

ВОЕННО-МЕДИЦИНСКАЯ АКАДЕМИЯ

В.И.Юнкеров, С.Г.Григорьев

МАТЕМАТИКО-СТАТИСТИЧЕСКАЯ
ОБРАБОТКА ДАННЫХ
МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ



1892



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

©Юнкеров В.И., Григорьев С.Г., 2002
© ВМедА, 2002
© ЭЛБИ-СПб, 2002

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 поделиться 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:

$$\hat{y} = b_0 + b_1x_1 + \dots + b_kx_k$$

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 не количественного 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:

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

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

v_j - коэффициент j -го эффекта основного неколичественного фактора F_j ($j = 1, 2, \dots, k$);

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

факторов;

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

$\sum_{s=1}^p \beta_s(F_i) x_{is}$ - 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:

$$\text{ЛКФ} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k,$$

где b_0 - константа;

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

x_1, x_2, \dots, 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:

$$\hat{y} = \frac{\exp(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k)}{1 + \exp(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k)},$$

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

b_0 - константа;

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

x_1, x_2, \dots, 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 (категоризованные) 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;

M_o, M_e - 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;

A_s - коэффициент асимметрии переменной в выборке;

E_x - коэффициент эксцесса переменной в выборке;

H_0, H_1 - нулевая и альтернативная статистические гипотезы;

p - уровень значимости - вероятность нулевой гипотезы H_0 ;

1-p - доверительная вероятность альтернативной гипотезы H_1 ;

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 $p=0,05, p=0,01, p=0,001$;

$\chi^2, \chi_{05}^2, \chi_{01}^2, \chi_{001}^2$ - 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;

$\bar{P} \pm t_{0,5} m_{\bar{P}}$ - 95%-й доверительный интервал частоты признака в выборке;

δp - поправка Йетса для частоты признака;

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

$\text{Can } r$ - коэффициент канонической корреляции;

R - коэффициент множественной корреляции;

R^2 - коэффициент множественной детерминации;

a, b_0 - константы в уравнениях регрессии;

b_1, b_2, \dots, b_k - коэффициенты регрессии;

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

S_0 - средняя квадратичная ошибка прогноза по уравнению регрессии;

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

$m_{\hat{y}}$ - 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_{η} - показатель чувствительности метода диагностики;

P_{α} - вероятность ошибки диагноза первого рода;

P_{ϵ} - показатель специфичности метода диагностики;

P_{β} - вероятность ошибки диагноза второго рода;

P_{∞} - 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.

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 X_1, X_2, \dots, X_k 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 Y_1, Y_2, \dots, Y_l , 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:

X_1 - возраст - V , лет;

X_2 - время доставки - D , ч;

X_3 - частота пульса - P , уд./мин.;

X_4 - артериальное давление систолическое - ADS , мм рт. ст.;

X_5 - тип дыхания - TD , в баллах: 1 - нормальное, 2 - частое, 3 - патологическое;

X_6 - речевой контакт - R , в баллах: 1 - нормальный, 2 - нарушенный, 3 - отсутствует;

X_7 - анизокория - A , в баллах: 0 - нет, 1 - есть;

X_8 - светлый промежуток - SP , в баллах: 0 - нет, 1 - есть;

X_9 - кровопотеря - KP , мл.

Выходными параметрами могут быть:

Y_1 - возникновение осложнений - OSL , в баллах: 0 - нет, 1 - есть;

Y_2 - срок лечения - $SRLEC$, дней;

Y_3 - 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

$$n \times (k+1), \quad (1.1)$$

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

k - число входных контролируемых факторов;

l - число выходных параметров;

$(k+1)$ - 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 $83 \times (9+3)$, 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 сутки 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 аномальных 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).

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).

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 A_s (skewness);

- эксцесса также в натуральном и стандартизованном виде E_x (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

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

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:

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

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.

