ВОЕННО-МЕДИЦИНСКАЯ АКАДЕМИЯ

В.И.Юнкеров, С.Г.Григорьев

МАТЕМАТИКО-СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Санкт-Петербург
2002

Юнкеров В.И., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований.– СПб.: ВМедА, 2002.– 266 с.

В книге просто и наглядно изложены назначение и сущность, как широко используемых, так и малоизвестных широкой научной общественности, математико-статистических методов описания, анализа и моделирования результатов медико-биологических исследований. Вполне строгое математическое описание каждого рассматриваемого метода сопровождается иллюстрацией конкретных оригинальных примеров преимущественно из практики авторов. Первая часть книги посвящена одномерной описательной статистике и оценке значимости различия признаков. Во второй части приведены многомерные методы анализа медицинских процессов и систем. Книга станет надежным подспорьем для врачей-исследователей в обработке результатов исследования. Издание ориентировано на студентов медицинских ВУЗов, врачей различных специальностей, научных сотрудников.

ISBN 5-94277-011-5

ISBN 5-94277-011-5

9 785942 770112 5

ПРЕДИСЛОВИЕ

Имея богатый многолетний опыт преподавания математической статистики курсантам, слушателям, адъюнктам и аспирантам Военно-медицинской академии, а также математико-статистического сопровождения многих научно-исследовательских работ, выполняемых в академии, авторы посчитали возможным поделится этим опытом на страницах, предлагаемого Вашему вниманию издания.

Статистическая обработка данных, полученных как в эксперименте, так и путем повседневного медицинского учета, необходима для проверки степени достоверности результатов, правильного их обобщения и выявления закономерностей медицинских процессов. Особенно важна роль статистических методов в моделировании медицинских систем и процессов с последующим использованием этих моделей для принятия верного решения в условиях неопределенности. Важно понимать, что каждый из методов математической статистики имеет свои возможности и ограниченную область применения. Только цель исследования и характер полученных данных определяют выбор математического аппарата для обработки этих данных.

Книга адресована главным образом студентам и аспирантам медицинских вузов, изучающим математическую статистику, а также научным медицинским работникам. Она не предназначена для изучения теоретических основ методов статистического анализа, поэтому в ней отсутствуют последовательное и деталь-

ное изложение алгоритмов анализа так как предполагается, что читатель имеет базовую подготовку в вопросах математической статистики и теории вероятностей. Математический комментарий дается только в самых необходимых случаях в доступном виде для понимания сущности методов и результатов анализа для врача, не имеющего специальной математической подготовки.

Практическая ценность книги состоит в большом числе содержательных примеров применения методов статистического анализа в медицинских исследованиях с помощью персональных компьютеров и пакетов программ по статистической обработке данных Excel и Statistica for Windows.

Первая часть книги посвящена методам статистического описания количественных и качественных переменных, определения значимости различия законов распределения и производных величин, а также изучения связи между переменными.

Во второй части книги описываются многомерные методы обработки данных с задачей создания математической модели изучаемого явления, объекта, процесса, что во многих случаях является конечной целью исследования. Основными элементами математических моделей являются признаки, которыми описываются объекты наблюдения. Такие признаки обычно подразделяют на контролируемые факторы, воздействующие на объекты - факторы-причины и параметры, характеризующие состояние изучаемой системы - показатели-отклики. Моделирование показателей-откликов в зависимости от значений воздействующих на них факторов - одна из основных и сложных задач статистического анализа в медицинских исследованиях.

В тех случаях, когда исследуемые факторы-причины и показатели-отклики измерялись в количественных шкалах и между ними установлена сильная и значимая корреляционная связь, моделирование выполняется методами ***регрессионного анализа*** (Дрейпер Н., Смит Г., 1986; Кувакин В.И., 1993; Григорьев С.Г, и др., 1998; Юнкеров В.И., 2000). В результате получается мо-

дель показателя в виде уравнения регрессии, с помощью которой решаются задачи прогнозирования исходов лечения, поиска оптимальных методов лечения, оценки степени влияния лечебных и профилактических мероприятий на отдаленные исходы. Линейное уравнение регрессии имеет вид:

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_i x_i + \ldots + b_k x_k ,$$

где $\hat{y}$ - прогнозируемое значение признака-отклика;

b_0, b_1, b_i, b_k – коэффициенты регрессионной модели;

x_1, x_i, x_k – значения факторов-причин изучаемого объекта.

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

Иногда результаты эксперимента включают как количественные, так и качественные факторы, воздействующие на объекты наблюдения. В этих условиях для моделирования показателя-отклика не только в зависимости от основных качественных факторов, но и с учетом влияния сопутствующих количественных факторов адекватным и эффективным методом является ***ковариационный анализ*** (Шеффе Г., 1980). Модель, полученная с помощью ковариационного анализа, имеет вид:

$$\hat{y}_i = \sum_{j=1}^{k} \nu_j F_{ij} + \sum_{s=1}^{p} \beta_s (F_i) x_{is}$$

где $\hat{y}_i$ - прогнозируемое значение показателя Y для i-го объекта (i=1, 2, …, n);

ν_j- коэффициент j-го эффекта основного неколичественного фактора F_j (j = 1,2,…,k);

$\sum_{j=1}^{k} \nu_j F_{ij}$ - сумма k эффектов основных неколичественных факторов;

$\beta_s(F_i)$ – коэффициент регрессии показателя на изменение сопутствующего количественного фактора x_s (s = 1, 2, …, p);

$\sum_{s=1}^{p} \beta_s(F_i) x_{is}$ - сумма линейных эффектов p сопутствующих количественных факторов X_s;

k – число эффектов (линейных и взаимодействия) основных неколичественных факторов;

p – число сопутствующих количественных факторов (ковариат).

Для решения задач классификации (распознавания образов) и отнесения объекта с определенным набором признаков к одному из известных классов используется ***дискриминантный анализ*** (О-Ким Дж., Мьюллер Ч.У., Клекк У.Р. и др., 1989). В медицине дискриминантный анализ применяется для решения диагностических, прогностических, экспертных задач, выбора методов и схем лечения. Для классификации определяется линейная комбинация (линейная дискриминантная функция), которая максимизирует различия между классами, но минимизирует дисперсию внутри классов. В итоге определяются линейные классификационные функции для каждого класса:

$$ЛКФ=b_0+b_1x_1+b_2x_2+\ldots+b_kx_k,$$

где b_0 - константа;

$b_1, b_2,\ldots b_k$- коэффициенты, которые определяются на основе данных обучающей информации;

$x_1, x_2,\ldots x_k$ – значения признаков изучаемого объекта.

Объект относится к классу с наибольшим значением ЛКФ.

Результаты исследования, в котором все переменные являются только качественными, традиционно сводятся в таблицы сопряженности. Моделировать по таким таблицам лучше всего посредством процедур ***логлинейного анализа*** (Аптон Г., 1982; Елисеева И.И., Рукавишников В.О., 1982; Григорьев С.Г., и др., 1998).

Логлинейный анализ обеспечивает установление силы и значимости связей между признаками с учетом их взаимодействия, определение степени влияния входных факторов на выходные результирующие признаки-отклики, прогнозирование ожидаемых частот наблюдений при определенных сочетаниях уровней факторов.

Анализ результатов эксперимента, содержащих качественные факторы и количественный признак-отклик, оценивающий продолжительность жизни (продолжительность ремиссии хронического заболевания, многолетней выживаемости онкологических больных после оперативного лечения, химиотерапии, лучевого лечения и др.) и построение модели функции продолжительности жизни проводится методом ***анализа времени выживания*** (Беляев Ю.К., 1987; Ермаков С.П., Гаврилова Н.С., 1987; Кокс Д.Р., Оукс Д., 1988; Григорьев С.Г. и др., 1998).

В последнее время в иностранной и отечественной литературе все чаще встречаются сведения о методе моделирования с помощью логистической регрессии (Bates, D. M., & Watts, D. G., 1988; Григорьев С.Г., Юнкеров В.И., Клименко Н.Б., 2001). Показаниями к применению метода являются:

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

факторы-причины - преимущественно качественные.

Математическое описание модели имеет вид:

$$\hat{y} = \exp(b_0+b_1x_1+b_2x_2+\ldots b_kx_k)/ [1+ \exp(b_0+b_1x_1+b_2x_2+\ldots b_kx_k)],$$

где $\hat{y}$ – вероятность исхода прогнозируемой градации признака-отклика;

b_0 - константа;

$b_1, b_2, \ldots b_k$ - коэффициенты для k симптомов $x_1, x_2, \ldots x_k$;

$x_1, x_2, \ldots x_k$ - возможные значения k симптомов.

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

Каждый рассматриваемый математико-статистический метод сопровождается примерами из реальных исследований, в которых авторы непосредственно участвовали.

СПИСОК УСЛОВНЫХ ОБОЗНАЧЕНИЙ

A, B, C - неколичественные (категорированные) факторы в многомерном дисперсионном анализе;

X, Y, Z - количественные переменные в многофакторном анализе;

x, y, z - частные значения количественных переменных;

$\bar{x}$, $\bar{y}$, $\bar{z}$ - средние арифметические значения переменных X, Y, Z;

$\bar{\Delta}x$ - среднее арифметическое значение разности переменной X в двух связанных выборках;

x_{min}, x_{max} - минимальное и максимальное значение переменной X в выборке;

n - количество наблюдений в выборке;

n' - число степеней свободы;

SS - сумма квадратов отклонений переменной от среднего арифметического значения;

S - среднее квадратическое (стандартное) отклонение переменной в выборке;

Mo, Me - мода и медиана переменной в выборке;

$m_{\bar{x}}$ - средняя квадратичная (стандартная) ошибка среднего арифметического значения переменной X;

$\bar{x} \pm t_{95} m_{\bar{x}}$ - 95%-й доверительный интервал среднего значения переменной X;

As - коэффициент асимметрии переменной в выборке;

Ex - коэффициент эксцесса переменной в выборке;

H0, H1 - нулевая и альтернативная статистические гипотезы;

p - уровень значимости - вероятность нулевой гипотезы H0;

1-p - доверительная вероятность альтернативной гипотезы H1;

t, t_{05}, t_{01}, t_{001} - критерий Стьюдента и его критические значения для уровня значимости p=0,05, p=0,01, p=0,001;

F, F_{05}, F_{01}, F_{001} - критерий Фишера и его критические значения для уровня значимости p=0,05, p=0,01, p=0,001;

χ^2, χ^2_{05}, χ^2_{01}, χ^2_{001} - критерий Пирсона и его критические значения для уровня значимости p=0,05, p=0,01, p=0,001;

$\bar{P}$ - относительная величина частоты (частость) признака в выборке;

$m_{\bar{p}}$ - средняя квадратичная ошибка частоты признака в выборке;

$\bar{P} \pm t_{95} m_{\bar{p}}$ - 95%-й доверительный интервал частоты признака в выборке;

δp - поправка Йетса для частости признака;

$\varphi = 2 \arcsin \sqrt{\bar{P}}$ - переменная Фишера - радианная мера частости $\bar{P}$;

m_{φ} - средняя квадратичная ошибка переменной Фишера;

r_{xy} - коэффициент парной корреляции переменных X и Y Пирсона;

r_s - ранговый коэффициент корреляции Спирмена;

Can r - коэффициент канонической корреляции;

R - коэффициент множественной корреляции;

R^2 - коэффициент множественной детерминации;

a, b_0 - константы в уравнениях регрессии;

$b_1, b_2, ..., b_k$ - коэффициенты регрессии;

$\beta_1, \beta_2, ..., \beta_k$ - стандартизованные коэффициенты регрессии;

S_0 - средняя квадратичная ошибка прогноза по уравнению регрессии;

$\hat{y}$ - прогнозируемое значение параметра-отклика по уравнению регрессии;

$m_{\hat{y}_x}$ - средняя квадратичная ошибка прогноза среднеожидаемого значения параметра-отклика;

K_i% - степень влияния i-го фактора на параметр-отклик;

ЛДФ - линейная дискриминантная функция;

ЛКФ - линейная классификационная функция;

КЛДФ - каноническая линейная дискриминантная функция;

$P_ч$ - показатель чувствительности метода диагностики;

P_α - вероятность ошибки диагноза первого рода;

P_c - показатель специфичности метода диагностики;

P_β - вероятность ошибки диагноза второго рода;

$P_{бо}$ - вероятность безошибочного диагноза;

S(t) - функция вероятности выживания на время t;

h(t) - функция интенсивности срыва жизни на время t.

ЧАСТЬ I

ОДНОМЕРНАЯ ОПИСАТЕЛЬНАЯ СТАТИСТИКА И ОЦЕНКА ЗНАЧИМОСТИ РАЗЛИЧИЯ ПРИЗНАКОВ

Математико-статистическое описание данных медицинских исследований и оценка значимости различия производных величин, характеризующих эффективность профилактических, диагностических и лечебных мероприятий и процедур являются одним из основополагающих разделов доказательной медицины. Именно этому разделу и посвящена первая часть книги.

Глава 1. ПЕРВИЧНАЯ СТАТИСТИЧЕСКАЯ ОБРАБОТКА КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ, ОЦЕНКА ЗНАЧИМОСТИ ИХ РАЗЛИЧИЯ

Характеристика биологических объектов, как сложных стохастических систем

Изучаемые в медицине объекты - трудовые коллективы, отдельные здоровые или больные люди, лабораторные животные, микроорганизмы и т.п., являются сложными стохастическими системами (рис.1.1), функционирующими при воздействии на них множества входных факторов. Часть таких факторов X1, X2, ... Xk являются контролируемыми, измеряемыми количественно, оцениваемыми в баллах или номинально. Другая часть относится к группе неконтролируемых, случайных факторов и зачастую неизвестных, они не поддаются измерению, но оказывают воздействие на систему, результатом которого является случайность ее состояния и функционирования.

Рис.1.1.Представление объекта исследования в виде «черного ящика».

Состояние системы характеризуется множеством выходных параметров Y1, Y2, ... Yl, которые также измеряются количественно или в баллах и представляют собой случайные величины, следующие нормальному или иному закону распределения с соответствующими числовыми характеристиками.

Количество входных контролируемых факторов и выходных параметров, описывающих объект исследования, определяется в зависимости от цели и задачи исследования. Так, например, для исследования связи между факторами тяжести состояния и факторами и параметрами, характеризующими эффективность лечения пострадавших с сочетанной механической травмой при ведущем повреждении головы, в качестве входных контролируемых факторов целесообразно иметь:

X1 - возраст - V, лет;

X2 - время доставки - D, ч;

X3 - частота пульса - P, уд./мин.;

X4 - артериальное давление систолическое - ADS, мм рт. ст.;

X5 - тип дыхания - TD, в баллах: 1 - нормальное, 2 - частое, 3 - патологическое;

X6 - речевой контакт - R, в баллах: 1 - нормальный, 2 - нарушенный, 3 - отсутствует;

X7 - анизокория - A, в баллах: 0 - нет, 1 - есть;

X8 - светлый промежуток - SP, в баллах: 0 - нет, 1 - есть;

X9 - кровопотеря - KP, мл.

Выходными параметрами могут быть:

Y1 - возникновение осложнений - OSL, в баллах: 0 - нет, 1 - есть;

Y2 - срок лечения - SRLEC, дней;

Y3 - исход лечения - I, в баллах: 0 - выжил, 1 - умер.

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

Статистический анализ сложной системы включает:

– статистическое описание переменных;

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

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

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

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

Для проведения многомерного статистического анализа изучаемого явления на ПК с применением ППП исследователь должен иметь достоверную базу данных (БД), представляющую собой матрицу наблюдений с достаточным числом случаев наблюдений по всем k входным факторам и l выходным параметрам.

Выборочный метод наблюдения - основной метод научного исследования

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

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

– случайностью отбора объектов однородной генеральной совокупности в выборку, когда каждый объект генеральной совокупности должен иметь одинаковую вероятность попадания в выборку;

– выборка должна иметь достаточную численность независимых наблюдений.

Число случаев наблюдений в выборке n называется объемом выборки. Выборочный метод наблюдения является основным при выполнении конкретных целей и задач исследования с применением различных методов многомерного статистического анализа.

По данным выборочного наблюдения объектов генеральной совокупности в соответствии с целью и задачами исследования формируется база данных (БД), представляющая матрицу наблюдений размером

$$n\times(k+l), \qquad (1.1)$$

где n - число строк в матрице равное числу случаев наблюдавшихся объектов;

k - число входных контролируемых факторов;

l - число выходных параметров;

(k+l) - число столбцов в матрице наблюдений.

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

Все элементы матрицы наблюдений должны иметь количественные значения по интервальной или порядковой шкале. Так, например, матрица наблюдений n=83 пострадавших с сочетанной механической травмой при ведущем повреждении головы и результатами наблюдений k=9 входных факторов и l=3 выходных параметров будет иметь размер 83×(9+3), т.е. 83 строки и 12 столбцов.

Для удобства статистического описания переменных различных групп объектов наблюдения в матрицу наблюдений необходимо ввести группировочные переменные. Например, Gr1- признак контрольной и опытных групп на соответствующем числе уровней (например, 0 - контрольная группа, 1 - опытная группа в день поступления, 2 на 7-е сути лечения и т.д.); Gr2 - признак тяжести состояния на четырех уровнях (0 - легкая, 1 - средняя, 2 - тяжелая, 3 - крайне тяжелая степень).

Достоверность БД определяет качество статистического анализа и, следовательно, выводов и рекомендаций по результатам исследований. Поэтому, важнейшей обязанностью исследователя является тщательная проверка БД и ее редактирование (исправление грубых технических ошибок, исключение явно анормальных наблюдений, дополнение пропущенных данных, введение дополнительных группировочных признаков для отличия, например, контрольной и опытных групп и т.п.). БД под именем соответствующего теме файла хранится на дискете или на жестком диске своего ПК. Создание и редактирование БД может выполняться с помощью одного из пакетов прикладных программ (ППП): табличного редактора (Excel), пакета статистического анализа данных (Statgraphics, Statistica, SPSS и др) по модулю Data Management или системы управления базой данных (СУБД).

Задачи статистического описания переменных

Для определения методов статистического анализа БД необходимо знать характер распределения и числовые характеристики всех переменных, входящих в матрицу наблюдений. Наилучшие результаты многомерного статистического анализа данных медико-биологических исследований получают тогда, когда распределение входных контролируемых факторов и выходных параметров нормальное или близкое к нему.

Основными задачами статистического описания переменных являются:

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

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

–оценка значимости различия показателей в независимых и связанных выборках.

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

Статистический ряд распределения дает представление о виде распределения показателя в диапазоне полученных наблюдений и является основой для оценки его соответствия с тем или иным теоретическим законом распределения. Графической иллюстрацией статистического ряда распределения является гистограмма и кумулятивная линия.

Оценка значимости различия показателей в независимых и связанных выборках - одна из основных задач решаемых исследователями при сравнении методов профилактики, лечения различных заболеваний, состояния работоспособности членов трудовых коллективов в различных условиях и в других подобных ситуациях.

ППП статистического анализа представляют широкие возможности для статистического описания переменных БД. Наиболее полное описание можно получить по ППП Statistica 5.0 for Windows, однако, более просто и с достаточной полнотой переменные БД можно охарактеризовать с помощью ППП Excel 7.0 (пример 1.1).

Определение числовых характеристик случайных переменных по результатам выборочного наблюдения

Числовые характеристики переменных подразделяются на три вида:

–характеристики положения;

–характеристики рассеяния;

–характеристики вида распределения.

К характеристикам положения относятся:

–среднее арифметическое значение - $\bar{x}$ (mean);

–медиана - Me (median);

–мода - Mo (mode);

–среднее геометрическое значение - $\bar{x}_g$ (geometric mean);

–среднее гармоническое значение - $\bar{x}_h$ (harmonic mean).

К характеристикам рассеяния значений переменной относятся:

–минимальное - x_{min} (minimum) и максимальное - x_{max} (maximum) значение;

–размах вариационного ряда - $R=x_{max}-x_{min}$(range);

–дисперсия - S^2 (variance);

–среднее квадратичное (стандартное) отклонение S (standard deviation);

–25%-й (LQ) и 75%-й (UQ) квартили и межквартильный размах (RQ = UQ - LQ);

–средняя квадратичная ошибка среднего значения m_x (standard error);

–95%-й доверительный интервал истинного среднего значения.

Вид распределения характеризуют коэффициенты:

–асимметрии в натуральном и стандартизованном виде As (skewness);

–эксцесса также в натуральном и стандартизованном виде Ex (kurtosis).

Аналитические выражения числовых характеристик и их сущность даны в справочной литературе и в частности в [3, 6]. Они реализованы в модулях ППП. Примеры расчета и анализа числовых характеристик даны с помощью ППП Microsoft Excel - в примере 1.1.

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

–среднее арифметическое, геометрическое и гармоническое значения незначительно различаются друг от друга, а также с модой и медианой;

–минимальные и максимальные значения примерно равноудалены от среднего значения;

–стандартизированные коэффициенты асимметрии и эксцесса по абсолютной величине меньше |2|.

Расчет числовых характеристик продемонстрирован в примере 1.1.

Оценка точности и надежности числовых характеристик

Числовые характеристики переменных, рассчитанные по выборке, содержат ошибки по отношению к аналогичным в генеральной совокупности. Характеристикой ошибок, следующих нормальному распределению, является средняя квадратичная ошибка, например, для среднего значения показателя $m_{\bar{x}}$.

Любое исследование должно включать элемент оценки точности и надежности числовых характеристик. Оценкой точности и надежности является 95%-й доверительный интервал истинного среднего значения. Например, истинное среднее значение показателя или по другому среднее значение генеральной совокупности находится в доверительном интервале

$$M_{95} = \bar{x} \pm t_{95} \times m_{\bar{x}}, \qquad (1.2)$$

где t_{95} - табличное значение t - критерия Стьюдента, отвечающее доверительной вероятности 95% по числу степеней свободы n'=n-1;

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

$$m_{\bar{x}} = \frac{S_x}{\sqrt{n}}, \qquad (1.3)$$

где S_x - среднее квадратичное отклонение показателя в выборке.

Из формулы (1.3) следует, что ошибка уменьшается с увеличением объема выборки. Так, чтобы уменьшить ошибку в два раза, число наблюдений следует увеличить в четыре раза.

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

$$X = \bar{x} \pm t_{95} \times S_x. \qquad (1.4)$$

Определение статистического ряда распределения случайной переменной по результатам выборочного наблюдения

Эмпирическое распределение переменной представляется в виде статистического ряда распределения, характеризующего связь между возможными значениями переменной и частостью их наблюдения в выборке. Для построения статистического ряда распределения в выборке необходимо иметь несколько десятков и более наблюдений, которые группируются в m интервалов (разрядов).

Выбор числа интервалов группировки возможен:

–по формуле Стерджеса:

$$m=1+3{,}32\times\lg n, \qquad (1.5)$$

– по эмпирически выработанным рекомендациям:

Объем выборки, n	Число интервалов, m
25 - 40	5 - 6
40 - 60	6 - 8
60 - 100	7 - 10
100 - 200	8 - 12
более 200	10 - 15

Подготовку данных для статистического ряда распределения выполняют в следующем порядке:

–в зависимости от числа наблюдений n выбирают число интервалов ряда m;

–определяют размах вариационного ряда $R=x_{max}-x_{min}$;

–рассчитывают длину интервала h=R/m;

–определяют границы и средние точки интервалов;

–подсчитывают частоту наблюдений, частость и накопленные частоту и частость для каждого интервала.

По данным статистического ряда распределения строят гистограмму и кумулятивную линию распределения. По виду гистограммы и кумулятивной линии делают предварительные выводы о характере и соответствии эмпирического распределения определенному теоретическому распределению.

Закон нормального распределения случайной переменной

Множество биологических и медицинских показателей, ошибки их измерения следуют нормальному распределению. Он адекватно описывает случайные величины, формирующиеся под влиянием большого числа статистически независимых факторов, когда ни один из них не доминирует над остальными. Распределениям близким к нормальному следуют показатели физического развития, составляющие плазмы крови и др. показатели.

Основные свойства закона нормального распределения (рис.1.2):

–равенство числовых характеристик $\bar{x}$=Mo=Me;

–симметричность отклонений от среднего значения;

–малые отклонения более вероятны, большие - менее вероятны;

–практические пределы отклонений от среднего значения ± 3S (с вероятностью 99,7%);

–вероятность значений переменной на интервалах равных одному среднему квадратичному отклонению дана на рис.1.2;

–качественно значения переменной оценивают по величине их отклонений от среднего значения, как показано на рис.1.2.

Рис.1.2. Кривая нормального распределения.

Для приведения любых переменных к одному масштабу применяют нормирование (стандартизацию):

$$z = \frac{x - \bar{x}}{S}. \qquad (1.6)$$

При этом z принимает значение для практических пределов рассеяния от -3 до +3.

В справочной литературе даются функции плотности f(z) рис.1.3 и интегральной функции F(z) нормального распределения (рис.1.4).

Плотность нормального распределения:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \qquad (1.7)$$

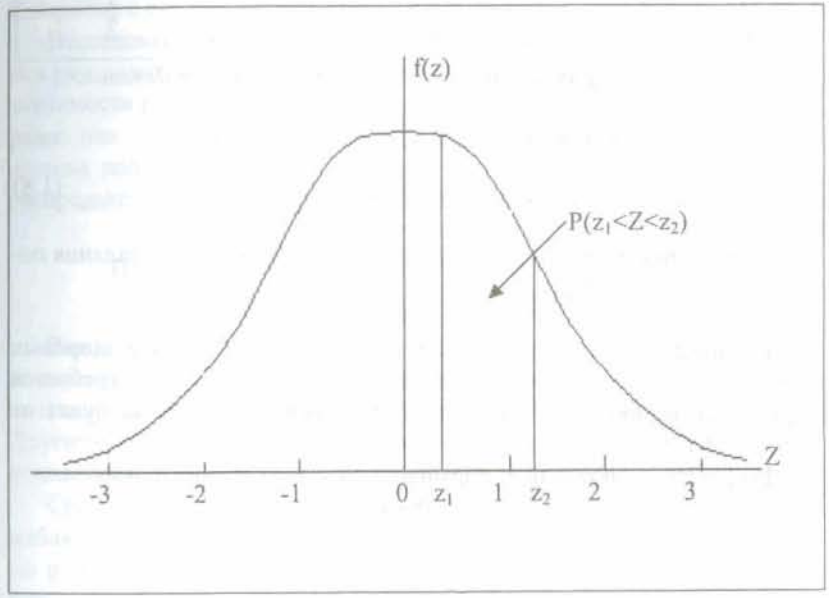

Рис.1.3.Кривая плотности нормального распределения.

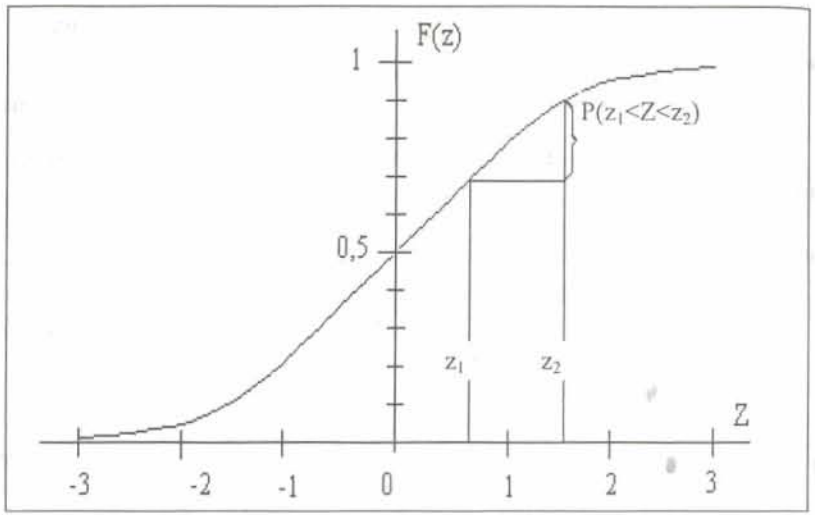

Рис.1.4.Интегральная функция нормального распределения.

Интегральная функция нормального распределения:

$$F(z) = P(X < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-\frac{z^2}{2})dz \cdot \qquad (1.8)$$

Функция применяется для определения вероятностей попадания показателя в интервал (z_1, z_2)

$$P(z_1<Z<z_2)=F(z_2)-F(z_1).$$

Например, при оценке содержания гемоглобина в крови здоровых мужчин в возрасте 20-30 лет получено $\bar{x}$ =145 г/л, S=5 г/л; требуется определить вероятность того, что содержание гемоглобина будет от 135 до 155 г/л.

Для решения определяем нормированное значение интервала:

$$z_1 = \frac{135-145}{5} = -2;$$

$$z_2 = \frac{155-145}{5} = +2$$

По таблице функции распределения:

$$F(z_1)=F(-2)=0,023;$$

$$F(z_2)=F(+2)=0,977;$$

$$P(-2<Z<2)=F(2)-F(-2)=0,977-0,023=0,954 \text{ или } 95,4\%.$$

Оценка соответствия эмпирического и теоретического законов распределения случайной переменной

Предварительные выводы о виде распределения переменной можно сделать по статистическому ряду распределения, гистограмме и кумулятивной линии, являющимися аналогами теоретических функций плотности распределения (рис.1.3) и интегральной функции распределения (рис.1.4). Для окончательного суждения о соответствии эмпирического распределения определенному теоретическому закону распределения применяют специальные критерии Пирсона, Колмогорова-Смирнова и др.

Алгоритм решения задач на ПК с ППП Statistica 5.0. предусматривает расчет по статистическому ряду распределения χ^2-критерия Пирсона и его уровня значимости p, а также d-критерия Колмогорова-Смирнова с его уровнем значимости p.

Исследователь выдвигает гипотезу H0 (нулевую) о соответствии законов распределения. Эту гипотезу принимают, если ее вероятность (уровень значимости p) будет больше 0,05, и отвергают, если ее вероятность будет равна или меньше 0,05 (т.е. $p \le 0,05$). В последнем случае исследователь должен подыскать для описания переменной более подходящий закон распределения (экспоненциальный, γ - распределения и т.п.).

Проверка статистических гипотез по результатам выборочного наблюдения

Важное место в медицинских исследованиях занимает сравнение показателей состояния организма в норме и при патологии, до лечения и после лечения или при применении различных методов лечения. Другими словами, теория проверки статистических гипотез является основным инструментом доказательной, а не интуитивной медицины.

Сравнение показателей выполняется по результатам выборочного наблюдения. При этом надежные результаты можно получить только по репрезентативным выборкам достаточного объема. При сравнении показателей, например, в контрольной (здоровые) и опытной (с патологией) группах выдвигают статистические гипотезы:

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

H0 - нулевую гипотезу - о равенстве (соответствии) показателя в опытной и контрольной группах.

Гипотезу H1 принимают, если ее вероятность имеет значение равное или больше 95% и отклоняют, если ее вероятность будет меньше 95%. В этом случае принимают гипотезу H0, а ее вероятность, как альтернативной, будет p>0,05.

Вероятность H0 p называют уровнем значимости, а величину 1-p называют доверительной вероятностью гипотезы H1.

Отметим жесткий подход к принятию гипотезы о существенном различии показателей, характеризующих состояние организма, при сравнении вновь предлагаемых и традиционных методов лечения. Практически задача проверки статистических гипотез решается либо графически (приближенно), либо с помощью специальных критериев, среди которых наибольшее значение приобрел t-критерий Стьюдента.

Оценка значимости различия средних значений показателя в независимых выборках

Независимыми называются выборки, в каждой из которых наблюдаются различные объекты, например первая контрольная группа (здоровые) и вторая опытная группа (больные, получающие определенную схему лечения).

Исходными данными для решения являются числовые характеристики показателя, полученные по исходной матрице наблюдений. К таким характеристикам относятся:

выборка 1: $n_1, \overline{x}_1, S_1, m_{\overline{x}_1}$, 95%-й доверительный интервал

$$M_1 = \overline{x}_1 \pm t_{95} \times m_{\overline{x}_1};$$

выборка 2: $n_2, \overline{x}_2, S_2, m_{\overline{x}_2}$, 95%-й доверительный интервал

$$M_2 = \overline{x}_2 \pm t_{95} \times m_{\overline{x}_2}.$$

По 95%-м доверительным интервалам дается приближенное графическое решение. Если доверительные интервалы не перекрывают друг друга или их перекрытие не превышает 1/3, можно считать, что имеет место значимое различие средних значений показателя в двух выборках.

Если перекрытие доверительных интервалов превышает 1/3, следует признать, что различие средних значений показателя в этих выборках незначимое (несущественное, недостоверное). Однако, приближенный метод оценки значимости различия по доверительным интервалам может использоваться в качестве экспресс метода, он хорош для графической демонстрации средних значений признаков и 95%-х доверительных интервалов их истинных значений. Более обоснованное решение получают по t-критерию

Стьюдента. При этом в результате решения с использованием ППП Excel или Statistica исследователь получает значение t-критерия и уровень значимости p - вероятность гипотезы H0 о соответствии средних значений показателя. Формулы расчета t даны в [3, 4, 5] и др.

Надежные значения t-критерия можно получить по формуле:

$$t = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{(S_1^2 \times (n_1 - 1) + S_2^2 \times (n_2 - 1)) \times (n_1 + n_2)}{(n_1 + n_2 - 2) \times n_1 \times n_2}}}. \qquad (1.9)$$

Двухсторонний уровень значимости p рассчитывают по функции распределения t-критерия (рис.1.5).

При p≤0,05 - различие значимо; при p>0,05 - различие незначимо.

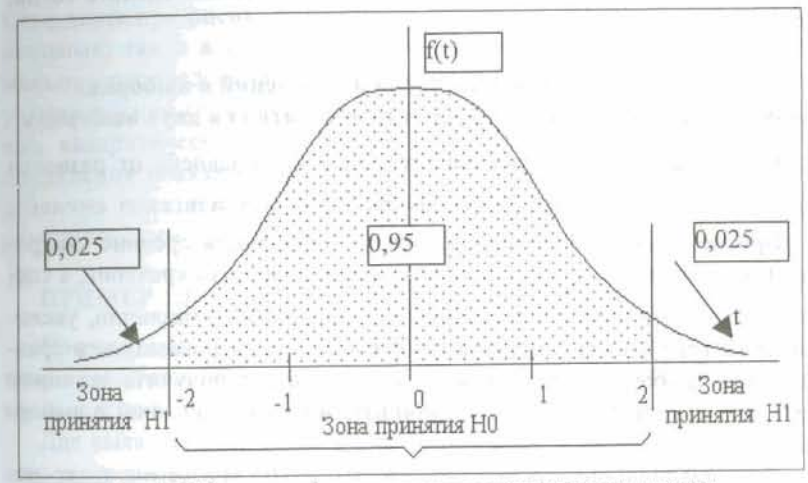

Рис.1.5.Величина двухстороннего уровня значимости p=0,025+0,025=0,05 отвечающего критерию t_{05}.

Оценка значимости различия показателей в связанных выборках

Связанными называются выборки, состоящие из одних и тех же объектов, наблюдающихся в различных условиях, например, до некоторого воздействия и после него, или в период разгара заболевания и на 3-й, 9-й и т.д. день лечения. В примере 1.1 связанными являются группы 2, 3, 4.

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

$$t = \frac{|\bar{\Delta}x|}{m_{\bar{\Delta}x}}, \qquad (1.10)$$

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

$m_{\bar{\Delta}x}$ – средняя квадратичная ошибка средней разности показателя, $m_{\bar{\Delta}x} = S_{\Delta x}/\sqrt{n}$, где $S_{\Delta x}$ – среднее квадратичное отклонение разности показателей.

По итогам расчета вывод о том, что различие показателя в сравниваемых парах связанных выборок значимо, можно сделать тогда, когда $p \le 0{,}05$.

Определение требуемого числа наблюдений в выборках для получения значимого различия показателя в двух выборках

Величина t-критерия значимости различия p зависит от разности $|\bar{x}_1 - \bar{x}_2|$, $\bar{\Delta}x$ и числа наблюдений в выборке n_1, n_2 и n.

При небольших объемах выборки увеличиваются средние квадратичные ошибки $m_{\bar{x}_1}, m_{\bar{x}_2}, m_{\bar{\Delta}x}$ и уменьшается величина t-критерия, а следовательно уменьшается вероятность гипотезы H1 о различии, увеличивается вероятность гипотезы H0 о соответствии показателя в сравниваемых выборках. При желании исследователя получить значимое различие показателя следует увеличивать число наблюдений в выборках.

Так, при заданном уровне значимости p=0,05 требуемое число наблюдений должно удовлетворять следующим требованиям:

– при сравнении независимых выборок

$$n_1 \text{ и } n_2 \ge \frac{t_{05}^2 \times (S_{\bar{x}_1}^2 + S_{\bar{x}_2}^2)}{(\bar{x}_1 - \bar{x}_2)^2}; \qquad (1.11)$$

– при сравнении связанных выборок

$$n \ge \frac{t_{05}^2 \times S_{\Delta x}^2}{(\bar{\Delta}x)^2}. \qquad (1.12)$$

Например, при получении незначимого различия показателя в независимых выборках:

1) $\bar{x}_1 = 10$, $S_1=5$, $n_1=25$;

2) $\bar{x}_2 = 12$, $S_1=6$, $n_1=25$,

для получения значимого различия с $p<0{,}05$ (при $t_{05}=2{,}00$ $n'=n_1+n_2-2=48$). Следует иметь число наблюдений:

$$n_1 \text{ и } n_2 \ge \frac{2^2 \times (5^2 + 6^2)}{(10-12)^2} = 61 \text{ набл.}$$

Таким образом, в каждой выборке надо иметь не менее 61 наблюдения. В этом случае возможно принятие гипотезы H1 о значимом различии показателей если таковая имеет право на существование.

В заключении заметим, что корректное применение t-критерия Стьюдента при оценке значимости различия показателя, как в независимых, так и в связанных выборках, можно получить при нормальном распределении показателя после расчета параметров этого распределения (средних значений, стандартных отклонений, средних квадратических ошибок). В случаях значимого отличия распределения показателя от нормального, задача оценки значимости различия показателя в сравниваемых выборках решается по непараметрическим критериям.

ПРИМЕР 1.1

Постановка задачи. Исследовали динамику нарушения ритма по типу желудочковой экстрасистолии у больных острым инфарктом миокарда при их комплексном лечении в условиях клиники.

Для выявления нарушений ритма наблюдался показатель - количество экстрасистол X (1/ч) с помощью ритмокардиоскопа РКС-02:

– в контрольной группе наблюдалось 15 больных ишемической болезнью сердца (ИБС);

– в опытной группе - 10 больных острым инфарктом миокарда на 1, 3 и 9-й день от начала развития острого инфаркта миокарда.

Таблица 1.1

Количество экстрасистол в группах X (1/ч)

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

***Требуется*:**

1. Определить числовые характеристики показателя в каждой группе.
2. Оценить значимость различий показателя в независимых и связанных выборках.
3. Сформулировать выводы.

Решение дано с помощью персонального компьютера с использованием электронной таблицы Microsoft Excel.

Числовые характеристики показателя X (1/ч) для четырех групп приведены в машинограмме 1.1.

Результаты расчетов t-критерия для оценки значимости различия показателя в контрольной и опытной группах - как независимых выборках: X1 и X2, X1 и X3, X1 и X4 приведены в машинограмме 1.2. Для оценки значимости различия показателя в опытной группе - как в связанных выборках X2 и X3, X2 и X4, X3 и X4 - в машинограмме 1.3. Итоговые результаты сведены в таблицы 1.2 и 1.3. Графическое представление - на рисунке 1.6.

Машинограмма 1.1

Числовые характеристики переменных

Числовые характеристики	Переменные			
	X1	X2	X3	X4
Среднее	3,27	30,20	13,40	4,70
Стандартная ошибка	0,64	2,39	1,63	0,73
Медиана	3	29,5	15	4,5
Мода	2	28	15	5
Стандартное отклонение	2,49	7,55	5,15	2,31
Дисперсия выборки	6,21	57,07	26,49	5,34
Эксцесс	-0,75	-0,80	-0,60	-1,37
Асимметричность	0,48	0,12	-0,84	0,39
Интервал	8	23	14	6
Минимум	0	19	5	2
Максимум	8	42	19	8
Счет	15	10	10	10

Машинограмма 1.2

Двухвыборочный t-тест с одинаковыми дисперсиями

Характеристики	X1	X2	X3	X4
Среднее арифметическое значение	3,3	30,2	13,4	4,7
Число наблюдений	15	10	10	10
df (число степеней свободы)		23	23	23
t-статистика		-12,91	-6,6	-1,449
P(T<=t) двустороннее		5E-12	1E-06	0,1608
t критическое двустороннее		2,0687	2,0687	2,0687

Таблица 1.2

Значение t-критерия и уровня значимости p при сравнении показателя X в контрольной и опытных группах

Сравниваемые группы	t-критерий	Уровень значимости	Выводы
X1 и X2	12,91	p<0,001	Различие значимо
X1 и X3	6,6	p<0,001	Различие значимо
X1 и X4	1,45	p>0,05	Различие незначимо

Машинограмма 1.3

Парный двухвыборочный t-тест для средних

Характеристики	X2 × X3		X2 × X4		X3 × X4	
Среднее	30,2	13,4	30,2	4,7	13,4	4,7
Наблюдения	10	10	10	10	10	10
df	9		9		9	
t-статистика	11,57		12,27		6,15	
P(T<=t) двусторонее	1,0E-06		6,4E-07		1,7E-04	
t критическое двустороннее	2,262		2,2622		2,262	

Таблица 1.3

Значение t-критерия и уровня значимости p при сравнении показателя X в опытных группах

Сравниваемые группы	t-критерий	Уровень значимости	Выводы
X2 и X3	11,57	p<0,001	Различие значимо
X2 и X4	12,27	p<0,001	Различие значимо
X3 и X4	6,15	p<0,001	Различие значимо

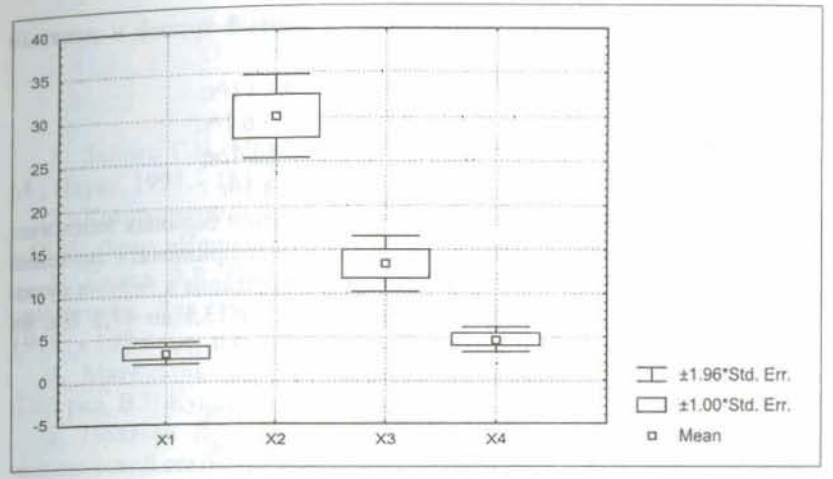

Рис.1.6. Средние значения показателя X с указанием 95% доверительных интервалов.

Выводы:

1.Из машинограммы 1.1 следует, что желудочковая экстрасистолия является патогномоничным признаком ишемической болезни сердца и острого инфаркта миокарда.

2.Среднее арифметическое значение числа экстрасистол у больных ишемической болезнью сердца составляет 3,27 в час. Встречаются больные, у которых за период наблюдения экстрасистолы не возникали, в то же время у некоторых больных число экстрасистол в час достигало 8. Размах вариационного ряда составил 8 экстрасистол. С развитием острого инфаркта миокарда среднее число экстрасистол увеличивается до 30,2 в час при минимальном их числе 19, а максимальном 42 в час и с размахом в 23 экстрасистолы. К третьему дню после возникновения инфаркта миокарда под воздействием комплексного лечения в условиях стационара среднее число экстрасистол уменьшилось до 13,4, минимальное их число составляло 5, максимальное 19, а размах 14. К девятому дню у больных этой группы среднее число экстрасистол уменьшилось до 4,7, минимальное их число составляло 2, максимальное 8, а размах 6.

3.95%-ые доверительные интервалы для истинных средних значений числа экстрасистол у больных ишемической болезнью сердца и

больных острым инфарктом миокарда на первый третий и девятый день лечения составит:

M_1=3,27±2,14×0,64=1,9 - 4,64 1/ч;

M_2=30,2±2,26×2,39=24,8 - 35,6 1/ч;

M_3=13,4±2,26×1,63=9,7 - 17,1 1/ч;

M_4=4,7±2,26×0,73=3,1 - 6,3 1/ч.

4.С вероятностью 95% можно утверждать, что у больных ишемической болезнью сердца число экстрасистол может принимать значение от 0 до 8,6 1/ч, у больных острым инфарктом миокарда в первые сутки число экстрасистол может принимать значение от 13,1 до 47,3 1/ч; на третьи сутки от 1,8 до 25,0 1/ч; на девятые от 0 до 9,9 1/ч.

X_1=3,27±2,14×2,49=0 - 8,6 1/ч;

X_2=30,2±2,26×7,55=13,1 - 47,3 1/ч;

X_3=13,4±2,26×5,15=1,8 - 25,0 1/ч;

X_4=4,7±2,26×2,31=0 - 9,9 1/ч.

5.Распределение показателя X во всех группах следует признать близким к нормальному т.к. имеет место примерное равенство средних значений (среднего арифметического и медианы), примерная симметричность минимальных и максимальных значений относительно среднего значения, коэффициенты асимметрии и эксцесса не превышают 2 по абсолютной величине. Следовательно, для оценки значимости различия показателя в группах можно применить параметрический t-критерий Стьюдента.

6.Показатель нарушения ритма – количество экстрасистол у больных острым инфарктом миокарда на 1-й и 3-й дни от начала его развития значимо увеличен по сравнению с этим показателем у больных ИБС ($p<0,001$). К девятому дню при комплексном лечении больных в условиях клиники количество экстрасистол существенно снижается и незначимо отличается от показателя в контрольной группе больных ИБС ($p>0,05$).

7.В динамике течения острого инфаркта миокарда при комплексном лечении больных в условиях клиники отмечается значимое уменьшение количества экстрасистол на 3-й и на 9-й дни по сравнению с 1-м и на 9-й день по сравнению с 3-м днем ($p<0,001$).

8.Полученные результаты свидетельствуют об эффективном воздействии комплексного лечения больных в условиях клиники на нарушение ритма при остром инфаркте миокарда.

Литература

1. Зайцев Г.Н. Математический анализ биологических данных. – М.: Наука, 1991. - 183 с.

2. Компьютерная биометрия: пакет CSS 3.1 /Под ред. В.Р.Лядова. – СПб.: Фонд «Инициатива», 1997. - 155 с.

3. Лядов В.Р. Основы теории вероятностей и математической статистики: Для студентов мед.ВУЗов. - СПб.: Фонд «Инициатива», 1998. - 107 с.

4. Математико-статистические методы в клинической практике /Под ред. В.И.Кувакина. - СПб.:Б.и, 1993. - 199 с.

5. Поляков Л.Е., Игнатович Б.И, Лашков К.В. Основы военно-медицинской статистики /Под ред. Л.Е.Полякова - Л.: Б.и, 1977. -336 с.

6. Урбах В.Ю. Статистический анализ в биологических и медицинских исследованиях. - М.: Медицина, 1975. - 295 с.

Глава 2. СТАТИСТИЧЕСКИЙ АНАЛИЗ КАТЕГОРИРОВАННЫХ ДАННЫХ

Задачи анализа категорированных данных медицинских исследований

Значительное число признаков, описывающих объекты медицинских исследований, как входных факторов, воздействующих на объект исследования, так и выходных параметров-откликов на воздействия, определяются качественно по номинальной шкале. Например, категории тяжести состояния (легкая, средняя, тяжелая, крайне тяжелая степень), пол, исход лечения (выжил, умер) и т.д. Данные о частотах наблюдения изучаемого признака и уровнях неколичественных переменных получили название категорированных. Такие данные сводятся в таблицы, получившие название частотных таблиц или таблиц сопряженности (см. табл.9.1 в главе Логлинейный анализ).

При наличии частотной таблицы исследователь может решать основные задачи исследования:

– определение относительных величин частоты наблюдений исследуемого признака и оценка их точности и надежности;

– проверка гипотез о значимости различия относительных величин частоты в различных группах, т.е. для различных категорий сочетаний уровней факторов;

– моделирование частот методами логлинейного анализа с целью их прогноза для различных сочетаний уровней факторов и др.

Постановка таких задач и порядок их решения будут рассмотрены в этой главе.

Относительные величины в медицинской статистике

Для характеристики заболеваемости, эффективности деятельности медицинских учреждений в медицинской статистике применяются относительные величины различного назначения.

1.Относительные величины частоты (интенсивные коэффициенты).

2.Относительные величины распределения, структуры (экстенсивные коэффициенты).

3.Относительные величины соотношения.

4.Относительные величины динамики изучаемых процессов и др.

Их назначение, сущность и порядок определения дан в [1].

Следует различать два понятия - частоту и частость. Под **частотой** понимают абсолютное число, показывающее, сколько раз (как часто) встречается в совокупности то или иное значение признака или, что то же самое, сколько единиц в совокупности обладают тем или иным значением признака. **Частость** - это относительная величина частоты, определяющая долю частот отдельных вариантов в общей сумме частот. Сумма всех частостей равна единице. Частости могут выражаться в процентах, промилле, продецимилле и т.д. Наиболее важной является относительная величина частоты случаев заболеваний, госпитализации, увольняемости, смертности, дней трудопотерь по болезни и т.п. Она, как правило, рассчитывается в промилле (‰), т.е. на 1000 человек жителей района или сотрудников предприятия из расчета на год. Например, в течение года на предприятии при средней численности сотрудников n наблюдалось m случаев заболеваний. Относительная величина частоты случаев заболеваний, которую принято называть уровнем заболеваемости, определяется в ‰, по формуле (2.1):

$$I = \frac{1000 \times m}{n}. \qquad (2.1)$$

Если наблюдение осуществлялось за период t, дн, то для того чтобы получить относительную величину частоты случаев заболеваний на 1000 человек из расчета на год, ее следует рассчитать по формуле (2.2):

$$I = \frac{365000 \times m}{n \times t}, \qquad (2.2)$$

где m - число случаев заболеваний за период t.

Например, на предприятии, работающем вахтовым методом в условиях Севера, численностью n=200 человек за 80 дней вахтового периода наблюдалось m=2 случая авитаминоза С, относительная величина частоты случаев заболеваний будет:

$$I = \frac{365000 \times 2}{200 \times 80} = 45{,}7^{0}/_{00}.$$

Полученный результат означает, что для этого предприятия следует ожидать 45,7 случаев заболеваний авитаминозом С из расчета на год на 1000 человек.

Такой порядок расчета необходим для правильного сравнения уровней заболеваемости в различных районах, на различных предприятиях и в различных условиях.

Определение относительных величин частоты по результатам выборочных наблюдений

В практике выборочных исследований часто изучаются неколичественные признаки, такие как осложнения заболевания, исходы лечения, оказание первой врачебной, квалифицированной или специализированной помощи и т.п. В таких исследованиях по выборке n наблюдений рассчитывается число случаев m интересующего признака и определяется относительная их частота (или частость)

$$\overline{P} = \frac{m}{n}. \qquad (2.3)$$

Порядок расчета относительной величины частоты летальных исходов приведен в примере 2.1. В таблице 2.1 представлены исходные данные, в таблице 2.2 в первой строке рассчитаны относительные величины частоты, в %.

В ряде случаев (особенно при малом числе наблюдений) число наблюдений изучаемого события m может оказаться равным либо 0, либо n и частость соответственно будет либо $\overline{P}$=0%, либо $\overline{P}$=100%.Такой результат при числе наблюдений до 10 должен квалифицироваться случайным и требующим коррекции на поправку, обоснованную Йетсом:

$$\delta p = \frac{100}{2n} \text{ (в \%)}. \qquad (2.4)$$

При получении m=0 относительная величина частоты принимается равной $\overline{P} = \delta p$, при получении m=n относительную величину частоты принимают $\overline{P} = 100 - \delta p$. Именно такая поправка для группы №4 с n=3 введена при определении $\overline{P}_4 = 100 - \delta p = 100 - \frac{100}{2 \times 3} = 100 - 16{,}7 = 83{,}3\%$ (табл.2.2).

Оценка точности и надежности относительных величин частоты

Вследствие случайности, допускаемой при формировании выборки, а также ограниченного ее объема, рассчитанные по выборке относительные величины частоты содержат погрешность. Ошибка относительной величины частоты следует закону нормального распределения и характеризуется средней квадратичной ошибкой

$$m_{\overline{p}} = \sqrt{\frac{\overline{P} \times (100 - \overline{P})}{n}}, (\text{в}\%) \cdot \qquad (2.5)$$

В научных работах принято указывать результат расчета относительной величины частоты, как $\overline{P} \pm m_{\overline{p}}$. Более целесообразно давать оценку точности и надежности относительной величине частоты 95%-м доверительным интервалом ее истинного значения

$$P = \overline{P} \pm t_{95} \times m_{\overline{p}}, \qquad (2.6)$$

где t_{95} - табличное значение t-критерия Стьюдента, отвечающее доверительной вероятности 95% и числу степеней свободы n'=n-1.

Оценка доверительного интервала по (2.6.) корректна при $25 \le \overline{P} \le 75\%$. Такой расчет доверительного интервала для относительной величины частоты летальных исходов в группе №3 дан в таблице 2.2. В случае $\overline{P} \le 25\%$ или $\overline{P} \ge 75\%$ более точная оценка точности и надежности дается с применением вспомогательной переменной Фишера в радианной мере:

$$\varphi = 2 \arcsin \sqrt{\overline{P}}, \qquad (2.7)$$

где $\overline{P}$ - относительная величина частоты от 0 до 1.

Ошибки переменной φ следуют закону нормального распределения и характеризуются средней квадратичной ошибкой:

$$m_{\varphi} = \frac{1}{\sqrt{n}}. \qquad (2.8)$$

95% доверительный интервал для истинного значения вспомогательной переменной определяют по (2.9).

$$\Phi = \varphi \pm t_{95} m_{\varphi}. \qquad (2.9)$$

От рассчитанных нижней и верхней границ доверительного интервала $\varphi_н$ и $\varphi_в$ переходят к соответствующим границам для относительной величины частоты по (2.10).

$$P_н = \sin^2 \frac{\varphi_н}{2}; \quad P_в = \sin^2 \frac{\varphi_в}{2}. \qquad (2.10)$$

Последовательность расчета 95% доверительных интервалов по вспомогательной переменной Фишера дана в примере 2.1 для групп №1, 2 и 4 – в табл.2.2 и иллюстрирована наглядно на рис.2.1.

Оценка значимости различия относительных величин частоты в независимых выборах по t-критерию Стьюдента

Сравнение относительных величин частоты и оценка значимости их различий в независимых выборках - одна из наиболее часто решаемых задач медицинскими исследователями. Исходными данными являются

результаты расчета относительных величин частоты и оценка их точности и надежности в двух независимых выборках.

К примеру, имеем:

Выборка №1: $n_1, \overline{P}_1, m_{\overline{P}_1}$, 95% доверительный интервал ($P_{1н}$, $P_{1в}$).

Выборка №2: $n_2, \overline{P}_2, m_{\overline{P}_2}$, 95% доверительный интервал ($P_{2н}$, $P_{2в}$).

На основе этих данных построен график доверительных интервалов. Требуется оценить значимость различия относительных величин частоты интересующего события в двух выборках. Приблизительное решение можно дать по графику, точное по результатам расчета t-критерия Стьюдента.

