Валуев В.Е., Волчек А.А., Пойта П.С., Шведовский П.В.

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПРИРОДОПОЛЬЗОВАНИИ

Допущено Министерством образования Республики Беларусь в качестве учебного пособия для студентов высших учебных заведений по специальности «Мелиорация и водное хозяйство»

Брест 1999

ББК 20.1 - 05 я 73
В 15
УДК 91 (476) (075.8)+519.95 : 330.115+631.6

Статистические методы в природопользовании. Учебное пособие для студентов высших учебных заведений по специальности "Мелиорация и водное хозяйство".

В.Е. Валуев, профессор, к. т. н.
А.А. Волчек, доцент, к. г. н.
П.С. Пойта, доцент, к. т. н.
П.В. Шведовский, профессор, к. т. н.

Брест: Брестский политехнический институт, 1999.-252 с., илл.

В учебном пособии освещаются теоретические и прикладные вопросы географического (природного) прогнозирования, использования методов математической статистики при решении задач из области природопользования. Пособие предназначено для студентов высших учебных заведений по специальности "Мелиорация и водное хозяйство". Представит определенный интерес для практических работников в области водохозяйственного строительства, экологов, биологов. Может служить пособием для преподавателей вузов, аспирантов и студентов специальностей - водоснабжение, водоотведение, очистка природных и сточных вод, других природопользовательских и географических профилей.
Табл. 65, илл. 24, библ. 38 назв.

Рецензенты: Профессор кафедры почвоведения и геологии Белорусского государственного университета - д. г. н., проф. Н.К. Чертко; Заведующий лабораторией физики Земли института геологических наук Национальной академии наук Беларуси д. г.-м. н., проф. Г.И. Каратаев.

ISBN 985-6584-02-7

ББК 20.1 - 05 я 73

Валуеў У.Я., Воўчак А.А., Пойта П.С., Швядоўскі П.У.
Статыстычныя метады ў прыродакарыстанні. Вучэбны дапаможнік для студэнтаў вышэйшых навучальных устаноў па спецыяльнасці "Меліярацыя і водная гаспадарка". - Брэст: Брэсцкі політэхнічны інстытут, 1999.- 252 с., іл.

У вучэбным дапаможніку асвятляюцца тэарэтычныя і прыкладныя пытанні геаграфічнага (прыроднага) прагназіравання, выкарыстыння метадаў матэматычнай статыстыкі пры рашэнні задач з галіны прыродакарыстання. Дапаможнік прызначаецца студэнтам вышэйшых навучальных устаноў па спецыяльнасці меліярацыя і водная гаспадарка. Выклікае пэўную цікавасць у практычных работнікаў у галіне водагаспадарчага будаўніцтва, эколагаў, біёлагаў. Можа служыць дапаможнікам для выкладчыкаў ВНУ, аспірантаў і студэнтаў спецыяльнасцей - водазабеспячэнне, водаадвядзенне, ачыстка прыродных і сцекавых вод, іншых прыродакарыстальніцкіх і геаграфічных профіляў.
Табл. 65, іл.24, бібл. 38 назв.

Vladimir Valuyev, Alexandr Volchek, Pyotr Poyta, Pyotr Shvedovsky.
Statistical methods in the use of natural resources. The manual for college students of higher educational establishments of soil conservation and water industries. - Brest : Brest Polytechnic Institute, 1999. - 252 p., ill.

In this manual theoretical and applied matters of geographical prognostication, application of mathematical and statistical methods to solving problems in the field of the use of natural resources are dealt with. This manual is meant for students of higher educational establishments specializing in land improvement and water industries. It can also be of interest for practical workers in the field of water industry engineering, for ecologists, biologists. It can be used as a teaching aid for teachers of higher educational establishments, post-graduates and students of the following specialties: water-supply and drainage engineering, water and waste water purification and other similar disciplines.
Tabl. 65, ill. 24, bibl 38 nam.

ОГЛАВЛЕНИЕ

ОСНОВНЫЕ БУКВЕННЫЕ ОБОЗНАЧЕНИЯ

ξ	- истинная величина
x	- измеренная (случайная) величина
δx	- абсолютная погрешность
δ	- относительная погрешность
μ	- математическое ожидание
$\bar{x}$	- эмпирическое, или выборочное среднее
n	- объем выборки
N	- объем генеральной совокупности
σ^2	- дисперсия случайной величины (x)
$\bar{S}^2$	- выборочная дисперсия
s	- среднеквадратическое отклонение
V	- коэффициент вариации
m_1, m_2, m_3, m_4	- статистические моменты, соответственно, первого, второго, третьего и четвертого порядков
$F(x)$	- функция распределения в точке (x)
t	- критерий Стьюдента
F	-распределение Фишера
ν, ℓ, ϑ	- число степеней свободы
r	- коэффициент парной корреляции
P	- символ вероятности события
α	- уровень значимости
β_0, β_1	- коэффициенты регрессии
b_0, b_1	- оценки параметров (β_0) и (β_1)

ВВЕДЕНИЕ

Интенсивное развитие материального производства, крупномасштабное хозяйственное освоение территорий, стихийный социально-технический прогресс и потребительский бум обусловили разноскоростное протекание естественных процессов в природно-ландшафтных комплексах, нарушение динамического равновесия человека и природы. Понятийно-логическая схема техноприродного объекта (рисунок 1) подтверждает тот факт, что изучение и прогнозирование изменений природной Среды под влиянием хозяйственной деятельности человека является актуальной и одной из важнейших проблем современной науки. Оптимизация антропогенных воздействий на природный комплекс сопряжена с решением конкретных задач, связанных с охраной окружающей Среды (Среды обитания человека) и рациональным использованием природных ресурсов в жизнедеятельности людей.

При этом, статистические методы в географических (природных) исследованиях находят самое широкое применение. Нередко эти методы служат единственным средством количественной оценки различных природных процессов, явлений, а также проявлений климата, погоды и аномалий. Исследователь имеет дело с многофакторностью формирования всех природных процессов и невозможностью учета, в полной мере, степени влияния определяющих факторов. Математическое описание таких явлений возможно лишь статистическими методами.

Однако, при географическом (природном) прогнозировании с использованием статистических методов, адаптированных к решению прикладных задач, студенты, изучающие комплекс специальных дисциплин по природопользованию, не имеют учебников и учебных пособий, вообще, или вынуждены изыскивать их аналоги, считающиеся библиографической редкостью.

Что касается пособий для общих курсов по теории вероятностей и математической статистики (число которых велико), то в них, как правило, не учитывается специфика использования статистических методов для природных процессов. Она связана, прежде всего, с наличием в экспериментальных данных внутрирядной связности, нарушающей принцип случайности отбора, а также с необходимостью учета, в некоторых случаях, неоднородности в их рядах. Эти и другие обстоятельства значительно ос-

ложняют описание и анализ информационных рядов классическими методами математической статистики. Общие курсы математической статистики, кроме того, довольно сложны по своему построению.

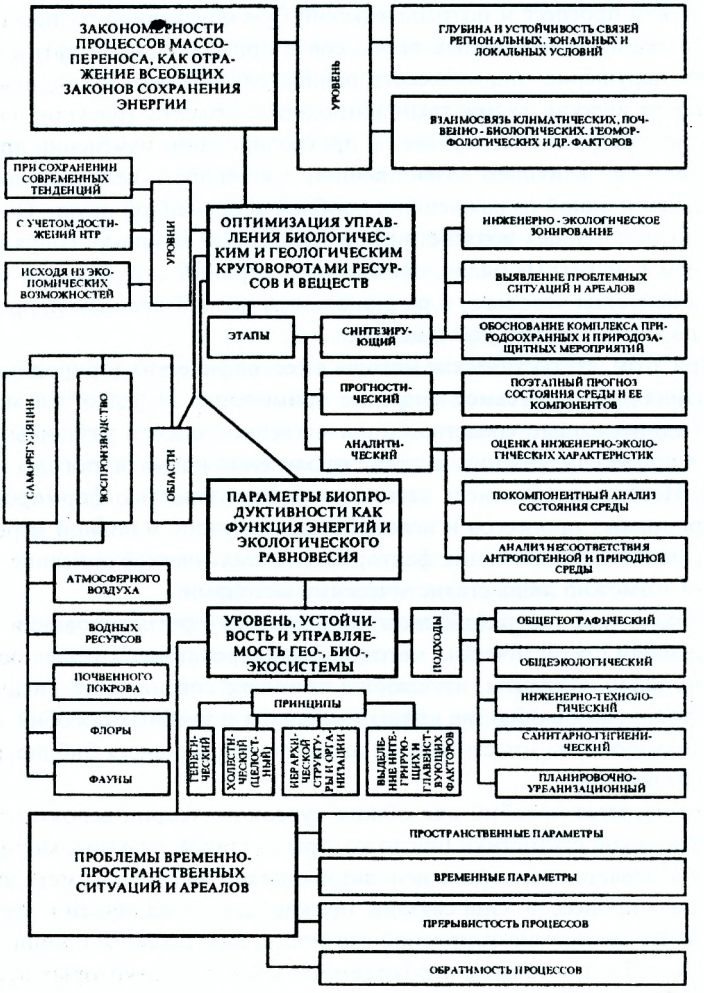


Рисунок 1 Понятийно-логическая схема техноприродного объекта.

Таким образом, необходимость составления данного пособия авторы видели в том, чтобы, исходя из достаточности материала и доступности его изложения, дать практическое руководство по прикладному использованию методов математической статистики для решения задач в области природопользования и географических наук.

С учетом практической направленности данного пособия, приоритет в изложении отдается описанию "технологии" статистических расчетов и интерпретации полученных результатов в их приложении к конкретным примерам. Теоретические аспекты используемых методов затрагиваются лишь в минимально необходимых пределах. При этом, предполагается, что пользователь имеет начальные знания из курса математической статистики.

В настоящее время имеет хождение большое количество компьютерных программ для статистического анализа информации (например, "Statgraphics", "Statistica", "Microsoft Excel" и другие). Эффективность использования программных продуктов при решении прикладных задач зависит не только от обеспеченности исполнителя исходными данными, но и от используемых им методов, влияющих на достоверность и точность расчетных характеристик.

Учебное пособие предназначено для студентов водохозяйственных и природопользовательских специальностей как практическое руководство в деле решения прикладных задач математической статистики и прогнозирования, органично входящих в циклы лабораторно-практических работ, в курсовые и дипломные проекты, в научное обоснование мелиоративных и водохозяйственных мероприятий.

Книга будет полезна практическим работникам в области водохозяйственного строительства, экологам, биологам и может служить пособием для преподавателей вузов, аспирантов и студентов смежных специальностей.

Авторы с благодарностью примут замечания, которые будут направлены на постоянное совершенствование программ читаемых курсов и связанного с ними учебного пособия.

1 ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Предварительная обработка результатов измерений (наблюдений) необходима для того, чтобы в дальнейшем с наибольшей эффективностью, а главное, корректно использовать статистические методы для построения эмпирических зависимостей, раскрывающих количественные и качественные связи в исследуемых процессах и явлениях.

Задача предварительной обработки в основном состоит в отсеивании грубых погрешностей, связанных с прямым измерением интересующей величины, или ошибок, неизбежно имеющих место при переписывании цифрового материала, вводе информации в ЭВМ и т.п.

Другим важным *моментом предварительной обработки* данных является *проверка соответствия* распределения результатов измерения (наблюдения) *закону нормального распределения*. Если эта *гипотеза неприемлема*, то следует *определить, какому закону* распределения *подчиняются* опытные *данные* и, если это возможно, *преобразовать* данное *распределение к нормальному*. Только *после выполнения* перечисленных выше *операций* можно *перейти к построению эмпирических формул*.

1.1 Элементы общей теории ошибок

Погрешности (ошибки) - неприменный *спутник любых измерений*. Несмотря на кажущуюся простоту и обыденность самого понятия "погрешность - ошибка", оно относится к разряду явлений, весьма сложных для теоретического осмысления и исключительно важных с точки зрения практических целей любого измерения. *Существует несколько* различных *подходов к классификации погрешностей*, частично перекрывающих друг друга, каждый из которых основан на рассмотрении отдельных аспектов понятия "погрешность". Ниже коротко перечислим *наиболее распространенные* варианты ***классификации погрешностей*** с указанием главного принципа, положенного в их основу:

1) **по способу выражения** (вычисления) погрешности принято делить на-

абсолютные $$\delta x = |x - \xi|, \qquad (1.1)$$

где x- измеренная величина; ξ - истинная величина;

относительные $$\delta = \pm \frac{\delta x}{x} \cdot 100\% ; \qquad (1.2)$$

2) **в зависимости от того, завышают или занижают** погрешности **результат** измерения (в среднем) по сравнению с истинным или средним значением, их можно подразделить на *положительные и отрицательные*. Примером систематического занижения истинных значений может быть измерение атмосферных осадков с помощью осадкомера, который дает значительные ошибки, особенно в зимние месяцы. Для устранения такого рода погрешностей вводятся специальные (поправочные) коэффициенты (на ветровой недоучет, смачивание осадкомерного ведра и испарение из него и т.п.);

3) **по типу связи между погрешностью и измеряемой величиной** различаются *постоянные*, значение которых не зависит от самой измеряемой величины, и *пропорциональные* погрешности, значение которых пропорционально измеряемой величине. Очевидно, что пропорциональные погрешности становятся постоянными при оценке их в относительной шкале;

4) **в зависимости от характера причин**, которые их вызывают, различаются *случайные, систематические* погрешности и *промахи* (грубые ошибки). Для оценки случайных погрешностей используются различные меры, - например, среднеарифметическая или среднеквадратическая погрешности, которым отвечают несколько отличные значения ширины интервала вариации при соответствующей доверительной вероятности;

5) **погрешность** может быть **оценена относительно** *единичного* измерения, среднего из нескольких параллельных определений ("серийная" или "генерализованная" погрешность) или ***относительно метода анализа, в целом,*** (погрешность метода);

6) **по источникам происхождения** погрешности (ошибки) подразделяются на *инструментальные, методические, методологические* и т.п.;

7) в зависимости от того, производится ли оценка **непосредственно измеряемой величины** или **величины, расчет которой опосредован** через ряд других экспериментальных величин с помощью определенной математической зависимости, различаются погрешности *прямых* и *косвенных* определений (измерений).

Опыт показывает, что ни одно *измерение*, как бы тщательно оно ни проводилось, *не* может быть совершенно *свободно от ошибок*. Случайные погрешности иногда можно надежно оценить, если повторить измерение несколько раз. И, естественно, предположить, что *наилучшей оценкой* из-

меряемой величины *будет среднее значение из всех измерений*. Кроме того, представляется довольно разумным предположение, что *правильное значение* измеряемой величины *лежит* где-то *между наименьшим и наибольшим значением*. Таким образом, *корректный способ* представления результата любого измерения *состоит в том, чтобы экспериментатор указывал* свою *наилучшую оценку* измеряемой величины *и интервал*, в котором, как он уверен, лежит это значение.

В общем случае, результат любого измерения величины (x) приводится как

$$x = \bar{x} \pm \delta x, \qquad (1.3)$$

где x - измеренная величина; $\bar{x}$ - большая - наилучшая оценка измеряемой величины, т.е., в нашем случае, среднее из всех измерений; δx - погрешность или ошибка в измерениях (x).

Это утверждение означает, что, во-первых, наилучшая оценка экспериментатора для измеряемой величины есть число ($\bar{x}$) и, во-вторых, он до определенной степени уверен, что эта величина лежит где-то между ($\bar{x} - \delta x$) и ($\bar{x} + \delta x$). Погрешность (δx) принято считать положительной величиной, так что ($\bar{x} + \delta x$) есть всегда наибольшее вероятное значение измеряемой величины, а ($\bar{x} - \delta x$) - наименьшее. К сожалению, для большинства научных измерений очень затруднительно сделать такое утверждение. В частности, если мы уверены в том, что измеряемая величина лежит между ($\bar{x} - \delta x$) и ($\bar{x} + \delta x$), то обычно необходимо выбрать для (δx) такое значение, которое слишком велико, чтобы представлять практический интерес. Чтобы избежать этого, мы можем иногда выбирать такое значение (δx), для которого вероятность того, что действительное значение лежит в заданном интервале будет равна, например, 67% или 95%. Однако, этого нельзя сделать без детального знания статистических законов, которым подчиняются процессы измерения, рассматриваемые ниже.

Следует отметить *несколько основных правил записи погрешностей. Во-первых*, поскольку величина (δx) служит оценкой погрешности, ее, очевидно, нельзя приводить с очень большой точностью. Например, глубину канала можно записать - h=1,85±0,05 м. *Во-вторых*, числа должны, как правило, содержать на одну значащую цифру больше, чем это оправдано. Это уменьшит неточности, возникающие при округлении чисел. *В конце расчета* окончательный результат следует округлить и избавиться от этой добавочной (и незначащей) цифры.

Общее правило приведения результатов выражается следующим образом: *последняя значащая цифра в любом приводимом результате обычно должна быть того же порядка (находиться в той же десятичной позиции), что и погрешность.*

Большинство физических величин обычно *невозможно измерять* непосредственно (суммарное испарение с почвы, испаряемость, расчетный расход воды в реке, незаиляющая скорость движения воды в мелиоративном канале и т. д.), и их определение включает два различных этапа. Сначала измеряют одну или более величин, которые могут быть непосредственно измерены, и с помощью которых можно вычислить интересующую нас величину. Затем, используя измеренные значения, вычисляют саму искомую величину. Кроме того, измеренные значения могут использоваться для косвенного определения искомого параметра или характеристики.

Если измерение включает эти два этапа, то и оценка погрешностей также включает их. Сначала надо оценить погрешности в величинах, которые измеряются непосредственно, а затем - определить, как эти погрешности "распространяются" в расчетах и приводят к погрешности, в конечном результате.

Рассмотрим ***погрешности*** при вычислении ***искомой величины по измеренным составляющим***:

а) погрешности в суммах и разностях

Если несколько величин (x,...,ω) измерены с погрешностями (δx,....,δω) и используются для дальнейшего вычисления

$$q = x+...+z-(u+...+\omega), \qquad (1.4)$$

то погрешность в рассчитанной величине (q) есть сумма погрешностей в определении всех составляющих

$$\delta q \approx \delta x+...+\delta z+\delta u+...+\delta\omega. \qquad (1.5)$$

Другими словами, когда складываются или вычитаются любые числа измеренных величин - *погрешности* в определении этих величин *всегда складываются*;

б) погрешности в произведениях и частных

Для удобства запишем числа с помощью относительных погрешностей- $x = \bar{x} \cdot (1 \pm \frac{\delta x}{|\bar{x}|})$ и $y = \bar{y} \cdot (1 \pm \frac{\delta y}{|\bar{y}|})$. Тогда наилучшая оценка произведения будет определяться как

$$q = \bar{x} \cdot \bar{y} \cdot (1 \pm \frac{\delta x}{|\bar{x}|}) \cdot (1 \pm \frac{\delta y}{|\bar{y}|}) . \qquad (1.6)$$

Наименьшее вероятное значение для (q) дает выражение со знаком (-) и наибольшее со знаком (+). Теперь результат произведения в скобках (1.6) может быть представлен для наибольшего вероятного значения

$$(1 + \frac{\delta x}{|\bar{x}|}) \cdot (1 + \frac{\delta y}{|\bar{y}|}) = 1 + \frac{\delta x}{|\bar{x}|} + \frac{\delta y}{|\bar{y}|} + \frac{\delta x \cdot \delta y}{|\bar{x}| \cdot |\bar{y}|} . \qquad (1.7)$$

Поскольку две относительные погрешности $(\frac{\delta x}{|\bar{x}|})$ и $(\frac{\delta y}{|\bar{y}|})$, - малые числа, то их произведение очень мало. Следовательно, последним членом в (1.7) можно пренебречь. Возвращаясь к (1.6), мы получим наибольшее вероятностное значение

$$\bar{q} = \bar{x} \cdot \bar{y} (1 + \frac{\delta x}{|\bar{x}|} + \frac{\delta x}{|\bar{y}|}) . \qquad (1.8)$$

Наименьшее вероятностное значение дается аналогичным выражением с двумя знаками минус. Таким образом, относительная погрешность (q) равна сумме относительных погрешностей $(\bar{x})$ и $(\bar{y})$

$$\frac{\delta q}{|\bar{q}|} \approx \frac{\delta x}{|\bar{x}|} + \frac{\delta y}{|\bar{y}|} . \qquad (1.9)$$

В общем случае, погрешность в произведениях и частных можно сформулировать следующим образом.

Если несколько величин (x,...,ω) измерены с малыми погрешностями (δx,...,δω) и измеренные значения используются для расчета

$$q = \frac{x \cdot \ldots \cdot z}{u \cdot \ldots \cdot \omega}, \qquad (1.10)$$

то относительная погрешность рассчитанной величины (q) равна сумме относительных погрешностей в (x,...,ω)

$$\frac{\delta q}{|q|} \approx \frac{\delta x}{|x|} + \ldots + \frac{\delta z}{|z|} + \frac{\delta u}{|u|} + \ldots + \frac{\delta \omega}{|\omega|}. \qquad (1.11)$$

Итак, при умножении (или делении) величин относительные погрешности складываются.

Рассмотрим *два важных частных случая применения правила* (1.10) - -(1.11).

Во-первых, предположим, что мы измеряем величину (x) и используем ее для вычисления произведения (q=bx) ,где (b) не содержит погрешности.

Сформулируем *общее правило.*

Если величина (x) измеряется с погрешностью (δx) и используется для вычисления произведения

$$q = b \cdot x^n, \qquad (1.12)$$

в котором (b) не имеет погрешности, то погрешность в (q) равна $|b|$ *, умноженному на погрешность в (x)*

$$\delta q \approx |b| \cdot \delta x. \qquad (1.13)$$

Второй частный *случай* применения правила (1.10)-(1.11) касается оценки степени некоторой измеряемой величины.

Погрешность при возведении в степень.

