Программа Статистика (FStat)

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 2
УСТАНОВКА ПРОГРАММЫ 3
НАДЕЖНОСТЬ ПРОГРАММЫ 3
ИСТОРИЯ ВЕРСИЙ 4
ДАННЫЕ 4
ФОРМАТ 4
ПОДГОТОВКА 5
ГРАФИКИ МОНИТОРИРОВАНИЯ 6
РАСПРЕДЕЛЕНИЕ 7
ВЫБОР ПЕРЕМЕННЫХ 8
СТАТИСТИКА 9
КОРРЕЛЯЦИОНЫЙ АНАЛИЗ 9
СРАВНЕНИЕ ВЫБОРОК 10
ДИАГНОСТИКА 11
КЛАСТЕРНЫЙ АНАЛИЗ 12
ОБЩАЯ СХЕМА АНАЛИЗА 12
РАССТОЯНИЕ МЕЖДУ ОБЪЕКТАМИ 13
РАССТОЯНИЕ МЕЖДУ ГРУППАМИ 14
КАЧЕСТВО ГРУППИРОВКИ 14
РАЗНОЕ 15
СОВЕТЫ 15
ЛИТЕРАТУРА 16

ВВЕДЕНИЕ

О ПРОГРАММЕ

Программа "Fwstat" была создана в давние времена для обработки научных данных многодневных наблюдений, поскольку имевшиеся математические программы не позволяли решать поставленные задачи. Первая версия программы была написана в 1988 году на языке Fortran IV для микрокомпьютера величиной с небольшой шкаф и оперативной памятью в 28 кб, вводом программ через перфоленту или с диска диаметром полметра. Вначале статистика программировалась по известным руководствам, а позднее были адаптированы исходные коды авторитетных математических библиотек.

Программа активно использовалась в научных исследованиях автора по разработке системы краткосрочного динамического прогнозирования обострений болезни с использованием физиологических и средовых факторов. Тщательно анализировались графики многодневных данных, особенно в период, предшествующий ухудшению состояния пациента.

В связи работой над новыми приоритетными проектами с 2004 года программа не развивается, но используется для тестов и отработки математических аспектов программирования, быстрого предварительного анализа данных.

В программе, наряду с традиционными статистическими процедурами (оценка близости к нормальному распределению, параметрические и непараметрические методы корреляционного анализа, сравнения средних значений), имеются возможности, которые отсутствуют в известных программах статистической обработки ("SPSS", "S-Plus", "Statictica"), либо их реализация достаточно трудоемка.

Представляется удобной **таблица подготовки данных**, позволяющая автоматически оценивать выскакивающие значения (выбросы, всплески) и близость распределений переменных к нормальному. Кроме того, имеется возможность логарифмирования и нормирования данных, удаления и выбора всплесков, смещения значений переменных друг относительно друга, создание приростов значений. После преобразования данные могут выводиться в файл, и далее использоваться в других статистических или графических программах.

Особенностью данной программы является возможность графического вывода мониторных данных одновременно, **одного графика под другим** . Имеется возможность выделить интересующий интервал наблюдения, задать пороговый уровень для оценки выраженных изменений величины данных, провести сглаживание данных, а также изменить масштаб.

Для быстрого анализа данных исследователям может помочь **таблица сопряженности "2x2"**, не требующая организации данных в файле или электронной таблице. Для оценки корректности сравнения переменных может быть полезна процедура расчета t-критерия Стьюдента по имеющимся значениям среднего арифметического, ошибки репрезентативности или среднеквадратического отклонения.

В программе реализованы иерархические аггломеративные методы кластерного анализа. Программа позволяет оценить оптимальность кластеризации на основе оценки прироста межкластерного расстояния. После кластеризации данных номера кластеров добавляются к исходному массиву и далее имеется возможность посмотреть в динамике изучаемую переменную в сопоставлении с динамикой кластеров. В программе проводится

автоматический расчет общестатистических характеристик (средняя арифметическая, среднеквадратическое отклонение, ошибка репрезентативности) сформированных кластеров, а также выводятся исходные значения, сгруппированные в соответствии с результатами кластеризации. Последние массивы могут быть использованы для последующей статистической и графической обработки.

Автор: Белялов Фарид Исмагильевич, **e-mail** : fbelyalov@yandex.ru

УСТАНОВКА ПРОГРАММЫ

Запустите программу установки setup.exe. После установки программы вызовите подпрограмму конфигурации и задайте каталоги расположения программы, хранения данных и вывода результатов, подберите параметры для оптимального графического представления данных и печати. При установке программы без инсталляционного диска нужно отредактировать файл fstat.ini.

НАДЕЖНОСТЬ ПРОГРАММЫ

Надежность программы обеспечивается использованием в программе стандартных математических библиотек IMSL и SSP. Результаты тестирования сопоставлялись с примерами из руководств по статистике и решениями известных статистических программ "SPSS" и "Statistica". И тем не менее, программа не проходила полного стандартного тестирования, обязательного для коммерческих статистических программ**.** Поэтому для расчетов рекомендуется использовать вышеуказанные или аналогичного уровня статистические программы (но не "Excel"). Современные программы математической обработки чаще разрабатываются на интерпретируемом языке Python (для запуска на компьютере должна быть установлена система Python с необходимыми библиотеками), имеющем огромную библиотеку мощных математических методов. Автор использует Python для исследовательских целей.

NB! Разные статистические программы могут дать несколько отличные результаты. Например, группировка данных при кластерном анализе (Евклидова метрика) на некоторых массивах существенно различается при использовании известных статистических программ.
В случае обнаружения ошибки в программе просьба сообщить об этом автору по адресу с подробным описанием ситуации ее появления и приложением файла данных, с которым в этот момент работала программа.