Если доверительные интервалы на графике не перекрываются или перекрываются на величину не более 1/3, можно утверждать, что имеется значимое различие относительных величин частоты. При большем перекрытии доверительных интервалов констатируется отсутствие значимого различия.

t-критерий рассчитывают по формуле (2.11)

$$t = \frac{|\overline{P}_1 - \overline{P}_2|}{\sqrt{m^2_{\overline{P}_1} + m^2_{\overline{P}_2}}} \qquad (2.11)$$

или по (2.12) при применении вспомогательной переменной Фишера

$$t = \frac{|\varphi_1 - \varphi_2|}{\sqrt{m^2_{\varphi_1} + m^2_{\varphi_2}}}. \qquad (2.12)$$

Гипотеза о значимом различии относительных величин частоты принимается при ее вероятности равной или большей 95%. Если вероятность этой гипотезы будет меньше 95%, принимается нулевая гипотеза об отсутствии значимого различия или о соответствии относительных величин частоты в двух выборках. Вероятность нулевой гипотезы при этом будет $p>0,05$. Величина этой вероятности p называется уровнем значимости. Решение о значимости различия относительных величин частоты в двух выборках принимают в результате сравнения рассчитанного значения t-критерия с критическими значениями t_{05}, t_{01}, t_{001}, которые берут из таблицы по соответствующим уровням значимости $p=0,05$; 0,01; 0,001 и числу степеней свободы $n'=n_1+n_2-2$.

Если $t<t_{05}$ - различие незначимо ($p>0,05$).

Если $t\geq t_{05}$ -различие значимо ($p\leq 0,05$).

Статистическая значимость различия возрастает, если $t>t_{01}$ или $t>t_{001}$, при этом уровни значимости будут соответственно $p<0,01$ и $p<0,001$.

Последовательность оценки значимости различия относительных величин частоты летальных исходов и варианты выводов даны в примере 2.1. Там же приведен расчет требуемого числа наблюдений в выборках для получения значимого различия относительных величин частоты по формуле (2.13)

$$n_{тр} \geq \frac{t^2_{05} \times [\overline{P}_1 \times (1-\overline{P}_1) + \overline{P}_2 \times (1-\overline{P}_2)]}{(\overline{P}_1 - \overline{P}_2)^2}, \qquad (2.13)$$

где $\overline{P}_1$ и $\overline{P}_2$ - относительная величина частоты от 0 до 1.

ПРИМЕР 2.1

Постановка задачи. Исследуется уровень летальности при различных формах острых гнойных деструкций легких (Вестник хирургии им.М.И.Грекова №1, 1986). В хирургической клинике сформированы данные о количестве наблюдений и случаев летальности для четырех форм острых гнойных деструкций легких (табл.2.1).

Таблица 2.1

Число случаев летальных исходов при острых гнойных деструкциях легких

Номер группы	Форма заболевания	Число больных	Число летальных исходов
1	Гнойный абсцесс	140	4
2	Гангренозный абсцесс	48	11
3	Гангрена доли	8	3
4	Тотальная гангрена	3	3

Требуется:

1.Определить относительные величины частоты (частости) летальных исходов; оценить их точность и надежность.

2.Построить график частости летальных исходов с указанием 95% доверительных интервалов.

3.Определить уровни значимости различия частостей летальных исходов для различных форм заболевания.

чимости различий уровней летальности.

5. Сформулировать выводы.

Решение выполнено с помощью ПК с использованием электронной таблицы Microsoft Excel 7.0.

1. Относительные величины частоты летальных исходов и оценки их точности и надежности приведены в табл.2.2.

Расчет проведен в следующем порядке. Для групп №1, №2, и №4, в которых относительные величины частоты оказались меньше 25 и больше 75%, потребовалось ввести переменную Фишера φ, а для группы №4 – поправку Йетса δρ.

Таблица 2.2

Относительные величины частоты летальных исходов и оценки их точности и надежности

Величины	Группы			
	№ 1	№ 2	№ 3	№ 4
Относительные величины частоты (ОВЧ) летальных исходов $\bar{P}$, %	2,9	22,9	37,5	100 поправка Йетса $\delta\rho$= 0,167, или 16,7% $\bar{P}_4$=83,3%
Средняя квадратическая ошибка ОВЧ летальных исходов m_p, %			17,1	
Переменная Фишера φ	0,342	0,998		2,300
Средняя квадратическая ошибка переменной Фишера m_φ	0,085	0,144		0,577
95% доверительный интервал для φ: $\varphi_н$ - $\varphi_в$	0,175-0,509	0,716-1,28		1,168-3,432
95% доверительный интервал для P: $P_н$—$P_в$, %	0,8-6,3	12,2-35,4	0 - 77,9	30,4-97,9

1-я группа. Гнойный абсцесс: n_1=140, m_1=4, $\bar{P}_1 = \frac{4}{140} = 0{,}029$, или 2,9%.

Т.к. $\bar{P}_1 < 25\%$, для оценки точности и надежности $\bar{P}_1$ следует применить вспомогательную переменную Фишера $\varphi_1 = 2 \arcsin \sqrt{0{,}029} = 0{,}342$

Средняя квадратическая ошибка $m_{\varphi_1} = \frac{1}{\sqrt{140}} = 0{,}085$.

95% доверительный интервал для φ_1 (при t_{95}=1,96):

φ_1=0,342±1,96×0,085=0,342±0,167 , или $\varphi_{1н}$= 0,175, $\varphi_{1в}$= 0,509.

95% доверительный интервал для P_1:

$P_{1н} = \sin^2 \frac{0{,}175}{2} = 0{,}008$, или 0,8% , $P_{1в} = \sin^2 \frac{0{,}509}{2} = 0{,}063$, или 6,3% .

С вероятностью 95% можно утверждать, что вероятность летальности при гнойном абсцессе легких находится в интервале от 0,8 до 6,3%.

2-я группа. Гангренозный абсцесс: n_2=48, m_2=11, $\bar{P}_2 = \frac{11}{48} = 0{,}229$, или 22,9 % .

Т.к. $\bar{P}_2 < 25\%$, для оценки точности и надежности $\bar{P}_2$ следует применить вспомогательную переменную Фишера

$\varphi_2 = 2\arcsin \sqrt{0{,}229} = 0{,}998$.

Средняя квадратическая ошибка $m_{\varphi_2} = \frac{1}{\sqrt{48}} = 0{,}144$.

95% доверительный интервал для φ_2 (при t_{95}=1,96):

φ_2 =0,998±1,96×0,144=0,998±0,282, или $\varphi_{2н}$= 0,716, $\varphi_{2в}$= 1,280.

95 % доверительный интервал для P_2:

$P_{2н} = \sin^2 \frac{0{,}716}{2} = 0{,}122$, или 12,2%, $P_{2в} = \sin^2 \frac{1{,}280}{2} = 0{,}354$, или 35,4% .

С надежностью 95% можно утверждать, что вероятность летальности при гангренозном абсцессе легких находится в интервале от 12,2 до 35,4%.

3-я группа. Гангрена доли: n_3=8, m_3=3, $\bar{P}_3 = \frac{3}{8} = 0{,}375$, или 37,5% .

Т.к. $\bar{P}_3 > 25\%$; для оценки точности и надежности $\bar{P}_3$ можно применить среднюю квадратическую ошибку частости

$$m_{\bar{P}_3} = \sqrt{\frac{0{,}375 \times (1-0{,}375)}{8}} = 0{,}171, \quad \text{при} \quad n'=7; \quad t_{95}=2{,}36$$

$P_3=0{,}375\pm2{,}36\times0{,}171=0{,}375\pm0{,}404.$

95% доверительный интервал для P_3:

$P_{3н}=0$, или 0%, $P_{3в}=0{,}779$, или 77,9%.

Доверительный интервал очень большой в связи с малым числом наблюдений.

4-я группа. Тотальная гангрена. $n_4=3$, $m_3=3$, $\bar{P}_4 = \frac{3}{3} = 1$, или 100%.

Требуется учесть поправку Йетса

$$\delta\rho = \frac{1}{2\times n} = \frac{1}{2\times 3} = 0{,}167.$$

Уточненная относительная величина частоты летальных исходов

$$\bar{P}_4 = 1-\delta\rho = 1-0{,}167 = 0{,}833, \text{ или } 83{,}3\%.$$

Т.к. $\bar{P}_4>75\%$, следует применить переменную Фишера $\varphi_4=2\arcsin\sqrt{0{,}833}=2{,}300$, $m_{\varphi_4} = \frac{1}{\sqrt{3}} = 0{,}577$.

95% доверительный интервал для φ_4:

$$\varphi_4=2{,}300\pm1{,}96\times0{,}577 = =2{,}300\pm1{,}132,$$
$$\text{или } \varphi_{4н}=1{,}168, \ \varphi_{4в}=3{,}432.$$

95% доверительный интервал для P_4:

$$P_{4н}=\sin^2\frac{1{,}168}{2}=0{,}304, \text{ или } 30{,}4\%,$$
$$P_{4в}=\sin^2\frac{3{,}432}{2}=0{,}979, \text{ или } 97{,}9\%.$$

Доверительный интервал очень большой в связи с малым числом наблюдений.

2. График частости летальных исходов с указанием 95% доверительных интервалов дан на рис.1.2.1.

3.Уровни значимости различия летальности для четырех форм гнойных деструкций легких даны в табл.1.2.3.

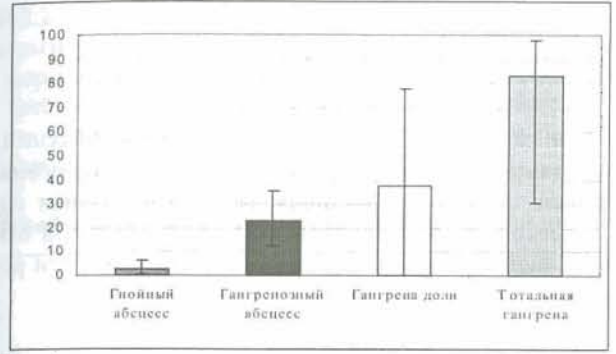

Рис.2.1. График относительных величин частоты летальных исходов для 4 форм заболеваний с указанием 95% доверительных интервалов.

Таблица 2.3

Уровни значимости p для сравнения относительных величин летальности в четырех группах

Сравниваемые группы	№ 2	№ 3	№ 4
№ 1	p<0,001	p<0,001	p<0,001
№ 2		p>0,05	p<0,01
№3			p>0,05

Выводы:

1.Уровень летальности имеет существенные различия при различных формах острых гнойных деструкций легких. Наименьший уровень летальности 2,9% получен при гнойном абсцессе (1-я группа), наибольший 100,0% – при тотальной гангрене (4-я группа). При гангренозном абсцессе уровень летальности – 22,9%, при гангрене доли – 37,5%.

2.Различие уровня летальности в 1-й группе по сравнению с тремя другими группами, а также уровня летальности во 2-й группе по сравнению с 4-й значимо (p<0,01).

3.Вследствие малого числа наблюдений различие уровней летальности во 2-й и 3-й группах и в особенности в 3-й и 4-й группах оказалось незначимым (p>0,05). Для получения значимого различия уровня летальности необходимо иметь число наблюдений в группах более $n_{тр}$:

для 2 и 3-й групп

$$n_{тр} \geq \frac{1{,}95^2[0{,}229(1-0{,}229)+0{,}375(1-0{,}375)]}{(0.229-0{,}375)^2} = 77;$$

для 3 и 4-й групп

$$n_{тр} \geq \frac{1{,}95^2[0{,}375(1-0{,}375)+0{,}833(1-0{,}833)]}{(0{,}375-0{,}833)^2} = 8.$$

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

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

$$\chi^2 = \sum \frac{(n_{1i} - n_{2i})^2}{n_{2i}}, \qquad 2.14$$

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

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

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

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

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

$$n'=(m-1)\times(s-1),$$

где m - число сравниваемых групп,

s - число уровней изучаемого признака.

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

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

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

ПРИМЕР 2.2

По данным исхода лечения больных при наличии у них одной из четырех форм острой гнойной деструкции легких (см. пример 2.1 табл.2.1) требуется оценить значимость различия между группами по числу случаев летальных исходов с помощью χ^2-критерия Пирсона. Исходные данные в таблице 2.4.

Таблица 2.4

Исходы лечения острых гнойных деструкций легких

Номер группы	Форма заболевания	Число случаев		Число больных
		летальных исх.	выздоровления	
1	Гнойный абсцесс	4	136	140
2	Гангренозный абсцесс	11	37	48
3	Гангрена доли	3	5	8
4	Тотальная гангрена	3	0	3

В таблице лишь в трех из восьми ячеек число наблюдений больше 5. Это указывает на то, что решение по χ^2 можно получить только приблизительно в отличие от точной оценки значимости различия групп по летальности, данной по t-критерию Стьюдента в табл.2.2 и 2.3 и по графику на рис.2.1.

χ^2=7,88+6,83+6,05+24,30+0,93+0,81+0,67+2,70=50,17

По числу степеней свободы n'=(4-1)×(2-1)=3 и уровням значимости p из таблицы критических значений:

-при p=0,05 χ^2_{05}=7,82;

-при p=0,01 χ^2_{01}=11,34;

-при p=0,001 χ^2_{001}=16,27.

Так как $\chi^2 > \chi^2_{001}$ различие числа летальных исходов для различных групп больных следует считать значимым (p<0,001).

Сделанный вывод имеет слишком общий характер и не содержит конкретных оценок при сравнении групп попарно. Отсюда следует, что оценка значимости различия групп по χ^2 может быть только предварительной и нуждается в уточнении.

Литература

1. Поляков Л.Е., Игнатович Б.И, Лашков К.В. Основы военно-медицинской статистики /Под ред. Л.Е.Полякова - Л.: Б.и, 1977. -336 с.

Таблица 2.5

Расчет χ^2-критерия Пирсона

№пп	Форма заболевания	Число летальных исходов			Число выздоровевших			Число больных
		наблюдавшееся n_{1i}	теоретическое n_{2i}	$(n_{1i}-n_{2i})^2$ / n_{2i}	наблюдавшееся n_{1i}	теоретическое n_{2i}	$(n_{1i}-n_{2i})^2$ / n_{2i}	
1	Гнойный абсцесс	4	14,8	7,88	136	125,2	0,93	140
2	Гангренозный абсцесс	11	5,1	6,83	37	42,9	0,81	48
3	Гангрена доли	3	0,8	6,05	5	7,2	0,67	8
4	Тотальная гангрена	3	0,3	24,03	0	2,7	2,7	3
	Всего: абсолютное число	21			178			199
	относительная величина	0,106			0,894			1,000

Глава 3. НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ ОЦЕНКИ ЗНАЧИМОСТИ РАЗЛИЧИЙ

Условия применения непараметрических методов

Непараметрические методы проверки статистических гипотез находят широкое применение в медицинских и биологических исследованиях. Они отличаются простотой проведения, для них не требуется вычислять какие-либо параметры распределения (средние значения, стандартные отклонения и др.).

Применение непараметрических методов статистического анализа целесообразно в следующих случаях:

–на этапе разведочного анализа;

–при малом числе наблюдений (до 30);

–когда нет уверенности в соответствии данных закону нормального распределения.

Однако, если данных много (например, $n > 100$), то не имеет смысла использовать непараметрические статистики.

По существу, для каждого параметрического критерия имеется, по крайней мере, один непараметрический аналог. Эти критерии можно отнести к одной из следующих групп:

–критерии различия между независимыми группами;

–критерии различия между зависимыми группами;

–критерии зависимости между переменными (изучение связи между переменными).

Различия между независимыми группами. Обычно, когда имеются две выборки (например, мужчины и женщины), которые необходимо сравнить относительно среднего значения некоторой изучаемой переменной, используется t-критерий Стьюдента для независимых выборок (см.главу 1). Непараметрическими альтернативами этому критерию являются: критерий серий Вальда-Вольфовица, U критерий Манна-Уитни и двухвыборочный критерий Колмогорова-Смирнова (примеры 3.1 –3.3).

Различия между зависимыми группами. Если есть необходимость сравнить две переменные, относящиеся к одной и той же выборке (например, биохимические показатели у больных с диагнозом гепатит А при поступлении в инфекционную клинику и перед выпиской из нее), то обычно используется t-критерий Стьюдента для связанных выборок

(см.главу 1). Альтернативными непараметрическими тестами являются: Z-критерий знаков и Т-критерий Вилкоксона парных сравнений (пример 3.4, 3.5).

Зависимости между переменными. Для того, чтобы оценить зависимость (связь) между двумя переменными, обычно вычисляется коэффициент корреляции Пирсона. Непараметрическими аналогами коэффициента корреляции Пирсона являются ранговые коэффициенты Спирмена R, тау Кендалла и коэффициент Гамма (см. главу 4. Ранговые коэффициенты корреляции). Если две рассматриваемые переменные по природе своей категорированы, подходящими непараметрическими критериями для тестирования зависимости будут: Хи-квадрат и точный критерий Фишера (примеры 2.2 и 3.6).

Примеры реализации непараметрические методов рассмотрим с помощью модуля Nonparametrics/Distrib ППП Statistica for Windows.

Проверка гипотезы о различии в независимых выборках

ПРИМЕР 3.1

Изучается систолическое артериальное давление (САД) (в мм рт.ст.) в двух однородных группах здоровых мужчин:

–лица с многолетним стажем работы в условиях нарушенного ритма сна и бодрствования (работа, связанная с ночными дежурствами) – группа 1;

–лица без нарушения суточного ритма сна и бодрствования – группа 2.

Требуется оценить значимость различия систолического артериального давления в двух независимых группах по критерию Вальда-Вольфовица. Исходные данные в таблице 3.1.

Таблица 3.1

Результаты измерения систолического артериального давления

№№ пп	ГРУППА	САД	№№ пп	ГРУППА	САД
1	1	90	11	1	145
2	1	95	12	2	110
3	1	100	13	2	115
4	1	105	14	2	115
5	1	120	15	2	122
6	1	135	16	2	122
7	1	135	17	2	125
8	1	135	18	2	125
9	1	140	19	2	130
10	1	140	20	2	150

Результаты решения с помощью процедуры Wald-Wolfowitz runs test в машинограмме 3.1.

Машинограмма 3.1

By variable ГРУППА
Group 1: 1 Group 2: 2

	Valid N	Valid N	Mean	Mean		
	Group 1	Group 2	Group 1	Group 2	Z	p-level
САД	11	9	121,82	123,78	-2,28	0,023

		No. of	No. of
Z adjstd	p-level	Runs	ties
2,043	0,041	6	0

Анализ результатов решения. Количество серий (No. of Runs) в упорядоченном по величине ряде значений показателя равно 6, уровень значимости различия p=0,023, а достоверность различия показателя в двух исследуемых группах 1-p=1-0,023=0,977 или 97,7%. Таким образом, многолетняя работа в условиях нарушения суточного ритма сна и бодрствования значимо влияет на повышение систолического артериального давления.

ПРИМЕР 3.2

Определяется содержание сиаловой кислоты (в единицах) в крови больных инфарктом миокарда, поступивших на стационарное лечение в срок до 3 дней (группа 1 – 7 человек) и позднее 6 дней (группа 2 – 12 человек) от начала заболевания.

Требуется оценить значимость различия содержания сиаловой кислоты в двух независимых группах по критерию Манна-Уитни. Исходные данные в таблице 3.2.

Таблица 3.2

Результаты определения содержания сиаловой кислоты

№№ пп	ГРУППА	СИАЛ_К	№№ пп	ГРУППА	СИАЛ_К
1	1	240	11	2	226
2	1	235	12	2	230
3	1	270	13	2	305
4	1	280	14	2	278
5	1	185	15	2	210
6	1	287	16	2	228
7	1	148	17	2	335
8	2	314	18	2	305
9	2	270	19	2	335
10	2	220			

Результаты решения с помощью процедуры Mann-Whitney U test приведены в машинограмме 3.2.

Машинограмма 3.2

Mann-Whitney U Test (pr_1_4_2.sta)
By variable ГРУППА
Group 1: 1 Group 2: 2

	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level
СИАЛ_К	57,5	132,5	29,5	-1,05644	0,290774

Z adjusted	p-level	Valid N Group 1	Valid N Group 2	2*1sided exact p
-1,05784	0,290138	7	12	0,299119

Анализ результатов решения. Критерий U=29,5, что соответствует уровню значимости p=0,29077 и достоверности различия в содержании сиаловой кислоты 1-p=1-0,29077=0,70923 или 70,9%. Следовательно различия в содержании сиаловой кислоты у больных инфарктом миокарда с различными сроками госпитализации незначимы (p>0,05).

ПРИМЕР 3.3

Изучается влияние на поглотительные способности ретикулоэндотелиальной системы витамина B_{12}, определяя величину конгорот-индекса у кроликов после 8 дневного введения препарата витамина B_{12} (группа - 1) и физиологического раствора (группа - 2). Сравнение двух распределений рассмотрим с помощью критерия Колмогорова-Смирнова. Исходные данные в таблице 3.3.

Таблица 3.3

Результаты определения значений конгорот-индекса у кроликов после введения им витамина B_{12} и физиологического раствора

№ пп	Группа	Конгорот	№ пп	Группа	Конгорот
1	1	28	14	2	40
2	1	29	15	2	48
3	1	33	16	2	50
4	1	34	17	2	50
5	1	35	18	2	51
6	1	36	19	2	53
7	1	39	20	2	55
8	1	48	21	2	59
9	1	50	22	2	60
10	1	53	23	2	60
11	1	54	24	2	62
12	1	57	25	2	84
13	1	59			

Результаты решения задачи в машинограмме 3.3.

Машинограмма 3.3

Kolmogorov-Smirnov Test (kolm_sm.sta)
By variable ГРУППА
Group 1: 1 Group 2: 2

	Max Neg Differnc	Max Pos Differnc	p-level	Mean Group 1	Mean Group 2
КОНГОРОТ	-0,5385	0	p < .10	42,692	56,000

Std.Dev. Group 1	Std.Dev. Group 2	Valid N Group 1	Valid N Group 2
11,093	10,821	13	12

Анализ результатов решения. Средняя величина конгоротиндекса в опытной группе составила 42,7%, а в контрольной – 56,0%. Уровень значимости различия распределения двух сравниваемых выборок p<0,1, а достоверность различия показателя более 90%. Следовательно с надежностью 90% можно утверждать, что введение витамина B_{12} способствует усилению поглотительной функции ретикулоэндотелиальной системы.

Проверка гипотезы о различии между зависимыми выборками

ПРИМЕР 3.4

У 12 работающих на ультразвуковых установках изучалось содержание сахара в крови натощак до работы и через три часа после работы. Исходные данные в таблице 3.4.

Таблица 3.4

Содержание сахара в крови обследованных натощак до работы и после 3 часов работы на ультразвуковых установках

№ пп	САХ_ДО	САХ_ПОС	№ пп	САХ_ДО	САХ_ПОС
1	112	54	7	64	66
2	82	67	8	70	66
3	101	96	9	88	48
4	72	59	10	81	50
5	79	79	11	66	61
6	82	76	12	88	61

Решение выполним с помощью непараметрического критерия знаков (Sign test). Результаты решения в машинограмме 3.4.

Машинограмма 3.4

Sign Test (pr_1_4_4.sta)

	No. of Non-ties	Percent v < V	Z	p-level
САХ_ДО & САХ_ПОС	11	9,0909	2,412091	0,015861

Анализ решения. Снижение уровня содержания сахара в крови через 3 часа работы на ультразвуковых установках по сравнению с его уровнем натощак существенное с уровнем значимости p=0,016, а достоверность различия 1-p=1-0,016=0,984 или 98,4%.

ПРИМЕР 3.5

Для сравнения двух методов определения времени свертываемости крови, каждая проба оценивается этими двумя методами:

по Бюркеру - появление нитей фибрина при комнатной температуре;

по Ли-Уайту - при опрокидывании пробирки в термостате при 37 градусах Цельсия кровь не выливается.

Исходные данные в таблице 3.5.

Таблица 3.5

Время свертывания крови при его определении двумя методами

№ пп	BURKER	LIWITE	№ пп	BURKER	LIWITE
1	10	10	7	5	6
2	9	8	8	5	6
3	8	9	9	6	7
4	8	10	10	6	7
5	7	6	11	7	9
6	7	10			

Решение задачи осуществлено с помощью критерия Вилкоксона, который является непараметрической альтернативой t-критерию Стьюдента для парных сравнений количественных данных в зависимых выборках. Результаты решения в машинограмме 3.5.

Машинограмма 3.5

Wilcoxon Matched Pairs Test (t_w.css)

	Valid N	T	Z	p-level
BURKER & LIWITE	11	8	1,988	0,047

Анализ результатов решения. Уровень значимости различия исследуемого показателя p= 0,047, а достоверность его различия 1-p=1-0,047=0,953 или 95,3%, что свидетельствует о значимом различии времени свертывания крови при использовании исследуемых методов.

Можно предположить, что механизм методов связан с различными звеньями процесса свертывания крови.

Оценка значимости различия частот наблюдений по четырехпольной таблице с помощью χ^2-критерия Пирсона

При сравнении двух независимых групп по альтернативному признаку, принимающему два значения (либо есть, либо нет) исходные данные о числе наблюдений сводятся в четырехпольную таблицу. При обозначении ячеек частотной таблицы так, как показано в таблице 1.3.6, расчет критерия можно выполнить по формуле (3.1).

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}. \qquad 3.1$$

Критические значения критерия в этом случае надо брать по числу степеней свободы

$$n' = (m-1)\times(s-1) = (2-1)\times(2-1) = 1.$$

ПРИМЕР 3.6

По данным об исходах лечения острых гнойных деструкций легких в виде гнойных и гангренозных абсцессов (таблица 3.6) необходимо дать оценку значимости различия групп по летальности с помощью χ^2-Пирсона. Решение дано в таблице 3.7 и машинограмме 3.6.

Таблица 3.6

Число случаев летальных исходов при острых гнойных деструкциях легких

Номер группы	Форма заболевания	Число больных	Число летальных исходов
1	Гнойный абсцесс	140	4
2	Гангренозный абсцесс	48	11

Таблица 3.7

Расчет χ^2-критерия Пирсона

Номер группы	Число случаев		Всего больных
	летальных исходов	выздоровления	
1	a 4	b 136	a+b 140
2	c 11	d 37	c+d 48
	a+c **15**	b+d **173**	n=a+b+c+d **188**

$$\chi^2 = \frac{(4\times 37 - 136\times 11)^2 \times 188}{140\times 48\times 15\times 173} = 19{,}62$$

Из таблицы по n'=1 и p=0,05 χ^2_{05}=3,84;

p=0,01 χ^2_{01}=6,64;

p=0,001 χ^2_{001}=10,83.

Так как $\chi^2 > \chi^2_{001}$ различие групп №1 и №2 по летальности значимо (p<0,001).

Простота решения приведенного примера не исключает применения ПК и его программного обеспечения. Расчет критерия χ^2-Пирсона реализован в модуле Nonparametrics/Distrib. ППП Statistica процедурой 2×2 Tables XI/VI/Phil, McNemar, Fisher exact. Исходные данные задаются натуральными значениями частоты наблюдений в четырехпольную таблицу.

Результаты - в машинограмме 3.6, в которой даны как абсолютные, так и относительные величины распределения больных по видам деструкции и по исходам, а также различные критерии значимости различия частоты летальных исходов. В частности по критерию χ^2-критерию Пирсона (Chi-square) уровень значимости различия p=0,0000, достоверность 1-p=1-0,000=1 или практически 100%. Следует отметить, что при частоте изучаемого события менее 5 наблюдений в одной из ячеек использование χ^2-критерия Пирсона является не корректным. В таком случае необходимо воспользоваться точным критерием Фишера (Fisher exact), который в нашем случае демонстрирует уровень значимости различия p=0,0001.

Машинограмма 3.6

2 x 2 Table (pr_1_4_6.sta)

	Column 1	Column 2	Row Totals
Frequencies, row 1	4	136	140
Percent of total	2,13%	72,34%	74,47%
Frequencies, row 2	11	37	48
Percent of total	5,85%	19,68%	25,53%
Column totals	15	72,34%	74,47%
Percent of total	7,98%	92,02%	
Chi-square (df=1)	19,59	p= ,0000	
V-square (df=1)	19,49	p= ,0000	
Yates corrected Chi-square	16,95	p= ,0000	
Phi-square	0,1042		
Fisher exact p, one-tailed		p= ,0001	
two-tailed		p= ,0001	
McNemar Chi-square (A/D)	24,98	p= ,0000	
Chi-square (B/C)	104,6	p= ,0000	

Получено адекватное решение, как и по t-критерию Стьюдента (см. пример 2.1). Таким образом, применение критерия χ^2-Пирсона для сравнения групп по четырехпольной таблице целесообразно.

О выборе непараметрического метода оценки значимости различия

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

Литература

1. Боровиков В. STATISTICA: искусство анализа данных на компьютере. Для профессионалов. – СПб,: Питер, 2001. –656 с.: ил.

2. Компьютерная биометрия: пакет CSS 3.1 /Под ред. В.Р.Лядова. – СПб.: Фонд «Инициатива», 1997. - 155 с.

3. Лядов В.Р. Основы теории вероятностей и математической статистики: Для студентов мед.ВУЗов. - СПб.: Фонд «Инициатива», 1998. - 107 с.

4. Статистические методы исследования в медицине и здравоохранении, под редакцией Л.Е.Полякова. Л.,: Медицина, 1971. –200 с.

5. Юнкеров В.И. Основы математико-статистического моделирования и применения вычислительной техники в научных исследованиях: Лекции для адъюнктов и аспирантов / Под ред. В.И.Кувакина. - СПб, 2000. –140 с.

Глава 4. ОДНОФАКТОРНЫЙ КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Сущность функциональной и корреляционной связи

Одной из важных задач медицинского исследования является изучение связи между фактором, воздействующим на организм, и параметром-откликом на это воздействие, а также моделирование этого параметра в зависимости от действующего фактора. Эта задача решается методами корреляционного и регрессионного анализа.

Связь между переменными величинами может быть функциональной и вероятностной или корреляционной. При функциональной связи заданному значению фактора X соответствует строго определенное значение параметра Y, что свойственно строго детерминированным процессам (связь температуры и объема, давления и объема).

При корреляционной связи заданному значению фактора X может соответствовать множество возможных значений параметра Y. Например, заданному уровню потребления пресной воды на санитарно-бытовые нужды x в л/чел.сут. в n населенных пунктах соответствует множество значений уровня общей заболеваемости y в $^0/_{00}$ (рис.4.1). При этом отмечается, что с ростом x наблюдается уменьшение y. Это - обратная, отрицательная корреляционная связь.

Существует и прямая, положительная корреляционная связь, когда с увеличением фактора X возрастает параметр Y. Примером такой связи является возрастание уровня инфекционной заболеваемости Y в $^0/_{00}$ при увеличении плотности рабочих мест в производственном помещении X, чел (рис.4.2).

Линейная тенденция изменения параметра Y при изменении фактора X на рис.4.1 – 4.2 показана прямой, называемой линией регрессии (отклика).

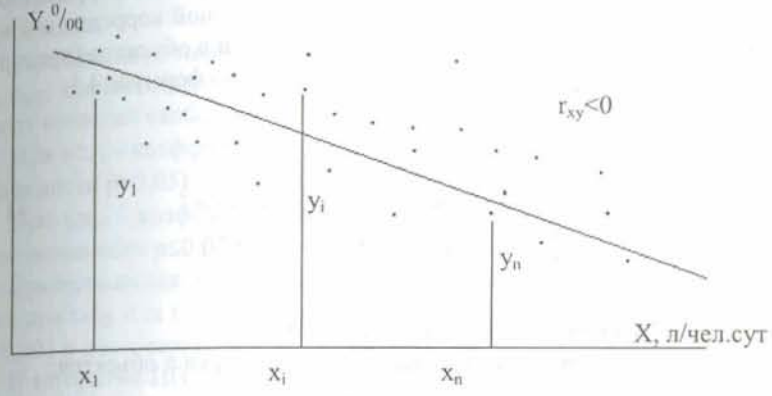

Рис.4.1. Поле наблюдений ($i = \overline{1,n}$) при обратной корреляционной связи между фактором X и параметром Y.

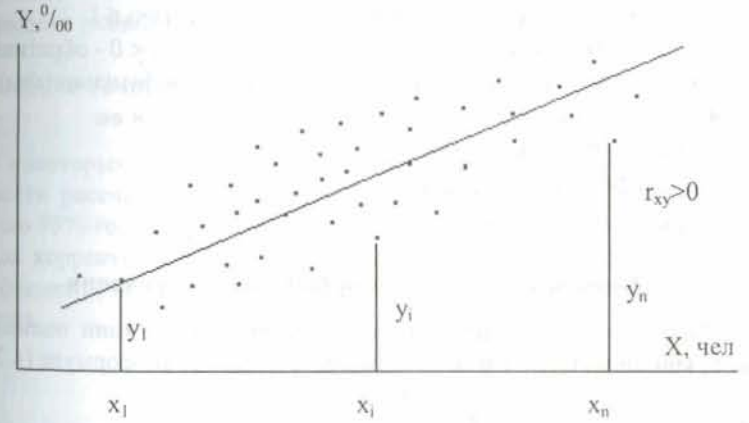

Рис.4.2. Поле наблюдений ($i=\overline{1,n}$) при прямой корреляционной связи между фактором X и параметром Y.

Коэффициент корреляции и его свойства

Направление (прямая или обратная) и сила (теснота) корреляционной связи характеризуется коэффициентом линейной корреляции Пирсона, который рассчитывают по данным выборки n объектов (предприятий, дошкольных учреждений, больных и т.д.) по формуле 4.1.

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}}, \quad (4.1)$$

где x_i, y_i - значения переменных для i-го объекта;

$\bar{x}$, $\bar{y}$ - средние значения переменных для выборки n объектов;

n - количество наблюдений в выборке.

В качестве исходной для расчета коэффициента корреляции является матрица наблюдений с размером n×2 (см. пример 4.1).

Свойства коэффициента корреляции.

1. Коэффициент корреляции величина относительная; он принимает значение от минус единицы до плюс единицы, т.е. $-1 \le r_{xy} \le 1$.
2. При $r_{xy} > 0$ связь оценивается, как прямая, при $r_{xy} < 0$ - обратная.
3. При $r_{xy} = 0$ - связь отсутствует, при $|r_{xy}| = 1$ - связь функциональная.
4. Сила связи оценивается:

– при $|r_{xy}| < 0{,}3$ - как слабая,

– при $0{,}3 \le |r_{xy}| \le 0{,}7$ - умеренная,

– при $|r_{xy}| > 0{,}7$ - сильная.

Оценка значимости коэффициента корреляции

Достоверность, значимость коэффициента корреляции оценивают по t - критерию Стьюдента, который рассчитывают по формуле (4.2):

$$t = \frac{|r_{xy}|}{\sqrt{\frac{1 - r_{xy}^2}{n-2}}}, \quad (4.2)$$

где $\sqrt{\frac{1 - r_{xy}^2}{n-2}} = m_r$, есть средняя квадратичная ошибка коэффициента корреляции.

Рассчитанные значения t-критерия сравнивают с критическими t_{05}, t_{01}, t_{001}, соответствующими уровням значимости p=0,05; 0,01; 0,001 и числу степеней свободы n'=n-2.

При $t < t_{05}$ - коэффициент корреляции считают незначимым (уровень значимости $p > 0{,}05$).

При $t \ge t_{05}$ - коэффициент корреляции считается значимым (с уровнем значимости $p \le 0{,}05$, достоверностью $1\text{-}p \ge 0{,}95$).

Статистическая значимость коэффициента корреляции увеличивается при $t > t_{01}$ или $t > t_{001}$ (с уровнем значимости соответственно $p < 0{,}01$ и $p < 0{,}001$ и достоверностью $> 0{,}99$; $> 0{,}999$).

В алгоритме ППП Statistica 5.0 оценка достоверности коэффициента корреляции дается по F-критерию Фишера, связанного функционально с t-критерием.

Так, в примере 4.1, в машинограмме 4.2 видно, что доза облучения Y (Гр) прямо, сильно и значимо связана с долей аберрантных клеток костного мозга X (%), т. к. коэффициент корреляции R=0,97 с уровнем значимости $p < 0{,}00000$, т. е. с достоверностью близкой 1.

Оценка точности и надежности коэффициента корреляции по вспомогательной переменной Фишера

В некоторых исследованиях обращаются к оценке точности и надежности рассчитанного коэффициента корреляции, т.е. к определению его 95%-го доверительного интервала.

Для корректного решения этой задачи от рассчитанного значения коэффициента корреляции переходят к вспомогательной переменной Фишера

$$z = \frac{1}{2} \ln \frac{1 + r_{xy}}{1 - r_{xy}}, \quad (4.3)$$

которая имеет нормальное распределение со средней квадратичной ошибкой

$$m_r = \frac{1}{\sqrt{n-3}}. \quad (4.4)$$

Далее определяют 95 %-й доверительный интервал для Z:

$$Z = z \pm 1,96 \times m_z, \quad (4.5)$$

а по нижней и верхней границам этого интервала находят соответствующие границы доверительного интервала для коэффициента корреляции по формуле (4.6):

$$r_{xy} = \frac{e^{2z} - 1}{e^{2z} + 1}. \quad (4.6)$$

Например, при решении задачи идентификации объектов по выборке n =20 получен коэффициент корреляции r_{xy}=0,87. Соответствующее ему значение переменной $z = \frac{1}{2}\ln\frac{1+0,87}{1-0,87} = 1,333$, средней квадратичной ошибки $m_z = \frac{1}{\sqrt{20-3}} = 0,234$, и 95%-го доверительного интервала Z=1,333±1,96×0,234=0,857÷1, 809.

Переходя от нижней границы переменной Z 0,857 и верхней 1,809 к соответствующим границам коэффициента корреляции по (4.6), получим:

$$r_{xyн} \frac{e^{2\cdot 0,857} - 1}{e^{2\cdot 0,857} + 1} = 0,70;$$

$$r_{xyв} = \frac{e^{2\cdot 1,809} - 1}{e^{2\cdot 1,809} + 1} = 0,95.$$

Таким образом, 95%-й доверительный интервал для коэффициента корреляции будет (0,70÷0,95).

При некорректном определении 95%-го доверительного интервала для коэффициента корреляции по его средней квадратичной ошибке $m_z = \sqrt{\frac{1-r_{xy}^2}{n-2}}$ можно получить верхнюю границу больше единицы.

Так, для этого примера $m_r = \sqrt{\frac{1-0,872}{20-2}} = 0,116$; 95%-й доверительный интервал при t_{05}=2,10 (по n'=20-2=18) r_{xy}=0,87±2,10×0,116=0,63÷1,11.

Это решение абсурдно, оно не соответствует правильному решению, полученному с помощью переменной Z Фишера.

Ранговые коэффициенты корреляции

При нелинейности связи между признаками, отсутствии данных о характере их распределения, небольшом числе наблюдений сравниваемых пар признаков, а также в случаях, когда эти признаки носят приближенный количественный или порядковый характер, целесообразно использовать непараметрические коэффициенты связи - коэффициент ранговой корреляции Спирмена или коэффициент ранговой корреляции Кендэла.

Идея коэффициента Спирмена проста. Нужно упорядочить данные по возрастанию и заменить реальные значения их рангами. Рангом значения называется его номер в упорядоченном ряду. Например, в ряду 1, 4, 8, 8, 12 ранг числа 1 равен 1, 4 – 2, 8 и 8 по 3,5, а 12 - 5. Затем, беря вместо самих значений их ранги, рассчитывается коэффициент ранговой корреляции Спирмена, который обозначается - r_s.

$$r_s = 1 - \frac{6\Sigma d^2}{n^3 - n}, \quad (4.7)$$

где d - разность рангов для каждого объекта выборки.

Достоверность коэффициента ранговой корреляции Спирмена оценивается на основе рассчитанного t-критерия Стьюдента (4.2).

Нулевая гипотеза о незначимости коэффициента ранговой корреляции Спирмена отвергается, если вычисленный t-критерий превысит значение t-критерия, указанное в таблице для выбранного уровня значимости и числа степеней свободы n'=n-2.

Вариант расчета коэффициента ранговой корреляции Спирмена приведен в примере 4.2.

Коэффициент и уравнение регрессии

Важной задачей изучения связи между фактором, воздействующим на объект, и параметром-откликом, является построение модели для параметра Y в зависимости от входного фактора X.

Модель для параметра y=f(x) может быть построена методом регрессионного анализа. Простейшей является линейная модель - уравнение регрессии вида:

$$\hat{y} = a + bx, \quad (4.8)$$

где $\hat{y}$ - прогнозируемое значение параметра Y;

a - свободный член;

b - коэффициент регрессии.

Свободный член определяется по формуле:

$$a = \bar{y} - b\bar{x}, \tag{4.9}$$

где $\bar{x}, \bar{y}$ - средние значения фактора X и параметра Y по выборке n наблюдений.

Коэффициент регрессии рассчитывается по формуле:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x\,y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{4.10}$$

Коэффициент регрессии показывает на сколько в среднем изменяется параметр Y при изменении фактора X на одну единицу.

Уравнение (4.8) является уравнением прямой линии регрессии (см. рис.4.1 и 4.2). По этому уравнению:

–изучают характер изменения параметра Y при изменении фактора X;

–прогнозируют значение параметра Y при заданном значении фактора X;

–определяют оптимальное значение фактора X для получения параметра Y на требуемом уровне.

Моделирование связи входного фактора и выходного параметра может быть выполнено на ПК с помощью процедуры «Регрессия» модуля «Анализ данных» ППП Excell, а также с помощью процедуры «Correlation matrices» из модуля «Basic Statistics» или модуля «Multiple Regression» ППП Statistica.

Оценка значимости коэффициентов уравнения регрессии

Коэффициенты уравнения регрессии, рассчитанные по случайно сформированной выборке ограниченного объема, содержат погрешность и нуждаются в оценке значимости (достоверности). Такая оценка дается по t-критерию Стьюдента. В модуле ППП Statistica 5.0 - предусмотрен расчет средних квадратичных ошибок коэффициентов m_a и m_b и на их основе критерия

$$t_a = \frac{a}{m_a}; t_b = \frac{b}{m_b} \tag{4.11}$$

и соответствующих им уровней значимости.

Значимыми признаются коэффициенты с уровнем значимости $p \le 0{,}05$ (достоверностью $1\text{-}p \ge 0{,}95$).

В примере 4.1. и в машинограмме 4.2. даны значения коэффициентов уравнения регрессии, критерия t и уровня значимости p: a=0,045, m_a=0,202, t=0,224, p=0,826; b=0,052, m_b=0,003; t=15,314 и p=0,000.

Из этих данных следует, что свободный член a незначим и его не следует включать в модель; коэффициент регрессии b значим. Поэтому модель для параметра Y будет содержать только линейный эффект фактора X:

$$\hat{y} = 0{,}052\,x\,.$$

График линейной регрессии дан на рис.4.1, 4.2 и 4.3.

Дисперсионный анализ, оценка информативности и значимости уравнения регрессии

Для оценки информативности и значимости модели выполняется ее дисперсионный анализ. В результате дисперсионного анализа рассчитывают:

- коэффициент детерминации

$$R^2 = \frac{SS_R}{SS}, \tag{4.12}$$

где SS_R - сумма квадратов отклонений рассчитанных значений $\hat{y}_i$ от среднего $\bar{y}$;

SS - сумма квадратов отклонений наблюдавшихся значений y_i от среднего $\bar{y}$.

- F - критерий Фишера

$$F = \frac{S_R^2}{S_0^2}, \tag{4.13}$$

где S_R^2 - дисперсия отклонений $\hat{y}_i$ от среднего $\bar{y}$;

S_0^2 - дисперсия отклонений y_i от $\hat{y}_i$.

Модель считают информативной при $R^2 > 0{,}5$ и значимой, достоверной при уровне значимости по F-критерию $p \le 0{,}05$.

В примере 4.1 в машинограмме 4.2 приведена таблица дисперсионного анализа уравнения регрессии. По данным таблицы видно, что R^2=0,947, а уровень значимости по F-критерию p=1,07×E-09. Из этого следует, что модель информативна ($R^2 > 0{,}5$) и значима (достоверность

1-р близка 1). Модель можно применять для решения всех задач исследования.

Прогноз по уравнению регрессии и оценка его точности и надежности

Для прогноза параметра Y при заданном значении фактора X можно применить либо аналитическое выражение модели, либо графики линии регрессии.

Прогноз содержит погрешность и поэтому нуждается в оценке точности и надежности, т.е. в расчете 95%-го доверительного интервала:

$$y_k = \hat{y}_k \pm t_{05} \times m_{\hat{y}_k}, \qquad (4.14)$$

где: $\hat{y}_k$ - прогнозируемое значение параметра при заданном значении фактора x_k;

$m_{\hat{y}_k}$ - средняя квадратичная ошибка прогноза, определяемая по формуле:

$$m_{\hat{y}_k} = S_0 \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}, \qquad (4.15)$$

где S_0 - среднее квадратичное отклонение параметра вследствие ошибок модели.

Так в примере 4.1 S_0=0,36 (см. машинограмму 4.2), $m_{\hat{y}_k}$ = 0,09, 95%-й доверительный интервал значений Y_k = (2,4÷2,8) Гр.

95% -й доверительный интервал для прогнозируемых значений параметра Y указан на рис. 4.3.

Вариант решения задачи корреляционного анализа и построения математической модели в виде уравнения регрессии, а также дисперсионный анализ этого уравнения и оценки его информационной способности и статистической значимости приведен в примере 4.1.

Особенности построения нелинейных уравнений регрессий

Ранее в этой главе рассмотрена линейная корреляционная связь и линейное однофакторное уравнение регрессии. На практике нередко встречается нелинейная зависимость моделируемого параметра Y от воздействующего фактора X. Нелинейное изменение параметра Y, по

желанию исследователя, можно описать функциями: показательной (экспоненциальной)

$$y=e^{a+bx}, \qquad (4.16)$$

степенной

$$y=a\times x^b, \qquad (4.17)$$

обратной (реципрокальной)

$$y=a+\frac{b}{x} \qquad (4.18)$$

и другими.

Для построения нелинейных моделей для параметра Y можно применять:

–модуль Multiple Regression с нестандартными опциями Fixed nonlinear построения нелинейной модели при трансформировании переменных Y и X соответствующими функциями, например, для фактора X: x^2, x^3, x^4, x^5, $\sqrt{x}$, lnx, lgx, e^x, 10^x, $\frac{1}{x}$;

–модуль Nonlinear Estimation, предлагающий построения моделей:

–user-specified regression - по заданию исследователя,

–logistic regression - логистическую модель,

–probit regression - модель линии пробитов,

–exponential growth regression - экспоненциальную модель,

–piecewise linear regression - линейно кусочную (разрывную) модель.

Из множества возможных видов нелинейного моделирования рассмотрим построение наиболее часто применяемых экспоненциальной и степенной моделей.

Сущность решения заключается:

–в преобразовании исходных данных о факторе X и параметре Y путем логарифмирования, в результате этого нелинейная модель приводится к линейной;

–в решении задачи с преобразованными данными по модулю линейного регрессионного анализа с получением коэффициентов и дисперсионного анализа модели;

–в обратном преобразовании коэффициентов путем потенцирования с целью построения требуемой нелинейной модели.

Так, для построения экспоненциальной модели $y=e^{a+bx}$ проведем ее логарифмирование:

$$\ln y = a + bx. \quad (4.19)$$

Обозначая $Z=\ln y$, получим линейное уравнение:

$$Z = a + bx \quad (4.20)$$

Это уравнение нетрудно построить по модулю линейной регрессии, вводя в качестве исходных данных:

— зависимая переменная $Z=\ln y$;

— независимая переменная x.

Получив коэффициенты a и b в модели $\hat{z} = a + bx$, путем потенцирования переходят к требуемой экспоненциальной модели $\hat{y} = e^{a+bx}$.

Для построения степенной модели $y=a\times x^b$ также проводят ее логарифмирование

$$\ln y = \ln a + b \times \ln x.$$

Обозначая $Z=\ln y$; $b_0=\ln a$; $t=\ln x$, получим линейное уравнение:

$$Z = b_0 + b \times t,$$

где Z - зависимая переменная $Z=\ln y$;

t - независимая переменная $t=\ln x$;

b_0 - свободный член $b_0=\ln a$;

b - коэффициент регрессии.

По модулю линейной регрессии определяют коэффициент b_0 и b, оценивают их значимость, проводят дисперсионный анализ и оценку эффективности модели: $\hat{z} = b_0 + b \times t$.

Путем потенцирования этого уравнения находят требуемую степенную модель: $y=a\times x^b$.

Последовательность построения степенной модели приведена в примере 4.3.

ПРИМЕР 4.1

Постановка задачи. Исследуется связь между поглощенной дозой облучения Y, Гр и долей аберрантных клеток костного мозга X, % у подопытных животных. С целью построения модели для определения поглощенной дозы облучения по доле аберрантных клеток костного мозга с 15 подопытными животными (белые мыши) проведен эксперимент (кафедра токсикологии ВМедА, 1990), результаты которого даны в табл.4.1.

Таблица 4.1

Доля аберрантных клеток X, % и доза облучения Y, Гр в эксперименте

Номер наблюдения	Доля аберрантных клеток костного мозга X, %	Доза облучения Y, Гр
1	59	3,2
2	44	2,5
3	85	4,5
4	70	4,0
5	52	3,0
6	21	0,8
7	26	1,3
8	79	4,0
9	41	3,1
10	67	3,5
11	32	1,8
12	18	0,7
13	90	4,3
14	12	0,3
15	100	5,0

Требуется определить:

1. Числовые характеристики переменных.
2. Коэффициент корреляции r_{xy} и оценить его точность и надежность.
3. Коэффициенты модели $\hat{y} = a + bx$ и оценить их достоверность.
4. Дать дисперсионный анализ и оценить достоверность и эффективность модели.
5. Построить график линии регрессии с указанием 95 % доверительных интервалов для возможных значений и среднего ожидаемого значения параметра.
6. Дать прогноз дозы облучения при количестве аберрантных клеток костного мозга $X_k = 50\%$. Оценить его точность и надежность.
7. Сформулировать выводы.

Решение дано с помощью персонального компьютера с использованием ППП Statistica 5.0.

1-р близка 1). Модель можно применять для решения всех задач исследования.

Прогноз по уравнению регрессии и оценка его точности и надежности

Для прогноза параметра Y при заданном значении фактора X можно применить либо аналитическое выражение модели, либо графики линии регрессии.

Прогноз содержит погрешность и поэтому нуждается в оценке точности и надежности, т.е. в расчете 95%-го доверительного интервала:

$$y_k = \hat{y}_k \pm t_{05} \times m_{\hat{y}_k}, \qquad (4.14)$$

где: $\hat{y}_k$ - прогнозируемое значение параметра при заданном значении фактора x_k;

$m_{\hat{y}_k}$ - средняя квадратичная ошибка прогноза, определяемая по формуле:

$$m_{\hat{y}_k} = S_0 \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}, \qquad (4.15)$$

где S_0 - среднее квадратичное отклонение параметра вследствие ошибок модели.

Так в примере 4.1 S_0=0,36 (см. машинограмму 4.2), $m_{\hat{y}_k}$ = 0,09, 95%-й доверительный интервал значений Y_k = (2,4÷2,8) Гр.

95% -й доверительный интервал для прогнозируемых значений параметра Y указан на рис. 4.3.

Вариант решения задачи корреляционного анализа и построения математической модели в виде уравнения регрессии, а также дисперсионный анализ этого уравнения и оценки его информационной способности и статистической значимости приведен в примере 4.1.

Особенности построения нелинейных уравнений регрессий

Ранее в этой главе рассмотрена линейная корреляционная связь и линейное однофакторное уравнение регрессии. На практике нередко встречается нелинейная зависимость моделируемого параметра Y от воздействующего фактора X. Нелинейное изменение параметра Y, по

желанию исследователя, можно описать функциями: показательной (экспоненциальной)

$$y=e^{a+bx}, \qquad (4.16)$$

степенной

$$y=a\times x^b, \qquad (4.17)$$

обратной (реципрокальной)

$$y=a+\frac{b}{x} \qquad (4.18)$$

и другими.

Для построения нелинейных моделей для параметра Y можно применять:

–модуль Multiple Regression с нестандартными опциями Fixed nonlinear построения нелинейной модели при трансформировании переменных Y и X соответствующими функциями, например, для фактора X: x^2, x^3, x^4, x^5, $\sqrt{x}$, lnx, lgx, e^x, 10^x, $\frac{1}{x}$;

–модуль Nonlinear Estimation, предлагающий построения моделей:

–user-specified regression - по заданию исследователя,

–logistic regression - логистическую модель,

–probit regression - модель линии пробитов,

–exponential growth regression - экспоненциальную модель,

–piecewise linear regression - линейно кусочную (разрывную) модель.

Из множества возможных видов нелинейного моделирования рассмотрим построение наиболее часто применяемых экспоненциальной и степенной моделей.

Сущность решения заключается:

–в преобразовании исходных данных о факторе X и параметре Y путем логарифмирования, в результате этого нелинейная модель приводится к линейной;

–в решении задачи с преобразованными данными по модулю линейного регрессионного анализа с получением коэффициентов и дисперсионного анализа модели;

–в обратном преобразовании коэффициентов путем потенцирования с целью построения требуемой нелинейной модели.

Так, для построения экспоненциальной модели $y=e^{a+bx}$ проведем ее логарифмирование:

$$\ln y = a + bx. \quad (4.19)$$

Обозначая $Z=\ln y$, получим линейное уравнение:

$$Z = a + bx \quad (4.20)$$

Это уравнение нетрудно построить по модулю линейной регрессии, вводя в качестве исходных данных:

— зависимая переменная $Z=\ln y$;

— независимая переменная x.

Получив коэффициенты a и b в модели $\hat{z} = a + bx$, путем потенцирования переходят к требуемой экспоненциальной модели $\hat{y} = e^{a+bx}$.

Для построения степенной модели $y=a\times x^b$ также проводят ее логарифмирование

$$\ln y = \ln a + b \times \ln x.$$

Обозначая $Z=\ln y$; $b_0=\ln a$; $t=\ln x$, получим линейное уравнение:

$$Z = b_0 + b \times t,$$

где Z - зависимая переменная $Z=\ln y$;

t - независимая переменная $t=\ln x$;

b_0 - свободный член $b_0=\ln a$;

b - коэффициент регрессии.

По модулю линейной регрессии определяют коэффициент b_0 и b, оценивают их значимость, проводят дисперсионный анализ и оценку эффективности модели: $\hat{z} = b_0 + b \times t$.

Путем потенцирования этого уравнения находят требуемую степенную модель: $y=a\times x^b$.

Последовательность построения степенной модели приведена в примере 4.3.

ПРИМЕР 4.1

Постановка задачи. Исследуется связь между поглощенной дозой облучения Y, Гр и долей аберрантных клеток костного мозга X, % у подопытных животных. С целью построения модели для определения поглощенной дозы облучения по доле аберрантных клеток костного мозга с 15 подопытными животными (белые мыши) проведен эксперимент (кафедра токсикологии ВМедА, 1990), результаты которого даны в табл.4.1.

Таблица 4.1

Доля аберрантных клеток X, % и доза облучения Y, Гр в эксперименте

Номер наблюдения	Доля аберрантных клеток костного мозга X, %	Доза облучения Y, Гр
1	59	3,2
2	44	2,5
3	85	4,5
4	70	4,0
5	52	3,0
6	21	0,8
7	26	1,3
8	79	4,0
9	41	3,1
10	67	3,5
11	32	1,8
12	18	0,7
13	90	4,3
14	12	0,3
15	100	5,0

Требуется определить:

1. Числовые характеристики переменных.
2. Коэффициент корреляции r_{xy} и оценить его точность и надежность.
3. Коэффициенты модели $\hat{y} = a + bx$ и оценить их достоверность.
4. Дать дисперсионный анализ и оценить достоверность и эффективность модели.
5. Построить график линии регрессии с указанием 95 % доверительных интервалов для возможных значений и среднего ожидаемого значения параметра.
6. Дать прогноз дозы облучения при количестве аберрантных клеток костного мозга $X_k = 50\%$. Оценить его точность и надежность.
7. Сформулировать выводы.

Решение дано с помощью персонального компьютера с использованием ППП Statistica 5.0.

1. Числовые характеристики переменных даны в машинограмме 4.1.

Машинограмма 4.1

Descriptive Statistics (korrel.sta)

	Valid N	Mean	Confid. -95%	Confid. 95%	Median	Minimum
X	15	53,07	37,46	68,68	52	12
Y	15	2,80	1,97	3,63	3,1	0,3

	Maximum	Std.Dev.	Standard Error	Skewness	Kurtosis
X	100	28,19	7,28	0,13	-1,25
Y	5	1,50	0,39	-0,35	-1,15

Распределения переменных X и Y следует признать близким к нормальному распределению, т.к. имеют место примерное равенство средних значений (среднего арифметического и медианы), примерная симметричность минимальных и максимальных значений относительно среднего значения, коэффициенты асимметрии и эксцесса не превышают 2 по абсолютной величине. Препятствий к применению корреляционного и регрессионного анализа нет. По характеру зависимости параметра Y от фактора X можно предположить, что модель для параметра Y может быть линейной.