Если величина (x) измеряется с погрешностью (δx) и измеренное значение используется для вычисления степени этого числа

$$q = x^2, \qquad (1.14)$$

то относительная погрешность в (q) в n*-раз больше относительной погрешности в (x)*

$$\frac{\delta q}{|q|} = n\frac{\delta x}{|x|}. \qquad (1.15)$$

Приведенные формулы позволяют получить крайние пределы для (δq) и, очевидно, это может случиться, если мы недооценили (x) на полную величину (δx) и недооценили (y) на полную величину (δy). Однако, это весьма маловероятно. Если (x) и (y) измеряются независимо, и выявленные ошибки случайны по природе, то в 50% - х случаев недооценка (x)

будет сопровождаться переоценкой (y), или наоборот. Тогда ясно, что вероятность недооценки как (x), так и (y) на полные величины (δx) и (δy) довольно мала, следовательно, значение ($\delta q \approx \delta x + \delta y$) переоценивает нашу возможную ошибку. Если измерения (x) и (y) выполняются независимо, и если они оба подчиняются нормальному закону распределения, то погрешность в ($\bar{q} = \bar{x} + \bar{y}$) определяется выражением

$$\delta q = \sqrt{(\delta x)^2 + (\delta y)^2} \ . \qquad (1.16)$$

Таким образом, возвращаясь к вопросу погрешности при вычислении искомой величины по измеренным составляющим, можно сформулировать два главных правила:

а) погрешность в суммах и разностях

Если известно, что погрешности в (x,...,ω) независимы и случайны, то погрешность в сумме или разности (q) равна квадратичной сумме исходных погрешностей

$$\delta q = \sqrt{(\delta x)^2 + ... + (\delta z)^2 + (\delta u)^2 + ... + (\delta \omega)^2} \ . \qquad (1.17)$$

В любом случае, (δq) никогда не больше, чем их обычная сумма

$$\delta q \le \delta x + ... + \delta z + \delta u + ... + \delta \omega \ ; \qquad (1.18)$$

б) погрешности в произведениях и частных

Если погрешности (x,...,ω) независимы и случайны, то относительная погрешность в произведениях и частных (q) равна квадратической сумме исходных относительных погрешностей

$$\frac{\delta q}{|q|} = \sqrt{(\frac{\delta x}{\bar{x}})^2 + ... + (\frac{\delta z}{\bar{z}})^2 + (\frac{\delta u}{\bar{u}})^2 + ... + (\frac{\delta \omega}{\bar{\omega}})^2} \ . \qquad (1.19)$$

В любом случае, она никогда не больше, чем их обычная сумма

$$\frac{\delta q}{|q|} = \frac{\delta x}{\bar{x}} + ... + \frac{\delta z}{\bar{z}} + \frac{\delta u}{\bar{u}} + ... + \frac{\delta \omega}{\bar{\omega}} . \qquad (1.20)$$

Мы рассмотрели особенности оценки независимых и зависимых погрешностей для сумм, разностей, произведений и частных. Однако, инженерные расчеты сопряжены и с более сложными операциями, такими, как вычисление тригонометрических функций, квадратного корня и т.п. Обычно функция (q(x)) известна в явном виде, и погрешность (δq) может быть выражена аналитически, т.е.

$$\delta q = q(\bar{x} + \delta x) - q(\bar{x}) . \qquad (1.21)$$

Теперь, согласно основному приближенному выражению математического анализа, для любой функции (q(x)) и любого достаточно малого приращения (u) можно записать

$$q(x+u)-q(x)=\frac{dq}{dx}\cdot u. \qquad (1.22)$$

Таким образом, при условии, что погрешность (δx) мала, можно переписать (1.21) и получить

$$q(x)=\frac{dq}{dx}\cdot u. \qquad (1.23)$$

Итак, чтобы найти погрешность (δq), необходимо вычислить производную (dq/dx) и умножить ее на погрешность (δx). Следовательно, *любой расчет* может быть представлен как *последовательность* определенных *шагов, каждый из которых* включает только *одну из операций*:1) нахождение сумм и разностей; 2) расчет произведений и частных; 3) вычисление функции одной переменной.

Данный метод получил название в литературе "*шаг за шагом*", так как *вычисление* сложной *функции* может быть *разбито на* отдельные *элементы и погрешность* в рассчитываемой *функции можно оценить также последовательно,* используя рассмотренные правила. Однако, *имеется ряд задач, когда функция включает одну и ту же величину более чем один раз.* Поэтому, *некоторые из ошибок могут* взаимно *компенсироваться* (эффект, который иногда называют эффектом компенсирующихся ошибок). Если это происходит, расчеты погрешности методом "шаг за шагом" могут привести к переоценке конечной погрешности. *В этом случае*, обычно, *используют обобщенную формулу*, которая получается *аналогично (1.23).* Опуская сам вывод формулы, запишем конечный результат.

Погрешность функции нескольких переменных

Если погрешности в (x,...,z) независимы и случайны, то погрешность в (q) равна

$$\delta q=\sqrt{\left(\frac{\partial q}{\partial x}\cdot\delta x\right)^2+\ldots+\left(\frac{\partial q}{\partial z}\cdot\delta z\right)^2}. \qquad (1.24)$$

В любом случае, *она* никогда *не больше, чем обычная сумма*

$$\delta q\le\left|\frac{\partial q}{\partial x}\right|\delta x+\ldots+\left|\frac{\partial q}{\partial z}\right|\delta z. \qquad (1.25)$$

Важнейшей особенностью обобщенного правила является то, что *из него следуют все перечисленные выше правила.* Например, (q(x,y)=x+y), т.е. величина (q) просто равна сумме (x+y). Обе частные производные равны единице ($\partial q/\partial x=\partial q/\partial y=1$) и, тогда, используя выражение (1.24), получим $\delta q=((1\cdot\delta x)^2+(1\cdot\delta y)^2)^{0,5}=(\delta x^2+...+\delta y^2)$, как и по выражению (1.16). *Непосредственное использование общего правила* довольно *трудоемко и*, если возможно, *в расчетах проще продвигаться шаг за шагом, используя* более *простые правила.* Однако, если функция - q(x,...,z) включает любую переменную более одного раза, могут возникнуть компенсирующиеся ошибки, в этом случае, вычисления методом "шаг за шагом" могут, как отмечалось, привести к переоценке окончательной погрешности, и тогда лучше выполнять вычисления за один прием, непосредственно используя формулы (1.24) и (1.25).

В качестве примера, *рассмотрим определение площади живого сечения мелиоративного канала трапецеидальной формы (грунты устойчивые)*

$$\omega = (b + m\cdot h)\cdot h, \qquad (*)$$

где ω - площадь живого сечения потока воды в канале, м²; b - ширина канала по дну, м; h -глубина воды в канале, м; m -заложение откосов. Пусть результаты измерения живого сечения канала следующие: h=2,00±0,05 м; b=0,40±0,03 м; m=2.

Для нахождения ошибки вычисления воспользуемся формулой (1.24), т.е. $\delta\omega=((\partial\omega/\partial h\cdot\delta h)^2+((\partial\omega/\partial b\cdot\delta b)^2)^{0,5}$. *Возьмем производные уравнения (*) по (∂h) и (∂b), т.е.* $\partial\omega/\partial h=(b+2\cdot m\cdot h)$; $\partial\omega/\partial b=h$; *подставляя эти выражения в формулу (1.24), имеем* $\delta\omega=((b+2\cdot m\cdot h)\cdot\delta h)^2+((h\cdot\delta b)^2)^{0,5}$; *используя соответствующие численные значения, получим*

$\delta\omega=(((0,4+2\cdot2\cdot2)0,05)^2+((2\cdot0,03)^2)^{0,5}=0,17$ *м², а также* $\omega=(0,4+2\cdot2)\cdot2=8,80$ *м².*

Площадь живого сечения *мелиоративного канала, полученная расчетным путем с использованием материалов прямых измерений гидравлических элементов потока воды составляет*

$$\omega=8,80\pm0,17 \text{ м}^2.$$

1.2 Генеральная совокупность и выборка. Их числовые характеристики

Важнейшей *задачей статистической обработки* экспериментальных *данных является установление* (выявление) ряда статистических *параметров, которые* в комплексе достаточно *полно характеризуют свойства* исследуемой *генеральной совокупности. Генеральной совокупностью* называ-

ется *совокупность всех* возможных *данных* наблюдений, *которые могли быть получены* в соответствии с программой исследования. Общее число членов генеральной совокупности называется *объемом генеральной совокупности.* Число членов в генеральной совокупности может быть конечным или бесконечным.

***Например**, конечным числом членов генеральной совокупности являются все реки Беларуси (более 20 тысяч), по которым исследуется максимальный сток воды весеннего половодья. Бесконечным числом членов генеральной совокупности может быть расход воды реки Припять, величина которого колеблется по месяцам, годам, столетиям, тысячелетиям и т.д. В бесконечной генеральной совокупности максимальных расходов воды Припяти можно вычленить дискретные промежутки времени (определенное десятилетие или столетие), в которых экспериментальные данные по стоку могут рассматриваться как генеральная совокупность. Однако, полное исследование генеральной совокупности, обычно, практически невозможно или неэкономично. При расчете гидрологических характеристик, расход воды реки берется за определенный период (период наблюдений).*

С целью экономии времени и средств, *прибегают к подбору характерных ключей или точек*, пространственных или временных *ограничений, которые принято называть выборкой из генеральной совокупности. Выборочной совокупностью, или выборкой,* называется совокупность результатов n-наблюдений, полученных с целью характеристики генеральной совокупности. Число членов выборочной совокупности называется *объемом выборки.* С помощью выборки оценивается генеральная совокупность по вероятностным свойствам. Чтобы оценки были достоверными, выборка должна быть *представительной (репрезентативной)*, т.е. ее *вероятностные свойства должны совпадать* или *быть близкими к* свойствам *генеральной совокупности.* Репрезентативные совокупности формируются путем отбора: рандомизации*) (случайного), направленного (типичного) и смешанного.

При *случайном отборе* все объекты имеют одинаковую возможность попасть в выборку. Иногда случайная выборка не отвечает задачам исследования из-за неоднородности условий. Тогда производится *направ-*

*) - **рандомизация** - статистическая процедура, в которой решение принимается случайным образом.

ленный отбор типичных характеристик. Правила отбора, при этом, остаются те же, что при случайном отборе.

Смешанный отбор производится в тех случаях, когда необходимо получить характеристику неоднородного объекта, например, влагозапасов почвенного покрова на большой территории. Водосбор или сельскохозяйственное угодье делится на локальные участки, характеризующиеся однородными типами (подтипами) почв. Для каждого выделенного участка производится случайный отбор проб, и полученные величины почвенных влажностей объединяются в одну выборку.

Если иметь ввиду характеристики распределений вероятностей, то для теоретических распределений их можно рассматривать в контексте генеральной совокупности, а для эмпирических распределений - как выборочные характеристики. Можно встретить иную терминологию, когда характеристики распределения вероятностей в генеральной совокупности называются *параметрами*, а выборочных (эмпирических) значений - -*оценками* или *статистиками*.

Параметры обозначаются *буквами греческого алфавита*, а *оценки* - -соответствующими *буквами латинского алфавита*.

Наиболее полной характеристикой случайной величины является ее *функция распределения*. Как правило, это довольно сложный объект. Поэтому, *в ряде задач* при описании случайных величин *ограничиваются* простыми их *характеристиками* или иными *параметрами функций распределения*. Важнейшими из таких параметров (характеристик) являются *математическое ожидание* (μ) и *дисперсия* (σ^2) случайной величины (x).

Обычно, в задачах математической статистики параметры распределения случайной величины (x) неизвестны. В распоряжении исследователя имеется лишь выборка объемом (n). В этом случае, по выборке находятся выборочные параметры, которые и служат приближением к теоретическим - генеральным параметрам. Приближение тем лучше, чем больше объем выборки (n). В практических расчетах, без большой погрешности, можно принять, что уже при ($n>50$) выборочные параметры совпадают с генеральными.

Рассмотрим более подробно эти *параметры и их свойства. Математическое ожидание* - значение переменной, вокруг которого сгруппирована выборка; математическое ожидание непрерывной случайной величины задается интегралом

$$\mu = \int_{-\infty}^{\infty} x \cdot \varphi(x) \cdot dx, \qquad (1.26)$$

где $\varphi(x)$ - функция плотности распределения. Для дискретной конечнозначной случайной величины

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i , \qquad (1.27)$$

где x_i и p_i - отдельные значения и соответствующие им вероятности случайной величины

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad i = 1,n , \qquad (1.28)$$

следовательно, понятие математического ожидания совпадает с понятием среднего арифметического. Таким образом, $(\bar{x})$ представляет собой эмпирическое, или выборочное среднее.

Рассмотрим основные свойства математического ожидания:

1) математическое ожидание постоянной величины есть сама постоянная величина, т.е. $\mu(c)=c$;

2) математическое ожидание суммы постоянной и случайной равно сумме постоянной и математического ожидания случайной величины, т.е. $\mu(x + c) = \mu(x) + c$;

3) математическое ожидание суммы случайных величин равно сумме их математических ожиданий, т.е. $\mu(x+y)= \mu(x)+\mu(y)$;

4) математическое ожидание произведения постоянной величины на случайную равно произведению постоянной величины и математического ожидания случайной величины, т.е. $\mu(c \cdot x)=c \cdot \mu(x)$;

5) математическое ожидание произведения двух случайных величин равно произведению математических ожиданий этих величин, т.е. $\mu(x \cdot y) = \mu(x) \cdot \mu(y)$; *утверждение справедливо только для независимых случайных величин, когда каждая из них принимает то или иное значение независимо от того, какое значение приняла другая величина.*

Дисперсией (σ^2) случайной величины *называется мера "рассеивания"* или *"разброса"* выборки:

для *непрерывной случайной* величины (x_i) -

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot \varphi(x) \cdot dx; \qquad (1.29)$$

для *конечнозначной случайной* величины, имеющей N-значений,-

$$\sigma^2 = \sum_{i=1}^{n} p_i \cdot (x_i - \bar{x})^2 ; \qquad (1.30)$$

дисперсия выборочной совокупности, состоящей из n-значений случайной величины, вычисляется по формуле

$$\bar{S}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 . \qquad (1.31)$$

Эту величину принято называть *выборочной дисперсией* и обозначать символом ($\bar{S}^2$) - от англ. Standard. *Дисперсия* является удобной естественной мерой рассеивания случайной величины, поскольку в равной степени *учитывает отклонения* отдельных *результатов от среднего* как в большую, так в меньшую сторону и одновременно *усредняет их по всем результатам*.

Основные свойства дисперсии следующие:

1) *дисперсия постоянной величины равна нулю*, т.е. $\sigma^2(c)=0$;

2) *дисперсия суммы постоянной* величины и *случайной* величины *равна дисперсии случайной величины*, т.е. $\sigma^2(x+c)= \sigma^2(x)$;

3) *дисперсия случайной величины, умноженной на постоянный множитель, изменяется пропорционально квадрату* этого *множителя*, т.е. $\sigma^2(c \cdot x)= c^2\sigma^2(x)$, в частности, $\sigma^2(-x)= \sigma^2(x)$;

4) *дисперсия суммы двух переменных* величин *равна сумме дисперсий* этих величин, т.е. $\sigma^2(x+y)= \sigma^2(c) + \sigma^2(y)$; утверждение справедливо для независимых случайных величин.

Выборочное среднеквадратическое отклонение может быть найдено по формуле

$$\bar{S} = \sqrt{\bar{S}^2} . \qquad (1.32)$$

Из других выборочных характеристик чаще всего используется *коэффициент вариации* (v), являющийся мерой относительной изменчивости наблюдаемой случайной величины; вычисляется по формуле

$$v = \frac{\bar{S}}{\bar{x}} . \qquad (1.33)$$

Кроме того, часто используются *статистические моменты третьего и четвертого порядков:*

$$m_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^3 ; \qquad (1.34)$$

$$m_4 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^4 . \qquad (1.35)$$

Все сказанное выше относится к *равноточным* измерениям и наблюдениям, т.е. к измерениям, которые содержат только случайную погрешность, подчиняющуюся закону нормального распределения вероятностей. В практике обычно пользуются результатами измерений различного происхождения, в т.ч. выполненными разными инструментами и разноточными методами. Результаты таких измерений называются *неравноточными*. В этом случае, при статистической обработке экспериментальных данных используется метод получения средневзвешенных величин из имеющихся в наличии серий наблюдений.

В качестве примера, *выполним статистический анализ величин атмосферных осадков в контексте изучения режима их выпадения на территорию Беларуси. Для этого располагаем в алфавитном порядке 354 метеостанции и, используя из Приложения к настоящей книге таблицу "Случайные числа" (таблица П.1), выбираем 100 метеопунктов с соответствующими им годовыми нормами атмосферных осадков. Начав с любой колонки случайных чисел и двигаясь по столбцам сверху-вниз (снизу-вверх), выписываем те первые (последние) цифры четырехзначного числа, которые по величине укладываются в общее число метеопунктов и будут представлять номера метеопунктов, включаемых в формируемую выборку (таблица 1.1).*

Таблица 1.1 Выборка годовых норм атмосферных осадков Беларуси (мм)

736	719	788	719	735	720	695	813	814	739
731	777	787	746	733	785	799	763	786	781
753	824	736	792	763	772	753	721	769	699
722	782	716	734	800	760	756	731	707	686
706	695	737	768	784	758	658	684	676	744
841	784	756	763	750	751	727	812	731	828
747	731	690	836	749	773	625	774	726	679
753	667	713	663	758	768	655	708	703	661
656	630	658	638	677	673	676	706	716	737
741	758	696	684	747	642	671	739	664	624

Составив выборку годовых норм атмосферных осадков для территории Беларуси, определим основные характеристики данной выборки:

математическое ожидание *выборки, рассчитанное по формуле (1.28)-*

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i) = 73217/100 = 732{,}2 \text{ мм};$$

выборочную дисперсию-

***а) смещенную** (по формуле, аналогичной - 1.31)-*

$$S^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{100}\cdot 24462{,}1 = 244{,}6\ ;$$

***б) несмещенную** -*

$$\bar{S}^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{100-1}\cdot 24462{,}1 = 2471{,}3\ ;$$

***выборочные среднеквадратические отклонения**, получаемые по формулам типа (1.32)-*

***а) смещенное** -*

$$S = \sqrt{S^2} = \sqrt{2446{,}6} = 49{,}5\ ;$$

***б) несмещенное** -*

$$\bar{S} = \sqrt{\bar{S}^2} = \sqrt{2471{,}3} = 49{,}7\ ;$$

***коэффициент вариации**, рассчитываемый по формуле (1.33)-*

$$v = \frac{\bar{S}}{\bar{x}} = \frac{49{,}7}{732{,}2} = 0{,}07\ \ (7\%);$$

***основные статистические моменты** 1, 2, 3, 4 - порядков-*

$$m_1 = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{100}\cdot 0{,}0027 \approx 0\ ;$$

$$m_2 = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{100}\cdot 244662{,}1 = 2446{,}6\ ;$$

$$m_3 = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{100}\cdot(-1506860) = -15068{,}6\ ;$$

$$m_4 = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^4 = \frac{1}{100}\cdot 1500749000 = 15007490{,}0\ .$$

Полученные выше, основные (выборочные) статистические характеристики годовых величин атмосферных осадков для территории Беларуси сведены в таблицу 1.2.

Таблица 1.2 Выборочные характеристики распределения атмосферных осадков на территории Беларуси

$\bar{x}$	S^2	$\bar{S}^2$	S	$\bar{S}$	m_1	m_2	m_3	m_4	v
732,2	2446,6	2471,3	49,5	49,7	0	2446,6	-15068,6	15007490	0,07

***Таким образом**, величина среднего по территории Беларуси слоя осадков в средний многолетний год (норма) составляет 732,2 мм при коэффициенте вариации (изменчивости), равном 0,07 или 7 %.*

1.3 Эмпирические и теоретические распределения

Основными объектами изучения в теории вероятностей являются *события и случайные величины. Событие* - факт, который имеет место или может иметь место в ходе некоторого эксперимента. Выделяются события *случайные* и *достоверные*. *Случайным событием* называется то, которое при данных условиях может произойти или не произойти (например, выпадение дождя 22 июня 2001 года в городе Бресте). *Достоверное событие* - которое при данных условиях обязательно должно произойти (например, смена времен года).

Случайной величиной называется такая переменная величина, которая может принимать то или иное, заранее неизвестное значение (например, расход воды в реке в годовом или многолетнем ходе). Случайные величины, получаемые в эксперименте бывают *дискретными и непрерывными*. К *дискретным* (прерывным) относятся те, которые принимают конечное или бесконечное множество значений и между которыми нет и не может быть переходов (например, количество дней с осадками в месяц, количество озер на водосборе и др.). *Непрерывные* случайные величины могут принимать любые значения из некоторого конечного или бесконечного промежутка их значений. При рассмотрении непрерывных случайных величин, говорят не о конкретных значениях, а о промежутках и вероятности "попадания" в них. Между вариантами возможны различные переходы, все зависит от того, какая степень точности принимается для характеристики данного количественного признака (например, поступление коротковолновой радиации на сельскохозяйственное поле можно измерять с точностью до ±5%).

Случайные величины, представленные рядом количественных показателей, образуют *статистическую (выборочную) совокупность*. Каждый член этой совокупности называется *вариантой* или *датой*. Варианты в статистической совокупности подвергаются обработке. Для этого составляется *вариационный ряд*, т.е. варианты в нем располагаются в возрастающем или убывающем порядке. Варианты в выборке, относящиеся к одному и тому же признаку, практически не совпадают между собой, или *варьируются*. В вариационном ряду всегда есть максимальная (x_{max}) и минимальная (x_{min}) варианты. Разность между ними составляет *размах варьирования*, или *амплитуду изменчивости*. Те варианты, которые резко отличаются от вариант статистической совокупности и вызывают сомне-

ние у исследователя, определяются как *артефакт.* Артефакт, обычно, исключается из статистической совокупности. После анализа вариационного ряда на репрезентативность, приступают к статистической обработке экспериментальных данных.

Вариационный ряд может быть *представлен графически* в *виде полигона* (кривая распределения частот) или *гистограммы. При построении вариационной кривой* по оси абсцисс откладываются значения вариант или средин классов, по оси ординат - частоты. Величина *классового интервала* (i), которая зависит от принятого числа классов (k) и объема выборки (n) *определяется* как

$$i = \frac{x_{max} - x_{min}}{k}. \qquad (1.36)$$

Число классов, в зависимости от выборки, характеризуется формулой

$$k \approx 1 + 3{,}32 \cdot \lg(n). \qquad (1.37)$$

Исходя из зависимости (1.37), можно определить класс выборки по объему:

n	30...50	51...100	101...400	401...1000	1001...2000
k	4-6	6-8	8-9	9-11	11-12

Величина классового интервала должна быть одинаковой на протяжении всего вариационного ряда. *Границы классов* выбираются такими, *чтобы каждая варианта* могла быть отнесена только *к одному классу.* Первый и последний классы могут быть неполными. Границы классов желательно выбирать так, чтобы крайние варианты (x_{max}; x_{min}), по возможности, оказались ближе к середине интервала своего класса.

При построении гистограммы по оси абсцисс откладываются границы классов, а число вариант каждого класса обозначается высотой или площадью соответствующего прямоугольника. *При сравнении изменчивости* одинаковых условий или признаков, полученные вариационные кривые распределения частот наносятся на один график. *Группировка вариант* в классы для сравниваемых выборок *должна быть одинаковой.* Если объем выборок одинаков, все частоты должны быть выражены в процентах от объема выборки по каждой совокупности отдельно.

Гистограмма и *полигон* распределений являются графическим отображением частот, которые, в свою очередь, представляют собой оценки

плотностей вероятностей. *Кумулятивная линия* - график накопленных частот, в свою очередь, оценивающих функцию распределения F(x) в точке (x).

При работе с выборочной совокупностью, иногда необходимо описывать вариационную кривую с помощью математических функций, т.е. ряда математических зависимостей. *Описание распределения* проводится подбором подходящей математической модели функции распределения, с установлением параметров функции и проверкой ее соответствия эмпирическому распределению. При обработке вариационных рядов наблюдений за природными явлениями часто получаются колоколообразные полигоны распределения.

Если распределение случайной величины подчиняется определенному закону и может быть, хотя бы приближенно, описано кривой

$$Y_i = \alpha \cdot \exp(-b \cdot x^2), \qquad (1.38)$$

то такое распределение называется *нормальным (гауссовым)*. Так как к коэффициентам (α) и (b) предъявляется требование: α, b>0, то можно говорить о семействе кривых нормального распределения. С увеличением коэффициента (α), кривая "вытягивается" в высоту, при уменьшении - -кривая "сплющивается". Нормальное распределение обладает и другими важными свойствами, которые позволяют считать это распределение основой математической статистики.