РАСПРОСТРАНЕНИЕ

Данная программа "FStat" распространяется бесплатно и в форме "как есть", в соответствии с общепринятой международной компьютерной практикой. Это означает, что за проблемы, возникающие в процессе инсталляции или эксплуатации программы "FStat", разработчик и распространитель ответственности не несут. Разработчик прилагает все усилия для того, чтобы данные проблемы никогда не возникали на компьютерах пользователей.

ИСТОРИЯ ВЕРСИЙ

версия 6.0
+ адаптация программы для Windows 10 и 11
+ вывод данных в кодировке UTF-8

версия 5.0
+ ввод данных из файлов Excel
+ написана справка
+ разработка калькуляторов

версия 4.0
+ разработка кластерного анализа
+ ввод из dbf-файла
+ фрагменты на ассемблере для ускорения работы

версия 3.0
+ программа переписана на Delphi в системе Турбо Pascal 3.0
+ графика

версия 2.0
+использование математических библиотек SSP и IMSL

версия 1.0
Разработка программы на Fortran в 1988 году на основе статистических руководств
+элементарная статистика

ДАННЫЕ

ФОРМАТ

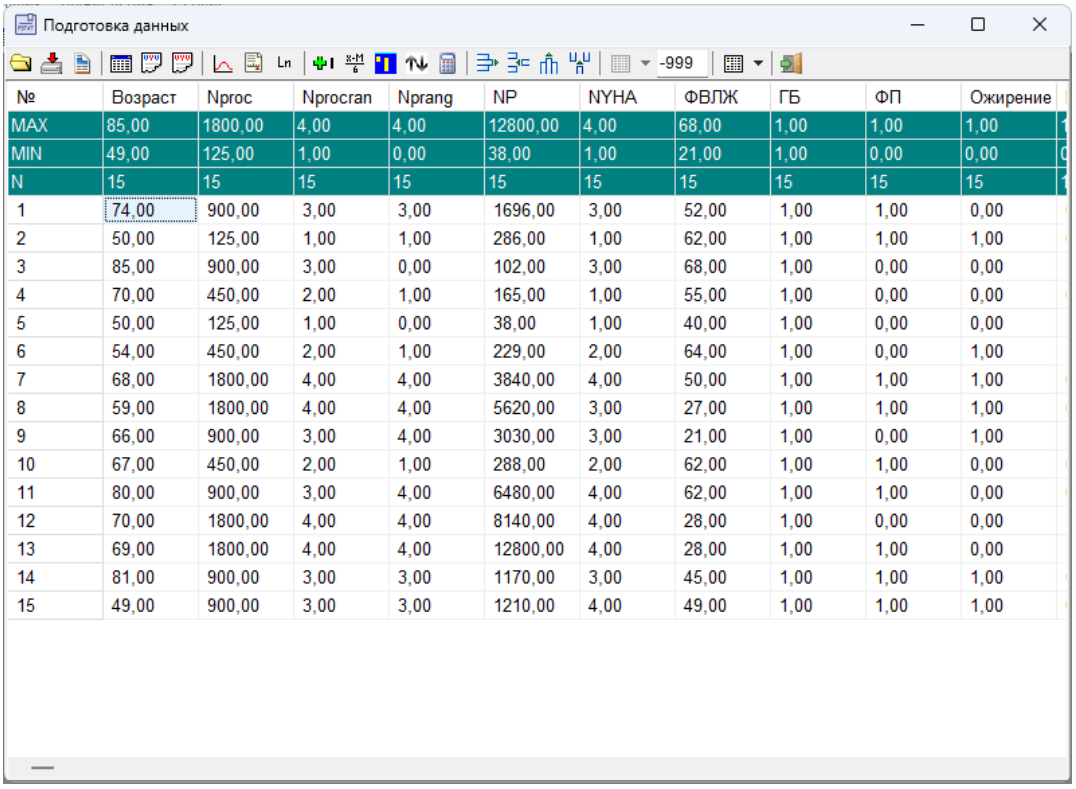
Данные могут быть представлены в текстовых форматах ASCII или XLS

Текстовый файл в популярных кодировках ASCII или UTF-8 можно сделать в различных текстовых редакторах, например, Notepad++. Данные организуются в виде матрицы с пробелом(ами) между столбцами. В первой строке указываются названия параметров, которые не должны превышать 8 символов и начинаться с цифры. Заметим, что десятичные значения включают точку, а не запятую как это принято в файлах DBF и XLS. Ниже показан пример данных:

АДс АДд ЧСС
120 60 55.6
140 70 88.9

Файл типа **XLS** создается в программе "Microsoft Excel" и должен в первой строке содержать названия параметров и соответствующие данные в столбцах.