2. Для оценки линейной связи между переменными X и Y рассчитан коэффициент корреляции r_{xy}=0,973 (см. машинограмму 4.2), что свидетельствует о прямой сильной корреляционной связи между долей аберрантных клеток костного мозга и дозой облучения подопытных животных. По t-критерию $t=15,31>t_{001}=4,07$, корреляционную связь следует считать значимой (p<0,001).

3. Коэффициенты модели и оценки их значимости даны в машинограмме 4.2. Свободный член a=0,045 с уровнем значимости p=0,83. Коэффициент регрессии $b=R_{y/x}$=0,052 с уровнем значимости p=0,00000 и достоверностью 1-p=1 определяет характер изменения параметра Y, а именно увеличение доли аберрантных клеток на 1% свидетельствует об увеличении дозы облучения на 0,052 Гр.

Машинограмма 4.2

Regression Summary for Dependent Variable: Y (korrel.sta)
R= ,97338588 RI= ,94748007 Adjusted RI= ,94344008
F(1,13)=234,53 p<,00000 Std.Error of estimate: ,35753

	BETA	St. Err. of BETA	B	St. Err. of B	t(13)	p-level
Intercpt			0,045	0,202	0,224	0,826
X	0,973	0,064	0,052	0,003	15,314	0,000

Analysis of Variance; DV: Y (korrel.sta)

	Sums of Squares	df	Mean Squares	F	p-level
Regress.	29,98	1	29,97827	234,5251	1,07E-09
Residual	1,66	13	0,127825		
Total	31,64				

Так как свободный член уравнения регрессии a=0,045 не оказывает существенного влияния на параметр Y, модель можно представить в виде $\hat{y} = 0,052 \times x$.

4. Дисперсионный анализ модели дан в машинограмме 4.2. По коэффициенту детерминации R^2=94,75% модель можно считать в высокой степени информативной. Фактор X и модель на 94,74% объясняет дисперсию параметра Y. По критерию F=234,5 и уровню значимости p=0,000 модель следует признать значимой, ее достоверность близка к 1 (100%). Таким образом, число наблюдений в эксперименте оказалось вполне достаточным для построения информационно способной (R^2=0,95) статистически значимой (p<0,001) модели.

5. График линии регрессии дан на рисунке 4.3. Точками обозначены результаты наблюдений, пунктирными линиями – 95% доверительный интервал для ожидаемого среднего значения параметра Y.

6. Прогноз дозы облучения при установленной дозе аберрантных клеток костного мозга X_k=50%.

По модели $\hat{y} = 0,052 \times 50 = 2,6$Гр. Эту же величину можно определить по графику на рис.4.3.

95% доверительный интервал:

– для возможных значений дозы облучения при

$S_0 = \sqrt{S_0^2} = \sqrt{0{,}128} = 0{,}36$ $t_{95}=2{,}16$ $Y_k=2{,}6\pm0{,}36\times2{,}16=1{,}8\div3{,}4$ Гр;

– для среднеожидаемого значения дозы облучения при

$$m_{\bar{y}_k} = \sqrt{\frac{S_0^2}{n}} = \sqrt{\frac{0{,}128}{15}} = 0{,}09 \qquad \bar{y} = 2{,}6 \pm 0{,}09 \times 2{,}16 = 2{,}4 \div 2{,}8 \text{Гр.}$$

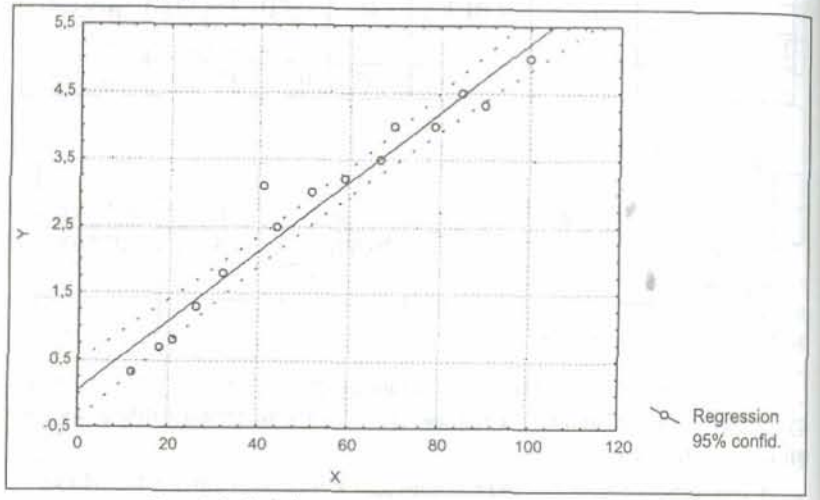

Рис.4.3. График линии регрессии $\hat{y} = 0{,}052 \times x$.

Выводы:

1. Между дозой облучения и долей аберрантных клеток костного мозга установлена сильная, прямая, значимая корреляционная связь ($r_{xy}=0{,}97$, $p<0{,}001$).

2. По доле аберрантных клеток костного мозга (x, %) можно определять полученную дозу облучения y (Гр) по модели $\hat{y} = 0{,}052x$. Полученная модель высоко информативна ($R^2=0{,}95$) и значима ($p<0{,}001$).

3. Увеличение доли аберрантных клеток на 1% свидетельствует об увеличении дозы облучения на 0,052 Гр.

4. Поскольку модель для параметра Y в высокой степени информативна, можно считать, что количество наблюдений в эксперименте вполне достаточно.

ПРИМЕР 4.2

Изучается связь между стажем работы с промышленными ядами и заболеваемостью токсическим гепатитом по данным выборочного наблюдения на предприятии химической промышленности. Данные по уровню заболеваемости токсическим гепатитом (Y, %) для сотрудников химического предприятия с различным стажем работы (X, годы) представлены в табл.4.2.

Таблица 4.2

Стаж работы и уровни заболеваемости

№ наблюдения	Стаж работы (X, годы)	Уровень заболеваемости (Y,%)
1	до года	2
2	1-2 года	9
3	2-3 года	7
4	3- 5 лет	12
5	5-10 лет	10
6	более 10 лет	14

Данные о стаже работы измерены в порядковой шкале, число наблюдений мало, неизвестен закон распределения признаков. Поэтому в рассматриваемом примере корректно оценивать статистическую связь с помощью процедуры рангового коэффициента корреляции Спирмена (Correlations (Speerman, Kendall tau, gamma)) из модуля "Nonparametrics/Distrib ППП Statistica. Результаты решения в машинограмме 4.3.

Машинограмма 4.3

Spearman Rank Order Correlations (correl.sta)
MD pairwise deleted

	Valid N	Spearman R	t(N-2)	p-level
X & Y	6	0,886	3,816	0,019

Анализ результатов

Установлено, что между стажем работы с промышленными ядами и заболеваемостью токсическим гепатитом имеется сильная прямая (R=0,886), статистически значимая (p=0,019) корреляционная связь. Таким образом, при увеличении стажа работы с промышленными ядами повышается уровень заболеваемости токсическим гепатитом.

ПРИМЕР 4.3

По данным токсикологического эксперимента «доза-эффект» необходимо построить степенную модель. В результате опыта сформирована матрица наблюдений для 8-ми групп лабораторных животных (табл. 4.3). В таблице указаны дозы активного вещества D_i, мкг/кг и относительные величины числа лабораторных животных с требуемым эффектом P_i, % для каждой группы.

Таблица 4.3

Результат эксперимента

№ группы	D_i мкг/кг	P_i, %
1	3,4	9
2	4,0	12
3	5,6	18
4	7,0	24
5	10,2	40
6	11,4	47
7	13,5	59
8	15,4	70

Построив график точек наблюдавшихся значений отклика в зависимости от дозы D_i (рис. 4.4), можно предположить, что пригодной моделью является степенное уравнение регрессии $P=a\times D^b$.

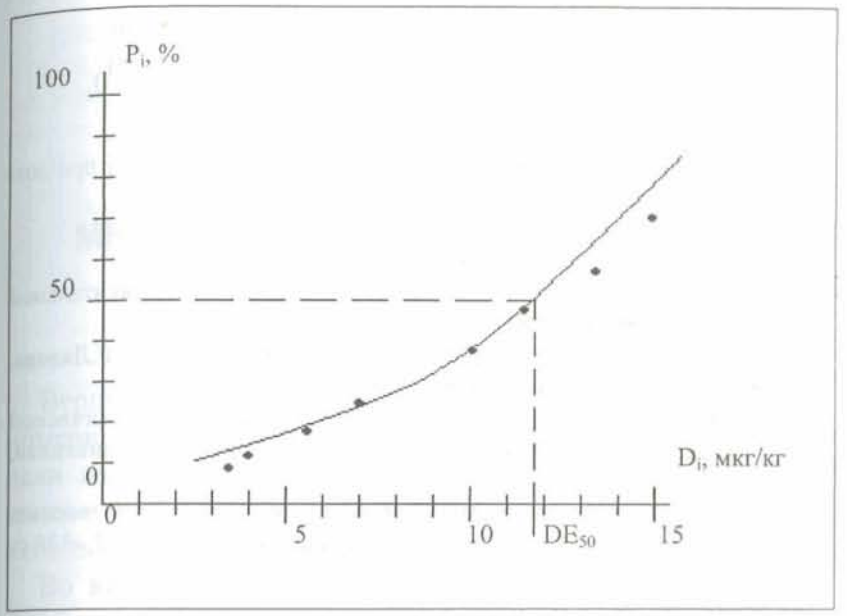

Рис.4.4.Поле наблюдений по данным эксперимента "доза-эффект".

Для получения коэффициентов степенной модели a и b переменные D и P следует трансформировать: Z=lnP, t=lnD. По модулю Linear Regression при вводе в качестве зависимой переменной Z и независимой t, получили линейное уравнение регрессии:

$$Z=b_0+b\times t,$$

где b_0=0,59, b=1,34.

После потенцирования находят требуемое степенное уравнение регрессии, в котором коэффициент $a=e^{b_0}=e^{0,59}=1,81$, коэффициент b=1,34.

$$P=1,81\times D^{1,34}.$$

Этому уравнению соответствует кривая линия (рис.4.4).

Дисперсионный анализ подтвердил его информативность и значимость. По этому уравнению можно найти среднюю эффективную дозу DE_{50}, при которой доля объектов с требуемым эффектом будет 50%.

$$50 = 1{,}81 \times DE_{50}^{1{,}34};\ \ln\frac{50}{1{,}81} = 1{,}34 \ln DE_{50};\ DE_{50} = \exp\left(\frac{\ln\frac{50}{1{,}81}}{1{,}34}\right)$$

и DE_{50}=11,9 мкг/кг. (рис. 3.4).

Таким образом, 50% объектов получает требуемый эффект при дозе 11,9 мкг/кг (рис. 4.4).

Литература

1. Зайцев Г.Н. Математическая статистика в экспериментальной ботанике. - М.: Наука, 1984. - 424 с.
2. Компьютерная биометрия: пакет CSS 3.1 /Под ред. В.Р.Лядова. – СПб.: Фонд «Инициатива», 1997. - 155 с.
3. Лядов В.Р. Основы теории вероятностей и математической статистики: Для студентов мед.ВУЗов. - СПб.: Фонд «Инициатива», 1998. - 107 с.
4. Поляков Л.Е., Игнатович Б.И, Лашков К.В. Основы военно-медицинской статистики /Под ред. Л.Е.Полякова - Л.: Б.и, 1977. -336 с.

ЧАСТЬ II

МНОГОМЕРНЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ И МОДЕЛИРОВАНИЯ МЕДИЦИНСКИХ СИСТЕМ

Вершиной медико-биологического исследования, закономерным его финалом очень часто является создание модели изучаемого явления, процесса. Наиболее объективными моделями являются модели, для создания которых используются математические методы.

Во второй части книги детально рассматриваются вопросы подготовки данных исследования к обработке с помощью многомерных методов математико-статистического моделирования, последовательность разработки, оценки качества и эксплуатации моделей на примерах материалов реальных исследований.

МНОГОФАКТОРНЫЙ КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Задачи исследования сложных систем

Известно, что объекты исследования в медицине представляют собой сложные вероятностные (стохастические) системы. Сложные системы функционируют при воздействии на них множества входных факторов. Часть из них является контролируемыми X1, X2 ,..., Xk, измеряемыми количественно или оцениваемыми в баллах. Другая часть входных факторов относится к группе неконтролируемых, случайных факторов; они не поддаются измерению, но оказывают воздействие на систему, результатом которого является случайность ее функционирования. Состояние системы характеризуется множеством выходных параметров Y1, Y2 ,..., Yl, которые также измеряются количественно или в баллах и представляют собой случайные величины, следующие нормальному или иному закону распределения с соответствующими числовыми характеристиками. Наилучшие результаты многомерного статистического анализа данных медицинских исследований получают тогда, когда распределение входных факторов и выходных параметров нормальное или близкое к нему.

Наблюдавшиеся значения k факторов и l параметров для n объектов сводятся в матрицу наблюдений размером n×(k+l). По матрице наблюдений на ПК с помощью ППП Statistica 5.0 проводится:

– статистическое описание переменных;
– корреляционный анализ;
– канонический корреляционный анализ;
– регрессионный анализ.

В результате статистического описания устанавливают законы распределения переменных и определяют их числовые характеристики, строят графики основных зависимостей между факторами и параметрами.

Корреляционный анализ обеспечивает оценку связей всех переменных попарно.

Канонический корреляционный анализ дает оценку связи всего множества входных факторов со всеми выходными параметрами в совокупности.

На основе канонического корреляционного анализа можно судить о достаточности связи входных факторов, включенных в матрицу наблюдений, и выходных параметров, характеризующих состояние системы.

Моделирование каждого выходного параметра методами регрессионного анализа дает возможность построить линейные или нелинейные модели. Модели используются для решения основных задач системного анализа:

– изучения характера изменения выходных параметров при изменении входных факторов;
– оценки степени влияния факторов на параметры;
– прогнозирования параметров при заданных значениях факторов;
– поиска оптимальных уровней факторов для получения требуемых значений параметров;
– оценки информативности параметров при заданной совокупности воздействующих факторов.

Требования к базе данных для многомерного статистического анализа

Матрица наблюдений с n строками по числу наблюдавшихся объектов в выборке и (k+l) столбцами по числу наблюдавшихся k входных факторов и l выходных параметров должна содержать только количественные данные в натуральных единицах измерения или баллах.

При отсутствии данных по какому либо признаку его заменяют средним значением признака для всей выборки, хотя это приводит к искажению исходной информации. Следует также иметь в виду, что некоторые статистические пакеты не рассчитывают корреляционной матрицы, в случае если число переменных превышает число наблюдений. Надежное решение можно получить, если в матрицах наблюдений число строк n в 3-5 раз превышает число столбцов (k+l).

Все данные должны быть тщательно проверены: устраняются грубые ошибки, исключаются явно анормальные результаты наблюдения.

Выборка должна быть безусловно репрезентативной по отношению к исследуемой генеральной совокупности.

В соответствии с целью и задачами исследования в матрицу необходимо ввести дополнительные столбцы с группировочными признаками, например, группировочный признак G1 - контрольная группа с

кодом 1, опытная группа с кодом 2; группировочный признак пола G2 - мужчины с кодом 1, женщины с кодом 2 и т.п.

Задачи и содержание многомерного корреляционного анализа

Многомерный корреляционный анализ проводится для количественной оценки направления, силы и значимости линейной связи между всеми переменными базы данных попарно. Такая связь характеризуется коэффициентом корреляции Пирсона.

В результате решения по опциям Descriptive statistics и Correlations на экран выводятся следующие результаты:

– таблица числовых характеристик переменных;

– корреляционная матрица, содержащая коэффициенты корреляции и уровни их значимости для всех пар переменных.

По таблице числовых характеристик анализируется соответствие распределений каждой переменной нормальному закону.

По корреляционной матрице, представляющей собой квадратную симметричную таблицу с размером (k+l)×(k+l), судят о направлении, силе и значимости корреляционной связи переменных попарно, в особенности о связи входных факторов с выходными параметрами.

Вариант расчета числовых характеристик переменных и корреляционной матрицы, а также интерпретации результатов приведен в примере 5.1.

Назначение и содержание канонического корреляционного анализа

Канонический корреляционный анализ предназначен для изучения связи между входными факторами и выходными параметрами в их совокупности.

Для проведения канонического корреляционного анализа в исходной матрице наблюдений с размерами n×(k+l), где n - число наблюдавшихся объектов, k - число входных факторов и l - число выходных параметров, выделяют две группы переменных:

1. Left set - группа выходных параметров;
2. Right set - группа входных факторов.

Алгоритмом предусмотрено.

1.Определение ограниченного числа канонических переменных, обобщающих выходные параметры 1-ой группы, и такого же количества канонических переменных, обобщающих входные факторы 2-й группы. При этом первая пара канонических переменных обобщает наибольшую часть дисперсии переменных, вторая пара - большую долю из оставшейся части дисперсии и т.д. Количество пар канонических переменных зависит от размерности матрицы наблюдений. Практика показала, что 2-3 пар канонических переменных достаточно для надежного представления всей совокупности переменных.

2.Формирование полей рассеяния объектов в координатах первой, второй, третьей пары канонических переменных для 1-ой и 2-ой группы, а после их формирования - расчет канонических коэффициентов корреляции: Can r1 - по паре первых, Can r2 - по паре вторых, Can r3 - по паре третьих канонических переменных.

По величине канонических коэффициентов корреляции судят о силе связи между совокупностями входных факторов и выходных параметров.

Квадраты коэффициентов (Eigenvalue) характеризуют степень детерминации совокупности параметров совокупностью факторов для каждой пары канонических переменных.

Значимость канонических коэффициентов корреляции и детерминации оценивают по χ^2 - критерию Пирсона. Коэффициенты считают значимыми при вероятности равной и более 0,95 или при уровне значимости $p \leq 0,05$;

3.Расчет факторной структуры канонических переменных (Factor structure), т.е. коэффициентов корреляции, характеризующих направление и силу корреляционной связи канонических переменных с наблюдавшимися входными факторами и выходными параметрами. В результате дается оценка важности входных факторов и информативности выходных параметров.

Такой анализ на начальном этапе исследования позволяет оценить достаточность связи между входными факторами и выходными параметрами с целью построения для них надежных моделей, а также выделить наиболее значимые факторы и информативные параметры-отклики на воздействия.

Назначение и содержание многомерного регрессионного анализа. Построение линейного уравнения регрессии

Многомерный регрессионный анализ (Multiple Regression) применяется для построения уравнения регрессии для параметра Y в зависимости от факторов X1 – Xk. Модель может быть линейной и нелинейной. Наиболее простой, содержащей только линейные эффекты факторов, является линейная модель.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k, \qquad (5.1)$$

где $\hat{y}$ - прогнозируемое значение выходного параметра;

b_0 – свободный член;

$b_1 b_2 \ldots, b_k$ - коэффициенты регрессии;

$x_1 x_2 \ldots, x_k$ - возможные значения факторов X1,X2,…Xk;

$b_1 x_1, b_2 x_2, \ldots b_k x_k$ – линейные эффекты факторов.

Коэффициенты модели получают методом наименьших квадратов по исходной матрице наблюдений n×(k+1), где: n – число строк в матрице, равное числу наблюдаемых объектов, k+1 – число столбцов, равное числу независимых переменных (k факторов X1,X2,…Xk) и одной зависимой переменной (моделируемый параметр Y).

Значимость коэффициентов оценивают по t-критерию Стьюдента. При построении модели в ответственных случаях, например, для прогноза параметра Y, в модели сохраняют только значимые коэффициенты с доверительной вероятностью больше или равной 0,95 или с уровнем значимостью p≤0,05. В поисковом исследовании с целью изучения характера изменения параметра Y, при изменении факторов и степени влияния их на параметр, допускают сохранение в модели эффектов с коэффициентами, при уровне их значимости p≤0,30 (доверительной вероятностью равной или больше 0,70).

Стандартный алгоритм регрессионного анализа предусматривает расчет:

– числовых характеристик переменных;

– корреляционной матрицы;

– коэффициентов модели с оценками их значимости;

– результатов дисперсионного анализа модели и оценки коэффициентов множественной корреляции и детерминации, средней квадратичной ошибки прогноза параметра Y по модели;

– графика линии регрессии с указанием 95%-го доверительного интервала для прогноза значений параметра Y.

Вариант многофакторного регрессионного анализа с целью построения линейного уравнения регрессии дан в примере 5.1.

Сущность пошагового регрессионного анализа

Стандартный алгоритм многомерного регрессионного анализа обеспечивает получение коэффициентов модели для всех независимых переменных X1 – Xk. Исходя из уровней значимости, исследователь решает, какие коэффициенты должны быть включены в модель, как значимые, достоверные. Для автоматического включения значимых эффектов в модель и исключения незначимых предлагается пошаговый регрессионный анализ в двух вариантах:

– Forward – поочередное включение в модель наиболее значимых эффектов;

– Backward – поочередное исключение из полной модели наименее значимых эффектов.

Отбор значимых эффектов реализуется по критерию F – Фишера.

В ответственных исследованиях для получения коэффициентов с уровнем значимости p≤0,05 задается значение критерия F=3-4. В поисковых исследованиях значение F=1-2 обеспечивает включение в модель коэффициентов с уровнем значимости p≤0,30.

В примере 5.1 методом пошагового регрессионного анализа при F=1 получены коэффициенты модели, приведенные в машинограмме 5.3. Исключенным из модели оказался эффект фактора X1.

Дисперсионный анализ и оценка эффективности модели

Дисперсионный анализ модели выполняется для оценки ее эффективности. Под эффективностью модели понимают ее информативность и значимость (достоверность). Модель считают информативной, если ее коэффициент детерминации $R^2>0,5$; значимой, достоверной при уровне значимости по F – критерию p≤0,05 (достоверности ≥0,95).

В примере 5.1 в машинограмме 5.4. дан дисперсионный анализ модели, из которого следует, что модель достоверна (p=1,93E–0,5); из машинограммы 5.3. видно, что модель информативна, т.к. $R^2=0,819$ (значительно больше 0,5), а стандартная ошибка прогноза возможных значений параметра S_0=26,528 нКи/кг и среднеожидаемых значений параметра:

$$m_{\hat{y}} = \frac{S_0}{\sqrt{n}} = \frac{26{,}528}{\sqrt{20}} = 5{,}9 \text{нКи / кг}.$$

Такую модель можно применять для решения задач исследования.

Оценка степени влияния факторов на моделируемый параметр

Степень влияния факторов на параметр Y рассчитывается по величине стандартизованных коэффициентов регрессии BETA по формуле (5.2.):

$$K_j = \frac{100 \times BETA_j}{\sum_{(j)} |BETA_j|} \times R^2, \text{в\%}, \qquad (5.2)$$

По данным примера 5.1 оценка степени влияния факторов X5, X3, X2 и X4 *дана в таблице 5.2.*

Прогноз по модели и оценка его точности и надежности

Прогноз среднеожидаемых значений параметра может быть дан по модели (5.1.) или графику линии регрессии (рис.5.1).

Точность и надежность прогнозируемого значением параметра *оценивается 95%-м доверительным интервалом:*

$$Y = \hat{y} \pm t_{95} \times m_{\hat{y}}. \qquad (5.3)$$

В примере 5.1 для заданных значений факторов получен прогноз *y=68,4 нКи/кг и его 95%-й доверительный интервал от 54 до 83 нКи/кг.*

Достаточно большой доверительный интервал характерен для поискового исследования из-за приблизительной оценки значений факторов и параметра в матрице наблюдений.

Особенности нелинейного регрессионного анализа

При нелинейной зависимости моделируемого параметра от входных факторов линейное уравнение регрессии (5.1) может оказаться неинформативным. В таких случаях следует построить нелинейное уравнение регрессии по операции Fixed non-linear модуля Multiple Regression или по модулю Nonlinear Estimation.

Вид нелинейного уравнения можно предположить, проведя предварительно графическое исследование зависимости параметра Y от факторов X1 – Xk путем построения соответствующих графиков. Наибо-

лее пригодным, как показывает практика моделирования, являются степенная и экспоненциальная модель.

Для построения степенной модели:

$$Z = b_0 \times u_1^{b_1} \times u_2^{b_2} \times ... \times u_k^{b_k}, \qquad (5.4)$$

где Z – зависимая переменная,

$u_1, u_2 \ldots, u_r$ – независимые переменные,

$b_0, b_1, b_2 \ldots, b_k$ – коэффициенты модели,

ее трансформируют в линейную путем логорифмирования:

$$\ln Z = \ln b_0 + b_1 \ln u_1 + b_2 \ln u_2 + \ldots + b_k \ln u_k$$

Обозначив $\ln Z = y$; $\ln b_0 = a$; $\ln u_1 = x_1$; $\ln u_2 = x_2$;...; $\ln u_k = x_k$, получим обычное линейное уравнение:

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k, \qquad (5.5)$$

которое можно построить и оценить стандартным линейным регрессионным анализом.

Исходные данные для построения модели формируются в матрицу по шаблону в таблице 5.1.

После определения *коэффициентов* модели (5.5) и оценки их значимости, дисперсионного анализа и оценки информативности и значимости модели выполняют ее потенцирование *и приходят* к требуемой степенной модели (5.4), в которой константа $b_0 = e^a$, а коэффициенты $b_1, b_2, \ldots, b_k$ - полученные в модели (5.5).

Таблица 5.1

Шаблон матрицы наблюдений

№ наблюдения	y=lnZ	$x_1=\ln u_1$	$x_2=\ln u_2$	...	$x_k=\ln u_k$

Для построения экспоненциальной модели:

$$Z = e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k}, \qquad (5.6)$$

где Z – зависимая переменная;

$x_1, x_2 \ldots, x_r$ – независимые переменные;

$b_0, b_1, b_2 \ldots, b_k$ – коэффициенты модели,

модель трансформируют в линейную также путем логарифмирования:

$$\ln Z = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k. \qquad (5.7)$$

Обозначив $\ln Z = y$ получим привычное линейное уравнение:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k, \qquad (5.8)$$

которое *можно построить* и оценить методом линейного регрессионного анализа по исходной матрице наблюдений (шаблон дан в таблице 5.2).

Таблица 5.2

Шаблон матрицы наблюдений

№ наблюдения	y=lnZ	x_1	x_2	…	x_k

После определения коэффициентов модели (5.8), оценки их значимости, дисперсионного анализа и оценки информативности и значимости модели, выполняют ее потенцирование и приходят к требуемой экспоненциальной модели (5.6). Все коэффициенты этой модели те же, что и в модели (5.8).

Вариант построения экспоненциальной модели дан в примере 5.2.

ПРИМЕР 5.1

Постановка задачи. Исследуется влияние пяти факторов на уровень загрязнения почвы радиоактивными веществами (РВ). Параметр, характеризующий уровень загрязнения почвы – удельная активность почвы Y, нКи/кг.

Факторы, влияющие на уровень загрязнения почвы РВ:

X1 – расстояние от источника РВ и ионизирующих излучений, км;

X2 – *относительная частота накрытия местности* 30-градусной зоной облака РВ под действием ветра, относительный коэффициент;

X3 – время эксплуатации объекта – источника РВ и ионизирующих излучений, число лет;

X4 – среднее число случаев приведения энергетических установок в рабочее положение на объекте в течение года;

X5 – относительная величина обеспеченности объекта санпропускниками, *относительный коэффициент.*

Результаты 20 наблюдений на местности в районе объектов – источников РВ и ионизирующих излучений даны в матрице наблюдений (таблица 5.3).

Матрица наблюдений

Номер наблюдения	Параметр Y, *нКи/кг*	Факторы				
		X1, км	X2, отн.коэф.	X3, лет	X4, случ.	X5,отн. коэф.
1	90	0,5	0,2	24	40	0,2
2	186	4,0	0,3	35	120	0,2
3	26	15,0	0,3	10	100	0,8
4	200	1,0	0,2	34	168	0,2
5	82	3,0	0,1	30	100	0,2
6	50	11,0	0,05	28	180	0,4
7	8	20,0	0,02	18	25	0,6
8	20	0,5	0,2	20	18	0,8
9	45	0,5	0,3	30	40	0,7
10	15	5,0	0,05	30	80	1,0
11	100	0,5	0,2	26	190	0,2
12	74	5,0	0,3	19	130	0,3
13	12	18,0	0,05	12	40	0,6
14	120	2,0	0,1	26	160	0,2
15	38	15,0	0,3	20	70	0,4
16	64	10,0	0,2	14	80	0,3
17	100	2,0	0,1	28	130	0,3
18	6	24,0	0,05	20	38	0,8
19	118	13,0	0,1	28	120	0,2
20	72	5,0	0,2	20	70	0,3

Требуется:

1. *Определить числовые характеристики* переменных и корреляционную матрицу.
2. Определить коэффициенты модели методом пошагового регрессионного анализа. Дать дисперсионный анализ модели и оценить ее информативность и значимость, а также степень влияния факторов на параметр Y.
3. Построить график для оценки точности и надежности прогноза параметра Y по модели.
4. *Дать прогноз параметра* Y *для* заданных значений факторов: X2=0,15; X3=25 лет; X4=100 случ.; X5=0,5. Оценить точность и надежность прогноза.

Решение произведено с помощью персонального компьютера с использованием ППП Statistica 5.0.

1.В машинограмме 5.1 приведены числовые характеристики переменных: средние значения, 95% - доверительный интервал для средних значений, минимальные и максимальные значения, размах, стандартные отклонения и ошибки, а также коэффициенты асимметрии и эксцесса, анализ которых убеждает, что все переменные имеют распределения близкие к нормальному.

Из данных корреляционной матрицы (машинограмма 5.2) пяти факторов X1 - X5 с параметром Y следует, что прямая связь параметра Y установлена с факторами X2 , X3 и X4 , обратная – с факторами X1 и X5. Параметр Y имеет сильную корреляционную связь с фактором X5 и умеренную - с факторами X4 , X1 , X3. Связь с фактором X2 - слабая. Значимыми являются коэффициенты корреляции r_{x1y} ($p<0,05$), r_{x3y} ($p<0,01$), r_{x4y} ($p<0,01$), r_{x5y} ($p<0,001$).

Машинограмма 5.1

Descriptive Statistics (mn_regr.sta)

	n	Mean	Confid. -95,000%	Confid. 95,00%	Minimum	Maximum
Y	20	71,30	45,37	97,23	6	200
X_1	20	7,75	4,27	11,23	0,5	24
X_2	20	0,16	0,11	0,21	0,02	0,3
X_3	20	23,60	20,30	26,90	10	35
X_4	20	94,95	69,91	119,99	18	190
X_5	20	0,44	0,31	0,56	0,2	1

Range	Std.Dev.	Standard Error	Skewness	Kurtosis
194	55,41	12,39	0,93	0,51
23,5	7,43	1,66	0,82	-0,55
0,28	0,10	0,02	0,25	-1,45
25	7,06	1,58	-0,29	-0,70
172	53,51	11,97	0,28	-1,06
0,8	0,26	0,06	0,82	-0,72

Машинограмма 5.2

Correlations (mn_regr.sta)
n=20 (Casewise deletion of missing data)

	Y	X1	X2	X3	X4	X5
Y	1	-0,55	0,29	0,66	0,64	-0,77
	p= ---	p=,012	p=,222	p=,002	p=,002	p=,000
X1	-0,55	1	-0,37	-0,58	-0,35	0,38
	p=,012	p= ---	p=,104	p=,008	p=,134	p=,097
X2	0,29	-0,37	1	0,03	0,06	-0,14
	p=,222	p=,104	p= ---	p=,884	p=,810	p=,564
X3	0,66	-0,58	0,03	1	0,45	-0,35
	p=,002	p=,008	p=,884	p= ---	p=,046	p=,133
X4	0,64	-0,35	0,06	0,45	1	-0,57
	p=,002	p=,134	p=,810	p=,046	p= ---	p=,009
X5	-0,77	0,38	-0,14	-0,35	-0,57	1
	p=,000	p=,097	p=,564	p=,133	p=,009	p= ---

2.Модель для параметра Y получена методом пошагового регрессионного анализа. Отбор значимых факторов для включения в модель проведен при уровне F=1, что обеспечивает уровень значимости коэффициентов $p<0,30$, а достоверность $1-p>0,70$. Таблица коэффициентов модели (B) для факторов, включенных в модель, и их значимости (p-level) дана в машинограмме 5.3. В этой же машинограмме приводятся оценки качества модели:

– коэффициент детерминации (RI=,81905816), определяющий ее информационную способность;

– значение F-критерия (F(4,15)=16,975) и уровень значимости модели (p<,00002), определяющие статистическую значимость модели;

– стандартная ошибка (Std.Error of estimate 26,528), используемая для построения 95% доверительного интервала прогнозируемого значения Y.

Из данных, приведенных в машинограмме 5.3, видно, что коэффициенты модели X5, X3 являются значимыми, достоверными ($p<0,001$). Коэффициенты X2, X4 значимы в пределах 70% уровня надежности (в соответствии с заданным F=1 для пошагового отбора в модель).

Машинограмма 5.3

Regression Summary for Dependent Variable: Y (mn_regr.sta)
R= ,90501832 RI= ,81905816 Adjusted RI= ,77080701
F(4,15)=16,975 p<,00002 Std.Error of estimate: 26,528

	BETA	St. Err. of BETA	B	St. Err. of B	t(15)	p-level
Intercpt			5,945	34,141	0,174	0,864
X_5	-0,504	0,137	-106,457	28,933	-3,679	0,002
X_3	0,418	0,124	3,281	0,977	3,359	0,004
X_2	0,206	0,112	114,420	62,181	1,840	0,086
X_4	0,155	0,141	0,160	0,146	1,095	0,291

Фактор X1 в модель не включен, как недостаточно значимый. Модель для параметра Y имеет вид:

$$\hat{y} = 5{,}9 + 114{,}4 \times x_2 + 3{,}3 \times x_3 + 0{,}16 \times x_4 - 106{,}5 \times x_5.$$

Знаки коэффициентов модели отражают прямую связь параметра Y с факторами X2, X3 и X4 и обратную - с X5.

Дисперсионный анализ модели и оценки ее информативности и значимости даны в машинограмме 5.4. Вклад факторов, включенных в модель (Regress=47784,02), составляет 82% от общей суммы квадратов отклонений прогнозируемого параметра Y (Total=58340,2), а 18% вклада вносят неучтенные (случайные) факторы (Residual=10556,18), что свидетельствует об информационной способности модели. По величине F-критерия F=16,97 с уровнем значимости p=0,000019 модель можно считать значимой, достоверной.

Машинограмма 5.4

Analysis of Variance; DV: Y (mn_regr.sta)

	Sums of Squares	df	Mean Squares	F	p-level
Regress.	47784,02	4	11946	16,97489	1,93E-05
Residual	10556,18	15	703,7455		
Total	58340,2				

Степень влияния факторов на параметр Y рассчитывается по величине стандартизованных коэффициентов регрессии BETA из машинограммы 5.3 с помощью формулы 5.2.

Таблица 5.2

Степень и значимость влияния факторов на параметр Y

Фактор	BETA	Степень влияния K_j, %	p-level
X_5	-0,50	32,06	0,002
X_3	0,42	26,80	0,004
X_2	0,21	13,21	0,086
X_4	0,15	9,93	0,291

Наибольшее влияние на параметр Y имеет фактор X5, затем факторы X3, X2, X4.

3.График линии регрессии прогнозируемых и наблюдавшихся значений параметра Y для оценки его точности и надежности дан на рисунке 5.1.

4.Прогноз параметра Y для заданных значений факторов X2=0,15; X3=25лет; X4=100 случ.; X5=0,5 дается по модели

$$\hat{y} = 5{,}9 + 114{,}4 \times 0{,}15 + 3{,}3 \times 25 + 0{,}16 \times 100 - 106{,}5 \times 0{,}5 = 68{,}4 \text{нКи/кг}.$$

По графику определяется 95% доверительный интервал для прогнозируемого значения y_k=54,0÷82,8 нКи/кг.

Полученный интервал достаточно большой из-за приблизительности оценок величин факторов в матрице наблюдений. Он с вероятностью 95% показывает значение удельной активности почвы при заданных значениях факторов X2-X5.

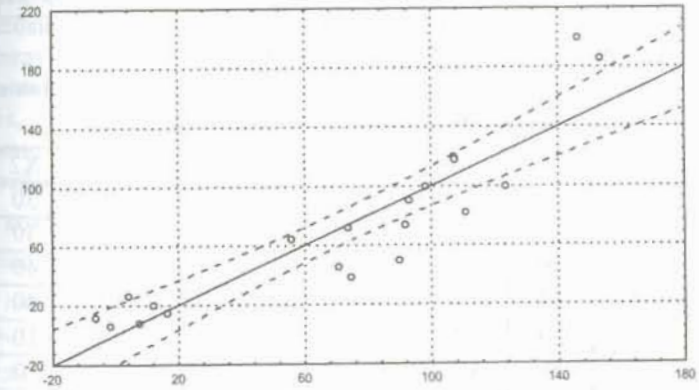

Рис.5.1. График линии регрессии прогнозируемых и наблюдавшихся значений параметра Y. (Точками обозначены результаты наблюдений, пунктирными линиями – 95% доверительный интервал для ожидаемого среднего значения параметра Y).

ПРИМЕР 5.2

Исследуется зависимость избыточного теплонакопления от факторов микроклимата и метаболизма. С этой целью планируется двухфакторный эксперимент в климатической камере со здоровыми мужчинами в возрасте 19-22 лет при выполнении ими физической работы в течение одного часа.

Фактор X1 - энерготраты (расход энергии в 1 мин) исследовался на семи уровнях: 5, 10, 15, 20, 25, 30 и 35 кДж/мин. Энерготраты определялись в ходе эксперимента газометрическим методом при выполнении работы на велоэргометре на семи адекватных уровнях. Поэтому в матрице наблюдений значения фактора X1 несколько отличаются от указанных выше уровней из-за различий в максимальном потреблении кислорода испытуемым при выполнении одинаковой работы (эти различия обычно составляют до 15-20%).

Фактор X2 - эффективная температура воздуха на четырех уровнях: 10, 20, 30, 40°С. Эффективная температура учитывает температуру, влажность и скорость движения воздуха в камере.

Параметром, характеризующим тепловой баланс организма, приведенный к единице поверхности тела, служит избыточное теплонакопление - Y кДж/м^2.

В эксперименте реализованы опыты на 28 сочетаниях уровней факторов X1 и X2 (число опытных точек N=7×4=28). В каждой опытной точке наблюдалось по два испытуемых, всего 56 наблюдений. Исходные данные приведены в таблице 5.3.

Таблица 5.3

Результаты активного эксперимента

№ пп	Y	X1	X2	№ пп	Y	X1	X2
1	4	5,1	10	29	150	20	30
2	5	5,2	10	30	160	21	30
3	20	5	20	31	380	19	40
4	24	5,3	20	32	372	21	40
5	59	4,8	30	33	20	25,2	10
6	50	4,7	30	34	17	24,8	10
7	360	4,9	40	35	159	25	20
8	340	5	40	36	162	25,1	20
9	10	9,9	10	37	158	24,9	20

№ пп	Y	X1	X2	№ пп	Y	X1	X2
10	12	10,1	10	38	170	25	20
11	54	9,8	20	39	314	26	30
12	50	10,2	20	40	300	24	30
13	154	11	30	41	50	30,1	10
14	148	9	30	42	55	30,3	10
15	360	9,8	40	43	143	29,7	20
16	370	10,2	40	44	154	29,9	20
17	10	15,1	10	45	134	30	20
18	9	14,8	10	46	148	30	20
19	56	14,9	20	47	260	30,9	25
20	60	15,2	20	48	248	29,1	25
21	149	15	30	49	64	35	10
22	140	14,9	30	50	70	36	10
23	380	15,1	40	51	200	34	20
24	370	15	40	52	197	35,2	20
25	19	19,9	10	53	210	34,7	20
26	17	19,7	10	54	190	35,3	20
27	60	20,1	20	55	380	34,7	25
28	64	20,3	20	56	370	35	25

Следует отметить, что при верхних значениях энерготрат (25-35 кДж/мин), соотвествующих тяжелой и очень тяжелой физической работе, эффективная температура задавалась не выше 25-30^0 С во избежание тепловых поражений.

Требуется определить:

– коэффициенты экспоненциальной модели $\hat{y} = b_0 \times \exp(b_1 x_1 + b_2 x_2)$ и оценки их значимости;

– дать дисперсионный анализ модели и оценить ее эффективность;

– дать прогноз параметра Y при: а) X1=34 кДж/мин (соответствует тяжелой физической работе) и X2=26^0 С, б) X1=15 кДж/мин (легкая физическая работа) и X2=18^0 С.

Решение. Есть некоторые основания предположить, что между параметром Y и факторами X1 и X2 имеется экспоненциальная зави-

симость вида: $\hat{y}=b_0\times\exp(b_1x_1+b_2x_2)$. Для построения экспоненциального уравнения регрессии следует прологарифмировать левую и правую части уравнения.

Для решения задачи методом многомерного регрессионного анализа зависимую переменную следует задать как ln (Y).

Ход решения в целом совпадает с описанным выше примером 5.1. Однако в начальном меню модуля Multiple Regression в окне Mode выбирается метод fixed non-linear в результате чего открывается окно Non-linear Components Regression, в котором следует выбрать *шестую функцию natural log. Затем в качестве зависимой* переменной задается LN-Y1, а в качестве независимых переменных X1 и X2. Дальнейшее решение в обычном порядке, а его результаты - в машинограммах 5.5, 5.6.

Машинограмма 5.5

Regression Summary for Dependent Variable: LN-V1
R= ,96979579 RI= ,94050388 Adjusted RI= ,93825875
F(2,53)=418,91 p<,00000 Std.Error of estimate: ,30942

	BETA	St. Err. of BETA	B	St. Err. of B	t(53)	p-level
Intercpt			0,34	0,15	2,23	0,0302
X1	0,61	0,03	0,07	0,00	17,45	0,0000
X2	0,94	0,03	0,12	0,00	26,91	0,0000

Машинограмма 5.6

Analysis of Variance; DV: LN-V1 (regr_non.css)

	Sums of Squares	df	Mean Squares	F	p-level
Regress.	80,22	2	40,11	418,91	3,342E-33
Residual	5,07	53	0,10		
Total	85,29				

В машинограмме 5.5. приведены коэффициенты модели, коэффициент детерминации, уровень значимости модели и стандартная ошибка прогноза, из которых следует, что нами получено вполне информационно способное (RI=,94) статистически значимое (p<,00000) уравнение прогноза уровня избыточного теплонакопления в зависимости от конкретных значений микроклимата и метаболизма следующего вида:

$$Ln(Y)=0,34+0,07\times x_1+0,12\times x_2.$$

Литература

1. Григорьев С.Г., Перфилов А.М., Левандовский В.В., Юнкеров В.И. Пакет прикладных программ STATGRAFICS на персональном компьютере. - СПб.: Б.и., 1992. - 104 с.

2. Зайцев Г.Н. Математический анализ биологических данных. – М.: Наука, 1991. - 183 с.

3. Компьютерная биометрия: пакет CSS 3.1 /Под ред. В.Р.Лядова. – СПб.: Фонд "Инициатива", 1997. - 155 с.

4. Математико-статистические методы в клинической практике / Под ред. В.И.Кувакина. - СПб,: Б.и, 1993. - 199 с.

5. Ферстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа: Руководство для экономистов / Пер. с нем. -М.: Финансы и статистика, 1983. - 302 с.

Глава 6. ДИСПЕРСИОННЫЙ АНАЛИЗ РЕЗУЛЬТАТОВ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Назначение и сущность дисперсионного анализа результатов медицинских исследований

Дисперсионный анализ - это метод математической статистики предназначенный для моделирования количественного выходного параметра-отклика на воздействующие входные факторы, уровни которых оцениваются качественно, по номинальной шкале. Например, фактор А - тяжесть заболевания на трех уровнях: (легкая, средняя, тяжелая).

В зависимости от количества контролируемых факторов различают одно- двух- и многофакторный дисперсионный анализ (ДА). Каждый контролируемый фактор может фиксироваться на двух, трех и более уровнях. Моделируемый параметр-отклик на воздействующие факторы оценивается количественно по интервальной или порядковой шкале для каждого сочетания уровней факторов.

Сущность ДА заключается в разложении дисперсии параметра Y на составляющие:

– дисперсию вследствие влияния контролируемых факторов;

– дисперсию, вызываемую действием неконтролируемых, случайных факторов и ошибками измерения.

По доле дисперсии, обусловленной контролируемыми факторами, определяется степень и значимость влияния входных факторов на параметр Y.

По средним значениям параметра Y на различных уровнях факторов изучается характер изменения параметра при изменении уровней воздействующих факторов, дается прогноз ожидаемых значений параметра при заданных уровнях факторов.

По результатам моделирования множества выходных параметров дается оценка их информативности на воздействующие факторы, что имеет большое значение при оценке весомости параметров и выработке комплексной интегральной оценки состояния объекта исследования.

Решение задач дисперсионного анализа может быть выполнено на ПК по модулю ANOVA/MANOVA ППП Statistica for Windows.

Содержание дисперсионного анализа полного факторного эксперимента (ПФЭ)

Для проведения ДА планируется и проводится эксперимент, в котором выходной параметр Y наблюдается на определенных уровнях контролируемых факторов. При числе контролируемых факторов не более 4 обычно проводят полный факторный эксперимент (ПФЭ), в котором параметр Y наблюдается на всех возможных сочетаниях уровней факторов. Например, при трехфакторном эксперименте с числом уровней для факторов A, B, C, соответственно, 3, 2, 2, число сочетаний уровней факторов (опытных точек) в ПФЭ будет N=3×2×2=12.

В каждой опытной точке для исключения возможных ошибок измерения параметра Y необходимо наблюдать несколько объектов. Число наблюдений берут n=2-5 при малой и средней вариабельности параметра Y и n=10 и более при значительной его вариабельности.

Если при N=12 наблюдать в каждой опытной точке n=5 объектов, то всего будет N×n=12×5=60 наблюдений.

Модуль ANOVA/MANOVA ППП Statistica 5.0 допускает обработку данных при неравном числе наблюдений в опытных точках. Результаты наблюдений сводятся в исходную матрицу наблюдений, вариант такой матрицы для двухфакторного эксперимента приведен в примере 6.1 (таблица 6.3, машинограмма 6.1).

Алгоритм ДА ПФЭ обеспечивает:

– оценку линейных эффектов и эффектов взаимодействия факторов на выходной параметр; на основе этих данных рассчитывают степени влияния факторов;

– расчет средних значений параметра Y для различных уровней факторов и их сочетаний; на основе этих данных получают графики средних значений с указанием 95%-х доверительных интервалов;

– оценку значимости различия средних значений параметра Y для различных уровней факторов по критериям LSD, Дункана, Шеффе, Тьюки и др.

В примере 6.1 решены все перечисленные задачи.

Оценка степени влияния линейных эффектов факторов и их взаимодействий на моделируемый параметр

Степень влияния контролируемых факторов на изучаемый выходной параметр рассчитывается по величине сумм квадратов отклонений этого параметра от средних значений:

$$K_j = \frac{100 \times SS_j}{\sum SS_j} \qquad (6.1)$$

где K_j - степень влияния j-го фактора на параметр Y, в %;

SS_j - сумма квадратов отклонений параметра Y от среднего значения вследствие влияния на него j-го фактора;

SS - общая сумма квадратов отклонений параметра Y от среднего значения вследствие влияния на него всех контролируемых, неконтролируемых, случайных факторов и ошибок измерения.

Значимость влияния факторов на параметр Y оценивается по F-критерию Фишера. Значимыми считают эффекты, вероятность которых равна или более 0,95 (при уровне значимости $p \le 0,05$).

В результате решения задачи на ПК выдается таблица (машинограмма 6.2), содержащая: суммы квадратов отклонений параметра Y, входящие в формулу (6.1) - Sum of Squares; числа степеней свободы - df (Degree of Freedom); средние квадраты отклонений или дисперсии - Mean Square; F (F-критерий Фишера) и его уровни значимости - p-level.

Оценка значимости различий средних значений параметра для различных уровней факторов

Так как на каждом сочетании уровней факторов проводилось n наблюдений параметра Y, имеется возможность определить его средние значения и оценить их точность и надежность 95%-ми доверительными интервалами. Модуль ANOVA/MANOVA ППП Statictica 5.0 дает возможность получить оценку средних занчений (Means) и стандартных отклонений (Std.Dev.) параметра Y для уровней каждого фактора и для всех сочетаний уровней факторов (машинограмма 6.3).

По данным машинограммы 6.3 могут быть получены графики средних значений параметра Y с 95%-ми доверительными интервалами (рис.6.1, 6.2 и 6.3). По этим графикам дается предварительное суждение о значимости различия средних значений параметра Y для различных уровней факторов.

Окончательные выводы о значимости различия средних значений параметра Y для различных уровней факторов даются по нескольким критериям. В примере 6.1 (машинограмма 6.4) даны таблицы уровней значимости различия по наиболее надежному и часто применяемому критерию LSD (Liqt Square Difference). Значимыми признаются различия при уровне значимости $p \le 0,05$.

Таким образом, цель и задачи исследования, поставленные в примере 6.1, в результате ДА двухфакторного ПФЭ выполнены полностью.

Ковариационный анализ результатов медицинских исследований

Нами изучены два метода моделирования выходного параметра Y: регрессионный анализ, когда все входные контролируемые факторы задаются количественно и дисперсионный анализ, когда входные контролируемые факторы в ПФЭ задаются на ограниченном числе качественных уровней. На практике встречаются случаи, когда в ПФЭ наряду с неколичественными факторами возможно измерять некоторое число сопутствующих количественных факторов (ковариат). В результате формируется матрица наблюдений с числом строк n, равным числу наблюдавшихся объектов и числом столбцов k+p+l, равным сумме числа неколичественных факторов k, числа количественных факторов (ковариат) p и числа показателей откликов на воздействие l. По этой матрице наблюдений в рамках модуля ANOVA/MANOVA ППП Statictica 5.0, с целью построения модели для исследуемого показателя-отклика Y, проводится ковариационный анализ, в результате которого получается уравнение, включающее эффекты, как неколичественных факторов, так и ковариат.

Ковариационный анализ выполняется в два этапа. На первом этапе проводится стандартный многомерный дисперсионный анализ показателей-откликов в зависимости от основных неколичественных факторов. Дисперсионный анализ заключается в разделении дисперсии показателей-откликов на составные части:

– дисперсию, обусловленную влиянием основных факторов и их взаимодействием;

– дисперсию, обусловленную влиянием других сопутствующих факторов и, в том числе, неконтролируемых и случайных факторов.

В результате дисперсионного анализа ПФЭ исследователь получает:

– оценки величины и значимости линейных эффектов и эффектов взаимодействия основных факторов на моделируемый показатель;

– средние значения показателя для различных уровней основных факторов и их сочетаний;

– оценки значимости различия средних значений показателя для различных уровней основных факторов и их сочетаний по множеству критериев (Фишера, Дункана, Тьюки, LSD и др.)

По данным первого этапа решения можно дать оценку степени влияния основных факторов на моделируемые показатели. Влияние сопутствующих количественных факторов на показатели-отклики на первом этапе не оценивается. Дисперсия показателей, обусловленная влиянием сопутствующих факторов, попадает в категорию дисперсии случайных и неконтролируемых факторов.

На втором этапе проводится многомерный регрессионный анализ с целью:

– определения коэффициентов линейных эффектов сопутствующих количественных факторов (ковариат) на показатель-отклик на фоне действия основных качественных факторов;

– оценки величины и значимости вклада в дисперсию показателя отклика сопутствующих количественных факторов (ковариат).

В результате дисперсия показателя-отклика вследствие неконтролируемых и случайных факторов существенно сокращается, т.к. из нее на втором этапе исключается дисперсия показателя за счет влияния сопутствующих количественных факторов (ковариат).

Таким образом, в итоге применения ковариационного анализа исследователь получает более содержательную информацию, на основе которой можно сделать более объективные выводы о характере изменения показателей-откликов на воздействие как основных, так и сопутствующих факторов, о степени влияния факторов, об оптимальных уровнях факторов, при которых показатели принимают требуемые значения.

Результаты ковариационного анализа позволяют исследователю решать основные задачи многомерного моделирования:

– оценивать характер изменения показателей-откликов при изменении воздействующих факторов;

– прогнозировать показатели-отклики при заданных значениях воздействующих факторов;

– определять степень влияния воздействующих факторов на моделируемые показатели-отклики;

– находить оптимальные уровни факторов, при которых показатели-отклики принимают желаемые или требуемые значения.

Характер изменения показателей изучают по таблицам (графикам) средних значений показателей для различных уровней факторов и их сочетаний. По этим таблицам определяют оптимальный уровень факторов, при которых показатели принимают требуемые значения. Степень влияния на показатели-отклики воздействующих факторов оценивают по результатам дисперсионного анализа.

Прогноз показателей-откликов для заданных уровней основных неколичественных факторов и значений сопутствующих количественных факторов (ковариат) выполняют по модели:

$$\hat{y}_{\mu} = \overline{y}_{\mu} + \sum_{s=1}^{p} b_s \times (x_s - \overline{x}_{s\mu}), \qquad (6.2)$$

где $\hat{y}_{\mu}$ – прогнозируемое значение показателя для μ-го сочетания основных неколичественных факторов;

$\overline{y}_{\mu}$ – среднее значение показателя для μ-го сочетания уровней факторов;

b_s - коэффициент регрессии показателя на изменение j-го сопутствующего количественного фактора x_j ;

x_s - заданное значение j-го сопутствующего фактора для прогноза выходного параметра;

$\overline{x}_{s\mu}$ - среднее значение j-го соответствующего фактора для данного μ-го сочетания уровней основных факторов;

p - число количественных факторов.

Методика проведения ковариационного анализа рассматривается в примере 6.2.

Содержание дисперсионного анализа дробного факторного эксперимента (ДФЭ) по планам латинских квадратов

ДФЭ целесообразно планировать когда число входных факторов k≥3. ПФЭ при таком количестве факторов становится неэкономич-

ным. Напротив, в ДФЭ значительно сокращается число опытных точек, т.к. опыты планируются не на всех сочетаниях уровней факторов, как в ПФЭ.

При числе факторов k=3 (A, B, C) применяется план по схеме латинского квадрата первого порядка. Вариант такого плана для трех факторов (k=3) каждый на трех уровнях (p=3) дан в табл.6.1.

Таблица 6.1

План ДФЭ по схеме латинского квадрата первого порядка

Уровни факторов		Фактор A		
		A_1	A_2	A_3
Фактор B	B_1	C_2	C_1	C_3
	B_2	C_3	C_2	C_1
	B_3	C_1	C_3	C_2

При числе факторов k=4 (A, B, C, D) применяется план по схеме латинского квадрата второго порядка. Вариант такого плана для четырех факторов (k=4) каждый на четырех уровнях (p=4) дан в табл.6.2.

При числе факторов k=5 (A, B, C, D, E) *применяется план по схеме* латинского квадрата третьего порядка.

В планах латинских квадратов число уровней p должно быть одинаковым для всех факторов. Количество ячеек плана (число опытных точек) в ДФЭ по схемам латинских квадратов $N=p^2$. Например, при p=4 $N=4^2=16$ ячеек как показано в табл.6.2. Количество наблюдений в каждой опытной точке может быть от 2 до 5 при малой и средней вариабельности параметра Y, а при значительной вариабельности параметра Y от 5 до 10 и более.

Таблица 6.2

План ДФЭ по схеме латинского квадрата второго порядка

Уровни факторов		Фактор A			
		A_1	A_2	A_3	A_4
Фактор B	B_1	C_1D_1	C_4D_2	C_3D_3	C_2D_4
	B_2	C_2D_3	C_1D_4	C_4D_1	C_3D_2
	B_3	C_3D_4	C_2D_3	C_1D_2	C_4D_1
	B_4	C_4D_2	C_3D_1	C_2D_4	C_1D_3

Планирование ДФЭ и его реализация выполняются по модулю Experimental Design ППП Statistica 5.0.