Рассмотрим эти свойства:

1) ордината (y), которая характеризует высоту кривой для каждой точки оси (0x) (абсциссы), представляет собой плотность вероятности некоторого значения переменной (x) и определяется по следующей формуле

$$y_i = f(x_i) = \frac{1}{\sigma\sqrt{2\cdot\pi}} exp(-0{,}5(\frac{x-\mu}{\sigma})^2), (-\infty < x_i < +\infty, \sigma > 0) , \qquad (1.39)$$

где σ - среднеквадратическое отклонение теоретического распределения;
μ - математическое ожидание теоретического распределения. Из формулы (1.39) следует, что нормальное распределение полностью определяется величинами (μ) и (σ) (π=3,141593...; e=2,718282... - математические постоянные). Математическое ожидание определяет положение кривой распределения относительно оси (0x). Среднеквадратическое отклонение (σ) задает форму кривой. Чем больше (σ) (разброс данных), тем кривая положе (ее основание более широкое);

2) кривая нормального распределения симметрична относительно среднего значения, следовательно, среднее, мода и медиана совпадают;
3) максимум ординаты кривой определяется как - $y_{max}=1/\sigma\sqrt{2\pi}$, что, при $\sigma=1$, составляет примерно 0,4; если $x\to\pm\infty$, то $y\to 0$, другими словами, очень большие и очень малые значения переменной (x) маловероятны;
4) примерно 68,3% всех случаев наблюдений лежит в площади, отсекаемой перпендикулярами к оси (0x), ($\mu\pm\sigma$); соответственно, в пределах от (-2σ) до ($+2\sigma$) находятся 95,5% вариант, в пределах (-3σ) до ($+3\sigma$) - 99,7% (рисунок 1.1).

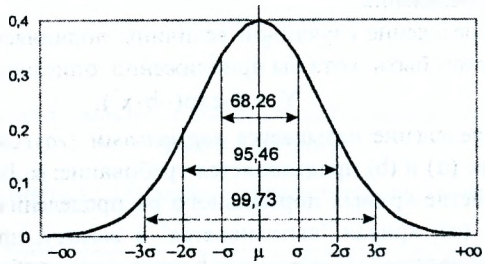


Рисунок 1.1 Процент случаев наблюдений (площадь), ограниченный кривой нормального распределения, для различных значений (σ).

Отклонение данных от нормального закона распределения указывает на влияние какого-либо другого фактора на статистическую совокупность.

При изучении природных процессов некоторые распределения имеют выраженную асимметрию. Поэтому, представляет практический интерес преобразование асимметричного распределения в симметричное (нормальное). Иногда это возможно, если каждую варианту выборки выразить в виде логарифма ($x_i'=\lg(x_i\pm\alpha)$). В тех случаях, когда логарифм случайной величины (x_i) подчиняется нормальному закону распределения, а сами значения случайных величин распределены асимметрично, распределение случайной величины принято называть *логарифмически нормальным*, или *логнормальным*.

В отдельных случаях, можно принимать и другие преобразования:

обратную величину ($x_i'=1/x_i$); *обратное значение квадратного корня* $x_i'=(x_i)^{-0,5}$.

Преобразование логнормальное используется для распределений, *у которых крутая левая ветвь* полигона и *пологая правая.*

Преобразование "*обратная величина*" является наиболее "сильным". Среднее положение между логарифметическим преобразованием и "обратной величиной" занимает преобразование "*обратное значение квадратного корня*". Для нормализации смещенного вправо распределения служат *тригонометрические преобразования*, а также *степенные преобразования* ($x_i'=x_i^{\alpha}$). При этом, $\alpha=1,5$ - при умеренном, $\alpha=2,0$ - при сильновыраженном правом смещении.

Закон нормального распределения проявляется при n>20...30. Однако, исследователь часто имеет дело с ограниченным числом данных и ему приходится основывать свои выводы на малых выборках. При небольшом числе наблюдений, результаты обычно близки и редко появляются большие отклонения. Это легко объяснить законом нормального распределения, согласно которому вероятность появления малых отклонений больше, чем отклонений значительных. Так, вероятность отклонений, превышающих по абсолютной величине (±2S), равна 0,05, или один случай на 20 вариант, а отклонений (±3S) - 0,01 (один случай на 100). Поэтому, *стандартное отклонение* (S), рассчитанное *по малой выборке*, в большинстве случаев будет *меньше, чем по всей генеральной совокупности* (σ). Следовательно, *в этих случаях полагаться на критерии нормального распределения* в своих выводах *нельзя.*

Огромная роль нормального закона в прикладных исследованиях определяется тем, что анализ многих природных явлений подтверждает правомерность его использования; даже в случаях, когда закон распределения вероятностей не может быть установлен теоретически, практика часто на его стороне. То обстоятельство, что в последнем случае нормальный закон распределения обосновывается не теоретическими построениями, а своим хорошим согласием со статистическими данными, относящимися к изучаемому явлению, не умаляет возможности использовать этот закон для дальнейшего анализа и расчетов. Но кроме всего этого, нормальный закон может служить основой для получения других законов распределения вероятностей, такие законы могут быть получены из нормального, если предположить, что нормальное распределение име-

ет не сама случайная величина (x), а некоторая ее функция (f(x)). Из обоснованных, таким образом, законов получил наибольшую известность, так называемый, логарифмически нормальный закон, применяемый для решения различных задач *природопользования*. Он получается из предположения, что нормально распределенной случайной величиной является не (x), а (Lnx). Естественно, что логарифмически нормальное распределение вероятностей может иметь только такая случайная величина (x), которая принципиально не может принимать отрицательных значений (например, расход воды в реке). Плотность логарифмически нормального распределения для ($x<0$) равна нулю, а для ($x>0$) она имеет вид

$$f(x) = \frac{1}{\gamma \cdot \sqrt{2 \cdot \pi}} \cdot \exp\left(- \frac{(\ln x - \ln b)^2}{2 \cdot \gamma^2} \right), \qquad (1.40)$$

где Lnb и γ^2 - суть математическое ожидание и дисперсия случайной величины (Lnx).

Параметры (b) и (γ^2) нетрудно выразить через математическое ожидание (μ) и дисперсию (σ^2) случайной величины (x):

$$b = \frac{\mu^{1.5}}{(\sigma^2 + \mu^2)^{\frac{2}{3}}}; \qquad (1.41)$$

$$\gamma^2 = \mathrm{Ln}\left(\frac{(\sigma^2 + \mu^2)^{\frac{2}{3}}}{\mu} \right). \qquad (1.42)$$

Логарифмически нормальное распределение, в отличие от нормального, имеет несимметричный график плотности вероятности (рисунок 1.2).

Большое практическое значение для экспериментальной работы имеет t-распределение, получившее название *распределение Стьюдента*, характеризуемое *для выборочных средних* равенством

$$t = \frac{\bar{x}_i - \mu}{\frac{S}{\sqrt{n}}}. \qquad (1.43)$$

Числитель формулы (1.43) означает *отклонение выборочной средней от средней всей совокупности (μ), а знаменатель является показателем оценки величины* ($\sigma / \sqrt{n} = \sigma_\mu$) *или стандартной ошибки средней* генеральной совокупности. Таким образом, величина (t) измеряется отклонением

выборочной средней ($\bar{x}$) от средней совокупности (μ), выраженным в долях ошибки выборки ($\bar{S}$), принятой за единицу.

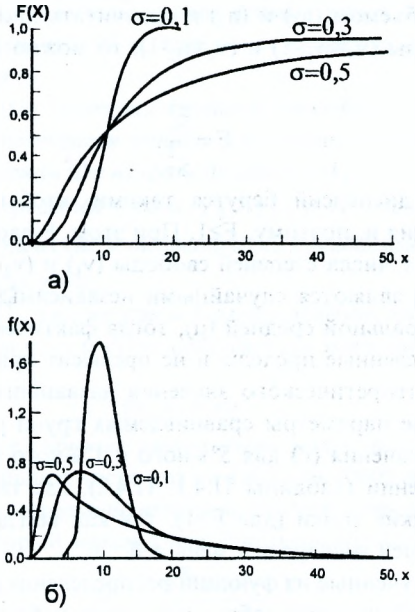


Рисунок 1.2 Логарифмически нормальное распределение вероятностей в дифференциальных (а) и интегральных (б) видах.

Распределение t - критерия Стьюдента представлено в Приложении (таблица П.2). Максимумы частот нормального и t-распределения совпадают, но форма кривой t-распределения зависит от числа степеней свободы. При очень малых значениях степеней свободы, она принимает вид плосковершинной кривой, причем площадь, ограниченная кривой, больше, чем при нормальном распределении, а при увеличении числа наблюдений (n>30), t-распределение приближается к нормальному и переходит в него при $n \to \infty$.

Распределение Стьюдента имеет важное значение при работе с малыми выборками и позволяет определить доверительный интервал, накрывающий среднюю совокупности (μ), а также проверить ту или иную гипотезу относительно генеральной совокупности. При этом, нет необ-

ходимости знать параметры совокупности (μ) и (σ), достаточно иметь их оценки ($\bar{x}$) и ($\chi^2_\Phi < \chi^2_T$) для определенного объема выборки.

Если из нормально распределенной совокупности взять две *независимые выборки* объемом (n_1) и (n_2) и подсчитать дисперсии (s_1^2) и (s_2^2) со степенями свободы ($\nu_1=n_1-1$) и ($\nu_2=n_2-1$), то можно определить *отношение дисперсий* (F)

$$F = \frac{\bar{S}_1^2}{\bar{S}_2^2}. \qquad (1.44)$$

Отношения дисперсий берутся такими, чтобы в числителе была большая дисперсия и, поэтому, $F \geq 1$. При этом, F-распределение Фишера зависит только от числа степеней свободы (ν_1) и (ν_2). Когда две сравниваемые выборки являются случайными независимыми (из общей совокупности) с генеральной средней (μ), тогда фактическое значение (F) не выйдет за определенные пределы и не превысит критического, для данных (ν_1) и (ν_2), теоретического значения названного критерия ($F_\phi \leq F_T$). Если генеральные параметры сравниваемых групп различны, то $F_\phi > F_T$. Теоретические значения (F) для 5%-ного и 1%-ного уровней значимости даны в Приложении (таблицы П.4.1, П.4.2), где табулированы только правые критические точки (для $F \geq 1$), так как всегда принято находить отношение большей дисперсии к меньшей.

Кривые, полученные из функции распределения для всех возможных значений (F), особенно при небольшом числе наблюдений, имеют асимметричную форму - длинный "хвост" больших значений и большую концентрацию малых величин.

Отметим, что *t-распределение Стьюдента является частным случаем F-распределения Фишера,* при числе степеней свободы ($\nu_1=1$) и ($\nu_2=\nu$), т.е. равно числу степеней свободы для t-распределения. В этом случае, наблюдается следующее соотношение между (F) и (t)

$$F_{(\nu_1=1,\nu_2)} = t^2_{(\nu=\nu_2)}, \text{ т.е. и } t = \sqrt{F}. \qquad (1.45)$$

Закон распределения χ^2 *(хи-квадрат)* открыл К. Пирсон. Кривая распределения, полученная из функции хи-квадрат имеет вид

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{f_i - F_i}{F_i} \right), \qquad (1.46)$$

где f_i - фактические и F_i - гипотетические частоты численности объектов выборки.

Ее вид в сильной степени зависит от числа степеней свободы (ν). Для малого числа степеней свободы кривая асимметрична, но с увеличением (ν) - асимметрия уменьшается и, при $\nu \to \infty$, кривая принимает вид нормальной гауссовской.

Критерий χ^2, *или критерий согласия (подобия), используется для оценки степени соответствия эмпирических данных* определенным *теоретическим предпосылкам или нулевой гипотезе* (H_0).

Гипотеза опровергается - если $\chi^2_\Phi \geq \chi^2_T$, и *не опровергается* - если $\chi^2_\Phi < \chi^2_T$. Когда *фактические и теоретически ожидаемые* частоты полностью *совпадают*, тогда $\chi^2=0$.

Распределение χ^2 , так же как и t-распределение, частной случайной F-распределения, при $\nu_1=\nu$ и $\nu_2=\infty$, описывается зависимостью

$$F_{(\nu_1,\nu_2=\infty)} = \frac{\chi^2}{\nu}. \qquad (1.47)$$

Когда некоторое *событие имеет очень малую вероятность свершения*, например, небольшое число раз на 1000 или 10000 обычных явлений, тогда *распределение случайной величины следует* определенному *закону редких событий*, который *выражается формулой Пуассона*

$$P_{x_i} = \frac{\alpha^{x_i} \exp(-\alpha)}{x_i!}, \qquad (1.48)$$

где P_{x_i} -вероятность значения (x_i); x_i- число редких событий, происходящих в каждой большой группе ($x_i=0,1,2,3$, и т.д.); α - среднее число редких событий на каждую большую группу; $x_i!$- произведение чисел от 1 до (факториал).

Распределение Пуассона определяется одним параметром - средней, и *дисперсия* этого распределения *равна средней*, т.е. $S^2=\alpha$.

Отсюда следует, *что все теоретические распределения можно построить только на основании одной выборочной средней.*

Распределение Пуассона является частным случаем биномиального распределения, когда в бином $(p+q)^n$ значение (p) очень мало, а (q) стремится к бесконечности. *Графически* распределение редких событий *представляет асимметричную кривую, и асимметрия тем больше, чем меньше вероятность события.*

Рассмотрим технологию построения полигона на примере *данных таблицы 1.1. Если данные таблицы 1.1 разделить на классы, то можно построить полигон и гистограмму частот. Разбивка на классы осуществляется по формуле (1.37), т.е.* $k=1+3,32 \cdot lg(n)=1+3,32 \cdot lg(100)=7,64$. *С другой стороны, разница между* (x_{max}) *и* (x_{mim}) *(размах варьирования) составляет 841-624=217 мм; исходя из этого, принимаем число классов, равным 8, со ступенями, равными 30 мм. Разбивка на классы приведена в таблице 1.3. Рассмотрим методику подсчета частот. Таблица 1.1 просматривается по порядку (от первой до последней строчки) и при чтении каждого результата соответствующая метка заносится в класс, к которому относится данное наблюдение.* ***Кумулятивная линия*** *(1),* ***гистограмма*** *(2) и* ***полигон*** *(3) распределений, построенные по данным таблицы 1.3,* ***даны на рисунках 1.2-1.3. Гистограмма и полигон*** *распределений* ***являются графическим отображением частот,*** *которые, в свою очередь,* ***представляют оценки плотностей вероятностей. Кумулятивная линия - график накопленных частот,*** *в свою очередь, оценивающих функцию распределения* $(F(x))$ *в точке* (x).

Таблица 1.3 Разбивка массива исходных данных на классы, вычисление частот

N п/п	*Классы (величины атмосферных осадков, мм)*	*Середины интервалов*	*Абсолютные частоты*	*Относительные частоты*	*Относительные накопленные частоты*
1	*2*	*3*	*4*	*5*	*6*
1	*610-640*	*625*	*4*	*0,04*	*0,04*
2	*641-670*	*655*	*9*	*0,09*	*0,13*
3	*671-700*	*685*	*15*	*0,15*	*0,28*
4	*701-730*	*715*	*21*	*0,21*	*0,49*
5	*731-760*	*745*	*26*	*0,26*	*0,75*
6	*761-790*	*775*	*16*	*0,16*	*0,91*
7	*791-820*	*805*	*7*	*0,07*	*0,98*
8	*821-850*	*835*	*2*	*0,02*	*1,00*
			$\Sigma 100$		

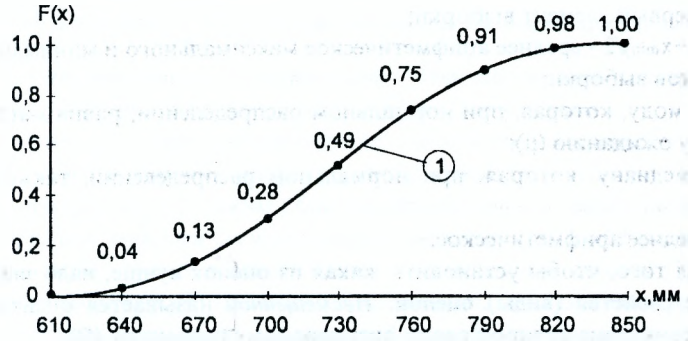


Рисунок 1.3 Кумулятивная кривая (1) распределения вариационного ряда норм атмосферных осадков Беларуси.

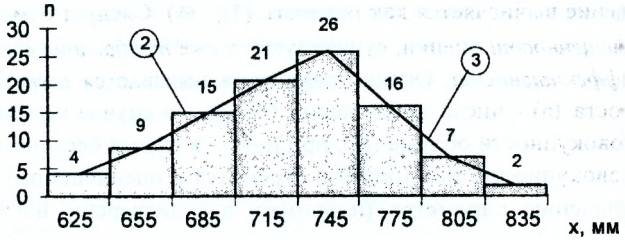


Рисунок 1.4 Графическое представление распределения вариационного ряда норм атмосферных осадков Беларуси: 2-гистограмма; 3-полигон.

1.4 Теория оценок

Генеральные совокупности описываются некоторыми *постоянными* числовыми ***характеристиками*** *распределения*. По выборкам *можно найти оценки* этих *характеристик*. Вследствие случайности выборок, *значения оценок* одной числовой характеристики, *вычисленные по разным выборкам* из одной генеральной совокупности, *бывают*, как правило, *различными*.

Обозначим неизвестный параметр распределения, т.е. *числовую характеристику* генеральной совокупности (x), *через* (Θ), а *оценку* неизвестного параметра - *через* (T_n). Оценка (T_n) - функция от выборки. Оценки неизвестного параметра можно находить различными способами. Например, если нужно оценить среднее значение ($\Theta=\mu$) нормального распределения, то *можно использовать следующие оценки* (T_n):

1) x_1 - первый элемент выборки;
2) $(x_{max}+x_{min})/2$ - среднее арифметическое максимального и минимального элементов выборки;
3) M_0 - моду, которая, при нормальном распределении, равна математическому ожиданию (μ);
4) M_e - медиану, которая, при нормальном распределении, также равна (μ);
5) $\bar{x}$ - среднее арифметическое.

Для того, чтобы установить, какая из оценок лучше, **надо знать** основные **свойства (виды) оценок**. *Несмещенной* называется оценка (T_n), среднее значение которой равно оцениваемому параметру (Θ)

$$\overline{T}_n = \mu \ . \qquad (1.49)$$

Если это условие не выполняется, то оценку называют *смещенной,* при этом смещение вычисляется как разность ($\overline{T}_n - \Theta$). Следует отметить, что *кроме несмещенности* оценки, существуют также *требования состоятельности и эффективности.* Оценка параметра называется *состоятельной,* по мере роста (n) - числа наблюдений ($n \to N$) - в случае конечной генеральной совокупности объема (N); при $n \to \infty$ - в случае бесконечной генеральной совокупности, она (оценка) стремится к оцениваемому теоретическому значению параметра (например, для дисперсии $\lim\limits_{n\to\infty} S^2(n) = \sigma$).

Следует четко разделять понятия состоятельности и несмещенности. Рассмотрим две оценки для дисперсии:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ ; \qquad (1.50)$$

$$\overline{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ . \qquad (1.51)$$

Эти оценки состоятельны, но вторая является несмещенной, первая - - смещенной, так как в ней содержится систематическая отрицательная погрешность - (σ^2/n), поскольку математическое ожидание $\left(\overline{S}^2 = (\sigma^2 - \sigma^2)/n\right)$, с ростом (n) монотонно убывает. Из этого следует, что требование несмещенности особенно важно при малом количестве данных наблюдений. Оценка параметра называется *эффективной,* если среди прочих оценок того же параметра она обладает наименьшей дисперсией. Если среднее оценивать как $(x_{max}+x_{min})/2$ (пункт 2, см. начало

параграфа) или как $\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$ (пункт 5), то *эти оценки обладают свойствами состоятельности и несмещенности* . Однако, можно показать, что *дисперсия во втором* случае оценки равна ($\sigma^2/2$), а в первом - $\pi^2\sigma^2/(24\cdot\ln(n))$, т.е. существенно больше, так как вторая оценка подвержена меньшим случайным колебанием вокруг неизвестного значения оцениваемого параметра. Таким образом, *второй способ оценки теоретического среднего является состоятельным, несмещенным и эффективным, а первый* способ - только *состоятельным и несмещенным. Оценки (T_n) неизвестного* параметра (Θ) *называются точечными*, так как они определяют одно значение, одну точку на числовой оси. *Все точечные оценки* параметров распределения генеральной совокупности *вычисляются по выборкам, но* из-за случайности выборок оценки *являются случайными величинами*, отличающимися от постоянного истинного значения параметра (Θ). Обозначим точность оценки через $\Delta(\Delta>0)$; $|\Theta - T_n| \le \Delta$. Чем меньше ($\Delta$), тем точнее оценки. Любую точность можно получить с определенной вероятностью (надежностью)

$$P(|\Theta - T_n| \le \Delta) = \gamma. \qquad (1.52)$$

Если преобразовать это выражение, то получим

$$P(-\Delta \le \Theta - T_n \le \Delta) = \gamma \qquad (1.53)$$

или

$$P(T_n - \Delta \le \Theta \le T_n + \Delta) = \gamma. \qquad (1.54)$$

Условие (1.54) означает, что интервал $[T_n-\Delta, T_n+\Delta]$ *покрывает значение параметра* (Θ) *с заданной доверительной вероятностью* (γ). Точность оценки (Δ) фактически определяет длину доверительного интервала (2Δ). Доверительная вероятность задается обычно значением, близким к единице, например, 0,95; 0,97; 0,99 и т.д. Доверительная *вероятность* (γ), *точность* оценки (Δ) и *объем выборки* (n) *связаны* между собой. Если *определены две* величины, то тем самым *будет определена и третья*. Рассмотрим определение доверительного интервала для среднего значения (μ) нормального распределения. Так как (σ^2) неизвестна, то непосредственно воспользоваться нормальным распределением нельзя. Однако, известно (формула 1.43), что случайная величина,

$$t = \frac{\bar{x} - \mu}{S}\sqrt{n}\ , \qquad (1.55)$$

имеет распределение Стьюдента (t-распределение) *с числом степеней свободы* (n-1). При этом *числом степеней свободы* считается число независимых отклонений отдельных вариант от среднего. Из отклонений независимыми считаются все варианты, кроме последней, величина которой уже определена остальными отклонениями и, поэтому, это отклонение не будет независимым. *Для* получения *интервальной оценки* необходимо выполнение условия

$$P\left(\left|\frac{\bar{x}-\mu}{\bar{S}}\sqrt{n}\right| \le t_\gamma\right) = \gamma \ . \qquad (1.56)$$

Величина (t_γ) определяется по таблицам распределения Стьюдента (Приложение, таблица П.2). На основании условия ($P(|x|>t_\gamma)=\gamma$) определяется (t_γ), но в данном случае, имеем противоположное неравенство, значит, необходимо использовать условие

$$P\left(\left|\frac{\bar{x}-\mu}{\bar{S}}\sqrt{n}\right| > t_\gamma\right) = 1-\gamma \ . \qquad (1.57)$$

Число степеней свободы равно (n-1). Преобразовав условие (1.56), имеем

$$P\left(\bar{x} - t_\gamma \frac{\bar{S}}{\sqrt{n}} \le \mu \le \bar{x} + t_\gamma \frac{\bar{S}}{\sqrt{n}}\right) = \gamma \ , \qquad (1.58)$$

где *доверительный интервал указан в скобках* и

$$t_\gamma \frac{\bar{S}}{\sqrt{n}} = \Delta \ . \qquad (1.59)$$

При определении доверительного интервала для дисперсии нормального распределения (σ^2) используется χ^2-распределение, т.е.

$$\chi^2 = \frac{(n-1)\cdot\bar{S}}{\sigma^2} \ . \qquad (1.60)$$

Случайная величина с χ^2-распределением *принимает* только *неотрицательные значения*. По таблицам χ^2 - распределения (Приложение, таблица П.3) можно найти (x_α), удовлетворяющее следующему условию: $P(\chi^2 f(x_\alpha)) = \alpha$ (рисунок 1.5). По таблицам χ^2-распределения всегда можно найти такие два числа, которые удовлетворяли бы условию

$$P(u_1 \le \chi^2 \le u_2) = \gamma. \qquad (1.61)$$

Таких пар чисел (u_1) и (u_2) существует бесконечное множество. Чтобы зафиксировать одну пару (u_1, u_2), выделим дополнительное условие

(симметричность по вероятности) (рисунок 1.6)

$$P(\chi^2 < u_1) = P(\chi^2 > u_2) = 0{,}5(1-\gamma). \qquad (1.62)$$

Из таблицы П.3. (Приложение), согласно условию (1.62), получаем (u_2). Для нахождения (u_1), используем вероятность противоположного события

$$P(\chi^2 > u_1) = 0{,}5(1+\gamma)\ . \qquad (1.63)$$

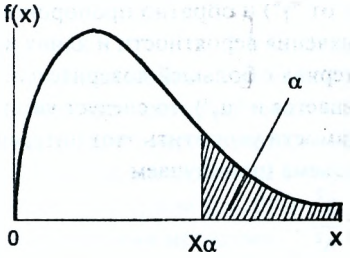


Рисунок 1.5
Использование таблицы χ^2 - распределения (таблица П.3).

