 x7 x6        x9 x8 x7 x6
           x9 x8 x7
              x9 x8
                 x9
```

9.1.1. Специальные виды корреляции

Для двух признаков вычисляются коэффициенты корреляции Пирсона (классический), Фехнера, корреляции рангов Спирмена и Кендалла; для трех признаков вычисляются коэффициенты парной, частной и множественной корреляции. Коэффициенты сопровождаются критериями достоверности (Стьюдента, Фишера, Z-критерием).

Коэффициент корреляции Фехнера вычисляется по формуле из [10, стр. 237]:

$\mathbf{R_f = (C–H)/(C+H)}$, где C – число совпадающих по знаку отклонений от средних переменных X и Y, H – число несовпадающих по знаку отклонений от средних.

Для оценки вариабельности коэффициентов корреляции можно использовать модификацию метода “Bootstrap”, предложенного Б.Эфроном в 1977 г.

Например, для двумерной выборки из N пар значений получено значение коэффициента парной корреляции. Если по какой-то причине исследователя не удовлетворяет значение стандартной ошибки, полученное на основе классической теории, можно оценить вариабельность следующим образом.

1. Задается некоторое большое число (500..1000..10000), определяющее, сколько раз генерировать случайные двумерные выборки (того же размера N) из значений имеющегося массива данных.

2. Методом Монте-Карло генерируются эти выборки, и каждый раз вновь вычисляется коэффициент корреляции. Эта оценка накапливается в массиве.

3. Для этого массива вычисляются среднее, экстремумы, квантили, на основании которых можно судить о вариабельности коэффициента корреляции.

Аналогичным образом выполняется оценка вариабельности других коэффициентов корреляции (Фехнера, Спирмена, Кендалла).

Коэффициент бисериальной корреляции, также разработанный К.Пирсоном, предназначен для тех случаев, когда только одна из переменных непрерывна и имеет приемлемо нормальное распределение, а другая искусственно дихотомизирована (предполагается, что она тоже непрерывна и нормально распределена, но представлена в бинарной форме, например, "поступил/не поступил"), связь между этими двумя переменными также можно выразить числом. В этом случае коэффициент корреляции обозначается через R_{bis}. Как и обычный коэффициент корреляции, он изменяется в диапазоне от +1,00 (прямая функциональная связь) через 0,00 (отсутствие связи) до -1,00 (обратная функциональная связь). Метод бисериальной корреляции оказался весьма полезным в процедурах анализа заданий, так как он измеряет связь между результатами выполнения каждого задания теста, выраженными в бинарной форме ("справился/не справился"), и общей оценкой по данному тесту. Пример массива данных (Пирсон, 1909, абитуриенты университета):

поступили	не поступили	возраст
583	563	16
666	980	17
525	868	18
383	814	20
214	439	25
40	81	33

Коэффициент точечно-бисериальной корреляции показывает связь между двумя переменными, одна из которых предположительно непрерывна и нормаль-

но распределена, а другая является дискретной в точном смысле слова. Точечно-бисериальный коэффициент корреляции обозначается через R_{pbis} Поскольку в R_{pbis} дихотомия отражает подлинную природу дискретной переменной, а не является искусственной, как в случае R_{bis}, его знак определяется произвольно. Поэтому для всех практических целей R_{pbis} рассматривается в диапазоне от 0,00 до +1,00. Пример массива данных для вычисления R_{pbis}:

да/нет	рост
0	153
0	164
1	175
0	186
1	190
0	192
1	196

Для тестирования программы можно использовать массивы KEND2x6.dat [48], стр. 411, KEND3x6.dat, [48], стр. 415, AN22x410.dat (тест R_{pbis}).

В случае трех признаков вычисляются:

1/ коэффициенты парной корреляции: степень линейной связи между парой признаков;

2/ коэффициенты множественной корреляции: наличие множественной линейной связи между одним признаком и двумя оставшимися;

3/ частные корреляции; отличаются от обычных парных корреляций тем, что характеризуют наличие линейной связи между двумя признаками в "очищенном" виде, так как из-за множественности связей между совокупностью признаков возможны не прямые зависимости, а косвенные – через третий признак.

Достоверность корреляций проверяется по критерию Фишера. 0-гипотеза формулируется следующим образом: отсутствует линейная связь между соответствующими признаками, коэффициент корреляции отличается от нуля вследствие действия случайных факторов. Для каждого F-критерия печатается "вероятность ошибки в случае отклонения 0-гипотезы". Если

P<=0,01 коэффициент значим на уровне 1%,

P<=0,05 коэффициент значим на уровне 5%,

P>0,10 корреляционная связь не доказана.

Достоверность бисериальных корреляций определяется приближенно, по аналогии с обычным коэффициентом корреляции. Достоверность точечно-бисериальной корреляции определяется Т-критерием [48], стр. 417.

9.1.2. Матрица сходства/различия объектов

Анализ сходства/различия объектов может быть полезен в самых различных ситуациях, например:

– выявление пар идентичных объектов, или “практически идентичных”;

– выявление аномальных объектов (”периферийных”), определяемых по значительному количеству больших расстояний в столбце или строке матрицы различия.

Матрица различия объектов является основой кластерного анализа.

Мера сходства объектов обычно определяется метрикой многомерного пространства, в котором расположены объекты, в других случаях (при несоблюдении стандартных аксиом метрического пространства) сходство/различие объектов выражается некоторыми функциями.

Стандартной метрикой является пространство Эвклидовых расстояний (Distances) между объектами; двумерное и трёхмерное Эвклидово пространство нам хорошо знакомо. Расстояния в пространствах большей размерности – это всего лишь обобщение всем хорошо известной формулы Пифагора:

двумерное расстояние: $\mathbf{D_2 = \sqrt{x^2 + y^2}}$

трёхмерное: $\mathbf{D_3 = \sqrt{x^2 + y^2 + z^2}}$

четырёхмерное: $\mathbf{D_4 = \sqrt{x^2 + y^2 + z^2 + p^2}}$ и т.д.

Исследователь вправе выбрать другие типы пространств, если из каких-либо свойств изучаемой системы следует предположение о некоторой специфике расположения объектов в пространстве признаков.

Например, имеет место существенная нелинейность взаимосвязей, явно дискретный или категорийный характер некоторых признаков, особая роль экстремальных значений некоторых признаков.

1. Эвклидово расстояние – очевидная геометрическая дистанция между объектами, обычно вычисляется по исходным данным без какой-либо нормализации/стандартизации. Желательно, чтобы характер измерения признаков был в ка-

кой-то степени однороден, в противном случае одиночные признаки с очень большими значениями будут сильно влиять на результат.

$$\mathbf{D_{ij} = \left(\sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2 \right)^{1/2}}$$

2. Расстояние Махаланобиса – обобщённое Эвклидово расстояние, учитывает не только значения признаков конкретной пары объектов, но и всего массива данных, из которого взяты эти два объекта – на основе ковариационной матрицы массива данных. С точки зрения теории, расстояние Махаланобиса должно эффективно отражать положение объекта в совокупности с принадлежностью к определённой системе объектов.

3. Манхэттенское расстояние – сумма абсолютных значений разностей:

$$\mathbf{D_m = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n|}$$

Влияние экстремальных значений некоторых признаков в этом случае уменьшается, поэтому эту метрику можно рекомендовать в случае подозрения на наличие артефактов, ошибок измерения. В некоторых публикациях эта метрика называется вариационным расстоянием Колмогорова.

4. Расстояние Чебышева – максимальная разница для какого-то признака:

$$\mathbf{D_{ch} = Max_k |x_{ik} - y_{ik}|}$$

Метрика Чебышева может быть применена, когда исключительное влияние должны иметь различия объектов по одной координате пространства.

5. Степенное расстояние Минковского (power metric):

$$\mathbf{D_p = \left(|x_1 - y_1|^p + |x_2 - y_2|^p + ... + |x_n - y_n|^p \right)^{1/r}, \quad p > 1, \quad r > 1}$$

Параметр **p** влияет на значимость вкладов разностей по некоторым координатам, параметр **r** – усиливает значимость больших расстояний между объектами. Эти параметры можно задать перед анализом.

6. Мера сходства – процент (доля) несогласия:

$$\mathbf{D_n = \left(число \ \ x_i \neq y_i \right) / n}$$

используется в случае признаков категориального типа, переведённых в числовую форму – например, в целые числа.

7. В ряде случаев, когда объекты числового (или нечислового) характера нельзя представить в виде точек многомерного пространства, но коэффициенты парной корреляции (Пирсона) между объектами могут быть рассчитаны практи-

чески всегда, тогда меру сходства/различия между объектами можно представить формулой:

$$\mathbf{D_{xr} = 1{,}00 - R_{xy}}$$

При линейной связи объектов D = 0, при отсутствии какой-либо связи D –>1,0. Вообще говоря, коэффициент корреляции объектов не является пространственной метрикой, но хорошо отражает сходство ”профилей” объектов на графике “номера признаков – значения признаков”.

8. Мера сходства: дивергенция объектов (различие, расхождение). Вычисляется по формуле (m = число признаков):

$$\mathbf{D_{ij} = \left(\frac{1}{m} \sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2 \Big/ \left(x_{ik} + x_{jk} \right) \right)^{1/2}}$$

Очевидно, при равных объектах D=0,0, для разных объектов D>0,0, таким образом, дивергенция объектов может считаться мерой сходства.

9. Мера сходства: коэффициент Гауэра:

$$\mathbf{D_{ij} = \sum_{i=1}^{m} W_{ijk} * S_{ijk} \Big/ \sum_{i=1}^{m} W_{ijk}}$$

$\mathbf{S_{ijk}}$ – сходство между состояниями признака, $\mathbf{W_{ijk}}$ – вес, приписанный этому признаку, а **m** – число признаков; допускается одновременное наличие дихотомических признаков (значения 0 или 1), признаков дискретного характера (ранги, частоты) и обычных признаков из непрерывных распределений. При идентичности объектов D=1.0, при максимальном различии D –> 0.0, поэтому значение коэффициента Гауэра вычитается из единицы, аналогично коэффициенту корреляции, чтобы получилась оценка различия объектов, а не сходства.

9.2. MCOR: Парные, множественные, частные корреляции

Программа MCOR предназначена для обработки экспериментальных данных, представляющих собой массив из M признаков и N объектов, различными видами корреляционного анализа.

Ограничения на размер массива: M может быть не более 100, N – любым, но при соблюдении условия M*N <= 100000. N в любом случае должно быть больше M не менее чем на 2. Если какой-либо признак имеет нулевую дисперсию, корреляции с этим признаком не могут быть рассчитаны; следует исключить этот признак.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. В качестве примера формирования массива для программ, обрабатывающих массивы "признаки-объекты" можно посмотреть файл SSP6x30.dat (6 признаков, 30 объектов).

Пороговые значения коэффициентов корреляции (на уровнях значимости 1%, 5% и 10%) вычисляются на основе критерия Стьюдента и приводятся под таблицами. Достоверность корреляций определяется по наличию звездочки.

Для теста значимости множественных корреляций вычисляется критерий Фишера-Снедекора. 0-гипотеза формулируется следующим образом: отсутствует линейная связь между соответствующим признаком и совокупностью всех прочих признаков, коэффициент корреляции отличается от нуля вследствие наличия случайных факторов. Для каждого F-критерия вычисляется вероятность ошибки в случае отклонения 0-гипотезы.

Результатом работы программы является:

1. Коэффициенты парной корреляции: степень линейной связи между парой признаков;

2. Коэффициенты множественной корреляции: наличие множественной линейной связи между одним признаком и совокупностью оставшихся;

3. Частные корреляции максимального порядка (M-2). Отличаются от обычных парных корреляций тем, что характеризуют наличие линейной связи между двумя признаками в "очищенном" виде, так как из-за множественности связей между совокупностью признаков возможны не прямые зависимости, а косвенные – через другие признаки. Математическим способом достигается максимальное очищение связи пары признаков от посторонних связей.

4. Частные корреляции 1-го порядка. Очищение связи между парой признаков минимальным образом – от влияния только одного какого-либо третьего признака. Анализ таких корреляций может прояснить картину взаимодействия совокупности признаков между собой. Эти корреляции программа вычисляет по желанию пользователя. Список признаков для вычисления этих корреляций формируется вводом символа "*" (или любого другого) в табличке справа на листе "Параметры", появляющемся перед выполнением анализа.

Если во время счета появляется сообщение программы "Матрица корреляций сингулярна!", это означает, что некоторые признаки связаны между собой **линейной функциональной зависимостью**, и рассчитать коэффициенты множе-

ственной или частной корреляции невозможно. Следует исключить сильно связанные признаки, или ограничиться только парными корреляциями, которые можно получить с помощью программы MATRIX.

Для случая очень больших матриц корреляций вывод на печать производится частями – с учетом значения параметра Long из файла CONFIG.sdc.

Имеется возможность устанавливать точность представления коэффициентов корреляции при выводе на дисплей или принтер. Для этого в форме "Параметры", открывающуюся перед анализом, следует выбрать нужную точность.

10. Анализ многомерных данных

В программах многомерного анализа следует различать и использовать два типа массивов. Первый тип – обычный массив “признаки-объекты”, второй – расширенный, по сравнению с первым, на 2 столбца, которые размещаются как 1-й и 2-й признак (столбец) массива данных. 1-й столбец – это номера объектов (обычно значения натурального ряда), 2-й столбец – номера групп, к которым принадлежат объекты.

С массивами второго типа работают две программы – MCOMP и DISCRYM, тогда как аналогичные программы MCOM и DISCRIM (и все прочие) работают с массивами первого типа. Версии программ для работы с массивами второго типа сделаны специально для тех пользователей, которым необходима дополнительная информация об объектах.

Массивы 2-го типа могут использоваться любыми программами, учитывая, что первые два столбца – не данные, а информация о структуре данных.

10.1. CANCOR: Канонические корреляции

Программа CANCOR предназначена для обработки экспериментальных данных, представляющих собой массив "признаки-объекты", с вычислением канонических корреляций.

Все множество признаков разбивается пользователем на две группы, исходя из некоторых представлений об общности признаков внутри этих групп. Например, массив из 8-и признаков: первая группа с 1-го по 3-й признак (X1..X3), вторая группа с 4-го по 8-й (X4..X8). Для формирования этих двух групп можно использовать матрицу обычных парных корреляций. Анализируя ее структуру, следует выявить группу признаков, имеющих значительные значения коэффици-

ентов корреляции между собой (как положительные, так и отрицательные); с помощью мышки нужно "перетащить" эти признаки в левую часть Табличного Редактора, оставшиеся признаки, естественно, окажутся в правой части Редактора, и составят вторую группу.

Программа вычисляет новые "признаки" – канонические переменные; количество их определяется числом исходных признаков в меньшей группе. Каждая каноническая переменная определяется зависимостью типа множественной линейной регрессии между всеми признаками "своей" группы и этой канонической переменной. Например, канонические переменные Y1..Y3 первой группы:

$Y1 = A_1*X1 + A_2*X2 + A_3*X3;$

$Y2 = B_1*X1 + B_2*X2 + B_3*X3;$

$Y3 = C_1*X1 + C_2*X2 + C_3*X3;$

канонические переменные Z1..Z3 второй группы:

$Z1 = D_1*X4 + D_2*X5 + D_3*X6 + D_4*X7 + D_5*X8;$

$Z2 = E_1*X4 + E_2*X5 + E_3*X6 + E_4*X7 + E_5*X8;$

$Z3 = F_1*X4 + F_2*X5 + F_3*X6 + F_4*X7 + F_5*X8;$

Анализируя вклады исходных признаков в канонические переменные, можно сделать продуктивные выводы о сущности неких факторов, действующих в исследуемой системе, математическим представлением которых и являются канонические переменные.

Коэффициенты "регрессий" ($A_1..A_3$, $B_1..B_3$, ... $F_1..F_5$) выводятся в результате работы программы, и на их основе вычисляются канонические переменные по центрированным/нормированным значениям признаков X1..X8 из массива данных. Канонические корреляции – это парные корреляции между каноническими переменными разных групп (в данном случае, корреляции Y1xZ1, Y2xZ2, Y3xZ3).

Ограничения на размер массива: M может быть не более 100, N – не более 4000, но при соблюдении условия M x N <= 100000, то есть максимальный размер массива – 100 тысяч элементов.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. В качестве примера формирования массива для программ, обрабатывающих массивы "признаки-объекты" можно посмотреть файлы SSP6x30.dat (6 признаков, 30 объектов), ABC8x50.dat.

Результатом работы программы является:

1. Матрица парных корреляций: служит для анализа структуры связности признаков, существенности разбиения на две группы.

2. Канонические корреляции: степень линейной связи между двумя группами признаков (квадратный корень из собственных значений некоторой матрицы), представленных парой новых признаков – канонических переменных.

3. λ-критерий Уилкса и критерий Hi^2 – для определения достоверности собственных значений (и канонических корреляций). Для критерия Hi^2 вычисляется вероятность ошибки в случае отклонения гипотезы об отсутствии связи между каноническими переменными. Если вероятность менее 0,05 – достоверность корреляции доказана.

4. Значения коэффициентов для каждого признака, отражающие вклад признаков в канонические переменные. Поскольку важны для анализа только первые канонические переменные, их число ограничивается при выводе – в случае превышения 6-ти. Значения коэффициентов нормированы: сумма квадратов коэффициентов меньшей группы равна 1.0.

5. Графическое представление связанности канонических переменных в виде поля корреляции между каноническими переменными.

10.1.1. Выбор групп признаков

Программа предоставляет возможность выбрать группы признаков для анализа тремя способами:

1 способ. Первая группа признаков формируется в той же последовательности, как они находятся в массиве данных, начиная с первого признака. Вторая группа признаков формируется из всех оставшихся признаков, вслед за последним признаком первой группы.

2 способ. Первая группа признаков выбирается кликами мышки в ячейках таблицы, появляющейся при выборе 2-го метода формирования групп. “Плюс” означает включение признака в первую группу, пустая ячейка – включение этого признака во вторую группу.

3 способ. Полный перебор всех возможных комбинаций признаков для формирования первой группы – от 2-х до максимально 10-и признаков в группе, вторая группа формируется из всех прочих признаков, не попавших в первую группу. В этом случае в результатах формируются только таблицы значений канонических корреляций с критериями достоверности. В конце текста результатов приводится комбинация признаков в группах с максимальным значением канонической

корреляции. При желании можно получить эту комбинацию в полном виде, выбрав 2-й способ формирования групп.

10.2. MCOM: Анализ данных методом главных компонент

Программа MCOM предназначена для обработки экспериментальных данных, представляющих собой массив “признаки–объекты”, методом **главных компонент**. Суть метода – формирование пространства новых стандартизированных, некоррелирующих между собой признаков, в этом пространстве облако объектов находится в центре координатных осей, и первые координатные оси проходят по наиболее протяженным размерностям этого облака. Это позволяет визуально анализировать структуру данных – однородность, присутствие групп/кластеров, наличие аномальных объектов и т.п.

В программе реализовано 4 способа вычисления главных компонент.

1. Стандартный способ: центрирование/нормирование исходного массива данных.

2. Только центрирование признаков.

3. Только нормирование признаков.

4. “Центрирование в группах”. В случае, когда массив данных состоит из нескольких групп объектов, можно сделать обработку данных по В.М.Ефимову (Институт цитологии и генетики СО РАН):

Ограничения на размер массива: M может быть не более 200, N не более 10000, но при соблюдении условия M x N <= 200000. Входной массив может состоять из одного признака (например, временной ряд), в этом случае программа "размножает" его в M признаков, число которых (лаг) задается пользователем. Для анализа временных рядов методом ГК имеется специализированная программа PROGNOZ.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х признаков и 11-и объектов в текстовом файле (две группы объектов 4 + 7=11):

```
6  11  4  7
12,3 22,5 34,2 0,34 1,45 3,11
8,34 23,7 33,1 0,23 1,66 3,65
9,23 24,6 31,6 0,45 1,89 2,79
7,12 20,9 30,3 0,23 1,73 3,09
8,27 19,4 32,4 0,78 1,77 3,35
6,21 18,5 31,6 0,98 1,85 3,69
5,67 17,2 30,6 0,75 1,57 3,51
8,55 16,3 33,9 0,77 1,33 3,40
7,23 17,6 32,1 0,82 1,21 3,22
6,47 15,5 31,7 0,79 1,42 3,74
5,18 16,0 31,2 0,78 1,63 3,71

Данные 1997 г.
```

<- начало файла

массив данных:
строки = объекты,
столбцы = признаки

<- необязательный комментарий

В качестве примера формирования массива для программ, обрабатывающих массивы "признаки-объекты" можно посмотреть файл SSP6x30.dat.

В случае данных, представленных (временным) рядом, после указания числа желаемых "признаков" массив для обработки формируется следующим образом (например, ряд из 10 элементов, лаг=4):