ПОДГОТОВКА



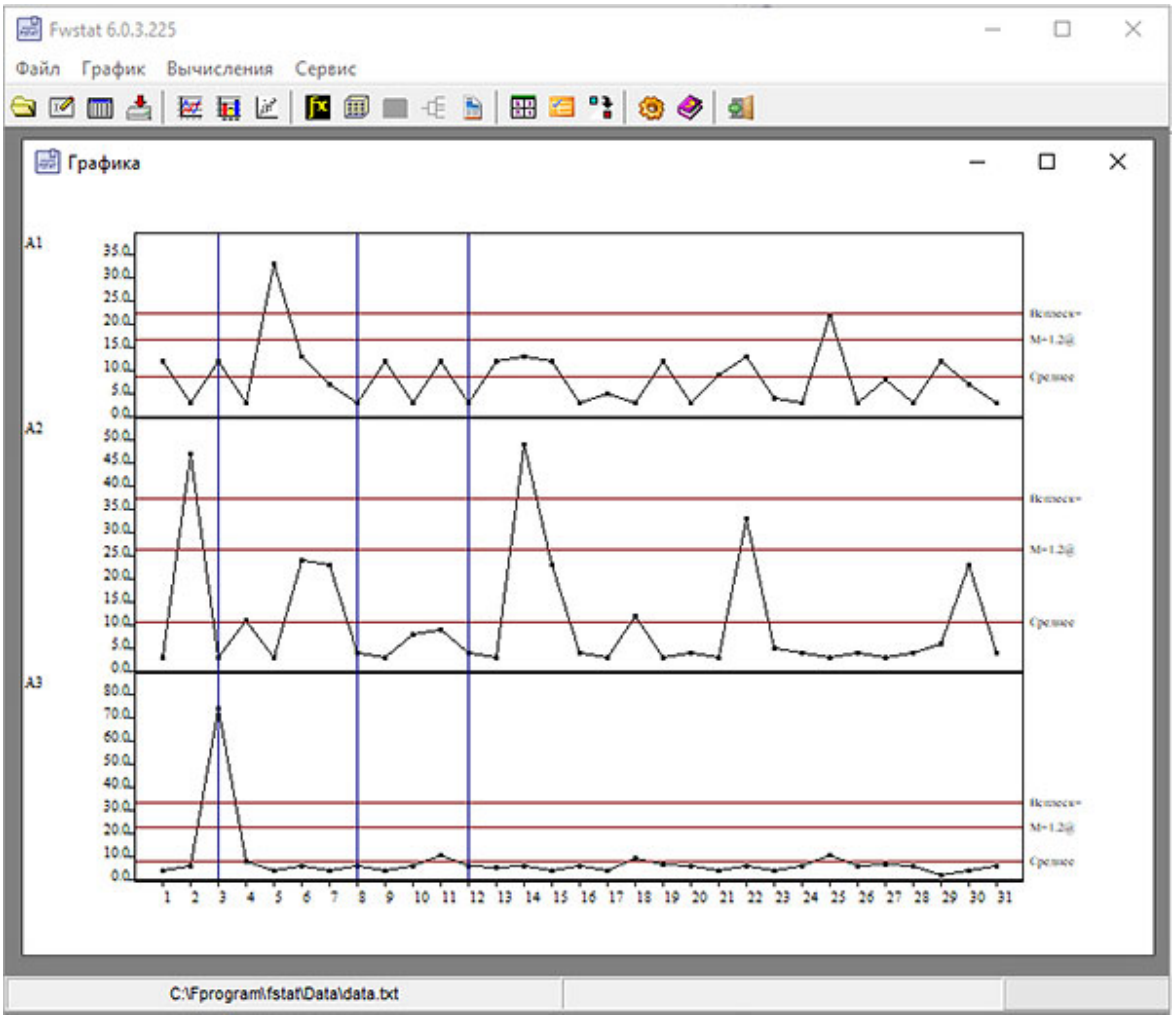
№	Возраст	Nproc	Nprocran	Nprang	NP	NYHA	ФВЛЖ	ГБ	ФП	Ожирение
MAX	85,00	1800,00	4,00	4,00	12800,00	4,00	68,00	1,00	1,00	1,00
MIN	49,00	125,00	1,00	0,00	38,00	1,00	21,00	1,00	0,00	0,00
N	15	15	15	15	15	15	15	15	15	15
1	74,00	900,00	3,00	3,00	1696,00	3,00	52,00	1,00	1,00	0,00
2	50,00	125,00	1,00	1,00	286,00	1,00	62,00	1,00	1,00	1,00
3	85,00	900,00	3,00	0,00	102,00	3,00	68,00	1,00	0,00	0,00
4	70,00	450,00	2,00	1,00	165,00	1,00	55,00	1,00	0,00	0,00
5	50,00	125,00	1,00	0,00	38,00	1,00	40,00	1,00	0,00	0,00
6	54,00	450,00	2,00	1,00	229,00	2,00	64,00	1,00	0,00	1,00
7	68,00	1800,00	4,00	4,00	3840,00	4,00	50,00	1,00	1,00	1,00
8	59,00	1800,00	4,00	4,00	5620,00	3,00	27,00	1,00	1,00	1,00
9	66,00	900,00	3,00	4,00	3030,00	3,00	21,00	1,00	0,00	1,00
10	67,00	450,00	2,00	1,00	288,00	2,00	62,00	1,00	1,00	0,00
11	80,00	900,00	3,00	4,00	6480,00	4,00	62,00	1,00	1,00	0,00
12	70,00	1800,00	4,00	4,00	8140,00	4,00	28,00	1,00	0,00	0,00
13	69,00	1800,00	4,00	4,00	12800,00	4,00	28,00	1,00	1,00	0,00
14	81,00	900,00	3,00	3,00	1170,00	3,00	45,00	1,00	1,00	1,00
15	49,00	900,00	3,00	3,00	1210,00	4,00	49,00	1,00	1,00	1,00

Проблема **пропусков** данных является очень сложной. Пропущенные значения могут восстанавливаться, например, средними значениями. Такой подход не уменьшает количества значений, но вносит определенные искажения в данные. В большинстве случаев корректнее удалять пропуски. При корреляционном анализе или парном сравнении рекомендуется удаление соответствующих пар значений двух вычисляемых переменных, сохраняющее больше данных по сравнению с удалением рядов.