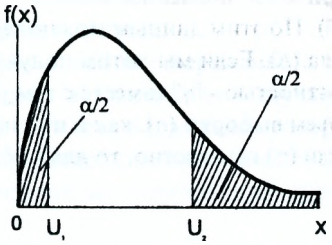


Рисунок 1.6
Нахождение чисел (u_1) и (u_2).

Заменяя (χ^2) в формуле (1.61) его значением из формулы (1.60), после преобразований, получаем

$$P\left(\frac{(n-1)\cdot\bar{S}^2}{u_2} \le \sigma^2 \le \frac{(n-1)\cdot\bar{S}^2}{u_1}\right) = \gamma\ , \qquad (1.64)$$

где в скобках задан доверительный интервал для дисперсии (σ^2). Извлекая квадратный корень из обеих сторон неравенства, *определяем доверительный интервал* для среднего квадратического (стандартного) отклонения (σ)

$$\frac{\bar{S}\sqrt{n-1}}{\sqrt{u_2}} \le \sigma \le \frac{\bar{S}\sqrt{n-1}}{\sqrt{u_1}}\ . \qquad (1.65)$$

До сих пор, *мы рассматривали обработку* готовых *выборок с фиксированным объемом (n)*. По закону больших чисел *предпочтение отдается выборкам с большим объемом*. Но обычно *большой объем* выборки *требует* и *больших затрат* для ее получения *и обработки*. Поэтому, *на практике,*

целесообразно использовать тот *минимальный объем, который позволяет получить удовлетворительные результаты. При вычислении* доверительных *интервалов среднего значения* нормального распределения, можно определить *объем выборки*

$$n = u_\gamma^2 \frac{\sigma^2}{\Delta^2} . \qquad (1.66)$$

Таким образом, объем выборки (n) прямо пропорционален (σ^2) и (u_γ^2) (при этом, последнее значение зависит от "γ") и обратно пропорционален (Δ^2). По этим данным можно задать значения вероятности и длину интервала (Δ). Если мы хотим получить интервал с большей доверительной вероятностью - "γ" (вместе с тем увеличивается и "u_γ"), то следует увеличить объем выборки (n), как и при необходимости укоротить этот интервал. Если (σ) неизвестно, то для выборки объема (n) получаем

$$n = t_\gamma^2 \frac{\bar{S}^2}{\Delta^2} . \qquad (1.67)$$

По формуле (1.66), *задавая* (σ), *можно определить* соответствующий *объем выборки (n) до получения* самой *выборки*. По формуле (1.67) можно определить *нужный объем* выборки (n) *после обработки* уже имеющейся *пробной выборки*, по которой *вычисляется несмещенная оценка дисперсии* генеральной совокупности ($\bar{S}^2$).

В качестве примера, *определим доверительные интервалы для математического ожидания и дисперсии, а также объем выборки, чтобы получить названные характеристики с заданной точностью, используя данные таблицы 1.1. Выборочные характеристики распределения норм атмосферных осадков на территории Беларуси представлены в таблице 1.2.*

Из таблицы П.2 (Приложение) t - распределения Стьюдента, при n-1=99 - степенях свободы и доверительной вероятности γ=0,95, найдем t_γ=1,9944. Используя формулу (1.59), имеем Δ=1,9944·49,7/(100)^{0,5}=9,91 и математическое ожидание запишется, как 732,2-9.9≤μ≤732,2+9,9, или окончательно, μ=732,2±9,9 мм, при 722,3≤μ≤742,1 мм. Таким образом, с вероятностью 95% математическое ожидание (μ) находится в диапазоне 722,3≤μ≤742,1 мм. Из Приложения (таблица П.3) χ^2-распределения u_1=82,36 и u_2=135,8. Используя формулу (1.64), имеем

$$\frac{2471{,}3(100-1)}{135{,}8} \le \sigma^2 \le \frac{2471{,}3(100-1)}{82{,}36}$$

1801,6≤ σ^2≤2970,6 или 42,4≤ σ≤54,5.

Таким образом, с вероятностью 95% дисперсия (σ) лежит в диапазоне 42,4≤ σ≤54,5. Для определения объема выборки, воспользуемся формулой (1.67). При γ=95 и n=100, по Приложению (таблица П.2) для t-распределения имеем t_γ=1,98 и, тогда, приняв Δ=10 мм, имеем $n = 1{,}98^2 \frac{2446{,}6}{10^2} = 95{,}9 \approx 96 < 100$, *а при* Δ=9 *мм -* $n = 1{,}98^2 \frac{2446{,}6}{9^2} = 118 > 100$.

***Таким образом, чтобы получить математическое ожидание** с точностью до 10 мм слоя атмосферных осадков при 95% - ной вероятности **достаточно 96 вариант**, а с точностью до 9 мм - уже **118 вариант**.*

1.5 Статистические гипотезы

Статистической гипотезой называется любое утверждение о виде или свойствах распределения наблюдаемых в эксперименте случайных величин. Например, случайная величина (x) имеет нормальное распределение, случайная величина с нормальным распределением имеет среднее значение (μ=10) или (μ≠10) и т. д. Статистические гипотезы проверяются статистическими методами.

Гипотезы о неизвестном параметре (Θ) распределения бывают простые и сложные; простая гипотеза утверждает, что параметр (Θ) имеет одно конкретное значение ($\Theta=\Theta_0$), а сложная гипотеза утверждает, что параметр (Θ) имеет значение из совокупности значений ($\Theta< \Theta_0$, $\Theta>\Theta_0$, $\Theta\neq\Theta_0$).

Проверяемую гипотезу обозначим (H_0). Обычно выбирают еще альтернативную гипотезу (H_1), отрицающую или исключающую основную гипотезу (H_0). Таким образом, в результате проверки можно принять только одну из гипотез (H_0) или (H_1), отвергая в то же время другую.

Гипотезу проверяют на основании выборки, полученной из генеральной совокупности. Из-за случайности выборки, в результате проверки могут возникнуть ошибки, ведущие к неправильным решениям. В принципе, возможны ошибки первого и второго рода. *Ошибка первого рода* имеет место тогда, *когда отвергается правильная гипотеза* (H_0). При ошибке *второго рода принимается неправильная гипотеза* (H_0).

Таким образом, по одним выборкам принимается правильное решение, по другим - неправильное. *Решение принимается по* значению *некоторой функции* выборки, называемой *статистикой* или *статистической характеристикой. Множество значений* этой *статистики* можно разделить на *два* непересекающихся *подмножества*:

а) *значений статистики*, при которых *гипотеза* (H_0) *принимается* (не отклоняется), называемых областью принятия гипотезы (*допустимой областью*);

б) *значений статистики*, при которых *гипотеза* (H_0) *отвергается* (отклоняется) и *принимается гипотеза* (H_1), называемых *критической областью*.

При проверке гипотез разумно *уменьшать вероятности* принятия *неправильных решений. Допускаемая вероятность ошибки первого рода* обозначается через (α) и *называется уровнем значимости*. Значение (α) обычно мало, но *уменьшение* вероятности *ошибки первого рода вызывает увеличение* вероятности *ошибки второго рода* (β).

Статистика выбирается так, чтобы вероятности (α) и (β) были бы минимальными. Для определения критической области статистики используется уровень значимости (α) и учитывается вид альтернативной гипотезы (H_1). Основная гипотеза (H_0) о значении неизвестного параметра (Θ) распределения выглядит так

$$H_0: \Theta = \Theta_0. \qquad (1.68)$$

Альтернативная гипотеза (H_1) может, при этом, иметь следующий вид

$$H_0: \Theta < \Theta_0, \; H_0: \Theta > \Theta_0 \text{ или } H_0: \Theta \neq \Theta_0. \qquad (1.69)$$

Соответственно, можно получить левостороннюю, правостороннюю или двустороннюю критические области. Граничные точки критических областей определяются по таблицам распределения статистики.

Проверка статистической гипотезы состоит из следующих этапов:

1) определение гипотез (H_0) и (H_1); 2) выбор статистики и задание уровня значимости (α); 3) определение (по таблицам) критической области по уровню значимости (α) и по альтернативной гипотезе (H_1); 4) вычисление по выборке значения статистики; 5) сравнение значения статистики с критической областью; 6) принятие решения: если значение статистики не входит в критическую область, то принимается гипотеза (H_0) и отвергается гипотеза (H_1); если значение статистики входит в критическую область, то отвергается гипотеза (H_0) и принимается гипотеза (H_1).

Итак, если в результате проверки статистической гипотезы приняли гипотезу (H_1), то можно считать ее доказанной; если приняли гипотезу (H_0), то признали, что она не противоречит результатам наблюдений.

Проверка нескольких статистических гипотез:

а) ***гипотеза о среднем значении нормального распределения при известном*** (σ)

Предположим, что генеральная совокупность имеет нормальное распределение $x \in N(\mu,\sigma)$, где σ - известно. При уровне значимости (α) нужно проверить гипотезу $(H_o : \mu = \mu_o)$. В качестве альтернативной можно использовать одну из следующих гипотез $H_1 : \mu < \mu_o$, $H_1 : \mu > \mu_o$ или $H_1 : \mu \neq \mu_o$. В качестве статистики воспользуемся случайной величиной

$$Z = \frac{\bar{x} - \mu_o}{\sigma}\sqrt{n}\ , \qquad (1.70)$$

которая при истинной гипотезе (H_o) имеет нормальное распределение $Z \in N(0,1)$. Критическая область определяется с помощью таблицы функции распределения (Приложение, таблица П.5). Если *альтернативная гипотеза имеет вид* $(H_1 : \mu < \mu_o)$, то *используется левосторонняя критическая область* (рисунок 1.7), которая удовлетворяет следующему условию

$$P(z < -z_\alpha) = \Phi(-z_\alpha) = \alpha. \qquad (1.71)$$

Таблицы составлены только для положительных значений аргумента, поэтому из таблиц находится (z_α), с учетом того, что

$$\Phi(z_\alpha) = 1 - \alpha. \qquad (1.72)$$

Отсюда следует, что критическая область - это множество таких (z), для которых

$$z < -z_\alpha. \qquad (1.73)$$

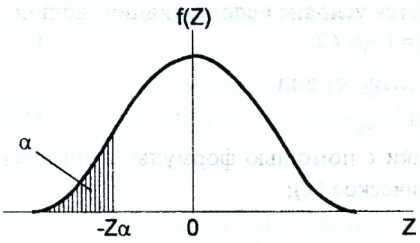


Рисунок 1.7 Левосторонняя критическая область.

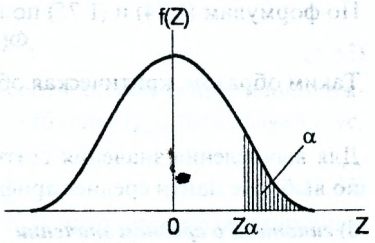


Рисунок 1.8 Правосторонняя критическая область.

Если *альтернативная гипотеза имеет вид* ($H_1 : \mu > \mu_o$), то *используется правосторонняя критическая область* (рисунок 1.8), которая удовлетворяет условию

$$P(z > z_\alpha) = \alpha. \qquad (1.74)$$

Из Приложений (таблица П.5) получается значение, с учетом того, что

$$P(z < z_\alpha) = \Phi(z_\alpha) = 1 - \alpha. \qquad (1.75)$$

Отсюда находится критическая область

$$z > z_\alpha. \qquad (1.76)$$

И наконец, *при альтернативной гипотезе* ($H_1 : \mu \neq \mu_o$), *используется двусторонняя критическая область* (рисунок 1.9), удовлетворяющая условию

$$P(|z| > z_\alpha) = \alpha. \qquad (1.77)$$

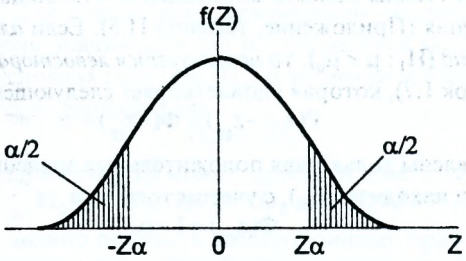


Рисунок 1.9 Двусторонняя критическая область.

Учитывая результаты определения абсолютной величины, находится

$$P(z < z_\alpha) = P(z > z_\alpha) = \alpha / 2. \qquad (1.78)$$

По формулам (1.74) и (1.75) получается условие использования таблицы

$$\Phi(z_\alpha) = 1 - \alpha / 2. \qquad (1.79)$$

Таким образом, критическая область имеет вид

$$|z| > z_\alpha. \qquad (1.80)$$

Для вычисления значения статистики с помощью формулы (1.70) нужно по выборке найти среднее арифметическое ($\bar{x}$);

б) *гипотеза о среднем значении нормального распределения при неизвестном (σ)*

Предположения те же, что и в предыдущем пункте, но только (σ) неизвестно. В этом случае, в качестве статистики используется случайная

величина

$$t = \frac{\bar{x} - \mu_o}{\bar{S}}\sqrt{n}\ , \qquad (1.81)$$

которая, если верна гипотеза (H_o), имеет t-распределение Стьюдента с числом степеней свободы (n-1), где n - объем выборки. Критические области определяются так же, как и в предыдущем пункте, но использование таблицы t-распределения Стьюдента проще, так как она составлена именно для этих целей. При нахождении левосторонней или правосторонней критических областей используется верхняя головка таблицы, а для двусторонней - нижняя. Перед вычислением по формуле (1.81) значения статистики, нужно по выборке вычислить ($\bar{x}$) и ($\bar{S}$);

в) *гипотеза о дисперсии нормального распределения*

Предположим, что генеральная совокупность имеет нормальное распределение - $x \in N(\mu,\sigma)$, где параметр (σ) неизвестен. Требуется при уровне значимости (α) проверить гипотезу ($H_o : \sigma^2 = \sigma_o^2$). В качестве статистики используется случайная величина

$$\chi^2 = \frac{(n-1)\cdot \bar{S}^2}{\sigma_0^2}\ . \qquad (1.82)$$

Если гипотеза (H_o) верна, то случайная величина (χ^2) имеет χ^2-распределение Пирсона с числом степеней свободы (n-1). Критическая область определяется в зависимости от альтернативной гипотезы (H_1) по Приложению (таблица П.3) для χ^2-распределения. Если альтернативная гипотеза имеет вид ($H_1 : \sigma^2 < \sigma_o^2$), то используется левосторонняя критическая область, удовлетворяющая условие

$$P(\chi^2 < x_\alpha) = \alpha. \qquad (1.83)$$

Таблица χ^2-распределения составлена в соответствии с противоположным условием. Затем, для нахождения из таблицы (χ_α), используется условие

$$P(\chi^2 > x_\alpha) = 1 - \alpha. \qquad (1.84)$$

При альтернативной гипотезе ($H_1 : \sigma^2 > \sigma_o^2$) находится правосторонняя критическая область, исходя из условия

$$P(\chi^2 > x_\alpha) = \alpha, \qquad (1.85)$$

по которому (χ_α) можно найти непосредственно из таблицы.

При альтернативной гипотезе ($H_1 : \sigma^2 \neq \sigma_o^2$) находится двусторонняя

критическая область согласно условию

$$P((\chi^2 < x'_\alpha) \cup (\chi^2 < x''_\alpha)) = \alpha. \quad (1.86)$$

Обычно принимается симметричная по вероятности критическая область, удовлетворяющая условие

$$P((\chi^2 < x'_\alpha) = P(\chi^2 \geq x''_\alpha)) = \alpha / 2. \quad (1.87)$$

При этом, можно сразу из таблицы найти (x_α''), а для получения (x_α') следует сделать преобразование

$$P(\chi^2 > x_\alpha) = 1 - \alpha / 2. \quad (1.88)$$

По выборке нужно вычислить несмещенную оценку дисперсии генеральной совокупности ($\bar{S}^2$), а затем по формуле (1.82) - найти значение статистики (χ^2);

г) ***гипотеза о равенстве двух средних значений***

Предположим, что заданы две генеральные совокупности с нормальным распределением - $X_1 \in N(\mu_1,\sigma_1)$, $X_2 \in N(\mu_2,\sigma_2)$, при этом, стандартные отклонения (σ_1) и (σ_2), неизвестны, но должны быть равными. Сначала нужно проверить гипотезу о равенстве дисперсий. Из обеих генеральных совокупностей сделаны независимые выборки с параметрами - ($n_1, \bar{x}_1, \bar{S}_1$); ($n_2, \bar{x}_2, \bar{S}_2$), соответственно. Обозначим разность средних через ($\delta = \mu_1 - \mu_2$). Зафиксировав уровень значимости (α), проверим гипотезу ($H_o: \delta = \delta_o$), используя статистику

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\frac{(n_1 - 1)\bar{S}_1^2 + (n_2 - 1)\bar{S}_2^2}{n_1 + n_2 - 2} \cdot + (\frac{1}{n_1} + \frac{1}{n_2})}}. \quad (1.89)$$

Если гипотеза (H_o) верна, то случайная величина имеет t-распределение с числом степеней свободы ($n_1 + n_2 - 2$). Обычно, ($\delta_o = 0$), т.е. проверяется гипотеза о равенстве средних значений генеральных совокупностей. Критическая область определяется, в зависимости от вида альтернативной гипотезы $H_o(\delta < \delta_o, \delta > \delta_o$ или $\delta \neq \delta_o)$, по Приложению (таблица П.2 для t-распределения);

д) гипотеза о равенстве двух дисперсий

Предположим, как и в предыдущем пункте, что заданы две генеральные совокупности (X_1) и (X_2) с нормальным распределением - $X_1 \in N(\mu_1,\sigma_1)$ и $X_2 \in N(\mu_2,\sigma_2)$. Из этих генеральных совокупностей образованы независимые выборки с параметрами - ($n_1, \bar{x}_1, \bar{S}_1$) и ($\bar{x}_2, \bar{S}_2$), соответственно. Требу-

ется при уровне значимости (α) проверить гипотезу ($H_0: \sigma_1^2 = \sigma_2^2$) при альтернативной гипотезе ($H_0: \sigma_1^2 > \sigma_2^2$). Обычно, здесь другие альтернативные гипотезы не используются. Предполагая, что ($\bar{S}_1^2 > \bar{S}_2^2$), принимаем в качестве статистики величину

$$F = \frac{\bar{S}_1^2}{\bar{S}_2^2}. \qquad (1.90)$$

Если гипотеза (H_0) верна, то случайная величина (F) имеет F-распределение Фишера с числом степеней свободы (n_1-1) и (n_2-1). Критическая область только правосторонняя; определяется условием

$$P(F > f_\alpha) = \alpha. \qquad (1.91)$$

Величины (f_α) находятся из Приложений для F-распределения (таблицы П.4.1 - П.4.2), которые зависят от трех величин: уровня значимости (α) и степеней свободы (ν_1) и (ν_2), поэтому, таблицы составлены отдельно для каждого значения (α). В таблицах число степеней свободы большей дисперсии (ν_1) находится в верхней их части;

е) χ^2-*критерий согласия*

Рассмотрим, как можно проверить гипотезу о распределении генеральной совокупности (X). Пусть генеральная совокупность имеет какое-то неизвестное распределение. Сделаем выборку из генеральной совокупности. На основании выборки или, учитывая какие-то другие соображения, *составим гипотезу о конкретном распределении генеральной совокупности*, выраженной через *функцию распределения* - F(x). Это распределение *является теоретическим*. По выборке *можно найти эмпирическую функцию распределения* - $F^*(x)$. *Гипотеза* (H_o) о распределении генеральной совокупности *принимается* тогда, *когда эмпирическое распределение хорошо согласуется с теоретическим*. Полного совпадения, конечно, ожидать не стоит. При использовании χ^2-критерия согласия вся область изменения генеральной совокупности (X) делится на k - интервалов, которые могут быть различной длины. По выборке составляется вариационный ряд по этим же интервалам. Если в некотором интервале частота (n_i) слишком мала (меньше 5), то этот интервал объединяется с соседним интервалом. При дискретной генеральной совокупности интервал может содержать только одно значение генеральной совокупности. По выборке вычисляются оценки параметров теоретического распределения. Тем самым, теоретическое распределение будет полностью определено. Теперь по теоре-

тическому распределению вычисляются вероятности (p_i) того, что случайная величина (x) принимает значение из i-го интервала, при этом, $\sum_{i=1}^{k} p_i = 1$. Затем, находятся теоретические частоты ($m_i = n_i \cdot p_i$). Гипотеза (H_o) верна, если теоретические и эмпирические частоты (m_i) и (n_i) достаточно мало отличаются друг от друга. Для проверки гипотезы (H_o) используется следующая статистика

$$Q^2 = \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i} . \qquad (1.92)$$

Случайная величина (Q^2) имеет χ^2-распределение с числом степеней свободы (k-v-1), где k - количество интервалов, v - количество параметров теоретического распределения, оценки которых вычисляются по выборке. Чем больше (Q^2), тем хуже согласованы теоретическое и эмпирическое распределения. При достаточно большом значении (Q^2) нужно отвергнуть гипотезу (H_o). Поэтому, используется только правосторонняя критическая область;