```
x0        x0
x1        x1 x0                    обрабатываемый массив
x2        x2 x1 x0
x3        x3 x2 x1 x0              x3 x2 x1 x0
x4   =>   x4 x3 x2 x1    =>        x4 x3 x2 x1
x5        x5 x4 x3 x2              x5 x4 x3 x2
x6        x6 x5 x4 x3              x6 x5 x4 x3
x7        x7 x6 x5 x4              x7 x6 x5 x4
x8        x8 x7 x6 x5              x8 x7 x6 x5
x9        x9 x8 x7 x6              x9 x8 x7 x6
             x9 x8 x7
                x9 x8
                   x9
```

Результатом работы программы является:

1. Элементарные статистики признаков: средние, ср.кв. отклонения и т.п.;

2. Матрица коэффициентов парной корреляции: степень линейной связи между парами признаков;

3. Собственные значения матрицы корреляций;

4. Собственные векторы матрицы корреляций;

5. Координаты объектов в пространстве ГК, вычисляются по формуле:

$\mathbf{G_i = X_1*V_{i1}+X_2*V_{i2}+...+X_m*V_{im}}$, где

G_i – i-я главная компонента,

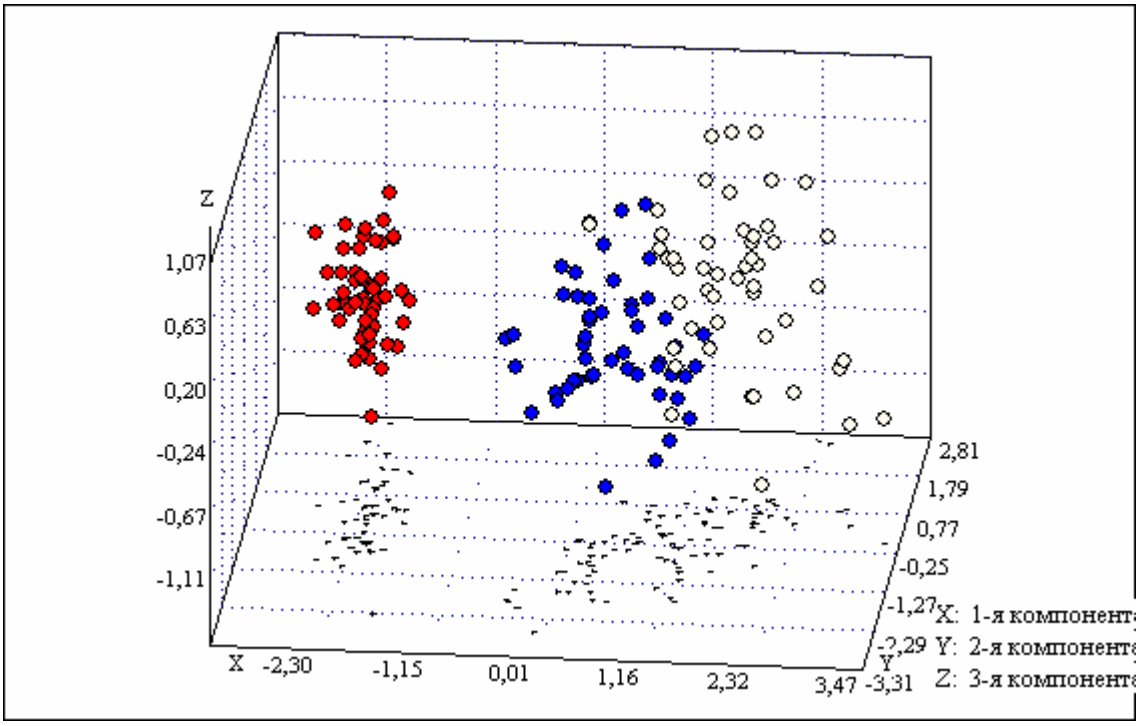
V_i – i-й собственный вектор,

$X_1, X_2, ... X_m$ – центрированные/нормированные признаки,

m – число признаков в массиве данных.

6. Матрица коэффициентов парной корреляции между ГК и исходными признаками; значения корреляций, значимые на уровне 5%, помечаются звездочкой *.

7. Графическое представление проекций объектов на плоскость какой-либо пары ГК (например 1-2, 1-3, 2-3, и т.д.), а также в пространстве трех гл. компонент. Смена ГК по осям описана в справке к форме "Графики" (массив FISH_IRI.dat):



Основной задачей исследователя является содержательная интерпретация вкладов признаков в значимые (первые 2-4) главные компоненты. По относительной величине, по знаку вкладов признаков – значений собственных векторов – необходимо выявить и логически непротиворечиво объяснить существование некоторых факторов – главных компонент – реально действующих в изучаемой системе, наблюдаемым следствием которых стали измеренные значения всех признаков (файл ABC8x50.dat):

Собственные векторы матрицы корреляций

	Соб.Векторы							
Признаки	V1	V2	V3	V4	V5	V6	V7	V8
Уран	0,2819	-0,5277	-0,3492	0,3273	-0,1914	-0,0173	-0,4134	-0,4529
Плутоний	0,4239	-0,1395	0,0696	-0,3499	0,8044	0,0014	-0,1120	-0,1179
Радий	0,3601	0,2504	-0,3757	-0,5700	-0,3646	-0,4312	-0,1205	0,0907
Стронций	0,3850	-0,3296	-0,3133	0,0538	-0,0707	0,2558	0,7003	0,2850
Цезий	0,4265	0,2355	0,1453	0,1125	-0,1506	0,5624	-0,4599	0,4224
Кобальт	0,3653	0,2215	0,4837	-0,0863	-0,2837	0,1535	0,2944	-0,6208
Калий	0,3282	0,4690	-0,1062	0,6474	0,2197	-0,4255	0,1073	0,0329
Натрий	0,1996	-0,4541	0,6063	0,0586	-0,1641	-0,4769	-0,0440	0,3560

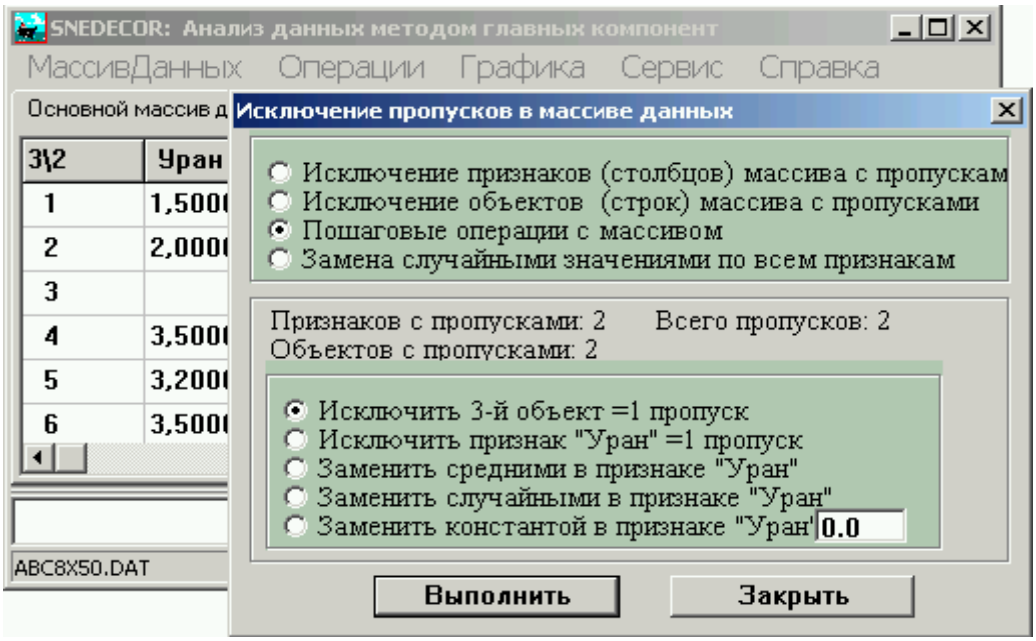
G(i) = X1*Vi(1)+X2*Vi(2)+...+Xm*Vi(m)

Точность представления чисел в результатах может быть задана перед анализом данных от 2 до 8 значащих цифр (по умолчанию 4 цифры). Увеличить точность до 5-6 цифр следует перед возможной записью массива ГК в виде файла данных или копированием массива в буфер Windows.

Программа может использовать для обработки массив данных с группировкой объектов (например, предназначенный для дискриминантного или многомерного 1-факторного дисперсионного анализа), в этом случае возможен 4-й способ вычисления главных компонент – по В.М.Ефимову.

Полученные в результате счета собственные векторы можно записать на диск в виде файла данных, передать через буфер Windows любой другой программе. Аналогичные операции возможны и для массива главных компонент (координаты всех объектов в пространстве главных компонент).

Если массив данных имеет значительное количество пропусков, можно исключить некоторые объекты или признаки в ручном пошаговом режиме, чтобы количество отсутствующих значений было не более 2-3%; далее программа заменит оставшиеся пропуски на средние в признаках:



Любой объект или группа некоторых дополнительных объектов может быть передана программе через буфер Windows для преобразования в ГК, число признаков в массиве дополнительных объектов должно быть в точности равно числу признаков исходного массива “признаки-объекты”, использованного для вычисления собственных векторов, в ином случае при копировании объектов из буфера массив будет либо усечен до требуемого числа признаков, либо заполнен значениями “NoValue” (-999,0). Результирующий массив ГК для дополнительных объ-

ектов может быть записан в виде файла данных, либо занесен в буфер Windows. Все операции с дополнительными объектами производятся с помощью Меню операций, вызываемого кликом ПРАВОЙ клавиши мышки в поле таблицы дополнительных объектов.

10.2.1. Методы вычисления главных компонент

Программа предлагает 4 способа вычисления главных компонент:

1-й способ: главные компоненты вычисляются на основе центрированного средними и нормированного среднеквадратическими отклонениями массива данных, в этом случае все главные компоненты имеют нулевые средние и примерно одинаковую дисперсию, что позволяет исследовать расположение объектов в координатах относительно однородного пространства

2-й способ: Только центрирование признаков. Этот способ может быть использован в тех случаях, когда все признаки измерены в одинаковой шкале. Главные компоненты вычисляются на основе центрированного средними массива данных и собственных векторов матрицы ковариаций.

3-й способ: Только нормирование признаков. Этот способ может быть использован в тех случаях, когда все признаки измерены в одинаковой шкале и имеют примерно одинаковые средние, но значительно различаются по размаху (дисперсии). Главные компоненты вычисляются на основе нормированного среднеквадратическими отклонениями массива данных и собственных векторов матрицы ковариаций.

4-й спобоб: В случае, когда массив данных состоит из нескольких групп объектов, можно сделать обработку данных по В.М.Ефимову (Институт цитологии и генетики СО РАН):

1-й шаг: вычисление общих и внутригрупповых средних;

2-й шаг: центрирование в группах внутригрупповыми средними;

3-й шаг: вычисление общих среднеквадратичных отклонений признаков такого центрированного в группах массива;

4-й шаг: масштабирование и поворот пространства – центрирование исходных признаков общими средними, нормировка среднеквадратичными отклонениями, вычисленными на шаге 3;

5-й шаг: на базе массива, полученного на шаге 4, вычисление матрицы корреляций, собственных векторов, главных компонент.

Этот метод используется для изучения внутри- и межпопуляционной изменчивости, фенотипическом анализе [61].

10.2.2. Bootstrap для главных компонент

Для оценки вариабельности собственных значений и собственных векторов матрицы корреляций используется модификация метода “Bootstrap”, предложенного Б.Эфроном в 1977 г.

Эта процедура выполняется следующим образом.

1. Задается некоторое большое число (500..1000..10000), определяющее, сколько раз генерировать случайные многомерные выборки (того же размера N) из значений имеющегося массива данных.

2. Методом Монте-Карло генерируются эти выборки, и каждый раз вновь вычисляется матрица корреляций, её собственные значения и собственные векторы. Эти оценки накапливаются в массивах.

3. Для этих массивов вычисляются средние, доверительные интервалы, экстремумы, на основании которых можно судить о вариабельности собственных значений и собственных векторов.

Число собственных векторов для анализа Bootstrap-методом рекомендуем ограничивать до 3-5, это обычная практика метода главных компонент, прочие собственные векторы как правило малоинформативны.

Перед выводом графики, записью массивов главных компонент и собственных векторов всегда вычисления повторяются, используется при этом массив исходных данных, а не какой-либо результат Bootstrap-анализа.

Bootstrap-процедура не может быть использована для 4-го метода анализа – с центрированием групп объектов по В.М.Ефимову.

10.3. MCOMP: Главные компоненты (номера объектов, коды групп)

Программа MCOMP предназначена для обработки экспериментальных данных, представляющих собой массив из M признаков и N объектов, методом главных компонент (ГК). Ограничения на размер массива: M может быть не более 100, N – не более 8000, но при соблюдении условия M x N <= 100000. В отличие от аналогичного модуля MCOM, массивы данных, передаваемые для обработки программе MCOMP, должны содержать номера объектов и коды групп в первых двух столбцах.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х признаков и 12-и объектов в текстовом файле:

численности групп

```
6  12 3 4 5
977 1 12,3 22,5 34,2 0,34
978 1 8,34 23,7 33,1 0,23
979 1 9,23 24,6 31,6 0,45
980 2 7,12 20,9 30,3 0,23
981 2 8,27 19,4 32,4 0,78
982 2 6,21 18,5 31,6 0,98
983 2 5,67 17,2 30,6 0,75
984 3 8,55 16,3 33,9 0,77
985 3 7,23 17,6 32,1 0,82
986 3 6,47 15,5 31,7 0,79
987 3 5,18 16,0 31,2 0,78
988 3 6,11 15,8 30,4 0,81

Данные за 1977-1988 г
```

<= начало файла
3 группы объектов: 3+4+5 = 12

массив данных:
12 строк = объекты,
3..6 столбцы = признаки;

1-й столбец – номера объектов;
2-й столбец – коды групп;

<= необязательный комментарий

В качестве примера формирования массива можно посмотреть файлы SSP8x30.dat (6 признаков, 30 объектов), FIS6x150.dat (4 признака, 150 объектов, 3 группы).

В пакете SNEDECOR есть специализированная программа для ввода массивов "признаки-объекты" – IODATA, но перед записью файла следует скорректировать его первую строку указанием размеров групп. Если отсутствует необходимость в разбиении объектов на группы, во втором столбце следует занести одно и то же число, например 1.0, а первую строку файла можно не корректировать. Если по каким-либо причинам объекты в группах (и, соответственно, коды объектов во втором столбце) неупорядочены, следует сделать сортировку объектов по возрастанию значений **второго** столбца.

Результатом работы программы является:

1. Матрица коэффициентов парной корреляции: степень линейной связи между парой признаков;

2. Собственные значения матрицы корреляций;

3. Все собственные вектора матрицы корреляций;

4. Значения ГК для всех объектов (только при выводе на печать или в текстовый файл); вычисляются по формуле:

$$\mathbf{G_i = X_1 * V_{i1} + X_2 * V_{i2} + ... + X_m * V_{im};}$$

G_i – i-я главная компонента,

V_i – i-й собственный вектор,

$X_1, X_2,... X_m$ – центрированные/нормированные признаки,

m – число признаков в массиве данных.

5. Графическое представление проекций объектов на плоскость какой-либо пары ГК (например, 1-2, 1-3, ... 1-9, 2-3, ... 2-9, и т.д.). ГК в парах указываются после ввода символа «C» (сменить номера ГК); появляется окно в нижней части экрана, далее необходимо указать номера ГК вводом с клавиатуры. Клавиша «N» модифицирует изображение, заменяя позиции объектов на их номера; клавиша «G» меняет на номера групп, если заголовок массива данных содержит разбиение множества объектов на группы.

Помимо графического изображения проекций объектов при значительном их числе (более 100) возможно получение распечатки проекций на принтере в текстовом режиме (или вывести в текстовый файл); для этого в соответствующем пункте Меню следует выбрать размеры распечатки, просмотреть результат на дисплее, затем отправить на принтер.

Если программе передан для обработки массив данных с группировкой (например, предназначенный для дискриминантного или многомерного дисперсионного анализов), в этом случае возможен специальный способ предварительной обработки – центрирование данных в группах по В.М.Ефимову. Полученные в результате счета собственные вектора можно записать на диск в виде файла данных, и передать затем любой другой программе. Аналогичная операция возможна и для массива "ГК – Объекты", но в этом случае добавляются два столбца – с номерами объектов и кодами групп.

10.4. FACTOR: Факторный анализ

Программа FACTOR предназначена для обработки экспериментальных данных, представляющих собой массив из M признаков и N объектов, различными методами факторного анализа. Ограничения на размер массива: M может быть не более 100, N – не более 4000, но при соблюдении условия MxN<= 100000.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. В качестве примера формирования массива для программ, обрабатывающих массивы "признаки-объекты" можно посмотреть файл SSP6x30.dat (6 признаков, 30 объектов).

Основная задача факторного анализа – вычисление матрицы факторных нагрузок [A_1, A_2, ..., A_p], с помощью которой исходные признаки могут быть описаны следующей моделью:

$$\mathbf{X_1 = A_{11}*F_1 + A_{12}*F_2 + ... + A_{1p}*F_p + V_1*H_1}$$

$$\mathbf{X_2 = A_{21}*F_1 + A_{22}*F_2 + ... + A_{2p}*F_p + V_2*H_2}$$

. . . . . .

$$\mathbf{X_m = A_{m1}*F_1 + A_{m2}*F_2 + ... + A_{mp}*F_p + V_m*H_m}$$

$F_1..F_p$ – некоторые новые гипотетические признаки (факторы), причем их число значительно меньше числа исходных признаков. Факторы могут быть некоррелированными (ортогональными) или же допускается некоторая степень линейной связи между ними. $H_1..H_m$ – факторы, присущие только "своим" входным переменным – "характерности" (или специфичности). $V_1..V_m$ – коэффициенты, отражающие долю этой характерности для каждой входной переменной – также определяются в ходе факторного анализа (равны квадратному корню из Специфичности, см. ниже).

Результатом работы программы является:

1. Матрица коэффициентов парной корреляции;
2. Собственные значения матрицы корреляций;
3. Собственные вектора матрицы корреляций;
4. Значения факторных нагрузок для всех признаков;
5. Значения Общностей/Специфичностей для каждого признака;
6. Значения факторных нагрузок после вращения факторов, если оно было выполнено;
7. Матрица остаточных корреляций, не определяемых факторной структурой.

Можно выбрать один из пяти методов факторного анализа:

1. Невзвешенный метод наименьших квадратов.
2. Обобщенный метод наименьших квадратов.
3. Метод максимального правдоподобия.
4. Факторизация по модели главных компонент.
5. Модель главных факторов с редуцированием матрицы корреляций.

Стандартным методом является модель Главных Факторов, модель Главных Компонент также можно рекомендовать для любых данных – особенно на начальном этапе работы с массивом. В последующем можно перейти к более сложным методам факторизации – 1, 2 или 3. Эти методы основаны на совре-

менных итерационных алгоритмах с большим объемом вычислений (К.Йореско), но они не всегда приводят к конечному результату за приемлемое время; для некоторых массивов данных возможно зацикливание или выход по ошибке вычислительного характера. При зацикливании (это можно увидеть по прекращению счета итераций) необходимо нажать любую клавишу, например пробел, это приведет к выходу в Меню.

В литературе описан еще один метод факторизации – минимальных остатков [47], однако решения, полученные этим методом, практически эквивалентны результатам факторизации методом Невзвешенных Наименьших Квадратов.

4-й и 5-й методы факторизации являются классическими, с относительно быстрыми алгоритмами, всегда приводящими к конечному результату. При выборе метода Главных Факторов программа формирует редуцированную матрицу корреляций – с заменой диагональных единиц оценками общностей – квадратами коэффициентов множественной корреляции соответствующих признаков с множеством остальных.

Необходимо также указать метод вращения факторов, используемый для получения "простой структуры". Как правило, цель всех вращений – еще более увеличить (по абсолютной величине) большие факторные нагрузки, и уменьшить малые нагрузки:

1. Ортогональное вращение Кайзера "ВариМакс" – стандартное, стремится увеличить нагрузки на относительно малое число переменных для каждого фактора. Максимизирует разброс квадратов нагрузок для каждого фактора.

Далее 4 метода косоугольного вращения по Харрису-Кайзеру, причем степень косоугольности можно варьировать с помощью параметра ортогональности Cf. При Cf=0 получается решение с максимальной коррелированностью факторов, при Cf=1 приближается к ортогональному решению.

2. Вращение "ВариМакс" – аналогично стандартному вращению, увеличивает нагрузки на относительно малое число переменных для каждого фактора.

3. Вращение "КвартиМакс" характеризуется стремлением дать каждой переменой большую нагрузку только на один или несколько факторов. Максимизирует общую дисперсию квадратов факторных нагрузок.

4. Вращение "ЭквиМакс" – компромисс этих двух методов.

5. Обобщенный метод вращения "ОртоМакс" требует указания еще одного параметра – W. Его значение может быть любым, но лучшие значения обычно находятся в интервале [1, 5*Nf], где Nf – число факторов.

Число факторов указывается пользователем. Рекомендуем начинать с двух факторов, и далее увеличивать при необходимости до требуемого числа факторов.

Здесь же, если выбрать косоугольное вращения факторов, программа запрашивает значение параметра вращения, Cf. Его значение должно быть от нуля до 1.00. Для данных с простой факторной структурой Cf=0.0 часто является наилучшим значением. Для более сложных данных Cf=0.5 возможно явится оптимальным. Реже, по-видимому, Cf>0.5 может быть приемлемым значением. По умолчанию программа устанавливает Cf=0.00. Но в этом случае результаты любого метода косоугольного вращения практически идентичны.

Программа выполняет проверку достоверности числа факторов по критерию Hi^2. Это в общем ориентировочный критерий, рекомендуемый при значительном числе объектов [(N-M)>50, по Лоули-Максвеллу]. 0-гипотеза – число факторов равно выбранному. Если вероятность ошибки в случае принятия 0-гипотезы больше 0.2, рекомендуется увеличить число факторов на 1, получить решение и вновь проанализировать по Hi^2. Вероятность ошибки менее 0,05 или отрицательное значение Hi^2 говорят о достаточности (или избыточности) числа факторов.

Перед выводом результатов счета на принтер или в текстовый следует скорректировать заголовок в тексте результатов (обычно 2-я строка "Комментарии").

Возможно вычисление массива "факторы-объекты" (шкалирование факторов) с последующей записью на диск в виде файла, пригодного для последующего использования другими программами пакета SNEDECOR. Векторы значений факторов формируются со средними, равными 10, и дисперсиями, близкими к единице. Программа предлагает 4 метода шкалирования факторов:

– метод множественной регрессии;

– метод наименьших квадратов;

– метод Бартлетта;

– метод Андерсона-Рубина.

Относительно выбора метода шкалирования рекомендуем [24], стр. 52-62. Методы Бартлетта и Андерсона-Рубина – более поздние разработки.

Проекции объектов на плоскость главных факторов используются для выявления объектов с экстремальными значениями факторов, а также для выявления кластеров. Номера объектов можно получить после нажатия комбинации

<Alt/N> (Numbers); номера факторов по вертикали меняются клавишами "1".."9", номера факторов по горизонтали – клавишами <Alt/1>..<Alt/9> (если факторов 9 или больше, но реальное число факторов обычно меньше).

10.5. CLASTER: Кластерный анализ

Программа CLUSTER предназначена для обработки экспериментальных данных в виде массива "признаки-объекты" различными алгоритмами кластерного анализа. Под этим названием подразумевается большая группа методов автоматической классификации множества объектов на некоторое, заранее неизвестное, число подмножеств – групп объектов (кластеров, популяций, таксонов). Определение кластера (Эверрит, 1980):

Кластеры – это «непрерывные» области (некоторого) пространства с относительно высокой плотностью точек, отделённые от других таких же областей областями с относительно низкой плотностью точек.

В программе реализованы различные методы кластеризации:

1/ группа методов, основанные на матрице сходства/различия (минимум Эвклидовых расстояний, максимум корреляций между объектами, максимум корреляций между Эвклидовыми расстояниями, максимум корреляций между корреляциями объектов, минимум коэффициентов дивергенции объектов;

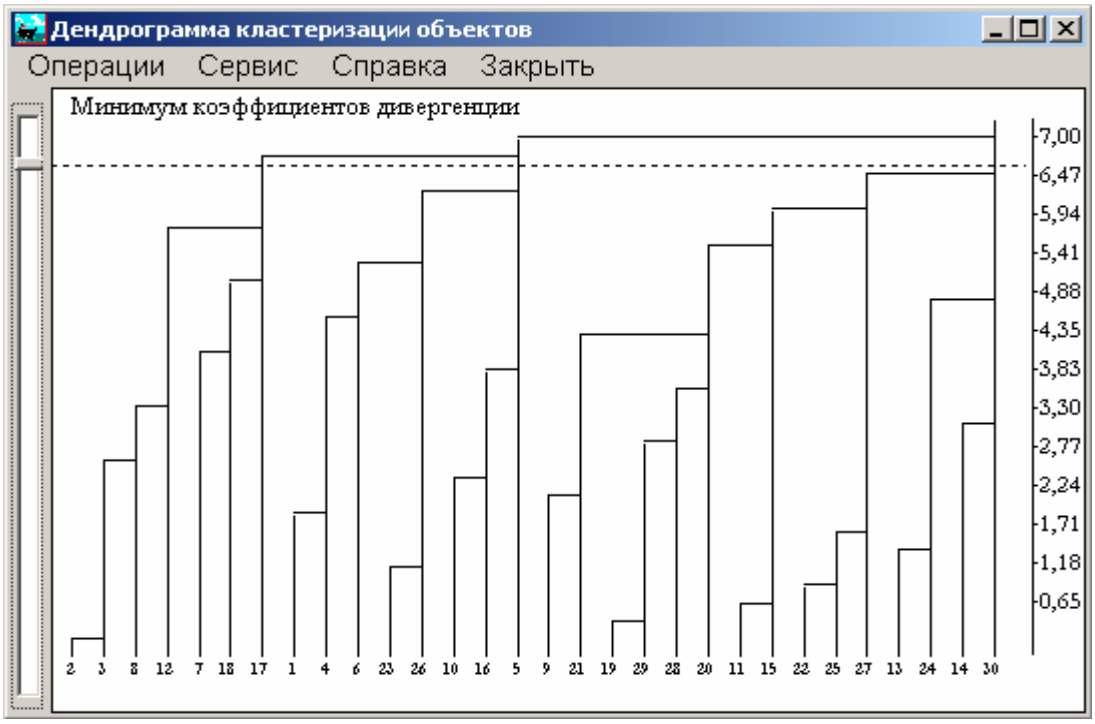
2/ группа методов, основанных на различных способах агломерации (объединения) объектов, общий принцип которых – определение минимального расстояния между парой объектов/кластеров; матрица расстояний при этом не вычисляется;

3/ алгоритм кластеризации объектов методом К-средних, с предварительным выбором числа кластеров.

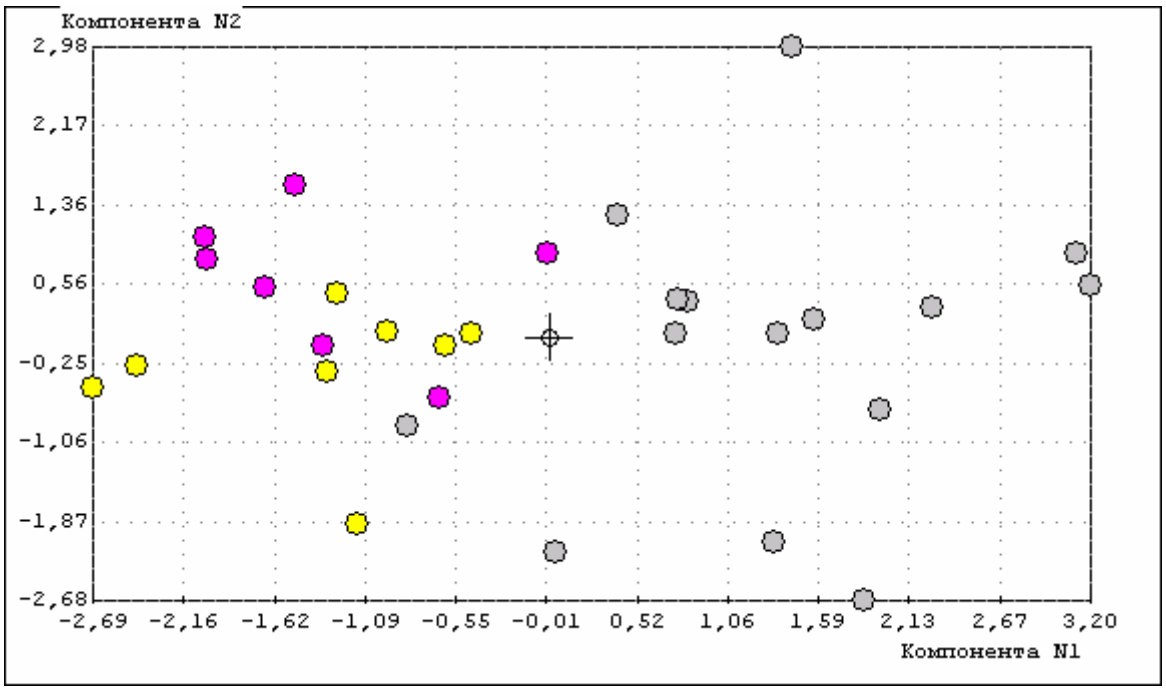
На графиках (дендрограмма кластеризации, проекции объектов на плоскость или в пространство главных компонент) исследователь должен визуально определить разумное число кластеров, руководствуясь опытом и принципом простой структуры: кластеров должно быть немного, группы должны быть как можно более компактны, не должно быть пересекающихся границ.

Максимальный массив данных – 100000 чисел. Число объектов не должно превышать 700 (признаки х объекты: до 140х700). Для иерархических методов и кластеризации методом «К-средних» число объектов может быть значительно больше (до 10х10000).

Группы кластеров на дендрограмме определяются пользователем с помощью мышки – перемещением по вертикали бегунка шкалы в левой части графика. Исходя из структуры дендрограммы, выбирается пересечение 2..5 или больше вертикальных линий, определяющих вершины кластеров:



На данной дендрограмме – три (Kd=6,5) или четыре кластера (Kd=6,3). Качество кластеризации можно посмотреть визуально, анализируя расположение объектов на плоскости или в пространстве главных компонент (стандартизированных некоррелирующих переменных – линейных комбинаций исходных признаков). Для этого следует одновременно с дендрограммой активизировать вторую графическую форму – с проекциями объектов в осях главных компонент; перемещение бегунка на дендрограмме будет автоматически менять распределение объектов по кластерам, определяемым цветом или номером кластера:



Клавишей “G” можно определить группы объектов, клавишей “N” – номера объектов. “Пробел” возвращает представление объектов в виде окружностей, клавиши “+” и “-” меняют диаметр этих окружностей. Проекции объектов в осях главных компонент замечательны тем, что располагают облако точек наиболее информативным образом – выстраивая его по максимально растянутым в N-мерном пространстве осям облака.

В методе кластеризации «К-средние» кластеры формируются автоматически, выбор делается только между кластеризациями, сделанными для различного числа кластеров.

Массив объектов, рассортированных по кластерам, можно записать в виде, приемлемом для использования программой дискриминантного анализа DISCRYM. Массив будет дополнен двумя столбцами (слева) – номерами исходных объектов и номерами кластеров (групп). Программа MCOMP (главные компоненты) также может использовать такой массив без каких-либо преобразований. В прочие программы также можно загружать такой массив, имея в виду наличие двух столбцов структурного характера.

Одним из важнейших пунктов работы с программой является выбор метода кластеризации данных – это необходимо делать, исходя из характера данных и опыта исследователя. Чаще всего используют методы «Минимум Эвклидовых расстояний между объектами», «К-средние».

Результатами работы программы являются:

1. Таблица последовательности кластеризации со значениями сходства или расстояния.

2. Дендрограмма кластеризации объектов в виде графика, с возможностью ручного выделения кластеров.

3. Анализ качества кластеризации – сумма внутригрупповых дисперсий кластеров (должна быть наименьшей при сравнении нескольких методов), критерии Фишера-Снедекора 1-мерного дисперсионного анализа групп каждого признака. Чем больше достоверных F-критериев (P<0,05), тем лучше разделение объектов на кластеры.

4. Матрица расстояний (сходств) между объектами. Эта матрица может быть сохранена либо в виде текстового файла, пригодного для просмотра с помощью какого-либо редактора текстов типа Блокнота/Windows, либо в виде массива данных в стандарте пакета Snedecor.

В случае необходимости можно транспонировать массив данных, (соответствующий подпункт в Меню «МассивДанных») и использовать любые методы для кластеризации ***признаков***. В этом случае число объектов (которые станут "признаками" после транспонирования) не должно превышать 1000.

10.5.1. Кластеризация методом К-средних

Этот метод относится к типу дивизимных методов кластеризации (разделяющих), так как исследователь должен сам задать число предполагаемых кластеров. Если нет никаких предположений о кластерной структуре данных, следует сделать несколько кластеризаций с возрастающим числом кластеров – 2, 3, 4, и так далее. Качество кластеризации анализируется визуально с помощью графиков «Дендрограмма кластеризации», «Проекции объектов на плоскость главных компонент» и особенно – «... в пространство главных компонент».

Метод К-средних работает следующим образом:

1/ случайно выбираются К объектов в качестве центров (центроидов) будущих кластеров;

2/ для всех остальных объектов последовательно вычисляются расстояния до К центроидов; метрика пространства может быть различной – выбираемой пользователем;

3/ минимальное расстояние определяет объект и кластер, к которому присоединяется этот объект;

4/ координаты центроида этого кластера пересчитываются в соответствии с новым числом объектов;

Цикл 2/ ... 4/ повторяется для всех остальных объектов, после этого вычисляется критерий качества кластеризации – сумма внутригрупповых дисперсий.

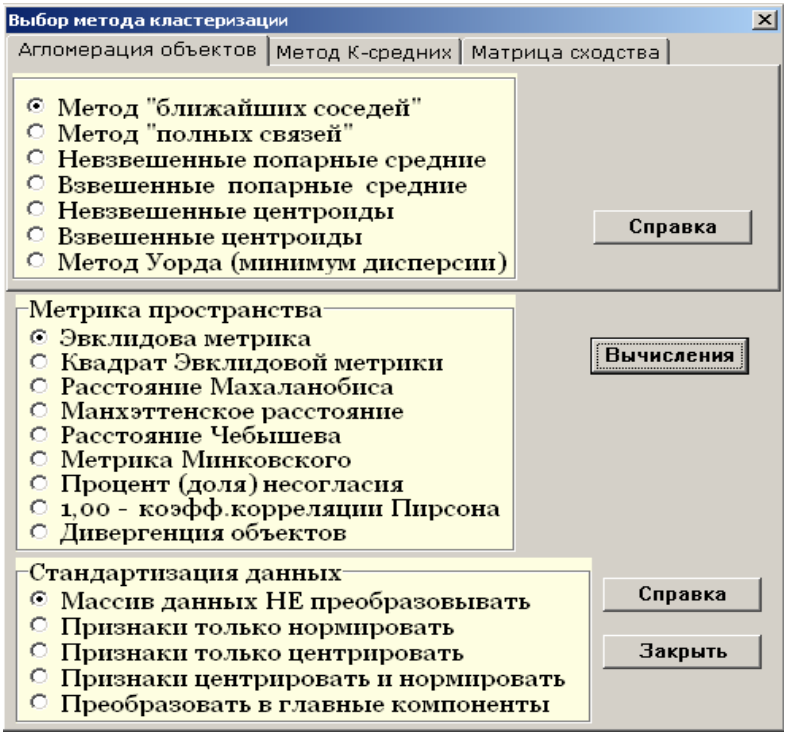
Так как кластеризацию можно повторять многократно (до 1000 раз и более, всё зависит от размера массива данных и мощности процессора), программа запоминает конфигурацию с минимальной суммой внутригрупповых дисперсий, и воспроизводит её в качестве результата кластеризации.

Можно дополнительно несколько улучшить результат кластеризации, пометив пункт «Выполнить процедуру ISODATA». В этом случае программа выполнит несколько дополнительных итераций, задавая в качестве стартовых центроидов векторы групповых средних, полученные на предыдущем этапе.

В случае подозрения на наличие артефактов – экстремально больших (по модулю) значений, рекомендуется использовать в качестве центроидов не векторы средних, а векторы медиан. Для этого следует пометить мышкой пункт «Центроиды: медианы вместо средних» на форме запроса метода кластеризации. Медианы значительно менее чувствительны к выбросам, нежели средние.

10.5.2. Выбор метрики пространства

Изменение метрики многомерного пространства (меры сходства объектов), в котором расположены объекты, может весьма существенно повлиять на результаты кластеризации.



Стандартной метрикой является пространство Эвклидовых расстояний (Distance) между объектами; двумерное и трёхмерное Эвклидово пространство нам хорошо знакомо. Расстояния в пространствах большей размерности – это всего лишь обобщение всем хорошо известной формулы Пифагора:

двумерное расстояние: $\mathbf{D_2 = \sqrt{x^2 + y^2}}$

трёхмерное: $\mathbf{D_3 = \sqrt{x^2 + y^2 + z^2}}$

четырёхмерное: $\mathbf{D_4 = \sqrt{x^2 + y^2 + z^2 + p^2}}$ и т.д.

Исследователь вправе выбрать другие типы пространств, если из каких-либо свойств изучаемой системы следует предположение о некоторой специфике расположения объектов в пространстве признаков.

Например, имеет место существенная нелинейность взаимосвязей, явно дискретный или категорийный характер некоторых признаков, особая роль экстремальных значений некоторых объектов.

1. Эвклидово расстояние – очевидная геометрическая дистанция между объектами, обычно вычисляется по исходным данным без какой-либо нормализации/стандартизации. Желательно, чтобы характер измерения признаков был в какой-то степени однороден, одиночные признаки с очень большими значениями будут сильно влиять на результаты кластеризации.

$$D_{ij} = \left(\sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2 \right)^{1/2}$$

2. Квадрат Эвклидова расстояния можно рекомендовать в тех случаях, когда из каких-либо соображений необходимо усилить влияние более удалённых друг от друга объектов.

3. Расстояние Махаланобиса – обобщённое Эвклидово расстояние, учитывает не только значения признаков конкретной пары объектов, но и всего массива данных, из которого взяты эти два объекта – на основе ковариационной матрицы массива данных. С точки зрения теории, расстояние Махаланобиса должно эффективно отражать положение объекта в совокупности с принадлежностью к определённой системе объектов, однако качество кластеризации по этой метрике обычно неудовлетворительное.

4. Манхэттенское расстояние – сумма абсолютных значений разностей:

$$D_m = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n|$$

Влияние экстремальных значений некоторых признаков в этом случае уменьшается, поэтому эту метрику можно рекомендовать в случае подозрения на наличие артефактов, ошибок измерения. В некоторых публикациях эта метрика называется вариационным расстоянием Колмогорова.

5. Расстояние Чебышева – максимальная разница для какого-то признака:

$$D_{ch} = Max_k |x_{ik} - y_{ik}|$$

Метрика Чебышева может быть применена, когда исключительное влияние должны иметь различия объектов по одной координате пространства.

6. Степенное расстояние Минковского (power metric):

$$D_p = \left(|x_1 - y_1|^p + |x_2 - y_2|^p + ... + |x_n - y_n|^p \right)^{1/r}, \quad p > 1, \quad r > 1$$

Параметр p влияет на значимость вкладов разностей по некоторым координатам, параметр r – усиливает значимость больших расстояний между объектами. Эти параметры можно задать перед кластеризацией.

7. Мера сходства – процент (доля) несогласия:

$$D_n = \left(\div \text{èñëî} \quad x_i \neq y_i \right) / n$$

используется в случае признаков категориального типа, переведённых в числовую форму – например, в целые числа.

8. В ряде случаев, когда объекты числового (или нечислового) характера нельзя представить в виде точек многомерного пространства, но коэффициенты парной корреляции (Пирсона) между объектами могут быть рассчитаны практически всегда, тогда меру сходства/различия между объектами можно представить формулой:

$$\mathbf{D_{xr} = 1{,}00 - R_{xy}}$$

При линейной связи объектов D = 0, при отсутствии какой-либо связи D –>1,0. Вообще говоря, коэффициент корреляции объектов не является метрикой, но допустимость использования коэффициента корреляции в кластерном анализе подтверждается многолетней практикой.

9. Мера сходства: дивергенция объектов (различие, расхождение). Вычисляется по формуле (m = число признаков):

$$\mathbf{D_{ij} = \left(\frac{1}{m} \sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2 / \left(x_{ik} + x_{jk}\right) \right)^{1/2}}$$

Очевидно, при равных объектах D=0,0, для разных объектов D>0,0, таким образом, дивергенция объектов может считаться мерой сходства.

Помимо выбора метрики пространства, исследователь может использовать различные типы стандартизации массива данных, во многих случаях существенно улучшающие результаты кластеризации – давая более выраженные группы объектов. Центрирование признака – вычитание среднего из всех значений, нормирование – деление на среднеквадратическое отклонение, центрирование и нормирование – соответственно вычитание среднего и деление на ср.-кв. отклонение по всем значениям. В последнем случае все признаки становятся стандартизованными, имеющими равные средние (=0,00) и равные дисперсии (=1,00).

Весьма полезным может быть преобразование исходных признаков, которые могут быть сильно коррелированными, в новые стандартизированные некоррелирующие признаки – главные компоненты. Такое преобразование сохраняет

относительное пространственное расположение объектов, и во многих случаях «улучшает» результаты кластеризации.

10.5.3. Методы иерархической кластеризации

Существует большое число методов иерархической кластеризации объектов, в программе реализованы лишь некоторые методы агломеративной (объединяющей) кластеризации. Во всех методах подразумевается, что на старте процесса все объекты массива данных – одиночные кластеры, и на каждом шаге делается выбор, какая пара кластеров должна быть объединена.

1. Метод одиночной связи («ближайшего соседа»). Вначале объединяются два объекта с минимальным расстоянием. Затем находится объект, имеющий минимальное расстояние либо с одним из объектов этой группы, либо с каким-либо объектом из оставшихся. Отсюда: либо объект присоединяется к имеющемуся кластеру, либо формируется новый кластер. Данная процедура повторяется то тех пор, пока не останется один кластер, содержащий все объекты массива данных. Возможный недостаток этого простейшего метода – образование больших удлинённых «цепочечных» кластеров.

2. Метод всех связей (complete linkage, «самого дальнего соседа»). Включение объекта в другой кластер определяется максимальным расстоянием до какого-либо объекта этого кластера, но минимальным среди всех имеющихся кластеров. Дополнительно делается проверка на *непревышение* этого расстояния некоторого порога D (фактически диаметра гиперсферы, окружающей все возможные объекты кластера). Если порог превышен, объекту (кластеру) отыскивается пара, формирующая новый кластер. При объединении кластеров проверка выполняется между всеми парами объектов обоих кластеров. Таким образом, в этом методе исследователь должен задать программе число – диаметр гиперсфер – исходя из характера данных и собственного опыта.

3. Невзвешенное попарное среднее. Расстояние между парой объект (кластер) – кластер определяется как среднее расстояние между всеми парами объек-

тов в них. Минимальное среднее расстояние по всему множеству пар кластер – кластер определяет объединение объектов в новый кластер.

4. Взвешенное попарное среднее. Аналогично предыдущему методу, определяется минимум среднего расстояния по множеству пар объекты/кластеры – кластеры, но затем среднее расстояние корректируется коэффициентом для воплощения очевидного принципа – больший кластер должен эффективно поглощать кластер меньшего размера. В программе коэффициент «поглощения» вычисляется следующим образом:

Co = 0,5 + Min(n1,n2)/(n1+n2), n1 и n2 – размеры кластеров.

При n1=n2 Co=1,00; при различных размерах кластеров Co стремится к 0,5, например, n1=1, n2=9, тогда Co=0,6, таким образом расстояние существенно уменьшается, облегчая выбор в пользу большого кластера.

5. Метод «Невзвешенные центроиды» (метод Кинга). Минимальное расстояние между векторами средних (центроидами) всех пар кластеров определяет очередной шаг кластеризации (объединения объектов/кластеров), вектор средних после этого рассчитывается заново. По-видимому, это наиболе естественный способ кластеризации данных, на который мало влияют экстремумы различного рода (артефакты, ошибочные значения).

6. Метод «Взвешенные центроиды». Аналогичная корректировка минимального расстояния между центроидами кластеров коэффициентом «поглощения», вычисляемым также, как в методе 4. Метод рекомендуется использовать, когда ожидаются серьёзные различия в размерах кластеров.

7. Метод Уорда (Ward). В этом методе решающим является не минимум расстояний, а минимум внутригрупповой дисперсии получающегося в результате предполагаемого объединения кластера. Внутригрупповая дисперсия – сумма квадратов расстояний между объектами кластера и центроидом.

Выбор того или иного метода кластеризации целиком лежит на совести исследователя. По-видимому, следует придерживаться одного и того способа кластеризации для однотипных данных, и экспериментировать с выбором метода для принципиально нового типа экспериментальных данных.

10.5.4. Методы кластеризации на базе матрицы сходства/различия

В программе имеются методы кластеризации, основанные на матрице сходства/различия. Это, в общем, несколько устаревший подход, так как требует большого ресурса оперативной памяти в случае массива со значительным числом объектов. Например, для 1000 объектов треугольная матрица сходства/различия требует около 2 мегабайт памяти, для 10000 объектов – 190 мегабайт.

В программе не вычисляется матрица сходства – только её элементы, когда они нужны для очередного шага кластеризации. Алгоритм кластеризации в этих методах следующий:

а/ в матрице сходства находится элемент с минимальным (для корреляций максимальным) значением, по нему определяется пара кластеров, которые объединяются на этом шаге;

б/ кластер с **меньшим** числом объектов включается в кластер с **большим** числом объектов;

в/ значения вектора средних кластера (центроида) пересчитывается как взвешенная сумма центроидов исходных кластеров, затем матрица сходства корректируется;

г/ столбец:строка меньшего кластера исключаются из матрицы сходства;

д/ значение сходства и номера кластеров текущего шага запоминаются для последующей сортировки и построения дендрограммы кластеризации.

Цикл а/ .. д/ повторяется до объединения всех объектов массива в один кластер. Методы кластеризации:

1. Минимум Эвклидовых расстояний. Стандартный метод кластеризации, который можно рекомендовать в большинстве случаев, когда все признаки массива данных – количественного характера.

2. Максимум корреляций между объектами. Спорный метод, так как корреляция объектов, равная 1.0, может означать как идентичность объектов, так и различие – в случае функционального характера линейной связи между значениями признаков. По-видимому, этот метод может быть использован в тех случаях,

когда значения признаков – кодированные числами номинальные характеристики объектов, или имеют порядковый тип (оценки, ранги и т.п.). В программе коэффициент корреляции преобразовывается в показатель различия объектов по формуле:

$$\mathbf{D_{ij} = 1{,}00 - R_{ij}}$$

это значение приводится в результатах кластеризации, поэтому диапазон изменения D = 0,00 .. 2,00. При выборе этого метода кластеризации программа автоматически переключает тип стандартизации данных на «центрирование+нормирование», в большинстве случаев это позволяет получить наилучшие результаты по сравнению с другими типами стандартизации данных.

3. Максимум корреляций между Эвклидовыми расстояниями. Критерий кластеризации – поиск наибольшего коэффициента корреляции Пирсона между **строками** матрицы Эвклидовых расстояний. Аналогично предыдущему методу, найденный коэффициент корреляции пересчитывается в показатель различия D. Поскольку матрица расстояний не вычисляется, этот метод требует значительного времени для завершения кластеризации при значительном числе объектов.

4. Максимум корреляций между корреляциями объектов. Критерий кластеризации – поиск наибольшего коэффициента корреляции Пирсона между **строками** матрицы парных корреляций. Интуитивно спорный метод, иногда может дать лучшие результаты. Аналогично, коэффициент корреляции пересчитывается в показатель различия D. Так же, как и в методе (2.), программа автоматически переключает тип стандартизации данных на «центрирование+нормирование». Так же, как в методе (3.), матрица корреляций не вычисляется, поэтому потребуется значительное время для завершения кластеризации при значительном числе объектов.

5. Минимум коэффициентов дивергенции. Дивергенция – расхождение, различие систем, множеств, объектов, в контексте задачи кластерного анализа – различие объектов в виде двух векторов – строк массива данных. В программе коэффициент дивергенции двух объектов вычисляется по формуле (m = число признаков):

$$D_{ij} = \left(\frac{1}{m} \sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2 / \left(x_{ik} + x_{jk} \right) \right)^{1/2}$$

Кластеризация объектов этим методом может быть выполнена на любых данных, и зачастую даёт наилучшие результаты в сравнении с другими методами. Так же, как в методе (2.), программа автоматически предлагает тип стандартизации данных «центрирование+нормирование», хотя могут быть использованы и другие типы стандартизации.

10.6. DISCRIM: Дискриминантный анализ

Программа DISCRIM предназначена для обработки экспериментальных данных различными методами дискриминантного анализа (ДА). Вначале программа использует эталонный "обучающий" массив для анализа групп объектов с хорошо известной принадлежностью, вычисления дискриминантных функций, затем, после ввода массива объектов для дискриминации, выполняется отнесение каждого объекта в одну из заданных в обучающем массиве групп с некоторой вероятностью, зависящей от "качества" групп. Основные предположения [24, с. 84], необходимые для корректной работы классических методов ДА:

– число объектов больше числа признаков (на 2 или больше),

– значения признаков измерены в интервальной шкале или шкале отношений,

– высокая корреляция между признаками (порядка >0.95) отсутствует,

– ковариационные матрицы групп обучающих объектов примерно равны,

– распределение значений в группах – многомерное нормальное.

При некоторых нарушениях этих предпосылок ДА в общем работает – с несколько меньшей эффективностью. Однородность групповых ковариационных матриц может быть проверена с помощью программы HOTELL, 1-мерная нормальность признаков в группах может быть проверена с помощью программы NORMAL.

Для линейного ДА (4-й метод анализа) не требуется равенство ковариационных матриц, но в этом случае для дискриминации объектов используется не аппарат дискриминантных функций, а максимум функции плотности вероятности нормального распределения значений объекта, вычисляемой для каждой группы.

Ограничения на размер обучающего массива: число признаков (M) может быть не более 200, число объектов (N) – не более 5000, но при соблюдении условия M x N <= 100000, (например, не более 100 x 1000, 200 x 500, 50 x 2000 и т.д.), число групп – не более 20. Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 6 признаков и 11 объектов в текстовом файле:

```
        Численности групп объектов, только для “обучающего” массива
          | | |                          3 группы:  4+4+3=11
6  11 4 4 3                                <- начало файла
12,3 22,5 34,2 0,34 1,45 3,11              \
8,34 23,7 33,1 0,23 1,66 3,65               1-я группа
9,23 24,6 31,6 0,45 1,89 2,79               /              массив данных:
7,12 20,9 30,3 0,23 1,73 3,09              /               строки  = объекты,
8,27 19,4 32,4 0,78 1,77 3,35              \               столбцы  = признаки
6,21 18,5 31,6 0,98 1,85 3,69               2-я группа
5,67 17,2 30,6 0,75 1,57 3,51               /
8,55 16,3 33,9 0,77 1,33 3,40              /
7,23 17,6 32,1 0,82 1,21 3,22              \
6,47 15,5 31,7 0,79 1,42 3,74               3-я группа
5,18 16,0 31,2 0,78 1,63 3,71              /
Лизин                                      <- названия признаков (необязательно)
Метионин
Триптофан
Треонин
Лейцин
Изолейцин

Данные 1997 г                              <- необязательный комментарий
```

Массив дискриминируемых объектов должен иметь то же число признаков, что и обучающий массив; последний, в частности, может быть использован для автодискриминации.

В качестве примера формирования массива можно посмотреть файл FISH_IRI.dat (4 признака, 150 объектов – классический пример с тремя видами ирисов). После загрузки обучающего массива программа вычисляет векторы групповых средних для каждого признака и заносит их в правый Табличный редактор в качестве объектов для дискриминации; на графиках "Проекции объектов" они отображаются в виде окружностей вблизи центров групп.

Программа предлагает четыре метода вычисления дискриминантных функций:

1/ метод Фишера-Андерсона-Рао [26]; [49], стр. 452, при котором число дискриминантных функций равно числу групп в обучающем массиве, а их последовательность привязана к последовательности групп; принадлежность объекта к

некоторой группе определяется максимумом из значений всех дискриминантных функций для этого объекта;

2/ метод Кульбака-Рао [13], [32] (канонический ДА), при котором общее число дискриминантных функций равно числу признаков; принадлежность объекта к группе определяется минимальным расстоянием в пространстве Махаланобиса от объекта до центров групп, вычисляемым по совокупности всех дискриминантных функций. Алгоритм метода предложен В.М.Ефимовым (Институт цитологии и генетики СО РАН);

3/ модификация метода Кульбака-Рао – "с конденсацией" групп; рекомендуем этот метод для случая обучающих массивов с сильно "размытыми" группами. Обучающий массив дополняется новыми признаками, генерируемыми методом Монте-Карло из групп объектов исходного массива. Это приводит к компактизации групп, позволяя более четко дискриминировать объекты. Например, имеем обучающий массив из 3-х групп, 3-х признаков, задаем "степень конденсации" массива 1 раз, получается обучающий массив из 3 групп, 6 признаков:

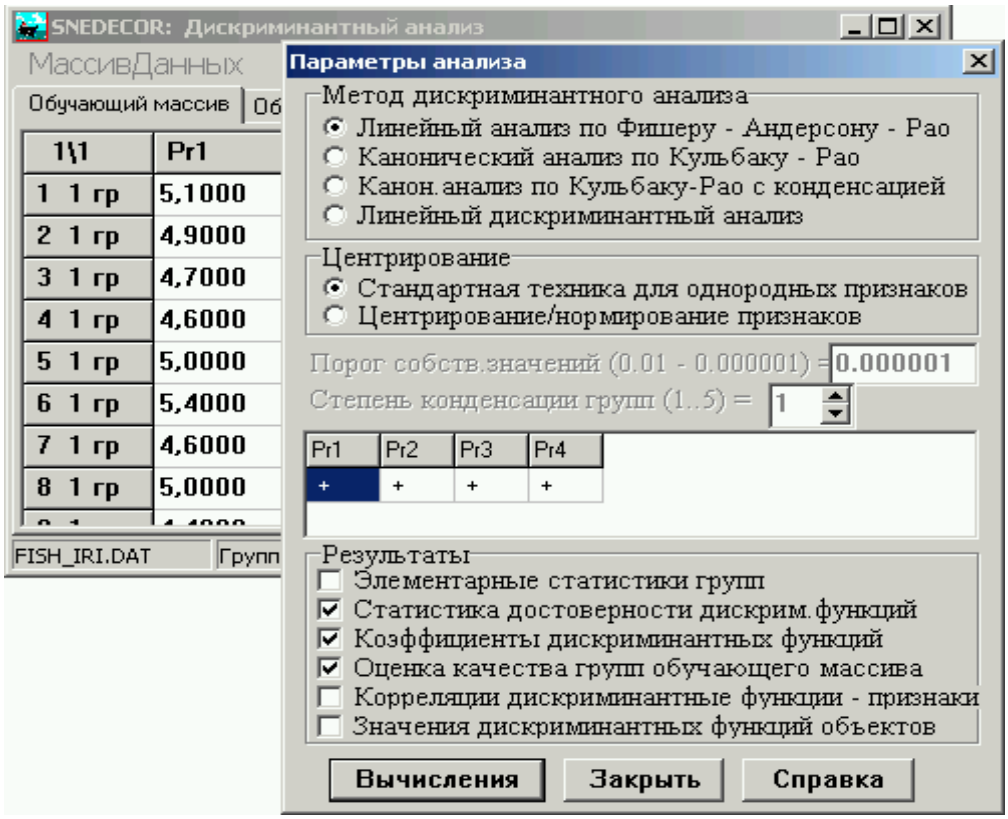
```
1    2    3        1    2    3    4    5    6
12,3 22,5 34,2     12,3 22,5 34,2 9,23 24,6 31,6
8,34 23,7 33,1     8,34 23,7 33,1 12,3 22,5 34,2     1-я группа
9,23 24,6 31,6     9,23 24,6 31,6 7,12 20,9 30,3
7,12 20,9 30,3     7,12 20,9 30,3 8,34 23,7 33,1
-------------      ---------------------------
8,27 19,4 32,4     8,27 19,4 32,4 6,21 18,5 31,6
6,21 18,5 31,6 --> 6,21 18,5 31,6 5,67 17,2 30,6     2-я группа
5,67 17,2 30,6     5,67 17,2 30,6 6,45 14,3 33,5
6,45 14,3 33,5     6,45 14,3 33,5 8,27 19,4 32,4
-------------      ---------------------------
7,23 15,2 32,1     7,23 15,2 32,1 8,34 14,9 30,5
6,78 16,2 31,9     6,78 16,2 31,9 7,23 15,2 32,1     3-я группа
8,34 14,9 30,5     8,34 14,9 30,5 6,78 16,2 31,9
 исходный          копия исходного  | новые признаки: объекты слу-
 массив            массива          | чайно выбраны из соответству-
                                    | ющих групп исходного массива
```

Объект для дискриминации формируется простым дублированием значений признаков:

```
5,87 17,7 28,9  --> 5,87 17,7 28,9 5,87 17,7 28,9
 исх. объект          копия объекта  |  копия объекта
```

Фактически это метод "сглаживания" обучающего массива и может быть использован ТОЛЬКО для предварительного анализа данных, для формирования рабочих гипотез при планировании экспериментов. Число дискриминантных функций равно общему числу признаков (исходных + добавленных).

4/ Метод линейного дискриминантного анализа [19], с. 113-134. В этом методе нет дискриминантных функций, а принадлежность объекта к группе определяется максимумом функции плотности вероятности значений объекта, вычисляемой с участием групповых ковариационных матриц и центроидов обучающего массива. Этот метод может быть более эффективен (около 5%) по показателю автодискриминации, чем классические методы ДА, но в некоторых случаях линейный ДА не может быть выполнен из-за вырожденности групповых ковариационных матриц, эта ситуация проверяется программой “Тестом сингулярности ковариационных матриц”.



Результатом работы программы являются:

1. Элементарные статистики для групп объектов: средние, среднеквадратические отклонения и коэффициенты вариации.

2. Дискриминантные функции – их эффективное количество определяется числом групп в методе Андерсона, или числом ненулевых собственных значений матрицы ковариаций (обычно на единицу меньше числа групп) в методе Кульбака-Рао:

$$\mathbf{DF = B_0 + B_1*X_1 + B_2*X_2 + ... + B_m*X_m;}$$

DF – значение дискриминантной функции для какого-либо объекта;

B_0 – свободный член (в методе Кульбака-Рао B_0=0,0); $B_1..B_m$ – коэффициенты дискриминантной функции;

$X_1..X_m$ – значения признаков для этого объекта.

Общая достоверность дискриминантных функций определяется по D-критерию Махаланобиса, точнее по аппроксимации этого значения критерием Hi^2. Чем ближе вероятность к нулю, тем достовернее разносятся объекты в заданные группы. В методе Кульбака-Рао различающая эффективность дискриминантных функций определяется критерием Hi^2 для собственных значений матрицы ковариаций (канонический анализ, массив FISH_IRI.dat):

Дискрим функция	Собственное значение	Критерий Уилкса	Критерий Hi^2	Степеней свободы	Вероятность ошибки 1 рода
Df4	0,0000	1,000	0,0000	1	1,00000
Df3	0,0000	1,000	0,0000	3	1,00000
Df2	0,2645	0,791	39,409	5	0,00000
Df1	31,955	0,024	4761,3	7	0,00000
Итог:	32,219	0,024	4800,7	8	0,00000

Дискриминантные функции: DF=B0+B1*X1+B2*X2+...+Bm*Xm

	Коэффициенты				
	B0	B1	B2	B3	B4
Df1	0,00000	-0,79959	-1,55453	2,26814	2,67244
Df2	0,00000	-0,01736	2,28991	-0,71212	2,41272
Df3	0,00000	2,06376	-2,55307	-2,92198	4,23749
Df4	0,00000	-2,46447	0,61144	0,46625	1,28679

Существенны 1-я и 2-я дискриминантные функции (P<0,00001).

3. Характеристики качества "обучающих" групп для цели дискриминации – в виде вероятностей отнесения соответствующих объектов (чем ближе средняя вероятность к 1, тем выше качество группы для дискриминации объектов) в методе Андерсона-Рао, или числом "своих" объектов в методе Кульбака-Рао.

4. Собственно результаты дискриминации – также с вероятностью попадания объекта в группу.

Программа позволяет визуально оценить качество групп для дискриминации с помощью графика "Проекции объектов на плоскость дискриминантных функций" (1-2, 1-3, 2-3 функции, и т.д.). Номер функции по оси X можно менять вводом цифры от 1 до 9, номер функции по оси Y меняется с помощью комбинации клавиш Alt/1..Alt/9. При вводе символа "N" вместо окружностей будут выво-

диться номера объектов, при вводе символа "G" будут выводиться коды групп, клавиша <пробел> возвращает представление объектов в виде окружностей, символ "O" (латинская буква) позволяет визуализировать позиции объектов, взятых для дискриминации, в виде неокрашенных окружностей большего размера:

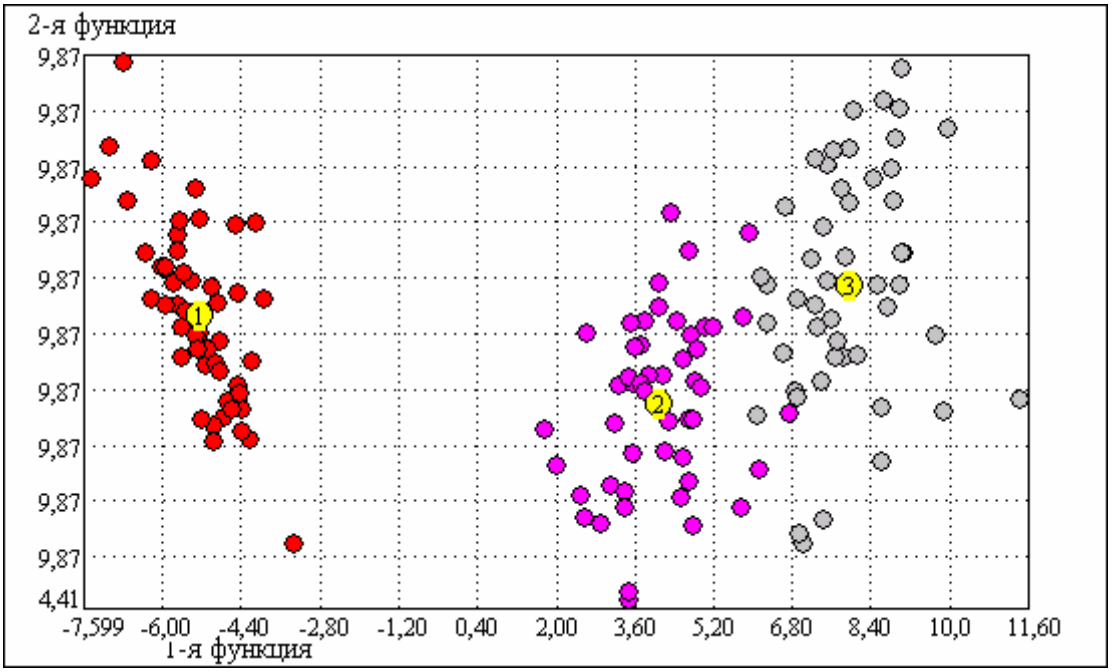


График "Гистограмма значений дискриминантной функции" также предназначен для визуальной оценки различимости групп; ширину столбиков можно менять клавишами <+> и <->, номер дискриминантной функции меняется клавишами <PgDown> и <PgUp>.

Программа тестировалась по [22], стр.254 (массив DISCSSP.dat); все вычисления выполняются с двойной точностью.

Результаты счета могут быть отредактированы непосредственно в среде программы при выводе результатов на дисплей (заголовок, комментарии, удаление ненужной информации и т.п.).

10.6.1 Дискриминирующая способность признаков

Одна из целей дискриминантного анализа – минимизация числа признаков для уверенной классификации объектов по группам. Для этого признаки обучающего массива могут быть протестированы операцией “Анализ дискриминирующей способности признаков”.

Вначале вычисляется матрица парных корреляций Пирсона; корреляции, большие чем 0.95 помечаются “*”, при наличии таких связей программа добавляет соответствующие рекомендации – какой-либо признак из связанной пары мож-

но удалить без ухудшения дискриминирующей способности обучающего массива. Отсутствие высоко коррелирующих признаков необходимо для нормального выполнения алгоритмов ДА (предпосылка невырожденности матрицы ковариаций).

Далее для каждого признака выполняется 1-факторный неравночисленный дисперсионый анализ типа “Fixed”. С помощью критерия Фишера-Снедекора доказывается достоверность различия межгрупповых средних тех признаков, которые могут быть использованы в качестве дискриминирующих переменных. Вероятность ошибки, вычисленная для F-критерия, (см. Help к программам дисперсионного анализа), должна быть порядка 0,1 и менее; по-видимому, можно безболезненно исключать те признаки, вероятность ошибки которых 0,4 – 1,0 (межгрупповые средние статистически равны, эти признаки непригодны для дискриминации). Остальные признаки, вероятность ошибки которых в диапазоне 0,4 .. 0,1, могут быть оставлены в анализе, исходя из некоторых дополнительных соображений (массив ARENS131.dat):

Признак	F-критерий	Вероятность	Рекомендации
Pr1	1,122	0,3452	исключить?
Pr2	1,277	0,3007	исключить?
Pr3	0,926	0,4126	исключить?
Pr4	0,033	0,9680	исключить!
Pr5	1,601	0,2265	исключить?
Pr6	10,999	0,0006	оставить!!
Pr7	6,368	0,0072	оставить!!
Pr8	1,160	0,3338	исключить?
Pr9	0,951	0,4031	исключить?
Pr10	24,058	0,0000	оставить!!

10.6.2. Анализ избыточности пространства признаков

С практической точки зрения одна из главных задач ДА – максимальное снижение числа признаков, требуемых для успешной дискриминации объектов. Например, для целей диагностики состояния пациентов в клиниках могут использоваться различные биохимические тесты, некоторые из них могут быть весьма дорогостоящими, но неэффективными для методов ДА.

Для анализа вклада признаков рекомендуем выполнить тест избыточности пространства параметров, основанном на последовательном исключении каждого признака и выполнении ДА с усеченными массивами. Сравнивая исходное значение Hi^2 для полного массива со значениями Hi^2 усеченных массивов, можно определить признак с минимальным снижением значения критерия. Если это сниже-

ние порядка единиц %, этот признак может быть исключен из обучающего массива, и далее тест избыточности повторен (ARENS131.dat):

Полный набор признаков: Критерий Hi^2= 199,76 c.c.=7

Исключенный признак	Критерий Hi^2	вероятность ошибки 1рода	% снижения
Pr1	192,25	0,00000	3,8
Pr2	174,26	0,00000	12,8
Pr3	185,03	0,00000	7,4
Pr4	197,88	0,00000	0,9
Pr5	199,47	0,00000	0,1*
Pr6	163,08	0,00000	18,4
Pr7	195,86	0,00000	2,0
Pr8	197,29	0,00000	1,2
Pr9	173,84	0,00000	13,0
Pr10	126,91	0,00000	36,5

5-й и, по-видимому, 4-й признаки можно исключить без снижения дискриминирующей способности обучающего массива.

Исключение неэффективных признаков следует проводить до той стадии, когда группы объектов достаточно хорошо разделяются при визуальном анализе, и вероятность для критерия Hi^2 имеет значение порядка 0,01 – 0,05.

10.7. DISCRYM: Дискриминантный анализ (номера объектов, коды групп)

Программа DISCRYM предназначена для обработки экспериментальных данных методом дискриминантного анализа. Вначале программа использует "обучающий" массив для вычисления дискриминантных функций, затем, после ввода массива дискриминируемых объектов, происходит отнесение каждого объекта в одну из заданных в "обучающем" массиве групп с некоторой вероятностью, зависящей от "качества" групп. В отличие от аналогичной программы DISCRIM, в данной программе массив с "обучающими" группами должен содержать два дополнительных столбца – в первый заносятся оригинальные номера объектов, во второй – коды групп в виде целых чисел. Такая же структура данных используется в программе MCOMP (главные компоненты).

Ограничения на размер массивов (обучающего и дискриминируемого): число признаков (M) может быть не более 80, число объектов (N) – не более 5000, но при соблюдении условия M x N <= 32000, (например, 80 x 400, 40 x 800, 20 x 1600 и т.д.), число групп – не более 40. Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в

среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример массива из 4-х признаков и 12-и объектов:

```
        Численности групп объектов, только для "обучающего" массива
          | | |              3 группы:  3+4+5=12
6  12 3 4 5                          <-  начало файла
977 1 12,3 22,5 34,2 0,34           \
978 1 8,34 23,7 33,1 0,23            1-я группа = 3 объекта
979 1 9,23 24,6 31,6 0,45           /         массив данных:
980 2 7,12 20,9 30,3 0,23           \         строки  = объекты,
981 2 8,27 19,4 32,4 0,78            \        3-6 столбцы  = признаки
982 2 6,21 18,5 31,6 0,98            2-я группа = 4 объекта
983 2 5,67 17,2 30,6 0,75           /         1-й столбец – номера объектов
984 3 8,55 16,3 33,9 0,77           \         2-й столбец – коды групп
985 3 7,23 17,6 32,1 0,82            \
986 3 6,47 15,5 31,7 0,79             3-я группа = 5 объектов
987 3 5,18 16,0 31,2 0,78            /
988 3 6,11 15,8 30,4 0,81           /

Данные 1977-88 г                    <- необязательный комментарий
```

Формировать группы объектов в массиве желательно едиными блоками, но при разбросе кодов во втором столбце программа переформирует "обучающий" массив по возрастанию значений этого столбца. Массив дискриминируемых объектов должен иметь то же число признаков, что и "обучающий" массив, и последний, в частности, по умолчанию используется для автодискриминации.

В качестве примера формирования массива можно посмотреть файл SSP8x30.dat (6 признаков, 30 объектов), а также FIS6x150.dat (4 признака, 150 объектов – классический пример с тремя видами ирисов). После загрузки обучающего массива программа вычисляет векторы групповых средних для каждого признака и заносит их в правый Табличный редактор в качестве объектов для дискриминации; на графиках "Проекции объектов" они отображаются в виде окружностей вблизи центров групп.

Программа предлагает три метода вычисления дискриминантных функций:

а/ метод Андерсона-Рао [26]; [49], стр. 452, при котором число дискриминантных функций равно числу групп в обучающем массиве, а их последовательность привязана к последовательности групп; принадлежность объекта к некоторой группе определяется максимумом из значений всех дискриминантных функций для этого объекта;

б/ метод Кульбака-Рао [13], [32], при котором общее число дискриминантных функций равно числу признаков; принадлежность объекта к группе определяется минимальным расстоянием в пространстве Махаланобиса от объекта до

центров групп, вычисляемым по совокупности всех дискриминантных функций. Алгоритм метода предложен В.М.Ефимовым (Институт цитологии и генетики СО РАН);

в/ модификация метода Кульбака-Рао – "с конденсацией" групп; рекомендуем этот метод для случая обучающих массивов с сильно "размытыми" группами. Обучающий массив дополняется новыми признаками, генерируемыми методом Монте-Карло из групп объектов исходного массива. Это приводит к компактизации групп, позволяя более четко дискриминировать объекты. Например, имеем обучающий массив из 3 групп, 3 признаков, задаем "степень конденсации" массива 1 раз, получается обучающий массив из 3 групп, 6 признаков:

1 2 3		1 2 3 4 5 6	
12,3 22,5 34,2		12,3 22,5 34,2 9,23 24,6 31,6	
8,34 23,7 33,1		8,34 23,7 33,1 12,3 22,5 34,2	**1-я группа**
9,23 24,6 31,6		9,23 24,6 31,6 7,12 20,9 30,3	
7,12 20,9 30,3		7,12 20,9 30,3 8,34 23,7 33,1	
8,27 19,4 32,4		8,27 19,4 32,4 6,21 18,5 31,6	
6,21 18,5 31,6	-->	6,21 18,5 31,6 5,67 17,2 30,6	**2-я группа**
5,67 17,2 30,6		5,67 17,2 30,6 6,45 14,3 33,5	
6,45 14,3 33,5		6,45 14,3 33,5 8,27 19,4 32,4	
7,23 15,2 32,1		7,23 15,2 32,1 8,34 14,9 30,5	
6,78 16,2 31,9		6,78 16,2 31,9 7,23 15,2 32,1	**3-я группа**
8,34 14,9 30,5		8,34 14,9 30,5 6,78 16,2 31,9	
исходный массив		копия исходного массива	новые признаки: объекты случайно выбраны из соответствующих групп исходного массива

Объект для дискриминации формируется простым дублированием значений признаков:

5,87 17,7 28,9	-->	5,87 17,7 28,9	5,87 17,7 28,9
исх. объект		копия объекта	копия объекта

Фактически это метод "сглаживания" обучающего массива и может быть использован ТОЛЬКО для предварительного анализа данных, для формирования рабочих гипотез при планировании экспериментов. Число дискриминантных функций равно общему числу признаков (исходных + добавленных).

Результатом работы программы являются:

1. Элементарные статистики для групп объектов: средние, среднеквадратические отклонения и коэффициенты вариации.

2. Дискриминантные функции – их эффективное количество определяется числом групп в методе Андерсона, или числом ненулевых собственных значений

матрицы ковариаций (обычно на единицу меньше числа групп) в методе Кульбака-Рао:

$$\mathbf{DF = B_0 + B_1*X_1 + B_2*X_2 + ... + B_m*X_m;}$$

DF – значение дискриминантной функции для какого-либо объекта;

B_0 – свободный член (в методе Кульбака-Рао B_0=0,0); $B_1..B_m$ – коэффициенты дискриминантной функции;

$X_1..X_m$ – значения признаков для этого объекта.

Общая достоверность дискриминантных функций определяется по D-критерию Махаланобиса, точнее по аппроксимации этого значения критерием Hi^2. Чем ближе вероятность Hi^2 к нулю, тем достовернее разносятся объекты в заданные группы. В методе Кульбака-Рао различающая эффективность дискриминантных функций определяется критерием Hi^2 для собственных значений матрицы ковариаций.

3. Характеристики качества "обучающих" групп для цели дискриминации – в виде вероятностей отнесения соответствующих объектов (чем ближе средняя вероятность к 1, тем выше качество группы для дискриминации объектов) в методе Андерсона-Рао, или числом "своих" объектов в методе Кульбака-Рао.

4. Собственно результаты дискриминации – также с вероятностью попадания объекта в группу.

Программа позволяет визуально оценить качество групп для дискриминации с помощью графика "Проекции объектов на плоскость дискриминантных функций" (1-2, 1-3, 2-3 функции, и т.д.). Номер функции по оси X можно менять вводом цифры от 1 до 9, номер функции по оси Y меняется с помощью комбинации клавиш <Alt/1>..<Alt/9>. При вводе символа «N» вместо окружностей будут выводиться номера объектов, при вводе символа «G» будут выводиться коды групп, клавиша <пробел> возвращает представление объектов в виде окружностей, символ «O» (латинская буква) позволяет визуализировать позиции объектов, взятых для дискриминации, в виде неокрашенных окружностей большего размера.

График "Гистограмма значений дискриминантной функции" также предназначен для визуальной оценки различимости групп; ширину столбиков можно менять клавишами <+> и <->, номер дискриминантной функции меняется клавишами <PgDown> и <PgUp>.

Программа тестировалась по [22], стр.254 (массив DISCSSP.dat); все вычисления выполняются с двойной точностью. Результаты счета могут быть отредак-

тированы непосредственно в среде программы при выводе результатов на дисплей (заголовок, комментарии, удаление ненужной информации и т.п.).

10.8. DIAGNOZ: Классификация объектов по Байесу

Программа DIAGNOZ предназначена для вероятностного отнесения объектов (классификации по Байесу) к одной из нескольких групп; это могут быть:

Объекты	Группы	Признаки
Пациенты	Заболевания	Симптомы
Особи	Животные/растительные виды	Фенотип
Растения	Сорта, линии	Фенотип/биохимия
Образцы	Горные породы, почвы	Минер. состав, химия
Предприятия	Группы риска, надежности	Показатели производства

Для этого необходимо, чтобы объект можно было охарактеризовать набором признаков – вектором из нулей и единиц (признак отсутствует/присутствует).

Предварительно должен быть подготовлен массив "группы-признаки", в котором содержатся частоты встречаемости признаков в известных (эталонных) группах объектов в виде дробных чисел от нуля до 1.0. Затем по бинарному вектору значений признаков для объекта вычисляются вероятности попадания в каждую группу.

Максимальная вероятность определяет отнесение объекта к группе; естественно, частоты встречаемости признаков должны как можно более различаться для групп объектов, признаки с близкими частотами не должны включаться в обработку.

Ограничения на размер массива: число групп (M) может быть не более 100, число признаков (N) – не более 4000, но при соблюдении условия M x N не более 32 тысячи элементов.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х групп и 9-и признаков в текстовом файле:

```
4 9
0,13 0,35 0,40 0,14
0,22 0,13 0,37 0,23
0,36 0,06 0,13 0,45
0,41 0,17 0,44 0,23
0,50 0,15 0,01 0,38
0,00 0,00 0,02 0,98
0,13 0,16 0,08 0,75
0,45 0,23 0,13 0,76
0,35 0,28 0,10 0,78
```

<- начало файла

массив данных:
4 столбца = 4 вида заболеваний,
9 строк = симптомов.

Пример массива "объекты-признаки":

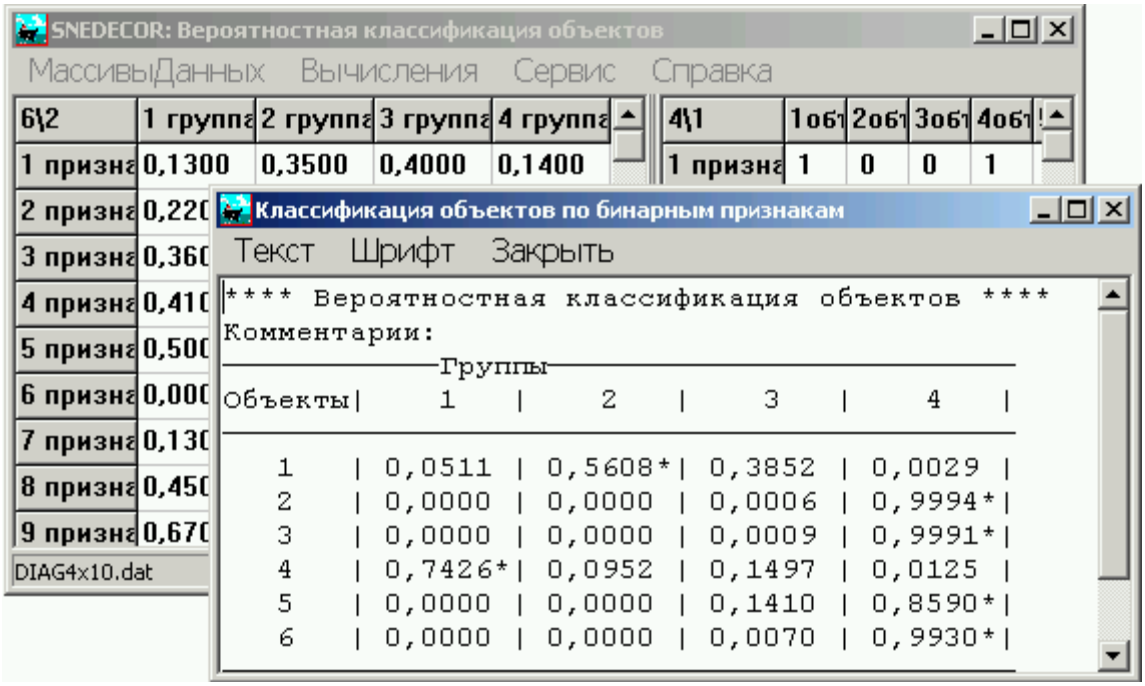
```
6 9
1 0 0 1 1 0
1 1 0 0 0 1
0 0 1 0 1 1
0 1 0 1 1 1
0 1 1 0 1 0
0 1 1 0 1 1
1 0 0 0 0 0
1 1 0 1 0 1
0 1 0 1 0 1
```

<- начало файла

массив данных:
6 столбцов = пациентов,
9 строк = симптомов.

В качестве примера формирования массива частот для программы DIAGNOZ можно посмотреть файл DIAG7x50.dat (7 групп, 50 признаков). В программу можно загрузить массив частот в виде целых чисел; в этом случае предполагается, что любое значение частоты получено из выборок фиксированного размера (например, всегда 100 объектов), тогда перед вычислением строки массива (признаки) нормализуются делением на максимальные элементы строк. Это, вообще говоря, не совсем правильно – надо нормализовать делением на размер выборки в каждом конкретном случае.

Результатом работы программы является таблица вероятностей попадания объектов в различные группы. Звездочками "*" помечено максимальное значение вероятности для объекта, которое и определяет отнесение к группе (массив DIAG4x10.dat, объекты для классификации – массив DIAG6x10.dat):



1-й объект с вероятностью 0,561 отнесен ко 2-й группе, 2-й, 3-й, 5-й и 6-й объекты с P>0,859 – к 4-й группе, 4-й объект с P=0,743 – к 1-й группе.

Формулы, использованные в программе, взяты из [4, стр. 340-342].

10.9. HOTELL: Анализ многомерных данных по Хотеллингу

Программа HOTELL предназначена для анализа массивов многомерных данных типа "признаки-объекты-группы":

1/ проверка массива данных на наличие многомерных выбросов по критерию Махаланобиса (в форме F-критерия);

2/ сравнение средних по всем признакам с вектором средних генеральной совокупности по критерию Т-квадрат Хотеллинга;

3/ сравнение двух векторов многомерных средних с одинаковым или различным числом объектов по критерию Т-квадрат Хотеллинга;

4/ проверка однородности нескольких эмпирических ковариационных матриц по критерию Hi^2;

5/ вычисление "многомерных" доверительных интервалов вектора средних.

Основные предположения, требуемые для корректности анализа многомерных средних – нормальность распределения данных во всех признаках и отсутствие выбросов. Эти предпосылки в одномерной статистике можно проверить с помощью программ NORMAL и IODATA. Для многомерных данных наличие выбросов можно выявить с помощью критерия D-квадрат (выборочное расстояние Махаланобиса), преобразованного в статистически эквивалентную форму – критерий Фишера-Снедекора (1-й метод обработки данных).

Ограничения на размер массива: число признаков (M) может быть не более 80, число объектов (N) – не более 5000, но при соблюдении условия M x N <= 100000, число групп – не более 20. Число объектов в любой группе должно быть больше числа признаков не менее чем на 2. Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 6-и признаков и 10-и объектов:

```
      Численности групп объектов
      | | |          3 группы:  4+3+3=10
6 10 4 3 3                                  <- начало файла
12,3 22,5 34,2 0,34 1,45 3,11               \
8,34 23,7 33,1 0,23 1,66 3,65                1-я группа
9,23 24,6 31,6 0,45 1,89 2,79               /          массив данных:
7,12 20,9 30,3 0,23 1,73 3,09               /          строки  = объекты,
8,27 19,4 32,4 0,78 1,77 3,35               \          столбцы  = признаки
6,21 18,5 31,6 0,98 1,85 3,69                2-я группа
5,67 17,2 30,6 0,75 1,57 3,51               /
7,23 17,6 32,1 0,82 1,21 3,22               \
6,47 15,5 31,7 0,79 1,42 3,74                3-я группа
5,18 16,0 31,2 0,78 1,63 3,71               /