Планы составляются так, чтобы каждый уровень фактора был только один раз в каждой строке (столбце) плана. Каждое сочетание уровней факторов может быть в плане только один раз. По данным ДФЭ определяются линейные эффекты факторов, а эффекты взаимодействия попадают в эффект неконтролируемых, случайных факторов.

Результаты ДА ДФЭ включают:

– оценку линейных эффектов факторов и их значимости;

– расчет средних значений параметра для различных уровней факторов;

– построение графиков средних значений параметра для различных уровней факторов.

Последовательность решения задач дробного факторного эксперимента с помощью дисперсионного анализа приведена в примере 6.3.

ПРИМЕР 6.1

Доктор Глазников Л.А. (кафедра ЛОР ВМедА) исследовал влияние на вестибуло-вегетативную устойчивость (ВВУ) здоровых мужчин в возрасте 20-30 лет двух факторов:

A – специальной физической и аутогенной тренировки на трех уровнях (p=3);

B – медикаментозных средств, предупреждающих укачивание, на четырех уровнях (q=4).

1. Уровни фактора A:
   - A1 – систематическая тренировка в течение более 3 месяцев;
   - A2 – систематическая тренировка в течение 1-3 месяцев;
   - A3 – несистематическая тренировка.
2. Уровни фактора B:
   - B1 – алмид;
   - B2 – амтизол;
   - B3 – бемитил;
   - B4 – гутимин.

Испытанию подвергались лица со слабой вестибуло-вегетативной устойчивостью (со временем укачивания на кресле двойного вращения до неприятных ощущений не более 3 мин.).

В полном факторном эксперименте (ПФЭ) с числом опытных точек p×q=12 на каждом сочетании уровней факторов наблюдали 3 человек (n=3). Всего опытов проведено N=p×q×n=36.

Параметром, характеризующим влияние факторов A и B, являлось время укачивания до появления неприятных ощущений, X, мин. Результаты ПФЭ даны в табл.6.3. В машинограмме 6.1 представлены исходные данные в форме удобной для ввода в электронную таблицу.

Таблица 6.3

Время укачивания X, мин. в ПФЭ

Фактор A		A_1				A_2				A_3			
Фактор B		B_1	B_2	B_3	B_4	B_1	B_2	B_3	B_4	B_1	B_2	B_3	B_4
Наблюдения	1	15	12	8	7	12	10	7	6	12	5	5	4
	2	14	8	9	10	14	9	5	3	8	7	4	3
	3	15	10	6	4	13	6	6	5	6	4	3	4

Машинограмма 6.1

Матрица исходных данных

№ пп	A	B	X	№ пп	A	B	X
1	1	1	15	19	2	3	7
2	1	1	14	20	2	3	5
3	1	1	15	21	2	3	6
4	1	2	12	22	2	4	6
5	1	2	8	23	2	4	3
6	1	2	10	24	2	4	5
7	1	3	8	25	3	1	12
8	1	3	9	26	3	1	8
9	1	3	6	27	3	1	6
10	1	4	7	28	3	2	5
11	1	4	10	29	3	2	7
12	1	4	4	30	3	2	4
13	2	1	12	31	3	3	5
14	2	1	14	32	3	3	4
15	2	1	13	33	3	3	3
16	2	2	10	34	3	4	4
17	2	2	9	35	3	4	3
18	2	2	6	36	3	4	4

Требуется:

1. Провести дисперсионный анализ параметра X и оценить степень влияния факторов A и B, их взаимодействия на дисперсию параметра X.
2. Рассчитать средние значения и стандартные отклонения параметра X для всех уровней факторов A и B и их сочетаний.
3. Построить графики средних значений параметра X.
4. Оценить значимость различия средних значений параметра X попарно на уровнях факторов A, B и при их сочетании.

Решение дано с помощью персонального компьютера с использованием модуля ANOVA/MANOVA ППП Statistica 5.0.

1. Результаты дисперсионного анализа параметра X в полном факторном эксперименте на всех сочетаниях уровней факторов A и B приведены в машинограмме 6.2.

Машинограмма 6.2

Вклад эффектов факторов и их взаимодействия в дисперсию параметра-отклика

	Sum of Squares	df	Mean Square	F	p-level
Effect A	118,167	2	59,083	18,991	0,00001
Effect B	265,194	3	88,398	28,414	4,55E-08
Effect A×B	10,722	6	1,787	0,574	0,75
Error	74,667	24	3,111		

Степени влияния факторов на параметр X, рассчитанные по величине сумм квадратов SS с использованием формулы (6.1) даны в табл. 6.4.

Таблица 6.4

Оценка степени влияния факторов

Факторы	SS	K_j, %	p
A	118,2	25,2	<0,001
B	265,2	56,6	<0,001
A×B	10,7	2,3	>0,05
I. Контролируемые факторы	394,1	84,1	<0,001
II. Неконтролируемые случайные факторы и ошибки	74,7	15,9	
Все факторы	468,8	100,0	

Из таблицы 6.4 следует, что контролируемые факторы A и B и их взаимодействие объяснили основную часть дисперсии параметра X на 84,1%. Степень их влияния значима (p<0,001). Из контролируемых факторов наибольшее влияние оказывает фактор B (56,6%) и в меньшем масштабе - фактор A (25,2%). Степень влияния взаимодействия факторов A и B на дисперсию параметра X мала (2,3% с уровнем значимости p>0,05). Доля ошибок в дисперсии параметра X составляет 15,9%. Такая степень влияния ошибок в эксперименте вполне приемлема.

2.Средние значения и стандартные отклонения параметра X даны в машинограмме 6.3. Данные таблицы использованы для построения графиков средних значений параметра для всех уровней факторов A, B и их сочетаний.

3.Графики средних значений параметра X и 95%-х доверительных интервалов на различных уровнях факторов и при их сочетаниях даны на рисунках 6.1 - 6.3

4.Оценка значимости различия средних значений параметра X попарно на уровнях контролируемых факторов и при их сочетаниях выполнена по критерию LSD. Результаты сравнения приведены в машинограмме 6.4.

Машинограмма 6.3

Средние значения параметра-отклика на различных уровнях факторов и их сочетаниях

Summary Table of Means (disp_an.sta)
N=36 (No missing data in dep. var. list)

	X Means	X N	X Std.Dev.
A_1	9,83	12	3,56
A_2	8,00	12	3,54
A_3	5,42	12	2,57
B_1	12,11	9	3,14
B_2	7,89	9	2,62
B_3	5,89	9	1,90
B_4	5,11	9	2,26
A_1 B_1	14,67	3	0,58
A_1 B_2	10,00	3	2,00
A_1 B_3	7,67	3	1,53
A_1 B_4	7,00	3	3,00
A_2 B_1	13,00	3	1,00
A_2 B_2	8,33	3	2,08
A_2 B_3	6,00	3	1,00
A_2 B_4	4,67	3	1,53
A_3 B_1	8,67	3	3,06
A_3 B_2	5,33	3	1,53
A_3 B_3	4,00	3	1,00
A_3 B_4	3,67	3	0,58
All Groups	7,75	36	3,66

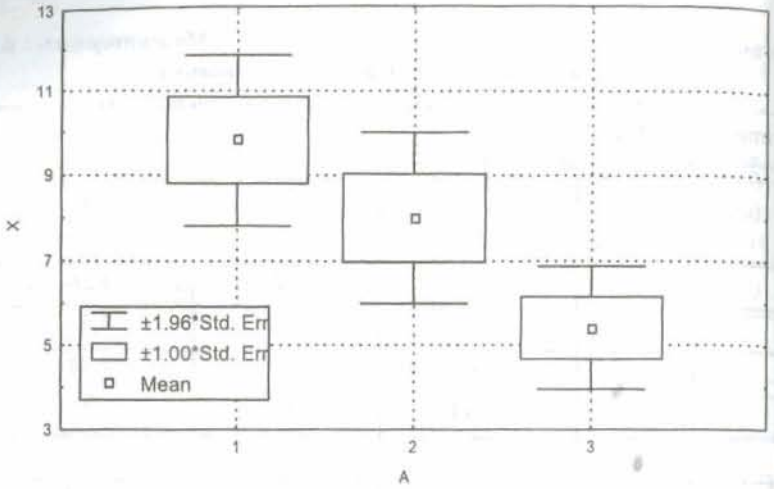

Рис.6.1. График средних значений вестибуло-вегетативной устойчивости на различных уровнях фактора А.

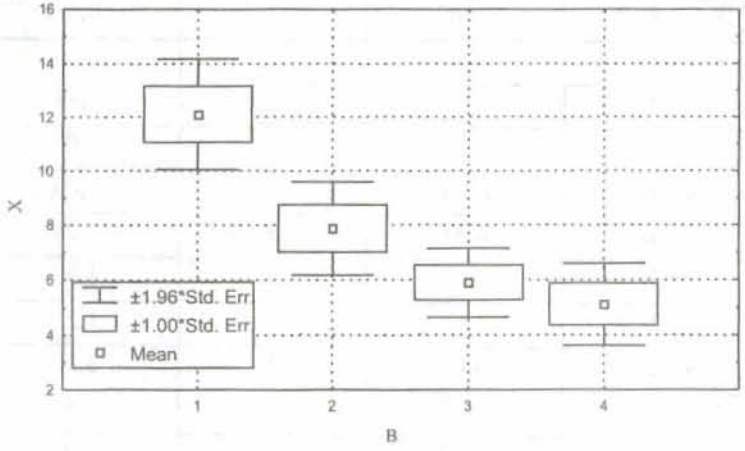

Рис.6.2. График средних значений вестибуло-вегетативной устойчивости на различных уровнях фактора В.

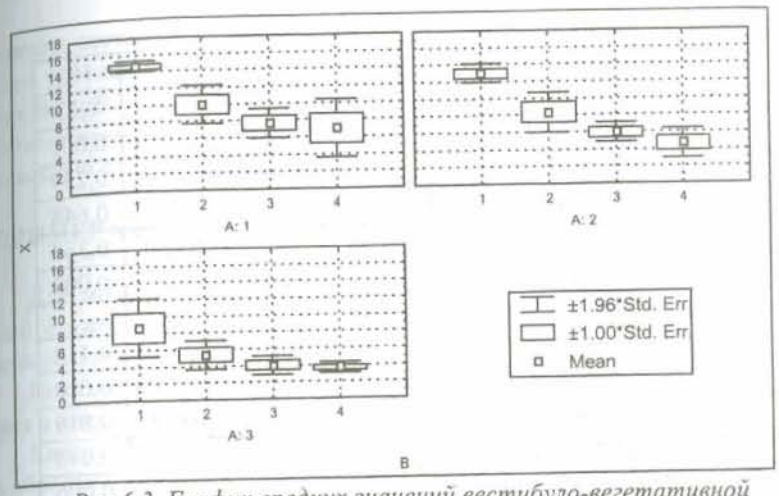

Рис.6.3. График средних значений вестибуло-вегетативной устойчивости на различных сочетаниях уровней факторов А и В.

Машинограмма 6.4

Оценка значимости различия средних значений параметра-отклика на различных уровнях факторов по критерию LSD

LSD test; variable X (disp_an.sta)
Probabilities for Post Hoc Tests
MAIN EFFECT: A

		{1}	{2}	{3}
		9,83	8,00	5,42
1	{1}		0,0177	2,46E-06
2	{2}	0,0177		0,0015
3	{3}	2,46E-06	0,0015	

MAIN EFFECT: B

		{1}	{2}	{3}	{4}
		12,11	7,89	5,89	5,11
1	{1}		3,41E-05	1,01E-07	1,26E-08
2	{2}	3,41E-05		0,0242	0,0027
3	{3}	1,01E-07	0,0242		0,3589
4	{4}	1,26E-08	0,0027	0,3589	

INTERACTION: 1 x 2

A	B	{1}	{2}	{3}	{4}	{5}	{6}
		14,67	10,00	7,67	7,00	13,00	8,33
1	1 {1}		0,003	5,9E-05	1,8E-05	0,259	0,000
1	2 {2}	0,003		0,118	0,048	0,048	0,259
1	3 {3}	5,9E-05	0,118		0,648	0,001	0,648
1	4 {4}	1,8E-05	0,048	0,648		0,000	0,364
2	1 {5}	0,259	0,048	0,001	0,000		0,003
2	2 {6}	0,000	0,259	0,648	0,364	0,003	
2	3 {7}	3,3E-06	0,010	0,259	0,494	5,9E-05	0,118
2	4 {8}	3,5E-07	0,001	0,048	0,118	5,8E-06	0,018
3	1 {9}	0,000	0,364	0,494	0,259	0,006	0,819
3	2 {10}	1,1E-06	0,003	0,118	0,259	1,8E-05	0,048
3	3 {11}	1,2E-07	0,000	0,018	0,048	1,9E-06	0,006
3	4 {12}	7,1E-08	0,000	0,010	0,030	1,1E-06	0,003

A	B	{7}	{8}	{9}	{10}	{11}	{12}
		6,0	4,67	8,67	5,33	4,00	3,67
1	1 {1}	3,3E-06	3,5E-07	0,000	1,1E-06	1,2E-07	7,1E-08
1	2 {2}	0,010	0,001	0,364	0,003	0,000	0,000
1	3 {3}	0,259	0,048	0,494	0,118	0,018	0,010
1	4 {4}	0,494	0,118	0,259	0,259	0,048	0,030
2	1 {5}	5,9E-05	0,000	0,006	1,8E-05	1,9E-06	1,1E-06
2	2 {6}	0,118	0,018	0,819	0,048	0,006	0,003
2	3 {7}		0,364	0,076	0,648	0,178	0,118
2	4 {8}	0,364		0,010	0,648	0,648	0,494
3	1 {9}	0,076	0,010		0,030	0,003	0,002
3	2 {10}	0,648	0,648	0,030		0,364	0,259
3	3 {11}	0,178	0,648	0,003	0,364		0,819
3	4 {12}	0,118	0,494	0,002	0,259	0,819	

Значимое различие средних арифметических значений по тесту LSD отмечено на всех трех уровнях фактора А и на трех уровнях фактора В. По данным таблиц и графиков видно, что значимую эффективность ($p<0,05$) в повышении вестибуло-вегетативной устойчивости

имеют: систематическая физическая и аутогенная тренировка (фактор А на уровнях А1 и А2) и применение медикаментов (фактор В на уровнях В1 и В2, т.е. алмида и амтизола). Совместное действие А и В увеличило время укачивания для лиц с вестибуло-вегетативной неустойчивостью от 3-4 до 10-15 мин.

ПРИМЕР 6.2

Для моделирования длительности лечения военнослужащих срочной службы с механической травмой в клиниках ВМедА выполнен ковариационный анализ по ПФЭ.

Показателем длительности лечения взят срок стационарного лечения в днях - SROKL.

Качественными факторами, влияющими на длительность лечения, выбраны:

– тяжесть состояния при поступлении в клинику - TIAJ, на трех уровнях:

1 - легкая,

2 - средняя,

3 - тяжелая;

– локализация травмы - MIKST, на трех уровнях:

1 - травма конечностей,

2 - травма груди или живота,

3 - сочетанная травма.

Сопутствующие количественные факторы:

–срок доставки в клинику с момента получения травмы - SROKD, ч;

–частота сердечных сокращений при поступлении в клинику - CHSS, уд./мин;

–систолическое артериальное давление при поступлении в клинику - AD, мм рт.ст.

Факторы TIAJ и MIKST в ПФЭ варьировались на 9 сочетаниях их уровней (3×3=9). На каждом сочетании уровней наблюдалось по три человека, всего 27 человек (9×3=27). Исходная матрица наблюдений 27 ×6 приведена в табл.6.5.

Таблица 6.5

Матрица исходных данных

case	TIAJ	MIKST	SROK_D	CHSS	AD	SROKL
1	2	1	2	100	110	58
2	1	2	2	98	125	43
3	2	3	4	130	80	110
4	1	2	2	84	120	48
5	3	1	3	115	80	96
6	1	1	1	84	120	28
7	1	2	2	80	115	41
8	2	2	2	96	120	64
9	2	2	2	78	110	78
10	2	3	4	95	70	115
11	1	1	1	64	140	15
12	2	2	2	84	120	64
13	1	1	2	68	110	35
14	2	1	2	110	125	49
15	1	1	1	78	140	28
16	3	2	4	130	70	112
17	2	3	3	120	80	88
18	2	1	2	110	85	77
19	1	3	1	78	140	41
20	1	3	1	72	130	36
21	3	3	4	140	45	120
22	3	3	3	110	65	100
23	2	3	3	105	65	98
24	1	3	2	90	130	45
25	2	1	2	84	110	58
26	3	1	3	120	60	98
27	3	2	4	130	65	100

Решение получено на ПК в модуле ANOVA/ MANOVA ППП Statistica 5.0 и дано в машинограммах 6.5-6.18.

Первый этап. Дисперсионный анализ двухфакторного ПФЭ (машинограмма 6.5).

Машинограмма 6.5

Эффекты основных факторов и их взаимодействия

Univar. Test	Sum of Squares	df	Mean Square	F	p-level
Effect TIAJ	18249,26	2	9124,631	106,64	0,00000
Effect MIKST	2314,039	2	1157,019	13,52	0,00026
Effect TIAJ×MIKST	1499,960	4	374,990	4,38	0,01196
Error	1540,08	18	85,560		

В машинограмме 6.5 показаны итоговые данные дисперсионного анализа всех эффектов. Из машинограммы следует, что значимыми ($p<0,01$) являются как линейные эффекты первого и второго факторов, так и эффект их взаимодействия. Эти данные позволяют оценить степень влияния основных факторов на показатель-отклик. Результаты расчетов по формуле (6.1) приведены в табл.6.6.

Таблица 6.6

Степень влияния факторов на показатель-отклик

Факторы	Сумма квадратов отклонений показателя SS	Степень влияния факторов k_j, %
TIAJ	18249,3	77,3
MIKST	2314,0	9,8
Взаимодействие факторов	1500,0	6,4
I. Контролируемые факторы	22063,3	93,5
II. Неконтролируемые и случайные факторы	1540,1	6,5
Все факторы	23603,4	100,0

В машинограмме 6.6 показаны средние значения срока лечения, полученные в эксперименте на различных уровнях основных факторов и их сочетаниях. Из этих данных следует, что среднее значение срока лечения увеличивается с возрастанием тяжести травмы и изменяется в зависимости от локализации (увеличение от 1-го к 3-му уровню фактора MIKST).

Машинограмма 6.6

Средние значения показателя SROKL на различных уровнях факторов и их сочетаний

TIAJ	MIKST	MEANS SROKL
1	.....	37,1
2	.....	77,3
3	.....	104,3
.....	1	61,3
.....	2	72,9
.....	3	84,5
1	1	26,5
1	2	44,0
1	3	40, 7
2	1	60,5
2	2	68, 7
2	3	102,7
3	1	97,0
3	2	106,0
3	3	110,0

Наименьшее среднее значение срока лечения (26,5 дн.) выявлено при легкой степени тяжести состояния пострадавших с травмой конечностей. Наибольшее (110 дн.) - при тяжелом состоянии пострадавших с сочетанной травмой конечностей, груди или (и) живота.

Значимость различия средних значений срока лечения для различных уровней основных факторов и их сочетаний оценена по критерию LSD TEST в машинограммах 6.7-6.9. Из машинограмм следует, что различия средних сроков лечения для трех уровней обоих факторов и большинства их сочетаний значимы ($p<0,01$).

Машинограмма 6.7

Результаты LSD-теста на уровнях фактора TIAJ

TIAJ	MIKST		{1} 37,0556	{2} 77,3056	{3} 104,3333
1	....	{1}		0,0000	0,0000
2	....	{2}	0,0000		0,0000
3	....	{3}	0,0000	0,0000	

Машинограмма 6.8

Результаты LSD-теста на уровнях фактора MIKST

TIAJ	MIKST		61,3333	72,8889	84,4722
....	1	{1}		0,0169	0,0000
....	2	{2}	0,0169		0,0190
....	3	{3}	0,0000	0,0190	

Машинограмма 6.9

Результаты LSD-теста на сочетаниях уровней факторов TIAJ и MIKST

TIAJ	MIKST		{1} 26,5000	{2} 44,0000	{3} 40,6667	{4} 60,5000	{5} 68,6667
1	1	{1}		0,0234	0,0602	0,0001	0,0000
1	2	{2}	0,0234		0,6642	0,0313	0,0043
1	3	{3}	0,0602	0,6642		0,0117	0,0016
2	1	{4}	0,0001	0,0313	0,0117		0,2628
2	2	{5}	0,0000	0,0043	0,0016	0,2628	
2	3	{6}	0,0000	0,0000	0,0000	0,0000	0,0001
3	1	{7}	0,0000	0,0000	0,0000	0,0002	0,0035
3	2	{8}	0,0000	0,0000	0,0000	0,0000	0,0003
3	3	{9}	0,0000	0,0000	0,0000	0,0000	0,0001

Продолжение машинограммы 6.9

TIAJ	MIKST		{6} 102,7500	{7} 97,0000	{8} 106,0000	{9} 110,0000
1	1	{1}	0,0000	0,0000	0,0000	0,0000
1	2	{2}	0,0000	0,0000	0,0000	0,0000
1	3	{3}	0,0000	0,0000	0,0000	0,0000
2	1	{4}	0,0000	0,0002	0,0000	0,0000
2	2	{5}	0,0001	0,0035	0,0003	0,0001
2	3	{6}		0,4821	0,6897	0,3774
3	1	{7}	0,4821		0,3435	0,1769
3	2	{8}	0,6897	0,3435		0,6706
3	3	{9}	0,3774	0,1769	0,6706	

Второй этап. Регрессионный анализ вклада сопутствующих факторов (машинограммы 6.10-6.17).

Машинограмма 6.10

Дисперсионный анализ уравнения регрессии со всеми ковариатами

Univar. Test	Sum of Squares	df	Mean Square	F	p-level
Effect	921,6674	3	307,223	7,451841	0,00278
Error	618,4159	15	41,228		

Машинограмма 6.11

Коэффициенты регрессии ковариат в модели со всеми ковариатами

general manova

Regression Results, Dependent Variable: SROKL
Multiple R: .7735974 R-Square: .5984530
F(3,15) = 7.451841 p = .00278

Variable	B-weight	Standard Error	Beta	t(15)	p-level
SROK_D	10,52251	4,133049	0,457921	2,54594	0,022377
CHSS	0,02478	0,134925	0,030969	0,18367	0,856736
AD	-0,39386	0,144281	-0,477603	-2,72984	0,015502

Машинограмма 6.12

Дисперсионный анализ уравнения регрессии с ковариатой SROK_D

Univar. Test	Sum of Squares	df	Mean Square	F	p-level
Effect	614,4381	1	614,438	11,2845	0,00372
Error	925,6453	17	54,450		

Машинограмма 6.13

Коэффициент регрессии ковариаты SROK_D в модели с одной ковариатой

Regression Results, Dependent Variable: SROKL
R: .6316361 R-Square: .3989642
F(1,17) = 11.28450 p = .00372

Variable	B-weight	Standard Error	Beta	t(17)	p-level
SROK_D	14,51429	4,320703	0,631636	3,359242	0,003722

Машинограмма 6.14

Дисперсионный анализ уравнения регрессии с ковариатой CHSS

Effect	33,417	1	33,417	0,377045	0,54732
Error	1506,667	17	88,627		

Машинограмма 6.15

Коэффициент регрессии ковариаты CHSS в модели с одной ковариатой

Regression Results, Dependent Variable: SROKL
R: .1473019 R-Square: .0216979
F(1,17) = .3770445 p = .54732

Variable	B-weight	Standard Error	Beta	t(17)	p-level
CHSS	0,117871	0,19196	0,147302	0,61404	0,547319

Машинограмма 6.16

Дисперсионный анализ уравнения регрессии с ковариатой AD

Univar. Test	Sum of Squares	df	Mean Square	F	p-level
Effect	626,6392	1	626,639	11,66231	0,00330
Error	913,4442	17	53,732		

Машинограмма 6.17

Коэффициент регрессии ковариаты AD в модели с одной ковариатой

Regression Results, Dependent Variable: SROKL

R: .6378766 R-Square: .4068865

F(1,17) = 11.66231 p = .00330

Variable	B-weight	Standard Error	Beta	t(17)	p-level
AD	-0,526035	0,154036	-0,637877	-3,41501	0,003299

В машинограмме 6.10 приведена оценка эффекта всех сопутствующих факторов, в машинограммах 6.12, 6.14, 6.16 - каждого в отдельности. По данным этих машинограмм рассчитана степень влияния ковариат на показатель - срок лечения. Результаты расчета по формуле (6.1) представлены в табл.6.7.

Таблица 6.7

Степень влияния ковариат на показатель-отклик

Факторы	Сумма квадратов отклонений показателя SS	Степень влияния факторов k_j, %
Совокупность сопутствующих факторов	921,7	3,5
SROKD	614,4	2,6
CHSS	33,4	0,0
AD	626,6	2,6

В машинограммах 6.11, 6.13, 6.15 и 6.17 даны коэффициенты регрессии ковариат и оценки их значимости (6.11 - модель со всеми ковариатами; остальные - модели с одной ковариатой). Значимыми являются коэффициенты регрессии ковариат SROKD и AD ($p<0,05$). Коэффициент регрессии ковариаты CHSS оказался незначимым ($p>0,05$).

В результате второго этапа установлено значимое влияние на срок лечения травмы сопутствующих факторов: срока доставки пострадавших в клинику и артериального давления при поступлении в клинику ($p<0,05$). Влияние частоты сердечных сокращений оказалось слабым и незначимым ($p>0,05$).

Дадим прогноз длительности лечения (в стационаре) для условий сочетания уровней основных неколичественных факторов, т.е., когда пострадавший имеет травму груди (уровень - 2), состояние при первичном осмотре оценено как средне тяжелое (уровень - 2). Значения сопутствующих факторов: срок доставки в клинику - 2 ч, частота сердечных сокращений - 78 уд./мин, артериальное давление - 110 мм рт. ст. Для данного сочетания уровней факторов: среднее значение срока лечения $\bar{y}_{22} = 68,7$дн.; среднее значение ковариат $\bar{x}_1 = 2$ ч.; $\bar{x}_2 = 86$уд/мин; $\bar{x}_3 = 117$мм рт. ст. (машинограмма 2.2.18); коэффициенты регрессии b_1=10,523; b_2=0,025; b_3=-0,394 (машинограмма 6.11).

Машинограмма 6.18

Средние значения показателя и ковариат на сочетаниях уровней основных факторов

general TIAJ	manova MIKST	MEANS			
		SROKL	Covar. SROK_D	Covar. CHSS	Covar. AD
1	1	26,5000	1,2500	73,5000	127,5000
1	2	44,0000	2,0000	87,3333	120,0000
1	3	40,6667	1,3333	80,0000	133,3333
2	1	60,5000	2,0000	101,0000	107,5000
2	2	68,6667	2,0000	86,0000	116,6667
2	3	102,7500	3,5000	112,5000	73,7500
3	1	97,0000	3,0000	117,5000	70,0000
3	2	106,0000	4,0000	130,0000	67,5000
3	3	110,0000	3,5000	125,0000	55,0000

По формуле (6.2) получаем

$\hat{y}_{22}$=68,7+10,523×(2-2)+0,025×(78-86)-0,394×(110-117)=71,3дн.

Заметим, что среднее значение срока лечения в данных условиях $\bar{y}_{22}$=68,7дн. Различие обусловлено влиянием на результат заданных значений ковариат.

Выводы:

1. На сроки лечения механической травмы военнослужащих срочной службы основное влияние оказывают факторы: тяжесть состояния (77,3%) и локализация повреждения (9,8%). С учетом эффекта взаимодействия (6,4%) степень влияния этих факторов возрастает до 93,5%.

2. Из числа исследованных трех сопутствующих факторов значимое влияние ($p<0,05$) на срок лечения травм установлено для факторов

срока доставки в клинику и артериального давления при поступлении пострадавшего в клинику (3,5%).

3. Сопутствующий фактор частоты сердечных сокращений при поступлении в клинику не оказал на срок лечения значимого влияния.

ПРИМЕР 6.3

Для исследования изменений белкового обмена в зависимости от рациона питания, уровня физической нагрузки и физического развития у здоровых молодых мужчин спланирован дробный факторный эксперимент по схеме латинского квадрата второго порядка (греко-латинского квадрата) (В.А.Майдан, 1990).

В эксперименте изучалось влияние четырех факторов на белковый обмен:

A - суточное потребление белка, г;

B - суточные энерготраты, ккал;

C - доля времени субмаксимальных физических нагрузок (>60% максимального потребления кислорода), %;

D - тощая масса тела, кг.

Все факторы испытывались на трех уровнях p=3 (табл.6.8). В качестве параметра, характеризующего белковый обмен, применялся показатель содержания общего азота в моче и поте Y, г/сут. В каждой опытной точке (ячейке плана ДФЭ) наблюдалось по 5 здоровых мужчин в возрасте 20-30 лет (n=5). В плане ДФЭ по схеме латинского квадрата при исследовании четырех факторов на трех уровнях, в каждом содержится 9 ячеек ($N=p^2=3^2=9$).

Выбранный исследователем вариант плана и результаты эксперимента приведены в табл.6.9 и машинограмме 6.19.

Таблица 6.8

Уровни варьирования факторов

Фактор	1	2	3
A	65-94 г	95-124 г	125-155 г
B	1800-3200 ккал	3300-4700 ккал	4800-6200 ккал
C	до 4%	5-9%	10 и более %
D	40-49 кг	50-59 кг	60-70 кг

Таблица 6.9

План и результаты эксперимента (параметр Y, г/сут) по схеме латинского квадрата второго порядка

	B_1		B_2		B_3	
A_1	C_2D_3	7	C_3D_1	8	C_1D_2	5
		10		10		7
		15		11		9
		10		12		11
		13		14		13
A_2	C_3D_2	16	C_1D_3	10	C_2D_1	17
		17		13		12
		11		19		13
		9		16		14
		15		12		9
A_3	C_1D_1	15	C_2D_2	16	C_3D_3	28
		17		18		23
		11		29		23
		9		22		30
		13		15		29

В машинограмме 6.19 показана матрица исходных данных. В столбцах A, B, C, D указаны уровни факторов, в столбце Y - результаты наблюдения параметра Y, г/сут. Число строк (45) соответствует общему числу наблюдений (в 9 опытных точках по 5 повторностей в каждой, всего N×n=45).

Машинограмма 6.19

Экранная форма матрицы наблюдений

№ пп	A	B	C	D	Y
1	1	1	2	3	7
2	1	2	3	1	8
3	1	3	1	2	5
4	2	1	3	2	16
5	2	2	1	3	10
6	2	3	2	1	17
7	3	1	1	1	15
8	3	2	2	2	16
9	3	3	3	3	28
10	1	1	2	3	10
11	1	2	3	1	10
12	1	3	1	2	7
13	2	1	3	2	17
14	2	2	1	3	13
15	2	3	2	1	12
16	3	1	1	1	17
17	3	2	2	2	18
18	3	3	3	3	23
19	1	1	2	3	15
20	1	2	3	1	11
21	1	3	1	2	9
22	2	1	3	2	11
23	2	2	1	3	19
24	2	3	2	1	13
25	3	1	1	1	11
26	3	2	2	2	29
27	3	3	3	3	23
28	1	1	2	3	10
29	1	2	3	1	12
30	1	3	1	2	11
31	2	1	3	2	9
32	2	2	1	3	16
33	2	3	2	1	14
34	3	1	1	1	9
35	3	2	2	2	22
36	3	3	3	3	30
37	1	1	2	3	13
38	1	2	3	1	14
39	1	3	1	2	13
40	2	1	3	2	12
41	2	2	1	3	12
42	2	3	2	1	9
43	3	1	1	1	13
44	3	2	2	2	15
45	3	3	3	3	29

Решение по разделу Latin squares. Greco-Latin squares модуля Experimental Design включает:

– дисперсионный анализ данных эксперимента (машинограмма 6.20);

– таблицу средних значений параметра Y (машинограмма 6.21);

– график средних значений параметра Y для уровней четырех факторов (рис.6.4).

Машинограмма 6.20

Экранная форма результатов дисперсионного анализа

Analysis of Variance (pr_4_1) 3 by 3 Greco Latin Square
Y; Mean = 14,5111 Sigma = 6,09628

	SS	Df	MS	F	p
A	712,84	2	356,42	28,95	3,201E-08
B	117,51	2	58,76	4,77	0,015
C	178,18	2	89,09	7,24	0,002
D	183,51	2	91,76	7,45	0,002
Residual	443,20	36	12,31		

Дисперсионный анализ (машинограмма 6.20) показал, что линейные эффекты четырех факторов на параметр Y значимы ($p<0,01$). Суммы квадратов отклонений параметра от среднего значения, вызываемых воздействием:

– контролируемых факторов $SS_I=SS_A+SS_B+SS_C+SS_D=1192,04$;

– неконтролируемых, случайных факторов и ошибок измерений $SS_{II}=443,20$;

– всех факторов $SS=SS_I+SS_{II}=1635,24$.

Степень влияния факторов на параметр Y определяется отношением (6.1). Например, для фактора A:

$$K_A = \frac{100 \times SS_A}{SS} = \frac{100 \times 712,84}{1635,24} = 43,6\%.$$

Величина степени влияния других факторов: $K_B=7,2\%$; $K_C=10,9\%$; $K_D=11,2\%$, в целом контролируемых факторов - $K_I=72,9\%$; неконтролируемых - $K_{II}=27,1\%$.

Следовательно, наибольшее влияние на белковый обмен оказывает фактор A - суточное потребление белка (43,6%) и в меньшей - факторы D - тощая масса тела (11,2%), C - доля времени субмаксимальных физических нагрузок (10,9%), B - суточные энерготраты (7,2%). Степень влияния неконтролируемых факторов (27,1%), в основном, может быть отнесена на индивидуальные особенности испытуемых и ошибки измерения остаточного азота в моче и поте испытателей.

Средние значения параметра Y (Means), отклонения от средних значений (Paramet. Estimate) и средние квадратичные отклонения средних значений в ячейках плана (Std. Dev.) для трех уровней каждого фактора представлены в машинограмме 6.21. Анализ этих показате-

лей свидетельствует, что средние значения общего азота в моче и поте испытателей, полученные в эксперименте, заключаются в интервале от 8 до 20 г/сут.

Машинограмма 6.2

Экранная форма результатов описательной статистики по уровням факторов

Means and Standard Deviations (pr_4_1) 3 by 3 Greco Latin Square
Y; Mean = 14,5111 Sigma = 6,09628

	Level	Means	Paramet. Estimate	Std.Dev.
A	1	10,33	-4,18	1,15
	2	13,33	-1,18	0,58
	3	19,87	5,36	6,80
B	1	12,33	-2,18	1,15
	2	15,00	0,49	4,58
	3	16,20	1,69	9,23
C	1	12,00	-2,51	2,65
	2	14,67	0,16	4,73
	3	16,87	2,36	8,49
D	1	12,33	-2,18	1,15
	2	14,00	-0,51	5,57
	3	17,20	2,69	8,28

С увеличением уровней всех факторов наблюдается рост параметра Y. Средний прирост общего азота Δy при изменении фактора на одну единицу зарегистрирован:

– при увеличении потребления белка на 1 г Δy_A=0,16 г/сут;

– при увеличении суточных энерготрат на 1 ккал Δy_B=0,0013 г/сут;

– при увеличении субмаксимальных физических нагрузок на 1 % Δy_C=0,49 г/сут;

– при увеличении тощей массы тела на 1 кг Δy_D=0,25 г/сут.

Наглядное представление о характере изменения средних значений параметра Y на различных уровнях факторов можно получить из рис.6.4. Наибольший размах средних значений от 10,3 до 19,9 г/сут для фактора А, подтверждает высокую степень влияния его на параметр Y.

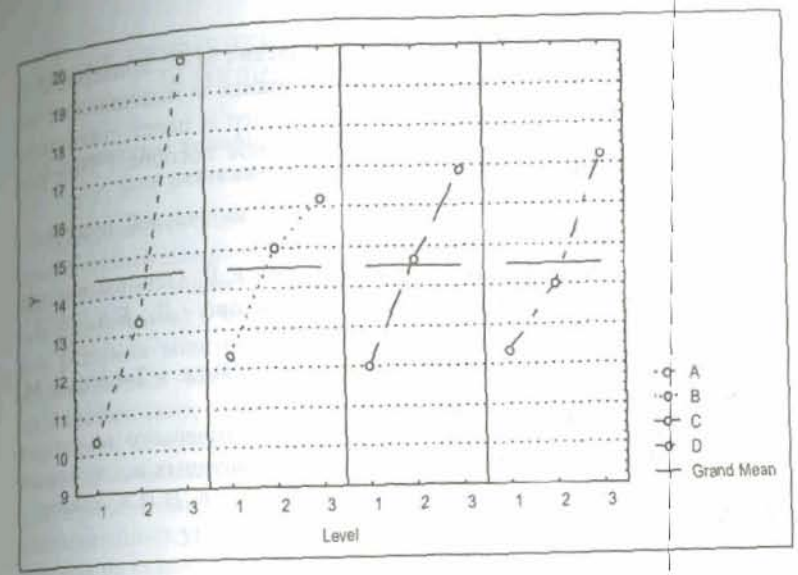

Рис.6.4. Средние значения параметра Y по уровням факторов.

Также, как наименьший размах от 12,3 до 16,2 г/сут для фактора В - низкую степень влияния этого фактора.

В данном случае применение дисперсионного анализа эксперимента по схеме латинского квадрата второго порядка с помощью ППП Statistica позволило решить следующие задачи исследования:

–изучен характер изменения белкового обмена в организме в зависимости от четырех факторов: суточного потребления белка в рационе питания, уровня физической нагрузки и физического развития испытуемых;

–оценена степень влияния факторов на параметр, характеризующий белковый обмен в организме.

Литература

1. Григорьев С.Г., Киреев О.В., Кувакин В.И., Левандовский В.В., Лядов В.Р., Юнкеров В.И., Мизерене Р.В., Николаевич М.С. Многомерные методы статистического анализа категорированных данных медицинских исследований. СПб, 1998. - 103 с.

2. Григорьев С.Г., Кувакин В.И., Николаевич М.С., Юнкеров В.И. Применение теории планирования эксперимента в медицинских исследованиях (включая персональный компьютер и пакет статистических программ Statistica for Windows): Учебное пособие / Под ред. В.И.Кувакина. - СПб, 1999. - 86 с.

3. Математико-статистические методы в клинической практике / Под ред. В.И.Кувакина. - СПб,: Б.и, 1993. - 199 с.

4. Поляков Л.Е., Игнатович Б.И, Лашков К.В. Основы военно-медицинской статистики /Под ред. Л.Е.Полякова - Л.: Б.и, 1977. - 336 с.

5. Шеффе Г. Дисперсионный анализ. Пер. с англ. - 2-е изд. - М.: Наука, 1980.- 512с.

6. Юнкеров В.И. Основы математико-статистического моделирования и применения вычислительной техники в научных исследованиях: Лекции для адъюнктов и аспирантов / Под ред. В.И.Кувакина. - СПб,: Б.и, 2000. - 140 с.

Глава 7. ПРИМЕНЕНИЕ ДИСКРИМИНАНТНОГО АНАЛИЗА В МЕДИЦИНСКОЙ ДИАГНОСТИКЕ

Сущность и условия применения дискриминантного анализа для решения задачи медицинской диагностики

Дискриминантный анализ - это метод многомерной статистики, применяемый для решения задач классификации (распознавания образов) и позволяющий отнести объект с определенным набором признаков (симптомов) к одному из известных классов. Метод применяется для решения многих медико-биологических задач. В медицине дискриминантный анализ используется для решения диагностических, прогностических, экспертных задач, задач профотбора, выбора методов и схем лечения.

Перечисленные выше задачи выполняются по решающим правилам, представляющим собой линейные классификационные функции в виде линейных уравнений, выработанным методами дискриминантного анализа на основе обучающей информации.

Обучающая информация формируется по результатам обследования объектов (пациентов), характеризующихся множеством признаков (симптомов) и достоверно установленным фактом принадлежности к одному из дифференцируемых состояний. Она представляет собой матрицу наблюдений размером n×(k+1), где n - число строк равных числу обследованных объектов (больных) с достоверно установленным состоянием (диагнозом определенного заболевания); k+1 - число столбцов, состоящих из k диагностических признаков (симптомов заболеваний) и 1 - группировочного признака G1, содержащего коды классов состояний (диагностируемых заболеваний) в виде чисел натурального ряда 1, 2 и т.д.

Отнесение объекта (больного) к определенному классу выполняется по набору его симптомов на основе расчета линейных дискриминантных функций (ЛДФ).

Надежность применения дискриминантного анализа обеспечивается достоверностью обучающей информации и достаточным количеством объектов в матрице наблюдений по каждому классу состояний от нескольких десятков до нескольких сотен. Число признаков в матрице наблюдений не ограничивается. Однако, для решения диагностической задачи по программе дискриминантного анализа автоматически отбирается ограниченное число наиболее

информативных признаков (обычно до 5-10 признаков). Признаки, включаемые в матрицу наблюдений, могут быть как количественными так и качественными, но при этом все они должны оцениваться количественно или в баллах по степени их выраженности. Важно чтобы коды степени проявления всех качественных признаков давались однотипно или по возрастанию, или по убыванию их выраженности. Например, число лейкоцитов в крови до 8 тыс/мкл обозначается 1; от 8 до 14 тыс/мкл - 2; свыше 14 тыс/мкл - 3.

Возможный набор симптомов для медицинской диагностики острого аппендицита и массив обучающей информации даны в примере 7.1 (табл.7.3 и 7.4).

Этапы применения дискриминантного анализа

Медицинская диагностика с применением дискриминантного анализа выполняется в три этапа.

На первом этапе формируется обучающая информация. Определение видов заболеваний для диагностики, диагностических признаков осуществляется врачом специалистом. Отбор объектов в матрицу наблюдений производится в клинике из историй болезни или специально разработанного для этой цели первичного учетного документа. Из историй болезни берут количественные значения признаков (в натуральных единицах измерения или баллах). Все данные должны быть тщательно проверены. Достоверность обучающей информации определяет надежность решающих правил диагностики.

На втором этапе вырабатываются решающие правила и дается оценка их информативности. Модуль Discriminant Analysis ППП Statistica обеспечивает пошаговый отбор информативных признаков и получение решающих правил в виде линейных классификационных функций (ЛКФ) и канонических линейных дискриминантных функций (КЛДФ). Качество выработанных правил оценивается сопоставлением результатов классификации с исходной классификацией объектов в обучающей матрице. Для наглядности выдается график положения объектов всех групп в координатах двух первых наиболее значимых КЛДФ.

На третьем этапе непосредственно решается задача медицинской диагностики по выработанным решающим правилам. После обследования больного определяются количественные значения симптомов, включенных в ЛКФ или КЛДФ, рассчитываются эти функции и по их

величинам дается решение об отнесении больного к той или иной группе заболеваний. Расчет ЛКФ или КЛДФ выполняется с применением персонального компьютера или микрокалькулятора.

Если используются ЛКФ, то отнесение больного к определенной группе выполняется по максимальному значению ЛКФ после их расчета по набору симптомов больного для каждой группы заболеваний.

При применении КЛДФ также производится расчет КЛДФ по значению симптомов конкретного больного. Отнесение больного к определенной группе выполняется после нанесения значений рассчитанных КЛДФ на диагностический график. Объект относят к той группе, для которой его удаление от соответствующего центроида окажется минимальным.

Отбор информативных симптомов для включения в модели ЛКФ и КЛДФ

Информативность симптомов содержащихся в матрице наблюдений, оценивается по F-критерию Фишера:

$$F = \frac{S_B^2}{S_W^2}, \qquad (7.1)$$

где S_B^2 - межгрупповая дисперсия симптома;

S_W^2 - внутригрупповая дисперсия симптома.

Очевидно, чем больше S_B^2 и меньше S_W^2 тем больше диагностическая информативность симптома.

В модель включаются симптомы, для которых уровень значимости по F-критерию $p \le 0{,}05$. Для выполнения этого требования величина критерия F при пошаговом дискриминантном анализе задается в пределах от 1 до 3.

В машинограммах, содержащих оценку информативности симптомов включенных и не включенных в ЛДФ, даются дополняющие сведения об информативности по критериям Wilks`Lambda, Partial Lambda и толерантности симптомов.

В примере 7.1 в машинограммах 7.1 и 7.2 указаны симптомы, включенные и не включенные в модели ЛДФ при заданном значении критерия для включения симптомов в модель F=2.

Решение диагностической задачи по линейным классификационным функциям (ЛКФ)

Для каждой группы заболеваний определяется обобщающая все включенные в модель симптомы линейная классификационная функция

$$ЛКФ_j=b_0+b_1x_1+b_2x_2+\ldots b_kx_k, \qquad (7.2)$$

где $ЛКФ_j$ - линейная классификационная функция для j-й группы заболеваний;

b_0 - константа;

$b_1, b_2, \ldots b_k$ - коэффициенты для симптомов X1, X2,...Xk;

$x_1, x_2,\ldots x_k$ - возможные значения k симптомов.

Количество таких функций определяется числом диагностируемых групп заболеваний (m).

Для решения диагностической задачи по симптомам больного производится расчет m ЛКФ. Больного относят к той группе, для которой ЛКФ примет максимальное значение.

Более предпочтительным по объему расчетов является решение диагностической задачи по каноническим линейным дискриминантным функциям.

Решение диагностической задачи по каноническим линейным дискриминантным функциям (КЛДФ)

Для всех групп заболеваний определяются 1-3 канонических ЛДФ, обобщающих данные о всех симптомах, включенных в модель, для всех больных, находящихся в обучающей матрице наблюдений. Первая КЛДФ1 охватывает наибольшую часть дисперсии симптомов, вторая КЛДФ2 - наибольшую часть из оставшихся дисперсий признаков и т.д. Вклад КЛДФ в межгрупповую дисперсию симптомов (Eigenvalue) оценивается по χ^2-критерию Пирсона. Этот вклад признается значимым при уровне значимости $p\leq0,05$.

В примере 7.1 значимыми получены две КЛДФ, обозначенные F1 и F2, о чем свидетельствуют данные собственных вкладов функций (машинограмма 7.4). В машинограмме 7.5 даны коэффициенты КЛДФ, их собственные вклады и кумулятивный вклад (Cum.Prop). Так, F1 и F2 обобщили дисперсию всех симптомов на 97,6% (0,9757). Там же приведены формулы для расчета F1 и F2.

По этим формулам программой предусмотрен расчет [illegible] всех объектов обучающей информации и построение полей (облаков) точек объектов всех групп заболеваний. График положения объектов четырех групп в координатах первой и второй КЛДФ дан на рис.7.1. По данным о координатах объектов в группах заболеваний производится расчет координат центроидов для каждой группы (Means of Canonical Variables - в машинограмме 7.7). По этим координатам центроиды наносят на график. От них измеряется удаление до точки обследуемого больного, которую наносят на график после расчета F1 и F2 по симптомам обследуемого больного. Больного относят к той группе, от центра которой получено наименьшее удаление.

По таблице факторной структуры КЛДФ (Factor Structure), данной в машинограмме 7.6, судят о корреляционной связи наблюдавшихся симптомов, включенных в модели, с каноническими ЛДФ. С первой канонической ЛДФ более тесно связаны симптомы X8, X6 и X7, со второй канонической ЛДФ - симптомы X2 и X3 . Данные о факторной структуре канонических ЛДФ могут использоваться для оценки коэффициентов весомости симптомов при решения диагностической задачи.

Применение решающих правил диагностики

Применение решающих правил диагностики острого аппендицита - КЛДФ F1 и F2 и диагностического графика (рис.7.1) покажем на примере для двух обследованных с жалобами на боли в животе. Для этого представим графики с положением центроидов четырех диагностируемых групп C_1 - C_4 в более крупном масштабе. Координаты центроидов возьмем из машинограммы 7.7.

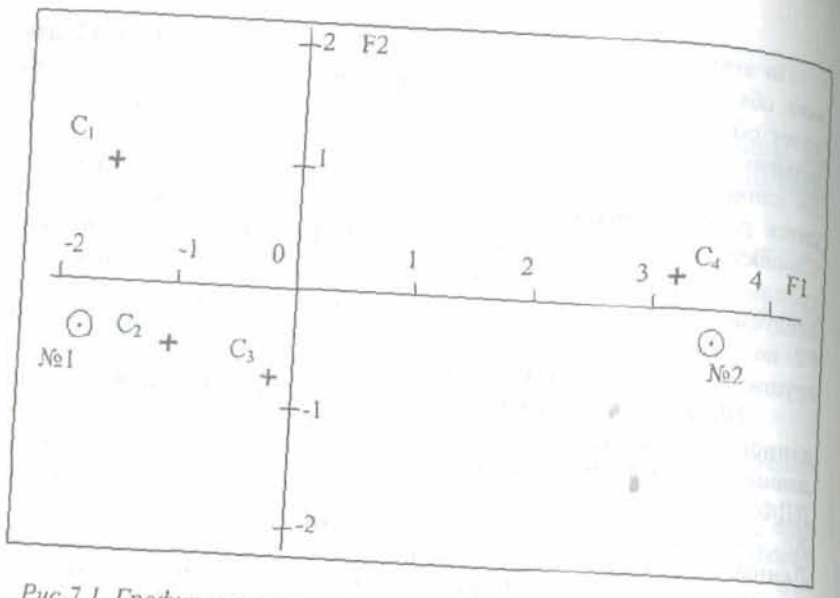

Рис.7.1. График положения центроидов C_1 - C_4 для четырех диагностируемых групп. В точке 1 - положение обследуемого №1, в точке 2 - положение обследуемого №2.

В результате обследования двух человек с жалобами на боли в животе получены оценки в баллах по всем семи симптомам, включенным в модели для двух канонических ЛДФ (табл.7.1).

Таблица 7.1

Оценка симптомов в баллах

№ обследованного	X_8	X_6	X_2	X_7	X_1	X_3	X_4
1	2	2	3	2	2	1	2
2	0	0	2	0	1	1	1

По формулам F1 и F2 (пример 7.1) рассчитаны координаты обследованных:

обследованного №1

F1=5,78-0,59×2-0,69×2-0,42×3-0,57×2-0,71×2-0,38×1-0,40×2=-1,78,

F2=-0,39+0,24×2+0,08×2-0,92×3-0,36×2+0,07×2+1,03×1+0,81×2=-0,44;

обследованного №2

F1=5,78-0,59×0-0,69×0-0,42×2-0,57×0-0,71×1-0,38×1-0,40×1=3,45,

F2=-0,39+0,24×0+0,08×0-0,92×2-0,36×0+0,07×1+1,03×1+0,81×1=-0,32.

Расчет может быть выполнен даже с помощью микрокалькулятора. Точки обследованных нанесены по F1 и F2 на график. По наименьшему удалению от центроидов установлено, что обследованных следует отнести к группам:

–первого обследованного - к группе №2 - флегмонозный аппендицит;

–второго обследованного - к группе №4 - неподтвержденный диагноз острого аппендицита.

Оценка эффективности решающих правил диагностики

Точность диагностики по решающим правилам, под которой понимают относительную частоту правильного отнесения объектов обучающей матрицы наблюдений к своей группе, показана в машинограмме 7.8. Из этой таблицы следует, что частость правильного диагноза для объектов четвертой группы (неподтвержденный диагноз острого аппендицита) составляет 100%. Для групп 1, 2, 3 частости правильного диагноза значительно меньше и равны соответственно 78,57%; 60,00%; 65,38%. Недостаточная точность диагностики для групп 2 и 3 объясняется значительным перекрытием симптомов для этих групп острого аппендицита, что хорошо видно на рис.7.2.

В практике оценки эффективности решающих правил диагностики применяют не только показатель относительной частоты правильного диагноза, предлагаемый модулем ППП Statistica, который получил название чувствительности, но и такие показатели, как специфичность, безошибочность, показатели ложно отрицательных ответов (ошибки первого рода) и ложноположительных ответов (ошибки второго рода).

Сущность названных показателей эффективности метода диагностики заключается в следующем.

Чувствительность - это относительная частота отнесения истинно больного к классу больных.

Специфичность - это относительная частота отнесения истинно здорового к классу здоровых.

Безошибочность - это относительная частота принятия безошибочных решений, как по отношению к истинно больным, так и истинно здоровым.

Ложноотрицательный ответ (ошибка первого рода) - это относительная частота отнесения истинно больного к классу здоровых.

Ложноположительный ответ (ошибка второго рода) - это относительная частота отнесения истинно здорового к классу больных.

Естественно требовать, чтобы ошибка первого рода была меньше чем ошибка второго рода.

Оценим эффективность полученных решающих правил по установлению диагноза об остром аппендиците. Для этого все объекты обучающей выборки разобьем на две группы: в первую группу отнесем объекты с заболеванием 1, 2, 3 групп; во вторую - объекты группы 4 с неподтвержденным диагнозом острого аппендицита. Результаты сведем в табл.7.2.

Таблица 7.2

Частотная таблица для двух диагностируемых групп (первая - заболевания 1, 2, 3; вторая - неподтвержденные заболевания 4)

Группы объектов обучающей выборки	Результаты классификации по решающим правилам				Всего объектов	
	отнесены к первой группе. (1, 2, 3)		отнесены ко второй группе (4)			
Первая (1, 2, 3)	a	77	b	2	a+b	79
Вторая (4)	c	0	d	24	c+d	24
Всего	**a + c**	**77**	**b + d**	**26**	**a+b+c+d**	**103**

Обозначим ячейки частотной таблицы строчными буквами латинского алфавита a, b, c, d.

При сделанных обозначениях частот наблюдений в табл.7.2 показатели эффективности диагностики по решающим правилам можно определить:

– чувствительность $\frac{100 \times a}{a+b} = \frac{100 \times 77}{79} = 97{,}5\%$;

– специфичность $\frac{100 \times d}{c+d} = \frac{100 \times 24}{24} = 100{,}0\%$;

– безошибочность $\frac{100 \times (a+d)}{(a+b+c+d)} = \frac{100 \times (77+24)}{103} = 98{,}1\%$;

– ложноотрицательный ответ (ошибка первого рода)

$$\frac{100 \times b}{a+b} = \frac{100 \times 2}{79} = 2{,}5\%;$$

– ложноположительный ответ (ошибка второго рода)

$$\frac{100 \times c}{c+d} = \frac{100 \times 0}{24} = 0\%.$$

На основе проведенных расчетов выработанные решающие правила диагностики следует признать весьма эффективными. Объекты с действительным заболеванием выявлены на 97,5%. Ошибки отнесения объектов с действительным заболеванием к группе с неподтвержденным заболеванием возможны с вероятностью 2,5%. Все объекты, не имеющие заболевания острого аппендицита, отнесены к соответствующей 4-й группе. Безошибочность диагностики очень высокая 98,1%. Только для 1,9% объектов обучающей выборки поставлен ошибочный диагноз.

Решающие правила диагностики с такими высокими показателями эффективности следует рекомендовать для практического применения.

ПРИМЕР 7.1

Для диагностики трех видов аппендицита (1-гангренозный, 2-флегмонозный, 3-катаральный) и другой патологии живота отобрано 8 симптомов (табл.7.3).

По данным 103 историй болезни больных с тремя видами острого аппендицита и неподтвержденным диагнозом сформирована обучающая информация. В матрице обучающей информации (табл.7.4) содержатся значения в баллах 8 симптомов и девятый – группировочный признак, указывающий, к какой группе относится больной. Группа с гангренозным аппендицитом состоит из 28 больных, ее группировочный признак – 1. Больные с флегмонозным аппендицитом – 25 человек - объединены в группу с признаком 2. Группа больных с катаральным аппендицитом содержит 26 наблюдений, ее группировочный признак – 3. С неподтвержденным диагнозом острого аппендицита – 24 больных, группировочный признак – 4.

Таблица 7.3

Симптомы острого аппендицита, степени выраженности и их коды

Симптом	Наименование симптома	Степени выраженности симптомов и их коды
X1	Боли в правой подвздошной области	1 – незначительные 2 – выраженные
X2	Продолжительность болей в правой подвздошной области	1 - свыше двух суток 2 – 25-48 ч. 3 – 13-24 ч. 4 – до 12 ч.
X3	Частота пульса	1 – до 80 2 – 81-100 3 – свыше 100 уд/мин.
X4	Лейкоциты крови	1 – до 8 2 – 8-14 3 – свыше 14 тыс/мкл.
X5	Изменение языка	0 – не обложен 1 – обложен.
X6	Симптом Щеткина-Блюмберга	0 – отсутствует 2 – выражен.
X7	Симптом Ровзинга	0 – отсутствует 2 – выражен.
X8	Защитное мышечное напряжение	0 – отсутствует 2 – выражен.