Всплески (выбросы, выскакивающие значения) являются нетипичными, резко выделяющимися наблюдениями могут существенно изменить результаты статистического анализа. Существуют различные критерии всплесков, например, всплесками считают значения, которые выходят за границы ±2 стандартных отклонений (и даже ±1.5 стандартных отклонений) вокруг выборочного среднего. В программе применен критерий Румшинского Л.З., зависящий от числа наблюдений и уровня достоверности (<0.05). Заметим, что всплески представляют особый интерес для исследователя и могут анализироваться отдельно.
Обработку пропусков и всплесков проводят в **таблице подготовки данных**, которая появляется после ввода данных или может быть вызвана кнопкой . Кроме того, здесь можно удалять и добавлять переменные и значения, заменять и удалять пропуски и всплески, создавать приросты, нормировать и логарифмировать данные, заменять значения, смещать ряды, а также проводить арифметические операции с переменными. Для изменения порядка переменных нажмите левую кнопку мыши на интересующем столбце и после появления

вертикальной линии переместите столбец в нужное место. Аналогичным образом можно перемещать ряды значений.

ГРАФИКИ МОНИТОРИРОВАНИЯ



В программе Fwstat, наряду с традиционной статистикой и кластерным анализом, имеется возможность оценки выраженных колебаний физиологических параметров, графического вывода мониторных данных с заданными пороговыми уровнями одновременно, одного графика под другим.

Программа активно использовалась в научных исследованиях автора по разработке системы краткосрочного динамического прогнозирования обострений болезни с использованием физиологических и средовых факторов. Тщательно анализировались графики многодневных данных, особенно в период, предшествующий ухудшению состояния пациента.

РАСПРЕДЕЛЕНИЕ

Для корректного применения статистических методов требуется оценить соответствие данных нормальному (Гауссову) распределению. С этой целью определяют следующие показатели: асимметрия, эксцесс и критерий X^2.

Асимметрия оценивает скошенность вариационной кривой распределения частот. При нормальном распределении показатель равен 0. Если вершина кривой лежит правее центра распределения, то говорят о правосторонней или отрицательной асимметрии. Если же вершина кривой лежит левее центра распределения, то говорят о левосторонней или положительной асимметрии.

Эксцесс оценивает максимум (вершину) распределения частот вариационного ряда. При нормальном распределении показатель равен 0. Если происходит чрезмерное накопление частот в центральных классах вариационного ряда, то максимум располагается выше, чем у нормального распределения. В этом случае говорят о положительном эксцессе. Если же происходит накопление частот преимущественно в крайних классах вариационного ряда (плосковершинная, многовершинная кривая), то максимум располагается ниже, чем у нормального распределения. В этом случае говорят об отрицательном эксцессе.

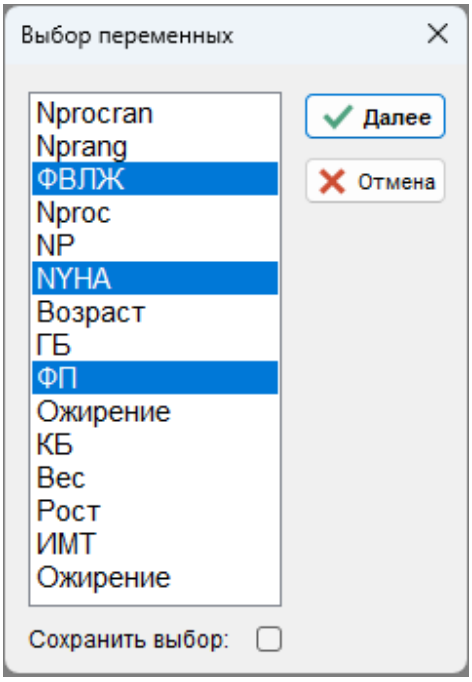
Критерий X^2 позволяет оценить гипотезу об отсутствии различий между эмпирическими частотами вариационного ряда и ожидаемыми при нормальном распределении частот. Критерий X^2 представляет собой сумму квадратических отклонений эмпирических частот от теоретических (нормальной кривой), отнесенную к теоретическим частотам. Для корректной работы критерия выборка должна содержать не менее 50 значений, а в крайних классах вариационного ряда находиться не менее пяти значений. В целях выполнения последнего условия в программе производится последовательное объединение крайних классов.

В случае существенного отличия распределения от нормального можно удалить всплески или логарифмировать данные (lnX). Последний прием используют в тех случаях, когда распределение ограничено слева от нуля, крутое слева и плоское справа (положительно-асимметричное), особенно если стандартное отклонение велико по сравнению со средним значением (коэффициент вариации >33%). Кроме того, для расчета ненормальных данных применяют **непараметрические статистические методы.**

NB! При малых объемах данных проверка нормальности ряда практически невозможна, а процедуры отбраковки грубых наблюдений бесполезны или малоэффективны, поэтому лучше применять непараметрические методы.

Оценка распределения данных проводится в **таблице подготовки данных**, которая появляется после ввода данных или может быть вызвана кнопкой . Если распределение существенно отличается от нормального по любому из вышеперечисленных показателей, то название переменной окрашено в вишневый цвет. Визуально распределение переменной можно посмотреть на графике (кнопка ), где представлены реальные и теоретические (нормальные) значения. Статистические показатели распределения (асимметрия, эксцесс и критерий X^2) можно рассчитать, нажав на кнопку  .