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

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 частотой 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:

$$m=1+3,32 \times \lg n, \quad (1.5)$$

– 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 $\bar{X} = M_0 = M_e$;
- симметричность отклонений от среднего значения;
- малые отклонения более вероятны, большие - менее вероятны;
- практические пределы отклонений от среднего значения $\pm 3S$ (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):

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

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:

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

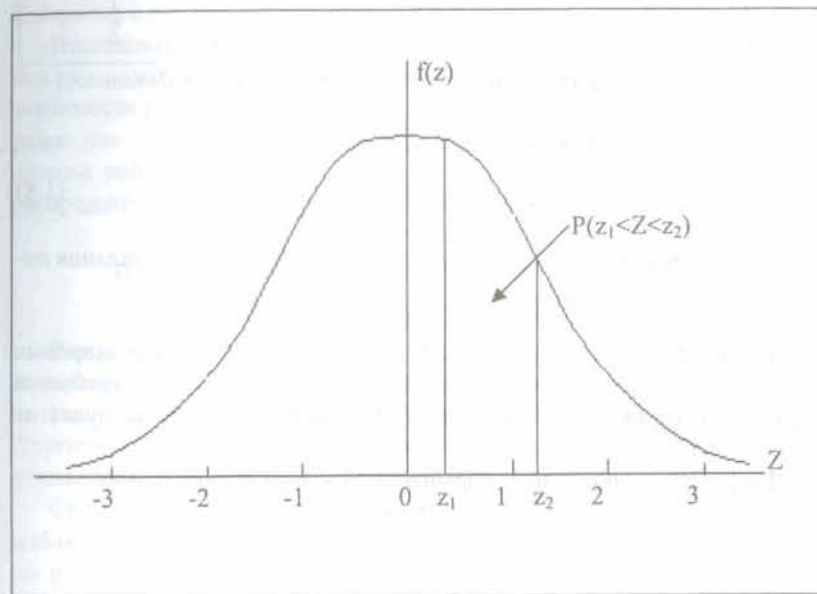


Рис.1.3.Кривая плотности нормального распределения.

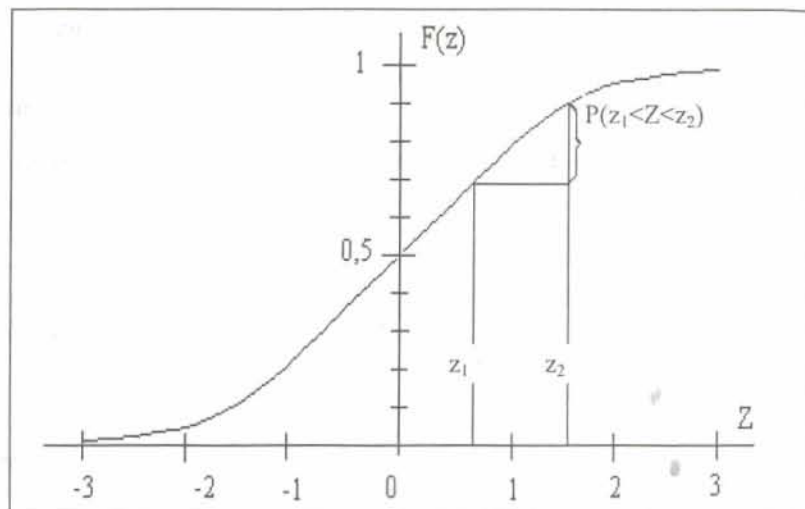


Рис. 1.4. Интегральная функция нормального распределения.

Интегральная функция нормального распределения:

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

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 H_0 (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. $p \leq 0,05$). В последнем случае исследователь должен подыскивать 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:

H_1 - о существенном различии показателя в опытной и контрольной группах;

H_0 - нулевую гипотезу - о равенстве (соответствии) показателя в опытной и контрольной группах.

Гипотезу H_1 принимают, если ее вероятность имеет значение равное или больше 95% и отклоняют, если ее вероятность будет меньше 95%. В этом случае принимают гипотезу H_0 , а ее вероятность, как альтернативной, будет $p > 0,05$.

Вероятность H_0 p называют уровнем значимости, а величину $1-p$ называют доверительной вероятностью гипотезы H_1 .

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 (приблизенно), 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: $n_1, \bar{x}_1, S_1, m_{\bar{x}_1}$, 95%-й доверительный интервал

$$M_1 = \bar{x}_1 \pm t_{0,95} \times m_{\bar{x}_1};$$

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

$$M_2 = \bar{x}_2 \pm t_{0,95} \times m_{\bar{x}_2}.$$

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 H_0 о соответствии средних значений показателя. Формулы расчета t даны в [3, 4, 5] и др.

Надежные значения t-критерия можно получить по формуле:

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

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

При $p \leq 0,05$ - различие значимо; при $p > 0,05$ - различие незначимо.