Таблица 7.4

Массив обучающей информации

Группировочный признак	Симптомы							
	X1	X2	X3	X4	X5	X6	X7	X8
1	2	3	1	2	1	2	2	2
1	2	2	2	2	1	2	0	2
1	2	3	1	3	1	2	2	2
1	2	2	3	1	1	0	2	2
1	2	3	2	2	1	2	2	0
1	2	3	1	3	0	0	2	2
1	2	2	2	2	1	2	0	2
1	2	4	1	3	1	2	2	2
1	1	2	2	3	1	2	2	2
1	2	3	2	2	1	2	2	2
1	2	1	1	3	1	2	2	0
1	2	3	2	2	1	2	2	2
1	2	3	1	3	0	2	0	2
1	2	3	2	2	1	0	2	2
1	2	4	2	2	1	2	2	2
1	2	2	1	3	1	2	2	2
1	2	3	3	2	1	2	0	2
1	1	1	2	2	0	2	2	2
1	2	3	2	3	1	2	2	2
1	2	1	1	3	1	0	2	2
1	2	3	3	2	1	2	2	2
1	2	3	2	3	1	2	2	2
1	2	2	1	2	1	2	0	2
1	2	3	2	2	0	2	2	2
1	2	3	1	2	1	2	2	2
1	2	3	2	3	1	2	2	2
1	2	3	1	3	1	2	2	2
1	2	3	1	2	1	2	2	0
2	2	3	1	2	1	2	2	2
2	1	4	2	1	0	2	0	2
2	2	3	1	3	1	0	2	2
2	1	4	2	2	1	2	2	2
2	2	4	1	2	0	2	2	0
2	2	4	2	2	1	2	0	2

Группировочный признак	Симптомы							
	X1	X2	X3	X4	X5	X6	X7	X8
2	1	2	1	2	1	2	2	2
2	2	4	2	3	0	0	2	2
2	1	3	1	1	1	2	0	0
2	2	4	1	2	1	2	2	2
2	2	4	1	3	0	2	2	2
2	1	2	1	2	1	0	0	0
2	2	3	1	3	1	2	2	2
2	1	4	1	1	1	2	2	2
2	2	4	1	2	0	2	0	0
2	2	3	1	2	1	0	2	2
2	1	4	2	2	1	2	2	2
2	2	4	1	3	0	2	2	0
2	2	3	1	2	1	2	0	2
2	1	4	2	1	1	0	2	2
2	2	3	1	2	0	2	2	2
2	2	4	1	2	1	2	2	2
2	2	4	2	2	1	2	2	2
2	2	4	2	3	1	0	2	2
2	1	3	2	2	1	2	2	2
3	1	3	1	2	1	0	2	2
3	2	4	1	1	0	2	0	0
3	2	3	1	2	1	0	2	2
3	2	4	2	2	1	2	0	0
3	1	2	1	1	0	0	2	2
3	2	3	1	3	1	2	2	0
3	2	4	1	2	1	2	2	2
3	2	1	1	1	1	2	2	0
3	1	4	1	2	0	0	0	2
3	2	1	2	2	1	2	2	0
3	2	3	1	1	1	2	0	2
3	2	4	1	2	1	0	0	0
3	1	3	1	1	0	2	2	0
3	2	4	1	2	1	0	2	2
3	2	3	2	2	1	2	2	2
3	1	4	1	1	0	0	2	0
3	2	3	1	2	1	2	2	0
3	2	4	2	2	1	2	0	2

Группировочный признак	Симптомы							
	X1	X2	X3	X4	X5	X6	X7	X8
3	2	3	1	3	0	0	2	2
3	2	4	1	2	1	0	0	0
3	1	3	1	1	1	2	2	0
3	2	3	1	2	0	2	2	2
3	2	4	1	2	1	2	2	2
3	1	4	2	1	1	2	2	0
3	2	3	1	2	1	2	2	0
3	1	4	1	2	1	2	2	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	0	0	0	0
4	1	3	1	1	1	0	0	0
4	2	1	1	2	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	1	1	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	1	0	0	0
4	1	2	1	2	0	0	0	0
4	2	1	1	1	0	0	0	0
4	1	2	1	2	1	2	0	0
4	1	2	1	2	1	0	0	0
4	1	1	1	2	0	0	0	2
4	1	1	2	1	0	0	2	0
4	1	4	1	1	0	0	0	0
4	1	3	1	1	0	0	0	0
4	2	1	1	2	1	0	0	0
4	1	4	1	1	0	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	1	2	1	0	0	0
4	1	2	1	1	0	0	0	0
4	1	1	2	1	0	0	0	0
4	2	1	1	1	0	0	0	0
4	1	2	1	1	0	0	0	0

Требуется определить:

1. Информативность симптомов, включенных и не включенных в линейные дискриминантные функции при F-enter=2,00 , F-remove=1,90.
2. Коэффициенты линейных классификационных функций (ЛКФ).
3. Вклад ЛДФ в дисперсию симптомов.
4. Коэффициенты канонических ЛДФ.
5. Факторную структуру канонических ЛДФ.
6. Координаты центроидов четырех групп.
7. График положения объектов четырех групп.
8. Классификационную матрицу с оценками чувствительности диагностики групп обучающей информации.

Решение дано с помощью персонального компьютера с использованием ППП Statistica 5.0 for Windows.

1. Оценки информативности симптомов, включенных в ЛДФ, показаны в машинограмме 7.1, а не включенных в ЛКФ в связи с недостаточной информативностью – в машинограмме 7.2.

2. Коэффициенты линейных классификационных функций приведены в машинограмме 7.3.

3. Вклады канонических ЛДФ в дисперсию симптомов с оценками вклада по Chi-Square даны в машинограмме 7.4.

4. Коэффициенты канонических ЛДФ приведены в машинограмме 7.5.

5. Факторная структура канонических ЛДФ дана в машинограмме 7.6.

Машинограмма 7.1

Оценка информативности симптомов, включенных в ЛДФ

Discriminant Function Analysis Summary (diskr_a.sta)
Step 7, N of vars in model: 7; Grouping: GR (4 grps)
Wilks' Lambda: ,14315 approx. F (21,267)=12,333 p< 0000

	Wilks' Lambda	Partial Lambda	F-remove (3,93)	p-level	Toler.	1-Toler. (R-Sqr.)
X8	0,1722	0,8314	6,2865	0,0006	0,9149	0,0851
X6	0,1819	0,7871	8,3870	5,4E-05	0,9605	0,0395
X2	0,1998	0,7163	12,2765	7,8E-07	0,9871	0,0129
X7	0,1745	0,8202	6,7948	0,0003	0,9459	0,0541
X1	0,1590	0,9004	3,4287	0,0203	0,8461	0,1539
X3	0,1607	0,8910	3,7914	0,0129	0,9070	0,0930
X4	0,1586	0,9027	3,3424	0,0226	0,7667	0,2333

Машинограмма 7.2

Оценка информативности симптомов, не включенных в ЛДФ

Variables currently not in the model (diskr_a.sta)
Df for all F-tests: 3,92

	Wilks' Lambda	Partial Lambda	F to enter	p-level	Toler.	1-Toler. (R-Sqr.)
X5	0,1374	0,9599	1,2824	0,2851	0,9489	0,0511

Машинограмма 7.3

Коэффициенты линейных классификационных функций (ЛКФ)

Classification Functions; grouping: GR (diskr_an.sta)

	G_1:1 p=,27	G_2:2 p=,24	G_3:3 p=,25	G_4:4 p=,23
X8	11,3	10,6	8,3	5,8
X6	14,3	13,5	12,3	7,8
X2	3,6	4,9	4,7	2,5
X7	11,8	11,7	12,0	7,0
X1	9,8	8,3	9,4	6,3
X4	5,2	4,3	3,0	2,8
X3	7,8	6,2	5,5	5,3
Constant	-63,0	-57,4	-49,6	-23,0

Машинограмма 7.4

Оценка вкладов канонических ЛДФ в дисперсию признаков

Chi-Square Tests with Successive Roots Removed (diskr_a.sta)

	Eigen-value	Canonicl R	Wilks' Lambda	Chi-Sqr.	df	p-level
0	3,2251	0,8737	0,1431	187,5828	21	1,23E-28
1	0,5126	0,5822	0,6048	48,5227	12	2,56E-06
2	0,0931	0,2918	0,9149	8,5859	5	0,126801

Машинограмма 7.5

Коэффициенты канонических ЛДФ

Raw Coefficients (diskr_a.sta)
for Canonical Variables

	Root 1	Root 2	Root 3
X8	-0,5922	0,2400	-0,6529
X6	-0,6940	0,0816	0,0225
X2	-0,4192	-0,9237	-0,3337
X7	-0,5674	-0,3626	0,5819
X1	-0,7109	0,0741	1,9359
X3	-0,3767	1,0349	-0,0002
X4	-0,3992	0,8067	-0,7622
Constant	5,7765	-0,3913	-0,7402
Eigenval	3,2251	0,5126	0,0931
Cum.Prop	0,8419	0,9757	1,0000

Машинограмма 7.6

Факторная структура канонических ЛДФ

Factor Structure Matrix (diskr_a.sta)
Correlations Variables - Canonical Roots
(Pooled-within-groups correlations)

	Root 1	Root 2	Root 3
X8	-0,4807	0,2371	-0,4999
X6	-0,4577	-0,0032	0,1303
X2	-0,3475	-0,7208	-0,3241
X7	-0,4510	-0,1362	0,3081
X1	-0,3627	0,1596	0,5903
X3	-0,1638	0,4197	-0,0599
X4	-0,3816	0,3320	-0,1349

6. Координаты центроидов для четырех групп – машинограмма 7.7.

Машинограмма 7.7

Координаты центроидов

Means of Canonical Variables (diskr_a.sta)

	Root 1	Root 2	Root 3
G_1:1	-1,4692	0,9503	0,1156
G_2:2	-1,0723	-0,4792	-0,4519
G_3:3	-0,2358	-0,8265	0,3732
G_4:4	3,0866	0,2859	-0,0685

7. График положения полей объектов четырех групп показан на рисунке 7.2.

8. Оценка чувствительности диагностики по решающим правилам объектов обучающей информации в классификационной матрице – машинограмма 7.8.

Машинограмма 7.8

Оценка чувствительности решающих правил

Classification Matrix (diskr_a.sta)
Rows: Observed classifications
Columns: Predicted classifications

	Percent Correct	G_1:1 p=,27184	G_2:2 p=,24272	G_3:3 p=,25243	G_4:4 p=,23301
G_1:1	78,57	22	4	2	0
G_2:2	60,00	4	15	4	2
G_3:3	65,38	2	7	17	0
G_4:4	100,00	0	0	0	24
Total	75,73	28	26	23	26

Выводы:

1. Из данных машинограммы 7.1 видно, что наиболее информативными симптомами (с уровнями значимости $p<0,01$) являются X8, X6, X2, X7, X1, X3 и X4.

2. Линейные классификационные функции (ЛКФ) (машинограмма 7.3) рассчитываются по формулам:

$$ЛДФ1=-63,0+9,8x_1+3,6x_2+7,8x_3+5,2x_4+14,3x_6+11,8x_7+11,3x_8;$$

$$ЛДФ2=-57,4+8,3x_1+4,9x_2+6,2x_3+4,3x_4+13,5x_6+11,7x_7+10,6x_8;$$

$$ЛДФ3=-49,6+9,4x_1+4,7x_2+5,5x_3+3,0x_4+12,3x_6+12,0x_7+8,3x_8;$$

$$ЛДФ4=-23,0+6,3x_1+2,5x_2+5,3x_3+2,8x_4+7,8x_6+7,0x_7+5,8x_8.$$

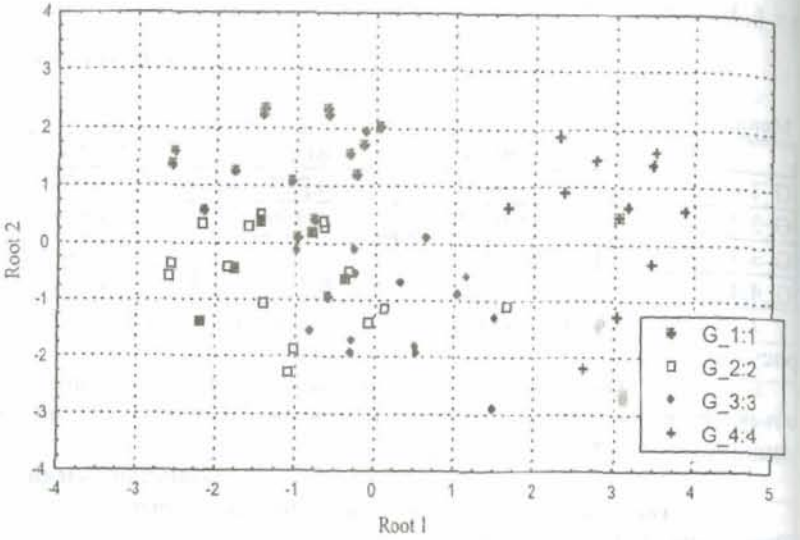

Рис.7.2. Положение объектов четырех групп в координатах первой и второй канонических ЛДФ.

3. Для решения задачи медицинской диагностики следует применить первые две канонические ЛДФ (с уровнем значимости $p<0,001$) с суммарным вкладом в дисперсию симптомов 97,6% (машинограммы 7.4 и 7.5).

4. Канонические ЛДФ F1 и F2 (машинограмма 7.5) рассчитываются по формулам:

$$F1=5,8-0,71x_1-0,42x_2-0,38x_3-0,4x_4-0,69x_6-0,57x_7-0,59x_8;$$

$$F2=-0,39+0,07x_1-0,92x_2+1,03x_3+0,81x_4+0,08x_6-0,36x_7+0,24x_8.$$

5. Решение диагностической задачи выполняется по графику на рис.7.2, на котором нанесены центроиды четырех диагностируемых групп (машинограмма 7.7). Больного, для которого по его симптомам определены F1 и F2 (Root1 и Root2), следует отнести к группе по *минимальному расстоянию от соответствующего центроида.*

6. Из данных машинограммы 7.8 видно, что точность диагностики по решающим правилам в среднем характеризуется достоверностью 75,7%, для первой группы – 78,6%, второй – 60,0%, третьей – 65,4% и четвертой – 100%. Недостаточная точность диагностики для второй и

третьей групп объясняется значительным перекрытием симптомов для этих групп острого аппендицита.

Литература

1. Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р, Олдендерфер М.С, Блэшфилд Р.К. Факторный, дискриминантный и кластерный анализ: Пер. с англ. Под ред. И.С. Енюкова.- М.: Финансы и статистика, 1989.

2. Математико-статистические методы в клинической практике / Под ред.В.И.Кувакина. - СПб.: Б.и, 1993. - 199 с.

Глава 8. АНАЛИЗ СООТВЕТСТВИЙ

Назначение и содержание анализа соответствий

Анализ соответствий содержит описательные и разведочные методы изучения структуры категорированных данных, представляемых в виде матрицы наблюдений или частотной таблицы сопряженности. В матрице наблюдений число строк n равно количеству наблюдавшихся объектов, число столбцов k равно количеству признаков, включающему как группировочные, так и качественные признаки.

В частотной таблице сопряженности число строк m равно количеству сочетаний уровней (категорий) признаков, а число столбцов (k+1), где (k+1) – количество группировочных и качественных признаков и столбец частоты наблюдений на различных сочетаниях уровней (категорий) признаков.

Основные задачи анализа.

1.*Снижение многомерности исходной информации о множестве* исследуемых признаков путем выделения 1-3 ортогональных (независимых) базисных векторов, адекватно отражающих представляемую информацию.

2.Оценка связи группировочных и качественных признаков по величине инерции (мере связи, называемой в статистике показателем *сопряженности) и оценка значимости этой связи по критерию* χ^2, *связанному с коэффициентом сопряженности* ($\varphi^2 = \chi^2/n$). Поэтому наряду с показателем инерции φ^2, критерий χ^2 также характеризует силу и устойчивость связи признаков.

3.Определение относительных расстояний между группами объектов (категориями признаков), по которым оценивают соответствие групп объектов (категорий признаков) и правильность группировки и *шкалирования признаков.*

4.Оптимизация группировки изучаемых признаков.

5.Наглядное представление результатов анализа на 1-3-х мерных диаграммах расстояний между группами объектов (категориями признаков) в координатах базисных векторов.

6.Оценка качества анализа по специфическим для этого метода *критериям: масса, инерция, качество, квадрат косинуса и др.*

Применение метода анализа соответствий и обсуждение результатов решения по модулю Correspondence Analysis ППП Statistica 5.0 for Windows рассмотрим на примерах.

ПРИМЕР 8.1. Исследование связи между должностными группами сотрудников учреждения и категориями их пристрастия к курению

В результате проведенного исследования 193-х сотрудников (Greenacre, 1984) сформирована двухвходовая таблица, содержащая частоты случаев четырех категорий пристрастия к курению пяти групп сотрудников (табл.8.1).

Таблица 8.1

Сводные данные обследования сотрудников фирмы об их отношении к курению

Группы сотрудников	Некурящие	Категории курящих			Всего по строке
		слабо	средне	сильно	
Старшие менеджеры	4	2	3	2	11
Младшие менеджеры	4	3	7	4	18
Старшие сотрудники	25	10	12	4	51
Младшие сотрудники	18	24	33	13	88
Секретари	10	6	7	2	25
Всего по столбцу	61	45	62	25	193

Для решения на ПК с помощью ППП Statistica 5.0 по модулю анализа соответствий данные таблицы 8.1 возможно представлять в виде частотной таблицы сопряженности – матрицы размером 20×3 (табл.8.2). Число строк матрицы равно числу возможных сочетаний уровней двух переменных: пяти групп сотрудников и четырех категорий курения (5×4=20). Три столбца матрицы – это столбцы кодов групп и категорий и столбец числа наблюдений.

Таблица 8.2

Частотная таблица сопряженности

№ пп	Группы сотрудников	Категории курящих	Число наблюдений
1	1	1	4
2	1	2	2
3	1	3	3
4	1	4	2
5	2	1	4
6	2	2	3
7	2	3	7
8	2	4	4
9	3	1	25
10	3	2	10
11	3	3	12
12	3	4	4
13	4	1	18
14	4	2	24
15	4	3	33
16	4	4	13
17	5	1	10
18	5	2	6
19	5	3	7
20	5	4	2

Алгоритм решения предусматривает представление абсолютных частот наблюдений относительными частотами, так чтобы сумма всех относительных частот была равна 1 (каждый элемент абсолютной частоты делится на общее число наблюдений - 193).

Полагая относительные частоты четырех категорий пристрастия к курению для каждой группы как координаты в четырехмерном пространстве, рассчитывают координаты пяти точек групп, евклидовы расстояния между ними и матрицу расстояний. Далее исследуется структура матрицы расстояний. В результате этого исследования выделяются три ортогональных (независимых), вектора, так что первый базисный вектор вносит наибольший вклад в объяснение связи между

группами сотрудников и категориями пристрастия к курению; второй – значительно меньший вклад; третий – минимальный вклад. В таблице 8.3 даны показатели, характеризующие величину меры связи (общая инерция, критерий χ^2 и его уровень значимости), а также базисные векторы их сингулярные и собственные значения, процент инерции и накопленный (кумулятивный) процент инерции, оценка вклада базисных векторов по критерию χ^2.

Для оценки соответствия между группами сотрудников по пристрастию к курению выполняется расчет положения групп в двумерном пространстве координат двух первых базисных векторов и построение соответствующей диаграммы (табл.8.4 и рис.8.1).

Таблица 8.3

Собственные значения и инерция для базисных векторов

Таблица ввода (Строки x Столбцы): 5 x 4 Общая инерция =0,085 χ^2=16,44 df=12, p=0,17					
Базисные векторы	Сингул. значения	Собств. значения	Процент инерции	Кумул. процент	χ-квадрат
1	0,273	0,075	87,756	87,756	14,428
2	0,100	0,010	11,759	99,514	1,933
3	0,020	0,000	0,855	100,000	0,080

Качество решения оценивается по специфическим критериям масса, качество, относительная инерция, квадрат косинуса (табл.8.4).

Важная информация для оценки сходства групп сотрудников представлена в таблице распределения групп сотрудников по их отношению к курению (табл.8.5). Частоты в этой таблице стандартизованы так, что их сумма по каждой строке равна 100%.

Таблица 8.5

Распределение групп сотрудников по их отношению к курению, в %

Группа сотрудников	Неку-рящие	Категории курящих			Всего по строке
		слабо	средне	сильно	
Старшие менеджеры	36,36	18,18	27,27	18,18	100,00
Младшие менеджеры	22,22	16,67	38,89	22,22	100,00
Старшие сотрудники	49,02	19,61	23,53	7,84	100,00
Младшие сотрудники	20,45	27,27	37,50	14,77	100,00
Секретари	40,00	24,00	28,00	8,00	100,00

Аналогично проведенному анализу соответствия пяти групп сотрудников можно провести исследование соответствия четырех категорий пристрастия к курению пятью группами сотрудников, которое включает определение координат четырех категорий курения в двумерном пространстве (табл.8.6) и построение соответствующей диаграммы (рис.8.2). С их помощью дается оценка соответствия категорий курения и правильности шкалирования этого признака.

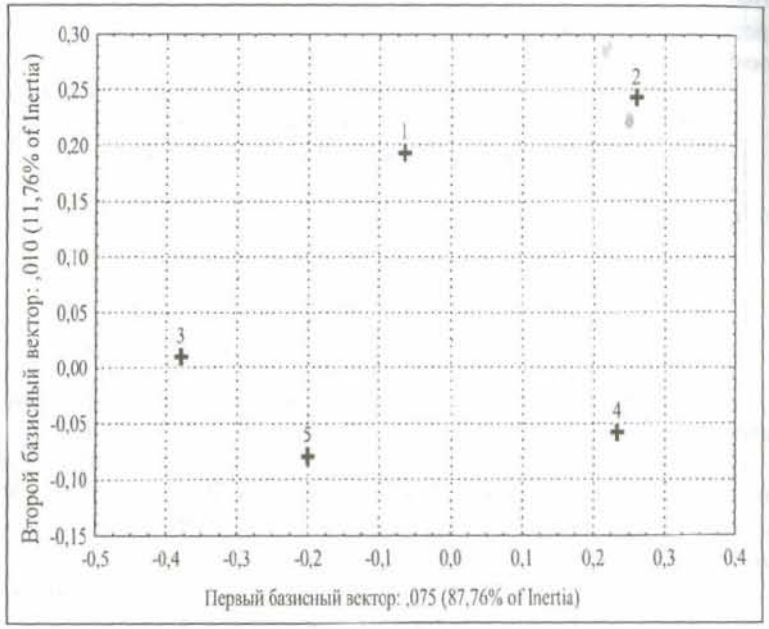

Рис.8.1.Положение групп сотрудников в координатах первых двух базисных векторов.

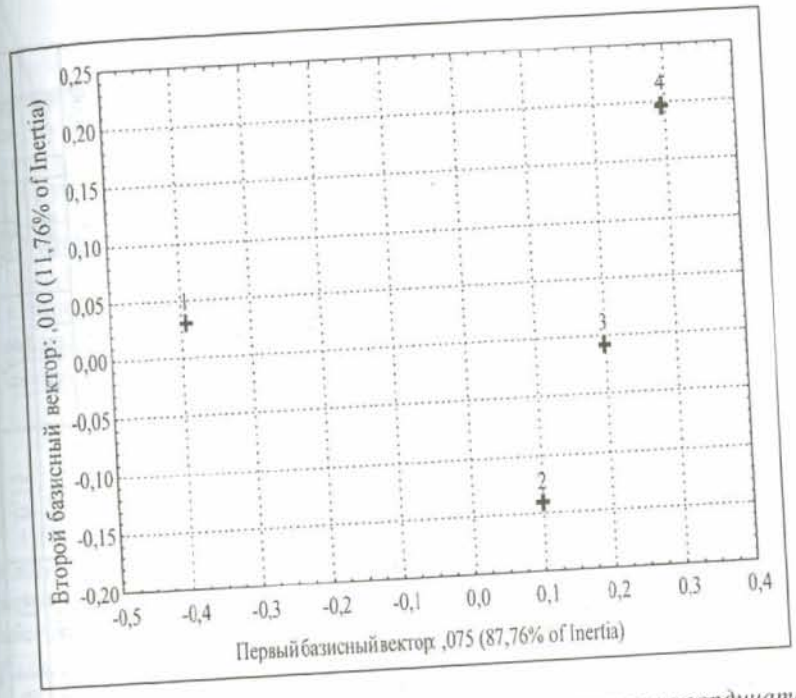

Рис.8.2.Положение категорий пристрастия к курению в координатах первых двух базисных векторов.

Таблица 8.4

Координаты двух первых базисных векторов для групп сотрудников

Группа сотрудников	Координаты по 1 вект.	Координаты по 2 вект.	Масса	Качество	Относ. инерция	Инерция по 1 вект.	Квадрат косинуса по 1 вект.	Инерция по 2 вект.	Квадрат косинуса по 2 вект.
Ст. менеджеры	-0,07	0,19	0,06	0,89	0,03	0,00	0,09	0,21	0,80
Мл. менеджеры	0,26	0,24	0,09	0,99	0,14	0,08	0,53	0,55	0,46
Ст. сотрудники	-0,38	0,01	0,26	1,00	0,45	0,51	1,00	0,00	0,00
Мл. сотрудники	0,23	-0,06	0,46	1,00	0,31	0,33	0,94	0,15	0,06
Секретари	-0,20	-0,08	0,13	1,00	0,07	0,07	0,87	0,08	0,13

Таблица 8.6

Координаты двух первых базисных векторов для категорий пристрастия к курению

Категория курящих	Координаты по 1 вект.	Координаты по 2 вект.	Масса	Качество	Относ. инерция	Инерция по 1 вектору	Квадрат косинуса по 1 вект.	Инерция по 2 вектору	Квадрат косинуса по 2 вект.
Некурящие	-0,39	0,03	0,32	1,00	0,56	0,65	0,99	0,03	0,01
Слабо курящие	0,10	-0,14	0,23	0,98	0,08	0,03	0,33	0,46	0,66
Средне курящие	0,20	-0,01	0,32	0,98	0,15	0,17	0,98	0,00	0,00
Сильно курящие	0,30	0,20	0,13	0,99	0,19	0,15	0,68	0,51	0,31

Таблица 8.7

Распределение некурящих и различных категорий курящих среди групп сотрудников, в %

Группа сотрудников	Некурящие	Категории курящих		
		слабо	средне	сильно
Старшие менеджеры	6,56	4,44	4,84	8,00
Младшие менеджеры	6,56	6,67	11,29	16,00
Старшие сотрудники	40,98	22,22	19,35	16,00
Младшие сотрудники	29,51	53,33	53,23	52,00
Секретари	16,39	13,33	11,29	8,00
ВСЕГО	100,00	100,00	100,00	100,00

Анализ результатов решения примера

1.По данным проведенного исследования установлена слабая связь между пятью группами сотрудников и четырьмя категориями пристрастия к курению. Она оценена (табл.8.3) общей инерцией 0,085, что соответствует коэффициенту связи (коэффициент сопряженности) $\varphi = \sqrt{0,085} = 0,292$, не достигшему порога умеренной связи 0,30. Значимость этой связи недостаточна, т.к. ее уровень по критерию χ^2=16,44 с числом степеней свободы $f = (r-1)(c-1) = (5-1)(4-1) = 12$ составил p=0,172, что соответствует достоверности связи 1-p=1-0,172=0,828, ниже критического значения 0,95.

2.Многомерная исходная информация, представленная в частотной таблице сопряженности размером 20×3 охарактеризована тремя базисными ортогональными векторами, объясняющими четыре категории интенсивности курения пятью группами сотрудников на 87,76% - первым, 11,76% - вторым и 0,48% третьим векторами, а в сумме на 100,0% (табл.8.3). Полагая вклад третьего вектора незначительным, дальнейший анализ следует провести по двум первым векторам с вкладом 99,51%; результаты анализа соответствия групп сотрудников и категорий пристрастия к курению наглядно представить на двумерных диаграммах в координатах двух первых базисных векторов.

3.Соответствие групп сотрудников по пристрастию к курению определяется по диаграмме (рис.8.1), на которой по координатам двух векторов (табл.8.4) нанесены точки пяти групп. Степень соответствия оценивается по расстоянию между точками групп по оси первого век-

тора, имеющего вклад в объяснение соответствия примерно в 7 раз больше, чем вклад второго вектора (табл.8.3). Наименьшее расстояние, а значит наибольшее соответствие по пристрастию к курению, отмечается между группами: 2 – младшие менеджеры и 4 – младшие сотрудники; 1 – старшие менеджеры и 5 секретари. Наибольшее расстояние, а значит минимальное соответствие по пристрастию к курению, отмечается между группами: 3 – старшие сотрудники и группами 2 и 4 – соответственно младшие менеджеры и младшие сотрудники. Высокая степень соответствия между старшими менеджерами и секретарями объясняется наличием в этих группах сотрудников большей доли некурящих по сравнению с другими группами. Высокая степень соответствия между младшими менеджерами и младшими сотрудниками объясняется большей долей в этих группах категории средней степени пристрастия к курению.

Справедливость оценок соответствия групп подтверждают данные об относительных частотах наблюдений пристрастия к курению в пяти группах сотрудников (табл.8.7).

На основании проведенного анализа первый, наиболее весомый базисный вектор следует интерпретировать как интенсивность курения сотрудниками различных должностных групп.

Минимальную степень пристрастия к курению имеют старшие сотрудники (группа 3); максимальную – младшие менеджеры (группы 2) и младшие сотрудники (группа 4).

4.*Соответствие категорий пристрастия к курению и обоснованность* их шкалирования определяется по диаграмме (рис.8.2), на который по координатам двух первых векторов (табл.8.6) нанесены точки четырех категорий пристрастия к курению. По направлению первого более весомого вектора, наибольшие расстояния, а значит отличия, отмечаются между категорией 1 – некурящих и остальными категориями 2-4 курящих. Отличие между категориями 2-слабо, 3-средне и 4-сильно курящих выражено менее значимо. На диаграмме четко прослеживается направленность категорий пристрастия к курению: от категории 1-*некурящих, к категориям 2- слабо, 3-средне и 4-сильно курящих*. Это подтверждает правильность шкалирования категорий пристрастия к курению.

5.Обоснованность анализа в п.п. 1-4 доказывается с помощью дополнительных критериев, значения которых даны в табл.8.4 для должностных групп сотрудников и в табл.8.6 для категорий пристрастия к

курению. К ним относятся критерии: масса, качество, относительная инерция и квадрат косинуса.

Ниже дается интерпретация этих критериев для оценки вклада в базисные вектора должностных групп сотрудников по табл.8.4.

Критерий массы показывает относительные частоты наблюдений в группах. Наибольшее число наблюдений имеется в группе 4 (0,46 или 46,0% от общего числа наблюдений) и группе 3 (0,26 или 26,0%); наименьшее число наблюдений в группе 1 (0,06 или 6,0%) и 2 (0,09 или 9,0%).

Критерий качества характеризует вклад групп в формирование базисных векторов. Наибольший вклад вносят 2, 3 и 4 группы с большим числом наблюдений; наименьший вклад имеет самая малочисленная группа 1. Для достижения более высокого уровня качества в группах необходимо иметь достаточное число наблюдений.

Относительная инерция (мера связи) представляет долю общей инерции приходящуюся на каждую из пяти групп сотрудников. Наибольший вклад в оценку связей между группами сотрудников и категориями пристрастия к курению вносят большие по численности группы 3 и 4. Доля их вклада соответственно равна 0,45 и 0,31. Наименьший вклад дали малочисленные группы 1 и 5 (0,03; 0,07). В таблице даны относительные инерции групп в два базисных вектора.

Критерий ***квадрант косинуса*** дополняет оценку качества каждой группы сотрудников с соответствующим базисным вектором. Эта величина интерпретируется как квадрат косинуса угла между направлениями из центра масс в начале координат на точку группы и осью вектора. Эта величина показывает вклад (нагрузку) данной группы в базисный вектор.

Выводы:

1.Применение модуля анализа соответствий позволило перейти от многомерных данных о четырех категориях пристрастия к курению и пяти групп сотрудников к двум независимым базисным векторам, отражающим исходную информацию на 99,51%.

2.По двумерным диаграммам расстояний в координатах двух первых базисных векторов дана оценка соответствий между группами сотрудников и категориями пристрастия к курению и оценка правильности шкалирования категорий пристрастия к курению.

3.Установлено, что имеет место слабая связь между должностными группами сотрудников и категориями пристрастия к курению (коэффи-

циент связи 0,292). Эта связь недостаточно достоверна (достоверность 82,8%) вследствие малого числа наблюдений в первой и пятой группах сотрудников.

4.Проведенный описательный и разведочный анализ соответствий показал, что по исходным данным невозможно получать достоверные модели для прогноза, связывающие категории пристрастия к курению с должностными группами сотрудников. Для получения эффективных моделей необходимо учитывать не только фактор должностных категорий сотрудников, но и такие факторы, как пол, возраст, состояние здоровья сотрудников, их привязанность к здоровому образу жизни, физической культуре, туризму и т.п.

ПРИМЕР 8.2. Исследование связи между систолическим артериальным давлением у пострадавших с тяжелой черепно-мозговой травмой при поступлении в стационар и показателем жизненной активности при их убытии

Группировка наблюдений - один из основных методов статистического исследования, заключающийся в расчленении совокупности наблюдений на группы по определенным существенным признакам. Основными вопросами метода группировок являются выбор группировочного признака и определение числа групп. Правильный выбор группировочных признаков возможен лишь на основе анализа сущности явлений, базирующегося на особенностях развития явления в конкретных условиях места и времени. Учет конкретных условий приводит к тому, что один и тот же тип явления может быть выявлен в одних условиях по одному признаку, а в других - по другому. Число групп, на которые расчленяется изучаемая совокупность наблюдений, зависит от типа изучаемого явления, от характера вариации группировочного признака и задач исследования. С помощью метода группировок решаются многие задачи, которые можно свести к изучению типов исследуемого явления; структуры совокупности и ее изменений; взаимосвязи между признаками.

Под группировкой понимается процесс систематизации совокупности наблюдений, образования групп, однородных в каком-либо существенном отношении, а также имеющих одинаковые или близкие значения группировочного признака. Группировка - это не просто технический прием, позволяющий представить первичные данные в упоря-

виде, но и глубоко осмысленное действие, направленное на выявление связей между признаками изучаемого явления. От того, как группируется исходный материал, во многих случаях зависят выводы о природе изучаемого явления. Один и тот же материал дает диаметрально противоположные выводы при разных приемах группировки. Нельзя сводить в одну и ту же группу неоднородные по составу данные. Группировка должна соответствовать цели конкретного исследования, поставленным задачам и содержанию изучаемого явления

Для осуществления группировки устанавливают признак, по которому единицы наблюдаемой совокупности распределяют по группам, число групп и их обозначение (границы). Каждая единица совокупности в зависимости от значения у нее группировочного признака относится к соответствующей группе. Затем производится подсчет числа единиц в каждой группе, а также общих (итоговых) частот признаков, которыми они характеризуются (например, группировка больных по полу, по видам медицинских процедур, оперативных вмешательств, как по технике исполнения, так и по методу доступа и др.).

Под признаком в медицинской статистике понимается отличительная черта, свойство, качество, присущее единице совокупности. Например, пол, возраст больного, диагноз заболевания, метод лечения и т.п. Признаки могут быть качественными (категорированными) или количественными.

Вся совокупность признаков, описывающих единицу наблюдения, делится на признаки-причины (факторы) и результативные признаки (признаки-отклики).

С целью оптимизации группировки, наряду с привлечением к решению этой задачи опытных и заинтересованных экспертов, целесообразно использовать математико-статистический метод анализа соответствий.

Н.Б. Клименко в интересах создания модели прогноза раннего исхода тяжелой черепно-мозговой травмы по показателю жизненной активности (ПЖА) при убытии пострадавшего из стационара, исследовала ряд признаков, характеризующих состояние пострадавшего во время его поступления в приемное отделение Российского научно-исследовательского нейрохирургического института им. профессора А.Л. Поленова. В число предиктивных признаков, наряду с другими, было включено систолическое артериальное давление. Исходная обучающая матрица содержала данные о 300 пострадавших. Задачу груп-

пировки этих двух признаков решали: эксперт предметной области врач-невропатолог и специалист по статистической обработке данных медицинских исследований. Первичная группировка была произведена следующим, вполне логичным, образом.

ПЖА при последнем осмотре в стационаре: 1 - полностью работоспособен, 2 - ограниченно работоспособен, 3 - ходит по улице, 4 - ходит по отделению, 5 - ходит по палате, 6 - обслуживает себя сидя, нуждается в надзоре, 7 - обслуживает себя лежа, 8 - глотает, поворачивается, катетер, клизма, 9 - ИВЛ, не глотает, не двигается, катетер, клизма, 10 – умер.

Систолическое АД (мм рт.ст.): 1 - критическая гипотония (ниже 60), 2 - выраженная гипотония (60-80), 3 - умеренная гипотония (90-109), 4 - нормотония (110-140), 5 - умеренная гипертония (141-180), 6 - выраженная гипертония (181-220), 7 - грубая гипертония (свыше 220).

Частоты наблюдений представлены в двухвходовой таблице (табл.8.8). С первого взгляда обращают на себя внимание следующие особенности:

– группировка ПЖА слишком детализирована, имеются определенные трудности в отнесении пострадавшего к конкретной группе (например, 2 - ограниченно работоспособен, 3 - ходит по улице, 4 - ходит по отделению, 5 - ходит по палате);

– АД сгруппировано по принципу "от минимального до максимального значения" без учета его связи с тяжестью состояния пострадавшего и с результирующим признаком - ПЖА.

В результате математико-статистического анализа, вопреки ожиданиям, выявлена крайне слабая, отрицательная (ранговый коэффициент корреляции r=-0,01) и статистически незначимая (уровень значимости p=0,799, достоверность 1-p=0,201) связь анализируемых признаков.

Таблица 8.8

Двухвходовая таблица данных при исходной группировке

ПЖА	Систолическое АД							Всего
	критическая гипотония	выраженная гипотония	умеренная гипотония	нормотония	умеренная гипертония	выраженная гипертония	грубая гипертония	
Полностью работоспособен	0	0	1	7	2	0	0	10
Ограниченно работоспособен	0	0	4	19	9	0	0	32
Ходит по улице	0	0	3	34	7	0	0	44
Ходит по отделению	0	1	5	22	8	3	0	39
Ходит по палате	0	1	1	2	4	1	0	9
Обслуживает себя сидя	0	0	1	7	4	0	0	12
Обслуживает себя лежа	0	0	2	3	1	2	0	8
Глотает, поворачивается, катетер, клизма	0	1	0	1	2	0	0	4
ИВЛ, не глотает, не двигается, катетер, клизма	0	0	0	3	0	0	0	3
Умер	3	14	25	46	35	14	2	139
Всего	3	17	42	144	72	20	2	300

Используя метод анализа соответствия мы получили следующие данные (табл.8.9 - 8.11, рис.8.3, 8.4):

– два первых базисных вектора (табл.8.9) объясняют общую инерцию соответствия анализируемых признаков на 90,4%, при этом инерция в целом оказалась недостоверной (p=0,12), что свидетельствует о слабой связи изучаемых признаков;

– распределение групп пострадавших по признаку ПЖА в координатах двух первых базисных векторов (табл.8.10, рис.8.3) демонстрирует картину, неподдающуюся интерпретации. Группировку по признаку ПЖА необходимо упростить, объединяя группы близкие по расстоянию между ними;

– распределение групп по признаку систолического АД (табл.8.11 и рис.8.4) демонстрирует возможность оптимизации группировки и снижения ее размерности путем объединения групп пострадавших с диаметрально противоположными значениями АД. Группировка может быть следующей: группа 1 - нормотония (значение первого базисного вектора -0,37), группа 2 - умеренная гипотония и умеренная гипертония (значения первого базисного вектора в интервале от 0,14 до 0,29), группа 3 - критическая гипотония, выраженная гипотония, выраженная гипертония и грубая гипертония (значения первого базисного вектора в интервале от 0,67 до 0,88).

Таблица 8.9

Собственные значения и инерция для всех базисных векторов

Таблица ввода (Строки x Столбцы): 10×7

Общая инерция =,22102 χ^2=66,307 df=54 p=,12171

Базисные векторы	Сингул. значения	Собств. значения	Процент инерции	Кумулятивный процент	χ-квадрат
1	0,41	0,16	74,56	74,56	49,44
2	0,19	0,04	15,85	90,41	10,51
3	0,12	0,01	6,23	96,63	4,13
4	0,08	0,01	3,15	99,78	2,09
5	0,02	0,00	0,22	100,00	0,15
6	0,00	0,00	0,00	100,00	0,00

Таблица 8.10

Координаты двух первых базисных векторов для групп пострадавших по ПЖА

Группа ПЖА	Координаты по 1 вектору	Координаты По 2 вектору	Масса	Качество
1	-0,51	0,02	0,03	0,99
2	-0,36	0,11	0,11	0,82
3	-0,61	-0,01	0,15	0,97
4	-0,18	-0,12	0,13	0,82
5	0,43	0,30	0,03	0,66
6	-0,37	0,21	0,04	0,80
7	0,29	-0,75	0,03	0,86
8	0,45	0,96	0,01	0,88
9	-0,92	-0,17	0,01	0,81
10	0,36	-0,01	0,46	0,97

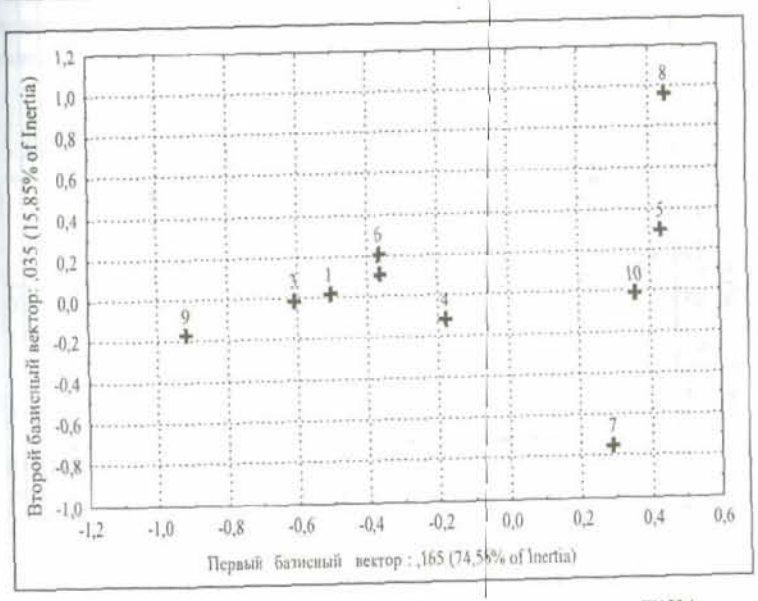

Рис.8.3.Распределение пострадавших по группам ПЖА.

Таблица 8.11

Координаты двух первых базисных векторов для групп пострадавших по систолическому АД

Систолическое АД	Координаты по 1 вектору	Координаты по 2 вектору	Масса	Качество
1	0,88	-0,05	0,01	0,67
2	0,83	0,32	0,06	0,91
3	0,29	-0,17	0,14	0,88
4	-0,37	-0,03	0,48	0,98
5	0,14	0,22	0,24	0,77
6	0,67	-0,45	0,07	0,95
7	0,88	-0,05	0,01	0,67

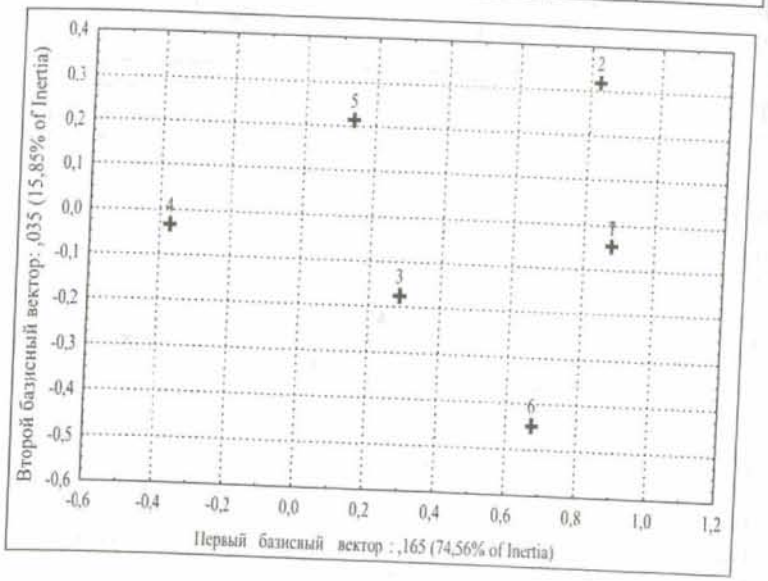

Рис.8.4.Распределение пострадавших по группам систолического АД.

Эксперт предметной области провел дополнительный логический анализ, увязал систолическое АД с морфологическими и патофизиологическими особенностями повреждения мозга и, в конечном счете, с исходами ТЧМТ. В итоге группировка изучаемых признаков приобрела вид:

ПЖА: 1– ПЖА удовлетворительный (вошли группы: 1, 2, 3, 4 первичной группировки), 2 – ПЖА низкий (5-9), 3 – пострадавший умер.

Систолическое АД: 1 – нормотония, 2 – умеренная гипертония, 3 – умеренная гипотония, 4 – выраженная гипертония, 5 – выраженная гипотония.

Частоты наблюдений при обновленной группировке признаков представлены в двухвходовой таблице (табл.8.12). Корреляционный анализ показал, что ранговый коэффициент корреляции увеличился до 0,35 с достоверностью более 99,9% (p=0,000), что свидетельствует об умеренной статистически значимой связи между изучаемыми признаками. Результаты анализа соответствия представлены в таблицах 8.13 - 8.15 и на рисунках 8.5, 8.6.

Из таблицы 8.13 следует, что вторичная группировка оказалась вполне удачной. Инерция (мера связи признаков) значимая (p=0,000) и объясняется двумя первыми базисными векторами на 100%.

Таблица 8.12

Двухвходовая таблица после повторной группировки данных

ПЖА	Систолическое АД					Всего
	нормотония	умеренная гипертония	умеренная гипотония	выраженная гипертония	выраженная гипотония	
Удовлетворительный	82	26	13	3	1	125
Низкий	16	11	4	3	2	36
Пострадавший умер	46	35	25	16	17	139
Всего	144	72	42	22	20	300

Таблица 8.13

Собственные значения и инерция для всех базисных векторов

Таблица ввода (Строки x Столбцы): 3×5 Общая инерция =,13094 χ^2=39,283 df=8 p=,00000					
Базисные векторы	Сингул. значения	Собств. значения	Процент инерции	Кумулятивный процент	χ-квадрат
1	0,36	0,13	96,94	96,94	38,08
2	0,06	0,00	3,06	100,00	1,20

Удалось создать вполне приемлемую группировку пострадавших по ПЖА, о чем свидетельствуют приблизительно равноудаленные друг от друга значения, координат групп по оси первого базисного вектора (табл.8.14, рис.8.5). Группы строго ранжированы по вариантам исхода и приблизительно равно удалены друг от друга: группа с удовлетворительными ПЖА имеет значение первого базисного вектора равное -0,40, с низкими ПЖА - +0,02, умершие - +0,36.

Таблица 8.14

Координаты двух первых базисных векторов для групп пострадавших по ПЖА после вторичной группировки

Группа ПЖА	Координаты по 1 вектору	Координаты по 2 вектору	Масса	Качество
1	-0,40	0,02	0,42	1,00
2	0,02	-0,17	0,12	1,00
3	0,36	0,02	0,46	1,00

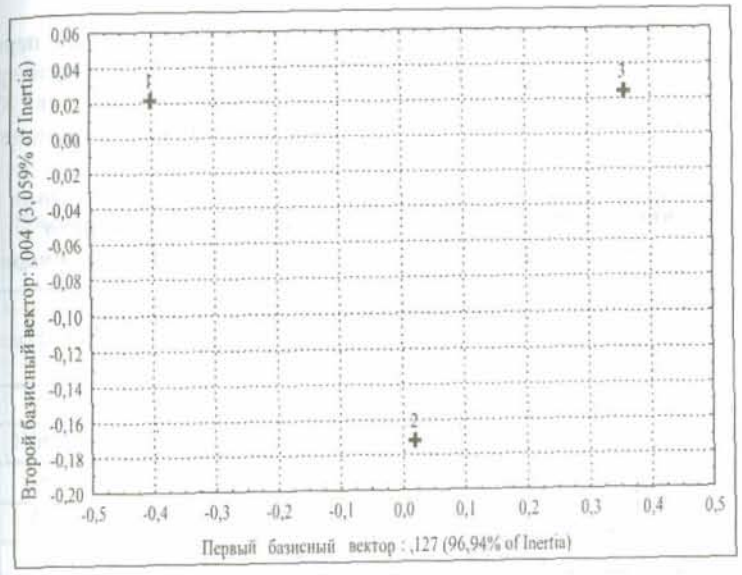

Рис.8.5.Распределение пострадавших по группам ПЖА после вторичной группировки.

Адекватно выглядит группировка пострадавших по показателю систолического АД (табл.8.15). Первый базисный вектор принял значение от -0,32 по возрастающей до 0,8 с интервалами между координатами от 0,16 до 0,41, что свидетельствует о равномерной градации признака в его привязке к показателю ПЖА. Наглядно вывод подтверждается диаграммой (рис.8.6).

Исходя из данных, представленных на рис.8.4 возможен вариант снижения размерности группировки показателя систолического АД до трех групп: 1 группа - нормотония, 2 группа - умеренная гипотония и умеренная гипертония, 3 группа - критическая гипотония, выраженная гипотония, выраженная гипертония и грубая гипертония. Исходные данные частот наблюдений приведены в двухвходовой таблице (табл.8.16). Исследуя этот вариант группировки с помощью метода анализа соответствия получены результаты, представленные в таблицах 8.17, 8.18 и на диаграмме (рис.8.7). Предложенная группировка АД обеспечивает высокую степень соответствия группировки этого признака с группировкой ПЖА о чем свидетельствует высокая значимость

общей инерции (p<0,0000) и высокая доля (99,55%) вклада первого базисного вектора. Координаты первого базисного вектора для групп АД распределились почти симметрично и приняли значение: -0,32, 0,15 и 0,69. Очень убедительно верность группировки продемонстрирована на рис.8.7.

Таблица 8.15

Координаты двух первых базисных векторов для групп пострадавших по систолическому АД после вторичной группировки

Систолическое АД	Координаты по 1 вектору	Координаты по 2 вектору	Масса	Качество
1	-0,32	0,02	0,48	1,00
2	0,09	-0,10	0,24	1,00
3	0,25	0,08	0,14	1,00
4	0,58	-0,04	0,07	1,00
5	0,80	0,08	0,07	1,00

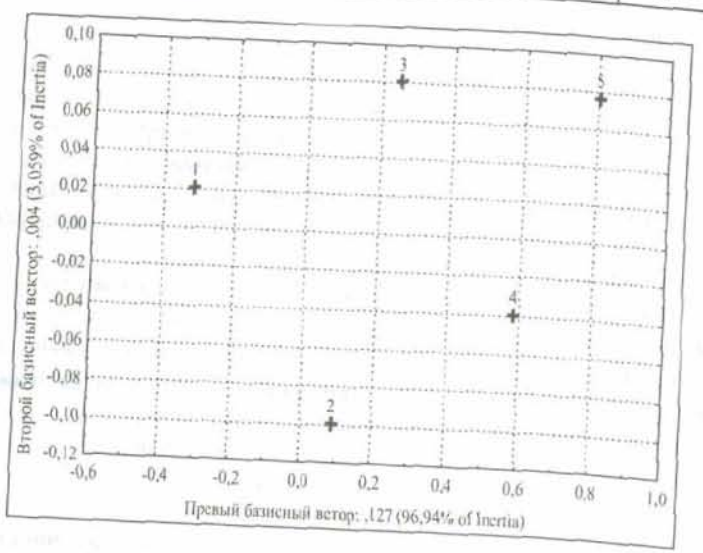

Рис.8.6.Распределение пострадавших по группам систолического АД после вторичной группировки.

Таблица 8.16

Двухвходовая таблица после повторной группировки систолического АД на трех уровнях

ПЖА	Систолическое АД			Всего
	нормотония	умеренная гипертония и умеренная гипотония	критическая гипотония, выраженная гипотония, выраженная гипертония и грубая гипертония	
1	82	39	4	125
2	16	15	5	36
3	46	60	33	139
Всего	144	114	42	300

Таблица 8.17

Собственные значения и инерция для всех базисных векторов

Таблица ввода (Строки × Столбцы): 3×3 Общая инерция =,12352 χ^2=37,056 df=4 p=,00000					
Базисные векторы	Сингул. значения	Собств. значения	Процент инерции	Кумулятивный процент	χ-квадрат
1	0,35	0,12	99,55	99,55	36,89
2	0,02	0,00	0,45	100,00	0,17

Таблица 8.18

Координаты двух первых базисных векторов для групп пострадавших по систолическому АД после группировки на трех уровнях

Группа ПЖА	Координаты по 1 вектору	Координаты по 2 вектору	Масса	Качество
1	-0,32	0,01	0,48	1,00
2	0,15	-0,03	0,38	1,00
3	0,69	0,04	0,14	1,00

Таким образом, с помощью математико-статистического метода анализа соответствия нами получена оптимальная группировка признаков ПЖА и систолического АД с минимальной размерностью 3×3.

Подобным образом анализируются все остальные предиктивные признаки, содержащиеся в исходной матрице наблюдений. В результате исследователь уточняет шкалирование, находит оптимальную группировку признаков, ранжирует их по мере связи (общей инерции) с признаком ПЖА и формирует обновленную матрицу наблюдений для моделирования результирующих признаков.

Построение моделей ПЖА для прогноза исхода и оценки эффективности лечения ЧМТ в зависимости от наиболее важных предиктивных признаков выполняется методами дискриминантного или регрессионного анализа.

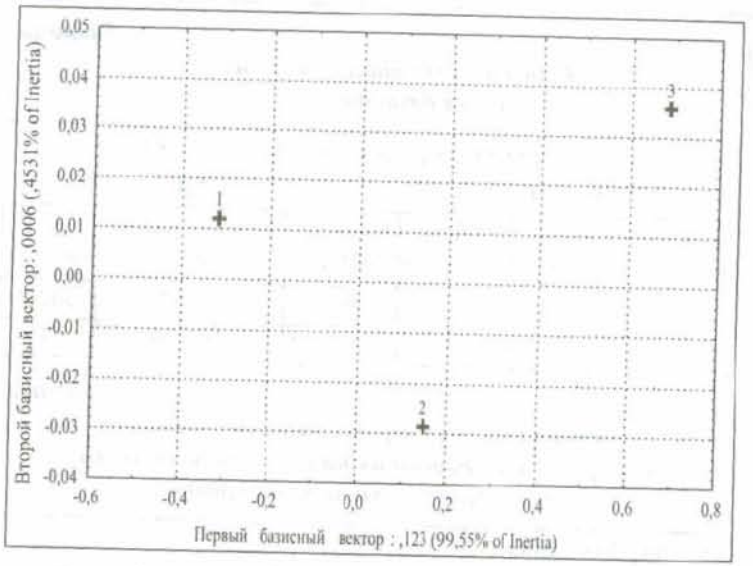

Рис.8.7.Распределение пострадавших по группам систолического АД после группировки на трех уровнях.

Литература

1. Benzecri, J. P. (1973). L'Analyse des Donnees: T. 2, l' Analyse des correspondances. Paris: Dunod.

2. Carrol, J. D., Green, P. E., and Schaffer, C. M. (1986). Interpoint distance comparisons in correspondence analysis. Journal of Marketing Research, 23, 271-280.