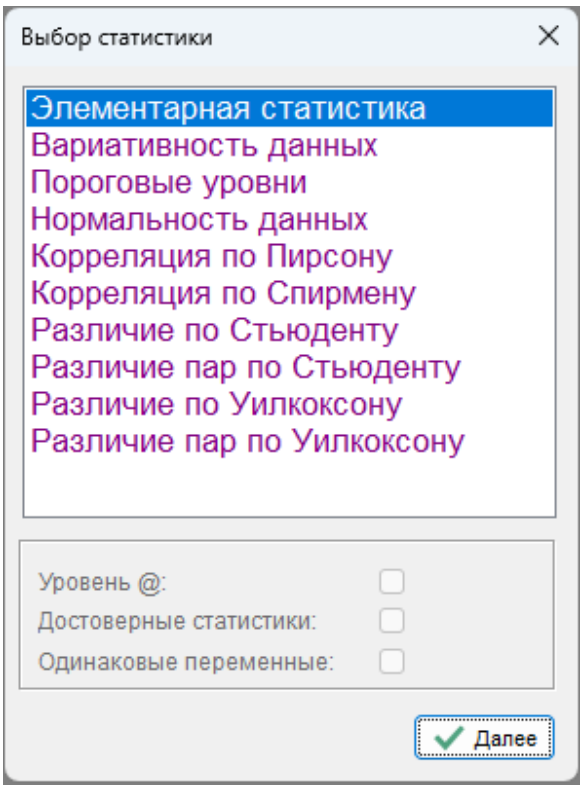
ВЫБОР ПЕРЕМЕННЫХ



Выбор переменных осуществляется в окне настройки переменных с помощью мыши и клавиатуры. Чтобы выбрать одну переменную нужно нажать клавишу Ctrl и левую кнопку мыши. Для выбора нескольких идущих подряд переменных нажмите клавишу Ctrl и не отпуская левую кнопку двигайте мышь.

В некоторых случаях желательна определенная последовательность вывода переменных. В этом случае порядок переменных можно изменить в таблице правки данных.

СТАТИСТИКА



КОРРЕЛЯЦИОНЫЙ АНАЛИЗ

Корреляционный анализ используется для оценки **линейной зависимости** между двумя переменными. Значения коэффициента корреляции лежат в пределах от -1 до +1. Положительная прямая связь (при значениях от 0 до +1) имеется когда большему значению одной переменной соответствуют большие значения другой. Отрицательная обратная связь (при значениях от -1 до 0) имеется когда большему значению одной переменной соответствуют меньшие значения другой. При значениях коэффициента корреляции в диапазоне 0-0.25 связь считается слабой, в диапазоне 0.25-0.5 - умеренной, в диапазоне 0.5-0.75 - сильной и более 0.75 - очень сильной. Оценку случайности отклонения коэффициента корреляции от 0 (достоверность) определяют по Р.А.Фишеру на основании t-распределения.

Выбор метода анализа существенно зависит от характера значений в исследуемых выборках. В случае нормального распределения используется **коэффициент корреляции Пирсона**, а при распределении отличном от нормального применяют **коэффициент ранговой корреляции Спирмена**.

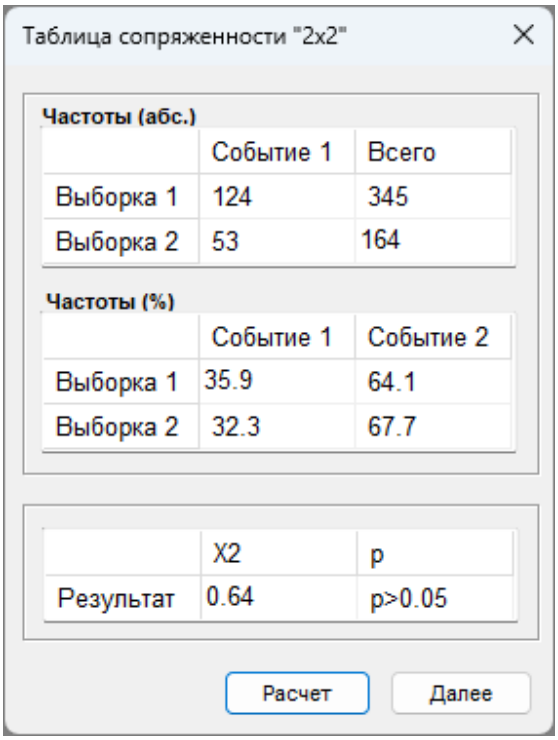
Визуально зависимость между двумя переменными можно представить в виде диаграммы рассеяния (скаттерограммы) с помощью кнопки  .
Для оценки нелинейной связи между двумя переменными используются **регрессионный анализ и нейронные сети**, которые не поддерживаются данной программой.

СРАВНЕНИЕ ВЫБОРОК

Выбор метода анализа определяется характером распределения значений в исследуемых выборках и зависимостью сравниваемых значений. Если каждому значению одной выборки соответствует значение другой выборки, то такие выборки называют **зависимыми**. Если такого соответствия нет, то выборки называются **независимыми** . При использовании методов сравнения проверяется гипотеза, что средние значения генеральных совокупностей не различаются.

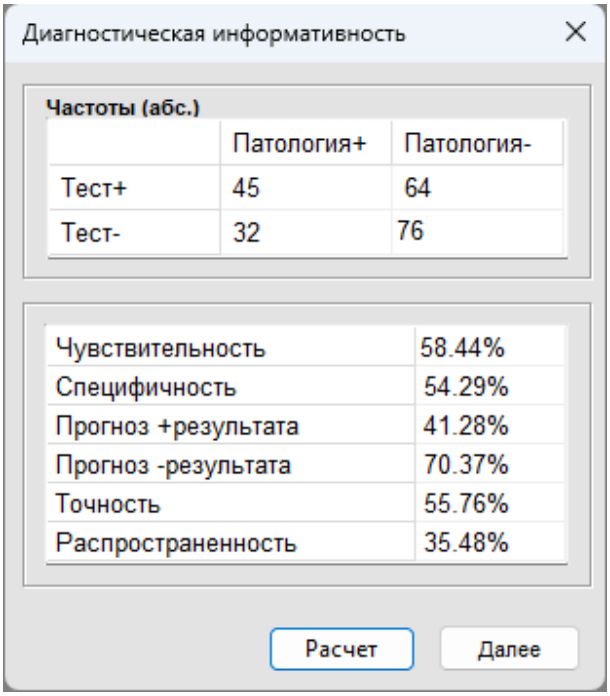
В случае нормального распределения значений в независимых выборках используют параметрический **t-критерий Стьюдента**, а для зависимых выборок предпочтение отдают **парному t-критерию Стьюдента.**