Данные 1997 г                               <- необязательный комментарий
```

В качестве примера формирования массива можно посмотреть файл FISH_IRI.dat (4 признака, 150 объектов – классический пример с тремя видами ирисов), а также файл DEV4x14.dat (4 признака, 14 объектов в 2-х группах). Аналогичная структура массивов данных используется в программах MCOM, DISCRIM и MANOVA1.

Пример теста на многомерные выбросы (массив ARENS131.dat):

Объект	D2	F-критерий	Вероятность	Выброс?
1	10,977	0,368	0,9164	-
2	69,645	2,332	0,1813	-
3	19,873	0,665	0,7275	-
4	640,942	21,460	0,0017	!!
5	27,468	0,920	0,5763	-
6	20,687	0,693	0,7101	-
7	71,998	2,411	0,1718	-
8	6,603	0,221	0,9796	-
9	21,022	0,704	0,7030	-
10	47,471	1,589	0,3178	-
11	17,924	0,600	0,7701	-
12	217,619	7,286	0,0203	!
13	21,174	0,709	0,6997	-
14	30,348	1,016	0,5270	-
15	41,954	1,405	0,3716	-
16	32,808	1,098	0,4885	-

4-й и 12-й объекты – явные артефакты или ошибки измерения.

Для анализа различия многомерных средних (2-й метод) в программе используется критерий Т-квадрат Хотеллинга (многомерный аналог Т-критерия Стьюдента), преобразованный аналогичным образом в критерий Фишера-Снедекора [4], стр. 321.

3-й метод – анализ различия векторов средних двух групп объектов как одномерным критерием Стьюдента, так и многомерным критерием Хотеллинга. Анализ по Хотеллингу выполняется только в том случае, когда число признаков меньше числа объектов.

В случае анализа однородности ковариационных матриц (4-й метод обработки данных) вычисляется М-статистика [18], стр. 468, и ее аппроксимация – критерий Hi^2. Это приближение удовлетворительно работает при числе объектов в каждой группе не менее 20. Проверка однородности двух ковариационных матриц – необходимая предпосылка корректности анализа различия двух векторов многомерных средних (3-й метод анализа данных программы HOTELL). Контр-гипотеза при анализе ковариаций формулируется следующим образом: по меньшей мере две ковариационных матрицы достоверно отличаются друг от дру-

га. В качестве числового параметра, характеризующего степень различия матриц ковариаций, можно рассматривать детерминанты матриц.

5-й метод анализа интересен тем, что при значительной коррелированности признаков позволяет получить более узкие, по сравнению со стандартными одномерными методами (на основе Т-критерия Стьюдента), доверительные интервалы для средних по совокупности признаков, характеризующих единый процесс или явление. Заданные доверительные интервалы (с вероятностью попадания истинных средних 90, 95 или 99%) вычисляются итерационным алгоритмом, применяя многомерную T^2-статистику Хотеллинга в форме F-критерия Фишера-Снедекора; в качестве начального приближения используются одномерные доверительные интервалы (массив FISH_IRI.dat):

****** Вычисление многомерных доверительных интервалов**

Признак	Среднее	Доверительные интервалы, 95%			
		одномерные	многомерные		
150 объектов в объединенном массиве					
X1	5,84333	± 0,1336	± 0,1020	5,7413 ...	5,9453
X2	3,05733	± 0,0703	± 0,0537	3,0036 ...	3,1110
X3	3,75400	± 0,2841	± 0,2169	3,5371 ...	3,9709
X4	1,19533	± 0,1226	± 0,0936	1,1017 ...	1,2890

10.10. MRAN: Многомерное ранжирование объектов (сортов, животных, предприятий)

Программа MRAN предназначена для многомерного ранжирования объектов исследования (сортов сельскохозяйственных культур, животных, образцов почвы и т.д.) по совокупности признаков, значения которых, возрастая от минимума к максимуму, отражают их хозяйственную (или иную) ценность. Фактически это метод автоматической классификации объектов в три группы – по принципу увеличения расстояния объектов от начала координатных осей, выраженного обобщенным рангом.

Поскольку признаки могут иметь различную степень полезности, можно задать некоторую систему весов для каждого признака – в виде чисел от –10,0 до 10,0 (отрицательное значение веса – для признаков, отражающих негативные, нежелательные свойства объектов); если не задавать веса, программа считает, что все признаки имеют равную степень полезности (=1,0). Для одного и того же массива данных можно сделать несколько ранжировок, варьируя систему весов.

Алгоритм множественного ранжирования разработан А.И.Южаковым (СибНИИЗХим СО РАСХН) [63].

Ограничения на размер массива: M может быть не более 100, N - не более 5000, но при соблюдении условия M x N <= 100000.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла.

Номера объектов автоматически формируются программой от 1 до N при вводе массива "признаки-объекты" с клавиатуры или вызове файла данных. Вся совокупность объектов распределяется программой на три равные группы – "лучшие", "средние" и "худшие" объекты. Размер группы может быть изменен пользователем в сторону уменьшения от 1/3 общего числа объектов, но если указать размер группы более этого числа, то разбиение объектов на 3 группы не будет производиться, и программы выведет одну общую таблицу пар “ранг объекта – номер объекта”.

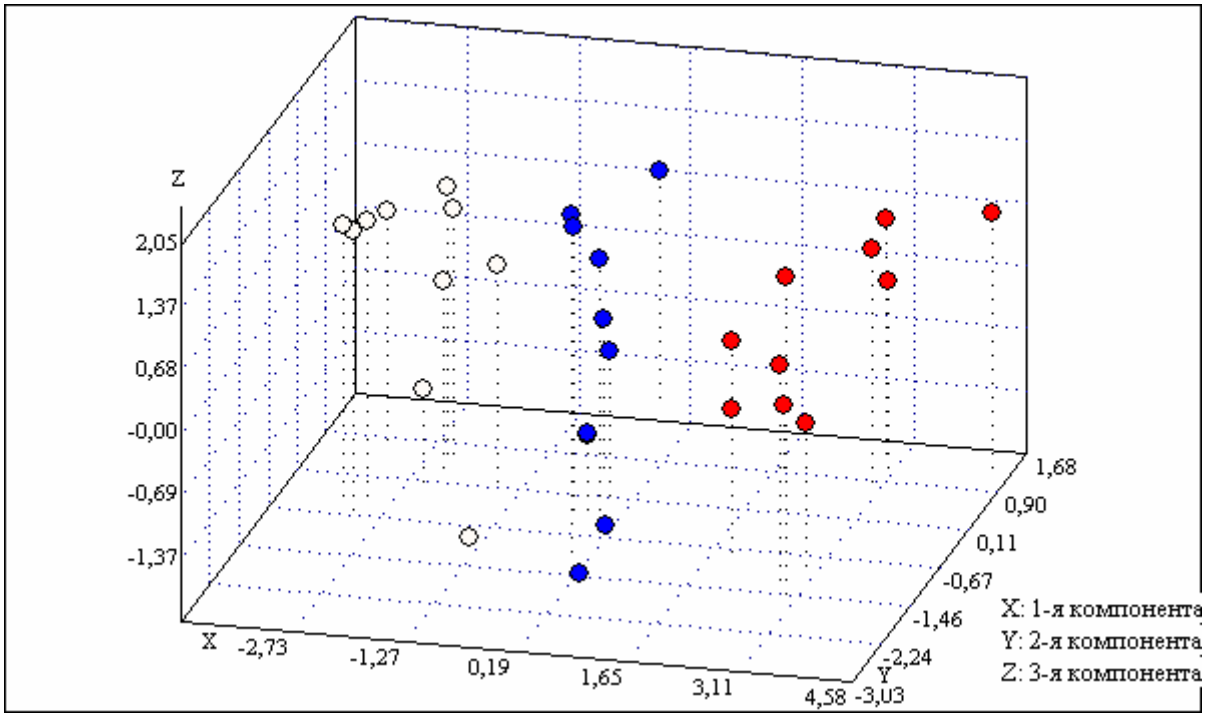
Результат работы программы – таблица пар "место объекта – исходный номер объекта" (массив SSP6x30.dat, заданы веса: 1, 2, 1, 2, 1, 2):

Место объекта	Исходный номер	Сумма рангов	Сумма взвешенных рангов	Эвклидово расстояние
Группа "лучших" объектов:				
1	29	144,0	227,0	499,6
2	28	130,0	209,5	720,8
3	19	130,0	209,5	458,9
4	21	136,0	205,5	427,7
5	9	121,0	182,5	453,5
Группа "средних" объектов:				
13	18	104,0	155,5	523,8
14	17	103,0	139,5	387,6
15	23	80,5	129,0	463,8
16	24	96,0	127,5	303,5
17	16	88,0	124,5	419,5
Группа "худших" объектов:				
26	14	66,5	96,5	357,3
27	30	71,0	94,0	284,8
28	10	67,5	86,0	320,4
29	4	44,5	69,0	405,4
30	1	34,5	56,0	372,1

Дополнительно вычисляются Эвклидовы расстояния от объектов до центра осей координат (длина вектора) в пространстве признаков. Эти значения некоторым образом коррелируют с результатами многомерного ранжирования – сумма-

ми рангов. Анализ Эвклидовых расстояний может помочь в принятии решения – оставлять конкретный объект в группе, или исключить.

Для визуального контроля разбиения множества объектов на 3 группы рекомендуем проанализировать графики “Объекты на плоскости главных компонент”, “Объекты в пространстве главных компонент”. При этом в многомерном пространстве формируются новые некоррелирующие стандартизованные признаки – “главные компоненты”, причем оси координат располагаются таким образом, что позволяют наиболее информативным образом повернуть облако объектов (вращение клавишами “влево”, “вправо”). Легко обнаруживаются группы сходных объектов, объекты с промежуточными свойствами, объекты-артефакты (массив ABC8x50.dat, группы по 10 объектов, все веса=1,0):



Слева направо: “худшие”, “средние”, “лучшие” объекты.

Клавишей “G” можно определить группы объектов, клавишей “N” – номера объектов. “Пробел” возвращает представление объектов в виде окружностей, клавиши “+” и “-” меняют диаметр этих окружностей.

Сформированные таким образом три группы объектов можно записать в виде файла данных, дополненного слева двумя столбцами – исходными номерами объектов и номерами групп (1..3). Такой массив может быть использован для обработки другими методами:

– программой MCOMP (главные компоненты);

– программой DISCRYM (дискриминантный анализ);

– программой MANOVA1 (многомерный дисперсионный анализ, после удаления первых двух столбцов);

– программой HOTELL (частные виды многомерного анализа по Хотеллингу).

10.11. METRIC: метрическое шкалирование по Торгерсону

Программа METRIC предназначена для обработки данных методом метрического шкалирования по Торгерсону. Анализ этим методом используется для поиска «структуры объектов» (по терминологии из психологии/социологии «структуры стимулов») в пространстве небольшой размерности (2-3-4). Подобно методам дискриминантного, кластерного, факторного анализов изучаются группы, общности объектов в пространстве шкал – новых переменных, аналогичных главным компонентам в факторном или компонентном анализах.

Массив данных должен быть квадратной, симметричной относительно диагонали матрицей различия/сходства объектов. Ограничения на размер матрицы: не более 400 объектов, максимальный размер массива – 120000 элементов (400 х 400 и менее).

Данные могут быть введены с клавиатуры непосредственно в среде программы, переданы через буфер Windows из программы MS Excel, из файла в стандарте пакета SNEDECOR, подготовленного заранее с помощью какой-либо программы пакета или любого редактора текстов типа Блокнота Windows. На диагонали матрицы должны быть нулевые значения. Пример формирования матрицы из 6 х 6 объектов в текстовом файле:

```
6 6
0,00 22,5 34,2 0,34 1,45 3,11
22,5 0,00 33,1 0,23 1,66 3,65
34,2 33,1 0,00 0,45 1,89 2,79
0,34 0,23 0,45 0,00 1,73 3,09
1,45 1,66 1,89 1,73 0,00 3,35
3,11 3,65 2,79 3,09 3,35 0,00

Данные 1997 г.
```

<- начало файла

массив данных:

<- необязательный комментарий

Одно из важных свойств матрицы различий – неотрицательность элементов; близкие к нулю значения говорят о сходстве объектов, с увеличением значений сходство объектов падает, отражая степень различия.

Таким образом, не всякая симметричная матрица может быть использована для обработки методом метрического шкалирования. Например, матрица различий D может быть получена из матрицы R парных корреляций объектов:

$$\mathbf{D_{ij} = 10{,}0 \times \left(1{,}0 - R_{ij}\right)}$$

$$\mathbf{D_{ij} = 14{,}142 \times \sqrt{\left(1{,}0 - R_{ij}\right)}}$$

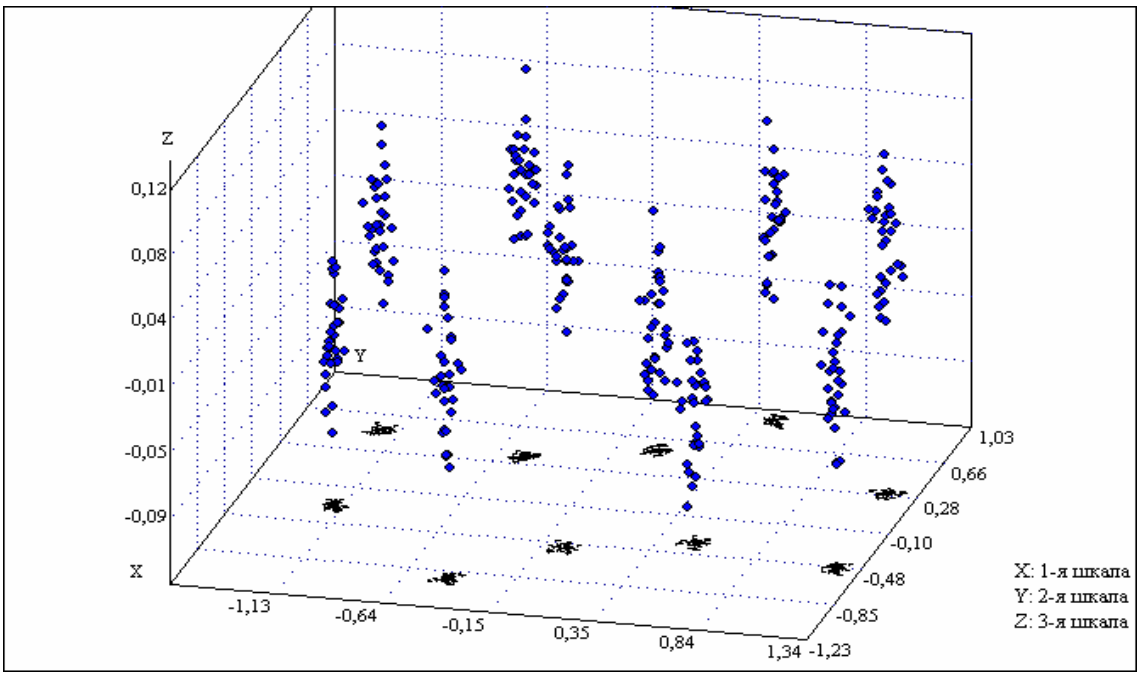
Тогда на диагонали матрицы – нулевые значения, внедиагональные значения – сходства/различия объектов. Множители (диапазон различий 0,0...20,0) выбраны из соображения более удобного представления результирующих значений координат объектов.

Также это может быть матрица эвклидовых расстояний между объектами в пространстве исходных признаков или в пространстве главных компонент. Второй вариант (главные компоненты) предпочтительнее вследствие стандартизации масштабов по всем осям пространства компонент.

Это могут быть матрицы мер сходства/различия, полученные на других принципах (вероятности попарных совпадений, расстояния Махаланобиса, Манхеттенское, в пространстве Минковского и т.п.).

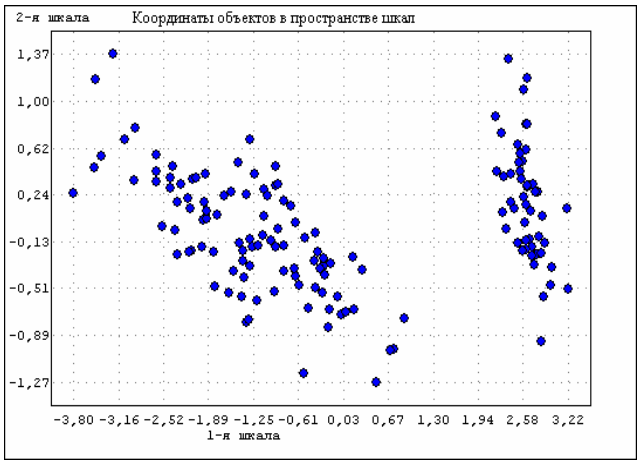
В качестве примера формирования матрицы различий для программы METRIC можно посмотреть файлы BART8x8.dat, GAID6x6.dat, Johnson6x6.dat.

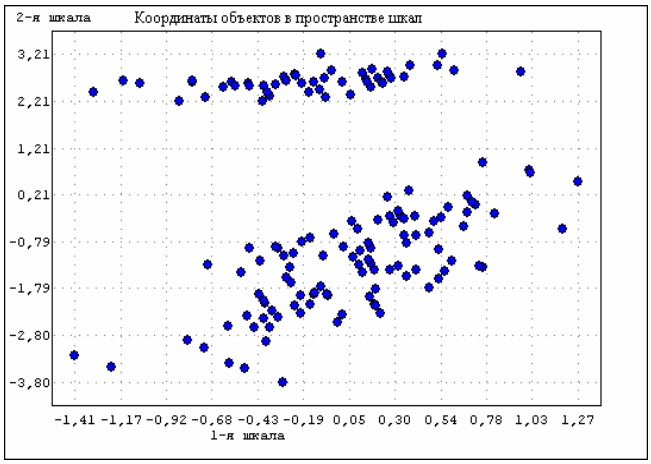
Результат работы программы – таблица координат объектов в пространстве новых, наиболее информативных шкал. Обычно рассматривают координаты объектов в плоскости 1-й и 2-й шкал, в большинстве случаев возможен анализ координат и в трехмерном пространстве (тестовый массив CLUST300x300.dat):



Специфика метода такова, что получаемое решение не является единственным, существует бесчисленное множество решений, отличающихся от базового решения только одновременным поворотом осей в каких-либо направлениях с сохранением ортогональности. Выбор решения остается за пользователем, который должен на основе своего понимания исследуемой системы выбрать наиболее информативное.

Для этого графическое представление координат в 2-мерном пространстве подвергают вращению для поиска такой конфигурации объектов, которая наиболее эффективно отражает сущность системы (тестовый массив Fisher150.dat):





После достижения желаемой конфигурации таблицу новых координат можно добавить к основным результатам шкалирования.

Для изучения метода метрического шкалирования можно использовать массивы Fisher150.dat (на основе классического примера Р.Фишера), Clust300x300.dat (10 четко выраженных групп по 30 объектов). В последнем случае время обработки массива – 10..20 секунд, все вычисления выполняются операциями с двойной точностью.

11. Временные ряды

11.1. TREND: Анализ временных рядов

Программа TREND предназначена для обработки экспериментальных данных, представленных временными рядами, различными статистическими методами:

1/ вычисление автокорреляционной и кросскорреляционной функций;

2/ сглаживание значений ряда различными методами;

3/ аппроксимация тренда различными функциями Методом Наименьших Квадратов и удаление тренда из ряда;

4/ вычисление различных статистический параметров ряда.

5/ определение главных гармоник ряда по Шустеру;

6/ прогнозирование значений ряда МНК-авторегрессией.

Для визуального анализа временных рядов имеются различные графические процедуры – выявление тренда, наличие цикличности.

Программа может использовать массив типа "признаки-объекты" с многомерным временным рядом, и использовать для анализа тренда любой из признаков. Ограничения на размер массива: число признаков (M) может быть не более 100, число объектов (N) – не более 10000, но при соблюдении условия M x N <= 100000.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 2-х признаков и 18 временных точек в текстовом файле, первым признаком здесь введен вектор отсчетов времени (необязателен): В качестве примера формирования массива можно посмотреть файл SSP6x30.dat (6 переменных, 30 временных точек), SUN2x72.dat, многолетние данные по числам Вольфа (солнечным пятнам).

```
2 17
1961 22,5
1962 23,7
1963 24,6
1964 20,9
1965 19,4
1966 18,5
1967 17,2
1968 17,1
1969 17,9
1970 18,3
1971 18,5
1972 18,7
1973 18,6
1974 18,8
1975 20,2
1976 17,0
1977 17,6
Годы
Урожай

Урожай зерновых
```

<- начало файла

массив данных:
строки = последов. Отсчеты времени,
столбцы = временные ряды, признаки

<- необязательные наименования признаков

<- необязательный комментарий

Собственно анализ тренда осуществляется выбором пункта Меню "Графика"; по умолчанию анализ начинается с первого признака; далее признаки меняются с помощью клавиш <СтрелкаВлево> и <СтрелкаВправо>. Автоматически выполняется вычисление функции, аппроксимирующей временной ряд линейной зависимостью (по умолчанию; можно выбрать из 20 функций).

В программе могут быть использованы усложненные формулы усреднения – с двойным сглаживанием, с увеличенным весом центральной точки, методом быстрого преобразования Фурье. По умолчанию программа использует обычные формулы усреднения – с равными весами всех дат. Простое усреднение хорошо работает в случае временных рядов без дрейфа (тренда); при наличии выраженного тренда сглаженные значения будут завышены при возрастающей зависимости, и занижены при снижающейся.

Наиболее эффективный метод сглаживания – МНК-сплайн с минимизацией Суммы Квадратов Отклонений и Интеграла Квадратов Вторых Производных в узлах сплайна. Доли СКО и ИКВП формируются коэффициентом, значение которого можно плавно менять от 0.0 до 1.0. Результирующий сплайн визуально подбирается (Меню дополнительных операций правой клавишей мышки) – наилучшим образом проходящий вблизи точек исходного временного ряда с исключением высокочастотных случайных гармоник.

11.1.1. Прогноз рядов методом авторегрессии

Метод прогноза разработан автором самостоятельно, возможно он имеет какое-то название, возможно это модификация какого-либо известного метода.

Последовательность операций при прогнозировании временных рядов:

1/ Если ряд зашумлен случайными процессами, выполняется сглаживание ряда каким-либо способом, сглаженный ряд заносится в массив данных как M+1 столбец (M = исходное число столбцов).

2/ Вводится величина лага, исходя из некоторых либо теоретических, либо интуитивных соображений относительно возможной главной гармоники, определяющей периодические изменения значений анализируемого признака, зашумленного случайными факторами. Оптимальное значение лага может быть определено из анализа графика ряда, или выбрано начальное значение около 5..7.

3/ Временной ряд после сглаживания и указания лага преобразуется программой в двумерный массив следующим образом (например, ряд из 10 элементов, лаг=4):

```
x0      x0
x1      x1 x0              обрабатываемый массив  4 x 7
x2      x2 x1 x0
x3      x3 x2 x1 x0         x3 x2 x1 x0
x4  ->  x4 x3 x2 x1   ->    x4 x3 x2 x1
x5      x5 x4 x3 x2         x5 x4 x3 x2
x6      x6 x5 x4 x3         x6 x5 x4 x3
x7      x7 x6 x5 x4         x7 x6 x5 x4
x8      x8 x7 x6 x5         x8 x7 x6 x5
x9      x9 x8 x7 x6         x9 x8 x7 x6
           x9 x8 x7
              x9 x8
                 x9
```

Таким образом, формируется массив “независимых” переменных, столбец значений “зависимой” переменной, например, для вышеприведенного примера данных Y – “зависимая” переменная, V1, V2 и V3 – “независимые” переменные:

```
          ----------------------
          |  Y    |  V1   V2   V3 |
          ----------------------
          |  x3   |  x2   x1   x0 |
          |  x4   |  x3   x2   x1 |
          |  x5   |  x4   x3   x2 |
          |  x6   |  x5   x4   x3 |
          |  x7   |  x6   x5   x4 |
          |  x8   |  x7   x6   x5 |
          |  x9   |  x8   x7   x6 |
          ----------------------
прогноз:  |  Pr   |  x9   x8   x7 |
          ----------------------
```

4/ Методом наименьших квадратов вычисляются коэффициенты B0..B3 множественной линейной регрессии:

Y = B0 + B1*V1 + B2*V2 + B3*V3;

5/ вычисляется прогнозное значение временного ряда по формуле:

Pr = B0 + B1*x9 + B2*x8 + B3*x7

и определяется доверительный интервал прогнозного значения с достоверностью 90%.

6/ Если задано несколько точек прогноза, вычисленное значение прогноза дополняет исходный временной ряд, и вычисление следующей точки повторяется. Не следует задавать число точек для прогнозирования больше значения лага.

Достоверность прогноза определяется значимостью коэффициентов регрессии по критерию Стьюдента, общей значимостью уравнения по критерию Фишера-Снедекора (массив AIRLINE.dat, после элиминации тренда, лаг=6):

Коэффициенты регрессии	Стандартная ошибка	Критерий Стьюдента T	вероятность	Корреляция между Y и Xi
B0 -0,824100	2,2245	0,370	0,7116	
B1 1,015001	0,0845	12,01	0,0000*	0,7275
B2 -0,501619	0,1230	4,078	0,0001*	0,2850
B3 0,066338	0,1306	0,508	0,6122	-0,0981
B4 -0,273537	0,1320	2,073	0,0402*	-0,3412
B5 0,307240	0,1270	2,419	0,0169*	-0,4221
B6 -0,339157	0,0903	3,758	0,0003*	-0,4576

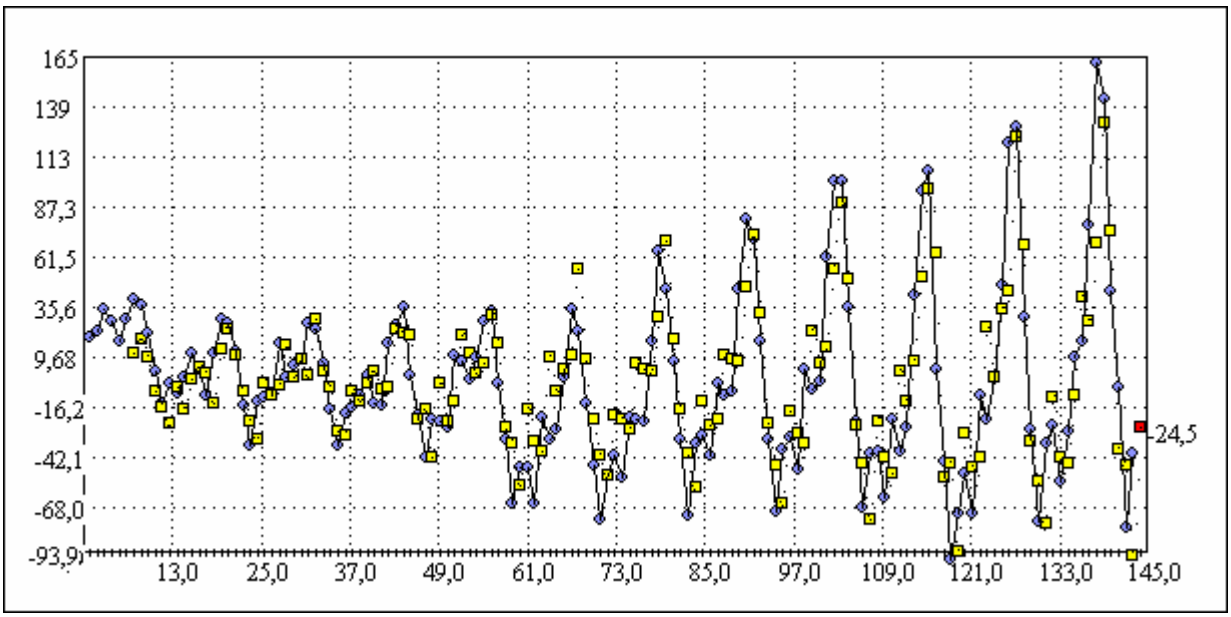
2. Достоверность регрессии
F-критерий = 50,728 c.c.=6, 131 Вероятность=0,00000
Относительная ошибка аппроксимации Er= 270,16%
Коэффициент множественной корреляции: R= 0,923654
Коэффициент детерминации: D= 0,853136

4. Прогноз ряда, 90% дов.интервал

No	Значение	Delta	Доверит. интервал
145	-24,50	23,67	-48,17 ..-0,837

Уравнение авторегрессии достоверно (P<0,00001), практически все коэффициенты регрессии также достоверны, прогнозному значению можно доверять.

Качество прогноза анализируется визуально на графике авторегрессии. На графике: положение исходных (возможно сглаженных) точек отражается синими окружностями, приближение ряда авторегрессией – желтыми квадратами, прогноз – красными квадратами, сопровождаемыми числовыми значениями (массив AIRLINE.dat, после элиминации тренда):



Прогноз в ряде случаев можно улучшить, исключая те “независимые переменные”, вклад которых недостоверен по критерию Стьюдента, но включая возможные квадратичные, кубические эффекты других “независимых переменных”. Для этого следует вручную сформировать массив для множественного регрессионного анализа с помощью программы IODATA, переместить “зависимую” переменную в крайний правый столбец, затем передать такой массив программе MLREG. В среде этой программы можно проверить значимость различных нелинейных эффектов, создавая дополнительные “независимые” переменные из исходных переменных. Прогноз в этом случае рассчитывается только на одну точку.

11.1.2. Методы сглаживания значений временных рядов

Программа предлагает сглаживание ряда различными алгоритмами:

1/ Методы простой, двойной и взвешенной скользящей средней, в последнем случае для увеличения влияния центральных членов коэффициенты вычисляются по формуле Миллера [46; стр. 39]. Степень сглаживания может изменяться перебором величины группы значений для усреднения от 3 до 9. При четных значениях все равно в счет берется нечетное число членов, но крайние включаются с весом 0,5.

2/ Весьма эффективным является метод сглаживания с помощью Быстрого Преобразования Фурье. Степенью сглаживания можно манипулировать в широком диапазоне значений коэффициента сглаживания, изменяемого с шагом 0,5.

3/ Сглаживание сплайн-регрессией с минимизацией суммы квадратов отклонений. Узлы сплайна располагаются на равных расстояниях друг от друга по

оси X и, соответственно, не совпадают с точками X исходных данных. Подгонка сплайн-функции заключается в выборе оптимального числа узлов; эффективность этого метода определяется сглаживанием переменной Y минимизацией суммы квадратов отклонений (МНК по Гауссу). Метод изложен в [54].

4/ Сглаживание сплайн-регрессией с минимизацией суммы квадратов отклонений и интеграла квадратов вторых производных в узлах сплайна. Узлы проходят по всем временным точкам, но степенью сглаживания можно манипулировать в широком диапазоне плавного изменения коэффициента сглаживания переменной Y – от 0,0000 (максимальное сглаживание в прямую линию) до 0,9999 (сплайн проходит практически по исходным экспериментальным точкам). Шаг изменения коэффициента устанавливается перед стартом операции (St=0,1), и далее автоматически уменьшается при приближении к границам – к 0,0 или к 1,0. Метод предложен К.Эбертом и Х.Эдерером [57].

Во всех случаях сглаженный ряд может быть занесен в какой-либо столбец массива данных, или же для него можно указать дополнительный (M+1) столбец массива, и затем использовать для последующего анализа.

Техника работы по сглаживанию ряда:

– указать столбец массива с рядом, столбец для записи сглаженных значений, выбрать мышкой метод сглаживания, можно выбрать число точек сглаживания для первой группы методов, для Фурье- и сплайн- методов начальные коэффициенты формируются автоматически;

– получить текстовые результаты сглаживания ряда;

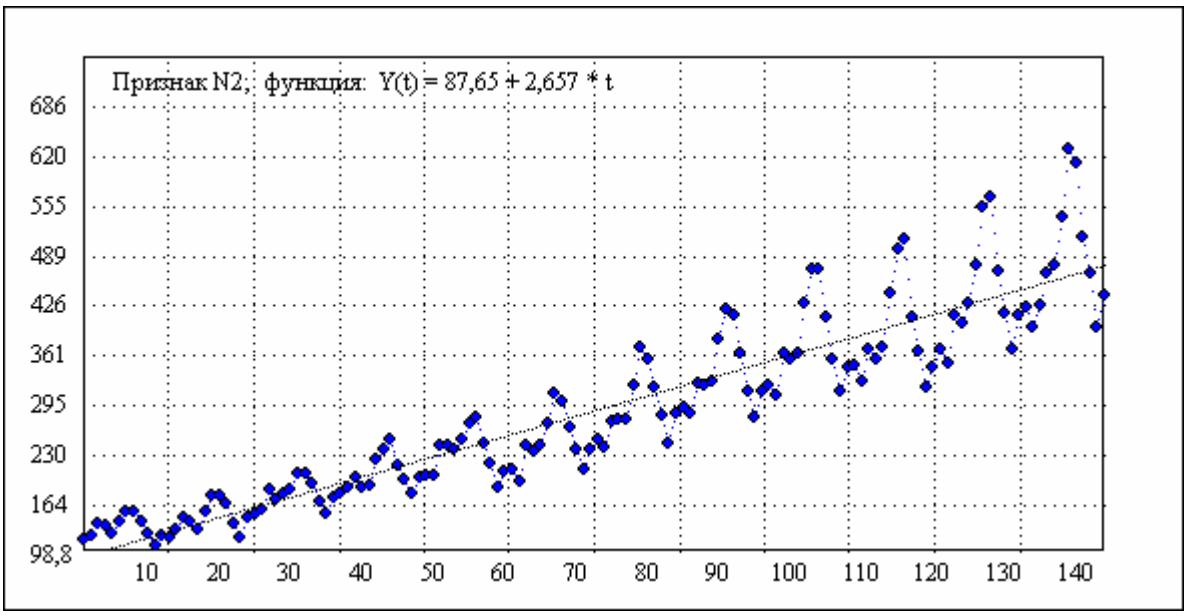
– получить график “Сглаженный временной ряд”;

– кликнуть **правой** клавишей мышки на графике – в Меню дополнительных операций выбрать пункт “Увеличить степень сглаживания” или “Уменьшить степень сглаживания”;

– добиться оптимального сглаживания временного ряда этими операциями, сглаженный ряд будет автоматически заноситься в заданный столбец массива, текст с результатами также будет обновляться.

11.1.3. Анализ и элиминация тренда

Исключение тренда – типовое условие дальнейшего анализа свойств временного ряда. Под трендом понимается длительное монотонное возрастание или убывание значений ряда, в идеале это линейная функция, хуже, когда это нелинейная функция неизвестного вида (массив AIRLINE.dat):



В последнем случае приходится подбирать некоторую функцию выбором из списка (20 видов), все эти функции выбраны потому, что с ними можно сделать преобразование к линейному виду и вычислить коэффициенты уравнения методом наименьших квадратов.

Выбор функции – исключительно визуальный, обычные статистические критерии типа Стьюдента или Фишера-Снедекора малоприменимы, так как практически всегда они говорят о недостоверности регрессии. На графике тренда нужно кликнуть **правой** клавишей мышки – в Меню дополнительных операций выбрать пункт “Следующая функция тренда”. В левом верхнем углу графика приводится вид функции. Перебирая функцию за функцией, следует выбрать оптимальную по характеру ряда.

Помимо элиминации тренда, следует делать стандартизацию значений ряда – приведение к нулевому среднему и единичной дисперсии, это несколько облегчает последующий анализ ряда.

Ряд после удаления тренда и стандартизации автоматически заносится в заранее указанный столбец массива данных.

11.2. PROGNOZ: Анализ временных рядов методом ГК

Программа PROGNOZ предназначена для обработки одномерных временных рядов методом главных компонент по Ефимову-Галактионову [35]. Временной ряд может иметь тренд линейного характера, или близкий к линейному.

Ограничения на размер массива первичных данных: число временных рядов (M признаков) может быть не более 100, число точек (N отсчетов) – не

более 4000, но при соблюдении условия M x N <= 100000. Один из признаков может быть отсчетами времени (годы, дни, секунды и т.п.); предполагается, что все отсчеты сделаны через равные промежутки. Входной массив может представлять из себя один столбец, в этом случае программа использует в качестве меток времени натуральный ряд (1, 2, ..., N).

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х признаков и 10-и объектов в текстовом файле:

```
4 10                          <- начало файла
1961 22,5 34,2 0,34
1962 23,7 33,1 0,23           массив данных:
1963 24,6 31,6 0,45           строки = последов. Отсчеты времени,
1964 20,9 30,3 0,23           столбцы = временные ряды
1965 19,4 32,4 0,78
1966 18,5 31,6 0,98
1967 17,2 30,6 0,75
1968 17,1 30,5 0,77
1969 17,9 31,2 0,82
1970 18,3 33,6 0,90
1971 18,5 33,5 0,89
1972 18,7 33,4 0,88
1973 18,6 33,4 0,86
1974 18,8 33,6 0,85
1975 20,2 33,8 0,84
Годы                          <- названия признаков
Пшеница                          (необязательно)
Ячмень
Кукуруза

Урожай зерновых               <- необязательный комментарий
```

В качестве примера формирования массива можно посмотреть файл SUN2x72.dat (2 признака, 30 объектов; 1-й признак – годы, 2-й – числа Вольфа).

Перед очередным циклом вычислений пользователь должен указать программе величину лага, исходя из некоторых либо теоретических, либо интуитивных соображений относительно возможной главной гармоники, определяющей периодические изменения значений анализируемого признака, зашумленного случайными факторами. Оптимальное значение лага пользователь может выбрать, анализируя графические представление проекций объектов ("фазовые портреты") на плоскость главных компонент (1-2, 1-3, 2-3 и т.д.) – должны формироваться повторяющиеся спиральные траектории. Временной ряд после указания лага (числа "новых признаков") преобразуется в двумерный массив следующим образом (например, ряд из 10 элементов, лаг=4):

```
z0      z0
z1      z1 z0              обрабатываемый массив  4 x 7
z2      z2 z1 z0
z3      z3 z2 z1 z0        z3 z2 z1 z0
z4  ->  z4 z3 z2 z1  ->    z4 z3 z2 z1
z5      z5 z4 z3 z2        z5 z4 z3 z2
z6      z6 z5 z4 z3        z6 z5 z4 z3
z7      z7 z6 z5 z4        z7 z6 z5 z4
z8      z8 z7 z6 z5        z8 z7 z6 z5
z9      z9 z8 z7 z6        z9 z8 z7 z6
           z9 z8 z7
              z9 z8
                 z9
```

Результаты счета могут быть отредактированы непосредственно в среде программы при выводе результатов на дисплей (заголовок, комментарии, удаление ненужной информации и т.п.).

Результатом работы программы является:

1. Матрица коэффициентов автокорреляции;

2. Собственные значения матрицы автокорреляций;

3. Все собственные векторы матрицы автокорреляций;

4. Прогнозы значения временного ряда на одну точку вперед – стандартный, по В.М.Ефимову, по суммам вкладов от главных компонент, критерии качества методов прогноза (суммы модулей отклонений и суммы квадратов отклонений).

5. Графическое представление проекций временных точек на плоскость какой-либо пары Главных компонент (например, 1-2, 1-3, ... 1-9, 2-3, ... 2-9, и т.д.). Главные компоненты в парах указываются вводом цифр от 1 до 9 – для компоненты по оси Y, и комбинациями клавиш от Alt/1 до Alt/9 – для компоненты по оси X. Обычно анализируются различные пары первых 4-х компонент, так как все последующие менее информативны. Временные точки соединяются плавной кривой (сплайном), формируя траекторию поведения исследуемого параметра в абстрактном многомерном пространстве главных компонент. Алгоритм построения сплайна разработан В.М.Ефимовым. Область прогноза следующей точки приводится в виде треугольника, метод прогноза для этой точки можно выбрать заранее.

Стандартный прогноз значения параметра на одну точку вперед выполняется методом множественного линейного авторегрессионного анализа следующим образом:

а/ формируется массив "независимых" переменных, столбец значений "зависимой" переменной, например, для вышеприведенного примера данных Y – зависимая переменная, X_1, X_2 и X_3 – независимые переменные:

Y	X1	X2	X3
z_3	z_2	z_1	z_0
z_4	z_3	z_2	z_1
z_5	z_4	z_3	z_2
z_6	z_5	z_4	z_3
z_7	z_6	z_5	z_4
z_8	z_7	z_6	z_5
z_9	z_8	z_7	z_6
Прогноз: z_{10}	z_9	z_8	z_7

б/ методом наименьших квадратов вычисляются коэффициенты $B_0..B_3$ множественной линейной авторегрессии:

$$\mathbf{Y = B_0 + B_1*X_1 + B_2*X_2 + B_3*X_3;}$$

в/ вычисляется прогнозное значение временного ряда по формуле:

$$\mathbf{Z_{10} = B_0 + B_1*Z_9 + B_2*Z_8 + B_3*Z_7;}$$

Помимо центральной прогнозной точки приводятся еще две точки для интервальной оценки погрешности прогноза. Крайние значения отличаются от центральной точки на среднеквадратическое отклонение исходного временного ряда:

$$\mathbf{Z_{10}-\sigma,\ Z_{10},\ Z_{10}+\sigma.}$$

Чтобы получить представление о характере разброса этих точек относительно прогноза следует проделать следующие действия:

г/ включать поочередно эти точки как N+1 значение временного ряда с помощью Табличного Редактора;

д/ проводить далее циклы вычислений с выводом графиков однотипных траекторий (например, 1-й и 2-й гл. компонент) на дисплей и затем на принтер;

е/ сравнить последние участки траекторий полученных графиков.

Прогноз по сумме вкладов главных компонент выполняется следующим образом:

а/ для каждой главной компоненты вычисляется прогнозное значение методом множественной авторегрессии с формированием трех "независимых" переменных (сдвигом этой компоненты на 1, 2 и 3 шага);

б/ начиная с первых двух (самых значимых) главных компонент, вычисляется последовательность прогнозов – с последующим добавлением вкладов всех остальных компонент, которые могут как улучшить, так и ухудшить прогноз из-за шума от случайных факторов;

в/ наименьшая сумма модулей отклонений или сумма квадратов отклонений (дисперсия) для последних 5-10 точек временного ряда определяет наилучший метод прогноза. Суммы отклонений по всему временному ряду (за исключением некоторого числа начальных точек) также может служить аргументом для выбора метода прогноза, но, вообще говоря, в последнем случае это всего лишь критерии качества подгонки некоторой функции к временному ряду, а не качества прогноза. Аналогичное замечание следует сделать и в отношении прогноза по авторегрессии.

Прогноз по В.М.Ефимову вычисляется следующим образом:

а/ так же, как и в предыдущем случае, вычисляются прогнозы главных компонент методом авторегрессии с формированием двух "независимых" переменных;

б/ вычисляется предварительный прогноз по сумме вкладов всех компонент, причем вклад компонент с собственными значениями, меньшими 0,0001, корректируется специальным образом;

в/ прогноз следующей точки временного ряда определяется точкой на некоторой допустимой траектории, расположенной на кратчайшем расстоянии от точки предварительного прогноза в пространстве Махаланобиса.

Решение о наиболее приемлемом значении прогноза остается за пользователем. Стандартный прогноз часто дает меньшую сумму невязок, но в большинстве случаев лучшим является прогноз по вкладам компонент, сформированный при включении 3-5 главных компонент. Число точек для вычисления сумм отклонений (тест качества метода прогноза) может быть изменено, по умолчанию используется 5 последних точек ряда. При изменении числа тестовых точек следует помнить, что время счета линейно зависит от этого значения, так как для каждой точки фактически выполняется полный цикл вычислений.

Пример рассчета прогноза числа Вольфа на 1922 год (массив SUN2x78.dat), лаг=9, сглаживание по 5 точкам:

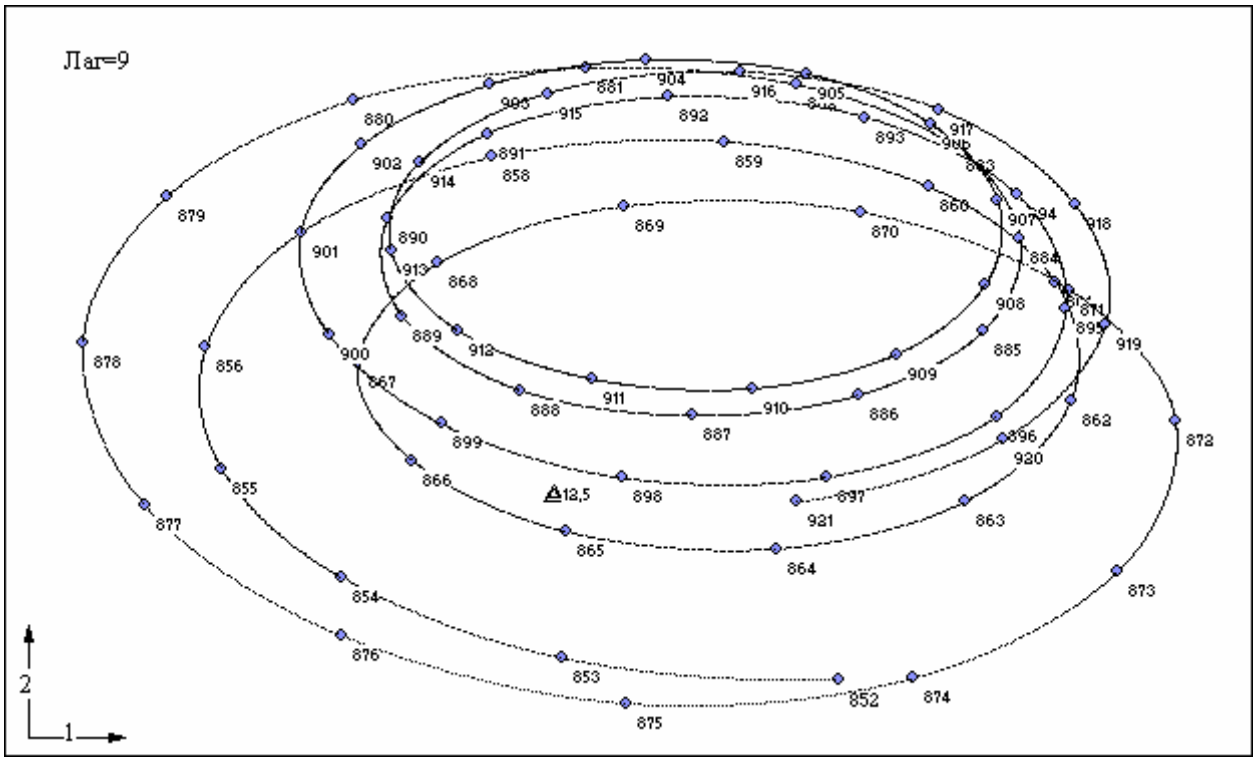
	Прогноз ряда	Сумма модулей отклонений по 5 точкам	Сумма квадр. отклонений по 5 точкам	Сумма модулей отклонений по 70 точкам	Сумма квадр. отклонений по 70 точкам
а/ Стандартный метод (множественная авторегрессия):	11,4996	58,060113	1272,3847	1237,2492	37781,331
б/ Прогноз по В.М.Ефимову:	10,3882	57,131157	1203,8482	-	-
в/ Прогноз по вкладам от главных компонент:					
No	Прогноз	Прогноз	Сумма модулей	Сумма квадр.	Сумма модулей	Сумма квадр.

гл. компоненты	ряда	отклонений по 5 точкам		отклонений по 70 точкам	
1	-0,9455	-	-	-	-
2	1,9114	28,5354	51,244227	752,32793	701,37143	11333,192
3	-1,3581	11,4404	21,989845	115,92835	237,22159	1412,9692
4	-0,0769	12,3420	17,810604	81,761561	182,27199	848,39495
5	0,0189	12,4932	18,306571	90,757207	176,66015	841,99243
6	-0,0617	12,8354	18,197657	90,318160	174,56130	832,82002
7	-0,0982	12,5227	17,809363	86,212599	173,73496	809,57967
8	0,0189	12,5703	18,002560	88,272728	174,62339	819,94784
9	-0,0204	12,5172	17,997903	88,190298	174,63362	822,28097

Наилучший прогноз – по 4-м главным компонентам, W=12,34, так как ему соответствует наименьшие суммы квадратов отклонений и модулей отклонений по пяти последним точкам (81,76 и 17,81).

При необходимости можно использовать различные виды предварительного сглаживания значений временного ряда перед вычислением матриц корреляций – собственных векторов – значений главных компонент. Обычно это несколько облегчает визуальный анализ траектории параметра в пространстве главных компонент. Поскольку главные компоненты фактически являются временными рядами, их, в свою очередь, можно анализировать точно таким же способом. Для этого массив главных компонент следует записать на диск и вновь передать программе для обработки.

Дополнительную информацию для выбора метода прогноза можно получить анализом графика временного ряда, сопровождаемого аппроксимацией ряда кубическим сплайном на точках, вычисленных различными методами прогнозирования – стандартной авторегрессией, по вкладам главных компонент:



В качестве тестового массива можно использовать файлы SUN2x78.dat – данные по динамике солнечных пятен с 1844 по 1921 г.г., с явной цикличностью 11 лет, NOVOSIB.dat – данные по средней урожайности зерновых в Новосибирской области с 1965 по 1993 год, с главным циклом 8..9 лет.

12. Специальные методы анализа

12.1. CONCORD: Согласованность экспертов по Кендаллу

Программа CONCORD предназначена для вычисления коэффициента конкордации Кендалла – характеристики согласованности (однонаправленности) нескольких признаков, измеренных в порядковой шкале (ранги, оценки от 1 до N, где N – максимальная оценка). В качестве признаков могут рассматриваться эксперты, для которых необходимо установить, согласуются ли их оценки какого-либо явления, процесса, набора экспериментальных образцов и т.д. [49]; стр. 218.

Массив данных может представлять из себя не только ранги (оценки), но и произвольные числовые характеристики типа "признаки-объекты". В этом случае программа автоматически отранжирует объекты в признаках.

На размеры массива данных из M признаков (экспертов) и N объектов (оценок) имеются ограничения: M может быть не более 100, N – не более 1000, но при соблюдении условия M x N <= 16000, то есть максимальный размер массива – 16 тысяч элементов (например, 100 x 160, 80 x 200, 20 x 800 и т.д. до 16 x 1000).

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла. Пример формирования массива из 4-х экспертов и 10-и оценок:

```
4 10
22 22 24 23
28 23 23 24
29 24 21 22
22 20 20 24
24 29 22 23
26 28 21 26
25 27 20 25
28 26 23 23
27 27 22 25
26 25 21 23

Данные 1997 г.
```

<- начало файла

массив данных:
строки = объекты (ранги, оценки),
столбцы = признаки (эксперты)

<- необязательный комментарий

В качестве примера формирования массива для программ, обрабатывающих массивы "признаки-объекты" можно посмотреть файл SSP6x30.dat (6 признаков, 30 объектов). Результат работы программы – коэффициент конкордации Кендалла – характеризует степень согласованности признаков (экспертов), изменяясь в диапазоне 0.0 – 1.0. Достоверность коэффициента проверяется по критерию Hi^2 в случае, когда число оценок N>8. При меньших значениях N следует пользоваться табличными значениями пороговых значений W-критерия, например, [15], стр. 98.

12.2. PEARSON: Анализ таблиц сопряженности

Программа PEARSON предназначена для анализа таблиц сопряженности различными методами:

1/ вычисления коэффициентов взаимной сопряженности между качественными признаками для нескольких групп объектов;

2/ энтропийного анализа таблиц частот по Кульбаку;

3/ вычисления коэффициента ассоциации таблиц 2 x 2.

Например, имеются данные по встречаемости сочетаний цвета глаз с цветом волос у человека [10, стр. 179]:

Имеется ли связь между признаками цветом глаз и цветом волос?

	Блондины	Шатены	Рыжие	Сумма
Голубоглазые	170	80	5	255
Сероглазые	70	152	8	230
Кареглазые	68	340	7	415
Сумма	308	572	20	900

Коэффициент сопряженности может быть вычислен различными способами:

– по классической формуле К.Пирсона;

– по формуле Чупрова;

– по Кендаллу, Стьюарту, Спирмену.

По умолчанию программа использует формулу Пирсона; в этом случае степень сопряженности выражается числом от 0 до 1. Пример формирования массива из 5-и групп и 6-и признаков в текстовом файле:

```
5 6                     <-  начало файла
12 22 34 35 45
 8 23 33 23 66
 9 24 31 45 89            массив данных:
 7 20 30 23 73             строки = признаки
 8 19 32 78 77             столбцы = группы, свойства
 6 18 31 98 85
Полевки                 <- названия групп (столбцов)
Сурки                      (необязательно)
Белки
Еноты
Крысы