ж) *отсев грубых погрешностей*

Предложим для практического использования наиболее простые методы отсева грубых погрешностей. Если в распоряжении экспериментатора имеется выборка небольшого объема ($n \le 25$), то можно воспользоваться методом вычисления максимального отклонения

$$\tau' = \frac{\left| x_{i\text{-}max(min)} - \bar{X} \right|}{\sqrt{(n-1)/n} \cdot \bar{S}} , \qquad (1.93)$$

где $x_{i\text{-}max(min)}$ - крайний (наибольший или наименьший) элемент выборки, по которой подсчитывались ($\bar{x}$) и ($\bar{S}$); $\tau_{1\text{-}P}$ - табличное значение статистики (τ), вычисленной при доверительной вероятности (q=1-P). Таким образом, для выделения аномального значения вычисляется

$$\tau = \frac{\left| x_{i\text{-}max(min)} - \bar{X} \right|}{\bar{S}} , \qquad (1.94)$$

которое, затем, сравнивается с табличным значением (τ_{1-p})

$$\tau \le \tau_{1-p}. \qquad (1.95)$$

Если неравенство соблюдается, то результат наблюдения не отсеивается, если не соблюдается, - он исключается. После исключения того или иного результата наблюдения (нескольких результатов), характеристики эмпи-

рического распределения должны быть пересчитаны по данным сокращенной выборки. Квантили распределения статистики (t), при уровнях значимости - P=0,10; P=0,05; P=0,01, или, соответственно доверительной вероятности (1-P)=q=0,90; 0,95; 0,975; 0,99 даны в Приложении (таблица П.2). На практике обычно используется уровень значимости- P=0,05 (с 95%-ной доверительной вероятностью). Проверку отсева можно повторить и для следующего по абсолютной величине максимального относительного отклонения, но предварительно необходимо пересчитать ($\bar{x}$) и ($\bar{S}$) для выборки нового объема (n-1). Рассмотрим другой метод отсева грубых погрешностей для малой выборки. В этом случае, вычисляется

$$\tau' = \frac{\left|x_{i\text{-}max(min)} - \bar{X}\right|}{\sqrt{(n-1)/n}\cdot\bar{S}}, \qquad (1.96)$$

и полученный результат сравнивается с критическим табличным значением (Приложение, таблица П.2) при следующих (n) и (1-P). Отсев грубых погрешностей для больших выборок лучше всего производить с использованием t-распределения Стьюдента. Известно, что критическое значение "τ_P" (p-процентная точка нормированного выборочного отклонения) выражается через ($t_{p;\ n-2}$) - критическое значение распределения Стьюдента (t - распределение. Приложение, таблица П.2)

$$\tau_{(p,n)} = \frac{t_{(p;n-2)}\sqrt{n-1}}{\sqrt{n-2+\left(t_{(p;n-2)}\right)^2}}. \qquad (1.97)$$

Рекомендуемый метод отсева грубых погрешностей удобен еще тем, что максимальные относительные отклонения в процессе вычисления могут быть разделены на группы: 1) $\tau \le \tau_{(5\%;n)}$; 2) $\tau_{(5\%;n)} < \tau < \tau_{(0,1\%;n)}$; 3) $\tau > \tau_{(0,1\%;n)}$. Результаты наблюдений, попавшие в первую группу, нельзя отсеивать ни в коем случае, из второй группы - можно отсеять, если в пользу этой процедуры имеются еще и другие соображения экспериментатора, из третей группы, по-видимому, результаты отсеиваются всегда.

Проиллюстрируем проверку статистических гипотез на примере выборки атмосферных осадков Беларуси:

а) проверим выборку из таблицы 1.1 на наличие в средних многолетних годовых величинах атмосферных осадков Беларуси грубых погрешностей

Применяем следующую процедуру отсева грубых погрешностей измерения атмосферных осадков для больших выборок:

1) из таблицы 1.1 выбираем результат наблюдения, имеющий наибольшее отклонение, - 841-732,2=109,2 мм;

2) по формуле (1.94) вычисляем - $\tau=|X_{i\cdot max(min)} - \bar{x}|/S=|(841-732,2)|/49,7=2,20$;

3) по Приложению (таблица П.2) находим процентные точки t-распределения Стьюдента ($t_{p;n-2}$) - $t_{5\%;98}=1,6615$; $t_{0,1\%;98}=3,1750$;

4) по формуле (1.97) вычисляем соответствующие точки $\tau_{5\%;98}$ *и* $\tau_{0,1\%;98}$:

$$\tau_{(5\%;98)} = \frac{1,6615\sqrt{100-1}}{\sqrt{100-2+(1,6615)^2}} = 1,647;$$

$$\tau_{(0,1\%;98)} = \frac{3,1750\sqrt{100-1}}{\sqrt{100-2+(3,1750)^2}} = 3,039.$$

Значение $\tau=2,20$ *находится между табличными критическими значениями:* $1,647 \le 2,20 \le 3,039$.

В этом случае, от отсева, выделяющегося наблюдения, лучше всего воздержаться;

б) проверим гипотезу нормального распределения на выборке средних многолетних годовых величин атмосферных осадков Беларуси (таблица 1.1)

Рассмотрим ***пять основных методик*** *проверки гипотезы нормальности распределения: по среднему абсолютному отклонению (САО), по размаху варьирования (Rv), по показателям асимметрии и эксцесса, по* χ^2*-критерию и по критерию Колмогорова-Смирнова (К - С - критерию).*

Методика проверки *нормальности распределения* ***по показателям асимметрии и эксцесса*** *хорошо иллюстрирует использование моментов, а также удобна при проведении расчетов на ЭВМ.* ***Проверка по К-С - критерию*** *проводится только в отдельных случаях. Для практического применения рекомендуются, в основном, две методики:* ***по размаху варьирования и по*** χ^2***-критерию,*** *причем, первая служит для быстрой "прикидочной" проверки, а вторая - для основательной проверки нормальности распределения.*

Для ***небольших выборок*** *(n<120)* ***можно использовать среднее абсолютное отклонение*** *(САО), выражаемое формулой*

$$САО=\sum|x_i - \bar{x}|/n. \qquad (1.98)$$

Для выборки, имеющей приближенно нормальный закон распределения , должно быть справедливо выражение

$$|САО/S - 0,7979| < 0,4\sqrt{n}. \qquad (1.99)$$

При соотношении, $|41,0 / 49,7 - 0,7979| < 0,4\sqrt{100}$, гипотеза нормальности распределения выборки данных, приведенных в таблице 1.1, принимается.

***Быструю проверку гипотезы нормальности распределения** можно выполнить с **помощью размаха варьирования Rv** (при 3<n<100). При этом, рассчитывается отношение $R / \overline{S}$ и сопоставляется с его критическими верхними и нижними границами. Если $R / \overline{S}$ меньше нижней или больше верхней границы, то нормального распределения нет. Особенно важно , чтобы это условие соблюдалось при P=0,10 (10%-ный уровень значимости). В нашем случае, $R / \overline{S} = 217 / 49,7 = 4,37$. При n=100 и p=0,10, нижняя и верхняя границы, соответственно, равны 4,44 и 5,65, т.е. условие не выполняется и, следовательно, с вероятностью P=90% можем отвергнуть гипотезу о нормальности распределения.*

Некоторое представление о близости эмпирического распределения к нормальному дает анализ показателей асимметрии и эксцесса.

***Показатель асимметрии** можно определить по формуле*

$$g_1 = \frac{m_3}{m_2^{1,5}} = \frac{-15068,6}{2446,6^{1,5}} = -0,12 \neq 0 \; ; \qquad (1.100)$$

*следовательно, некоторая **асимметрия имеет место**.*

***В качестве показателя эксцесса** принимается величина*

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{15007490}{2446,6^2} - 3 = -0,49 \neq 0; \qquad (1.101)$$

*как видно, **эксцесс также имеет место.***

***Несмещенные оценки** для показателей **асимметрии и эксцесса** определяются с использованием соответствующих формул:*

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 = \frac{\sqrt{100(100-1)}}{100-2}(-0,12) = -0,122 \; ; \qquad (1.102)$$

$$G_2 = \frac{(n-1)}{(n-2)(n-3)}((n+1)g_2 + 6) = \frac{100-1}{(100-2)(100-3)}((100+1)(-0,49)+6) = -0,453. \qquad (1.103)$$

*При проверке гипотезы нормальности распределения **следует также вычислить среднеквадратические отклонения для показателей асимметрии и эксцесса** :*

$$S_{G_1} = \sqrt{\frac{6 \cdot n(n-1)}{(n-2)(n+1)(n+3)}} = \sqrt{\frac{6 \cdot 100(100-1)}{(100-2)(100+1)(100+3)}} = 0,24 \; ; \qquad (1.104)$$

$$S_{G_2} = \sqrt{\frac{24 \cdot n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}} = \sqrt{\frac{24 \cdot 100(100-1)^2}{(100-3)(100-2)(100+3)(100+5)}} = 0,48 \; . \qquad (1.105)$$

Если выполняются условия:

$$|G_1| \leq 3S_{G_1}; \qquad (1.106)$$

$$|G_2| \leq 5S_{G_2}, \qquad (1.107)$$

то гипотеза нормальности исследуемого распределения может быть принята.
В нашем случае, $|-0,122| \le 3 \cdot 0,24$ *и* $|-0,45| \le 5 \cdot 0,48$, *следовательно, условия (1.106) - (1.107) выполняются и гипотеза нормальности распределения может быть принята.*

***Рассмотрим методику проверки гипотезы нормальности** распределения **по** χ^2**- критерию.** *Применение критерия* (χ^2) *предполагает также использование свойств, так называемого, стандартного нормального распределения. Уравнение кривой стандартного нормального распределения имеет вид*

$$y = f(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{z^2}{2}) \approx 0,4\, exp(-\frac{z^2}{2}), \qquad (1.108)$$

где $z = (x_i - \mu) / \sigma$. *Значения ординат кривой стандартного нормального распределения протабулированы и приведены в Приложении (таблица П.6). Расчеты выполняются в табличной форме с использованием данных таблицы 1.3. Методика (см. формулу 1.92) и результаты расчетов значений* (χ^2) *нашли свое отражение в таблице 1.4 и ниже по тексту*

$$\bar{x} = \sum_{i=1}^{n} n_i x_i / n = 72760 / 100 = 727,61;$$

$$\bar{S} = \sqrt{\frac{\sum n_i x_i^2 - \sum (n_i x_i)^2 / n}{n-1}} = \sqrt{\frac{53169096 - 72760^2 / 100}{100-1}} = 48,1;$$

$$k' = nb / \bar{S} = \frac{100 \cdot 30}{48,1} = 62,4;$$

b *– размер класса.*

Из таблицы 1.4 видно, что критерий $\hat{\chi}^2 = 3,857$. *Число степеней свободы* $\nu = n_{кл} - 1 - 2 = 6 - 1 - 2 = 3$, *так как оценивается два параметра:* $\bar{x}$ *и* $\bar{S}$ ($n_{кл}$ *- число классов интервалов).*

По Приложению (таблица П.3) находим значение - $\chi^2_{(3,10\%)} = 6,251$.

***Таким образом,** гипотеза о том, что наблюдаемые частоты распределены нормально, принимается на 10%-ном уровне.*

***Данные таблицы 1.4** можно **использовать и для проверки гипотезы нормальности распределения с помощью критерия согласия Колмогорова-Смирнова** (К - С - критерия), для этого вычисляется*

$$\hat{D} = \frac{max\left|F_{n_i} - F_{m_i}\right|}{n}, \qquad (1.109)$$

где F_{n_i} *– накопленная наблюдаемая частота;* F_{m_i} *- накопленная ожидаемая частота.*

Таблица 1.4 Результаты вычисления (χ^2)

№ класса	Середина интервалов (x_i)	Частоты (n_i)	x_i^2	$n_i x_i$	$n_i x_i^2$	$x_i - \bar{x}$
1	2	3	4	5	6	7
1	625	4	390625	2500	1562500	-102,6
2	655	9	429025	5895	3861225	-72,6
3	685	15	469225	10275	7038375	-42,6
4	715	21	511225	15015	10735725	-12,6
5	745	26	555025	19370	14430650	17,4
6	775	16	600625	12400	9610000	47,4
7	805	7	648025	5635	4536175	77,4
8	835	2	697225	1670	1394450	107,4
		Σ 100		72760	53169096	

→ *Продолжение таблицы 1.4*

$\left\|\frac{x_i - \bar{x}}{S}\right\| = z$	Ординаты $f(z)$	$f(z)k'$	m_i	$n_i - m_i$	$(n_i - m_i)^2$	$\frac{(n_i - m_i)^2}{m_i}$
8	9	10	11	12	13	14
2,133	0,0413	2,577				
1,509	0,1276	7,962	10,539	2,461	6,057	0,575
0,886	0,2685	16,754	16,754	-1,754	3,077	0,184
0,262	0,3857	24,068	24,068	-3,068	9,413	0,391
0,362	0,3739	23,331	23,331	2,669	7,124	2,669
0,985	0,2444	15,251	15,251	0,749	0,561	0,037
1,609	0,1092	6.814	8,886	0,114	0,013	0,001
2,233	0,0332	2,072				
						Σ 3,857

Результаты вычисления К - С - критерия приведены в таблице 1.5. Величины (n_i) и (m_i) получены накоплением их частот. Затем, выбрано максимальное значение $F_{n_i} - F_{m_i}$ и по нему определен критерий согласия Колмогорова-Смирнова ($\hat{D}$). Полученное значение сравнивается с критическим, взятым из Приложения (таблица П.9) ($\hat{D} \le D^T$). В нашем случае, $D_{(100;1,10)}=0,121>0,025=\hat{D}$, т.е. можно сделать тот же вывод, что и вы-

*ше: **гипотеза нормального распределения на достаточно "жестком" 10%-ном уровне принимается.***

***Таким образом**, четыре теста из пяти не отвергают гипотезу нормального распределения, следовательно, можно сделать заключение, что вариационный ряд годовых атмосферных осадков (таблица 1.1) подчиняется нормальному закону распределения вероятностей.*

Таблица 1.5 Результаты вычисления К - С - критерия

n_i	4	9	15	21	26	16	7	2
m_i	2,577	7,962	16,754	24,068	23,331	15,251	6,814	2,072
F_{n_i}	4	13	28	49	75	91	98	100
F_{m_i}	2,577	10,539	27,293	51,361	74,692	89,943	96,757	98,829
$\lvert F_{n_i} - F_{m_i} \rvert$	1,423	2,461	0,707	2,361	0,308	1,057	1,243	1,171

$D = 2,461/100 = 0,025$.

*Как известно, территория Беларуси находится на водоразделе Балтийского и Черного морей. **Рассмотрим различимы ли средние величины годовых атмосферных осадков для различных частей территории Беларуси,** при уровне значимости $\alpha = 0,05$. Рассчитав среднюю величину осадков и дисперсию для различных склонов, соответственно, имеем: $x_ч = 741,1$ мм; $S_ч^2 = 1272,1$; $n_ч = 39$ и $x_б = 737,5$ мм; $S_б^2 = 1940,7$; $n = 30$.*

*Сначала **проверим гипотезу о равенстве дисперсии** , т.е. $H_0: \sigma_ч^2 = \sigma_б^2$, при $H_1: \sigma_ч^2 > \sigma_б^2$. Статистика F (формула 1.90) имеет F-распределение с числом степеней свободы $(n_ч - 1)$ и $(n_б - 1)$. Найдем правостороннюю критическую область, согласно условию (1.91), по Приложению (таблица П.4.1); при F-распределении имеем*

$P(F > f_н) = 0,05$; $f_н = 1,85$; $F > 1,85$.

По формуле (1.90) вычислим значение статистики

$$F = \frac{1940,7}{1272,1} = 1,53.$$

Это значение не принадлежит критической области, поэтому, нет оснований отвергать гипотезу ($H_0: \sigma_ч^2 = \sigma_б^2$), т.е. можно считать дисперсии генеральных совокупностей равными.

***Проверим гипотезу о равенстве средних значений генеральных совокупностей** , т.е. $H_0: \delta = 0$, при $H_1: \delta < 0$.*

Статистика "t" (формула 1.89) имеет t-распределение с числом степеней свободы $(n_ч + n_б - 2)$. По альтернативной гипотезе (H_1), используя условие ($P(t < t_\alpha) = 0,05$) и таблицу t-распределения, найдем левостороннюю критическую область $t_\alpha = -1,67$; $t < -1,67$.

По формуле (1.89) вычислим значение статистики

$$t = \frac{737,5 - 741,1}{\sqrt{\frac{(30-1)\cdot 1940,7 + (39-1)\cdot 1272,1}{30+39-2}\cdot\left(\frac{1}{30}+\frac{1}{39}\right)}} = -0,37.$$

Полученное значение не принадлежит критической области, поэтому оснований отвергать гипотезу (H_0:δ =0) нет, т.е. ***считаем средние значения генеральных совокупностей равными.***

При уровне значимости ($\alpha = 0,05$) ***проверим гипотезу о среднем значении*** *(H_0:μ=750), при (H_1:μ<750) (таблица 1.2).*

Статистика "t" (формула 1.81) имеет t-распределение с числом степеней свободы- (n-1=100 - 1=99). При альтернативной гипотезе (H_1), найдем левостороннюю критическую область по условию ($P(t< t_\alpha)=0,05$).

Из Приложения (таблица П.2, t-распределение Стьюдента) получаем $t_{\alpha=0,05} = -1,6602$.

Значение статистики вычислим по формуле (1.81); $t = \frac{732,2 - 750}{49,7}\sqrt{100} = -3,5$.

Значение статистики принадлежит критической области. Следовательно, отвергаем гипотезу (H_0) и принимаем альтернативную гипотезу (H_1:μ<750).

При уровне значимости ($\alpha = 0,05$), ***проверим гипотезу о дисперсии,*** *т.е. (H_0:σ^2=2450), при (H:σ^2>2450).*

Статистика "χ^2" (формула 1.82) имеет χ^2- распределение с числом степеней свободы (n-1=100-1=99). По альтернативной гипотезе (H_1) найдем правостороннюю критическую область, используя условие (1.85) и таблицу χ^2-распределения Пирсона. При этом имеем: $P_{(\chi^2 > \chi_\alpha)}=0,05$; $\chi_\alpha=123,3$; $\chi^2>123,3$.

По формуле (1.82) вычислим значение статистики:

$$\chi^2 = \frac{(100-1)\cdot 2471,3}{2450} = 99,9.$$

Это значение принадлежит критической области, поэтому ***гипотеза*** *(H_0:$\sigma^2 = 2450$)* ***не будет отвергнута.***

Рассмотрим несколько примеров вычисления статистических параметров:

а) при определении содержания фосфора в растительном материале получены ы ***следующие результаты*** *(в г P_2O_5 на 100 г сухого вещества): 0,56; 0,53; 0,49; 0,57; 0,48.* ***Необходимо вычислить*** *($\bar{x}$), (s_x) - 95%-ные и 99%-ные доверительные интервалы для среднего значения совокупности.*

Решение

Целесообразно исходные данные преобразовать по соотношению $X_1=XK-A=X\cdot100-50$, т. е. умножить каждое значение (X) на 100, а затем отнять условную среднюю - A=50. В итоге, получим ряд однозначных цифр, удобных для вычисления статистических показателей. При наличии вычислительной машины, расчеты можно вести без подобного преобразования. В таблице 1.6 представлено три способа вычисления суммы квадратов отклонений, и легко убедиться в рациональности преобразования исходных дат. При вычислении статистических характеристик записи рекомендуется вести в такой последовательности:

$$\bar{x}=\frac{\Sigma x}{n}=\frac{2{,}63}{5}=0{,}526\,г;$$

$$s^2=\frac{\Sigma(x-\bar{x})^2}{n-1}=\frac{0{,}00652}{5-1}=0{,}0016;$$

$$s=\sqrt{s^2}=\sqrt{0{,}0016}=0{,}04\ г;\ V=\frac{s}{\bar{x}}\cdot100=\frac{0{,}04}{0{,}526}\cdot100=7{,}60\ \%;$$

$$s_{\bar{x}}=\sqrt{\frac{s^2}{n}}=\sqrt{\frac{0{,}0016}{5}}=0{,}018\ г;$$

$$s_{\bar{x}}\%=\frac{s_{\bar{x}}}{\bar{x}}\cdot100=\frac{0{,}018}{0{,}526}\cdot100=3{,}42\ (отн.)\%;$$

$$\bar{x}\pm t_{0,5}s_{\bar{x}}=0{,}526\pm2{,}776\cdot0{,}018=0{,}526\pm0{,}050(0{,}48...0{,}58)\ г;$$

$$\bar{x}\pm t_{0,1}s_{\bar{x}}=0{,}526\pm4{,}604\cdot0{,}018=0{,}526\pm0{,}083(0{,}44...0{,}61)\ г.$$

Теоретические значения (t) берутся из Приложения (таблица П.2) для 5%-ного и 1%-ного уровня значимости при степенях свободы (n–1=5–1=4).

***Итак**, средняя изучаемой совокупности с 95%-ным уровнем вероятности находится в интервале 0,48...0,58 и с 99%-ным уровнем - в интервале 0,44...0,61 граммов P_2O_5 на 100 граммов сухого вещества. Вероятность ошибочного заключения, в первом случае, составляет 5%, во втором, - 1%. Абсолютная ошибка средней $s_{\bar{x}}=0{,}018$ граммов и относительная ошибка $s_{\bar{x}}=3{,}42\%$; коэффициент вариации V=7,6% характеризует в данном примере, ошибку параллельных анализов;*

Таблица 1.6 Способы вычисления средней арифметической суммы квадратов отклонений

X	От истинной средней ($\bar{x}$)		По исходным датам (X)	По преобразованным датам (X_1)			
				$X_1=X-A$ ($A=0,50$)		$X_1=XK-A$ ($K=100$; $A=50$)	
	$X-\bar{x}$	$(X-\bar{x})^2$	X^2	X_1	X_2	X_1	X_2
1	2	3	4	5	6	7	8
0,56	0,034	0,001156	0,3136	0,06	0,0036	6	36
0,53	0,004	0,000016	0,2809	0,03	0,0009	3	9
0,49	0,036	0,001296	0,2401	–0,01	0,0001	–1	1
0,57	0,044	0,001936	0,3249	0,07	0,0049	7	49
0,48	0,046	0,002116	0,2304	–0,02	0,0004	–2	4
$\Sigma X = 2,63$	$\Sigma(X-\bar{x})=0$	$\Sigma(X-\bar{x})^2=$ $=0,00652$	$\Sigma X^2=$ $=1,3899$	$\Sigma X_1=$ $=0,13$	$\Sigma X_1^2=$ $=0,0099$	$\Sigma X_1=13$	$\Sigma X_1^2=99$
Средняя ($\bar{x}$)	$\frac{\Sigma X}{n}=\frac{2,63}{5}=0,526$			$A+\frac{\Sigma X_1}{n}=0,50+$ $+\frac{0,13}{5}=0,526$		$\left(A+\frac{\Sigma X_1}{n}\right):K=$ $=\left(50+\frac{13}{5}\right):100=0,526$	
Сумма квадратов $(X-\bar{x})^2$	0,00652	$\Sigma X^2-(\Sigma X)^2:n=1,3899-$ $-(2,63)^2:5=0,00652$		$\Sigma X_1^2-(\Sigma X_1)^2:n=0,0099-$ $-(0,13)^2:5=0,00652$		$\left[\Sigma X_1^2-(\Sigma X_1)^2:n\right]K^2=$ $=\left[99-(13)^2:5\right]100^2=$ $=0,00652$	

б) обследовано 113 полей озимой пшеницы на зараженность корневой гнилью (таблица 1.7). Существенно ли различие в пораженности пшеницы, высеянной по чистым и занятым парам ?