При распределениях отличных от нормального применяются непараметрические методы сравнения. Если выборки независимые, то для сравнения выборок применяют ранговый **U-критерий Манна-Уитни**. Если имеются зависимые выборки, то применяют **ранговый T-критерий Уилкоксона**.



Сравнение относительных частот событий в двух разных выборках можно провести с помощью анализа **таблицы сопряженности “2x2”**. Для проверки гипотезы принадлежности обоих выборок к общей генеральной совокупности используется двусторонний критерий X^2 . Для корректной работы данной статистики необходимо, чтобы частоты превышали 3. В ячейки таблицы заносят частоты двух событий, например, число выздоровевших и общее число пациентов. То же самое повторяют для другой выборки, например, число выздоровевших и общее число пациентов после применения исследуемого лекарственного препарата.

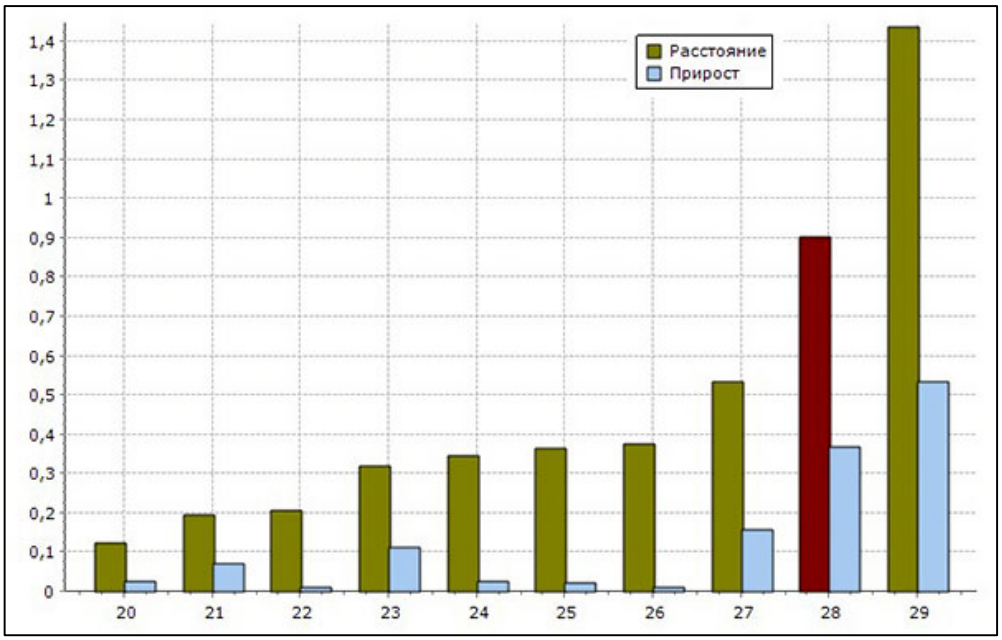
ДИАГНОСТИКА



Для оценки диагностической информативности различных методов применяют оценку показателей чувствительности, специфичности, тестов прогнозирования положительного и отрицательного результатов.

• Чувствительность (TP/(TP+FN)*100) показывает как эффективно тест выявляет заболевание.
• Специфичность (TN/(FP+TN)*100) показывает как хорошо тест определяет норму (отсутствие заболевания).
• Положительная прогностическая ценность (TP/(TP+FP)*100) определяет как часто пациент с положительным тестом имеет данное заболевание.
• Отрицательная прогностическая ценность (TN/(TN+FN)*100) определяет как часто пациент с отрицательным тестом не имеет данного заболевания.
• Точность прогноза ((TP+TN)/(TP+FN+FP+TN)*100) показывает как часто тест дает точный отрицательный или положительный результат.
• Распространенность заболевания ((TP+FN)/(TP+FN+FP+TN)*100) показывает частоту заболевания в данной группе пациентов.

КЛАСТЕРНЫЙ АНАЛИЗ



ОБЩАЯ СХЕМА АНАЛИЗА

Методы кластерного анализа предназначены для разбиения множества объектов на классы (группы, кластеры), так чтобы каждый объект принадлежал только одному классу и в один класс попадали наиболее сходные объекты.

1. Определить расстояние между объектами .

2. Выбрать методику оценки расстояния между близких объектов.

3. После группировки данных (кластеризации) нужно оценить качество, т.е. на каком шаге получены наиболее оптимальные группы (кластеры).