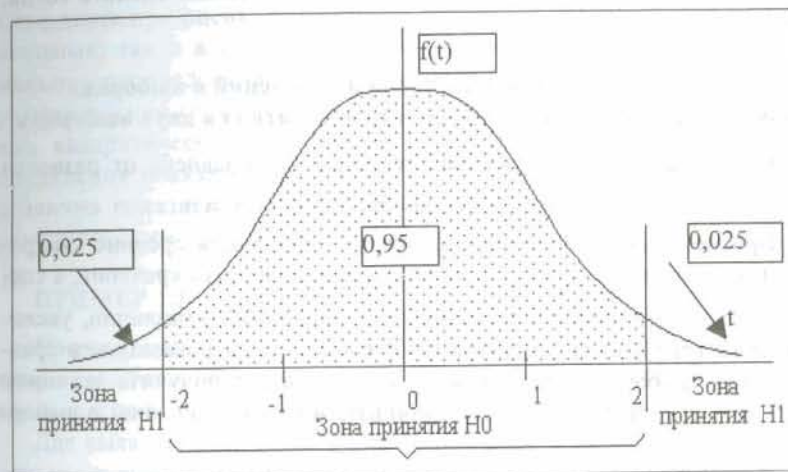


Рис.1.5. Величина двухстороннего уровня значимости $p = 0,025 + 0,025 = 0,05$ отвечающего критерию $t_{0,95}$.

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 показателей, получаемые по исходной матрице наблюдений. Расчет t-критерия проводится по формуле:

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

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

$m_{\Delta x}$ – средняя квадратичная ошибка средней разности показателя,

$m_{\Delta x} = \frac{S_{\Delta x}}{\sqrt{n}}$, 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 $p \leq 0,05$.

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 p .

При небольших объемах выборки увеличиваются средние квадратичные ошибки $m_{\bar{x}_1}, m_{\bar{x}_2}, m_{\Delta x}$ и уменьшается величина t-критерия, а следовательно уменьшается вероятность гипотезы H_1 о различии, увеличивается вероятность гипотезы H_0 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

$$n_1 \text{ и } n_2 \geq \frac{t_{05}^2 \times (S_{x_1}^2 + S_{x_2}^2)}{(\bar{x}_1 - \bar{x}_2)^2}; \quad (1.11)$$

– при сравнении связанных выборок

$$n \geq \frac{t_{05}^2 \times S_{\Delta x}^2}{(\Delta x)^2}. \quad (1.12)$$

Например, при получении незначимого различия показателя в независимых выборках:

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

$$2) \bar{x}_2 = 12, S_2 = 6, n_2 = 25,$$

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

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

Таким образом, в каждой выборке надо иметь не менее 61 наблюдения. В этом случае возможно принятие гипотезы H_1 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 экстракстол 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
Экссесс	-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) двустороннее	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

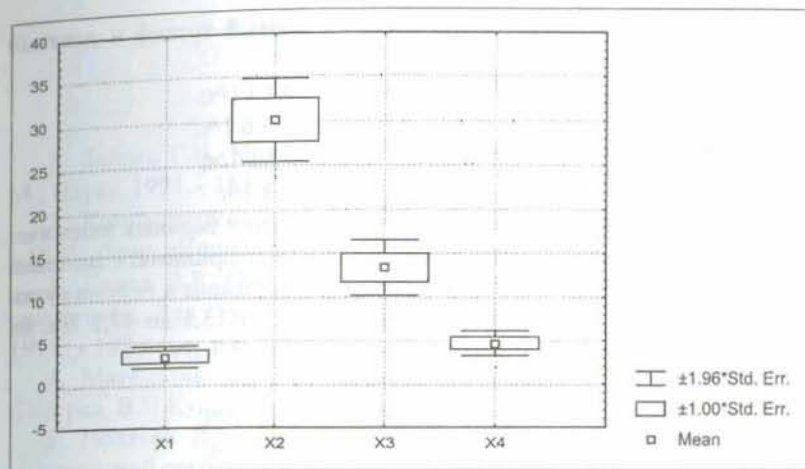


Рис.1.6. Средние значения показателя X с указанием 95% доверительных интервалов.