Решение

*Согласно нулевой гипотезы (H_0), вид пара не оказывает влияния на пораженность озимой пшеницы корневой гнилью и, следовательно, колебание соотношений сильно и слабо пораженных полей в каждой колонке таблицы 2*2 является случайным. На основании нулевой гипотезы, для каждой клетки таблицы вычисляем, каковы должны быть ожидаемые значения (F). Для вычисления ожидаемых частот общее число полей в каждой группе умножаем на ожидаемую долю слабо зараженных (58,4) или сильно зараженных*

*(41,6) полей. Ожидаемая численность слабо зараженных полей чистого пара- F_1=(42*58,4):100=24,5 и сильно зараженных - F_2=(42*41,6):100=17,5; в группе занятых паров ожидаемая численность слабо зараженных полей - F_3=(71*58,4):100=41,5 и сильно зараженных - F_4=(71*41,6):100=29,5. Эти числа, (ожидания) в таблице 1.7, заключены в скобки. После определения ожиданий, находим разности между фактическими и ожидаемыми частотами (таблица 1.8). Суммы всех разностей по колонкам и строчкам равны нулю.*

*Таблица 1.7 Пораженность озимой пшеницы в связи с видами паров и вычисление ожидаемой численности полей (F) по таблице 2*2*

Вид пара	Заражение		Сумма	Процент слабо зараженных полей
	слабое	сильное		
Чистый	30(24,5)	12(17,5)	42(42)	71,4
Занятой	36(41,5)	35(29,5)	71(71)	50,7
Сумма	66(66)	47(47)	113(113)	58,4

Таблица 1.8 Разности между фактическими и ожидаемыми численностями полей (f–F)

Вид пара	Заражение		Сумма
	слабое	сильное	
Чистый	5,5	–5,5	0
Занятой	–5,5	5,5	0
	0	0	0

Далее,

$$\chi^2 = \Sigma \frac{(f-F)^2}{F} = \frac{(5,5)^2}{24,5} + \frac{(-5,5)^2}{17,5} + \frac{(-5,5)^2}{41,5} + \frac{(5,5)^2}{29,5} =$$

$$= 1,23 + 1,74 + 0,73 + 1,02 = 4,72, \text{ при } (c-1)(k-1) = (2-1)(2-1) = 1 -$$

- степень свободы. Теоретическое значение - χ^2_{05}=3,84 (по Приложению, таблица П.3).

Вывод

*Наблюдается существенное увеличение зараженности посевов пшеницы при посеве ее по занятым парам ($\chi^2_{факт} > \chi^2_{0,5}$) и нулевая гипотеза о независимости заражения посевов от вида пара отвергается. Использование критерия (χ^2) при работе с таблицами состава 2*2 требует, чтобы ни одно из ожиданий не было меньше 5. Если теоретические численности невелики, то, до того как вычислять (χ^2), все разности (f–F) уменьшаются*

на 0,5, приближаясь к нулю. В нашем примере,

$$\chi^2 = \frac{5,0^2}{24,5} + \frac{(-5,0)^2}{17,5} + \frac{(-5,0)^2}{41,5} + \frac{5,0^2}{29,5} = 3,90;$$

в) в двух образцах почвы определено содержание гумуса в четырехкратной повторности и для каждого образца вычислена средняя ее ошибка (в %): $\bar{x}_1 \pm s_{\bar{x}_1} = 2,36 \pm 0,08\%$; $\bar{x}_2 \pm s_{\bar{x}_2} = 2,09 \pm 0,07$. *Число степеней свободы* $(\nu = n_1 + n_2 - 2 = 4+4-2=6)$. *В приложениях (таблица П.2) ему соответствует теоретическое* $t_{05}=2,45$ *и* $t_{01}=3,71^*$. *Фактическое значение критерия существенности находим по соотношению*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}} = \frac{2,36 - 2,09}{\sqrt{0,08^2 + 0,07^2}} = \frac{0,27}{0,106} = 2,55.$$

Сопоставляя фактическое значение (t) с теоретическим, ***приходим к выводу, что*** $(t_{факт} > t_{05} < t_{01})$. *Следовательно, разность существенна при 5%-ном уровне значимости. При более строгом подходе к оценке результатов, т. е. при 1%-ном уровне, разность несущественна,* ***образцы почвы по содержанию гумуса относятся к одной совокупности и другие выборки могут иметь одинаковые значения этого показателя;***

г) при просмотре 500 растений льна было обнаружено 50 растений, пораженных фузариозом. Определить 95%-ные и 99%-ные доверительные интервалы для генеральной доли пораженных растений в совокупности.

Решение

Исходные ***данные при альтернативной (двояковозможной) изменчивости распределяем по двум группам. Первая группа*** *- растения, имеющие признак; в нашем примере –* ***пораженные растения*** $(n_1=50)$, ***и вторая группа*** *– растения, у которых этот признак отсутствует, т. е.* ***здоровые растения*** $(n_2 = N - n_1 = 500 - 50 = 450)$.

Вычисления *свободных характеристик выборки* ***ведутся в такой последовательности:***

1) ***доля пораженных*** *(p) и* ***здоровых*** *(q) растений -*

$$p = \frac{n_1}{N} = \frac{50}{500} = 0,10 \text{ (или } 10\%);$$

$$q = 1 - p = 1 - 0,10 = 0,90 \text{ (или } 90\%);$$

2) ***стандартное отклонение*** *доли -*

$$s = \sqrt{pq} = \sqrt{0,10 \cdot 0,90} = 0,30 \text{ (или } 30\%);$$

---

* **Примечание**. Индексами при букве (t) записаны показатели уровня значимости: 5%-ный и 1% - ный (соответственно, 05; 01).

*3) **коэффициент вариации** (при k=2; $s_{макс}$=0,50) -*

$$V_p = \frac{s}{s_{макс}} \cdot 100 = \frac{0,30}{0,50} 100 = 60,0\ \%;$$

*4) **ошибка выборочной доли** -*

$$s_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{0,10 \cdot 0,90}{500}} = 0,013\ (\text{или } 1,3\%);$$

*5) **доверительный** (95%-ный) **интервал** генеральной доли пораженных фузариозом растений в совокупности (t_{05}=1,96, при N-1=500-1=499), $p \pm t_{05} s_p$=0,10±1,96·0,013=0,10±0,025 (0,075...0,125 или 7,5...12,5%).*

***Таким образом, генеральная доля растений, пораженных фузариозом** в изучаемой совокупности с 95%-ным уровнем вероятности, **составляет** 7,5...12,5%, **ошибка репрезентативности** - s_p=1,3%, **коэффициент вариации** - 60,0%.*

2 ТЕОРИЯ КОРРЕЛЯЦИИ И ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ КОРРЕЛЯЦИОННОГО АНАЛИЗА

В эколого-мелиоративных и гидрометеорологических исследованиях *редко приходится иметь дело с точными (функциональными) связями*, когда каждому значению одной величины соответствует строго определенное значение другой величины. Здесь *чаще* встречаются такие соотношения между переменными, когда каждому *значению признака (x) соответствует* не одно, а *множество* возможных *значений признака* (*y*), т.е. их *распределение*. *Такие связи*, обнаруживаемые лишь при массовом изучении признаков, в отличие от функциональных, *называются стохастическими (вероятностными)* или *корреляционными*.

Математический *анализ связей*, существующих *между случайными величинами (x) и (y)*, составляет *содержание корреляционного анализа*. Корреляционный анализ *сводится*, прежде всего, к *измерению степени тесноты сопряженности* между варьирующимися признаками, в качестве показателя которой наибольшее распространение получил линейный *коэффициент корреляции (r)*. Корреляционный анализ включает в себя также *определение формы и направления* существующей между (x) и (y) *связи* и др.

Однако, корреляцию не следует отождествлять с причинностью. Хотя необходимо иметь в виду, что *доказательство математической связи должно опираться на реальную зависимость между явлениями*, так как иногда можно установить несуществующие корреляции.

Например, *минерализация воды понижается с севера на юг Беларуси, в этом же направлении понижается содержание питательных веществ в почве. Между рассматриваемыми показателями может быть получена положительная достоверная зависимость. Однако, степень минерализации воды не определяет оптимальное содержание питательных веществ в почвенном покрове. Иначе в ландшафтах пустынь плодородие было бы максимальным, так как здесь самая высокая минерализация воды, а это противоречит истине. Поэтому, установление подобной связи при проведении теоретических исследований бессмысленно.*

Любой показатель связи служит приближенной оценкой рассматриваемой зависимости и не является гарантией существования жесткой (функциональной) соподчиненности. Отсутствие жестких зависимостей способствует саморегуляции процессов и явлений в природе. Вскрытие

корреляции в географической Среде позволяет предвидеть и прогнозировать закономерности развития природной Среды, в целом.

По форме корреляционная связь бывает *линейной* и *нелинейной* (криволинейной), по направлению - *прямой* и *обратной*, по величине - от 0 до ±1, по количеству коррелируемых признаков - *парной* и *множественной*.

Вообще, выделяется *несколько видов парной корреляционной связи*:

а) *параллельно-соотносительная*, или ассоциативная, когда оба признака изменяются сопряженно, частично под действием общих причин и следствий (приуроченность растительности и почв к определенным формам рельефа в лесостепной ландшафтной зоне);

б) *субпричинная*, когда один фактор выступает как отдельная причина сопряженного изменения признака (связь урожайности озимых зерновых культур с динамикой почвенных влагозапасов в вегетационный период);

в) *взаимоупреждающая*, когда причина и следствие, находясь в устойчивой взаимосвязи, последовательно влияют друг на друга (гидромелиорация и естественная увлажненность водосборов).

В теории корреляции принято выделять *две основные задачи. Первая задача* - установить форму корреляционной зависимости, или, как принято говорить в математической статистике, определить вид функции регрессии одной переменой (случайной) величины по другой. *Вторая задача* теории корреляции - оценить тесноту корреляционной зависимости.

2.1 Линейный коэффициент корреляции

Если *зависимость* между признаками *указывает на линейную корреляцию*, то обычно *рассчитывают коэффициент корреляции (r)*, который позволяет, с одной стороны, оценить тесноту связи переменных величин, с другой, - выяснить: какая доля изменений признака обусловлена влиянием основного фактора, какая - влиянием других факторов. При положительной зависимости величина коэффициента корреляции изменяется от 0 до +1, при отрицательной - от 0 до -1. Если r=0, то связь между признаками отсутствует. Принято считать, что при r<0,5 корреляционная связь слабая, при r=0,5...0,7 - средняя, при r=0,7...0,99 - сильная.

Коэффициент корреляции приближенно характеризует тесноту связи между признаками. Поэтому, иногда при высоком значении коэффици-

ента корреляции и небольшом объеме выборки связь между признаками может быть слабой. *Мерой корреляционной связи является величина (d_{xy})*, получившая название *коэффициента детерминации*, который определяется по формуле ($d_{xy}=r^2$).

Коэффициент детерминации *указывает на долю взаимной связи между признаками.* Например, если r=0,5, то d_{xy}=0,25, т.е. 25% всех изменений одного признака связано с изменением другого. Отсюда следует, что, при r≥0,70, истинная взаимообусловленность признаков составляет около 50%.

Одна и та же величина коэффициента корреляции будет по-разному определять достоверность зависимости признаков для малых и больших выборок. Например, при P=0,95, для n=5, достоверны значения r≥0,878, для n=20, - достоверной величиной будет r≥0,44, для n=100 - достоверны значения r≥0,196.

При работе с малыми выборками используется следующая формула для расчета коэффициента корреляции

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (2.1)$$

в которой (x_i-$\bar{x}$), (y_i-$\bar{y}$) - отклонения значений индивидуальных вариант (x_i; y_i) от их средних значений ($\bar{x}$; $\bar{y}$).

Любой выборочный (эмпирический) коэффициент корреляции (r_{xy}=r), являясь величиной случайной, может оказаться отличным от нуля даже при независимом варьировании признаков (x) и (y). Отсюда возникает *необходимость проверки надежности связи* (r_{xy}), тем более, что (r_{xy}) рассматривается в качестве оценки неизвестного (истинного) генерального параметра (ρ_{xy}). При оценке достоверности вычисленного коэффициента корреляции ($r_ф$) необходимо с помощью таблицы коэффициентов корреляции (Приложение, таблица П.8) сравнить его с табличным значением (r_T), а также установить достоверность коэффициента корреляции через t - критерий Стьюдента.

Если $r_ф > r_T$, то влияние фактора на признак достоверно; наоборот, если $r_ф < r_T$, то *коэффициент корреляции недостоверен и влияние фактора на признак несущественно.*

При использовании t-критерия Стьюдента для доказательства достоверности (r), вначале *определяется стандартная* (квадратическая) *ошибка* коэффициента корреляции по формуле

$$\sigma_r = \sqrt{\frac{1-r^2}{n-2}}, \qquad (2.2)$$

где n - число сопряженных пар в сравниваемых выборочных совокупностях. Значение коэффициента корреляции записывается с учетом его ошибки: ($r\pm\sigma_r$). Затем, проверяется нулевая гипотеза (H_o), применительно к оценке генерального (ρ_{xy}) по величине эмпирического коэффициента корреляции (r_{xy}), которая заключается в предположении, что между случайными величинами (x) и (y) корреляция отсутствует, т.е. (H_o:ρ_{xy}=0). Для проверки *нулевой гипотезы* используется t-критерий Стьюдента

$$t = \frac{|r|}{\sigma_r}. \qquad (2.3)$$

Выборочная функция (t) при условии (H_o) удовлетворяет распределению Стьюдента с (m=n-2) - степенями свободы. При уровне значимости (α) и (m) - степенях свободы по таблице t-распределения Стьюдента можно найти значение предела ($t_{\alpha;m}$). При $t \geq t_{\alpha;m}$, гипотеза должна быть отвергнута. Это значит, что выборочный коэффициент корреляции (r_{xy}) существенно отличен от нуля. Тогда можно принять, что в генеральной совокупности ($\rho_{xy}\neq 0$), т.е. случайные величины (x) и (y) не являются независимыми. При $t<t_{\alpha;m}$, гипотеза (H_o) не отвергается, отклонение (r_{xy}) от нуля - чисто случайного характера, $t_{\alpha;m}$ - двусторонний критерий. В случае одностороннего критерия, при уровне значимости (α), вычисленное значение t-критерия сравнивается с величиной ($t_{\alpha;m}$).

Для проверки надежности (r), вычисленных при достаточно больших n($n\geq 40$), можно воспользоваться рекомендациями В.И. Романовского

$$\frac{|r|}{\sigma_r} > t_{\alpha,m}, \qquad (2.4)$$

где

$$\sigma_r = \frac{1-r^2}{\sqrt{n}} \qquad (2.5)$$

рассматривается как ошибка вычисления (r).

Этот критерий менее жесткий, чем предыдущий, но во многих случаях является достаточным.

В тех случаях, когда эмпирический коэффициент корреляции мал ($|r| \leq 0{,}5$), возникает вопрос о том, не являются ли в действительности случайные величины некоррелированными, т.е. не объясняется ли наличие малого (r) случайными ошибками измерений (эксперимента). Фактически речь идет о проверке гипотезы ($H_o: \rho_{xy} = \rho_o = 0$).

Если для эмпирического значения (r) произведение $|r| \cdot \sqrt{n-1}$ окажется больше некоторого критического значения (таблица 2.1), то с надежностью (P) можно утверждать, что (r) истинный ($\rho_o \neq 0$), т.е. гипотеза (H_o) отвергается.

По В.И. Романовскому, если $|r| \cdot \sqrt{n-1} \geq 3$, можно считать (r) значимым и связь реальной. Если $|r| \cdot \sqrt{n-1} < 3$, то (r) отличается от нуля с большой вероятностью лишь случайно.

Таблица 2.1 Критические значения $|r| \cdot \sqrt{n-1}$

n	P=0,95	P=0,99	P=0,999	n	P=0,95	p=0,99	p=0,999
10	1,89	2,29	2,62	25	1,94	2,47	3,03
11	1,90	2,32	2,68	30	1,94	2,49	3,07
12	1,91	2,35	2,73	35	1,95	2,50	3,10
13	1,91	2,37	2,77	40	1,95	2,51	3,13
14	1,92	2,39	2,81	50	1,95	2,527	3,160
15	1,92	2,40	2,84	60	1,953	2,536	3,184
16	1,93	2,41	2,87	70	1,954	2,541	3,198
17	1,93	2,42	2,90	80	1,955	2,546	3,209
18	1,93	2,43	2,92	90	1,956	2,550	3,219
19	1,93	2,44	2,94	100	1,956	2,553	3,226
20	1,94	2,45	2,96	∞	1,960	2,576	3,291

Когда (n) достаточно велико и (r) не слишком близко к единице, распределение выборочных коэффициентов корреляции стремится к нормальному закону с центром распределения, равным ($\bar{r}$), и средним квадратическим отклонением $\left(\sigma_r = \frac{1-r^2}{\sqrt{n-1}}\right)$. При малом (n), когда $|r| \to 1$, распределение выборочных коэффициентов корреляции все более уклоняется от нормального. Коэффициент корреляции, рассчитанный по выборке конечного объема (n), в среднем всегда меньше коэффициента корреля-

ции генеральной совокупности, т.е. выборочные коэффициенты корреляции имеют отрицательное смещение. Это смещение уменьшается с увеличением (n).

Нормальное распределение выборочных коэффициентов корреляции приблизительно сохраняется, когда значение (n) не слишком мало, а (r) - не слишком велико. Во всех остальных случаях (при малом (n) и большом "r") распределение выборочных (r) асимметрично.

Для коэффициентов корреляции, полученных по выборкам из распределения, отличного от нормального, закон распределения выборочных (r), вообще говоря, неизвестен и, следовательно, оценка эмпирического коэффициента корреляции затруднена. При малых объемах выборок (n<50) и, особенно при больших (r), для оценки случайного рассеивания выборочных коэффициентов корреляции обычно используется преобразование Фишера, основанное, в свою очередь, на использовании специальной переменной (Z), функционально связанной с (r) выражением

$$Z_r = \frac{1}{2}\ln\frac{1+r}{1-r} = 1{,}1513\cdot\lg\frac{1+r}{1-r}\,. \qquad (2.6)$$

Величина (Z_r) распределена асимптотически нормально с дисперсией (σ_{Z_r})

$$\sigma_{Z_r} = \frac{1}{\sqrt{n-3}} \qquad (2.7)$$

и математическим ожиданием ($\overline{Z}_r$)

$$\overline{Z}_r = \frac{1}{2}\ln\frac{1+\rho_o}{1-\rho_o} + \frac{\rho_o}{2\cdot(n-1)}\,. \qquad (2.8)$$

Величина

$$U_r = \frac{Z_r - \overline{Z}_r}{\sigma_{Z_r}} \qquad (2.9)$$

приближенно удовлетворяет нормированному нормальному закону распределения. В силу основного свойства величины (Z_r), имеем

$$P(-x\cdot\sigma_{Z_r} < Z_r - \overline{Z}_r < x\cdot\sigma_{Z_r}) = 2\cdot\Phi(x) - 1\,, \qquad (2.10)$$

где P - символ вероятности события, описываемого содержанием круглых скобок; Φ(x) - функция нормированного нормального распределения

$$\Phi(x) = \frac{1}{2\cdot\sqrt{\pi}}\int_{-\infty}^{x}\exp\left(\frac{-t^2}{2}\right)\cdot dt\,. \qquad (2.11)$$

Задавая определенное (достаточно большое) значение разности $(2\cdot\Phi(x)-1)$ в формуле (2.10), по таблицам нормированного нормального распределения, найдем (x), после чего можно найти границы доверительного интервала неизвестного (ρ_o).

Нередко возникает необходимость сравнения двух коэффициентов корреляции (r_1) и (r_2), полученных по выборкам (n_1) и (n_2) - пар наблюдений одних и тех же признаков (x) и (y). Это достигается путем проверки нулевой гипотезы (H_o:$\rho_1=\rho_2$), которая гласит, что выборки взяты из двух совокупностей с одинаковыми коэффициентами корреляции. Для проверки нулевой гипотезы может быть использовано преобразование Фишера. Для оценки существенности расхождения между выборочными коэффициентами корреляции (r_1) и (r_2), вычисляется отношение

$$Z=\frac{|Z_1-Z_2|}{\sqrt{\sigma^2_{Z_1}+\sigma^2_{Z_2}}}, \qquad (2.12)$$

где Z_i и $\sigma^2_{Z_i}$ - определяются по формулам (2.6) и (2.7).

Выбрав по таблице 2.2 пороговое значение (Z_α), соответствующее уровню значимости (α), сравним его с величиной (Z), найденной по формуле (2.12). Если $|Z|<Z_\alpha$, делается вывод, что гипотеза (H_o) не отвергается и, следовательно, расхождение между (r_1) и (r_2) можно считать случайным. Это также значит, что в (100-α)% - случаях гипотеза (H_o) будет отвергнута.

Таблица 2.2 Значения (Z_α) для различных (α) при двустороннем ограничении

α	0,001	0,027	0,010	0,0455	0,050
P=1-α	0,999	0,9973	0,990	0,9545	0,950
Z_α	3,291	3,000	2,576	2,000	1,960

В таблице 2.2, наряду с (α), показаны также значения (P=1-α) - статистической достоверности гипотезы (H_o). Разность (Z_1-Z_2) имеет также нормальное распределение с дисперсией

$$\sigma^2_{Z_1\cdot Z_2}=\sigma^2_{Z_1}+\sigma^2_{Z_2} \qquad (2.13)$$

и средней, равной нулю. Поэтому, *вероятность (P) определяется как*

$$P(|Z_1-Z_2|)\geq Z\cdot\sigma_Z=2-2\cdot\Phi(Z), \qquad (2.14)$$

где Z - вычисляется по формуле (2.12); $\Phi(Z)=\Phi(x)$ - функция нормированного распределения - по (2.14).

При значительной вероятности (P), *расхождение* между Z_1 и Z_2, определяемое различием (r_1) и (r_2), имеет *случайный характер*, а *(r_1) и (r_2) несущественно отличаются друг от друга.*

Важной особенностью применения Z-критерия Фишера является то, что в основу расчета коэффициентов корреляции (r_1) и (r_2) должна быть положена одна и та же группировка данных.

В качестве примера, рассмотрим применение коэффициента корреляции при решении задачи выбора реки-аналога.

Имеется ряд наблюдений за годовым расходом воды в реке Птичь-с.Лучицы; необходимо подобрать реки с синхронными колебаниями водности. Исходя из комплексного анализа физико-географических, гидрографических характеристик, отбираем реки - кандидаты (Ясельда-с.Сенин, Оресса-с.Верхутино, Оресса-с.Андреевка). Гидрометрические данные по реке Птичь-с.Лучицы и по рекам - кандидатам приведены в таблице 2.3.