Заметим, что метод весьма субъективен и больше годится для генерации новых идей! Для получения содержательной классификации полезно воспользоваться следующими советами:

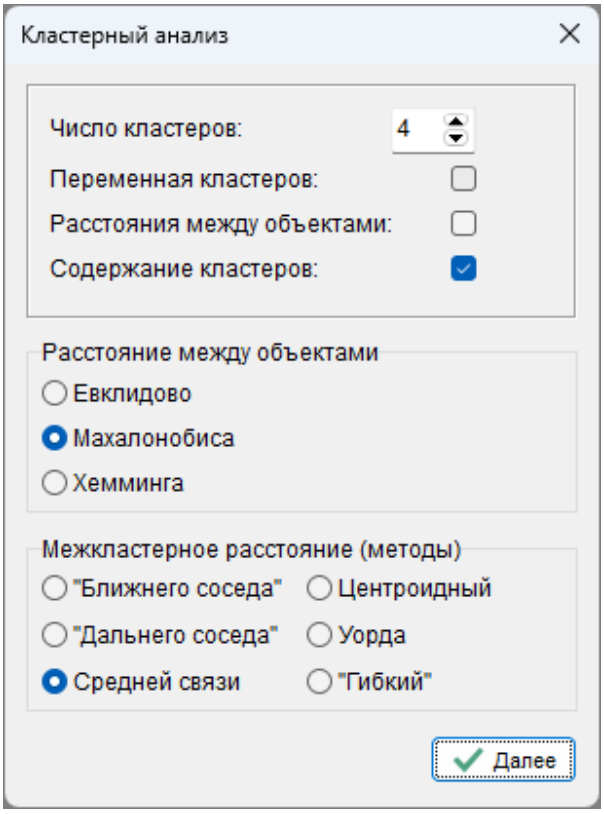
• Применяйте к данным несколько алгоритмов классификации с последующим сравнением результатов. Это позволяет выбрать устраивающий вас результат, т.е. подтверждающий вашу замечательную идею!

• Используйте несколько метрик и несколько вариантов параметров, управляющих работой алгоритма (гибкая стратегия), с последующим сравнением результатов. При этом выбираются вариант классификации наиболее устойчивый к вариации параметров.

• Результаты классификации тем устойчивее, чем больше объем выборки и чем меньше отношение размерности (числа параметров по которым проводится кластеризация) к объему выборки.

• Наличие аномальных наблюдений, как правило ухудшает результаты классификации, "сжимая" имеющиеся классы. Поэтому необходима проверка наличия таких наблюдений и их удаление.

РАССТОЯНИЕ МЕЖДУ ОБЪЕКТАМИ



Выбор метрики является узловым моментом исследования, от которого решающим образом зависит окончательный вариант разбиения объектов на классы. Ниже описаны критерии выбора той или иной метрики.

При использовании метрики Евклида проводится нормирование для устранения различий в масштабах параметров. Это оправдано, если разброс значений соответствует нормальной случайной дисперсии. Но если разброс связан с наличием подклассов, то нормирование бесполезно! Учтите, что нормирование преувеличивает роль резко выделяющихся наблюдений. Для применения данной метрики требуется соблюдение следующих условий:

• Признаки однородны по физическому смыслу.

• Все признаки важны при отнесении объекта к тому или иному классу (иначе требуется взвешенное Евклидово расстояние).

• Признаковое пространство совпадает с геометрическим пространством.

• Объекты извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с одной и той же матрицей ковариации.

• Компоненты объекта взаимонезависимы и имеют одну и ту же дисперсию.

Для применения метрики Махалонбиса требуется соблюдение следующих условий:

• Признаки объекта значительно коррелируют между собой.

• Признаки имеют различную значимость для классификации.

• Объекты извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с одной и той же матрицей ковариации.

• Нормирование не требуется.

Расстояние Хемминга используется для объектов, задаваемых дихотомическими (бинарными - есть, нет) признаками.

РАССТОЯНИЕ МЕЖДУ ГРУППАМИ

Метод **ближнего соседа** определяет расстояние по двум самым близким объектам групп. Данная стратегия монотонная, т.е. последовательность мер различий постепенно однонаправленно изменяется, и сильно сжимает пространство, что размывает группировку. Метод предпочтителен для групп вытянутой формы или с вытянутыми отростками, и если важно не разбиение на группы, а оценка меры смежности.

Метод **дальнего соседа** предложен P.MacNaughton-Smith (1965). Расстояние между группами определяется по двум самым дальним объектам этих групп. Данная стратегия монотонна и сильно растягивает пространство - группы различаются больше, чем на самом деле. Метод предпочтительнее использовать для компактных групп типа шаровых.

Центроидный метод предложен R.R.Sokal и C.D.Michener (1958). Расстояние определяется между центрами групп. Данная стратегия немонотонна, поэтому используется редко. Метод сохраняет метрику пространства и предпочтителен для естественного разбиения на группы не очень сложной формы.

Метод приращений внутригрупповой суммы квадратов отклонений предложен J.H.Ward (1963). Объединяются такие две группы, которые ведут к минимальному увеличению внутригрупповой суммы квадратов расстояний между каждым объектом и средней по группе, содержащему этот объект. Данная стратегия монототонна и растягивает пространство.

Метод группового среднего предложен G.N.Lance и W.T.Williams (1967). Расстояние между группами определяется как среднее расстояние между объектами двух этих групп. Данная стратегия монотонна и сохраняет метрику пространства. Метод предпочтителен для естественного разбиения на группы не очень сложной формы.
Гибкий метод изменяет метрику пространства в зависимости от значения константы b. Если b равно 0, то пространство не изменяется, если b принимает положительное значение - пространство сжимается, а отрицательное значение - растягивается. Обычно применяют константу равную -0.25. Данная стратегия монотонна.

КАЧЕСТВО ГРУППИРОВКИ

Использован показатель прироста расстояния на каждом шаге классификации.
Максимальный прирост показывает, что начинают объединяться дальние классы, и вероятно на предыдущем шаге была оптимальная классификация.

Программа имеет критерий оценки выбора шага, на котором формируется оптимальная группировка. Если происходит максимальный прирост расстояния при объединении с новой группой, причем присоединяется группа, содержащая более одного объекта, то классификация на предыдущем шаге является оптимальной.
Основным критерием качества и обоснованности полученного разбиения является содержательный анализ результатов, основанный на осмыслении исследователем возможных причинных механизмов осуществления и обособления полученных групп объектов.