3. Greenacre, M. J. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. Biometrica, 75, 457-467.

4. Lebart, L., Morineau, A., and Tabard, N. (1977). Techniques de la description statistique. Paris: Dunod.

5. StatSoft, Inc. (2001). Электронный учебник по статистике. Москва, StatSoft. WEB: http://www.statsoft.ru/home/textbook/ default.htm.

Глава 9. ЛОГЛИНЕЙНЫЙ АНАЛИЗ

Сущность, условия применения и задачи логлинейного анализа

Часто в медицинских исследованиях регистрируемые переменные величины являются категорированными, т.е. оцениваемыми качественно. Тогда, при наблюдении n объектов, каждый из k признаков, описывающих эти объекты, будет представлен в исходной матрице кодами их категорий (уровней). По данным исходной матрицы наблюдений получают таблицу сопряженности, включающую частоты наблюдений на всех сочетаниях уровней признаков (табл.9.1).

Многофакторный анализ таблиц сопряженности проводится для решения задач моделирования частот наблюдений в ячейках таблицы, т.е. на всех возможных сочетаниях уровней признаков, сравнения их и поиска оптимальных уровней, представляющих собой входные факторы, при которых признак-отклик получает требуемое значение и др.

Проведение анализа связанно с поиском ответов на следующие вопросы: как корректно обойтись с ограниченным объемом данных; как поступить с малым числом наблюдений или с отсутствием данных в некоторых ячейках; как решить проблему нелинейного изменения частот при изменении уровней входных факторов и результирующих признаков-откликов. В то же время применение традиционных методов многомерного статистического анализа, таких, как корреляционный, регрессионный, дисперсионный для исследования таблиц сопряженности, невозможно. Непараметрические методы анализа частотных таблиц сопряженности позволяют решать только некоторые частные задачи оценки связи и значимости различия частот наблюдения в исследуемых группах и не дают возможности решать главную задачу по моделированию частот наблюдений для различных сочетаний уровней факторов.

По нашему мнению, наиболее эффективным методом многомерного моделирования по таблицам сопряженности является логлинейный анализ, который разработан в последние десятилетия и реализован в ППП для ПК. Логлинейный анализ обеспечивает:

– установление силы и значимости связей между признаками с учетом их взаимодействия;

– определение степени влияния входных факторов на выходные результирующие признаки-отклики;

– прогнозирование ожидаемых частот наблюдений при определенных сочетаниях уровней факторов.

В основу метода положено утверждение, что выборочные частоты анализируемой таблицы сопряженности n_{ijk}... порождаются теоретическими частотами $\hat{n}_{ijk}$..., характеризующими генеральную совокупность. Теоретические частоты отвечают определенным гипотезам о связях, формируемым в виде моделей для каждой ячейки таблицы сопряженности. Доказано, что оптимальными являются логлинейные модели, параметры которых определяются методами максимального правдоподобия. Так, для трехфакторной таблицы сопряженности, в которой фактор A исследуется на I уровнях (i=1, 2, ... I), B - на J уровнях (j=1, 2, ... J), C - на K уровнях (k=1, 2, ... K), логлинейная модель для прогнозируемой частоты наблюдений имеет вид:

$$\ln \hat{n}_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}, \qquad (9.1)$$

где $\ln \hat{n}_{ijk}$ – логарифм теоретической частоты наблюдений на ijk сочетании уровней факторов A, B и C.

λ - логарифм константы;

$\lambda_i^A, \lambda_j^B, \lambda_k^C$ – логарифмы эффектов факторов A, B и C соответственно на i, j, k уровнях;

$\lambda_{ij}^{AB}, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}$ – логарифмы эффектов взаимодействия факторов A и B, A и C, B и C соответственно на ij , ik, jk сочетаниях уровней факторов;

λ_{ijk}^{ABC} – логарифм эффекта тройного взаимодействия факторов A, B и C на ijk сочетании уровней факторов.

Иногда в исходной частотной таблице сопряженности образуются пустые ячейки, и тогда задача не имеет решения. В подобных случаях программой предусматривается автоматическое увеличение числа наблюдений во всех ячейках таблицы сопряженности на очень малую величину Δ (дельта), которая по умолчанию принимается равной 0,5. Эта добавка учитывается должным образом в алгоритме решения задачи и поэтому совершенно не сказывается на результатах анализа.

Переходя от логарифмов к натуральным значениям анализируемых данных, получают теоретические частоты $\hat{n}_{ijk}$.

Гипотеза об адекватности модели, т.е. соответствии теоретических частот, полученных по модели $\hat{n}_{ijk}$ (Fitted Freguency), наблюдавшимся частотам n_{ijk} (Observed Freguency) проверяется по критерию χ^2 Пирсона.

$$\chi^2 = \sum_{IJK}(n_{ijk} - \hat{n}_{ijk})^2 / \hat{n}_{ijk}, \qquad (9.2)$$

где суммирование выполняется для всех IJK ячеек таблицы сопряженности. Адекватность модели проверяется также и по критерию χ^2_{ML} максимального правдоподобия (Likelihood Chi-square). Чувствительность этих двух способов проверки примерно одинакова. Модель признается адекватной при незначительном различии критериев χ^2 и вероятности соответствия (уровне значимости) $p>0,05$; при этом вероятность неадекватности будет $1-p<0,95$. Модель считается неадекватной, если вероятность соответствия (уровень значимости) p будет $\leq 0,05$, а вероятность неадекватности $(1-p)\geq 0,95$.

Адекватная модель со значимыми эффектами применяется для решения различных задач исследования, в частности:

– изучения характера изменения ожидаемых частот наблюдений при различных сочетаниях уровней факторов;

– определения степени влияния эффектов факторов и их взаимодействий на ожидаемые частоты наблюдений;

– поиска оптимальных уровней факторов для получения ожидаемых частот наблюдений на требуемых уровнях;

– прогноза соотношения ожидаемых частот наблюдений при задании уровней факторов и др.

Некоторые из перечисленных выше задач будут рассмотрены далее. Для их решения формируется файл данных в виде матрицы наблюдений размером n×m, где n - число наблюдавшихся объектов в выборке (например, больных), m - число признаков, характеризующих наблюдавшиеся объекты. В матрицу включаются только качественные признаки, оцениваемые по номинальной шкале. Исследуемые категории (уровни) признаков кодируются числами натурального ряда. Например, первому уровню присваивается - 1, второму - 2, третьему - 3 и т.д. Матрица исходных данных по числу наблюдаемых объектов n не ограничивается. Число признаков m, включающее как входные, воздействующие на объект факторы, так и результирующие признаки-отклики на воздействия, не должно быть большим и обычно, для удобства анализа, ограничивается 5-6.

Решение на ПК включает.

1. Расчет четырехпольных таблиц наблюдавшихся частот (Obs.Freq.) для всех сочетаний уровней признаков.

2. Оценка связи частот наблюдений с эффектами признаков их взаимодействия по величине коэффициентов парциальной и маргинальной ассоциации.

3. Определение начальной модели для дальнейшего автоматического поиска окончательной, адекватной модели (Best, initial model).

4. Определение лучшей адекватной модели, имеющей наименьшее число значимых эффектов ожидаемых частот наблюдений для различных сочетаний уровней признаков, методом пошагового исключения незначимых эффектов (Best model).

5. Расчет четырехпольных таблиц ожидаемых частот (Fitted Freg.) для всех сочетаний уровней признаков на основе полученной адекватной модели.

6. Расчет четырехпольных таблиц разностей наблюдавшихся и ожидаемых частот и стандартизованных разностей для всех сочетаний уровней признаков (Residual and Standardized Residual).

7. Построение графика, иллюстрирующего распределение наблюдаемых частот относительно ожидаемых (теоретических, рассчитанных по адекватной модели).

8. Расчет четырехпольных таблиц частот наблюдений для всех признаков попарно (Marg.tabl).

Степень влияния эффектов факторов и их взаимодействия на ожидаемые частоты наблюдений определяется по данным таблицы коэффициентов парциальной и маргинальной ассоциации с последующей оценкой их значимости по методу χ^2 для полной насыщенной модели (Tests of Marginal and Partial Association).

С помощью этой же таблицы (см. примеры) оценивается важность эффектов факторов и их взаимодействия для объяснения исследуемого явления.

О степени влияния того или иного эффекта судят по величине отношения критерия χ^2_m данного m-эффекта к сумме $\sum\chi^2_m$ всех эффектов в %

$$K_m = 100\chi^2_m / \sum\chi^2_m . \qquad (9.3)$$

При исследовании связи результирующего признака-отклика с действующими факторами, определяющими частоты наблюдений, из таблицы оценки значимости эффектов по насыщенной модели выбирают те эффекты, которые связаны с исследуемым признаком-откликом. Степень связи того или иного эффекта оценивается в % по отношению χ^2_m для m-го эффекта к сумме $\sum\chi^2_m$ для всех эффектов, связанных с исследуемым признаком-откликом (9.3).

ПРИМЕР 9.1. Исследование связи показателя устойчивости результатов лечения с факторами, характеризующими социально-бытовые условия, на основе пятифакторной логлинейной модели

Для практического изучения элементов и порядка применения логлинейного анализа проведено обследование 1561 пациента, лечившегося по поводу хронического алкоголизма и наблюдавшегося в течение двух лет после лечения в Международном институте по изучению резервных возможностей человека. Исследовались пять признаков, один из которых признак-отклик (показатель устойчивости лечения) и четыре других - входные признаки, характеризующие социально-бытовые условия. Все пять признаков дихотомические и регистрируются на двух уровнях: 1 и 2, а частотная таблица содержит 32 ячейки.

Введем следующие обозначения.

A - показатель устойчивости результата лечения: A1 - продолжительная ремиссия после лечения, A2 - рецидивы алкоголизма;

B - *фактор удовлетворенности семейной жизнью: B1 - оценивается* положительно, B2 - оценивается отрицательно;

C - образование больного: C1 - среднее и высшее, C2 - ниже среднего;

D - фактор признания и следования принципам христианской морали: D1 - оценивается положительно, D2 - оценивается отрицательно;

E - фактор удовлетворенности профессиональной деятельностью: E1 - оценивается положительно, E2 - оценивается отрицательно.

Результаты исследования представлены в массиве исходных данных - матрице наблюдений размером 1561x5, где: 1561 - *число объектов* наблюдения, 5 - количество признаков (A, B, C, D, E). Для каждого пациента кодами 1 и 2 указаны значения наблюдавшихся у него признаков. Эта матрица не приводится ввиду ее больших размеров. Сгруппированная матрица наблюдений представлена в виде частотной таблицы (табл.9.1), которая характеризует распределение частот наблюдений для всех возможных сочетаний уровней факторов.

При рассмотрении данной *матрицы видно*, что все 32 ячейки не содержат нулевых значений частот наблюдений. Поэтому введение добавки к этим частотам не обязательно. Тем не менее, эту добавку (Δ =0,5) можно и сохранить по умолчанию, поскольку она по существу не изменит результатов анализа.

Таблица 9.1

Частота наблюдений на различных сочетаниях уровней факторов

N ячейки	Уровни факторов					Число наблюдений
	A	B	C	D	E	
1	1	1	1	1	1	51
2	1	1	1	1	2	31
3	1	1	1	2	1	142
4	1	1	1	2	2	62
5	1	1	2	1	1	11
6	1	1	2	1	2	34
7	1	1	2	2	1	37
8	1	1	2	2	2	61
9	1	2	1	1	1	51
10	1	2	1	1	2	83
11	1	2	1	2	1	64
12	1	2	1	2	2	57
13	1	2	2	1	1	23
14	1	2	2	1	2	106
15	1	2	2	2	1	19
16	1	2	2	2	2	99
17	2	1	1	1	1	8

N ячей-ки	Уровни факторов					Число наблюдений
	A	B	C	D	E	
18	2	1	1	1	2	8
19	2	1	1	2	1	37
20	2	1	1	2	2	23
21	2	1	2	1	1	6
22	2	1	2	1	2	16
23	2	1	2	2	1	11
24	2	1	2	2	2	24
25	2	2	1	1	1	35
26	2	2	1	1	2	94
27	2	2	1	2	1	21
28	2	2	1	2	2	54
29	2	2	2	1	1	15
30	2	2	2	1	2	143
31	2	2	2	2	1	25
32	2	2	2	2	2	110

Решение задачи проведено на ПК с помощью ППП Statistica for Windows и имеет результаты (машинограммы 9.1-9.7), которые обсуждаются ниже.

В машинограмме 9.1 в 8 четырехпольных таблицах даны частоты наблюдений (Obs.Freg) для различных сочетаний уровней факторов.

В машинограмме 9.2 в таблице Results of Fitting all K-Factor Interaction приводится оценка значимости эффектов К-го порядка. Как известно, эффекты признаются значимыми при $p<0,05$ (так же, как и при оценке коэффициентов при построении уравнения регрессии), что соответствует достоверности $1-p>0,95$. Из таблицы видно, что значимыми являются эффекты первого и второго порядков, т.е. эффекты всех факторов по отдельности и эффекты их парных взаимодействий.

В этой же машинограмме в таблице Test of Marginal and Partial Association имеются оценки значимости ассоциации (связи) частот с *эффектами факторов 1-4-го порядков в полной насыщенной модели* для ожидаемых частот наблюдений. Кроме того, из содержания таблицы можно заключить, что значимыми, достоверными в порядке убывания важности, являются 14 эффектов. Ниже эти эффекты перечислены

(в скобках указаны оценки вкладов каждого): E (131,0); B (124,0); CE (116,1); BD (68,8); AB (63,7); A (58,4); BE (55,6); AE (13,6); BC (11,5); D и DE (11,0); AC (6,4); ABCD (5,0); C (4,2).

Степень влияния K_m на частоты наблюдений при сумме $\sum\chi^2 = 695,1$ для всех эффектов оценивается по (9.3). Результат расчета вклада значимых эффектов дан в таблице 9.2.

Таблица 9.2

Степень влияния эффектов факторов на частоты наблюдений

Эффекты факторов	K_m, %	Эффекты факторов	K_m, %	Эффекты факторов	K_m, %
E	18,9	A	8,4	DE	1,6
B	17,9	BE	8,0	AC	0,9
CE	16,7	AE	2,0	ABCD	0,7
BD	9,9	BC	1,7	C	0,6
AB	9,2	D	1,6		

Как видно из таблицы, значимые эффекты факторов в сумме объясняют наблюдавшиеся в исследовании частоты на 98,2%. Доля незначимых эффектов составляет всего 1,8%.

Степень связи факторов с показателем А, характеризующим результаты лечения (при сумме $\sum\chi^2 = 156,7$ для эффектов, связанных с показателем А) оценивается также по (9.3). Результаты расчета для значимых эффектов показаны в табл.9.3, из которой следует, что результат лечения на 37,3% зависит от эффективности собственно терапии и на 56,7% определяется факторами B, E, C и их взаимодействием. Напомним, что это - удовлетворенность семейной жизнью и профессиональной деятельностью, а также уровень образования пациента.

Таблица 9.3

Степень связи факторов с показателем результатов лечения

Эффекты факторов	K_m, %	Эффекты факторов	K_m, %
A	37,3	AC	4,1
AB	40,7	ABCD	3,2
AE	8,7		

В машинограмме 9.3 содержатся итоговые результаты автоматического поиска оптимальной адекватной модели для ожидаемых частот наблюдений. В первой части таблицы Automatic Selection of Best Model даны коды эффектов начальной модели (Best initial model). Во второй части - коды эффектов оптимальной модели с наименьшим числом значимых из них, но обеспечивающих, в то же время, достаточную ее адекватность (Best Model). Оптимальной оказалась модель, включающая эффекты взаимодействия факторов 21, 31, 32, 54, 42, 53, 51, 52 с вероятностью ее адекватности по критерию χ^2 =22,32, p=0,218, что больше требуемого уровня p>0,05. Прогноз ожидаемых частот наблюдений по этой модели вполне допустим.

В машинограмме 9.4 в 8 таблицах показаны результаты расчета ожидаемых (теоретических) частот наблюдений для различных сочетаний уровней факторов (Fitted Freg). Прогноз ожидаемых частот наблюдений для различных уровней факторов дается по таблицам этой машинограммы. Сравнение теоретических частот с наблюдавшимися (в машинограмме 9.1) показывает, что различия оказались незначительными, а это подтверждает адекватность оптимальной модели. Прогноз ожидаемых частот наблюдений для различных уровней факторов осуществляется по таблицам именно этой машинограммы.

Машинограммы 9.5 и 9.6 включают в себя по 8 таблиц каждая, с разностями и стандартизованными разностями (в средних квадратичных отклонениях разностей) наблюдавшихся и ожидаемых частот на всех сочетаниях уровней факторов. Анализ таблиц машинограммы 9.6 показывает, что стандартизованные разности не выходят за пределы интервала (-2; 2), т.е. закон их распределения близок к нормальному. Характер распределения отклонений наблюдавшихся частот от теоретических (по модели) изображен на рис.9.1. Точки, реально наблюдавшихся частот, незначительно рассеяны относительно прямой, описывающей теоретические частоты по модели.

Проведем содержательный разбор 8 таблиц наблюдавшихся частот, сгруппированных в четырехпольные таблицы для попарного сочетания уровней факторов. Эти таблицы используются для окончательного суждения об оценке результатов влияния факторов и их взаимодействия на исследуемые явления. Так, из первой таблицы следует: во-первых, при всем многообразии входящих факторов, положительный результат лечения (A1) преобладал над отрицательным (A2) в соотношении 931:630=1,48:1,00; во-вторых, положительная оценка удовлетворенности семейной жизнью (B1), также при всем разнообразии других изучаемых факторов, наблюдалась реже, чем отрицательная (B2) в соотношении 562:999=0,56:1,00; в-третьих, при отрицательном результате лечения (A2) отрицательная оценка удовлетворенности семейной жизнью (B2) значительно превосходила положительную (B1) в соотношении 497:133=3,74:1,00.

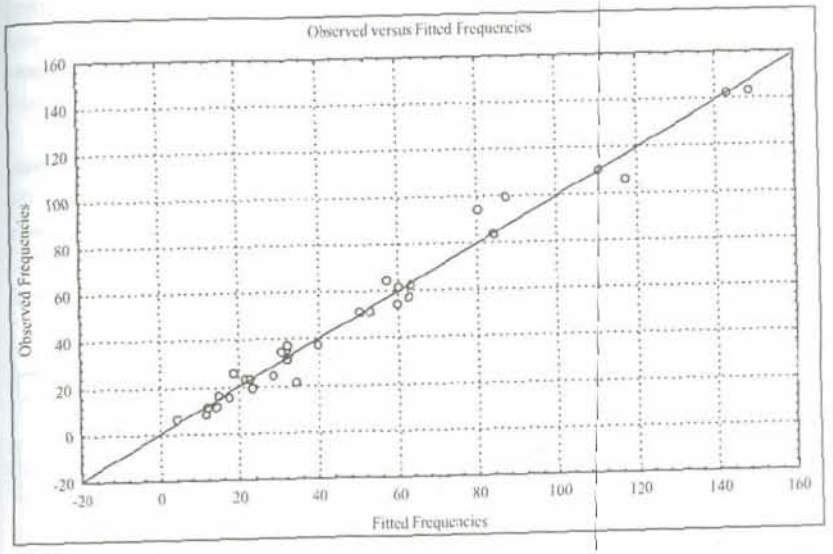

Рис.9.1.Распределение отклонений наблюдавшихся частот от теоретических (по модели).

Данные машинограммы 9.7 указывают на преобладание отрицательной оценки удовлетворенности профессиональной деятельностью (E2) над положительной (E1) при отрицательных результатах лечения в соотношении 472:158=2,99:1,00.

Аналогичный анализ можно выполнить и в отношении других исследованных факторов и их взаимодействия.

Таким образом, применение логлинейного анализа позволяет сделать следующие выводы.

1. При всем разнообразии исследуемых социально-бытовых условий, положительный результат лечения хронического алкоголизма составил 59,6%.

2. Результат лечения на 56,7% определяется влиянием следующих факторов: удовлетворенности семейной жизнью (40,7%), удовлетворенности профессиональной деятельностью (8,7%), уровнем образования пациента (4,1%), а также синергическим эффектом взаимосвязи всех перечисленных выше факторов (3,2%).

3. По полученной модели осуществлен прогноз исходов лечения для различных условий (по таблицам машинограммы 9.4). Например, при оценках на 2-м уровне факторов B, C, D и E (8 таблица машинограммы 9.4) положительный результат лечения ожидается в 86,7 случаях, отрицательный - в 109,8 случаях из 196,5, что составляет вероятность положительного – 44,1% и отрицательного – 55,9% соответственно.

4. Положительный исход лечения хронического алкоголизма значимо зависит от благополучия, удовлетворенности в семейной жизни, от профессиональной деятельности, образования и в меньшей степени - от признания принципов христианской морали и следования им.

Машинограмма 9.1

Таблица частот наблюдений для различных сочетаний уровней признаков

A	Obs. Freq.: C=1	A by B D=1	w/in vars: E=1
	B1	B2	Total
1	51	51	102
2	8	35	43
Total	59	86	145

A	Obs. Freq.: C=2	A by B D=1	w/in vars: E=1
	B1	B2	Total
1	11	23	34
2	6	15	21
Total	17	38	55

A	Obs. Freq.: C=1	A by B D=2	w/in vars: E=1
	B1	B2	Total
1	142	64	206
2	37	21	58
Total	179	85	264

A	Obs. Freq.: C=2	A by B D=2	w/in vars: E=1
	B1	B2	Total
1	37	19	56
2	11	25	36
Total	48	44	92

A	Obs. Freq.: C=1	A by B D=1	w/in vars: E=2
	B1	B2	Total
1	31	83	114
2	8	94	102
Total	39	177	216

A	Obs. Freq.: C=2	A by B D=1	w/in vars: E=2
	B1	B2	Total
1	34	106	140
2	16	143	159
Total	50	249	299

A	Obs. Freq.: C=1	A by B D=2	w/in vars: E=2
	B1	B2	Total
1	62	57	119
2	23	54	77
Total	85	111	196

A	Obs. Freq.: C=2	A by B D=2	w/in vars: E=2
	B1	B2	Total
1	61	99	160
2	24	110	134
Total	85	209	294

Машинограмма 9.2

Проверка значимости эффектов К-го порядка

crossprd K-Factor	Results of Fitting all K-Faktor Interactions These are simultaneous tests that all K-Factor Interactions are simultaneously Zero.				
	Degrs.of Freedom	Max.Lik. Chi-squ.	Probab. P	Pearson Chi-squ	Probab. P
1	5	328,5961	0,0000	227,6470	0,0000
2	10	543,0063	0,0000	703,7083	0,0000
3	10	9,3665	,4977	9,0996	,52268
4	5	7,4013	,1925	7,2765	,20090
5	1	3,0777	,0793	3,1148	,07759

Оценка значимости эффектов факторов и их взаимодействий в полной насыщенной модели

crossprd Effekt	Tests of Marginal and Partial Association				
	Degrs.of Freedom	Prt.Ass. Chi-squ.	Prt.Ass P	Mrg.Ass. Chi-squ	Mrg.Ass P
1	1	58,4056	,00000	58,4056	,000000
2	1	123,9879	,00000	123,9879	,000000
3	1	4,2050	,04031	4,2050	,040314
4	1	11,0065	,00090	11,0065	,000909
5	1	130,9913	,00000	130,9913	,000000
12	1	63,7195	,00000	105,6719	,000000
13	1	6,4255	,011254	28,1885	,000000
14	1	1,1258	,288680	14,2277	,000162
15	1	13,6471	,000221	52,3804	,000000
23	1	11,5149	,000691	49,7368	,000000
24	1	68,8490	,000000	97,9188	,000000
25	1	55,6353	,000000	126,7349	,000000
34	1	1,5040	,220060	2,3446	,125727
35	1	116,1453	,000000	157,0630	,000000
45	1	11,0314	,000897	33,9780	,000000
123	1	,0599	,806707	,0023	,961590
124	1	1,3927	,237960	1,1192	,290092

crossprd Effekt	Tests of Marginal and Partial Association				
	Degrs.of Freedom	Prt.Ass. Chi-squ.	Prt.Ass P	Mrg.Ass. Chi-squ	Mrg.Ass P
125	1	,8929	,344705	1,0789	,298954
134	1	,0681	,794193	,9742	,323647
135	1	2,4053	,120933	2,9763	,084504
145	1	,0162	,898833	,0233	,878798
234	1	3,0525	,080621	4,2220	,039910
235	1	,1284	,720074	,5773	,447390
245	1	,0142	,905304	,2635	,607695
345	1	,0013	,970759	,4832	,486975
1234	1	4,9823	,025615	4,5068	,033768
1235	1	,6327	,426369	,3891	,532802
1245	1	,0698	,791567	,0880	,766799
1345	1	2,5225	,112245	,7356	,391072
2345	1	,5566	,455632	,3606	,548188

Машинограмма 9.3

Результаты автоматического пошагового поиска оптимальной модели

Automatic Selection of Best Model

Table to be analized: (1) A x 2 x (2) B x 2 x (3) C x 2 x (4) D x 2 x (5) E 2

Minimum cell frequency: 6. Maximum: 143. Sum: 1561.

Best initial model: Chi-Square = 19,8454 df = 16 p= ,2273
21, 31, 32, 41, 42, 43, 51, 52, 53, 54

Best Model: Chi-Square=19,8454 df=16 p=,2183
21, 31, 32, 54, 42, 53, 51, 52

Press any key to continue

Note: Use option M from MODEL SPECIFICATION menu to further evaluate model.

Машинограмма 9.4

Таблица ожидаемых частот наблюдений, рассчитанных по модели для различных сочетаний уровней признаков

A	Fitted Freq.: C=1	A by B D=1	w/in vars: E=1
	B1	B2	Total
1	49,60427	52,49503	102,0993
2	10,95036	31,42865	42,3790
Total	60,55463	83,92368	144,4783

A	Fitted Freq.: C=2	A by B D=1	w/in vars: E=1
	B1	B2	Total
1	13,62829	21,13340	34,76169
2	3,98798	16,77178	20,75976
Total	17,61627	37,90518	55,52145

A	Fitted Freq.: C=1	A by B D=2	w/in vars: E=1
	B1	B2	Total
1	142,6680	56,52055	199,1885
2	31,4946	33,83870	65,3333
Total	174,1625	90,35925	264,5218

A	Fitted Freq.: C=2	A by B D=2	w/in vars: E=1
	B1	B2	Total
1	39,19664	22,75398	61,95061
2	11,46992	18,05791	29,52783
Total	50,66656	40,81189	91,47845

A	Fitted Freq.: C=1	A by B D=1	w/in vars: E=2
	B1	B2	Total
1	31,55356	83,5632	115,1168
2	11,11779	79,8516	90,9694
Total	342,67135	163,4149	206,0862

A	Fitted Freq.: C=2	A by B D=1	w/in vars: E=2
	B1	B2	Total
1	30,09944	116,8030	146,9024
2	14,05825	147,9534	162,0116
Total	44,15768	264,7563	308,9140

A	Fitted Freq.: C=1	A by B D=2	w/in vars: E=2
	B1	B2	Total
1	62,56671	62,0284	124,5951
2	22,04518	59,2733	81,3185
Total	84,61189	121,3018	205,9137

A	Fitted Freq.: C=2	A by B D=2	w/in vars: E=2
	B1	B2	Total
1	59,68338	86,7021	146,3855
2	27,87572	109,8249	137,7006
Total	87,55910	196,5270	284,0861

Машинограмма 9.5

Таблица разностей наблюдавшихся и ожидаемых частот наблюдений для различных сочетаний уровней признаков

A	Res. Freq.: C=1	A by B D=1	w/in vars: E=1
	B1	B2	Total
1	1,39573	-1,49503	-,099300
2	-2,95036	3,57135	-,620999
Total	-1,55463	2,07632	,521699

A	Res. Freq.: C=2	A by B D=1	w/in vars: E=1
	B1	B2	Total
1	-2,62829	1,86660	-.761692
2	2,01202	-1,77178	,240238
Total	-,61627	,09482	-,521455

A	Res. Freq.: A by B C=1	D=2	w/in vars: E=1
	B1	B2	Total
1	-,667984	7,4795	6,81147
2	5,505451	-12,8387	-7,33325
Total	4,837467	-5,3593	-,52179

A	Res. Freq.: C=2	A by B D=2	w/in vars: E=1
	B1	B2	Total
1	-2,19664	-3,75398	-5,95062
2	-,46992	6,94209	6,47217
Total	-2,66656	3,18811	,52155

A	Res. Freq.: C=1	A by B D=1	w/in vars: E=2
	B1	B2	Total
1	-,55356	-,56324	-1,11680
2	-3,11779	14,14837	11,03058
Total	-3,67135	13,58513	9,91378

A	Res. Freq.: C=2	A by B D=1	w/in vars: E=2
	B1	B2	Total
1	3,900564	-10,8030	-6,90239
2	1,941750	-4,9534	-3,01164
Total	5,842314	-15,7563	-9,91403

A	Res. Freq.: C=1	A by B D=2	w/in vars: E=2
	B1	B2	Total
1	-,566711	-5,0284	-5,59515
2	,954821	-5,2733	-4,31852
Total	,388109	-10,3018	-9,91366

A	Res. Freq.: C=2	A by B D=2	w/in vars: E=2
	B1	B2	Total
1	1,31662	12,29789	13,61451
2	-3,87573	,17509	-3,70063
Total	-2,55910	12,47298	9,91388

Машинограмма 9.6

Таблица стандартизованных разностей наблюдавшихся и ожидаемых частот наблюдений для различных сочетаний уровней признаков

A	Stdrd. Resid.: C=1	A by B D=1	w/in vars: E=1
	B1	B2	Total
1	,198172	-,206343	-,008172
2	-,891580	,637045	-,254535
Total	-,693408	,430702	-,262706

A	Stdrd. Resid.: C=2	A by B D=1	w/in vars: E=1
	B1	B2	Total
1	-,711956	,406038	-,305918
2	1,007525	-,432634	,574891
Total	,295569	-,026596	,268973

A	Stdrd. Resid.: C=1	A by B D=2	w/in vars: E=1
	B1	B2	Total
1	-,055925	,99487	,93895
2	,981014	-2,20706	-1,22605

A	Stdrd. Resid.: C=2	A by B D=2	w/in vars: E=1
	B1	B2	Total
1	-,350861	-,786979	-1,13784
2	-,138754	1,633640	1,49489
Total	-,489615	,846662	,35705

A	Stdrd. Resid.: C=1	A by B D=1	w/in vars: E=2
	B1	B2	Total
1	-,09855	-,061615	-,160161
2	-,93506	1,583305	,648248
Total	-1,03360	1,521690	,488087

A	Stdrd. Resid.: C=2	A by B D=1	w/in vars: E=2
	B1	B2	Total
1	,710965	-,99958	-,288611
2	,517878	-,40723	,110648
Total	1,228843	-1,40681	-,177962

A	Stdrd. Resid.: C=1	A by B D=2	w/in vars: E=2
	B1	B2	Total
1	-,071646	-,63847	-,71011
2	,203360	-,68495	-,48159
Total	,131714	-1,32341	-1,19170

A	Stdrd. Resid.: C=2	A by B D=2	w/in vars: E=2
	B1	B2	Total
1	,170426	1,320735	1,491161
2	-,734074	,016708	-,717366
Total	-,563648	1,337446	,773795

Таблицы частот наблюдений для сочетаний уровней факторов, включенных в модель

B	Marg. Tabl.: A1	A by B A2	Total
1	429	133	562
2	502	497	999
Total	931	630	1561

C	Marg. Tabl.: A1	A by C A2	Total
1	541	280	821
2	390	350	740
Total	931	630	1561

C	Marg. Tabl.: B1	B by C B2	Total
1	362	459	821
2	200	540	740
Total	562	999	1561

E	Marg. Tabl.: D1	D by E D2	Total
1	200	356	556
2	515	490	1005
Total	715	846	1561

D	Marg. Tabl.: B1	B by D B2	Total
1	165	550	715
2	397	449	846
Total	562	999	1561

E	Marg. Tabl.: C by E		
	C1	C2	Total
1	409	147	556
2	412	593	1005
Total	821	740	1561

E	Marg. Tabl.: A by E		
	A1	A2	Total
1	398	158	556
2	533	472	1005
Total	931	630	1561

E	Marg. Tabl.: B by E		
	B1	B2	Total
1	303	253	556
2	259	746	1005
Total	562	999	1561

ПРИМЕР 9.2. Построение и анализ трехфакторной логлинейной модели оценки профессиональной деятельности операторов

Рассмотрим вариант решения задачи на другом примере, в котором исследуемые факторы наблюдались на двух и трех уровнях, а частотная таблица содержит IJK ячеек (I, J и K - число уровней факторов A, B и C). Подходящим инструментом для решения задач такого класса также может служить логлинейный анализ.

Объектами наблюдения являются 103 оператора сложной электронной медицинской аппаратуры, которые описываются следующим набором факторов.

Фактор-отклик A - успешность профессиональной деятельности операторов - на двух уровнях:

A1 - посредственная оценка деятельности (по времени и количеству допущенных ошибок),

A2 - хорошая оценка деятельности.

Определяющий фактор B - образование операторов - на трех уровнях:

B1 - среднее специальное,

B2 - высшее,

B3 - высшее и курсы повышения квалификации по специальности.

Определяющий фактор C - опыт работы операторов по специальности - на трех уровнях:

C1 - до 1 года,

C2 - от 1 до 3 лет,

C3 - более 3 лет.

Распределение обследованных по уровням факторов дано в матрице наблюдений (табл.9.4).

Таблица 9.4

Матрица данных наблюдений

№ пп	A	B	C	№ пп	A	B	C
1	2	1	2	53	1	2	2
2	2	2	2	54	1	1	2
3	2	1	3	55	2	1	1
4	1	3	1	56	2	1	2
5	2	2	2	57	2	2	2
6	2	1	3	58	1	1	1
7	2	2	2	59	2	1	3
8	2	1	3	60	2	1	2
9	1	2	3	61	2	1	1
10	2	2	2	62	1	1	2
11	2	1	3	63	2	2	2
12	2	2	2	64	2	1	1
13	2	1	3	65	2	1	2
14	2	2	2	66	1	1	1
15	2	2	2	67	2	1	2
16	2	1	3	68	2	2	2
17	2	3	2	69	1	1	1
18	1	2	2	70	2	1	2
19	2	2	3	71	2	2	2
20	2	1	3	72	2	1	3
21	2	3	2	73	2	1	2

№ пп	A	B	C	№ пп	A	B	C
22	2	2	3	74	1	1	1
23	2	1	2	75	2	1	2
24	2	2	2	76	2	1	2
25	2	1	2	77	1	2	1
26	2	2	3	78	2	1	2
27	2	1	3	79	1	1	2
28	2	1	2	80	1	1	1
29	2	1	2	81	1	2	1
30	1	2	1	82	1	1	1
31	2	1	3	83	2	1	2
32	1	2	2	84	1	1	1
33	2	1	2	85	1	1	1
34	2	2	2	86	1	1	1
35	1	1	2	87	1	2	1
36	2	2	3	88	1	1	2
37	1	1	1	89	2	1	1
38	2	1	2	90	1	1	2
39	2	1	3	91	1	1	2
40	1	1	2	92	1	1	2
41	2	1	3	93	1	2	1
42	1	1	1	94	1	1	1
43	2	1	2	95	1	1	1
44	2	1	2	96	2	1	2
45	1	2	2	97	1	1	1
46	2	1	3	98	1	1	1
47	2	1	2	99	1	1	2
48	1	2	1	100	1	1	1
49	2	1	2	101	1	2	1
50	2	1	2	102	2	1	1
51	2	2	2	103	1	1	1
52	2	2	3				

Решение в машинограммах 9.8-9.15. Как видно из машинограммы 9.8 с таблицами частот наблюдений для различных сочетаний уровней факторов, ячейки некоторых сочетаний содержат нули, что делает невозможным применение логлинейного анализа. Чтобы избавиться от нулей, необходимо воспользоваться специальным приемом, заключающимся в увеличении частоты во всех ячейках таблицы сопряженности на некоторую небольшую величину, например $\Delta = 0{,}1$. В результате будут получены новые наблюдаемые частоты (машинограмма 9.9).

Машинограмма 9.10 содержит результаты оценки значимости эффектов факторов 1, 2 и 3-го порядков, а также значения критериев χ^2 для всех эффектов насыщенной модели. Результаты пошагового определения оптимальной адекватной модели для ожидаемых частот наблюдений для различных сочетаний уровней факторов представлены в машинограмме 9.11.

Из первой таблицы машинограммы 9.10 видно, что значимыми являются эффекты 1 и 2-го порядков ($p < 0{,}05$). Согласно данным второй таблицы этой же машинограммы, значимыми эффектами, определяющими наблюдавшиеся частоты, в порядке убывания важности являются эффекты следующих факторов: B (χ^2=73,3); AC (40,5); C (16,0); A (5,1). Степень их влияния на частоты наблюдений при $\sum\chi^2_m$ =137,4 определяется по (9.3) и составляет:

Эффект факторов	Степень влияния K_m, %
B	53,3
AC	29,5
C	11,6
A	3,7
Сумма	98,1

Следовательно, значимые эффекты объясняют наблюдавшиеся частоты на 98,1%, незначимые - на 1,9%. Степень связи факторов B и C с показателем успешности профессиональной деятельности операторов (A) при $\sum\chi^2_m$ =46,1 характеризуется такими числами:

– эффект фактора A 11,1%;

– эффект взаимодействия A и C 87,9%;

– эффект взаимодействия A и B 1,0 %.

Все это указывает на то, что успешность деятельности операторов существенно зависит от опыта работы и в меньшей степени - от образования.

В машинограмме 9.11, кроме того, имеются результаты отбора начальной модели для пошаговой селекции (эта модель содержит эффекты взаимодействия 2 1, 3 1, 3 2) и оптимальной адекватной модели, включающей эффекты 3 1 и 2, с большой вероятностью ее адекватности p=0,648 при χ^2=7,80 и числе степеней свободы df=10.

Машинограмма 9.12 содержит три таблицы с теоретическими частотами, полученными по модели, машинограмма 9.13 - разности наблюдавшихся и ожидаемых частот, машинограмма 9.14 - стандартизованные разности частот, а рис.9.2 наглядно характеризует разброс частот наблюдений относительно ожидаемых (теоретических) по модели.

Анализ перечисленных таблиц и рис.9.2 убеждает в адекватности модели с наблюдавшимися частотами и высокой сходимости прогноза частот с наблюдавшимися в выборочном исследовании. По таблицам машинограммы 9.12 можно установить следующее: у сотрудников с опытом работы до 1 года (признак C - на уровне 1) успешность профессиональной деятельности при всех трех уровнях фактора B (специальное образование) оценивается посредственно (A1) в 25,3 случаев, хорошо (A2) - в 5,63 случаев с соотношением 4,77:1,00; с опытом работы от 1 до 3 лет (C2) для A1 – 14,3 случаев, для A2 – 39,3 случаев и соотношением 1,00:2,75; с опытом работы более 3 лет (C3) для A1 – 1,3 случая, для A2 – 19,3 случая и соотношением 1,00:14,8.

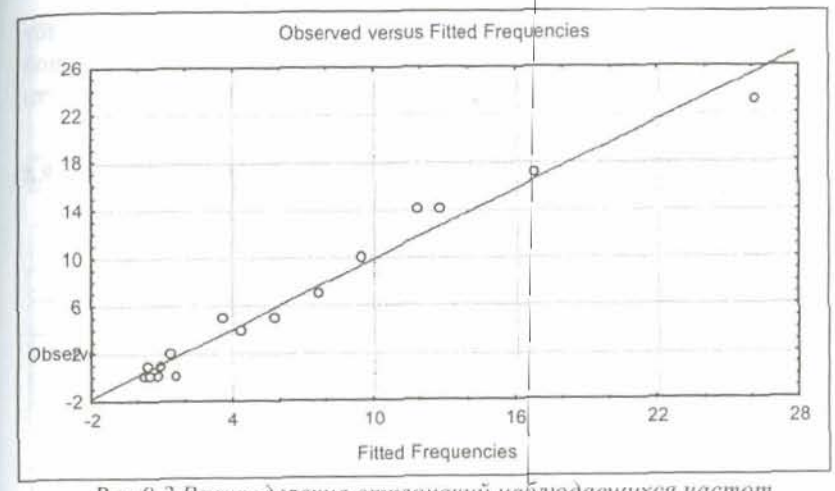

Рис.9.2.Распределение отклонений наблюдавшихся частот от теоретических (по модели).

Приведенные выше данные убеждают в том, что успешность деятельности операторов определяется в основном опытом их работы (конечно, при достаточном уровне специального образования, исследованного в эксперименте, т.е. среднего или высшего).

В машинограмме 9.15 представлены итоговые расчеты распределения операторов в зависимости от изученных факторов: по качеству профессиональной деятельности (фактор A) - посредственно (A1) оценивается 39,0%, хорошо (A2) – 61,0%; по образованию (фактор B) - со средним специальным образованием – 66,4%, с высшим образованием и курсом повышения квалификации по специальности – 36,4%; по опыту работы (фактор C) - с опытом работы до 1 года – 29,2%, от 1 до 3 лет – 51,1%, более 3 лет – 19,7%.

Выводы:

1. Распределение частот в таблице сопряженности трех исследованных переменных объясняется четырьмя значимыми эффектами факторов: B на 53,3%, AC на 29,5%, C на 11,6% и A на 3,7%.

2. Показатель успешности профессиональной деятельности операторов обусловлен главным образом фактором C - опытом работы (на 87,9%) и, в значительно меньшей степени, фактором B - образованием на исследованных уровнях (на 1,0%).

3. В результате исследования получена адекватная модель частот наблюдений для различных сочетаний уровней факторов, по которой можно прогнозировать успешность работы операторов в зависимости от их образования и опыта.

Машинограмма 9.8

Таблица частот наблюдений для различных сочетаний уровней факторов

A	Obs. Freq.:	A by B		w/in vars:C:1
	B1	B2	B3	Total
1	17	7	1	25
2	5	0	0	5
Total	22	7	1	30

A	Obs. Freq.:	A by B		w/in vars:C:2
	B1	B2	B3	Total
1	10	4	0	14
2	23	14	2	39
Total	33	18	2	53

A	Obs. Freq.:	A by B		w/in vars:C:3
	B1	B2	B3	Total
1	0	1	0	1
2	14	5	0	19
Total	14	6	0	20

Машинограмма 9.9

Таблица частот наблюдений, увеличенных на delta=0,1, для различных сочетаний уровней факторов

A	Obs. Freq.: (+delta)	A by B		w/in vars: C:1
	B1	B2	B3	Total
1	17,10000	7,100000	1,100000	25,30000
2	5,10000	,100000	,100000	5,30000
Total	22,20000	7,200000	1.200000	30,60000

A	Obs. Freq.: (+delta)	A by B		w/in vars: C:2
	B1	B2	B3	Total
1	10,10000	4,10000	,100000	14,30000
2	23,10000	14,10000	2,100000	39,30000
Total	33,20000	18,20000	2,200000	53,60000

A	Obs. Freq.: (+delta)	A by B		w/in vars: C:3
	B1	B2	B3	Total
1	,10000	1,100000	,100000	1,30000
2	14,10000	5,100000	,100000	19,0000
Total	14,20000	6,200000	,200000	20,60000

Машинограмма 9.10

Проверка значимости эффектов K-го порядка

crossprd K-Factor	Results of Fitting all K-Faktor Interactions These are simultaneous tests that all K-Factor Interactions are simultaneously Zero.				
	Degrs.of Freedom	Max.Lik. Chi-squ.	Probab.P	Pearson Chi-squ	Probab.P
1	5	94,37073	0,000000	103,4459	0,000000
2	8	42,23371	0,000000	35,5013	0,000022
3	4	5,67361	,224912	5,6043	,230740

Оценка значимости эффектов факторов и их взаимодействий в полной насыщенной модели.

crossprd Effekt	Tests of Marginal and Partial Association				
	Degrs.of Freedom	Prt.Ass. Chi-squ.	Prt.Ass P	Mrg.Ass. Chi-squ	Mrg.Ass P
1	1	5,08903	,024084	5,08903	,024084
2	2	73,25375	,000000	73,25375	,000000
3	2	16,02792	,000331	16,02792	,000331
12	2	,45053	,798307	,01426	,992894
13	2	40,54189	,000000	40,10564	,000000
23	4	2,11381	,714836	1,67755	,794790

Машинограмма 9.11

Результаты автоматического пошагового поиска оптимальной модели

Automatic Selection of Best Model

Table to be analized:

(1)		(2)		(3)
A	x	B	x	C
2	x	3	x	3

Minimum cell frequency: 0. Maximum: 23. Sum: 103.

Best initial model: Chi-Square = 5,672899 df = 4 p = ,2250
21, 31, 32

Best Model: Chi-Square = 7,801697 df = 10 p = ,6482
31, 2

Press any key to continue

Note: Use option M from MODEL SPECIFICATION menu to further evaluate model.

Машинограмма 9.12

Таблицы ожидаемых частот наблюдений, рассчитанных по модели для различных сочетаний уровней факторов

A	Fitted Freq.:	A by B		w/in vars: C:1
	B1	B2	B3	Total
1	16,80229	7,628626	,869084	25,30000
2	3,51985	1,598091	,182061	5,30000
Total	20,32214	9,226718	1,051145	30,60000

A	Fitted Freq.:	A by B		w/in vars: C:2
	B1	B2	B3	Total
1	9,49695	4,31183	,491221	14,30000
2	26,10000	11,8500	1,350000	39,30000
Total	35,59695	16,16183	1,841221	53,60000

A	Fitted Freq.:	A by B		w/in vars: C:3
	B1	B2	B3	Total
1	,86336	,391985	,044656	1,30000
2	12,81756	5,819466	,662977	19,30000
Total	13,68092	6,211451	,707634	20,60000

Машинограмма 9.13

Таблица разностей наблюдавшихся и ожидаемых частот наблюдений для различных сочетаний уровней факторов

A	Obs. Freq.:	A by B		w/in vars:C:1
	B1	B2	B3	Total
1	,297710	-,52863	,230916	,000000
2	1,580153	-1,49809	-,082061	,000000
Total	1,877863	-2,02672	-2,02672	,000001

A	Obs. Freq.:	A by B		w/in vars:C:2
	B1	B2	B3	Total
1	,60305	-,211833	-,391221	-,000001
2	-3,00000	2,250000	,750000	,000000
Total	-2,39695	2,038167	,358779	-,000001

A	Obs. Freq.:	A by B		w/in vars:C:3
	B1	B2	B3	Total
1	-,763359	,708015	,055344	-,000000
2	1,282442	-,719466	-,562977	-,000001
Total	,519083	-,011451	-,507634	-,000001

Машинограмма 9.14

Таблица стандартизованных разностей наблюдавшихся и ожидаемых частот наблюдений для различных сочетаний уровней факторов

A	Stdrd. Freq.:	A by B		w/in vars:C:1
	B1	B2	B3	Total
1	,072629	-,19139	,247699	,128935
2	,842243	-1,18505	-,192322	-,535131
Total	,914872	-1,37645	,055377	-,406197

A	Stdrd. Freq.: A by B			w/in vars: C:2
	B1	B2	B3	Total
1	,195688	-,102015	-,558192	-,464519
2	-,587220	,653617	,645497	,711894
Total	-,391533	,551603	,087305	,247375

A	Stdrd. Freq.: A by B			w/in vars:C:3
	B1	B2	B3	Total
1	-,821548	1,130858	,261893	,571203
2	,358208	-,298242	-,691419	-,631453
Total	-,463340	,832616	-,429526	-,060250

Машинограмма 9.15

Таблицы частот наблюдений для сочетаний уровней факторов, включенных в модель

C	Marg. Tabl. (freq+delta): A by C		
	A1	A2	Total
1	25,30000	5,30000	30,6000
2	14,30000	39,30000	53,6000
3	1,30000	19,30000	20,6000
Total	40,90000	63,90000	104,8000

	Marg. Tabl. (freq+delta): B			
	B1	B2	B3	Total
Freqncs.	69,.60000	31,60000	3,60000	104,8000

Литература

1. Алтон Г. Анализ таблиц сопряженности: Пер. с англ. - М.: Финансы и статистика, 1982. - 143 с.

2. Елисеева И.И., Рукавишников В.О. Логика прикладного статистического анализа. - М.: Финансы и статистика, 1982 - 192 с.

Глава 10. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Назначение и содержание логистической регрессии

Построение моделей для показателей состояния многомерных медицинских систем в зависимости от воздействующих на них факторов является важной задачей статистического анализа, выполняемого исследователями с применением современной информационной технологии.

По модели решают основные задачи исследования, среди которых: изучение характера изменения показателя при изменении действующих на систему факторов; оценка степени влияния факторов на величину показателя-отклика; прогнозирование показателя-отклика для заданных уровней факторов; определение оптимальных уровней факторов для получения требуемых или желаемых значений показателей состояния системы. Построение таких моделей проводится на основе выборочного наблюдения, по результатам которого формируется исходная база данных, представляющая собой матрицу наблюдений с числом строк, равным числу наблюдавшихся объектов и числом столбцов, равным числу контролируемых факторов и моделируемого показателя-отклика на воздействующие факторы.

В условиях количественного определения факторов и показателя-отклика для построения модели показателя – уравнения регрессии – применяют многомерный регрессионный анализ. Коэффициенты модели при этом определяются методом наименьших квадратов. В основу метода наименьших квадратов заложен принцип минимизации суммы квадратов отклонений прогнозируемых значений показателя по модели от наблюдавшихся значений в выборке.

В условиях наблюдения качественных оценок показателя – отклика всего на двух уровнях, например, выживание больного при тяжелой травме (код 1) и летальный исход (код 0), для построения модели вероятности благоприятного исхода применяют логистическую регрессию, представляющую собой нелинейную функцию распределения вероятностей. Коэффициенты уравнения логистической регрессии определяются методом максимального правдоподобия, в основу которого положен принцип максимизации вероятности соответствия, адекватности прогнозируемых по моделям уровней показателя–отклика с наблюдавшимися значениями этого показателя в выборке.

Для случая, когда исследуется некоторый положительный эффект при воздействии на объект только одного фактора, например, в эксперименте «доза-эффект», уравнение логистической регрессии имеет вид:

$$\hat{y} = \frac{\exp(b_0 + bx)}{1 + \exp(b_0 + bx)}, \qquad (10.1)$$

где $\hat{y}$ - вероятность положительного эффекта ($0 \le \hat{y} \le 1$);

b_0 – константа;

b – коэффициент фактора X;

x – текущее значение фактора X.

Кривая функции логистической регрессии, соответствующая уравнению (10.1), показана на рисунке 10.1. Она представляет собой кривую, непрерывно возрастающую от 0 до 1.

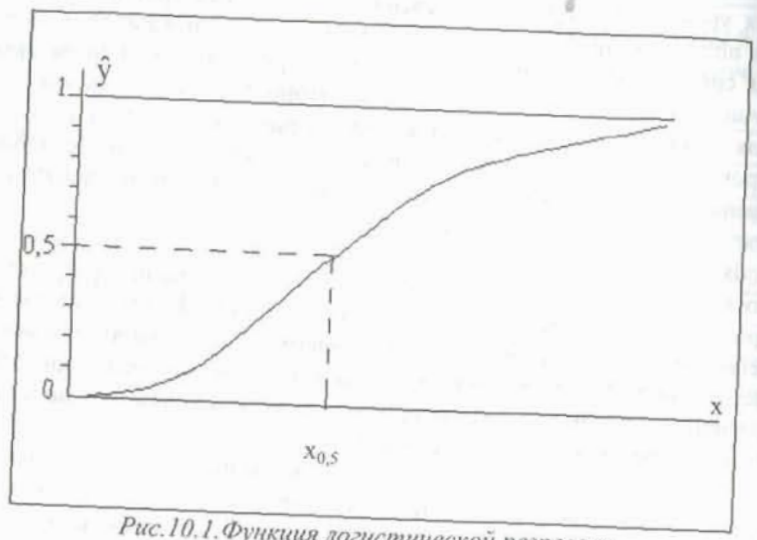

Рис.10.1.Функция логистической регрессии.

При малых дозах фактора X вероятность положительного эффекта незначительна; при больших дозах, например, при $x > x_{0,5}$ вероятность положительного эффекта возрастает от 0,5 до 1.

При прогнозе вероятности положительного эффекта по (10.1) принимают:

-положительный эффект при $\hat{y} > 0,5$;

-отрицательный результат при $\hat{y} \le 0,5$.

Для случая, когда исследуется положительный эффект при воздействии на объект множества контролируемых факторов X1, X2, … Xk, уравнение логистической регрессии принимает вид:

$$\hat{y} = \frac{\exp(b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k)}{1 + \exp(b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k)}, \qquad (10.2)$$

где $\hat{y}$ - вероятность положительного эффекта ($0 \le \hat{y} \le 1$);

b_0 – константа;

$b_1, b_2, \ldots b_k$ – коэффициенты X1, X2, … Xk факторов;

$x_1, x_2, \ldots x_k$ – текущие значения k факторов.

Из (10.2) следует, что при стремлении величины показателя экспоненты $b_0+b_1x_1+b_2x_2+ \ldots + b_kx_k$ к очень малому значению (в пределе - ∞) вероятность положительного эффекта стремиться к 0, и наоборот, при стремлении $b_0+b_1x_1+b_2x_2 + \ldots +b_kx_k$ к очень большому значению (в пределе + ∞) вероятность положительного эффекта стремиться к 1. Характер изменения показателя $\hat{y}$ аналогичен показанному на рис.10.1. Прогноз, положительного эффекта дается при $\hat{y} > 0,5$, отрицательного при $\hat{y} \le 0,5$.

Для определения модели (10.2) по модулю нелинейного моделирования ППП Statistica или по модулю регрессия ППП SPSS формируется обучающая исходная матрица наблюдений размером $n \times (k+1)$, где n – число наблюдавшихся объектов; k – число контролируемых факторов и 1 – показатель-отклик, выражаемый кодами: 1 – при положительном эффекте и 0 – при отрицательном исходе.

Результат решения на ПК включает:

– таблицу коэффициентов модели (10.2) с оценками их значимости;

– таблицы прогнозируемых по модели значений показателя $\hat{y}$, наблюдавшихся значений y и разностей $\hat{y}$ -y для всех объектов обучающей матрицы наблюдений;

– классификационную матрицу, характеризующую абсолютные величины и относительные частоты правильных прогнозов положительных и отрицательных эффектов по модели;

– уровень значимости модели по критерию хи-квадрат;

– оценку соответствия распределения остатков нормальному закону.

Модель признается значимой при уровне значимости $p \le 0,05$ (достоверности $1-p \ge 0,95$).

Следует отметить, что в процедуре Logistic Regression ППП SPSS возможен пошаговый отбор значимых коэффициентов для включения в модель 10.2. С этой целью задается критическое значение критерия Фишера $F \ge 4$, обеспечивающее уровень значимости коэффициентов $p \le 0,05$.

Примеры применения логистической регрессии:

– в экспериментах «доза-эффект»;

– в медицинской диагностике (есть заболевание или нет заболевания у обследуемого);

– в профотборе (пригодность или непригодность кандидата);

– при оценке исхода лечения больного (выжил, умер; без осложнения, с осложнением и др.).

ПРИМЕР 10.1

Прогнозирование ранних исходов тяжелой черепно-мозговой травмы (РИ ТЧМТ) (по данным Н.Б. Клименко).

Основой обучающей информации для создания логистической регрессионной модели РИ ТЧМТ по признаку выписан из стационара - умер в стационаре стали истории болезней 300 пострадавших с ТЧМТ, поступивших на стационарное лечение в клинику Российского научно-исследовательского нейрохирургического института им.профессора А.Л.Поленова и прошедших лечение до определившегося исхода.

Основной задачей моделирования является прогноз РИ ТЧМТ по данным первичного врачебного осмотра пострадавшего в приемном отделении стационара или в любых других условиях, например, на месте происшествия на автодороге, на производстве, в боевых условиях и т.п. По сути, такая модель является экспресс-прогнозом, т.к. строится на основании минимально достаточного числа наиболее простых и всегда исследуемых неврологических симптомов и синдромов, не требующих высокой квалификации врачебного персонала и не включающих применения специальных дополнительных инструментальных методов исследования.