Выводы:

1. Из машинограммы 1.1 следует, что желудочковая экстрасистолия является патогномичным 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:

$$M_1=3,27\pm 2,14\times 0,64=1,9 - 4,64 \text{ 1/ч;}$$

$$M_2=30,2\pm 2,26\times 2,39=24,8 - 35,6 \text{ 1/ч;}$$

$$M_3=13,4\pm 2,26\times 1,63=9,7 - 17,1 \text{ 1/ч;}$$

$$M_4=4,7\pm 2,26\times 0,73=3,1 - 6,3 \text{ 1/ч.}$$

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.

$$X_1=3,27\pm 2,14\times 2,49=0 - 8,6 \text{ 1/ч;}$$

$$X_2=30,2\pm 2,26\times 7,55=13,1 - 47,3 \text{ 1/ч;}$$

$$X_3=13,4\pm 2,26\times 5,15=1,8 - 25,0 \text{ 1/ч;}$$

$$X_4=4,7\pm 2,26\times 2,31=0 - 9,9 \text{ 1/ч.}$$

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, продцимилле и т.д. Наиболее важной является относительная величина частоты случаев заболеваний, госпитализации, увольнения, смертности, дней нетрудоспособности 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):

$$I = \frac{1000 \times m}{n}, \quad (2.1)$$

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):

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

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:

$$I = \frac{365000 \times 2}{200 \times 80} = 45,7 \text{ ‰}.$$

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)

$$\bar{P} = \frac{m}{n} \quad (2.3)$$

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 $\bar{P}=0\%$, либо $\bar{P}=100\%$. Такой результат при числе наблюдений до 10 должен квалифицироваться случайным и требующим коррекции на поправку, обоснованную Йетсом:

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

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

$$m_p = \sqrt{\frac{\bar{P} \times (100 - \bar{P})}{n}}, \text{ (в \%)} \quad (2.5)$$

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

$$P = \bar{P} \pm t_{95} \times m_p \quad (2.6)$$

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 $25 \leq \bar{P} \leq 75\%$. Такой расчет доверительного интервала для относительной величины частоты летальных исходов в группе №3 дан в таблице 2.2. В случае $\bar{P} \leq 25\%$ или $\bar{P} \geq 75\%$ более точная оценка точности и надежности дается с применением вспомогательной переменной Фишера в радианной мере:

$$\phi = 2 \arcsin \sqrt{\bar{P}} \quad (2.7)$$

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

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

$$m_\phi = \frac{1}{\sqrt{n}} \quad (2.8)$$

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

$$\Phi = \phi \pm t_{95} m_\phi \quad (2.9)$$

От рассчитанных нижней и верхней границ доверительного интервала ϕ_n и ϕ_v переходят к соответствующим границам для относительной величины частоты по (2.10).

$$P_n = \sin^2 \frac{\phi_n}{2}; \quad P_v = \sin^2 \frac{\phi_v}{2} \quad (2.10)$$

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 выборках 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, \bar{P}_1, m_{P_1} , 95% доверительный интервал ($P_{1л}, P_{1в}$).

Выборка №2: n_2, \bar{P}_2, m_{P_2} , 95% доверительный интервал ($P_{2л}, P_{2в}$).

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)

$$t = \frac{|\bar{P}_1 - \bar{P}_2|}{\sqrt{m_{P_1}^2 + m_{P_2}^2}} \quad (2.11)$$

или по (2.12) при применении вспомогательной переменной Фишера

$$t = \frac{|\varphi_1 - \varphi_2|}{\sqrt{m_{\varphi_1}^2 + m_{\varphi_2}^2}} \quad (2.12)$$

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 $t_{0,05}, t_{0,01}, t_{0,001}$, 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 $t < t_{0,05}$ - различие незначимо ($p > 0,05$).

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

Статистическая значимость различия возрастает, если $t > t_{0,01}$ или $t > t_{0,001}$, 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)

$$n_{гр} \geq \frac{t_{0,05}^2 \times [\bar{P}_1 \times (1 - \bar{P}_1) + \bar{P}_2 \times (1 - \bar{P}_2)]}{(\bar{P}_1 - \bar{P}_2)^2}, \quad (2.13)$$

где \bar{P}_1 и \bar{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 (частоты) летальных исходов; оценить их точность и надежность.
2. Построить график частоты летальных исходов с указанием 95% доверительных интервалов.
3. Определить уровни значимости различия частоты 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 $\delta\rho$.

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 : $\phi_{II} - \phi_{IV}$	0,175- 0,509	0,716 -1,28		1,168-3,432
95% доверительный интервал для P : $P_{II} - P_{IV}$, %	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 $\phi_1 = 2 \arcsin \sqrt{0,029} = 0,342$

Средняя квадратическая ошибка $m_{\phi_1} = \frac{1}{\sqrt{140}} = 0,085$.