РАЗНОЕ

СОВЕТЫ

“Есть ложь, наглая ложь и статистика.” (Дизраэли Дж.)

“Профессионализм пользователя как фактор успеха в решении задач анализа данных проявляется в качестве собранных им данных, в четкой формулировке целей проведения анализа и в личном участии в процессе обработки и интерпретации результатов.” (Александров В.В.,Алексеев А.И.,Горский Н.Д.)

“Первостепенное значение для выбора объема анализируемых данных (числа объектов, числа признаков) должны иметь нестатистические соображения, вытекающие из особенностей предметной области.” (Александров В.В.,Алексеев А.И.,Горский Н.Д.)

“Статистика ничего не может сказать о причинно-следственной связи признаков.” (Александров В.В.,Алексеев А.И.,Горский Н.Д.)

“Не следует гнаться за точностью измерения количественных признаков, избыточная точность не помогает, а часто мешает поиску решения задачи.” (Александров В.В.,Алексеев А.И.,Горский Н.Д.)

“В целом для показа общих качественных особенностей явления более подходящи графические методы.” (Кокс Д.,Снелл Э.)

“Все, что критерии значимости могут сделать, - это оценить, содержат ли рассматриваемые данные достаточно весомые свидетельства относительно направления определенного различия в указанный процессах”. (Кокс Д.,Снелл Э.)

““Мусор” на входе порождает “мусор” в результатах!”.

“Подбор и предобработка входной информации съедает до 80-90% времени аналитика”.

“Лучше быть точно неправым, чем приблизительно правым” (Тьюки Дж.).

“Среднее это задохнувшаяся индивидуальность” (Уайтхед А.Н.).

“От ложного знания к истинному незнанию”.

“Наблюдения и измерения никогда не бывают распределены по магической колоколообразной кривой” (Мостеллер Ф., Тьюки Дж.).

“Для слишком разбросанных “хвостов” вообще не получается хороших результатов ни в какой ситуации и ни для какого метода анализа” (Мостеллер Ф., Тьюки Дж.).

“Довольно простой анализ данных может пролить свет на глубокую проблему” (Мостеллер Ф., Тьюки Дж.).

“Сглаживание данных сглаживает и функцию, подлежащую оценке” (Мостеллер Ф., Тьюки Дж.).
“Практики часто разочаровываются в моделях типа уравнений множественной регрессии, так как “прогноз” удовлетворителен лишь для тех данных, которые участвовали в построении модели. Падение силы предсказания на “свежих” данных действует угнетающе” (Мостеллер Ф., Тьюки Дж.)

ЛИТЕРАТУРА

1. Александров В.В., Горский Н.Д. Алгоритмы и программы структурного метода обработки данных. Л.:Наука 1983;208 с.

2. Александров В.В., Алексеев А.И.,Горский Н.Д. Анализ данных на ЭВМ (на примере системы СИТО). М:Финансы и статистика. 1990;192

3. Вайнберг Дж., Шумекер Дж. Статистика. Пер. с англ. М:Статистика 1979;389.

4. Вирт Н. Алгоритмы+структуры данных = программы. Пер. с англ. М.:Мир 1985;99-103.

5. Готтсданкер Р. Основы психологического эксперимента. Пер. с англ. М 1982;464.

6. Джефферс Дж. Введние в системный анализ: применение в экологии.Пер. с англ. М:Мир 1981;256.

7. Дьяков В.П Справочник по алгоритмам и программам на языке Бейсик для ПЭВМ. М.:Наука 1987;144-145.

8. Дуда Р., Харт П. Распознавание образов и анализ сцен. Пер. с англ. М:Мир 1976;511.

9. Дэвис Дж.С. Статистический анализ данных в геологии. Пер. с англ. В 2 кн. М:Недра 1990.

10. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа. М.: Финансы и статистика 1986;232.

11. Закс Л. Статистическое оценивание. Пер. с нем. М.: Статистика 1976;598.

12. Кендалл М. Ранговые корреляции. Пер. с англ. М.:Статистика 1975;45-51.

13. Колкот Э. Проверка значимости.Пер с англ. М:Статистика 1978;128.

14. Кокс Д., Снелл Э. Прикладная статистика. Принципы и примеры. Пер.с англ. М:Мир 1984;200.

15. Лакин Г.Ф. Биометрия. М.:Высш.шк. 1990;352.

16. Петрович М.Л., Давидович М.И. Статистичесекое оценивание и прверка гипотез на ЭВМ. М.: Финансы и статистика 1989;191.

17. Румшинский Л.З. Математическая обработка результатов эксперимента. М.:Наука 1971;16-18,172-173.

18. Справочник по прикладной статистике. Под ред. Э.Ллойда, У.Ледермана. М: Финансы и статистика 1990.

19. Статистические методы для ЭВМ. Под ред. К.Энслейна и др. Пер. с англ. М.: Наука 1986;464.

20. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. М 1998;528.

21. Уиллиамс У.Т., Ланс Дж.Н. Методы иерархической классификации. В кн.:Статистические методы для ЭВМ. Под ред. К.Энлейна и др. Пер. с англ. М.:Наука 1986;.269-300.

22. Шураков В.В., Дайитбегов Д.М., Мизрохи С.В., Ясеновский С.В. Автоматизированное рабочее место для статистической обработки данных.М:Финансы и статистика 1990;190.
23.Эренберг А. Анализ и интерпретация статистических данных. Пер. с англ. М:Статистика 1981;406.