Таблица 2.3 Расходы воды рек, м³/с

N п/п	*Птичь-с.Лучицы*	*Ясельда-с.Сенин*	*Оресса-с.Верхутино*	*Оресса-с.Андреевка*
1	*2*	*3*	*4*	*5*
1	*46,6*	*11,5*	*2,52*	*12,8*
2	*48,8*	*17,6*	*2,43*	*16,3*
3	*33,1*	*19,0*	*1,68*	*10,4*
4	*27,3*	*14,6*	*2,07*	*9,85*
5	*49,3*	*21,4*	*3,06*	*13,8*
6	*24,1*	*9,91*	*1,64*	*9,14*
7	*41,3*	*17,4*	*2,38*	*13,4*
8	*17,6*	*4,0*	*0,64*	*6,52*
9	*40,8*	*14,2*	*2,22*	*16,3*
10	*54,5*	*14,8*	*3,06*	*20,0*
11	*42,6*	*18,3*	*2,61*	*16,2*
12	*90,8*	*38,1*	*2,58*	*36,0*
13	*41,9*	*12,4*	*1,97*	*15,2*
14	*36,2*	*14,9*	*1,49*	*13,5*
15	*55,2*	*18,5*	*3,76*	*20,4*
16	*38,2*	*12,2*	*2,54*	*14,1*

Продолжение таблицы 2.3

17	35,7	9,74	2,37	13,4
18	49,1	13,5	3,14	18,8
19	48,7	20,1	3,30	19,7
20	49,3	22,6	4,15	20,3
21	50,5	23,2	3,84	20,0
22	42,1	15,2	2,93	16,9
23	73,2	37,6	4,39	26,5
24	57,4	29,6	3,40	20,9
25	32,3	12,0	2,51	13,7
26	36,9	15,7	3,02	15,7
27	44,4	31,7	3,43	19,2
28	60,8	29,6	3,57	23,1
$\bar{x} = \sum_{i=1}^{28} (x_i)/28$	45,31	18,55	2,74	16,86
$\sum_{i=1}^{28} (x_i - \bar{x})^2$	5737,61	1854,75	19,52	917,15

Для вычисления коэффициентов корреляции расходов реки Птичь- с.Лучицы с расходами других рек (Ясельда-с.Сенин; Оресса-с.Верхутино; Оресса-с.Андреевка) и соответствующих статистик использовались формулы (2.1) и (2.12):

$$Z_\alpha = \frac{|1,83 - 1,16|}{\sqrt{0,19^2 + 0,19^2}} = 2,49; \qquad r_{(П.-Л.)-(О.-А.)} = \frac{2187,84}{\sqrt{5737,61 \cdot 917,15}} = 0,95;$$

$r_{(П.-Л.)-(Я.-С.)} = 0,82; \qquad r_{П.-Л.-О.-В.} = 0,65.$

Таким образом, предпочтение отдается реке Оресса-с.Андреевка, так как для реки Птичь-с.Лучицы она является лучшим аналогом, ($r_{(П.-Л.)-(О.-А.)}$=0,95), а колебания водности этих рек наиболее синхронные (рисунок 2.1). Коэффициент детерминации $d_{(П.-Л.)-(О.-А.)}$=0,90, т.е. 90% всех колебаний этих рек взаимообусловлены. При этом, стандартная ошибка коэффициента корреляции ($r_{(П.-Л.)-(О.-А.)}$), рассчитанная по формуле (2.2), составляет

$$\sigma_r = \sqrt{\frac{1 - 0,95^2}{28 - 2}} = 0,06,$$

и значение коэффициента корреляции запишется как r=0,95±0,06, т.е. он изменяется от 0,89 до 1,0.

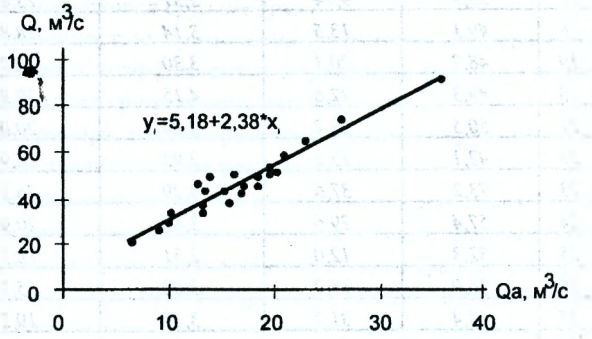


Рисунок 2.1 Связь годового стока рек Птичь-с.Лучицы и Оресса-с.Андреевка.

***Проверим нулевую гипотезу**, т.е. (H_0:ρ=0), при H_1:ρ>0, используя t-критерий Стьюдента (2.3)*

$$t = \frac{|0,95|}{0,06} = 15,83.$$

*Теоретическое значение t-критерия Стьюдента (при m=28-2=26 - степенях свободы и α=0,05) будет - $t^T(26;0,05)$=2,05. Так как 15,83>2,05 **отвергаем нулевую гипотезу и принимаем альтернативную** - коэффициент корреляции отличен от нуля. Для проверки нулевой гипотезы можно воспользоваться таблицей проверки коэффициента корреляции на значимость относительно нуля (Приложение, таблица П.8). В нашем случае, критическое значение коэффициента корреляции ($r_{кр.}$=0,374), что значительно меньше полученного. Следовательно, **нулевая гипотеза отвергается. Выполним еще одну проверку нулевой гипотезы**: определим величину -*

$|r| \cdot \sqrt{n-1} = 0,95 \cdot \sqrt{28-1} = 4,9$; *сопоставим эту величину с данными таблицы 2.1. - 4,9>1,94 (при P=0,95), 4,9>2,48 (при P=0,99) и 4,9>3,03 (при P=0,999), убедившись, что **нулевая гипотеза отвергается**. Как было сказано выше, **для малых выборок при значениях (|r|→1), распределение выборочных коэффициентов корреляции** заведомо отличается **от нормального**. Поэтому t - критерий Стьюдента становится ненадежным. Чтобы обойти это затруднение, воспользуемся преобразованием Фишера. В данном случае, согласно (2.6), (2.7), (2.8), соответственно, имеем:*

переменную - $$Z_r = \frac{1}{2} ln \frac{1+0,95}{1-0,95} = 1,83;$$

дисперсию - $\sigma_{Z_r} = \frac{1}{\sqrt{28-3}} = 0{,}19;$

математическое ожидание - $\overline{Z}_r = \frac{1}{2} ln \frac{1+0{,}95}{1-0{,}95} + \frac{0{,}95}{2 \cdot (28-1)} = 1{,}85;$

величину U_r - $U_r = \frac{|1{,}83 - 1{,}85|}{0{,}19} = 0{,}11.$

Находим интервал изменения (r), соответственно интервалу изменения (Z)

$$Z_r \pm \sigma_{Z_r} = 1{,}83 \pm 0{,}19 = (\frac{2{,}02 - \text{верхняя граница}}{1{,}64 - \text{нижняя граница}}).$$

По верхней и нижней границам (Z_r) находим границы изменения "r" (по специальным таблицам), откуда r=0,965 - верхняя граница и r=0,928 - нижняя граница, т.е. $r = (\frac{0{,}965}{0{,}928})$.

Сравним различимы ли между собой $r_{(П.-Л.)-(О.-А.)}$=0,95 *и* $r_{(П.-Л.)-(Я.-С.)}$=0,82. ***Проверим нулевую гипотезу*** (H_0:$\rho_1=\rho_2$), *которая гласит, что выборки должны быть взяты из двух совокупностей с одинаковыми коэффициентами корреляции.*

Для этого, используя формулы (2.6) и (2.7), вычислим следующие параметры:

$$Z_{r\,(П.-Л.)-(О.-А.)} = \frac{1}{2} ln \frac{1+0{,}95}{1-0{,}95} = 1{,}83;$$

$$\sigma_{Z\,(П.-Л.)-(О.-А.)} = \frac{1}{\sqrt{28-3}} = 0{,}19;$$

$$Z_{r\,(П.-Л.)-(Я.-С.)} = \frac{1}{2} ln \frac{1+0{,}82}{1-0{,}82} = 1{,}16;$$

$$\sigma_{Z\,(П.-Л.)-(Я.-С.)} = \frac{1}{\sqrt{28-3}} = 0{,}19.$$

Воспользовавшись формулой (2.12), имеем

$$Z_\alpha = \frac{|1{,}83 - 1{,}16|}{\sqrt{0{,}19^2 + 0{,}19^2}} = 2{,}49.$$

Выберем по таблице 2.2 пороговое значение (Z_α =1,96), *при уровне значимости* ($\alpha = 0{,}05$). *Так как 2,49>1,96, -* ***нулевую гипотезу отвергаем и, следовательно, расхождения между коэффициентами корреляции нельзя считать случайными в 95%-тах случаев.***

2.2 Корреляционное отношение

Если *мерой тесноты связи при линейной ее форме* служит *линейный коэффициент корреляции*, то *для криволинейной зависимости* такой мерой служит другой *показатель*, предложенный К. Пирсоном и *называемый*

корреляционным отношением. Корреляционное отношение имеет разные значения для корреляционной связи (y=f(x)) и для связи (x=f(y)). *Прямое корреляционное отношение* (y/x) вычисляется по формуле

$$\eta_{y/x} = \sqrt{\frac{k\sum_{j=1}^{m}(\frac{y_j}{x_j} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (2.15)$$

где (y_j/x_j) - точка эмпирической линии регрессии; k - число вариант в частной группе; $(y_j/x_j - \bar{y})$ - отклонение точек эмпирической линии регрессии от общего среднего по $(\bar{y})$. *Обратное корреляционное отношение* (x/y) для малых выборок вычисляется по формуле

$$\eta_{x/y} = \frac{k \cdot \sum_{j=1}^{m}\left(\frac{x_j}{y_j} - \bar{x}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (2.16)$$

где $(x_j/y_j - \bar{x})$ - отклонение точек эмпирической регрессии от общей средней по $(\bar{x})$; $(x_i - \bar{x})$ - отклонение индивидуальных вариант в выборке от общей средней по $(\bar{x})$.

Основные *свойства корреляционного отношения*:

1) корреляционное отношение всегда положительно и изменяется в пределах ($0 \le \eta \le 1$). При этом, если между (y) и (x) нет корреляционной связи, то ($\eta=0$), а если (y) связано с (x) функциональной связью, то ($\eta=1$);

2) корреляционное отношение всегда не меньше численного значения соответствующего коэффициента корреляции, т.е. ($\eta \ge |r|$);

3) если ($\eta_{y/x}=|r|$), - регрессия (y) на (x) точно линейна; если ($\eta_{x/y}=|r|$), регрессия (x) на (y) точно линейна;

4) корреляционные отношения ($\eta_{y/x}$) и ($\eta_{x/y}$) обычно не равны между собой ($\eta_{y/x} \ne \eta_{x/y}$); лишь при строго линейной связи между (x) и (y) наблюдается равенство - ($\eta_{y/x}=\eta_{x/y}$). Таким образом, чем больше связь между (x) и (y) приближается к прямолинейной, тем ближе по величине ($\eta_{y/x}$) и ($\eta_{x/y}$).

Для оценки достоверности полученных значений (η), можно *использовать t-критерий*

$$t = \eta \cdot \sqrt{\frac{n-2}{1-\eta^2}}. \qquad (2.17)$$

Нулевая гипотеза об отсутствии связи между признаками отвергается, если ($t \geq t_{\alpha;m}$), где α - уровень значимости; m=n-2 - число степеней свободы. Параметр ($t_{\alpha;m}$) определяется по таблицам t-распределения Стьюдента. *Коэффициент корреляции (r) характеризует только линейную связь, а корреляционное отношение - любую форму связи.* При строго линейной связи ($\eta_{y/x}=\eta_{x/y}=|r|$). При наличии нелинейной связи ($\eta_{y/x}\neq\eta_{x/y}$) и ($\eta\neq|r|$). Следовательно, по разности между этими показателями можно судить о форме корреляционной зависимости между варьирующимися признаками.

Для решения вопроса - является ли исследуемая зависимость линейной или криволинейной - используется *критерий криволинейности.* Существует несколько способов оценки степени криволинейности. Рассмотрим некоторые из них. *Наиболее простой*, но менее строгий, способ заключается в определении разности коэффициентов корреляции и корреляционного отношения; при этом, используется неравенство ($\eta^2-r^2\geq 0{,}1$). Корреляция считается криволинейной, если полученный результат соответствует этому неравенству. *Второй способ* оценки степени криволинейности связан с применением t-критерия Стьюдента

$$t = \gamma \cdot \sqrt{\frac{n-2}{\gamma - \gamma^2 \cdot (2-\eta^2-r^2)}}, \qquad (2.18)$$

где

$$\gamma = \eta^2 - r^2. \qquad (2.19)$$

Если при уровне значимости (α) и (m=n-2) - степенях свободы ($t<t_{\alpha;m}$), то корреляция между признаками с (P-α) - надежностью оценивается как прямолинейная. При ($t\geq t_{\alpha;m}$), зависимость между признаками следует считать заметно отличающейся от прямолинейной.

Можно также *использовать F-критерий* (Фишера) *для определения степени приближения криволинейной зависимости к прямолинейной*

$$F = \frac{(\eta^2 - r^2)\cdot(n-k_x)}{(1-\eta^2)\cdot(k_x-2)}, \qquad (2.20)$$

где n - объем выборки; k_x - число групп по ряду (x) .

Связь можно практически принять за линейную, если ($F<F_T$) и определять показатели для прямолинейной корреляции и регрессии. Корреляция нелинейна, если ($F \geq F_T$). Теоретические значения (F_T) берутся из Приложения (таблица П.4.1), для $\nu_1=(k_x-2)$ и $\nu_2=(n-2)$ - степеней свободы.

Для примера, рассмотрим вычисление корреляционного отношения ($\eta_{y/x}$) и ($\eta_{x/y}$), *используя данные о средних многолетних годовых расходах воды малых рек Беларуси Y(м^3/с) и площадях водосборов (км^2) (таблица 2.4).*

Таблица 2.4 Значения расходов малых рек Беларуси и площадей их водосборов

Y - расход (м^3/с)	*0,65*	*1,15*	*1,25*	*1,75*	*2,25*	*2,10*	*2,55*	*2,50*	*3,05*
X - площадь водосбора (км^2)	*50*	*25*	*75*	*50*	*100*	*150*	*150*	*200*	*250*
Y - расход (м^3/с)	*2,75*	*3,15*	*3,60*	*3,50*	*4,00*	*3,75*	*3,85*	*4,00*	*4,35*
X - площадь водосбора (км^2)	*300*	*400*	*450*	*550*	*650*	*700*	*800*	*900*	*1000*

Из графика видно, что кривая Q=f (F) близка к логарифмической, зависимость положительная (рисунок 2.2).

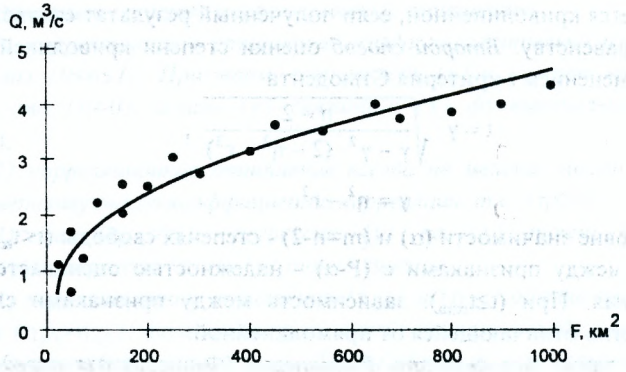


Рисунок 2.2 Зависимость расходов воды (Q=Y) малых рек Беларуси от их площадей водосборов (F=X).

На основании данных (таблица 2.4), рассчитываем корреляционное отношение ($\eta_{y/x}$) между (x) и (y). Выборки разделим на шесть частных групп, однако, дальней-

шие операции проводим с (y). Рассчитываем общее среднее ($\bar{y}$). Точки ***эмпирической*** *линии регрессии (y/x) представляют собой среднее арифметическое частных групп. Записываем отклонения точек эмпирической линии регрессии от общего среднего по (y), возводим эти отклонения в квадрат и суммируем. Далее, вычисляем отклонения индивидуальных показателей (y_i) от общего среднего ($\bar{y}$) и суммируем (сумма должна быть равна нулю или близка к нему). Каждое отклонение возводим в квадрат и суммируем. Результаты расчетов сводим в таблицу 2.5.*

Таблица 2.5 Результаты вычисления прямого (η y/x) корреляционного отношения для невзвешенных рядов

y_i (расход, $м^3/с$)	Σy_i (по группам)	y/x	$(y/x - \bar{y})$	$(y/x - \bar{y})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	2	3	4	5	6	7
1 группа						
0,65					-2,14	4,5796
1,15	3,05	1,02	-1,77	3,1447	-1,64	2,6896
1,25					-1,54	2,3716
2 группа						
1,75					-1,04	1,0816
2,10	6,10	2,03	-0,76	0,5725	-0,69	0,4761
2,25					-0,54	0,2916
3 группа						
2,50					-0,29	0,0841
2,55	8,10	2,70	-0,09	0,0081	-0,24	0,0576
3,05					0,26	0,0676
4 группа						
2,75					-0,04	0,0016
3,15	9,50	3,17	0,38	0,1419	0,36	0,1229
3,60					0,81	0,6561
5 группа						
3,50					0,71	0,5041
4,00	11,25	3,75	0,96	0,9216	1,21	1,4641
3,75					0,96	0,9216

Продолжение таблицы 2.5

6 группа						
3,85 *4,00* *4,35*	*12,20*	*4,07*	*1,28*	*1,6299*	*1,06* *1,21* *1,56*	*1,1236* *1,4641* *2,4336*
$\sum_{i=1}^{18} y_i = 50,2$ $\bar{y} = 2,79$			*Σ0*	*Σ6,4187*	*Σ-0,02≈0*	*Σ20,3978*

Полученные данные используем для расчета прямого корреляционного отношения по формуле (2.15)

$$\eta_{y/x} = \sqrt{\frac{3 \cdot 6,4187}{20,3978}} = 0,972.$$

Достоверность результатов определим по t-критерию Стьюдента, учитывая зависимость (2.17)

$$t_{\eta_{y/x}} = 0,972 \cdot \sqrt{\frac{18-2}{1-0,972^2}} = 14,33.$$

Поскольку *($t_Ф = 14,33 > t_T = 1,746$), при P=0,95, для ν=16 (по Приложению, таблица П.2),* ***значение корреляционного отношения следует признать доказанным; зависимость между расходами воды рек (Y=Q) и площадями водосборов (X=F) положительна и достоверна.*** *Аналогично* ***вычисляем обратное корреляционное отношение*** *($\eta_{x/y}$) и результаты расчета сводим в таблицу 2.6.*

Таблица 2.6 Результаты вычисления обратного ($\eta_{x/y}$) корреляционного отношения для невзвешенных рядов

x_i (площадь, км²)	Σx_i (по группам)	x/y	$(x/y - \bar{x})$	$(x/y - \bar{x})^2$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	*2*	*3*	*4*	*5*	*6*	*7*
1 группа						
50 *25* *75*	*150*	*50*	*-328*	*107584*	*-328* *-353* *-303*	*107584* *124609* *91809*
2 группа						
50 *100* *150*	*300*	*100*	*-278*	*77284*	*-328* *-278* *-228*	*107584* *77284* *51984*

Продолжение таблицы 2.6.

3 группа						
150 *200* *250*	*600*	*200*	*-178*	*31684*	*-228* *-178* *-128*	*51984* *31684* *16384*
4 группа						
300 *400* *450*	*1150*	*383*	*5*	*25*	*-78* *22* *72*	*6084* *484* *5184*
5 группа						
550 *650*	*1900*	*633*	*255*	*65025*	*172* *272*	*29584* *73984*
700					*322*	*103684*
6 группа						
800 *900* *1000*	*2700*	*900*	*522*	*272484*	*422* *522* *622*	*178084* *272484* *386884*
$\sum_{i=1}^{18} x_i = 6800$ $\bar{x} = 378$			Σ-2≈0	Σ554086	Σ-4≈0	Σ1717362

Получим корреляционное отношение

$$\eta_{x/y} = \sqrt{\frac{3 \cdot 554086}{1717362}} = 0{,}980,$$

а также t-критерий Стьюдента

$$t_{\eta_{y/x}} = 0{,}980 \cdot \sqrt{\frac{18-2}{1-0{,}980^2}} = 20{,}10.$$

Так как $t_Ф$=20,10>t_T=1,746, при P=0,95 для ν=16 (Приложение, таблица П.2), то ***значение корреляционного отношения следует признать доказанным.***

Оценим линейность связи между расходами воды рек и площадями водосборов *(на основе данных таблицы 2.4)*

$$r=0{,}90;\ \eta_{y/x}=0{,}972;\ \gamma=0{,}972^2-0{,}90^2=0{,}135;$$

$$t_{y/x} = 0{,}135 \cdot \sqrt{\frac{18-2}{0{,}135 - 0{,}135^2 \cdot (2 - 0{,}972^2 - 0{,}90^2)}} = 1{,}495.$$

По t-распределению Стьюдента (Приложение, таблица П.2) для α=0,10 и m=16 находим - $t_{0{,}05;16}$ =1,337. Полученное значение t - критерия превышает ($t_{\alpha,m}$), при α=0,10 и m=16, и это позволяет заключить, что зависимость между расходами воды рек и площадями водосборов отличается от прямолинейной.

Оценим также линейность связи между площадями водосборов и расходами воды рек

$$\gamma = 0{,}980^2 - 0{,}90^2 = 0{,}154;$$

$$t_{\eta_{x/y}} = 0{,}154 \cdot \sqrt{\frac{18-2}{0{,}154 - 0{,}154^2 \cdot (2 - 0{,}980^2 - 0{,}90^2)}} = 1{,}579.$$

Полученное значение также превышает теоретическое значение t - критерия Стьюдента, при α=0,10 и m=16, и зависимость между этими признаками также отличается от линейной.

*Проверим **степень приближения криволинейной зависимости к прямолинейной**, используя критерий Фишера (2.20), согласно формулы (2.20)*

$$F = \frac{(0{,}972^2 - 0{,}90^2) \cdot (18-6)}{(1 - 0{,}972^2) \cdot (6-2)} = 9{,}76.$$

*Используя материалы Приложения (таблицы П.4.1), для ν_1=6-2=4 и ν_2=18-2=16 --степеней получаем $F^T_{(4,16,5\%)}$=3,01 и так как это значение меньше F_Φ - **корреляция может быть признана нелинейной**.*

2.3 Множественный коэффициент корреляции

При изучении многофакторных природных процессов и, в частности, при построении их расчетных и прогностических моделей достаточно часто возникает *необходимость установления вида линейной корреляционной зависимости между несколькими переменными.* Для примера, на рисунке 2.3 приведена схема упорядоченности расположения переменных в анализе функций экологической надежности (F_2) и эколого-социальных последствий (F_1) при мелиоративном преобразовании больших территорий.

Для решения подобных задач целесообразно привлекать аппарат множественной корреляции. Сущность этого подхода состоит в распределении основных положений метода линейной корреляции двух переменных (y) на случай зависимости интересующей нас переменной, от произвольного числа аргументов (x).