95% доверительный интервал для ϕ_1 (при $t_{95}=1,96$):
 $\phi_{1II} = 0,342 \pm 1,96 \times 0,085 = 0,342 \pm 0,167$, или $\phi_{1II} = 0,175$, $\phi_{1IV} = 0,509$.

95% доверительный интервал для P_1 :

$P_{1II} = \sin^2 \frac{0,175}{2} = 0,008$, или 0,8%, $P_{1IV} = \sin^2 \frac{0,509}{2} = 0,063$, 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
 $\phi_2 = 2 \arcsin \sqrt{0,229} = 0,998$.

Средняя квадратическая ошибка $m_{\phi_2} = \frac{1}{\sqrt{48}} = 0,144$.

95% доверительный интервал для ϕ_2 (при $t_{95}=1,96$):
 $\phi_{2II} = 0,998 \pm 1,96 \times 0,144 = 0,998 \pm 0,282$, или $\phi_{2II} = 0,716$, $\phi_{2IV} = 1,280$.

95% доверительный интервал для P_2 :

$P_{2II} = \sin^2 \frac{0,716}{2} = 0,122$, или 12,2%, $P_{2IV} = \sin^2 \frac{1,280}{2} = 0,354$, 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 частоты

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

$$P_3 = 0,375 \pm 2,36 \times 0,171 = 0,375 \pm 0,404.$$

95% доверительный интервал для P_3 :

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

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

$$\varphi_4 = 2 \arcsin \sqrt{0,833} = 2,300, \quad m_{\varphi_4} = \frac{1}{\sqrt{3}} = 0,577.$$

95% доверительный интервал для φ_4 :

$$\varphi_4 = 2,300 \pm 1,96 \times 0,577 = 2,300 \pm 1,132,$$

или $\varphi_{4н} = 1,168, \quad \varphi_{4в} = 3,432.$

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 частоты 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.

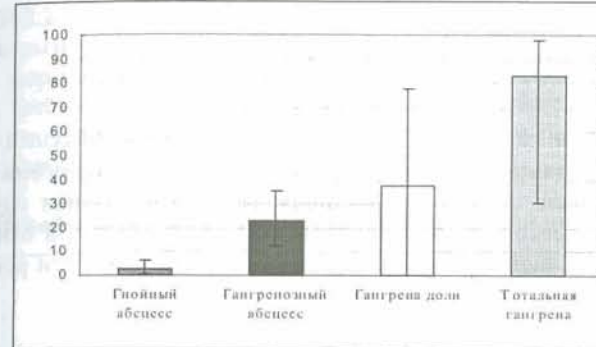


Рис.2.1. График относительных величин частоты летальных исходов для 4 форм заболеваний с указанием 95% доверительных интервалов.

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

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

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

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

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

n_{2i} - теоретическое (рассчитанное, как среднеожидаемое) число случаев признака 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 становится равной или больше 95% считают, что различие групп по данному критерию значимо.

Решение получают, сравнивая рассчитанное значение χ^2 с критическими значениями $\chi_{0,05}^2, \chi_{0,01}^2, \chi_{0,001}^2$, 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 $\chi^2 < \chi_{0,05}^2$ - различие групп по данным признакам незначимо ($p > 0,05$);

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

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.

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

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

-при $p=0,05$ $\chi_{0,05}^2 = 7,82$;

-при $p=0,01$ $\chi_{0,01}^2 = 11,34$;

-при $p=0,001$ $\chi_{0,001}^2 = 16,27$.

Так как $\chi^2 > \chi_{0,001}^2$ 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

Группа	Форма заболевания	Число летальных исходов			Число выздоровевших			Число больных
		наблюда- шеся n_{1j}	теорети- ческое p_{2j}	$(n_{1j}-p_{2j})^2$ p_{2j}	наблюда- шеся n_{1j}	теорети- ческое p_{2j}	$(n_{1j}-p_{2j})^2$ p_{2j}	
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 нормально-го 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 T-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 непараметрических 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.

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 ($p>0,05$).

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 В₁₂ 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).

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

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

Расчет χ^2 -критерия Пирсона

Таблица 3.7

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 $p=0,05$ $\chi_{05}^2=3,84$;

$p=0,01$ $\chi_{01}^2=6,64$;

$p=0,001$ $\chi_{001}^2=10,83$.

Так как $\chi^2 > \chi_{001}^2$ 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 2x2 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 $p=0,0001$.

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, ока-

зывает 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.

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 ‰ (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 ‰ 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).