Данные 1997 г.          <- необязательный комментарий
```

В программе диагностируется достоверность коэффициента сопряженности по критерию Hi^2, а также вычисляется вероятность ошибки в случае отклонения 0-гипотезы (о равенстве коэффициента сопряженности нулю), и на основе этой величины формируется один из вариантов текстового вывода:

P<=0,01 0-гипотеза отвергается на уровне значимости 1%,

P<=0,05 0-гипотеза отвергается на уровне значимости 5%,

P >0,10 0-гипотеза остается в силе.

12.2.1. Анализ таблиц 2 x 2

Коэффициент корреляции между двумя качественными признаками, варьирующими по двум альтернативным показателям, называется также коэффициентом ассоциации [10]. Например, данные по устойчивости сортов пшеницы к бурой ржавчине из книги [22, стр. 165]:

Сорта пшеницы	Ости имеются	Ости отсутствуют	Сумма
Иммунные	166	34	200
Неиммунные	1297	1078	2375
Сумма	1463	1112	2575

Имеется ли связь между устойчивостью к бурой ржавчине и наличием остей? Коэффициент ассоциации может быть вычислен различными способами:

– по классической формуле К.Пирсона;

– по формуле К.Пирсона, скорректированной Ф.Йетсом;

– по упрощенной формуле, приведенной в [28], стр. 213 (коэффициент Юла, Q);

– по формуле коэффициента коллигации Юла, Y

По формуле Пирсона-Йетса степень ассоциации выражается числом от 0 до 1,0, по другим методам – от –1,0 до +1,0.

В программе диагностируется достоверность коэффициента ассоциации по критерию Hi^2, а также вычисляется вероятность ошибки в случае отклонения 0-гипотезы (о равенстве коэффициента ассоциации нулю), и на основе этой величины печатается один из вариантов текстового вывода:

P<=0,01 0-гипотеза отвергается на уровне значимости 1%,

P<=0,05 0-гипотеза отвергается на уровне значимости 5%,

P >0,10 0-гипотеза остается в силе.

Вычисляется также точный критерий Фишера – вероятность получения частоты f_{11} в первой ячейке при фиксированных значениях частот в прочих ячейках.

12.3. BRASN: Однородность частот по Брандту-Снедекору

Одним из способов получения информации в различных биологических экспериментах является подсчет частот альтернативных признаков, наблюдаемых на группе однородных объектов. Это может быть число больных/здоровых растений на нескольких делянках, окрашенных/неокрашенных зерен в колосьях злака, соотношение самцы/самки в популяциях насекомых на различных полях, и т.д. Например, в опыте по сравнению действия нескольких типов гербицидов могут быть получены такие данные:

Вариант	Всего растений на делянке	Число пораженных растений	Доля пораженных растений	Число здоровых растений	Доля здоровых растений
Станд.гербицид	65	31	0,4559	34	0,5441
Гербицид «A»	59	28	0,4746	31	0,5254
Гербицид «B»	65	37	0,5692	28	0,4308
Гербицид «C»	70	50	0,7143	20	0,2857

Достоверно ли различаются гербициды по поражающей способности? Обычно в таких случаях применяется неравночисленный дисперсионный анализ, в котором повторности в вариантах представлены последовательностями из нулей и единиц, а доли (отношеня число единиц/общее число объектов) есть факторные средние, однако распределение таких данных очевидно не является нормальным (одна из предпосылок дисперсионного анализа).

Программа BRASN предназначена для проверки гипотезы однородности долей частот в вариантах (или для подтверждения различия) с помощью критерия Hi-квадрат, являющийся приближенным критерием. Если анализируются только

две выборки, может быть выполнен анализ вычислением точного значения вероятности такого сочетания частот.

Для анализа частот в программе используются:

– метод Пирсона, (Дж. Флейс, [65]) формулы на стр. 35, 151.

– точный критерий Фишера-Ирвина, вычисление вероятности для частот в таблицах 2x2;

– метод Брандта-Снедекора, изложенный в [7].

– метод В.М.Ефимова – анализ таблиц 2 x 2 с помощью критерия Стьюдента;

– методы Брандта-Снедекора и Кульбака-Лейблера для теста однородности двух рядов частот.

Доверительные интервалы для долей частот вычисляются по формулам, приведенным книге Флейса, стр. 25.

Мера степени связи между признаками в двуклеточных таблицах (фи-коэффициент) вычисляется по Флейсу, стр. 72.

Данные в виде двумерного массива "признаки-объекты" могут быть введены с клавиатуры непосредственно в среде программы, либо иными способами – через буфер Windows, из текстового файла.

В одном столбце должен быть массив частот одного из альтернативных признаков, в другом – массив общего числа обьектов (дат) по вариантам. Пример формирования массива из 6-и вариантов в текстовом файле:

```
2 6
121 154
146 231
101 248
154 207
119 192
102 186
```

<- начало файла: 2 столбца, 6 строк

массив данных:
строки = варианты,
1-й столбец – частота,
2-й столбец – общее число дат в варианте

В качестве примера формирования массива для программыы BRASN можно посмотреть файлы BRASN.dat (2 признаков, 10 объектов), Fleiss2x6.dat. Fleiss2x2.dat. В программе вычисляются критерий Hi^2 и вероятность ошибки в случае отклонения 0-гипотезы (об однородности долей частот), и на основе этой величины печатается один из вариантов текстового вывода:

P<=0,01 0-гипотеза отвергается на уровне значимости 1%,

P<=0,05 0-гипотеза отвергается на уровне значимости 5%,

P >0,10 0-гипотеза остается в силе.

12.3.1. Анализ таблиц частот 2 x M по Брандту – Снедекору

Анализ таблиц частот 2 x M по Брандту-Снедекору используется для проверки гипотезы: имеют ли доли частот в выборках неслучайные отклонения от средней доли частот, вычисленной по всем выборкам. Для этого вычисляется критерий χ-квадрат по формуле Брандта- Снедекора ([7], стр. 423):

$$\overline{\chi}^2 = \mathbf{n}_{\bullet\bullet}^2 \left((\sum_{i=1}^{m} \mathbf{x}_i^2 / \mathbf{n}_i) - \mathbf{x}^2 / \mathbf{n}_{\bullet\bullet} \right) \Big/ (\mathbf{x}(\mathbf{n}_{\bullet\bullet} - \mathbf{x}))$$

X – сумма частот наблюдаемого признака по всем выборкам, n**..** – общее число объектов по всем выборокам. Таким образом, критерием χ-квадрат проверяется нуль-гипотеза об однородности M выборок: все выборки принадлежат генеральной совокупности с соотношением частот p=x/n**..**

Для небольших таблиц (M=2..4) все ожидаемые частоты должны бытьне менее 2; если M>4, тогда все ожидаемые частоты должны быть больше 0. Если эти требования не выполняются, таблицу нужно сократить объединением мало заполненных клеток, и лишь тогда допустим расчёт статистики χ-квадрат.

12.3.2. Анализ однородности двух рядов частот по Брандту – Снедекору

Анализ однородности двух равновеликих рядов частот используется для проверки гипотезы: две независимые выборки частот относятся к одной и той же генеральной совокупности. Фактически это проверка согласия двух эмпирических распределений частот (примерно то же можно сделать с помощью критери Смирнова в программе TwoSamp). Наример, имеем следующие данные ([7], стр. 426):

Категория	Частоты 1-го ряда	Частоты 2-го ряда	Сумма
1	$\mathbf{x}_1$=60	$\mathbf{y}_1$=84	$\mathbf{n}_{1\bullet}$=108
2	$\mathbf{x}_2$=52	$\mathbf{y}_2$=50	$\mathbf{n}_{2\bullet}$=102
3	$\mathbf{x}_3$=30	$\mathbf{y}_3$=36	$\mathbf{n}_{3\bullet}$=66
4	$\mathbf{x}_4$=31	$\mathbf{y}_4$=20	$\mathbf{n}_{4\bullet}$=51
5	$\mathbf{x}_5$=10	$\mathbf{y}_5$=15	$\mathbf{n}_{5\bullet}$=25
6	$\mathbf{x}_6$=12	$\mathbf{y}_6$=10	$\mathbf{n}_{6\bullet}$=22
7	$\mathbf{x}_7$=5	$\mathbf{y}_7$=8	$\mathbf{n}_{7\bullet}$=13
Сумма	$\mathbf{n}_{\bullet 1}$=200	$\mathbf{n}_{\bullet 2}$=187	$\mathbf{n}_{\bullet\bullet}$=387

Анализ одродности двух рядов частот может быть выполнен либо по Брандту-Снедекору, либо по Кульбаку-Лейблеру. Критерий χ-квадрат по формуле Брандта-Снедекора ([7], стр. 423):

$$\bar{\chi}^2 = n_{\bullet\bullet}^2 \left(\left(\sum_{i=1}^{m} x_i^2 / n_i \right) - x^2 / n_{\bullet\bullet} \right) \Big/ \left(x(n_{\bullet\bullet} - x) \right)$$

В таблице частот не должно быть малых значений (ориентировочно <3), если такие частоты есть, их следует объединять с соседними частотами, при этом, однако, соответственно уменьшается число степеней свободы для критерия χ-квадрат. Для данной таблицы значение χ-квадрат=5.734, вероятность, вычисленная для этого значения, P=0,45358, гипотеза однородности двух рядов подтверждена.

Информационная статистика по Кульбаку-Лейблеру вычисляется по следующей формуле ([7], стр. 445):

$$2\hat{I} = \sum_{i=1}^{m} \sum_{j=1}^{2} 2n_{ij} \ln(n_{ij}) + 2n_{\bullet\bullet} \ln(n_{\bullet\bullet}) - \sum_{i=1}^{m} 2n_{i\bullet} \ln(n_{i\bullet}) - \sum_{j=1}^{2} 2n_{\bullet j} \ln(n_{\bullet j})$$

Статистика 2I при выполнении нуль-гипотезы об однородности распределена асиммптотически как χ-квадрат с m-1 степеней свободы. При не очень слабо заполненной таблице аппроксимация χ-квадрат статистики информационной статистикой вполне удовлетворительна. Если одна или несколько клеток нулевые, рекомендуется делать поправку – для каждого нуля вычитать из 2I единицу (программа делает это автоматически). Для данного массива частот 2I=5,7635, что ненамного отличается от классической статистики.

Обе статистики, по-видимому, малочувствительны к разнице в сдвиге центров распределений (средних значений), но хорошо дискриминируют по форме распределений (асимметрия, эксцесс).

Массив данных, передаваемый для анализа, должен содержать два столбца частот обоих рядов, в отличие от массивов для других методов анализа частот данной программы.

12.3.3. Анализ таблиц 2 X 2 методом Фишера-Ирвина

Для анализа частот в таблицах 2 X 2 предложено много различных методов, практически все они являются приближенными. Если маргинальные частоты малы в том смысле, одно или несколько значений менее 5, то проверять значимость

с помощью критерия χ-квадрат не рекомендуется. Метод Фишера-Ирвина позволяет вычислить точное значение вероятности появления таблицы с заданными значениями частот. Пример из книги Флейса (стр. 37):

	Признак В есть	**Признака В нет**	**Всего объ-**
Признак А есть	**a=2**	**b=3**	**5**
Признака А нет	**c=4**	**d=0**	**4**
Всего объектов	**6**	**3**	**9**

Для вычисления рассматриваются только те таблицы, в которых маргинальные частоты (суммы частот по вертикали и горизонтали) фиксированы, и имеют наблюдаемые значения, в данном случае 5 и 4 – суммы по горизонтали, 6 и 3 – суммы по вертикали. При этом условии точные вероятности появления таблиц с различными частотами a, b, c, d в клетках 2 x 2 могут быть вычислены по формуле гипергеометрического распределения вероятностей:

$$\mathbf{P(a,b,c,d) = \frac{(a+b)!\cdot(c+d)!\cdot(a+c)!\cdot(b+d)!}{(a+b+c+d)!\cdot a!\cdot b!\cdot c!\cdot d!}}$$

Точный тест Фишера-Ирвина заключается в вычислении вероятностей появления наблюдаемой таблицы (P=0,1190) и всех прочих таблиц, имеющих вероятности меньше, чем у наблюдаемой. Результирующее значение вероятности формируется как сумма этих меньших вероятностей плюс вероятность для таблицы, полученной в эксперименте.

Если сумма вероятностей не больше заданного уровня значимости (обычно 0,05), гипотеза независимости признаков отвергается, в противном случае – не отвергается – изменения частот исследуемых признаков не зависят друг от друга.

Например, для вышеприведенной таблицы возможны ещё три таблицы с теми же маргинальными частотами:

Таблицы	**Соответствующие вероятности реализации таблиц**
a=3 b=2 c=3 d=1	**0,4762**
a=4 b=1 c=2 d=2	**0,3571**
a=5 b=0 c=1 d=3	**0,0476**

Результирующая точная вероятность P=0,1190+0,0476=0,1666; гипотеза независимости признаков подтверждена.

Таким образом, для того, чтобы получить все нужные вероятности, программа перебирает все возможные таблицы от заданной, вычисляет вероятности, и суммирует меньшие значения, чем у заданной.

12.3.4. Анализ таблиц частот по Пирсону

Возможен двоякий подход к анализу частот в выборках. Обычно такие таблицы рассматривают с точки зрения анализа степени сопряжённости пары признаков (по вертикали и горизонтали таблиц), иными словами, зависит ли от изменения частоты одного признака соответствующее изменение частоты в градациях второго признака. В случае, когда второй признак имеет дихотомический характер (наличие/отсутствие = 1/0), анализ несколько упрощается, и сводится к анализу достоверности различий частот в нескольких выборках объектов. Аппарат анализа в первом случае также выполняется на базе критерия χ-квадрат (программа PEARSON пакета). Программа BRASN разработана для анализа выборок с частотами дихотомического признака. Например, имеем данные с дихотомическим признаком "Поражаемость растений":

Выборка, вариант	Всего растений на делянке	Число пораженных растений	Доля пораженных растений
Стандартный гербицид	n_1=68	31	p_1=0,4559
Гербицид "A"	n_2=59	28	p_2=0,4746
Гербицид "B"	n_3=65	37	p_3=0,5692
Гербицид "C"	n_4=70	50	p_4=0,7143

Анализ таблиц частот по Пирсону выполняется вычислением критерия χ-квадрат по формуле (дихотомический признак):

$$\chi^2 = \frac{1}{\bar{p}(1-\bar{p})}\sum_{i=1}^{m}(n_i(p_i-\bar{p})^2)$$

n_i – объем i-ой выборки, p_i – доля объектов с наблюдаемым признаком в i-ой выборке, p – средняя доля объектов в m выборках ([65], стр. 151). Число степе-

ней свободы критерия равно числу выборок без единицы. Предполагается, что последовательность выборок и частот в них не носит какой-либо характер монотонного снижения или увеличения степени воздействия на систему. Для вышеприведённой таблицы значение χ-квадрат равно 11,5, с числом степеней свободы 3, вероятность ошибки в случае отклонения 0-гипотезы P=0,0093. Гипотеза об однородности (равенстве) долей частот в исследуемом случае отклоняется на уровне значимости 1%, принимается контр-гипотеза – по крайней мере одна пара частот различается достоверно. Очевидно, что доли частот 1-го и 4-го вариантов могут считаться достоверно различными. Различие других пар частот может быть выявлено после исключения 1-го или 4-го вариантов и повторения анализа по Пирсону.

В случае таблиц 2 х 2 используется другая формула, более корректно работающая в такой ситуации (a,b,c,d – частоты признаков в ячейках, [65], стр. 23):

$$\chi^2 = \frac{n_{..}(|ad - bc| - 0{,}5n_{..})^2}{(a+b)(c+d)(a+c)(b+d)}; \; n_{..} = a + b + c + d;$$

Эта формула рекомендуется для использования в 1-м методе выбора объектов с целью получения таблицы частот 2 х 2. 1-й метод, называемый "перекрёстным отбором", состоит в том, что из некоторой совокупности выбирается n..объектов и для каждого объекта устанавливается, присутствуют или отсутствуют у него признаки. До сбора данных назначается лишь размер выборки n.. Целью исследования, основанного на использовании метода выбора 1, является выяснение, зависимы признаки или нет.

В случае числа вариантов >2, если монотонный характер изменения фактора имеет место, возможно уточнение метода обработки данных ([65], стр.159).

12.3.5. Анализ частот дихотомических признаков по В.М.Ефимову

В.М.Ефимов (ИЦИГ СО РАН) считает допустимым анализировать различие долей в двух выборках с помощью Т-критерия Стьюдента.. Его подход изложен в [66] в подразделе "Почему бинарные признаки можно обрабатывать так же, как количественные". Т-критерий Стьюдента обычно используется для анализа различия двух средних выборок из генеральных совокупностей с непрерывным законом распределения вероятностей.

Основной аргумент – дискретное биномиальное распределение вероятностей может быть аппроксимировано нормальным распределением с параметрами

среднее=p, σ=p*(1-p)/N, p – доля объектов с наблюдаемым признаком, N – размер выборки. Приближение применимо при N*p*(1-p)>10 (Корн, Корн, 1970). Поэтому грубое сравнение двух средних для бинарных признаков (долей частот в двух выборках) можно проводить, как и для количественных признаков, с помощью обычного Т-критерия. Того же мнения придерживается Г.Н.Зайцев [22].

Формула вычисления Т-критерия ([22], стр. 257):

$$\mathbf{T = |p_1 - p_2| / \sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)};}$$

Метод применяется в тех случаях, когда выборки достаточно велики по объему и принадлежат к разным совокупностям. Приближенный критерий применимости этого метода $p_i*n_i>4$ для обеих выборок; число степеней свободы для критерия Стьюдента $=(n_1+n_2-2)$.

В случае малых (и больших) выборок Зайцев рекомендует сравнивать доли после преобразования Фишера:

$$\mathbf{\varphi = 2 \cdot ArcSin(\sqrt{p});}$$

Сравнение преобразованных долей выполняется при помощи критерия Фишера-Снедекора, вычисляемого по формуле (Зайцев, стр. 258):

$$\mathbf{F = (\varphi_1 - \varphi_2)^2 n_1 n_2 / (n_1 + n_2);}$$

Число степеней свободы для критерия 1 и (n_1+n_2-2). Для Т-критерия и F-критерия вычисляются вероятности ошибки в случае отклонения 0-гипотезы, и на основании их значений делаются выводы об отклонении/принятии 0-гипотезы.

12.3.6. Анализ таблиц 2 x 2, полученных 2 и 3 методами отбора

Второй метод выбора, называемый "целевым отбором", состоит в том, что для анализа отбирается заранее установленное число объектов n_1, которые вероятно будут иметь некоторый признак и заранее установленное число n_2 объектов, которые возможно не будут иметь наблюдаемого признака.

	Признак есть	**Признака нет**	**Всего объектов**
Выборка-1	**a=15**	**b=135**	**$n_{1.}$=150**
Выборка-2	**c=10**	**d=40**	**$n_{2.}$=50**
Всего объектов	**25**	**175**	**$n_{..}$=200**

Статистическая значимость различия долей $p_1=a/n_1$ и $p_2=c/n_2$ оценивается следующей формулой:

$$Z = \frac{|p_2 - p_1| - 0{,}5(1/n_{1\bullet} + 1/n_{2\bullet})}{\sqrt{\bar{p}(1-\bar{p})(1/n_{1\bullet} + 1/n_{2\bullet})}};\ \bar{p} = (a+c)/(n_{1\bullet} + n_{2\bullet});$$

В этом случае в качестве критерия выступает квантиль нормального рапределения. Если вероятность, вычисленная по эмпирическому значению Z, меньше 0,05. делается вывод, что доли неравны.

Третий метод выбора данных также предполагает формирование выборок заранее определённого размера. Однако в этом случае требуется, чтобы значения выборк формировались случайно. Например, этот метод лежит в основе контроля эффективности способов клинического лечения: из общего числа больных n_1 пациентов отбирают случайно и лечат обычным способом, остальных n_2 больных лечат тем способом, эффективность которого исследуется.

Значимость различия пропорций $p_1=a/n_1$ и $p_2=c/n_2$ оценивается также Z-критерием, однако последующий анализ отличается от методов, используемых в случае данных, полученных во 2-м методе отбора ([65], стр. 35).

12.3.7. Мера степени связи дихотомических признаков – φ-коэффициент

Для оценки степени связи между парой дихотомических признаков в таблицах 2x2 используется φ-коэффициент ([65], стр. 72).

$$\varphi = \sqrt{\chi_n^2 / N};$$

Значение χ-квадрат вычисляется по стандартной формуле Пирсона, N - общее число объектов в обеих выборках

Его можно интерпретировать как коэффициент парной корреляции. В отличие от простых критериев связанности дихотомических признаков, его значение не зависит от размера выборок, но только от различия частот по диагонали двуклеточной таблицы. Значение φ, близкое к нулю, означает слабую связь или отсутствие связи. Если значение φ близко к 1, то связьсильная. Максимальное значение φ - единица. В качестве значений, соответствующих слабой связи, можно ориентировочно указать величины менее 0,3..0,35

φ-коэффициент используется преимущественно при анализе ответов в психологических тестах, .в педагогическом анализе, в факторном анализе дихотомических признаков.

φ-коэффициент не рекомендуется использовать в тех исследованиях, где важна сравнимость результатов для различных методов проведения исследований, так как имеются критические замечания математиков о плохой воспроизводимости значений коэффициента, получаемых в разных случаях.

12.3.8. Преобразование массива частот в массив для дисперсионного анализа

Помимо стандартного метода анализа сходства/различия частот критерием χ-квадрат существует возможность анализа таких данных методом классического неравночисленного дисперсионного анализа по Фишеру. Для этого массив частот должен быть преобразован следующим образом:

1/ число вариантов для дисперсионного анализа = числу выборок в анализе по χ-квадрат;

2/ общее число повторений = размеру выборки с максимальным числом объектов;

3/ в каждом варианте своё число число повторений, равное соответствующему размеру выборки;

4/ в варианте методом Монте-Карло формируется последовательность из нулей и единиц, причём число единиц равно частоте наблюдаемого признака в этой выборке.

Например, имеется массив частот:

Выборка	Всего объектов	Частота признака	Доля объектов
1	10	5	0,5
2	12	6	0.5
3	15	7	0,466667
4	11	8	0,727273

Формируется следующий массив для неравночисленного дисперсионного анализа (–999 = признак отсутствующего значния):

Вариант	Повторения	Средние вариантов
1	1 0 1 1 1 0 1 0 0 0 –999 –999 –999 –999 –999	0.5
2	0 1 0 1 0 0 1 1 0 0 1 1 –999 –999 –999	0.5
3	0 1 1 0 0 0 1 0 0 1 0 1 1 0 1	0,466667
4	1 0 1 0 1 1 0 1 1 1 1 –999 –999 –999 –999	0,727273

В дисперсионном анализе наличие/отсутствие различия средних выявляется критерием Фишера-Снедекора. Очевидно, что в этом случае не выполняется одна

из предпосылок классического дисперсионного анализа – нормальность распределения данных в выборках. Поэтому дисперсионный анализ может быть лишь вспомогательным инструментом анализа частот, особенно полезным в случае малых выборок, когда критерий χ-квадрат малочувствителен.

Идея такого анализа частот базируется на подходе В.М.Ефимова (ИЦИГ), изложенного в [66] в подразделе "Почему бинарные признаки можно обрабатывать так же, как количественные".

Для формирования такого массива "варианты/повторения" следует выбрать пункт "Записать в виде массива для Дисп. Анализа" в Меню "Массив данных".

12.4. SERIES: Анализ серий по Вальду-Волфовицу

Программа SERIES предназначена для проверки предположения об отсутствии какой-либо закономерности в последовательности элементов выборки из генеральной совокупности. Это может быть ряд нечисловой природы, например последовательность символов:

AABBBABBAAAABBBABABABBAABBBBBBAAABABBABA

Такая последовательность может быть перекодирована эквивалентным числовым рядом:

00111011000011101010110011111100010110 10

используемым для совместимости со структурой массивов данных пакета SNEDECOR. Серия – группа подряд идущих одинаковых элементов, группа может состоять из одного элемента, в этой выборке – 21 серия.

Обычные числовые выборки также могут быть проанализированы на отсутствие закономерностей в последовательности их элементов. Например, имеется выборка, полученная с помощью датчика псевдослучайных чисел со средним, равным нулю:

1,12 0,55 -0,23 -2,34 -1,48 0,02 0,93 -1,06 -0,23 0,46 0,89 1,62 -0,51

В этой выборке серии положительных и отрицательных значений могут быть представлены в следующем виде:

+1, +1, -1, -1, -1, +1, +1, -1, -1, -1, +1, +1, +1, -1 <= 6 серий

Закономерностью в данном контексте является наличие либо “слишком” частого чередования знака элементов (большое число серий), либо “слишком” длинных последовательностей одного знака (малое число серий). Перекодировка

произвольной числовой последовательности в дихотомический ряд может быть сделана относительно выборочного среднего или медианы.

Анализ серий может быть использован для проверки отсутствия дрейфа в показаниях измерительных приборов, качества уравнений регрессии (остатки должны быть случайной последовательностью), доказательства стационарности временных рядов, проверки качества генерации псевдослучайных чисел и т.д.

Для анализа серий в последовательности используется метод, изложенный в [59], стр. 71-73.

В качестве примера формирования массива для программ, обрабатывающих массивы "признаки-объекты" можно посмотреть файл SSP6x30.dat (6 признаков, 30 объектов).

Для анализа выборки используется преобразование в знаковый ряд либо относительно выборочного среднего, либо относительно медианы.

Еще один метод формирования знакового ряда определяет число серий “вверх” и “вниз” следующим образом:

$Z_i=1$ если $X_i>X_{i-1}$ $Z_i=-1$ если $X_i<X_{i-1}$

В программе вычисляется число серий (R-критерий) в знаковых последовательностях, и используется табличный W-критерий Вальда-Вольфовица на уровнях значимости 1%, 5% и 10% для выборок небольшого размера, и аппроксимация критерия R нормальным распределением для выборок с числом элементов одинакового знака >20.

Проверяется 0-гипотеза: последовательность элементов случайна.

Контр-гипотеза, двусторонний критерий: последовательность элементов неслучайна (либо слишком много серий, либо слишком мало);

Контр-гипотеза, односторонний критерий: последовательность элементов неслучайна, слишком много серий, $R>=W_{up}$;

Контр-гипотеза, односторонний критерий: последовательность элементов неслучайна, слишком мало серий, $R<=W_{down}$;

W_{up} и W_{down} - пара табличных значений критерия Вальда-Вольфовица на выбранном уровне значимости; для проверки двусторонней гипотезы используется половинный уровень значимости (5% для уровня 10%, 2% - для уровня 5%).

С помощью программы могут быть обработаны нечисловые данные дихотомической природы, которые могут быть представлены последовательностью символов (например, A-B), знаков (+ -) и т.п. При формировании массивов при вводе с клавиатуры следует заменять символы на последовательность 0 и 1.

В качестве теста использовались примеры из [58], массив RUNION1.dat, [15], [21].

12.5. PROBIT – анализ данных “доза – доля объектов”

Программа PROBIT предназначена для обработки экспериментальных данных, полученных в исследованиях с возрастающими уровнями некоторого фактора, обычно называемого "Доза". Это может быть лекарство, испытываемое на нескольких группах животных, ядохимикат в сельскохозяйственных опытах, силовое воздействие на детали машин в инженерных экспериментах и т.п. Отклик – доля объектов, которые в результате воздействия дозы будут, например, вылечены, или просто отреагируют определённым образом на действие фактора.

Обычно ставится задача по оценке параметра, называемого ED50 (Effective Dose) – определения достоверного значения дозы, при котором 50% животных будет вылечено, или LD50 (Letal Dose) – 50% популяции насекомых погибнет, или 50% конструкций будет разрушено. Помимо LD50, обычно оценивают и другие уровни дозы, например, LD10, LD95 и т.д.

Основная предпосылка, используемая в классическом пробит-анализе, заключается в предположении, что зависимость "доза фактора – доля объектов" может быть аппроксимирована функцией "интеграл нормального распределения", то есть доли – суть вероятности некоторого генерального распределения Гаусса с параметрами (среднее, сигма), подлежащими оценке.

Как справедливо заметил В.Ю.Урбах ([67], стр. 76), "Нормальность кривой смертности не имеет теоретического обоснования и является просто эмпирическим фактом, который должен для каждого объекта устанавливаться экспериментально. Если при изучении какого-нибудь нового объекта (вида животных, микроорганизмов и т.д.) окажется, что точки (x, y) явно не ложатся на прямую (т.е. имеет место систематическое уклонение), то LD50 нельзя найти с помощью пробит-метода. В этом случае следует попытаться подобрать какое-нибудь другое преобразование.".

В соответствии с этим замечанием программа предлагает четыре метода оценки параметров функции действия фактора:

1/ аппроксимация зависимости "доза фактора – доля объектов" с помощью функции "интеграл нормального распределения";

2/ аппроксимация зависимости "доза фактора – доля объектов" с помощью логистической функции;

3/ прямая оценка параметров функции "доза фактора – доля объектов" с помощью приближения зависимости полиномами малых степеней методом наименьших квадратов;

4/ оценка параметров .функции "доза фактора – доля объектов" полиномами малых степеней методом наименьших квадратов после пробит-преобразования долей в нормализованные отклонения; это несколько отличается от стандартного метода Финни [75].

Данные должны быть представлены тройками экспериментальной зависимости "общее число объектов – число погибших объектов – доза" по возрастанию дозы. Минимальное число уровней воздействия – 3.

Пример формирования массива из 3-х переменных, 5-и троек значений в текстовом файле:

```
3 5              <-- первая строка массива данных
50  6  2,6           3 столбца = переменных
48 16  3,8           5 строк = вариантов
46 24  5,1
49 42  7,7
50 44 10,2
Total            <-- наименования переменных
Fatal                (необязательно)
Doza

Finney test      <-- комментарий (необязательно)
```

В качестве примера формирования массива для программы PROBIT можно посмотреть файлы FINNEY.dat , SPY3x10.dat, BLISS.dat, DIAZINON.dat.

Программа предлагает пользователю выбрать один из четырёх методов обработки данных, обычно рекомендуют выбрать метод, дающий максимальное значение критерия Фишера-Снедекора, или минимальный доверительный интервал в области LD40 – LD60, однако интуиция исследователя должна подсказать наилучший метод – с помощью визуального анализа графиков зависимостей "доля – доза".

Для проверки достоверности полученных уравнений регрессии вычисляются критерии Фишера-Снедекора. Нуль-гипотеза формулируется следующим образом: отсутствует регрессионая связь между переменными, значения коэффициентов регрессии отличаются от нуля только вследствие действия случайных факторов.

Для F-критерия печатается "вероятность ошибки в случае отклонения нульгипотезы". Если

P <= 0.01 – уравнение регрессии значимо на уровне 1%,

P <= 0.05 – уравнение регрессии значимо на уровне 5%,

P > 0.10 – регрессионная связь не доказана.

В результатах анализа формируются:

1. Таблица исходных данных с анализом невязок отклика (долей) с аппроксимирующей функцией.

2. Стандартные таблицы дисперсионного анализа

3. Таблица LD10 .. LD90, или LD05 .. LD95, или же LD1 .. LD99 по выбору пользователя.

4. Дополнительно может быть сделан тест нормальности распределения значений долей – главной предпосылки классического пробит-анализа.

12.5.1. Некоторые замечания по методам анализа, структуре данных

Как было ранее замечено, предположение о нормальности распределения долей – не более чем эмпирическое правило. Исследователь не должен слепо следовать традициям классического пробит-анализа, так как неклассические методы могут дать более корректные оценки LD50 или других "летальных доз", более узкие доверительные интервалы. Главным критерием "качества" результатов следует считать оптимальное расположение графика регрессии поверх экспериментальных точек. Конечно, это довольно субъективный критерий, однако в дополнение к субъективному мнению следует присоединять значение критерия Фишера-Снедекора, а также минимальный "коридор", формируемый доверительными интервалами. Дополнительно может быть сделан анализ остатков, на графике остатков не должно быть явной зависимости в последовательности точек, отклонения от средней линии должны быть случайными.

Таким образом, обработку данных следует сделать несколько раз, меняя метод, меняя способ преобразования шкалы доз, каждый раз просматривая график зависимости. Только после этого следует выбрать наилучший способ анализа.

Структура массива данных для пробит-анализа должна соответствовать некоторым правилам, имеются также очевидные ограничения.

Минимальное число вариантов – три, причем все пары "дозы – доли" должны быть разными. При большем числе вариантов допустимы пары одинаковых

значений, хотя это не имеет особого смысла, лучше объединить эти варианты в один.

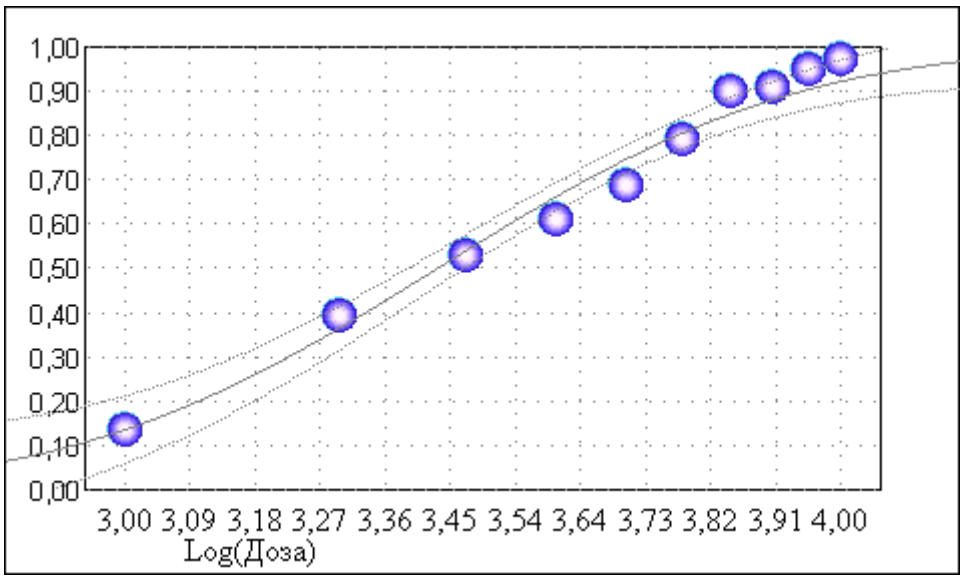
Первые два столбца (всего объектов в выборке, количество объектов, отреагировавших на воздействие) должны быть целые числа, третий столбец – "доза" – может быть и целого, и вещественного типа. Доза, вообще говоря, не может быть нулевой (возможный контрольный вариант), так как логарифм нуля не существует. Логарифмическое преобразование значений дозы рекомендовано во всех руководствах по пробит-анализу, хотя в большинстве случаев обработка может быть выполнена и без логарифмирования. Если всё-таки требуется логарифмирование шкалы доз, следует исключить строку с нулевой дозой комбинацией клавиш <Ctrl/Y>.

Если количество объектов во втором столбце нулевое (нет объектов, отреагировавших на минимальный уровень дозы), первый, второй и четвёртый методы, вообще говоря, неприменимы из-за проблем с математикой (деление на ноль, бесконечно большое значение квантиля и т.п.). В этом случае программа подставляет вместо нуля значение доли 0,001, что соответствует одному "отреагировавшему" объекту из тысячи. Аналогичная проблема возникает, когда все объекты "погибли". Программа подставляет значение доли в этом случае 0,999. Эти значения, как можно предположить, вполне могут отражать реальность при больших размерок выборок, позволяют применить эффективные методы пробит-анализа с минимальными искажениями результатов.

12.5.2. Аппроксимация интегралом нормального распределения

В этом методе предполагается, что S-образная зависимость "доза – доля выживших" может приближенно считаться функцией "интеграл нормального распределения вероятностей", параметры которого (среднее, среднеквадратическое отклонение) следует оценить каким-либо образом. Тогда LD50 и прочие дозы могут быть вычислены как квантили нормального распределения.

Известно, что оценки среднего (Sr) и сигмы (Sg) малых выборок (5..20 значений) стандартным методом могут быть некорректными по ряду причин. Для более точной оценки параметров распределения долей автором пакета разработан итерационный алгоритм минимизации суммы квадратов отклонений эмпирического распределения данных от функции "интеграл нормального распределения".



Довольно быстро процесс минимизации суммы квадратов сходится, в качестве стартовых значений используются оценки среднего и сигмы по обычным формулам. После этого по выборочным значениям доз вычисляются вероятности (в качестве долей) и невязки для оценки качества приближения эмпирической зависимости функцией "интеграл нормального распределения", в виде Y=Gauss((X – Sr)/Sg).

Далее формируется функция, обратная к функции "интеграл нормального распределения", которая по значению доли (в качестве вероятности) вычисляет квантиль нормального распределения и преобразует его в значение дозы умножением на среднеквадратическое отклонение (Sg) и прибавлением среднего (Sr). Для этой обратной функции Sr и Sg являются параметрами нелинейной регрессии и оптимизируются минимизацией суммы квадратов отклонений методом Ньютона-Рафсона.

Для объективной оценки качества аппроксимации зависимостей приводятся таблицы стандартного дисперсионного анализа функций регрессии Y=Gauss((X – Sr)/Sg) и X=Sr+Xnorm(Y)*Sg. Максимум значения критерия Фишера-Снедекора из второй таблицы может быть критерием выбора того или иного метода пробит-анализа. В таблице "Летальные дозы", вычисляемой на основе обратной функции, определяют значение LD50 и доверительный интервал (на стандартном уровне доверия 95%), минимальный размах которого также может быть критерием выбора метода пробит-анализа.

Для проверки предпосылки нормальности распределения долей может быть выполнен тест критериями Колмогорова-Смирнова, Уилка-Шапиро, Мизеса-Смирнова и Джири. Плюсы в графе "Выводы" говорят о нормальности данных,

Минусы – о значительных отклонениях от нормального закона распределения вероятностей.

12.5.3. Аппроксимация зависимости логистической функцией

S-образная зависимость "доза – доля выживших" может быть эффективно приближена логистической функцией. Формула логистической функции для таких данных:

$$\mathbf{Y = 1.0/(1.0 - Exp(A + B \cdot X)}$$

Параметры A и B логистической функции должны быть определены некоторым образом. Предварительная оценка параметров логистической функции выполняется методом наименьших квадратов, затем выполняется итерационная процедура минимизации суммы квадратов отклонений от экспериментальной зависимости методом Ньютона-Рафсона, в результате которой получаются оптимизированные значения параметров функции. По выборочным значениям доз вычисляются доли и невязки для оценки качества приближения эмпирической зависимости логистической функцией.

Затем формируется нелинейная функция регрессии, обратная к логистической, её параметры A и B также оптимизируются минимизацией суммы квадратов отклонений методом Ньютона-Рафсона:

$$\mathbf{X = (Ln(1.0 - 1.0/Y) - A)/B}$$

Тогда LD50 и прочие дозы могут быть вычислены по этой обратной функции и рассчитаны доверительные интервалы для доз с помощью оценки ошибки регрессии из таблицы дисперсионного анализа.

Для объективной оценки качества аппроксимации зависимостей приводятся таблицы стандартного дисперсионного анализа функций регрессий. Максимум значения критерия Фишера-Снедекора из второй таблицы может быть критерием выбора того или иного метода пробит-анализа. В таблице "Летальные дозы" определяют значение LD50 и доверительный интервал (на стандартном уровне доверия 95%), минимальный размах которого также может быть критерием выбора метода пробит-анализа.

Доверительный интервал для LD определяется по формуле ([4], стр. 172):

$$\mathbf{Y = y_x \pm T_{(1-\alpha/2,n-2)} \sqrt{S \cdot (1/n + d' A^{-1} d)}}$$

T – критерий Стьюдента, S – остаточный средний квадрат из таблицы дисперсионного анализа, A – матрица Грамма, d – отклонение X от среднего значения.

12.5.4. Прямая оценка параметров зависимости Методом НК

В большннстве случаев возможна прямая оценка параметров функции "доза фактора – доля объектов" с помощью приближения зависимости полиномами малых степеней методом наименьших квадратов. Качество аппроксимации можно легко оценить по графикам прямой и обратной зависимости, дополнительной возможностью улучшения качества подгонки является выбор логарифмирования значений доз или отказ от логарифмирования.

Аналогично другим методам пробит-анализа, по уравнению регрессии вычисляются значения долей и невязки с значениями, полученными в эксперименте. В этом методе легко оцениваются обе зависимости – прямая и обратная, приводятся два уравнения регрессии и две таблицы дисперсионного анализа. На основании обратной зависимости вычисляется таблица "Летальных доз" и доверительных интервалов.

Максимум значения критерия Фишера-Снедекора из второй таблицы может быть критерием выбора того или иного метода пробит-анализа. В таблице "Летальные дозы" определяют значение LD50 и доверительный интервал для него (на стандартном уровне доверия 95%), минимальный размах которого также может быть критерием выбора метода пробит-анализа.

В случае явной нелинейности, определяемой визуально на графике обратной зависимости, следует увеличить степень полинома с помощью клика ПРАВОЙ клавиши мышки в поле текста с результатами анализа, появится Меню дополнительных операций, в котором можно выбрать увеличение (или снижение) степени полинома.

Некоторым недостатком метода прямой оценки параметров является увеличение погрешности оценок на краях диапазона изменения доз.

12.5.5. Пробит-преобразование долей в нормализованные отклонения

Оценка параметров .функции "доза фактора – доля объектов" может быть выполнена полиномами малых степеней методом наименьших квадратов после пробит-преобразования долей в нормализованные отклонения распределения Гаусса; это несколько отличается от классического метода Финни [75].

Предполагается, что такое преобразование долей превращает S-образную зависимость в практически линейную, поэтому после этого зависимость легко приближается простой линейной регрессией. Для "хороших" данных это действительно линейная регрессия, однако в действительности чаще наблюдается некоторая нелинейность, которую можно скорректировать, увеличивая степень полинома до квадрата или куба. Более высокие степени нерациональны, как правило, достоверность коэффициентов больших степеней полинома быстро падает, уменьшается значение критерия Фишера-Снедекора

Аналогично третьему методу анализа, вычисляются две полиномиальные регрессии – прямая и обратная, и две таблицы разложения дисперсий. На основании обратной зависимости вычисляется таблица "Летальных доз" и доверительных интервалов.

13. О Джордже Снедекоре

Биография Джорджа В. Снедекора (1882 -1974)

George Waddell Snedecor родился в Мемфисе, штат Теннеси. Он изучал математику и физику в университетах Алабамы и Мичигана и стал профессором математики университета штата Айова. Работая там, он впервые ввёл курс статистики и начал знаменитое сотрудничество с Генри А. Вэлласом, ставшим редактором *Wallace's Farmer* в Des Moines, и позднее вице-президентом Соединённых Штатов. Они весьма интересовались сельскохозяйственными исследованиями, организовали семинар по изучению множественной регрессии, и делали пионерские работы по использованию перфокарт и перфокартных (счетных) машин. Совместно, они опубликовали в 1925 “Корреляции и машинные вычисления”, основали в 1927 Математико-статистическую Службу, и в 1933 ныне широко известную Статистическую лабораторию штата Айова. В 1931 Снедекор был приглашен сэром Рональдом А. Фишером в Эймс, и эта встреча стимулировала развитие исследова-

ний, привела к разработке множества методов, которые были обсуждены в этой сессии. Среди очевидных последствий этого события было основание в штате Айова первого в Соединённых Штатах Департамента статистики, председательство Снедекора в Американской Статистической ассоциации, и публикация Снедекором двух известных работ: “*Вычисление и интерпретация анализа дисперсий и ковариаций”* (1934) и “Статистические методы” (1937). Последняя работа, позднее в соавторстве с Вильямом Дж. Кочраном, выдержала семь изданий при жизни обоих авторов, было продано более 125000 экземпляров.

Источник: 0. Kempthorne, "George W. Snedecor," *International Statistical Review*, 42, 1974, pp. 319-321

Библиография

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1983. – 471 с.

2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: исследование зависимостей. – М.: Финансы и Статистика, 1985. – 487 с.

3. Айвазян С.А., Бухштабер В.М. и др. Прикладная статистика: классификация и снижение размерностей. – М.: Финансы и Статистика, 1989. – 607 с.

4. Афифи А., Эйзен С. Статистический анализ: подход с использованием ЭВМ. – М.: Мир, 1982. – 488 с.

5. Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983. – 518 с.

6. Лисенков А.Н. Математические методы планирования многофакторных медико-биологических экспериментов. – М.: Медицина, 1979. 344 с.

7. Закс Л. Статистическое оценивание. – М.: Статистика 1976. 598 с.

8. Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980. – 512 с.

9. Зедгинидзе И.Г. Планирование эксперимента для исследования многокомпонентных систем. – М.: Наука, 1976. – 390 с.

10. Лакин Г.Ф. Биометрия. – М.: Высш. школа, 1980. – 293 с.

11. Доспехов Б.А. Методика полевого опыта. – М.: Колос, 1979. – 416 с.

12. Хан Г., Шапиро С. Статистические модели в инженерных задачах. – М.: Мир, 1969. – 401 с.

13. Кульбак С. Теория информации и статистика. – М.: Наука, 1967. – 408 с.

14. Иванова В.М., Калинина В.Н. и др. Математическая статистика. Изд. 2-е. – М.: Высшая школа, 1981. – 371 с.

15. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. – 416 с.

16. Сильвестров Д.С., Семенов Н.А., Марищук В.В. Пакеты прикладных программ статистического анализа. – Киев: Тэхника, 1990. – 176 с.

17. Семенов Н.А. Программы регрессионного анализа и прогнозирования временных рядов. Пакеты ПАРИС и МАВР. – М.: Финансы и статистика, 1990. – 111 с.

18. Девис Дж. Статистика и анализ геологических данных. – М.: Мир, 1977. – 572 с.

19. Статистические методы для ЭВМ. /Под ред. Энслейна К., Рэлстоуна Э., Уилфа Г. – М.: Наука, 1986. – 464 с.

20. Снедекор Дж.У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. – М.: Сельхозиздат, 1961. – 503 с.

21. Ликеш И., Ляга Й. Основные таблицы математической статистики. – М.: Финансы и Статистика, 1985. – 356 с.

22. Зайцев Г.Н. Математическая статистика в экспериментальной ботанике. – М.: Наука, 1984. – 424 с.

23. Справочник по прикладной статистике. В 2-х томах. /Под ред. Э.Ллойда, У.Ледермана, Ю.Тюрина. – М., Финансы и Статистика, 1989.

24. Факторный, дискриминантный и кластерный анализ. Пер. с англ. /Д.Ким, Ч.Мьюллер, У.Клекка и др. – М.: Финансы и Статистика, 1989. – 215 с.

25. Шураков В.В., Дайитбегов Д.М. и др. Автоматизированное рабочее место для статистической обработки данных. – М.: Финансы и Статистика, 1990. – 190 с.

26. Сборник научных программ на Фортране. Вып. 1, 2. – М.: Статистика, 1974. – 316 с.,

27. Митропольский А.К. Техника статистических вычислений. М.: – Наука, 1971. – 576 с.

28. Статистический словарь. /Гл. ред. М.А.Королев. – М.: Финансы и Статистика, 1989. – 623 с.

29. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977. – 408 с.

30. Фишер Р. Статистические методы для исследователей. – М.: Госстатиздат, 1958. – 268 с.

31. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. – М.: Финансы и статистика, 1984. – 230 с.

32. Рао С.Р. Линейные статистические методы и их применения. – М.: Наука, 1968. – 548 с.

33. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. – М.: Мир, 1981. – 520 с.

34. Хикс Ч. Основные принципы планирования эксперимента. – М.: Мир, 1967. – 406 с.

35. Ефимов В.М., Галактионов Ю.К., Шушпанова Н.Ф. Анализ и прогноз временных рядов методом главных компонент. Новосибирск, 1988; изд. Наука, сиб. отделение, 70 с.

36. Томилов В.П. О статистической обработке многолетних данных полевых опытов. Земледелие, 1987. – Т. 3 – с. 48-51.

37. Южаков А.И., Сорокин О.Д. Пакет программ прикладной статистики «Snedecor V4» для обработки данных, полученных в биологических экспериментах// Информационные технологии, Информационные измерительные системы и приборы в исследованиях сельскохозяйственных процессов. Материалы региональной научно-практической конференции (Новосибирск, 2000 г.). с. 323-324.

38. Сорокин О.Д. Пакет прикладных программ СНЕДЕКОР. В сборнике: "Применение математических методов и ЭВМ в почвоведении, агрохимии и земледелии". Тезисы докладов 3-й научной конференции Российского общества почвоведов. Барнаул, 1992. – 97 с.

39. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. М.: Финансы и статистика, 1982, 545 с.

40. Литтл Р.Дж., Рубин Д.Б. Статистический анализ данных с пропусками. – М.: Финансы и статистика, 1991. –336 с.

41. Greenhouse S.W., Geisser S. On methods in the analysis of profile data. Psychometrika, 1959, vol. 24, no. 2, p. 95-112

42. Мардиа К., Земроч П. Таблицы F-распределения. М.: Наука, 1984 – 255 с.

43. Андерсон Т. Введение в многомерный статистический анализ. – М. Физматгиз, 1963. – 500 с.

44. Линник Ю.В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. – М.: Физматгиз, 1958. – 333 с.

45. Поллард Дж. Справочник по вычислительным методам статистики. М.: Фининсы и статистика, 1982, – 344 с.

46. Мудров А.Е. Численные методы для ПЭВМ на языках Бейсик, Фортран и Паскаль";, 1992, Томск.

47. Хартман Г. Современный факторный анализ. М.: Статистика, 1972.

48. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. М.: – Наука, 1976. – 736 с.

49. Хеттманспергер Т. Статистические выводы, основанные на рангах. – М.: Финансы и статистика, 1987. – 334 с.

50. Wichman B.A., Hill I.D. Algorithm 183. Apll. Statistics. 1982 V.31, NO.2, p. 188.

51. Уилкс С. Математическая статистика. М., 1967. – 632 с.

52. Худсон Д. Статистика для физиков. М. Мир, 1967. – 242 с.

53. Пагурова В.И. Критерий сравнения средних значений по двум нормальным выборкам. – М.: ВЦ АН СССР. 1968. – 58 с.

54. Химмельблау Д. Анализ процессов статистическими методами.

55. Благовещенский Ю.Н., Самсонова В.П., Дмитриев Е.А. Непараметрические методы в почвенных исследованиях.М. Наука, 1987, 96 с. Справочник по прикладной статистике.

56. Lawson, Hanson; Solving least square problems. Prentice Hall, 1974.

57. Эберт К., Эдерер Х. Компьютеры. Применение в химии. М.: Мир, 1988. - 416 с. (K.Ebert, H.Ederer; Computeranwendungen in der Chemie. 1985).

58. Рунион Р.. Справочник по непараметрической статистике. М.: Финансы и Статистика, 1982. – 1?? с.

59. Орлов А.И. Непараметрическое точечное и интервальное оценивание характеристик распределения. Заводская лаборатория. Диагностика материалов, №5, 2004, стр.65-70.

60. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М.: Инфра-М, 2003. – 544 с.

61. Васильев А.Г., Фадеев В.И., Галактионов Ю.К. и др. Реализация морфологического разнообразия в природных популяциях млекопитающих. 2-е изд. – Новосибирск: Издательство СО РАН, 2004. – 231 с.

62. А.Я.Жежер, А.М.Криков, А.Н.Власенко, О.Д.Сорокин. Стратегия и тактика исследований в земледелии на основе планирования эксперимента. Метод. рекомендации / РАСХН. Сиб. отд-ние. – Новосибирск, 1999. – 110 с.

63. А.И.Южаков. Многомерное ранжирование. Метод. рекомендации. РАСХН. Сиб. отд-ние. Краснообск. РПО, 2000. – 50 с.

64. С.П.Мартынов, О.Д.Сорокин. Пакет программ прикладной статистики «BIOGEN» для обработки данных, полученных в селекционно-генетических экспериментах// В сб. Информационные технологии, информационные измерительные системы и приборы в исследовании сельскохозяйственных процессов. Ч.1. АГРОИНФО-2003. Новосибирск, 2003. – 380 с.

65. Дж. Флейс, Статистические методы для изучения таблиц долей и пропорций. М., "Финансы и статистика", 1989, – 319 с.

66. В.М.Ефимов, В.Ю.Ковалева. Многомерный анализ биологических данных. Горно-Алтайск, РИО, 2007 – 75 с.

67. В.Ю.Урбах. Математическая статистика для биологов и медиков. М., Изд. АН СССР, 1963, – 323 с.

68. Д.А.Родионов, Р.И.Коган, В.А.Голубева и др. Справочник по математическим методам в геологии. М., Недра, 1987. – 335 с.

69. Компьютерная биометрика. /Под ред. В.Н.Носкова. М.: – Изд. МГУ,1990. – 222 с.

70. Бостанджиян В.А. Распределение Пирсона, Джонсона, Вейбулла и обратное нормальное. Оценивание их параметров. – Черноголовка. РИО ИПХФ РАН, 2009. 240 с.

71. Бостанджиян В.А. Пособие по статистическим распределениям / РАН, Ин-т проблем хим. физики. –Черноголовка: Изд-во ИПХФ РАН, 2000. 1007 с.

72. И.Гайдышев. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001. – 752 с.

73. М.Абрамовиц, И.Стиган. Справочник по специальным функциям. – М, Наука, 1979. – 832 с.

74. В.Глинский, В.Ионин. Статистический анализ. – М, Финансы и статистика, 2002. – 345 с.

75. D.J.Finney. Probit analysis. – Cambridge University Press. 1971.

76. ГОСТ Р ИСО 5725-2–2002. Точность (правильность и прецизионность) методов и результатов измерений. Часть 2. М.: Изд-во стандартов. – 51 с.