В качестве прогнозируемого показателя-отклика определен исход травмы (благоприятный - больной выписан из стационара - 1 и неблагоприятный - больной умер в стационаре - 0).

В качестве признаков, предшествующих исходу травмы, и включаемых в модель как независимые факторы-причины, определена совокупность клинических признаков, достоверно связанных с исходами и определяемых у больных на ранних этапах оказания медицинской помощи. В исходную обучающую матрицу было включено 59 признаков, получаемых анамнестически и с помощью непосредственного врачебного обследования и регистрируемых в приемном отделении. После логического анализа и оценки связей исходных данных с помощью корреляционного анализа для дальнейшего исследования в обучающей матрице осталось 25 признаков, которые имели сильную ($r>0,7$) или умеренную ($0,27<r<0,7$) и статистически значимую ($p<0,05$) корреляционную связь с РИ ТЧМТ.

Решение задачи логистического регрессионного анализа может быть реализовано с помощью процедуры Logistic Regression из пакетов прикладных программ по статистической обработке данных Statistica или SPSS. По нашему мнению, преимущество следует отдать ППП SPSS, так как он обеспечивает пошаговый отбор в модель статистически значимых факторов с заданным порогом значимости.

По итогам расчетов с помощью модуля Logistic Regression ППП SPSS, в модель включены 5 признаков, обладающих статистической надежностью не менее 80%. Перечень этих признаков и их коэффициенты приведены в табл.10.1.

Полученная методом логистического регрессионного анализа статистически значимая ($p<0,0001$) модель, имеет вид:

$$\hat{y} = \exp(7,54-0,59x_1-0,26x_2-0,79x_3+0,56x_4-1,12x_5)/ \quad (10.3)$$
$$(1 + \exp(7,54-0,59x_1-0,26x_2-0,79x_3+0,56x_4-1,12x_5)).$$

Расчеты прогноза по этой модели могут быть произведены на ПЭВМ или на программируемом микрокалькуляторе.

Любая синтезированная модель, логистическая в том числе, требует подтверждения - на сколько она соответствует наблюдавшимся данным, С этой целью используются:

– классификация данных обучающей информации с помощью полученной модели и оценка соответствия этой классификации с наблюдаемой в опыте (табл.10.2);

– сравнение опытных и прогнозируемых значений для каждого конкретного наблюдения (табл.10.3);

– оценка остатков (разности наблюдаемых величин и величин прогнозируемых с помощью модели) (рис.10.2-10.3).

Таблица 10.1

Признаки, включенные в логистическую регрессионную модель прогноза РИ ТЧМТ

№ пп	Наименования и градации симптомов	Коды	Коэффициенты модели	Уровень значимости, p
1	Возраст: 15-24 года – 1, 25-34 года – 2, 35-44 года – 3, 45-54 года – 4, 55-64 года – 5, 65-74 года – 6, 75 лет и старше – 7.	X1	-0,59	0,000
2	Систолическое АД: 110-140 мм рт.ст. – 1, 141-180 мм рт.ст. – 2, 90-109 мм рт.ст. – 3, 181мм рт.ст. и более – 4, 89 мм рт.ст. и менее – 5.	X2	-0,26	0,059
3	Уровень сознания: ясное – 1, легкое оглушение – 2, умеренное оглушение – 3, сопор – 4, кома I – 5, кома II – 6, кома III – 7.	X3	-0,79	0,000
4	Окулоцефалический рефлекс: отсутствует – 1, вызывается – 2.	X4	0,56	0,179
5	Иннервация зрачков: не нарушена – 1, нарушена – 2.	X5	-1,12	0,072
	Константа.		7,54	0,000

Результаты классификации исходов, полученные с помощью логистической регрессионной модели по данным обучающей информации, и их сравнение с классификацией, наблюдавшейся в опыте, приведены в табл.10.2, из которой следует, что в группе больных с благоприятным исходом исследуемая модель обеспечивает совпадение прогнозируемого результата с реальным в 83,2% случаев (у 134 из 161 выписанного больного); в группе умерших больных совпадение прогнозируемых и реальных исходов лечения отмечено в 77,0% наблюдений (в 107 из 139 случаев); информационная способность модели в целом составляет 80,3% (совпадение исходов лечения у 241 из 300 больных).

Таблица 10.2

Классификация пострадавших с ТЧМТ по признаку выжил - умер с помощью логистической регрессионной модели в сравнении с наблюдавшейся в опыте

	%	Исход лечения - благоприятный	Исход лечения - летальный	Всего в опыте
Исход лечения - благоприятный	83,2	134	27	161
Исход лечения - летальный	77,0	32	107	139
Всего в прогнозе	80,3	166	134	300

По строкам: классификация соответственно базе данных.
По столбцам: классификация соответственно прогнозу.

Из таблицы 10.3, в которой представлены наблюдаемые и прогнозируемые значения результирующего признака и разность между ними для конкретных наблюдений, видно, что прогнозируемая вероятность благоприятного исхода у больных выписанных из стационара преимущественно (83,2%) больше 0,5, хотя и встречаются значения меньше 0,5 (16,8%). В группе больных с неблагоприятным исходом - прогнозируемая вероятность благоприятного исхода в 77% случаев меньше 0,5, а в 23% случаев больше 0,5. Данные представленные в описываемой таблице перекликаются с данными таблицы 10.2, но дают возможность оценить прогноз для каждого конкретного случая наблюдения.

Важное значение в оценке адекватности модели данным исследования имеет характеристика остатков и в частности соответствие их распределения нормальному закону. На рис.10.2 представлена гистограмма остатков с наложенной плотностью нормального распределения, которая свидетельствует о достаточно близком распределении остатков к нормальному, а соответственно и о том, что модель неплохо описывает наши данные. К подобному выводу приводит анализ графика на рис.10.3. Когда распределение остатков точно соответствует нормальному закону, то они хорошо ложатся на прямую, что мы и наблюдаем в нашем случае.

Таблица 10.3

Наблюдаемые и прогнозируемые значения результирующего признака и разность между ними для 11-и конкретных наблюдений

	Наблюдаемые значения	Прогнозируемые значения	Остатки
Наблюдение 1	1	0,884	0,116
Наблюдение 7	1	0,355	0,645
Наблюдение 8	1	0,644	0,356
Наблюдение 12	1	0,515	0,485
Наблюдение 14	1	0,183	0,817
Наблюдение 16	0	0,191	-0,191
Наблюдение 17	0	0,037	-0,037
Наблюдение 18	0	0,548	-0,548
Наблюдение 23	0	0,288	-0,288
Наблюдение 25	0	0,035	-0,035
Наблюдение 26	0	0,629	-0,629

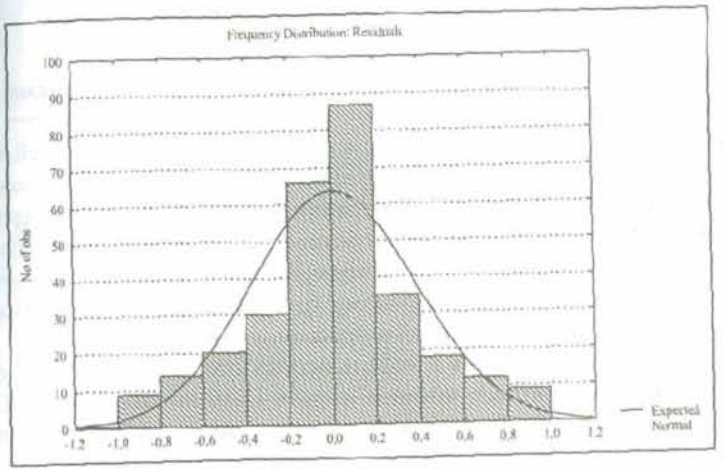

Рис.10.2.Гистограмма остатков с наложенной плотностью нормального распределения.

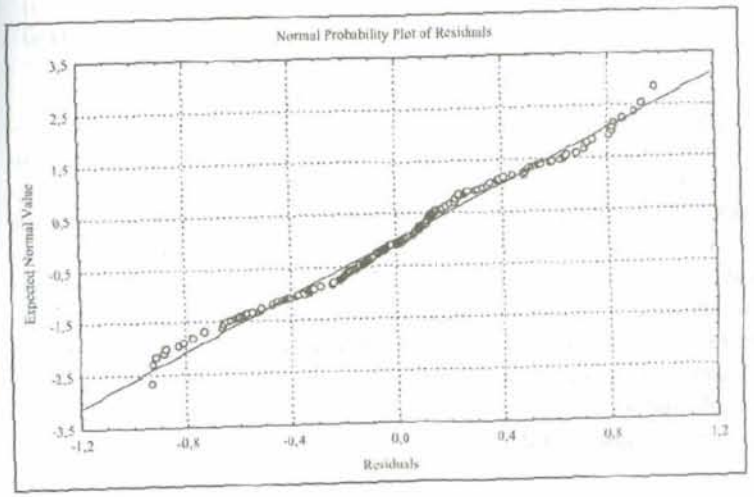

Рис.10.3.График остатков на нормальной вероятностной бумаге.

ПРИМЕР 10.2

Модель ранних исходов тяжелой черепно-мозговой травмы (РИ ТЧМТ), полученная с помощью дискриминантного анализа.

Проблемам и методологии создания математических моделей классификации на основе линейных дискриминантных функций посвящена глава 7. Мы обращаемся к демонстрации этого метода с целью сравнения возможностей двух методов построения информационно способных моделей, с одной стороны, и верификации относительно мало известного, по крайней мере широкому кругу медицинских научных работников, метода логистической регрессии, с другой стороны.

Для селекции признаков в модель использовалась процедура пошагового дискриминантного анализа модуля Discriminant Analysis ППП Statistica for Windows. В результате получены две линейные классификационные функции.

Линейные классификационные функции благоприятного исхода (ЛКФ1) и неблагоприятного исхода (ЛКФ2) рассчитываются по формулам:

$$ЛКФ1=-51{,}41+1{,}90x_1-0{,}19x_2+7{,}53x_3+23{,}97x_4+13{,}63x_5;$$
$$ЛКФ2=-58{,}27+2{,}51x_1+0{,}10x_2+8{,}40x_3+23{,}05x_4+14{,}41x_5. \qquad (10.4)$$

Модель построенная на основе 5 клинических признаков, выявляемых во время первичного обследования пострадавшего и представленных в табл.10.1, является статистически значимой ($p<0{,}00001$) и обладает достаточно высокой прогностической способностью. Отнесение больных к двум группам по исходу лечения выполняется по максимальному значению ЛКФ. Так, если ЛКФ1>ЛКФ2 - больного следует отнести к группе благоприятного исхода; если ЛКФ1≤ЛКФ2 - к группе неблагоприятного исхода. Классификационная матрица, в которой представлены результаты классификации пострадавших по данным обучающей информации и соотношение этой классификации с наблюдавшимися исходами ТЧМТ, представлены в таблице 10.4.

Таблица 10.4

Классификационная матрица по модели ЛДФ

	%	Исход лечения – благоприятный	Исход лечения – летальный	Всего
Исход лечения – благоприятный	82,61	133	28	161
Исход лечения – летальный	76,98	32	107	139
Всего:	80,00	165	135	300

По строкам: классификация соответственно базе данных.
По столбцам: классификация соответственно прогнозу.

Из таблицы следует, что в группе выписанных больных (благоприятный исход ТЧМТ) предлагаемая модель обеспечивает совпадение прогнозируемого исхода с реальным результатом в 82,61% случаев (совпадение результатов лечения у 133 из 161 реально выписанного больного); в группе умерших больных (летальный исход) - совпадение прогнозируемого исхода с реальными результатами составило 76,98% (совпадение результатов у 107 из 139 умерших больных).

Таким образом, дискриминантная модель прогноза РИ ТЧМТ по данным первичного осмотра, основанная на 5 простейших клинических признаках (возраст, систолическое артериальное давление, уровень сознания, окулоцефалический рефлекс, иннервация зрачков), обладает достаточно высокой информационной способностью (80,00%) и является статистически значимой ($p<0{,}0001$).

С целью дополнительной оценки качества полученных моделей воспользуемся такими ее характеристиками как чувствительность, специфичность, безошибочность, ложноотрицательный ответ и ложноположительный ответ. Расчет этих характеристик производится по формулам, изложенным в главе 7.

Таблица 10.5

Оценка качества классификаций

Характеристики модели	ЛРМ	ЛДФ
Чувствительность	83,2	82,6
Специфичность	77,0	77,0
Безошибочность	80,3	80,0
Ложноотрицательный ответ	16,8	17,4
Ложноположительный ответ	23,0	23,0

Из данных таблицы 10.5 следует, что применяя дискриминантный и логистический регрессионный анализы нами получены практически одинаковой информационной способности и статистической значимости математические модели прогноза РИ ТЧМТ, обладающие достаточной чувствительностью (≈83%), специфичностью (77%) и безошибочностью (≈80%) с допустимой долей ложноотрицательных (≈17%) и ложноположительных (≈23%) ответов. Это свидетельствует о доброт-ности данных, отобранных в матрицу обучающей информации и адекватности используемых методов моделирования данным исследования.

Литература

1. Боровиков В.П., Боровиков И.П. Statistica. Статистический анализ и обработка данных в среде Windows. – М.: Инф. издат. дом «Филинъ», 1997. - 608 с.
2. Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р, Олдендерфер М.С, Блэшфилд Р.К. Факторный, дискриминантный и кластерный анализ: Пер. с англ. Под ред. И.С. Енюкова.- М.: Финансы и статистика, 1989.
3. Лядов В.Р. Основы теории вероятностей и математической статистики: Для студентов мед.ВУЗов. - СПб.: Фонд «Инициатива», 1998. - 107 с.
4. Математико-статистические методы в клинической практике /Под ред. В.И.Кувакина. - СПб.: Б.и, 1993. - 199 с.
5. Юнкеров В.И. Основы математико-статистического моделирования и применения вычислительной техники в научных исследованиях: Лекции для адъюнктов и аспирантов / Под ред. В.И.Кувакина. – СПб, 2000. –140 с.

Глава 11. АНАЛИЗ ДАННЫХ ВРЕМЕНИ ЖИЗНИ

Назначение и содержание анализа данных времени жизни

В статистическом смысле понятие «время жизни» распространяется на любые данные, описывающие длительность пребывания объектов в интересующем исследователя состоянии. Примерами могут служить продолжительность инкубационного периода заболевания, лечения, жизни больных после лечения и другие явления, анализ которых позволяет расширить представления о закономерностях этиологии, патогенеза, клиники заболеваний, а также оценить эффективность лечебных и профилактических мероприятий.

Методы анализа данных времени жизни особенно хорошо отражают специфику информационных материалов в онкологии, радиологии, геронтологии, фармакологии и в целом ряде других областей медицины. Они абсолютно незаменимы при оценке результатов клинических испытаний.

Данные времени жизни имеют две характерные особенности, которые предопределяют специфику их анализа. Прежде всего возможна неполнота данных. Например, в клинических исследованиях больные по тем или иным причинам «уходят» из-под наблюдения, часть лабораторных животных может регулярно забиваться для проведения анализов. Реальное же время жизни таких объектов больше длительности наблюдения за ними. Описанный феномен в статистике называется цензурированием справа. Наличие цензурированных данных затрудняет оценку эффекта изучаемого воздействия на время жизни. Особенно эта трудность обнаруживает себя при характеристике отдаленных результатов лечения. Другая особенность данных времени жизни - неадекватность распределения времени жизни статистической модели нормального закона распределения. Конкретный же вид распределения, как правило, неизвестен. Поэтому аппроксимация распределения времени жизни нормальному закону, явная или неявная (при использовании параметрических методов анализа), представляет угрозу для корректности статистических выводов.

Алгоритм анализа данных времени жизни, адекватный их специфике, составляет определенную последовательность действий:

– анализ времени жизни в одной группе;

– сравнение времени жизни в двух или более группах;

– оценку влияния экзогенных или эндогенных факторов на время жизни объектов.

Анализ времени жизни в одной группе. По образному выражению авторов одной из биологических работ, «развязку узла проблем продолжительности жизни» необходимо начинать с анализа индивидуальных различий по срокам жизни. Характеристика вида и параметров кривой распределения времени жизни основывается на результатах выборочного наблюдения.

Данные о времени жизни n объектов обычно представлены набором соответствующих дат, ($t_1 < t_2 \ldots t_i \ldots t_N$), рассматриваемых как случайные величины. Для описания распределения данных времени жизни используются различные статистические функции.

Функция распределения случайной величины, называемая также кумулятивной функцией риска вероятности смерти, отражает вероятность того, что организм проживет время, меньшее чем t:

$$F(t) = P\ \{T<t\}. \qquad (11.1)$$

Вероятность противоположного события, т.е. что организм проживет время, большее, чем t, называется функцией дожития (выживания):

$$S(t) = 1 - F(t) = P\ \{T>t\}. \qquad (11.2)$$

Функция плотности вероятности описывает кривую распределения организмов по срокам жизни:

$$f(t) = dF(t) / dt. \qquad (11.3)$$

Интенсивность смертности (функция риска) характеризует риск смерти в момент t:

$$h(t) = -dS(t) / S(t)dt. \qquad (11.4)$$

В медико-биологических исследованиях наибольшей популярностью пользуется представление распределения данных времени жизни в виде функции дожития. Очевидной причиной такого предпочтения является интуитивная однозначность восприятия вероятности дожития до определенного момента времени, основанная на ментальном отношении к ценности человеческой жизни.

С точки зрения теории, венцом первого этапа анализа данных времени жизни должна быть аппроксимация эмпирического ряда времени жизни одним из известных теоретических законов распределения. В качестве таких распределений могут быть использованы распределения, сосредоточенные на положительной полуоси. Типичными распределениями для описания данных времени жизни являются экспоненциальное, гамма, Вейбулла, Гомперца-Мейкема, логарифмически нормальное и др.

Сравнение времени жизни в двух и более группах. После оценки статистических характеристик времени жизни в единичной группе закономерна постановка вопроса о сравнении времени жизни в двух или более группах. Такое сравнение может быть выполнено либо сопоставлением отдельных параметров распределений, полученных при аппроксимации эмпирических распределений теоретическими, либо сравнением функций дожития в различных группах. При сравнении двух выборок с помощью статистических критериев проверяется гипотеза о равенстве двух распределений времени жизни, называемая нулевой гипотезой:

$$H_o : F_1(t) = F_2(t). \qquad (11.5)$$

Поскольку вид распределения времени жизни в сравниваемых выборках, как правило, не соответствует нормальному закону, наиболее корректным оказывается использование непараметрических критериев. Распределение в нескольких (3 и более) группах оценивается по критерию χ^2 Пирсона, в двух группах - по другим непараметрическим критериям, семейство которых обширно и вполне заслуживает отдельного рассмотрения. Поэтому здесь ограничимся лишь комментариями относительно природы тестов, рассчитываемых в частности в ППП Statistica for Windows.

Тест Гехана-Вилкоксона (Gehan's Wilcoxon test), наиболее часто используемый, является модификацией непараметрического критерия Гехана-Вилкоксона. Его рекомендуют применять в тех случаях, когда различия в кривых дожития наиболее выражены в начальный период наблюдения, а также в случаях, когда не выполняется модель пропорциональных рисков, то есть отношение интенсивностей «отказов» в сравниваемых группах не остается постоянным в течение всего периода наблюдения: $h_1(t) / h_2(t) \neq const$. F-test позволяет осуществить проверку гипотезы о равенстве параметра интенсивности отказа в экспоненциально распределенных выборках. Мощность критерия Кокса и логрангового критерия (Cox's и Log-rank test) максимальна при выполнении модели пропорциональных рисков. Критерий Кокса особенно чувствителен к различиям в кривых выживания, обнаруживающимся на концах распределений. Это свойство может быть полезным при изучении отдаленных эффектов лечения. Логранговый критерий рекомендуется применять, когда наблюдаемое число смертей мало.

При интерпретации результатов следует иметь в виду, что, если рассчитанное значение критерия превышает критическое, гипотеза об *отсутствии различия между двумя функциями дожития отклоняется.*

Оценка влияния экзогенных или эндогенных факторов на время жизни объектов. Даже если различия времени жизни в двух или более группах выявлены, остается открытым вопрос, какие факторы и в какой степени влияют на время жизни. Решение этой задачи возможно путем построения регрессионных моделей, имеющих для данных времени жизни свою специфику.

Регрессионная модель времени жизни может быть построена исходя из предположения об экспоненциальном, нормальном или логнормальном распределении времени жизни. Несколько в стороне стоит модель пропорциональных интенсивностей, или пропорциональных рисков (Кокса).

Различие между названными моделями, вероятно, лучше всего может быть проиллюстрировано поведением функции интенсивности смерти объекта в момент времени t.

Экспоненциальная регрессия обосновывается предположением постоянства интенсивности смерти во времени. Уравнение имеет вид

$$\hat{y} = b \times \exp(ht). \qquad (11.6)$$

Нормальное распределение времени жизни (феномен достаточно редкий) имеет возрастающую интенсивность смерти. В определенных границах приближение к нормальному распределению может оказаться вполне приемлемым для реализации классической схемы регрессионного анализа.

Интенсивность смерти логарифмически нормального распределения сначала возрастает, а затем падает до нуля. Можно предположить, что эта модель будет адекватна, в частности, для описания времени наступления летальных исходов после травмы. Действительно, если по истечении критического периода пострадавший остается живым, то его шансы умереть (от травмы) с течением времени будут только убывать.

Если основная задача исследования состоит в изучении качественного влияния воздействующих факторов на время жизни, то выбор модели не имеет решающего значения. В случаях, когда задача связана с относительно «тонкими» вопросами зависимости времени жизни от воздействующих факторов, требуется выбор адекватной модели, для чего необходимо понимание сущности альтернативных моделей.

Рассмотрим уравнение регрессии для модели Кокса (*пропорциональных рисков*) как *наиболее специфичной* в методах регрессионного анализа данных времени жизни. Модель пропорциональных рисков задается через функцию интенсивности.

Основная идея модели состоит в том, что влияние фактора(ов) на интенсивность смерти соответствует умножению интенсивности смерти, существующей при стандартных условиях, на множитель, постоянный для всех t. Таким образом, задача построения регрессионной модели сводится к выбору вида зависимости указанного множителя от экзогенных или эндогенных факторов.

Наиболее важным частным случаем модели Кокса является предположение, что интенсивность смерти определяется соотношением:

$$h(t;X) = h_0(t)e^{\beta X}, \qquad (11.7)$$

где $h(t;X)$ - функции интенсивности смерти под влиянием фактора X,

$h_0(t)$ - функция интенсивности смерти при стандартных условиях,

β - коэффициент, подлежащий определению.

Нулевая гипотеза ($H_0 : \beta = 0$) означает, что множитель $\exp(\beta X)$ принимает значение, равное 1, что возможно только при отсутствии влияния фактора(ов) на интенсивность смерти.

Рассмотренные направления анализа данных времени жизни в целом образуют цепочку: описательные - аналитические - экспериментальные методы. Основной целью этой цепочки является выяснение механизмов, определяющих время жизни. После применения любого вида анализа данных остается вопрос о степени достижения цели. Ответ на него состоит из двух слагаемых: соответствия полученных моделей известным фактам и способность предсказывать новые закономерности.

ПРИМЕР 11.1

Построение модели продолжительности ремиссии при лечении хронического алкоголизма.

Для изучения связи показателя устойчивости результатов лечения с факторами, характеризующими социально-экономические и бытовые условия, состояние здоровья и методы лечения, Международным институтом резервных возможностей человека (МИРВЧ) проведено обследование 556 пациентов, лечившихся по поводу хронического алкоголизма и наблюдавшихся в течение трех лет. В результате анкетиро-

вания сформирована база данных - матрица наблюдений, в которой каждый пациент охарактеризован 16 признаками (табл.11.1). Исходная матрица наблюдений (файл mirvc_91.sta.) не приводится ввиду ее больших размеров (556×17). В ней исследуемый показатель - продолжительность ремиссии (REM_CENZ) указан в днях, остальные признаки определены кодами по порядковой шкале (1, 2, 3 и т.д.). Принятая система кодирования уровней признаков позволяет применять их как *группировочные при сравнении продолжительности ремиссии* в различных группах и как независимые факторы, влияющие на продолжительность ремиссии при ее моделировании.

Таблица 11.1

Кодировка признаков, учтенных в базе данных, и их описание

№ пп	Код признака	Характеристика признака
1	REM_CENZ	Ремиссия, в днях.
2	CENZOR	Факт цензурирования.
3	POL1	Пол:
		мужчина - 1;
		женщина - 2.
4	ZA	Закрепление зарока:
		было – 1;
		не было – 2.
5	KOD_L_R	Лечение больного в МИРВЧ до данного сеанса:
		было – 1;
		не было – 2.
6	ERSOV2	Шкала осознанности лечебной ситуации (анозогнозия, самооценка):
		высокая – 1;
		средняя – 2;
		низкая – 3.
7	LEC_DO_M	Лечение до обращения в МИРВЧ:
		в госнаркослужбе – 1;
		нигде не лечился – 2;
		в госнаркослужбе и кооперативе – 3;
		в кооперативе – 4.
8	KOD_VOZ	Код возраста:
		до 30 лет включительно – 1;
		более 30 лет – 2.

№ пп	Код признака	Характеристика признака
9	STADIJA	Стадия заболевания:
		первая – 1;
		вторая – 2;
		третья – 3.
10	KL_TIP	Клинический тип:
		экзогенно-конституциональный – 1;
		на патологической почве – 2;
		привитой – 3;
		смешанный – 4;
		наследственно-конституциональный – 5.
11	KOD_SROK	Код выбранного срока лечебного зарока:
		до 5 лет – 1;
		5 лет и более – 2.
12	KOD_ALK	Код срока трезвости до лечения:
		до 2 недель – 1;
		более 2 недель – 2.
13	OBRAZ	Образование:
		начальное – 1;
		среднее – 2;
		высшее – 3.
14	SEMJA	Семейное положение:
		семейный – 1;
		разведенный – 2;
		холост – 3.
15	UST_TREZ	Установка на трезвость:
		есть – 1;
		нет – 2.
16	PLATA_LC	Плата за лечение:
		своими деньгами – 1;
		деньгами из других источников – 2.

Отождествив продолжительность ремиссии с временем жизни, а срыв ремиссии - с ее завершением (смертью), для исследования применим эффективный метод - анализ данных времен *жизни* (Survival Analysis) ППП Statistica 5.0 for Windows.

Из большого количества возможных решений по модулю анализа данных времени жизни выберем выполнение трех задач исследования:

– статистическое описание продолжительности ремиссии;

– сравнение продолжительности ремиссии в различных группах;

– построение моделей функций распределения продолжительности ремиссии - в зависимости от факторов, влияющих на сохранение состояния ремиссии.

Решение:

Статистическое описание продолжительности ремиссии. Описание исследуемого показателя - продолжительности ремиссии основными функциями распределения вероятностей дано в таблице Life Table (машинограмма 11.1). Таблица продолжительности ремиссии содержит следующие функции, характеризующие динамику срыва и сохранения состояния ремиссии:

– частости срыва и сохранения состояния ремиссии;

– функции плотности вероятности и интенсивности срыва ремиссии;

– функцию сохранения состояния ремиссии.

Машинограмма 11.1

Таблица жизни

Life Table (mirvc_91.sta)
Log-Likelihood for data: -1612,26

Номер интервала	Нижняя граница интервала	Средняя точка интервала	Ширина интервала	Число наблюдений в начале интервала	Количество цензурированных случаев
	Interval Start	Mid Point	Interval Width	Number Entering	Number Withdrwn
Intno.1	0	45,63	91,25	556	32
Intno.2	91,25	136,9	91,25	442	11
Intno.3	182,50	228,1	91,25	389	9
Intno.4	273,75	319,4	91,25	344	10
Intno.5	365,00	410,6	91,25	317	36
Intno.6	456,25	501,9	91,25	265	5
Intno.7	547,50	593,1	91,25	253	4

Номер интервала	Нижняя граница интервала	Средняя точка интервала	Ширина интервала	Число наблюдений в начале интервала	Количество цензурированных случаев
	Interval Start	Mid Point	Interval Width	Number Entering	Number Withdrwn
Intno.8	638,75	684,4	91,25	245	0
Intno.9	730,00	775,6	91,25	234	0
Intno.10	821,25	866,9	91,25	224	2
Intno.11	912,50	958,1	91,25	215	1
Intno.12	1003,8	1049	91,25	206	0
Intno.13	1095,0	--	--	203	203

Продолжение машинограммы 11.1

Номер интервала	Среднее число наблюдений на интервале	Количество случаев срыва ремисии на интервале	Частость срыва ремисии на интервале	Частость сохранения состояния ремиссии на интервале
	Number Exposed	Number Dying	Proportn Dead	Proportn Survivng
Intno.1	540	82	0,1519	0,8481
Intno.2	436,5	42	0,0962	0,9038
Intno.3	384,5	36	0,0936	0,9064
Intno.4	339	17	0,0501	0,9499
Intno.5	299	16	0,0535	0,9465
Intno.6	262,5	7	0,0267	0,9733
Intno.7	251	4	0,0159	0,9841
Intno.8	245	11	0,0449	0,9551
Intno.9	234	10	0,0427	0,9573
Intno.10	223	7	0,0314	0,9686
Intno.11	214,5	8	0,0373	0,9627
Intno.12	206	3	0,0146	0,9854
Intno.13	101,5	0	0,0049	0,9951

Продолжение машинограммы 11.1

Номер интервала	Функция сохранения состояния ремиссии	Плотность вероятности срыва ремиссии на интервале	Интенсивность срыва ремиссии на интервале	Стандартная ошибка функции сохранения состояния ремиссии
	Cum.Prop Survivng	Problty Density	Hazard Rate	Std.Err. Cum.Surv
Intno.1	1	0,002	0,0018	0
Intno.2	0,8481	9E-04	0,0011	0,0154
Intno.3	0,7665	8E-04	0,0011	0,0184
Intno.4	0,6948	4E-04	0,0006	0,0202
Intno.5	0,6599	4E-04	0,0006	0,0209
Intno.6	0,6246	2E-04	0,0003	0,0215
Intno.7	0,6080	1E-04	0,0002	0,0219
Intno.8	0,5983	3E-04	0,0005	0,0220
Intno.9	0,5714	3E-04	0,0005	0,0225
Intno.10	0,5470	2E-04	0,0003	0,0228
Intno.11	0,5298	2E-04	0,0004	0,0230
Intno.12	0,5101	8E-05	0,0002	0,0232
Intno.13	0,5026	--	--	0,0232

Продолжение машинограммы 11.1

Номер интервала	Стандартная ошибка плотности вероятности срыва ремиссии на интервале	Стандартная ошибка интенсивности срыва ремиссии на интервале	Медиана продолжительности состояния ремиссии для наблюдений данного интервала
	Std.Err. Prob.Den	Std.Err. Haz.Rate	Median Life Exp
Intno.1	0,0002	2E-04	1095
Intno.2	0,0001	2E-04	1003,8
Intno.3	0,0001	2E-04	912,5

Номер интервала	Стандартная ошибка плотности вероятности срыва ремиссии на интервале	Стандартная ошибка интенсивности срыва ремиссии на интервале	Медиана продолжительности состояния ремиссии для наблюдений данного интервала
	Std.Err. Prob.Den	Std.Err. Haz.Rate	Median Life Exp
Intno.4	9E-05	1E-04	821,25
Intno.5	9E-05	2E-04	730
Intno.6	7E-05	1E-04	638,75
Intno.7	5E-05	9E-05	547,5
Intno.8	9E-05	2E-04	456,25
Intno.9	8E-05	2E-04	365
Intno.10	7E-05	1E-04	273,75
Intno.11	8E-05	1E-04	182,5
Intno.12	5E-05	9E-05	91,25
Intno.13	--	--	--

Частость срыва или сохранения состояния ремиссии определяется для каждого интервала как отношение количества случаев срыва (сохранения) ремиссии к среднему числу наблюдений на интервале с учетом выбывших из поля наблюдения (цензурированных) по разным причинам.

Функции плотности вероятности и интенсивности срыва ремиссии являются производными от функций распределения вероятностей срыва и сохранения состояния ремиссии в различные моменты времени.

Так, функция плотности вероятности срыва ремиссии есть производная (11.3): $f(t) = \frac{dF(t)}{dt}$, где F(t) - функция распределения времени срыва ремиссии, равная вероятности того, что срыв ремиссии произойдет за время меньше t (11.1).

Противоположной F(t) является функция распределения времени сохранения ремиссии $S(t) = 1 - F(t) = P(T > t)$, равная вероятности того, что состояние ремиссии сохраняется в течение времени больше t (11.2). Эта функция является основной и наиболее часто применяемой в анализе. Для краткости будем ее называть функцией сохранения состояния ремиссии.

Интенсивность срыва ремиссии есть производная

$h(t) = -\frac{1}{S(t)}\frac{dS(t)}{dt}$, где S(t) - функция сохранения состояния ремиссии. Она характеризует долю случаев срыва ремиссии, приходящуюся на один день наблюдения. Например, на втором интервале при среднем числе наблюдений 436,5 чел, срыв ремиссии произошел у 42 чел. Интенсивность срыва ремиссии за один день наблюдения при продолжительности интервала 91,25 дн. будет

$$h(t) = \frac{42}{436{,}5 \times 91{,}25} = 0{,}00\,1\text{сл/день}$$

(в таблице эта величина (Hazard Rate) тоже = 0,001).

Для построения таблицы продолжительности ремиссии наблюдавшееся время разбито на 12 интервалов с таким расчетом, чтобы ширина интервала составила один квартал (3 мес., или 91,25 дня). В таблице указаны:

– нижние границы, средние точки и ширина интервалов;

– число наблюдений в начале и среднее число наблюдений на интервале, количество случаев срыва ремиссии и ушедших из поля наблюдения (цензурированных наблюдений);

– частости срыва и сохранения состояния ремиссии на интервалах;

– функции сохранения состояния ремиссии, плотности вероятности и интенсивности срыва ремиссии на интервалах;

– медианы ожидаемой продолжительности ремиссий для пациентов каждого интервала;

– стандартная ошибка определения функций срыва и сохранения состояния ремиссии.

Данные машинограммы 11.1 свидетельствуют, что из 556 пациентов ушли из поля наблюдения (цензурированы) 313 чел, в том числе 3 года наблюдались 203 чел.

Функция сохранения состояния ремиссии за период наблюдения уменьшается от 1 до 0,5026 по закону распределения, близкому к экспоненциальному (рис.11.1). Интенсивность срыва ремиссии отличается существенной неравномерностью (рис.11.2): в течение первого квартала она имеет максимальные значения – 0,0018, второго - третьего кварталов – 0,0011, в последующие периоды стабилизируется и удерживается на уровне в 3-9 раз меньше – 0,0006 – 0,0002.

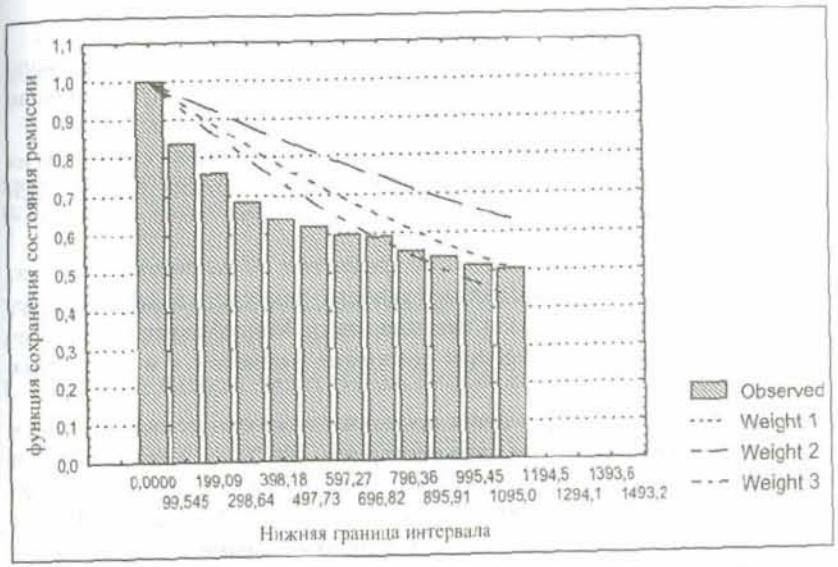

Рис.11.1.Функция сохранения состояния ремиссии.

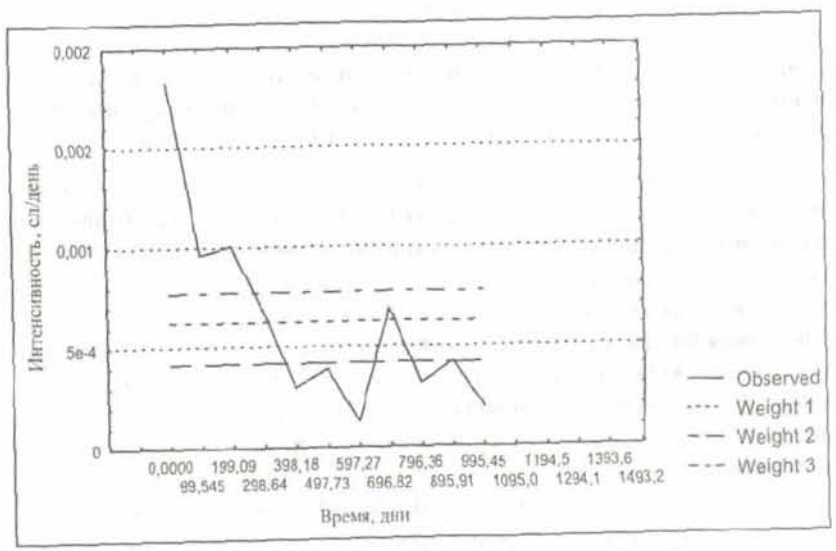

Рис.11.2.График интенсивности срыва ремиссии.

На рис.11.1 видно, что из общего числа получивших лечение около 75% пациентов сохранили состояние ремиссии более полугода. Около 65% - более года и около 50% - более трех лет наблюдения.

Ошибки оценки функции сохранения состояния ремиссии и других функций малы и оцениваются величинами стандартных ошибок 1,5-2,3% от оцениваемых ими величин.

Сравнение продолжительности ремиссии в различных группах. Сравнительная оценка продолжительности ремиссии в двух или нескольких группах (Comparing two samples or Multiple samples) выполняется по функциям сохранения состояния ремиссии в исследуемых группах.

Оценку значимости различия продолжительности ремиссии в двух группах можно получить по пяти критериям. Наиболее адекватным для условий примера при указании уровней признаков кодами по порядковой шкале является критерий Гехана-Вилкоксона (Gehan's Wilcoxon test), *представляющий собой модификацию непараметрического ран*гового критерия Вилкоксона для оценки значимости различия показателя в двух независимых выборках. Исходная база данных позволяет сравнивать продолжительность ремиссии в группах, различающихся по таким признакам, как пол, возраст, образование и др. Проведем исследование для трех групп по группировочному признаку - шкала осознанности лечебной ситуации (ERSOV2) и двух групп по группировочному признаку - плата за лечение (PLATA_LC).

Сравнение продолжительности ремиссии в группах по признаку осознанности лечебной ситуации (ERSOV2). Выдвигается гипотеза о влиянии на продолжительность ремиссии уровня осознанности больным лечебной ситуации (анозогнозии и самооценки). Чем выше уровень такой осознанности заболевания алкоголизмом, тем больше наблюдаемая продолжительность ремиссии после лечения. Принятие или отклонение этой гипотезы выполняется по результатам сравнения функций сохранения состояния ремиссии в трех группах группировочного признака ERSOV2: 1 - высокая, 2 - средняя и 3 - низкая осознанность.

Решение задач по модулю Comparing multiple samples включает:

– расчет критерия χ^2 Пирсона для оценки значимости нулевой гипотезы;

– таблицы функции сохранения состояния ремиссии для трех групп (машинограмма 11.2);

– график функций сохранения состояния ремиссии (рис.11.3);

– оценку значимости нулевой гипотезы о соответствии функций сохранения состояния ремиссии в двух группах по ранговому критерию WW- Гехана-Вилкоксона.

Машинограмма 11.2

Variables: REM_CENZ by ERSOV2 (3 groups) (mirvc_91.sta)
Censoring var.: CENZOR ChiI = 3,75092 df = 3 p = ,15330

Life Table for Group 1 (mirvc_91.sta, ERSOV2)

	No.Enter	No.Cnsrd	No.Dying	% Srvvng	Cum.%.Sr
0,000000	21	1	2	90,244	100
121,6667	18	0	1	94,444	90,244
243,3333	17	0	1	94,118	85,230
365,0000	16	1	0	100	80,217
486,6667	15	0	0	100	80,217
608,3333	15	0	0	100	80,217
730,0000	15	0	1	93,333	80,217
851,6667	14	0	0	100	74,869
973,3334	14	14	0	100	74,869
1095,000	0	0	0	0	74,869

Продолжение машинограммы 11.2

Life Table for Group 2 (mirvc_91.sta, ERSOV2)

	No.Enter	No.Cnsrd	No.Dying	% Srvvng	Cum.%.Sr
0,000000	462	27	83	81,494	100
121,6667	352	12	41	88,150	81,4939
243,3333	299	10	23	92,177	71,8371
365,0000	266	28	16	93,651	66,2172
486,6667	222	6	10	95,434	62,0129
608,3333	206	1	10	95,134	59,1813
730,0000	195	0	10	94,872	56,3014
851,6667	185	3	9	95,095	53,4141
973,3334	173	167	6	93,296	50,7944
1095,000	0	0	0	0	47,3892

Продолжение машинограммы 11.2

Life Table for Group 3 (mirvc_91.sta, ERSOV2)

	No.Enter	No.Cnsrd	No.Dying	% Srvvng	Cum.%.Sr
0,000000	73	8	13	81,159	100
121,6667	52	1	10	80,583	81,15942
243,3333	41	3	3	92,405	65,400
365,0000	35	9	0	100	60,433
486,6667	26	0	1	96,154	60,433
608,3333	25	0	1	96	58,109
730,0000	24	0	1	95,833	55,784
851,6667	23	0	1	95,652	53,460
973,3334	22	22	0	100	51,136
1095,000	0	0	0	0	51,136

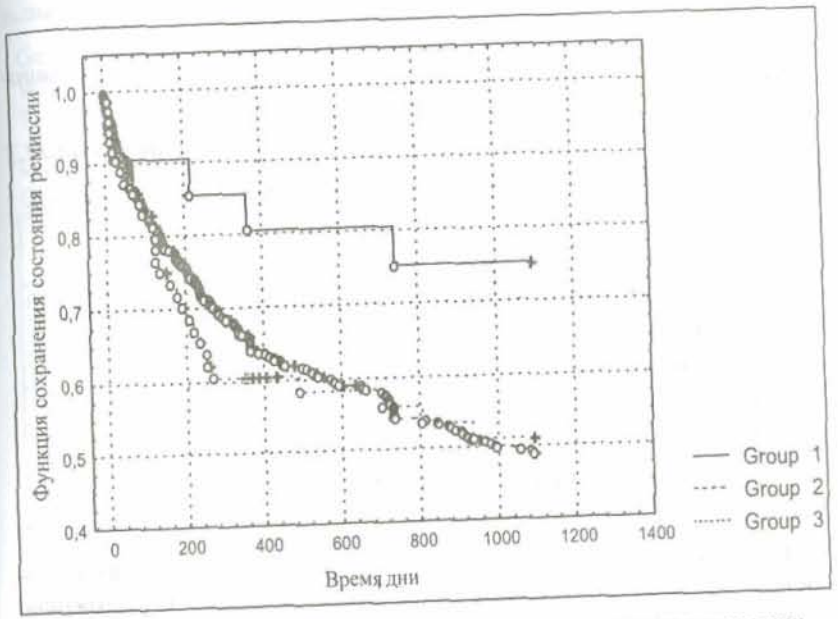

Рис.11.3.Функция сохранения состояния ремиссии в зависимости от осознанности своего состояния больным.

Анализ таблиц и графика функций сохранения состояния ремиссии опровергает выдвинутую гипотезу о влиянии уровня осознанности лечебной ситуации больным на продолжительность ремиссии.

В табл.11.2 даны квартили продолжительности ремиссии в трех группах, определенные по данным таблиц и графика.

Таблица 11.2

Квартили продолжительности ремиссии.

Вероятность сохранения состояния ремиссии, %	Квартили продолжительности сохранения состояния ремиссии (в днях)		
	Гр. 1	Гр. 2	Гр. 3
75	1100	200	150
50		1100	1000

Из таблицы видно, что 75% больных первой группы и 50% больных второй и третьей групп сохраняют состояние ремиссии в течение

всего трехлетнего периода наблюдения, в то время как 50% больных третьей группы удерживают это состояние 1000 дней.

Результаты расчета критерия Гехана-Вилкоксона даны в машинограмме 11.3.

Машинограмма 11.3

Критерий Гехана-Вилкоксона

Gehan's Wilcoxon Test (mirvc_91.sta, ERSOV2, Gr 1 - Gr 2)
WW = 1941,0 Sum = 2557E4 Var = 1065E3
Test statistic = 1,879930 p = ,06012

Gehan's Wilcoxon Test (mirvc_91.sta, ERSOV2Gr 1 - Gr 3)
WW = 293,00 Sum = 1688E2 Var = 29592,
Test statistic = 1,700350 p = ,08907

Gehan's Wilcoxon Test (mirvc_91.sta,ERSOV2, Gr 2 - Gr 3)
WW = 920,00 Sum = 3446E4 Var = 4068E3
Test statistic = ,4558812 p = ,64848

Тем не менее, с уровнем значимости p<0,10 (что значит с достоверностью 1-p>0,90) можно утверждать, что установлено существенное различие функций сохранения состояния ремиссии в 1 и 2, 1 и 3 группах. Различие между 2 и 3 группами незначимо (p=0,64848).

Сравнение продолжительности ремиссии в группах по признаку вида платы за лечение (PLATA_LC). Выдвигается гипотеза, что оплата лечения за свой счет определяет большую продолжительность сохранения ремиссии, чем в случаях оплаты за счет других источников. Принятие или отклонение этой гипотезы выполняется по результатам сравнения функций сохранения состояния ремиссии в двух группах по группировочному признаку PLATA_LC: 1 - своими деньгами, 2 - деньгами из других источников.

Решение по модулю Comparing two samples включает:

– таблицу функций сохранения состояния ремиссии для двух групп (машинограмма 11.4);

– график этих функций (рис.11.4);

– оценку значимости различий функций сохранения состояния ремиссии по критерию Гехана-Вилкоксона (машинограмма 11.4).

Gehan's Wilcoxon Test (mirvc_91.sta)
WW = 4768,0 Sum = 3848E4 Var = 3098E3
Test statistic = 2,708645 p = ,00676

Life Table for Group 1 and Group 2 (mirvc_91.sta)
Group 1: Code 1 Group 2: Code 2

	Group 1: No.Enter	Group 2: No.Enter	Group 1: No.Cnsrd	Group 2: No.Cnsrd	Group 1: No.Dying
0,000000	507	49	30	6	87
121,6667	390	32	12	1	43
243,3333	335	22	11	2	24
365,0000	300	17	35	3	14
486,6667	251	12	4	2	11
608,3333	236	10	1	0	11
730,0000	224	10	0	0	11
851,6667	213	9	3	0	8
973,3333	202	7	0	0	6
1095,000	196	7	196	7	0

Продолжение машинограммы 11.4

	Group 2: No.Dying	Group 1: % Srvvng	Group 2: % Srvvng	Group 1: Cum.%.Sr	Group 2: Cum.%.Sr
0,000000	11	82,317	76,087	100	100
121,6667	9	88,802	71,429	82,317	76,087
243,3333	3	92,716	85,714	73,099	54,348
365,0000	2	95,044	87,097	67,775	46,584
486,6667	0	95,582	100	64,416	40,573
608,3333	0	95,329	100	61,570	40,573
730,0000	1	95,089	90	58,695	40,573
851,6667	2	96,217	77,778	55,812	36,516
973,3333	0	97,030	100	53,701	28,401
1095,000	0	100	100	52,106	28,401

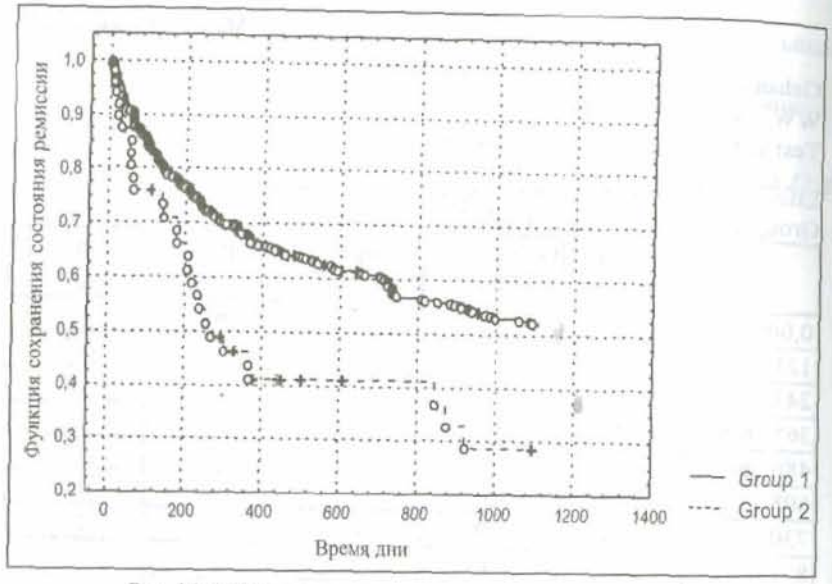

Рис.11.4.Функция сохранения состояния ремиссии в зависимости от вида платы за лечение.

Анализ графика на рис.11.4 и машинограммы 11.4 убедительно подтверждает выдвинутую гипотезу. Действительно, плата за лечение за свой счет значительно увеличивает продолжительность ремиссии ($p<0,001$). В табл.11.3 даны квартили продолжительности ремиссии в двух группах.

Таблица 11.3

Квартили продолжительности ремиссии

Вероятность сохранения состояния ремиссии, %	Квартили продолжительности сохранения состояния ремиссии (в днях)	
	Гр. 1	Гр. 2
75	210	120
50	1100	260

Из таблицы 11.3 следует, что продолжительность сохранения состояния ремиссии у больных, уплативших за лечение своими деньгами,

в 2-4 раза больше, чем у больных, за лечение которых уплачено из других источников.

Как было сказано выше, подобные исследования могут быть выполнены по многим другим признакам группировки больных в исходной матрице наблюдений.

Построение моделей функций продолжительности ремиссии в зависимости от факторов, влияющих на сохранение состояния ремиссии. Модели функций распределения продолжительности ремиссии могут быть построены методом регрессионного анализа при предположении об экспоненциальном, нормальном или логнормальном распределении времени ремиссии. Зависимой переменной в этих моделях выступает продолжительность ремиссии, независимые переменные - факторы, влияющие на нее. Имеется возможность построить модели для любой группы пациентов, задаваемой уровнем соответствующего группировочного признака. В случаях задания факторов кодами по порядковой шкале наиболее подходящей является модель интенсивности срывов ремиссии при допущении пропорциональности рисков:

$$h(t;x) = h_0(t;x) \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k), \qquad (11.8)$$

где $h(t;x)$ интенсивность срыва ремиссии на время t при воздействии множества факторов X1, X2, ... Xk ;

символом x обозначают центрированные значения факторов, т.е. разность задаваемого и среднего значения этого фактора;

$h_0(t;x)$ - интенсивность срыва ремиссии на время *t* при воздействии того же множества факторов, задаваемых средними значениями, т.е. $x_1 = x_2 = \ldots x_k = 0$;

$\beta_1, \beta_2, \ldots \beta_k$ - коэффициенты при независимых переменных X1, X2, ... Xk;

$\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$ - коэффициент пропорциональности между $h(t;x)$ и $h_0(t;x)$.

Для получения модели (11.8) следует указать способ моделирования: Proportional hazard (Cox) regression.

В результате решения по матрице наблюдений базы данных получены:

– таблица коэффициентов модели для девяти наиболее значимых факторов (машинограмма 11.5);

– график функции сохранения состояния ремиссии для средних значений факторов (рис.11.5).

Машинограмма 11.5

Dependent Variable: REM_CENZ (mirvc_91.sta)
Censoring var.: CENZOR
Chil = 60,5626 df = 9 p = ,00000

Номер фактора	Код фактора	Beta	Standard Error	t-value	exponent beta	Wald Statist.	p
4	POL1	0,532	0,229	2,319	1,702	5,379	0,020
5	ZA	0,417	0,202	2,067	1,518	4,274	0,039
6	KOD_L_R	0,427	0,254	1,681	1,532	2,824	0,093
7	ERSOV2	0,291	0,162	1,799	1,338	3,237	0,072
8	LEC_DO_M	0,247	0,072	3,445	1,280	11,870	0,001
11	KL_TIP	0,141	0,052	2,730	1,152	7,452	0,006
12	KOD_SROK	-0,455	0,134	-3,384	0,634	11,455	0,001
15	SEMJA	0,364	0,109	3,344	1,439	11,182	0,001
17	PLATA_LC	0,660	0,206	3,198	1,934	10,225	0,001

По данным машинограммы 11.5 построена модель для интенсивности срыва ремиссии, оцененная по критерию Хи-квадрат максимального правдоподобия как достоверная (χ^2=60,56, p=0,000, достоверность 1-p=1-0000=1). Все коэффициенты модели значимы с уровнем значимости p<0,10.

$$h(t;x) = h_0(t;x)\exp(0{,}532x_4 + 0{,}417x_5 + 0{,}427x_6 + 0{,}291x_7 + 0{,}247x_8 + 0{,}141x_{11} - 0{,}455x_{12} + 0{,}364x_{15} + 0{,}660x_{17}) \quad (11.9)$$

x_4, x_5, ... x_{17} - центрированные значения факторов, т.е. разности текущих и средних значений этих факторов, например:

$$x_4 = POL - \overline{POL}$$

$$x_5 = ZA - \overline{ZA}$$

.....

$$x_{17} = PLATA\_LC - \overline{PLATA\_LC}$$

По знакам коэффициентов модели (11.9) видно, что все факторы при возрастании уровней увеличивают интенсивность срыва ремиссии, кроме одного фактора KOD_SROK (x_{12}), с увеличением уровня которого происходит снижение интенсивности срыва ремиссии. Это значит, что интенсивность срыва ремиссии уменьшается при выборе большего срока лечебного зарока.

По данным exp(β) в машинограмме 11.5 можно оценить относительную величину степени влияния k,% 9-ти факторов, включенных в модель (11.9).

$$k_i = \frac{100\ \exp(\beta_i)}{\sum \exp(\beta_i)} \quad (11.10)$$

Для условий примера по машинограмме 11.5 $\sum \exp(\beta_i) = 12{,}529$, степень влияния, например, фактора x_4 (POL) $k_4 = \frac{100 \times 1{,}702}{12{,}529} = 13{,}59\%$.

В табл.11.4 даны степени влияния факторов на интенсивность срыва ремиссии.

Таблица 11.4

Относительные величины степени влияния факторов на интенсивность срыва ремиссии

№ фактора	Наименование фактора	Степень влияния k_i,%
4	POL	13,59
5	ZA	12,11
6	KOD_L_R	12,23
7	ERSOV2	10,68
8	LEC_DO_M	10,22
11	KL_TIP	9,19
12	KOD_SROK	5,06
15	SEMJA	11,48
17	PLATA_LC	15,44
		$\Sigma k_i = 100{,}0$

Из таблицы следует, что степень влияния на интенсивность срыва ремиссии факторов PLATA_LC, POL, составляет 29,02%, т.е. почти третью часть из всех, включенных в модель факторов.

Остальные факторы имеют примерно одинаковую степень влияния по 9-12% за исключением выбранного срока зарока (5,06%).

График функции сохранения состояния ремиссии, адекватный модели (11.9) интенсивности срыва ремиссии при средних значениях

факторов $x_4 = x_5 = \ldots x_{17} = 0$, дан на рис.11.5. По графику видно, что 75% пациентов сохраняют состояние ремиссии более 200 дней, 50% - более 1100 дней после лечения, т.е. в течение трех лет наблюдения.