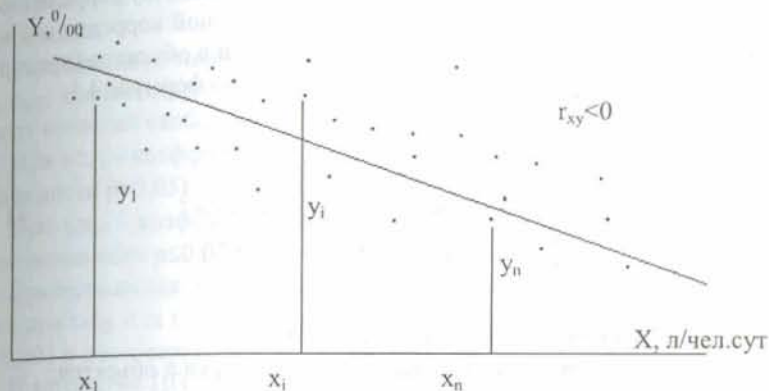


Рис.4.1. Поле наблюдений ($i = \overline{1, n}$) при обратной корреляционной связи между фактором X и параметром Y .

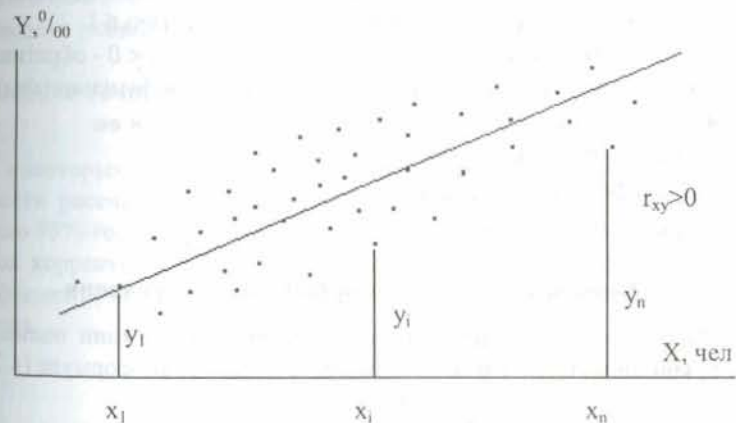


Рис.4.2. Поле наблюдений ($i = \overline{1, n}$) при прямой корреляционной связи между фактором X и параметром Y .

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 $n \times 2$ (см. пример 4.1).

Свойства коэффициента корреляции.

1. Коэффициент корреляции величина относительная; он принимает значение от минус единицы до плюс единицы, т.е. $-1 \leq r_{xy} \leq 1$.

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 $0,3 \leq |r_{xy}| \leq 0,7$ - 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 $t_{0,05}, t_{0,01}, t_{0,001}$, соответствующими уровням значимости $p = 0,05; 0,01; 0,001$ и числу степеней свободы $n' = n - 2$.

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

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

Статистическая значимость коэффициента корреляции увеличивается при $t > t_{0,01}$ или $t > t_{0,001}$ (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 aberrантных 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

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

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

$Z=1,333 \pm 1,96 \times 0,234 = 0,857 \div 1,809$.

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 $m_r = \sqrt{\frac{1-0,87^2}{20-2}} = 0,116$; 95%-й доверительный интервал при $t_{05}=2,10$ (по $n'=20-2=18$) $r_{xy}=0,87 \pm 2,10 \times 0,116 = 0,63 \div 1,11$.

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

При нелинейности связи между признаками, отсутствии данных о характере их распределения, небольшом числе наблюдений сравнимых 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 Кендалла.

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 .

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

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:

$$a = \bar{y} - b\bar{x}, \quad (4.9)$$

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

Коэффициент регрессии рассчитывается по формуле:

$$b = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad (4.10)$$

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

$$t_a = \frac{a}{m_a}; t_b = \frac{b}{m_b} \quad (4.11)$$

и соответствующих им уровней значимости.

Значимыми признаются коэффициенты с уровнем значимости $p \leq 0,05$ (достоверностью $1-p \geq 0,95$).

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:

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

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

$$R^2 = \frac{SS_R}{SS}, \quad (4.12)$$

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

$$F = \frac{S_R^2}{S_0^2}, \quad (4.13)$$

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 $p \leq 0,05$.

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 $p=1,07 \times E-09$. Из этого следует, что модель информативна ($R^2 > 0,5$) и значима (достоверность

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

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.