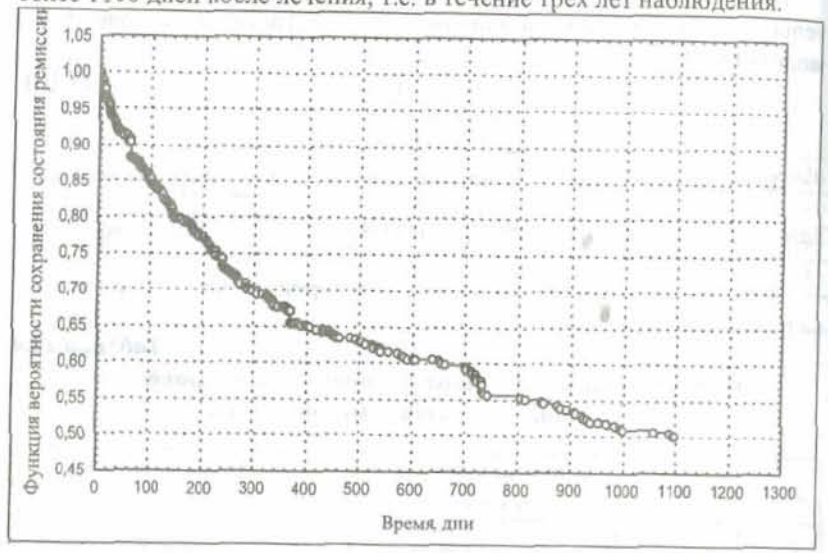

Рис.11.5.Функция вероятности сохранения состояния ремиссии при средних значениях, включенных в модель факторов.

Программой предусмотрена опция построения графиков функции вероятности сохранения состояния ремиссии для задаваемых исследователем значений факторов, детерминирующих продолжительность ремиссии.

На рис.11.6 дан график функции сохранения состояния ремиссии для самых благоприятных условий со значениями факторов, указанных в таблице 11.5, на рис.11.7 - график для крайне неблагоприятных условий со значениями факторов, указанными в той же табл.11.5.

При благоприятных условиях ожидается сохранение состояния ремиссии в течение трех лет наблюдения у 93,5% больных. При неблагоприятных условиях только 50% больных сохраняют состояние ремиссии в течение 1-1,5 мес., 20% - в течение 2 мес., а после года наблюдения уже не остается практически ни одного больного, сохранившего состояние ремиссии.

Таблица 11.5

Уровни факторов для построения функции сохранения ремиссии для средних, благоприятных, неблагоприятных условий и для больного k

№ признака	Код признака	Среднее значение	Благоприятные	Неблагоприятные	Больной k
4	POL	1,87	1	2	1
5	ZA	1,87	1	2	2
6	KOD_L_R	1,91	1	2	2
7	ERSOV2	2,09	1	3	2
8	LEC_DO_M	2,03	1	4	4
11	KL_TIP	2,86	1	5	4
12	KOD_SROK	1,69	2	1	3
15	SEMJA	1,24	1	3	2
17	PLATA_LC	1,09	1	2	2

По данным об уровнях факторов для каждого конкретного больного можно построить функции сохранения состояния ремиссии этим больным для прогноза ожидаемого результата лечения. В табл.11.5, для примера, даны уровни факторов для больного k, они близки к условиям ниже средних значений. Этим данным соответствует функция сохранения состояния ремиссии на рис.11.8.

По графику легко определить, что больной сохранит состояния ремиссии в течение одного года с вероятностью 48%, трех лет - с вероятностью 32%.

Аналогичные оценки влияния факторов на продолжительность ремиссии можно получить по модели (11.9) для интенсивности срыва ремиссии. Подставляя значения факторов в формулу (11.9) для различных условий, указанных в таблице 11.5, нетрудно определить, что интенсивность срыва ремиссии $h(t)$ по сравнению с $h_0(t)$ для средних значений факторов в 16 раз увеличивается при неблагоприятных условиях и в 10 раз уменьшается при благоприятных условиях.

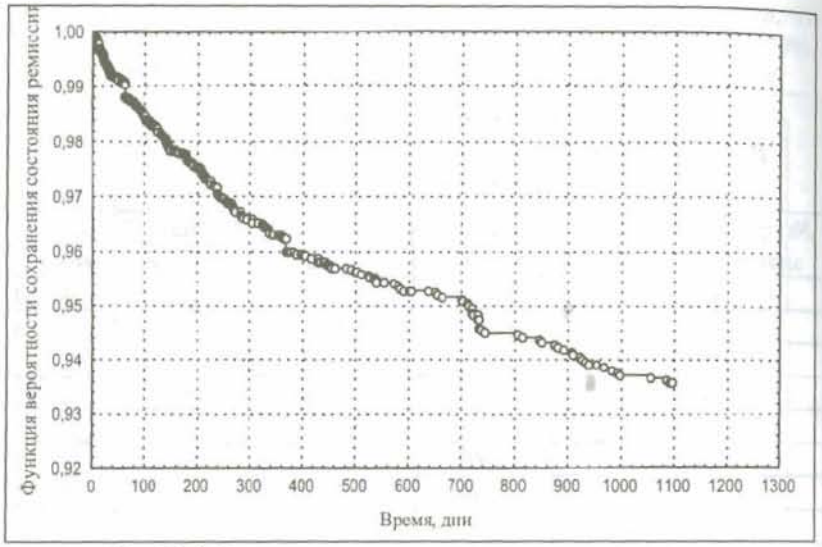

Рис.11.6.Функция вероятности сохранения состояния ремиссии при благоприятных значениях факторов.

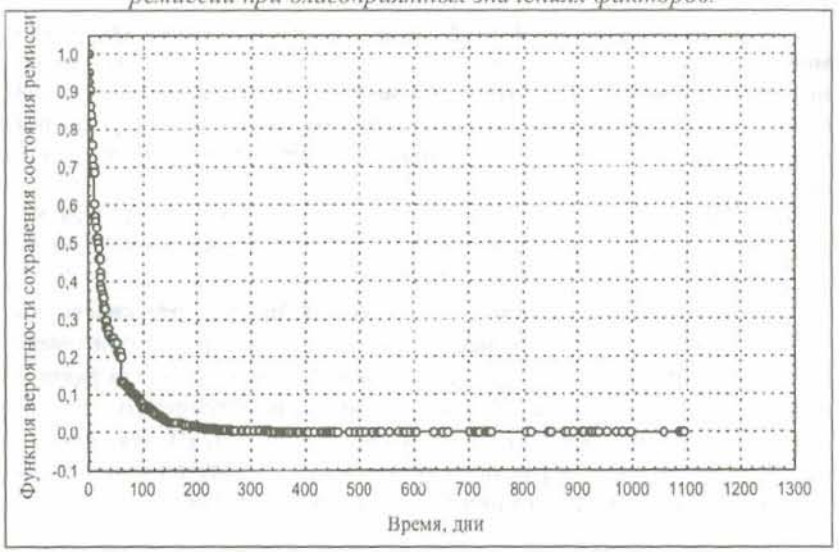

Рис.11.7.Функция вероятности сохранения состояния ремиссии при неблагоприятных значениях факторов.

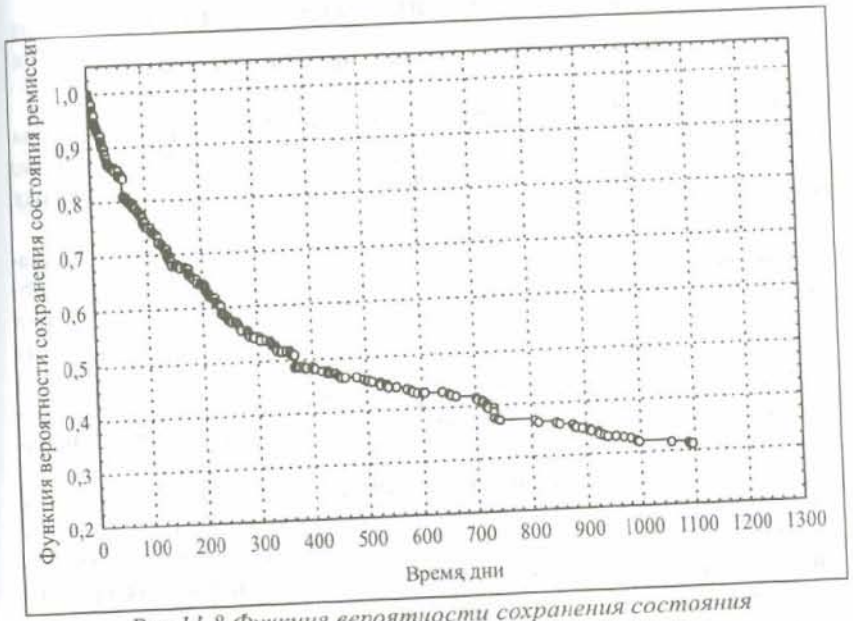

Рис.11.8.Функция вероятности сохранения состояния ремиссии для больного k.

Литература

1. Беляев Ю.К. Статистические методы обработки неполных данных о надежности изделий. - М.: Знание, 1987. - С. 3-55.

2. Ермаков С.П., Гаврилова Н.С. Первичная статистическая обработка данных по выживаемости организмов // Популяционная геронтология. - М: ВИНИТИ, 1987. - С. 230-276.

3. Кокс Д.Р., Оукс Д. Анализ данных типа времени жизни: Пер. с англ. - М.: Финансы и статистика, 1988. - 191 с.

4. Математико-статистические методы в клинической практике/ Под ред. В.И.Кувакина. - СПб.: ВМедА, 1993. - 199 с.

5. Флетчер Р., Флетчер С., Вагнер Э. Клиническая эпидемиология. Основы доказательной медицины. Пер. с англ. – М. Медиа Сфера, 1998. – 352 с.

Глава 12. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

Задачи и методы анализа временных рядов

Процесс изменения исследуемых показателей во времени описывают временными (динамическими) рядами. Например, временной ряд уровня заболеваемости, расхода лекарственных средств, показателя, характеризующего развитие эпидемического процесса и др.

Временным рядом называется последовательность зависимых значений показателя за некоторый период времени. Различают интервальные и моментные временные ряды. Интервальным называется ряд показателя, характеризующий его значения за определенные равные интервалы времени (год, квартал, месяц, декаду и т.д.). Моментным называется ряд показателя, характеризующий его значения на определенные моменты времени. Например, численность сотрудников предприятия на первое число каждого квартала в течение года.

Как интервальный, так и моментный временные ряды, относятся к дискретным рядам, в отличие от непрерывных, в которых значения показателя регистрируются непрерывно в течение времени наблюдения. В этой главе будет рассматриваться анализ дискретных временных рядов.

Анализ временных рядов – это совокупность методов математической статистики и теории случайных процессов предназначенных для вскрытия закономерностей изменения временной последовательности дискретных значений показателя за период наблюдения, а также для прогноза возможных его значений. На базе выявленных закономерностей и прогностических моделей разрабатывается план управления изученным процессом в последующие периоды.

Основное предположение при анализе временных рядов заключается в том, что случайный процесс, лежащий в основе ряда, является стационарным, обладающим свойством эргодичности. Под стационарным понимается процесс, вероятностные свойства которого не зависят от начала отсчета времени. Иначе говоря, сохраняют примерное постоянство среднее значение показателя и его дисперсия, а автокорреляция между парами значений показателя зависит только от временного сдвига (его называют лагом) между ними. Практически наблюдаемые реализации временных рядов сравнительно редко бывают стационарными. Обычно они содержат различные тренды (изменение среднего ожидаемого значения показателя со временем) и периодические компоненты.

При анализе временных рядов обычно выделяют три компоненты:

— тренд или систематическое изменение показателя u_t;

— периодическую компоненту, связанную, например, с сезонными изменениями показателя v_t;

— случайную компоненту вследствие влияния на значения показателя случайных факторов и ошибок ε_t.

Таким образом, значение показателя на время t может быть представлено в виде:

$$y_t=u_t+v_t+\varepsilon_t. \qquad 12.1$$

Чтобы привести временной ряд к стационарному виду из него исключают тренд, оставляя для анализа периодическую и случайные компоненты:

$$v_t+\varepsilon_t=y_t-u_t, \qquad 12.2$$

или применяют метод разностей между значениями показателя, что по существу означает переход к исследованию ряда прироста исследуемого показателя.

Свойством эргодичности обладают ряды с внутренней однородностью процесса генерирующего временную последовательность значений показателя. Свойство эргодичности позволяет при анализе временных рядов оценивать их выборочные характеристики по одной реализации достаточной продолжительности, такой, чтобы периодическая сезонная компонента в процессе наблюдения повторялась несколько раз. При исследовании временных рядов уровня заболеваемости желательно иметь наблюдения за период 4-10 и более лет.

Итак, получив временной ряд, исследователь должен описать многолетнюю тенденцию изменения показателя – тренд и периодическую компоненту – сезонную колеблемость показателя.

Тренд легко описать методом однофакторного регрессионного анализа. В качестве фактора или независимой переменной выступает индекс времени t = 1, 2,……, T, а зависимой переменной является уровень исследуемого показателя $y_1, y_2, \ldots\ldots, y_T$,
где T – общее количество наблюдений во временном ряду (период наблюдения, например, T=5 лет, или T=60 мес.).

Для описания тренда достаточно применить одну из четырех видов моделей:

- линейную $u_t=a+bt$; (12.3)
- *квадратичную (параболическую)* $u_t=a+bt^2$; (12.4)
- обратную (гиперболическую) $u_t=a+b/t$ (12.5)
- экспоненциальную (показательную) $u_t=ae^{bt}$. (12.6)

Вид уравнения тренда выбирается на основании качественного анализа исходного временного ряда после построения графика временной последовательности (рис.12.1, 12.2).

Понятие сезонности используют для описания периодических компонент временных рядов, которые связаны с причинами, внешними по отношению к основным механизмам, определяющим поведение системы. *Обычно сезонные компоненты связаны с каким-либо календарным* циклом (например, смена времен года, формирование детских дошкольных коллективов и др.).

В моделировании сезонной периодической компоненты применяют метод, разработанный Боксом и Дженкинсом (1976), основанный на авторегрессии и интегрированного скользящего среднего (Autoregressive Integrated Moving-Average, *сокращенно ARIMA*).

Построение модели временного ряда методом авторегрессии и интегрированного скользящего среднего (ARIMA)

Процедуры анализа и моделирования временных рядов методом ARIMA реализованы в современных статистических ППП для ПЭВМ. Применение метода обеспечивает получение:

— автокорреляционной функции, характеризующей зависимость временной последовательности значений исследуемого показателя;

— периодограммы для спектрального анализа временного ряда и выявления причин колеблемости показателя;

— оценочной прогностической модели показателя, учитывающей как непериодическую, так и периодическую компоненты временного ряда;

— прогнозируемых значений показателя и оценок их точности и надежности.

Автокорреляционной функцией $K_y(\tau)$ коэффициентов парной или парциальной (частной) автокорреляции характеризуют связь значений *показателя во временном ряду в зависимости от временного интервала* между ними (лага τ).

Коэффициенты парной автокорреляции рассчитывают по формуле:

$$K_{y(\tau)} = \frac{\sum_{i=1}^{n}\left[y_i(t)y_i(t+\tau)\right] - n\bar{y}(t)\bar{y}(t+\tau)}{\sqrt{\left[\sum_{i=1}^{n} y_i^2(t) - n\bar{y}^2(t)\right]\left[\sum_{i=1}^{n} y_i^2(t+\tau) - n\bar{y}^2(t+\tau)\right]}}, \qquad 12.7$$

где $y_i(t)$ – значение показателя на время t;

$y_i(t+\tau)$ –значение показателя на время $(t+\tau)$, т.е. при сдвиге времени на величину лага τ;

$\bar{y}(t)$ и $\bar{y}(t+\tau)$ – *средние значение* показателя на время t и на время $t+\tau$;

n – число пар значений показателя.

Значения коэффициентов парной автокорреляции *для различных* лагов в пределах заданного периода сводят в таблицу автокорреляционной функции (в отдельную таблицу не выделены, а представлены на рис.12.6 и 12.7). Там же даются оценки *точности коэффициентов* – средние квадратические ошибки $m_{K_{y(\tau)}\ S.E.}$ Для наглядности строится коррелограмма с указанием 95%-го доверительного интервала для значимых коэффициентов автокорреляции (рис.12.6 *и* 12.7). Значимость коэффициентов автокорреляции возрастает с увеличением числа пар значений показателей, т.е. продолжительности наблюдения.

Направление и сила связи между *значениями показателя*, разделенными соответствующими лагами, определяется по знаку и величине коэффициента автокорреляции. Если коэффициент имеет знак плюс – связь прямая положительная, минус *означает обратную* отрицательную связь. Если коэффициент по абсолютной величине меньше 0,3 – связь слабая, от 0,3 до 0,7 – связь умеренная, более 0,7 – связь сильная.

Быстрое затухание автокорреляционной *функции* с увеличением лага τ свойственно для стационарного временного ряда. Не затухание и периодический характер автокорреляционной функции, наоборот, свидетельствуют о нестационарности временного ряда: *наличии* в нем периодической компоненты (рис.12.6 и 12.7). По величине лагов между положительной (или отрицательной) автокорреляцией судят о периодах циклических колебаний показателя.

Наряду с коэффициентами парной автокорреляции выполняется расчет коэффициентов парциальной (частной) автокорреляции для более точной оценки связи между парами значений показателя, разделенными

соответствующими лагами, при исключении влияния на них промежуточных значений показателя в пределах лага. Автокорреляционная функция парциальной автокорреляции дополняет уже выполненную диагностику временного ряда по коэффициентам парной автокорреляции.

Для установления причин и закономерностей периодической колеблемости исследуемого показателя применяется спектральный анализ временного ряда. Известно, что между автокорреляционной функцией $K_y(\tau)$ и спектральной плотностью $S_y(w)$, описывающей плотность дисперсии показателя в зависимости от частоты колебаний w, имеется функциональная связь, носящая название преобразований Фурье:

$$K_y(\tau) = \int_0^\infty S_y(\tau) \cos w\tau dw; \qquad 12.8$$

$$S_y(w) = \frac{2}{\pi}\int_0^\infty K_y(\tau)\cos w\tau d\tau. \qquad 12.9$$

Зная автокорреляционную функцию сложной гармонической временной последовательности показателя $K_y(\tau)$, можно определить спектральную плотность дисперсий $S_y(w)$ простых гармоник, образующих временной ряд. Для удобства анализа, спектр дисперсий показателя в простейших составляющих гармониках выгодно представлять в виде графика в зависимости от относительной частоты f/f_0,

где f – частота простейшей гармоники;

f_0 – базовая частота колебаний, с периодом равным одному интервалу временного ряда (рис.12.9 и 12.10).

Такой график называется периодограммой. По нему устанавливают доминирующие колебания показателя, причины вызывающие их и периоды цикличности.

Повторимся, что оценочную прогностическую модель для показателя по данным временного ряда получают методом авторегрессии и интегрированного скользящего среднего (ARIMA). Модель для временного ряда может быть представлена, как линейная функция показателя на момент t на основе авторегрессии этого значения с предшествующими на момент t-1, t-2……t-p:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p}, \qquad 12.10$$

где a_0, a_1, a_2,……. a_p – коэффициенты модели, получаемые методом наименьших квадратов;

y_t y_{t-1} y_{t-2} y_{t-p} – значения показателя на момент t, t-1, t-2,……. t-p;

p - порядок авторегрессии, представляющий число предшествующих значений показателя, со значимыми коэффициентами.

Одновременно модель временного ряда может быть получена, как скользящего среднего порядка q, предоставляющего число значений показателя со значимыми коэффициентами.

Оказалось, что соединение двух видов моделей – авторегрессии и скользящего среднего в одну дает наиболее эффективный результат построения оценочной прогностической модели, учитывающей как непериодическую, так и периодическую компоненты изменения показателя во временном ряду. Получение надежных моделей для оценки показателя на момент t обеспечивается при учете нескольких предыдущих значений показателя, практически при порядке авторегрессии и скользящего среднего до 6.

Так как реальные временные ряды представляются, как правило, нестационарными, т.е. имеющими тренд и периодическую компоненту, то их необходимо приводить к стационарному виду. Для этого вместо временного ряда уровней показателя переходят к ряду приростов $\Delta y_t = y_t - y_{t-1}$, называемых разностями первого порядка d=1. Иногда при значительной нестационарности необходимо применять разности второго порядка (d=2). В этих случаях методом авторегрессии и интегрированного скользящего среднего определяют коэффициенты модели для разностей показателя первого или второго порядка.

Коэффициенты модели находят, применяя процедуру ARIMA методом нелинейных наименьших квадратов Макварта, путем последовательных приближений (итераций) к оптимальному решению, при котором коэффициенты модели имеют требуемые точность и надежность. Для реализации процедуры ARIMA необходимо задать параметры модели:

$$(p, d, q) \times (P, D, Q)\, s, \qquad (12.11)$$

где p и P – порядок авторегрессии для непериодической и периодической компонент;

q и Q – порядок скользящего среднего для непериодической и периодической компонент;

d и D – порядок разности для непериодической и периодической компонент;

s – величина периода сезонной колеблемости, т.е. количество наблюдений показателя за один период сезонных колебаний (например, при периоде сезонности равной одному году и помесячных данных временного ряда, s=12).

Как уже отмечено выше, порядок p, q, P и Q, в зависимости от степени нестационарности временного ряда, могут задаваться от 0 до 6. Порядок разности d и D может задаваться от 0 до 1, 2. Следует иметь в виду, что число определяемых коэффициентов оценочной прогностической модели обуславливается заданным порядком непериодической и периодической компонент авторегрессии и скользящего среднего. Наиболее простые компактные и достаточно эффективные модели могут быть получены при задании порядка до 1, 2 . Для их получения с помощью ПЭВМ не требуется большого времени. Число итераций определяется задаваемыми критериями точности коэффициентов и их числом.

Типичным при анализе временных рядов является получение нескольких вариантов моделей для различных заданных порядков p, q, P, Q, d и D. После анализа полученных вариантов исследователь может выбрать модель наиболее простую, но достаточно надежную по точности прогноза.

Результат решения задачи с помощью ПЭВМ включает таблицу прогнозируемых значений заболеваемости, их доверительных интервалов и стандартных ошибок (машинограммы 12.4 и 12.5). Наряду с таблицей программой предусмотрена выдача графика с результатом прогноза показателя на предстоящий период с указанием 50, 90, 95 или 99%-го доверительного интервала. Поэтому необходимость выражать модель аналитически, тем более, что она дает в большинстве случаев только значения разностей 1-го или 2-го порядка, отпадает. И прогноз, и оценку точности и надежности прогноза удобнее давать по таблице (машинограммы 12.4 и 12.5) или графику (рис.12.11 и 12.12). Для построения графика задается уровень вероятности доверительного интервала: в ответственных исследованиях 95-99%; при поисковых исследованиях 50-90%.

ПРИМЕР 12.1

На основе официальных данных о помесячном количестве острых вирусных гепатитов (Y, случаев) среди населения Самарской области за 17 лет в период с 1979 по 1995 г., исследуется динамика и структура заболеваемости вирусным гепатитом А (ГА) и гепатитом В (ГВ) (Торопов Д.Е., 1997 г.). Для решения задачи на ПК использован модуль Time Series ППП Statistica for Windows. Исходные данные в таблице 12.1.

Таблица 12.1

Данные помесячной заболеваемости населения Самарской области вирусными гепатитами А и В

Год	Месяц	ГА	ГВ
1979	1	361	67
1979	2	272	54
1979	3	238	39
1979	4	277	61
1979	5	267	55
1979	6	211	65
1979	7	189	43
1979	8	222	46
1979	9	398	43
1979	10	597	59
1979	11	652	63
1979	12	500	44
1980	1	423	52
1980	2	365	60
1980	3	317	63
1980	4	339	66
1980	5	267	73
1980	6	271	61
1980	7	255	54
1980	8	296	75
1980	9	506	58
1980	10	634	75
1980	11	592	66
1980	12	601	65
1981	1	381	71
1981	2	300	51
1981	3	296	42
1981	4	287	62
1981	5	208	51
1981	6	195	44
1987	7	200	79
1987	8	194	60
1987	9	334	60
1987	10	407	84
1987	11	449	96
1987	12	425	87
1988	1	318	96
1988	2	271	63
1988	3	238	73
1988	4	201	63
1988	5	191	94
1988	6	165	71
1988	7	155	50
1988	8	295	69
1988	9	450	69
1988	10	518	72
1988	11	586	90
1988	12	556	86
1989	1	374	79
1989	2	332	77
1989	3	318	76
1989	4	253	79
1989	5	239	44
1989	6	187	47
1989	7	155	58
1989	8	202	62
1989	9	527	62
1989	10	686	76
1989	11	723	109
1989	12	752	77

Год	Месяц	ГА	ГВ
1981	7	219	44
1981	8	258	55
1981	9	482	60
1981	10	563	49
1981	11	675	58
1981	12	446	42
1982	1	520	52
1982	2	463	63
1982	3	351	37
1982	4	292	62
1982	5	303	57
1982	6	248	70
1982	7	264	57
1982	8	320	76
1982	9	534	70
1982	10	805	69
1982	11	809	45
1982	12	645	73
1983	1	658	74
1983	2	578	61
1983	3	487	77
1983	4	419	67
1983	5	325	55
1983	6	316	73
1983	7	279	60
1983	8	450	76
1983	9	789	22
1983	10	1049	118
1983	11	907	59
1983	12	685	66
1984	1	594	59
1984	2	390	64
1984	3	366	67
1984	4	233	67

Год	Месяц	ГА	ГВ
1990	1	568	71
1990	2	408	71
1990	3	297	75
1990	4	328	47
1990	5	217	62
1990	6	149	61
1990	7	154	85
1990	8	216	82
1990	9	361	109
1990	10	515	99
1990	11	482	128
1990	12	444	79
1991	1	331	65
1991	2	274	58
1991	3	218	72
1991	4	128	59
1991	5	132	39
1991	6	143	56
1991	7	110	41
1991	8	200	58
1991	9	410	64
1991	10	458	101
1991	11	471	85
1991	12	540	82
1992	1	299	73
1992	2	223	69
1992	3	168	75
1992	4	160	81
1992	5	82	86
1992	6	59	64
1992	7	81	79
1992	8	96	84
1992	9	191	86
1992	10	309	125

Год	Месяц	ГА	ГВ
1984	5	260	51
1984	6	242	74
1984	7	230	53
1984	8	365	61
1984	9	655	61
1984	10	852	97
1984	11	1059	78
1984	12	890	85
1985	1	712	58
1985	2	433	56
1985	3	385	47
1985	4	342	80
1985	5	378	79
1985	6	306	70
1985	7	279	62
1985	8	383	84
1985	9	733	11
1985	10	1043	95
1985	11	918	91
1985	12	915	98
1986	1	718	97
1986	2	525	80
1986	3	334	62
1986	4	248	76
1986	5	247	70
1986	6	212	57
1986	7	233	53
1986	8	315	67
1986	9	568	84
1986	10	688	127
1986	11	857	94
1986	12	410	63
1987	1	311	83
1987	2	255	42

Год	Месяц	ГА	ГВ
1992	11	409	126
1992	12	342	90
1993	1	202	132
1993	2	177	145
1993	3	98	119
1993	4	75	136
1993	5	63	123
1993	6	69	141
1993	7	75	112
1993	8	119	154
1993	9	256	212
1993	10	408	203
1993	11	410	223
1993	12	420	191
1994	1	321	202
1994	2	273	212
1994	3	238	228
1994	4	129	187
1994	5	155	198
1994	6	117	168
1994	7	140	188
1994	8	186	187
1994	9	299	193
1994	10	465	188
1994	11	655	183
1994	12	703	216
1995	1	433	197
1995	2	432	235
1995	3	318	250
1995	4	243	198
1995	5	216	248
1995	6	201	213
1995	7	204	228
1995	8	377	212

Год	Месяц	ГА	ГВ
1987	3	215	91
1987	4	218	73
1987	5	184	73
1987	6	144	59

Год	Месяц	ГА	ГВ
1995	9	532	213
1995	10	620	216
1995	11	715	253
1995	12	773	273

Требуется:

1.Построить горизонтальные диаграммы динамики помесячного количества случаев заболеваний гепатитом А и гепатитом В.

2.Определить уравнения регрессии тренда количества случаев заболеваний гепатитом А и гепатитом В и оценить их информативность и значимость.

3.Определить автокорреляционные функции и построить графики автокорреляции с лагами 1-24 месяца для гепатита А и гепатита В.

4.Определить кросскорреляционные функции гепатита А и гепатита В с лагами от 0 до 12 месяцев и построить графики кросскорреляции.

5.Построить периодограммы разложения сложной периодичности заболеваемости гепатитом А и гепатитом В на простейшие многолетние и сезонные колебания.

6.Определить подходящие модели для описания динамики гепатита А и гепатита В и прогноза количества случаев заболеваний на 12 месяцев с 50%-м доверительным интервалом.

Решение:

1.Горизонтальные диаграммы динамики помесячного количества случаев заболеваний гепатитом А и гепатитом В - на рисунках 12.1 и 12.2.

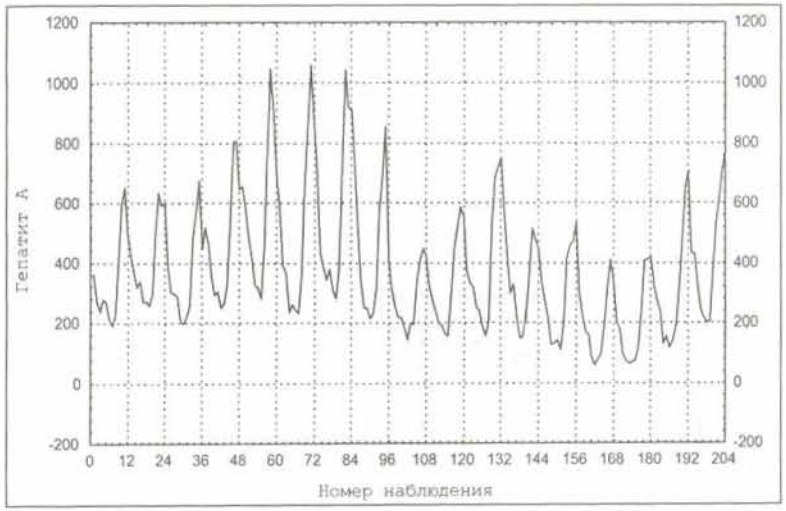

Рис.12.1. Горизонтальная диаграмма динамики помесячного количества случаев заболеваний гепатитом А.

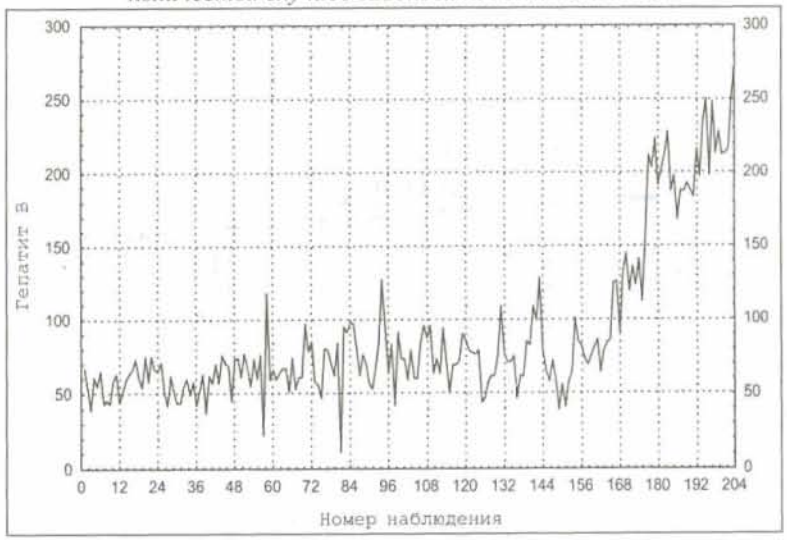

Рис.12.2. Горизонтальная диаграмма динамики помесячного количества случаев заболеваний гепатитом В.

2.Уравнения регрессии тренда количества случаев заболеваний гепатитом А и гепатитом В, оценка их значимости - в машинограммах 12.1 - 12.3, а графическое представление - на рисунках 12.3-12.5. Независимая переменная времени t в месяцах обозначена TIMS.

Машинограмма 12.1

Коэффициенты регрессионной модели прогноза числа случаев заболевания гепатитом А.

Regression Summary for Dependent Variable: ГА (tarop_1.sta)
R= ,22735631 RI= ,05169089 Adjusted RI= ,04699629
F(1,202)=11,011 p<,00107 Std.Error of estimate: 208,42

	BETA	St. Err. of BETA	B	St. Err. of B	t(202)	p-level
Intercpt			463,163	29,293	15,812	0,000
TIMS	-0,227	0,069	-0,822	0,248	-3,318	0,001

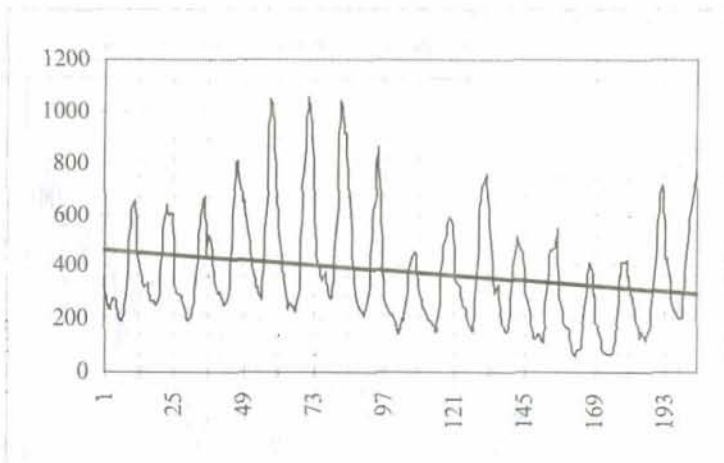

Рис.12.3. Горизонтальная диаграмма динамики помесячного количества случаев заболеваний гепатитом А и линия тренда типа $\hat{y} = 463{,}2 - 0{,}8 \times \text{TIMS}$.

Машинограмма 12.2

Коэффициенты регрессионной модели прогноза числа случаев заболевания гепатитом В за первые 13 лет сравнительно благополучного его распространения.

Regression Summary for Dependent Variable: ГВ (tarop_1.sta)
R= ,34829887 RI= ,12131210 Adjusted RI= ,11560633
F(1,154)=21,261 p<,00001 Std.Error of estimate: 16,974

	BETA	St. Err. of BETA	B	St. Err. of B	t(154)	p-level
Intercpt			56,583	2,731	20,718	0,000
TIMS	0,348	0,076	0,139	0,030	4,611	0,000

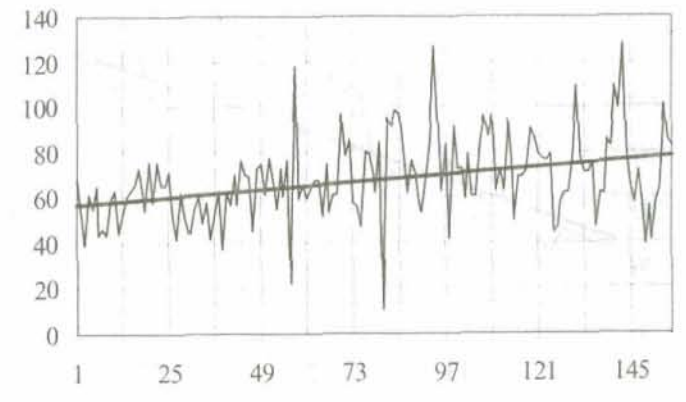

Рис.12.4. Горизонтальная диаграмма динамики помесячного количества случаев заболеваний гепатитом В и линия тренда типа $\hat{y} = 56{,}6 + 0{,}1 \times \text{TIMS}$.

Машинограмма 12.3

Коэффициенты регрессионной модели прогноза числа случаев заболевания гепатитом В за последние 4 года наблюдения.

Regression Summary for Dependent Variable: ГВ (tarop_1.sta)
R= ,90728583 RI= ,82316758 Adjusted RI= ,81932339
F(1,46)=214,13 p<,00000 Std.Error of estimate: 25,089

	BETA	St. Err. of BETA	B	St. Err. of B	t(46)	p-level
Intercpt			73,262	7,357	9,958	0,000
TIMS_1	0,907	0,062	3,825	0,261	14,633	0,000

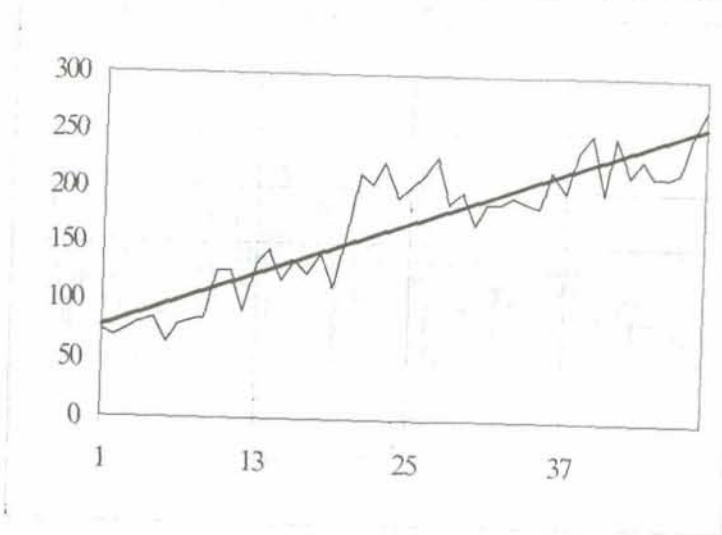

Рис.12.5. Горизонтальная диаграмма динамики помесячного количества случаев заболеваний гепатитом В и линия тренда типа $\hat{y} = 73{,}3 + 3{,}8 \times \text{TIMS\_1}$.

Временные ряды количества случаев заболевания острым вирусным гепатитом в месяц являются нестационарными с выраженной тенденцией снижения заболеваемости за период наблюдения гепатита А и подъема заболеваемости гепатита В, особенно за последние 4 года наблюдения. Обращает на себя внимание выраженная сезонная периодичность заболеваемости гепатитом А, связанная со сменой времени года. Модели тренда количества случаев заболеваний имеют вид:

– для гепатита А за период наблюдения $\hat{y} = 463{,}2 - 0{,}8 \times \text{TIMS}$;

– для гепатита В за первые 13 лет наблюдения $\hat{y} = 56{,}6 + 0{,}1 \times \text{TIMS}$;

– для гепатита В за последние 4 года наблюдения $\hat{y} = 73{,}3 + 3{,}8 \times \text{TIMS\_1}$.

Таким образом, заболеваемость гепатитом А имеет тенденцию незначительного снижения на 0,8 сл./мес. Заболеваемость гепатитом В имеет тенденцию незначительного роста в первые 13 лет наблюдения (0,1 сл./мес.) и существенный рост за последние 4 года наблюдения (с 1991 по 1995 г.г.) на 3,8 сл./мес. (около 46 сл./год).

По критерию F-Фишера полученные нами модели тренда оцениваются как значимые ($p<0{,}05$). Однако модели тренда гепатита А за весь период наблюдения, а также гепатита В за первые 13 лет наблюдения для прогноза непригодны из-за их недостаточной информационной способности (коэффициент детерминации $R^2<0{,}5$). В то же время модель тренда гепатита В за последние 4 года вполне приемлема для прогноза динамики его уровня, так как ее информационная способность составляет 82% ($R^2=0{,}82$).

3. Автокорреляционные функции количества случаев заболевания гепатитом А и гепатитом В представлены на рисунках 12.6 и 12.7.

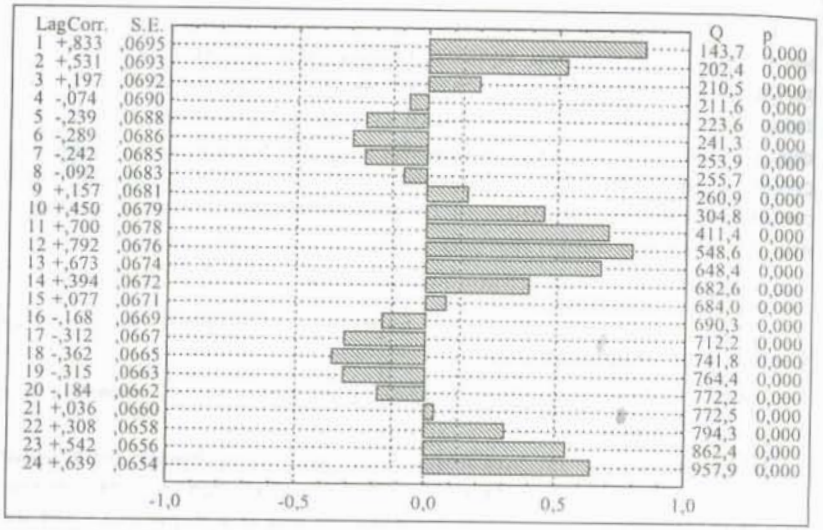

Рис.12.6. Автокорреляционная функция заболеваемости гепатитом А.

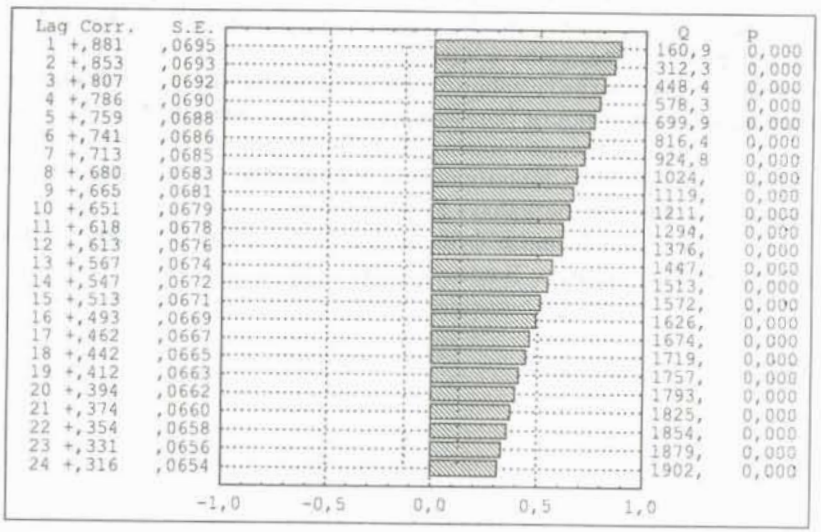

Рис.12.7. Автокорреляционная функция заболеваемости гепатитом В.

Автокорреляционный анализ подтверждает вывод о нестационарности временных рядов динамики заболеваемости гепатитом А и гепатитом В. Заболеваемость гепатитом А имеет резко выраженную значимую сезонную периодичность с периодом 12 месяцев. Заболеваемость гепатитом В такой периодичности не имеет и характеризуется значимой устойчивостью связи текущего уровня с предыдущими на протяжении всего периода наблюдения.

4. Кросскорреляционная функция гепатита А и гепатита В с лагом от 0 до 12 месяцев представлена на рисунке 12.8, из которого следует, что между двумя гепатитами установлена обратная, значимая кросскорреляционная связь на уровне до r = -0,3. После подъема уровня заболеваемости гепатитом А через 5-7 месяцев наблюдается снижение уровня гепатита В на 10,4% и, напротив, при предварительном снижении уровня заболеваемости гепатитом А, через 5-7 месяцев наблюдается подъем уровня заболеваемости гепатитом В на 21%.

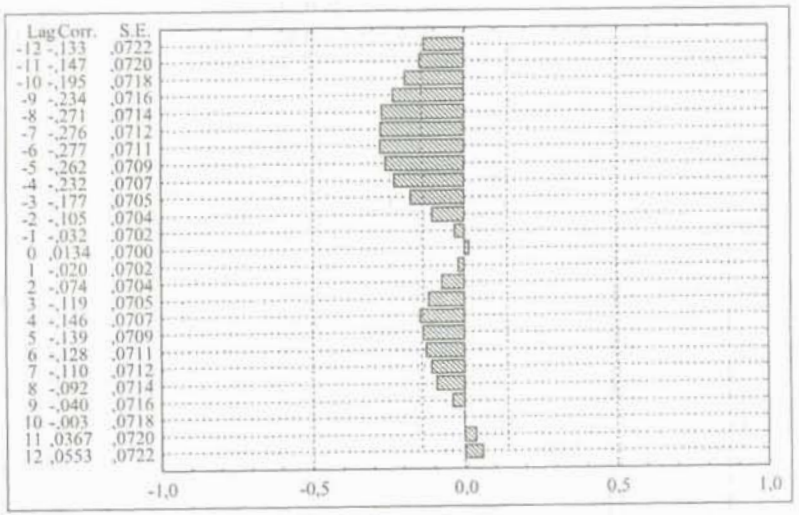

Рис.12.8. Кросскорреляционная функция заболеваемости гепатитом А и гепатитом В.

5. Периодограммы разложения сложной периодичности заболеваемости гепатитом А и гепатитом В на простейшие колебания представ-

лены на рисунках 12.9 и 12.10, из которых следует, что основными компонентами этой периодичности гепатита А являются:

–резко выраженная сезонная цикличность с периодом 12 мес. (с относительной частотой $f/f_0=T_0/T=1/12=0,08$), связанная со сменой времен года, с пиками амплитуды колебаний в июле–ноябре;

–слабо выраженную сезонную периодичность заболеваемости с периодом 6 мес. (с относительной частотой $f/f_0=T_0/T=1/6=0,17$) и значительно меньшей амплитудой;

–слабо выраженная периодичность с периодом 4 года ($f/f_0=T_0/T=1/48=0,02$) и амплитудой значительно меньше амплитуды годовой периодичности.

Основными компонентами наблюдавшейся периодичности заболеваний ГВ является:

–выраженная многолетняя периодичность с периодом сопоставимым с временем наблюдения;

–слабо выраженная двух волновая сезонная колеблемость с периодами 10-14 мес. и 5-7 мес. с пиками заболеваемости в летние и зимние месяцы.

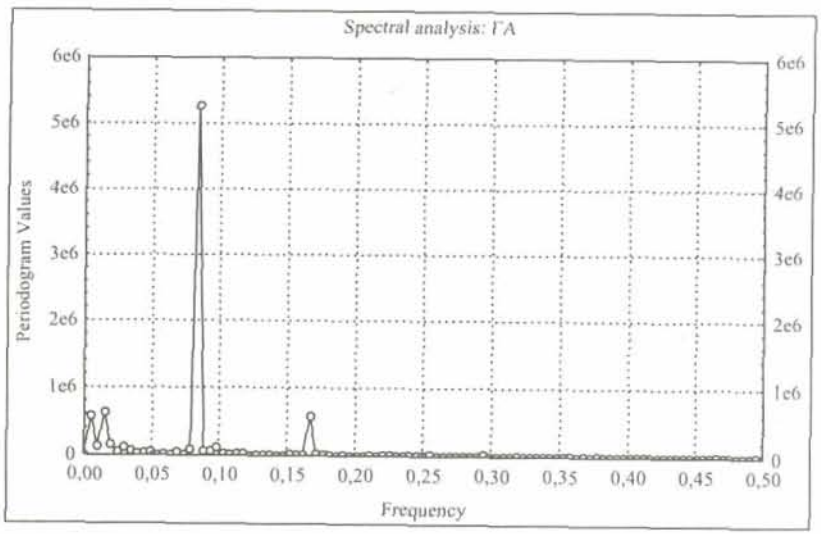

Рис.12.9. Периодограмма заболеваемости гепатитом А.

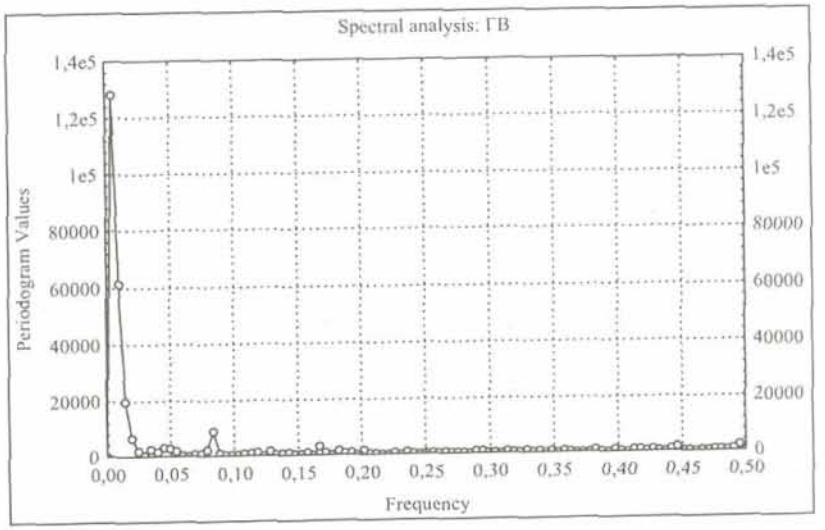

Рис.12.10. Периодограмма заболеваемости гепатитом В.

6. Результаты прогноза количества случаев заболеваний гепатитом А и гепатитом В по наиболее подходящим моделям ARIMA на 24 месяца после завершения периода наблюдения для гепатита А и на 12 месяцев для гепатита В – в машинограммах 12.4 и 12.5, а также на рисунках 12.11 и 12.12.

Машинограмма 12.4

Прогнозируемое число помесячных случаев заболевания гепатитом А на предстоящие 24 месяца.

Forecasts; Model:(1,0,1)(1,0,1) Seasonal lag: 12 (tarop_1.sta)
Input: ГА
Start of origin: 1 End of origin: 204

	Forecast	Lower 50,00%	Upper 50,00%	Std.Err.
205	567,4	509,6	625,1	85,5
206	510,0	432,4	587,5	114,8
207	414,3	324,1	504,6	133,6
208	332,7	233,5	431,9	146,8
209	298,8	193,1	404,5	156,4
210	264,4	153,8	375,0	163,7
211	257,1	142,8	371,4	169,2
212	335,1	217,8	452,3	173,5

	Forecast	Lower 50,00%	Upper 50,00%	Std.Err.
213	458,4	339,0	577,9	176,8
214	561,0	439,9	682,2	179,3
215	639,7	517,2	762,2	181,3
216	659,4	535,8	783,0	182,9
217	469,1	340,5	597,8	190,4
218	422,4	289,8	554,9	196,1
219	337,4	201,8	472,9	200,6
220	265,1	127,3	403,0	204,0
221	238,6	98,9	378,3	206,7
222	210,8	69,7	351,9	208,8
223	208,8	66,6	351,1	210,5
224	289,3	146,2	432,4	211,8
225	413,5	269,7	557,3	212,8
226	517,0	372,6	661,4	213,7
227	596,8	452,0	741,6	214,3
228	618,9	473,7	764,0	214,8

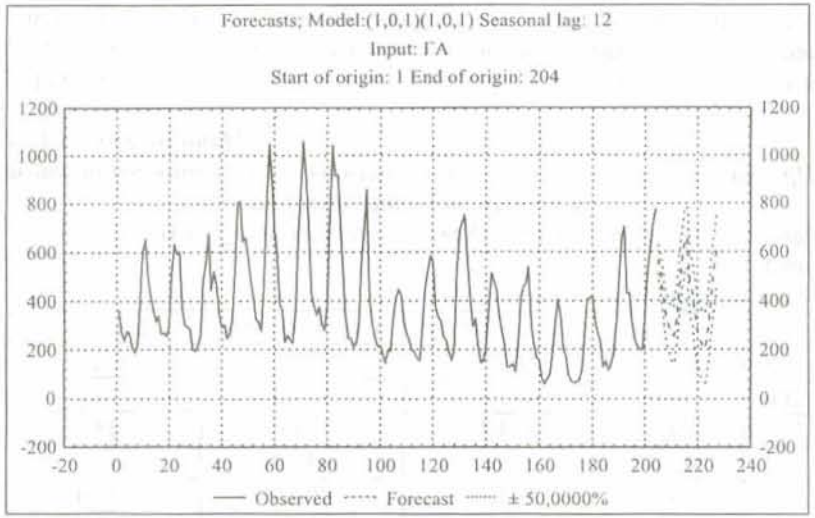

Рис.12.11.Прогностическая модель числа случаев заболевания гепатитом А на предстоящие 24 месяца.

Прогнозируемое число помесячных случаев заболевания гепатитом В на предстоящие 12 месяцев.

Forecasts; Model:(1,0,1)(1,0,1) Seasonal lag: 12 (tarop_1.sta)
Input: ГВ
Start of origin: 1 End of origin: 204

	Forecast	Lower 50,00%	Upper 50,00%	Std.Err.
205	247,45	234,16	260,74	19,67
206	257,98	243,70	272,26	21,13
207	262,40	247,19	277,60	22,50
208	247,29	231,21	263,37	23,80
209	260,99	244,08	277,90	25,02
210	250,77	233,07	268,46	26,19
211	255,29	236,84	273,74	27,31
212	251,11	231,93	270,28	28,38
213	251,74	231,87	271,62	29,41
214	252,39	231,84	272,94	30,41
215	262,23	241,02	283,43	31,38
216	268,45	246,61	290,29	32,32

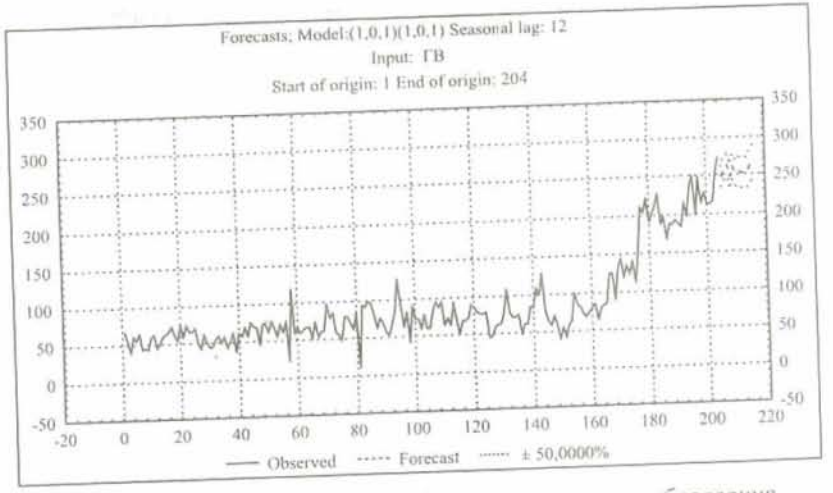

Рис.12.12.Прогностическая модель числа случаев заболевания гепатитом В на предстоящие 12 месяцев.

Из приведенных машинограмм и рисунков следует, что прогностические модели для гепатита А и гепатита В, построенные по данным наблюдения за 17 лет позволяют с достаточной точностью и надежностью прогнозировать уровень заболеваемости на 24 и 12 мес. Кроме этого следует иметь в виду, что регрессионная модель уровня заболеваемости гепатитом В (машинограмма 12.3) является статистически значимой и обладает вполне достаточной информационной способностью для надежного прогноза.

Литература

1. Боровиков В.П., Боровиков И.П. Statistica. Статистический анализ и обработка данных в среде Windows. - М.: Информационно-издательский дом "Филин". 1997. - 608 с.
2. Налимов В.В. Теория эксперимента. - М.: Наука. 1971. - 207 с.
3. Плескунин В.И., Воронина Е.Д. Теоретические основы организации и анализа выборочных данных в эксперименте. - Л. Изд. Ленинградского университета. 1979. - 232 с.
4. Торопов Д.Е. Эпидемиологическая характеристика острых вирусных гепатитов в Самарской области. Автореферат на соиск. уч. ст. канд. мед. наук. СПб 1997.
5. StatSoft, Inc. (2001). Электронный учебник по статистике. Москва, StatSoft. WEB: http://www.statsoft.ru/home/textbook/ default.htm.

ОГЛАВЛЕНИЕ

Юнкеров Виктор Иванович
Григорьев Степан Григорьевич

МАТЕМАТИКО-СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Издание Военно-медицинской академии
ИД № 02909 от 29.09.2000 г.

Компьютерная верстка С.Э. Ивановой

Подписано в печать 4.03.2002 г. Формат 60x84 1/16.
Объем 17,0 п.л. Тираж 1500 экз. Заказ № 67.

Отпечатано в типографии ООО «ИПК "Бионт"»
199026, Санкт-Петербург, Средний пр. ВО., д. 86,
тел. (812) 322-68-43