Корреляция называется множественной, если на величину результативного признака одновременно влияют несколько факториальных.

При множественном корреляционном анализе вычисляются два типа парных коэффициентов корреляции:

1) r_{yx_j} - коэффициенты, определяющие тесноту связи между функцией отклика и одним из факторов (x_j);

2) $r_{x_j x_m}$ - коэффициенты, показывающие тесноту связи между одним из факторов и фактором - $x_m(j,m=\overline{1,k})$, где k- число факторов.

Для вычисления коэффициентов корреляции используется формула (2.1). Значимость парных коэффициентов корреляции можно проверить способами, рассмотренными в разделе 2.1. Если один из коэффициентов ($r_{x_j x_m}$) окажется равным 1, то факторы (x_j) и (x_m) функционально (невероятностно) связаны между собой и тогда целесообразно один из них исключить из рассмотрения, причем, оставив тот фактор, у которого коэффициент (r_{y,x_j}) - больше.

После вычисления всех парных коэффициентов корреляции и исключения из рассмотрения того или иного фактора можно построить матрицу коэффициентов корреляции вида:

$$\begin{Vmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_j} & \dots & r_{yx_k} \\ r_{x_1y} & 1 & r_{x_1y_2} & \dots & r_{x_1x_j} & \dots & r_{x_1x_k} \\ r_{x_2y} & r_{x_2x_1} & 1 & \dots & r_{x_2x_j} & \dots & r_{x_2x_k} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{x_jy} & r_{x_jx_1} & r_{x_jx_2} & \dots & 1 & \dots & r_{x_jx_k} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{x_ky} & r_{x_kx_1} & r_{x_kx_2} & \dots & r_{x_kx_j} & \dots & 1 \end{Vmatrix} \qquad (2.21)$$

Используя матрицу (2.21), можно вычислить *частные коэффициенты корреляции,* которые показывают степень влияния одного из факторов (x_i) на функцию отклика (y) при условии, что остальные факторы имеют постоянный уровень (закреплены на нем). Формула для вычисления частных коэффициентов корреляции следующая

$$r_{y*x_1,x_2,x_3,\ldots,x_j,\ldots,x_k} = \frac{D_{1j}}{\sqrt{D_{11} \cdot D_{jj}}}, \qquad (2.22)$$

где D_{1j} -определитель матрицы, образованной из матрицы (2.21) вычеркиванием 1-й строки j-го столбца. Определители (D_{11}) и (D_{jj}) вычисляются аналогично, как и парные коэффициенты; частные коэффициенты корреляции изменяются от (-1) до (+1).

Значимость и доверительный интервал для коэффициентов частной корреляции определяются как для коэффициентов парной корреляции, только число степеней свободы вычисляется по формуле

$$\nu = n - k' - 2, \qquad (2.23)$$

где $k'=(k-1)$ - порядок частного коэффициента парной корреляции.

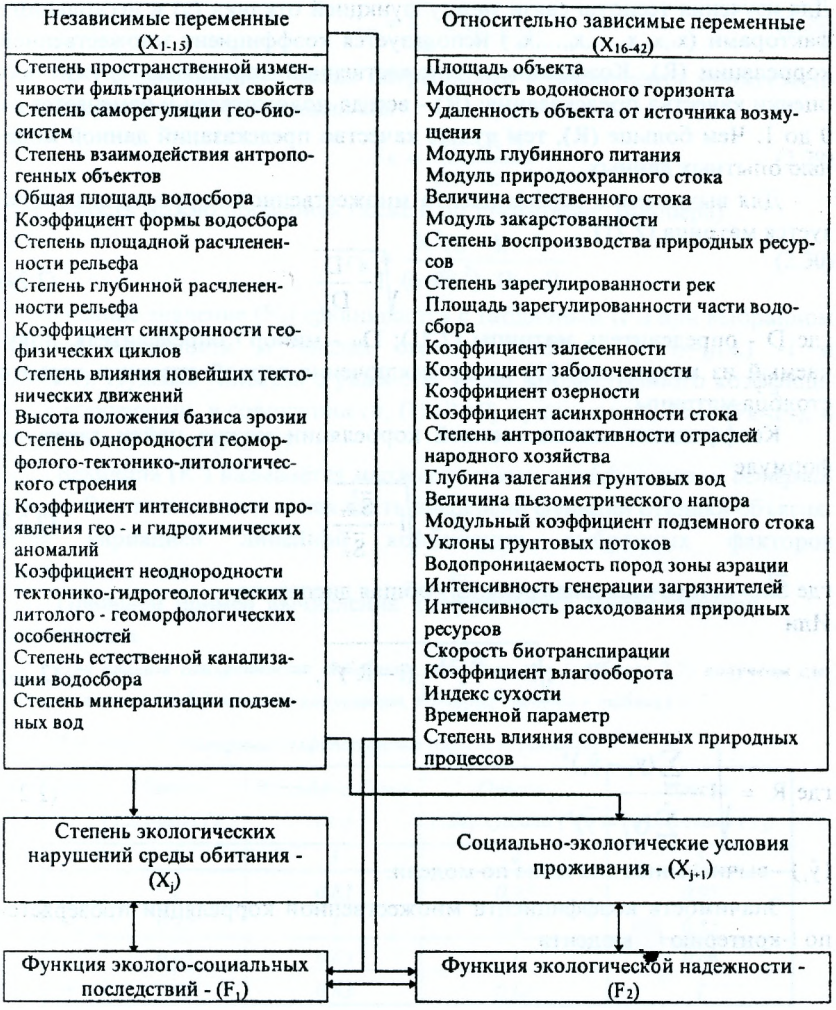


Рисунок 2.3 Схема упорядоченности расположения переменных в анализе функций (F_1) и (F_2) при мелиоративном преобразовании территорий.

Для изучения тесноты связи между функцией отклика (y) и несколькими факторами ($x_1, x_2, x_3, \ldots, x_j, \ldots, x_k$) используется коэффициент множественной корреляции (R). Коэффициент множественной корреляции служит для оценки качества предсказания; (R) - всегда положителен и изменяется от 0 до 1. Чем больше (R), тем лучше качество предсказаний данной моделью опытных данных.

Для вычисления коэффициента множественной корреляции используется матрица (2.21)

$$R = \sqrt{1 - \frac{D}{D_{11}}}, \qquad (2.24)$$

где D - определитель матрицы (2.21); D_{11} - минор*) определителя, получаемый из матрицы (2.21) путем исключения первой строки и первого столбца матрицы.

Коэффициент множественной корреляции можно найти также по формуле

$$R = \sqrt{1 - \frac{\bar{S}^2_{ост.}}{\bar{S}^2_y}}, \qquad (2.25)$$

где $\bar{S}^2_{ост.}$ - остаточная дисперсия; $\bar{S}^2_y$ - общая дисперсия.

Или

$$R = \sqrt{1 - \frac{n-1}{n-k-1}(1-R')^2}, \qquad (2.26)$$

где $R' = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$; (2.27)

($\hat{y}_i$) - вычисленное значение по модели.

Значимость коэффициента множественной корреляции проверяется по t-критерию Стьюдента

*) - **минор k-го порядка матрицы** - определитель матрицы, составленный из элементов данной матрицы, стоящих на пересечении произвольно выделенных ее (k) - строк и (k) - столбцов с сохранением их порядка, т.е. минор k-го порядка есть определитель квадратичной матрицы размера (k*k).

$$t_R = \frac{R}{\bar{S}_R} \geq t^T_{(n-k-1)}, \qquad (2.28)$$

где $\bar{S}_R$ - среднеквадратическая погрешность коэффициента множественной корреляции, определяемая как

$$\bar{S}_R = \frac{(1 - R^2)}{\sqrt{n - k - 1}}. \qquad (2.29)$$

Значимость можно проверить также и по F-критерию (Фишера)

$$F_R = \frac{R^2(n - k - 1)}{(1 - R^2) \cdot (k - 1)}. \qquad (2.30)$$

Полученное значение (F_R) сравнивается с табличным (F_T) при выбранном уровне значимости и числах степеней свободы - $\nu_1=(n-k)-1$ и $\nu_2=(k-1)$. Нулевая гипотеза о равенстве нулю множественного коэффициента корреляции, в совокупности, (H_o:R=0) принимается, если ($F_R<F_T$), и отвергается - если ($F_R \geq F_T$).

Величина (R^2) называется *множественным коэффициентом детерминации*. Она показывает, какая часть дисперсии функции отклика объясняется вариацией линейной комбинации выбранных факторов ($x_1, x_2, x_3, \ldots, x_j, \ldots, x_3$).

Приведем пример вычисления коэффициента множественной корреляции.

При анализе синхронности колебания водности рек *(таблица 2.3)* ***получены*** *следующие* ***парные коэффициенты корреляции,*** *которые сведены в таблицу 2.7.*

Таблица 2.7 Матрица коэффициентов парной корреляции

j|m	*Птичь-с.Лучицы (y)*	*Ясельда -с.Сенин (x_1)*	*Оресса-с.Верхутино (x_2)*	*Оресса-. с.Андреевка (x_3)*
1	*2*	*3*	*4*	*5*
y	*1*	*0,82*	*0,65*	*0,95*
x_1	*0,82*	*1*	*0,65*	*0,83*
x_2	*0,65*	*0,65*	*1*	*0,66*
x_3	*0,95*	*0,83*	*0,66*	*1*

Как было показано в разделе 2.1, критическое значение коэффициентов парной корреляции, при уровне значимости P=0,95 и ν=26, будет r_{kp} =0,374, таким образом, все значения коэффициентов в таблице 2.7 значимы.

***Необходимо выявить** степень синхронности колебания годового стока рек Птичь-с.Лучицы, Ясельда-с.Сенин, Оресса-с.Верхутино и Оресса-с.Андреевка, т.е. рассчитать коэффициент множественной корреляции. Для этого вычислим определитель (D) и его минор (D_{11}):*

$$D = \begin{vmatrix} 1 & 0,82 & 0,65 & 0,95 \\ 0,82 & 1 & 0,65 & 0,83 \\ 0,65 & 0,65 & 1 & 0,66 \\ 0,95 & 0,83 & 0,66 & 1 \end{vmatrix} = 0,010 ;$$

$$D_{11} = \begin{vmatrix} 1 & 0,65 & 0,83 \\ 0,65 & 1 & 0,66 \\ 0,83 & 0,66 & 1 \end{vmatrix} = 0,165$$

и тогда $R = \sqrt{1 - \frac{0,010}{0,165}} = 0,969.$

Значимость коэффициента (R), проверяем по формулам (2.28) и (2.29):

$$\overline{S}_R = \frac{1 - 0,969^2}{\sqrt{28 - 3 - 1}} = 0,012, \qquad t_R = \frac{0,969}{0,012} = 80,8.$$

Теоретическое значение t - критерия Стьюдента определяется как $t^T_{(23,5\%)} = 1,711$, *что значительно меньше* $t_R = 80,8$. *Вычислим критерий Фишера*

$$F_R = \frac{0,969^2 (28 - 3 - 1)}{(1 - 0,969^2) \cdot (3 - 1)} = 184,6.$$

Табличное значение (F^m), при $\nu_1 = 3$ *и* $\nu_2 = 28-4 = 24$ *- степенях свободы, будет* $F^T_{5\%} = 3,01$, *что меньше* $F_R = 184,6$.

***Таким образом**, взаимосвязь между колебаниями водности рассматриваемых рек R=0,969 значима на 5%-ном уровне:* $F^T_{5\%} < F_R$. *Судя по коэффициенту множественной детерминации* ($R^2 = 0,969^2 = 0,939$), *вариация водности реки Птичь-с.Лучицы на 93,9% связана с ее колебаниями других рек и только 6,1% вариации* ($1-R^2$) *вызваны индивидуальными особенностями самой реки Птичь (створ Лучицы).*

2.4 Корреляция между качественными признаками

В практике применения корреляционного анализа нередко встречаются случаи, когда признаки не поддаются измерению, не распределяются в вариационный ряд, а ряды различаются по качественным признакам (классам) и характер взаимосвязей между признаками оценивается с помощью методов, отличающихся от рассмотренных ранее. Например, для оценки тесноты связи между качественными признаками, используются

иные критерии. Рассмотрим использование в этих целях *коэффициента ассоциации* Дж. Юла и *коэффициента сопряженности*:

а) методика использования коэффициента ассоциации

Если учитываемые признаки группируются в четырехклеточную корреляционную решетку (таблица 2.8), степень сопряженности между ними измеряется с помощью *коэффициента ассоциации, называемого* также *тетрахорическим показателем связи* (коэффициентом корреляции) и *обозначаемого* символом (r)

$$r = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}}, \qquad (2.31)$$

где a, b, c, d - частоты признаков в ячейках таблицы 2.8. Квадратный корень в формуле (2.31) всегда берется со знаком плюс. Коэффициент ассоциации (r) принимается, когда два признака разделяются только на два класса. В таблице 2.8 символы ($\overline{A}$) и ($\overline{B}$) обозначают признаки, противоположные признакам (A) и (B), или указывают на их отсутствие. Тогда (a),(b),(c) - частоты комбинаций (AB), ($A\overline{B}$), ($\overline{A}B$), ($\overline{A}\,\overline{B}$). Коэффициент ассоциации (r) всегда заключен между (-1) и (+1). Когда (r=0), признаки (A) и (B) не зависят друг от друга (тогда a/c=b/d и "A" появляется относительно одинаково часто совместно с "B" и "$\overline{B}$"). Если (r=-1), при появлении (B), признак (A) не появляется, а при непоявлении (B) (событие $\overline{B}$) - появляется. Когда (r=±1), то (A) появляется и не появляется только одновременно с появлением и непоявлением (B). Таким образом, при (r=±1), имеется полная прямая или обратная связь между (A) и (B). Достоверность выборочного коэффициента ассоциации оценивается по его отношению к средней ошибке (σ_r), определяемой по формуле, аналогичной формуле (2.5). *Нулевая гипотеза* (H_o) заключается в предположении, что связь между учитываемыми альтернативными признаками отсутствует. Гипотеза (H_o) отвергается, если ($\frac{|r|}{\sigma_r} \cdot f \cdot t_{\alpha,m}$), при заданных уровне значимости (α) и m=(n-2) - степенях свободы. Для оценки достоверности выборочного коэффициента ассоциации можно воспользоваться также данными таблицы 2.1. Если величина $[|r| \cdot \sqrt{n-1}]$ превосходит указанные в таблице критические значения, для принятого уровня надежности P=(1-α), нулевая гипотеза отвергается;

Таблица 2.8 Таблица сопряженности категорий двух признаков

Признак	B	$\overline{B}$	Сумма
A	a	b	a+b
$\overline{A}$	c	d	c+d
Сумма	a+c	b+d	n

б) методика использования коэффициента сопряженности

Во многих случаях, данные по качественным признакам можно подразделить более, чем на два качественных класса. Мерой тесноты связи между качественными признаками, в этом случае, может служить *коэффициент сопряженности Пирсона*

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \qquad (2.32)$$

где n - число наблюдений, а χ^2 - определяется как

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - m_i')^2}{m_i'}, \qquad (2.33)$$

где m_i и m_i' - частоты комбинаций различных классов (градаций) коррелируемых признаков (фактические (m_i) и (m_i') - при полном отсутствии связи между признаками); k - число категорий в таблице сопряженности коррелируемых признаков. Коэффициент сопряженности (C), как мера тесноты связи между качественными признаками, изменяется от 0 до размеров таблицы сопряженности, т.е. ($0 \le C \le C_{max}$). Если число строк и столбцов в таблице сопряженности одинаково и равно (S), то

$$C_{max} = \sqrt{\frac{S-1}{S}}. \qquad (2.34)$$

В частности, при (S=3), коэффициент сопряженности, для случая идеальной связи, равен 0,817. Для того, чтобы верхнюю границу величины (C) привести к 1, вычисленное по формуле (2.34) значение (C) рекомендуется разделить на произведение значений (C') (таблица 2.9), в зависимости от числа классов (градаций), на которые разбит каждый коррелируемый признак.

Таблица 2.9 Поправка к (C) на число классов (S)

S	C'	S	C'	S	C'
2	0,798	5	0,943	10	0,985
3	0,859	6	0,959	12	0,989
4	0,915	8	0,976	15	0,999

Значимость коэффициентов сопряженности (C) может быть оценена с помощью критерия (χ^2). Нулевая гипотеза (H_o) заключается в предположении, что связь между качественными признаками отсутствует. Гипотеза (H_o) отвергается, если, вычисленное по формуле (2.33), значение (χ^2) окажется больше некоторого ($\chi^2_{\alpha;m}$), определяемого по Приложению (таблица П.3) при заданном уровне значимости (α) и (ν) - степенях свободы. Распределение (χ^2) непрерывно, а распределение частот - дискретно. Поэтому, в результаты вычисления (χ^2) по формуле (2.33), рекомендуется вводить поправки. Суть их состоит в том, что из каждого возведенного в квадрат абсолютного значения разности $|m_i - m_i'|$ вычитается 0,5.

Сходной, но не вполне сравнимой *мерой сопряженности* между качественными признаками, когда каждый из них делится более чем на два качественных класса, *является* также *коэффициент сопряженности* А.А. Чупрова (T), вычисляемый без поправочного коэффициента (C)

$$T = \sqrt{\frac{\chi^2}{n \cdot (k-1)^{0,5} \cdot (m-1)^{0,5}}}, \qquad (2.35)$$

где n - число наблюдений; k - число строк; m - число граф в таблице сопряженности.

Рассмотрим порядок вычисления корреляции между качественными признаками на конкретных примерах:

а) данные о расходах воды в реке Неман - створ Гродно (Неман - г. Гродно) за межень и год, в целом, в период с 1947 по 1981 годы представлены в виде их обеспеченности, причем, разделены на два класса: "выше нормы" ($P_г$<50%, $P_м$<50%), где "г" - индекс года, "м" - индекс межени, и "равно" и "ниже нормы" ($P_г$≥50%, $P_м$≥50%).

Имеется ли синхронная связь между годовыми и меженными расходами реки Неман в створе города Гродно?

Таблица 2.10 Сопряженность годовых и меженных расходов реки Неман - г. Гродно (1947-1981годы)

P	$P_г<50\%$	$P_г\geq50\%$	Сумма
$P_м<50\%$	$a=11$	$b=6$	$a+b=17$
$P_м\geq50\%$	$c=6$	$d=12$	$c+d=18$
Сумма	$a+c=17$	$b+d=18$	

Пользуясь данными таблицы 2.10 и формулой (2.31), находим

$$r=\frac{11\cdot12-6\cdot6}{\sqrt{(11+6)\cdot(6+12)\cdot(11+6)\cdot(6+12)}}=0{,}314,$$

что указывает на наличие слабой прямой связи между годовым и меженным стоком реки Неман в створе Гродно за период с 1947 по 1981 годы. Оценим ***достоверность выборочного коэффициента ассоциации.*** *Подставив числовое значение полученного (r) в формулу* $(|r|\cdot\sqrt{n-1})$, *находим* $(|0{,}314|\cdot\sqrt{35-1})=1{,}831$. *Эта величина меньше критического значения (1,831<1,950), для* $P=0{,}95$ *и* $n=35$ *(таблица 2.1).* ***Следовательно,*** *при принятом уровне достоверности (*$P=0{,}95$*),* ***связь*** *между рассматриваемыми признаками в генеральной совокупности* ***носит лишь случайный характер,*** *т.е. не является надежной;*

б) рассмотрим ***случай, когда качественные признаки можно разделить более чем на два качественных класса.***

Ежегодные обеспеченности стока *реки Припять-с. Туров за межень (*P_c*) и* ***обеспеченности мелиоративных норм за июль*** *в зоне метеостанции г.п. Житковичи (*$P_м$*) за период с 1947 по 1981 годы разделены на* ***три градации:*** *1) "норма" (Н) - рассматриваемые величины лежат в диапазоне обеспеченностей - (*$33\%\leq P\leq66\%$*); 2) "выше нормы" (ВН) для стока - (*$P<33\%$*) и для мелиоративных норм - (*$P>66\%$*); 3) "ниже нормы" (НН) для стока - (*$P>66\%$*) и для мелиоративных норм - (*$P<33\%$*); они сведены в таблицу сопряженности (таблицу 2.11).*

Какова связь между обеспеченностью стока реки Припять - створ Туров в межень и мелиоративной нормой за июль месяц в зоне метеостанции г.п. Житковичи.

Цифры в скобках (таблица 2.11) показывают частоты (m_i'*), если бы между рассматриваемыми характеристиками не было связи. Эти числа представляют собой произведения суммы столбца на сумму строки, деленные на общую сумму. Так, в нашем случае, для первой клетки имеем*

$$m_i'=(10\cdot9)/35=3.$$

Таблица 2.11 Сопряженность обеспеченностей стока реки Припять - ст.Туров за межень и обеспеченности мелиоративных норм в зоне г.п.Житковичи за июль (1947-1981 годы)

Обеспеченность стока в межень, P_c	Обеспеченность мелиоративных норм за июль, $P_м$			Сумма
	НН	Н	ВН	
НН	$m_1=2$	$m_2=5$	$m_3=3$	10
	$(m_1'=3)$	$(m_2'=4)$	$(m_3'=3)$	10
Н	$m_4=3$	$m_5=4$	$m_6=6$	13
	$(m_4'=3)$	$(m_5'=5)$	$(m_6'=4)$	12
ВН	$m_7=4$	$m_8=5$	$m_9=3$	12
	$(m_7'=3)$	$(m_8'=6)$	$(m_9'=4)$	13
Сумма	9	14	12	35
	9	15	11	(35)

Пользуясь данными таблицы 2.11 и рекомендациями по вычислению (χ^2), согласно (2.33), находим

$$\chi^2 = \frac{(|2-3|-0,5)^2}{3} + \frac{(|5-4|-0,5)^2}{4} + \frac{(|3-3|-0,5)^2}{3} +$$
$$+ \frac{(|3-3|-0,5)^2}{3} + \frac{(|4-5|-0,5)^2}{5} + \frac{(|6-4|-0,5)^2}{4} +$$
$$+ \frac{(|4-3|-0,5)^2}{3} + \frac{(|5-6|-0,5)^2}{6} + \frac{(|3-4|-0,5)^2}{4} = 0,99 .$$

По формуле (2.32), при (n=35), имеем

$$C = \sqrt{\frac{0,99^2}{35+0,99^2}} = 0,165 .$$

*По таблице 2.9, находим поправку (C') к (C) на число классов (таблица 2.11): $C'=0,859$. $0,859=0,737$. Деление ($C=0,165$) на ($C'=0,737$) дает исправленное значение коэффициента сопряженности: $C^*=0,165/0,737=0,224$. Найденное значение коэффициента сопряженности ($C^*=0,224$) указывает на определенную, хотя и не очень тесную, связь между обеспеченностями расходов воды в межень реки Припять - ст.Туров и обеспеченностями мелиоративных норм за июль месяц для зоны метеопункта Житковичи. Однако, при 4-х степенях свободы (таблица 2.11) и ($\alpha=10\%$), табличное значение ($\chi^2_{0,10,4}=7,78$) и превышает найденное ($\chi^2=0,99$). Таким образом, **нулевая гипотеза об отсутствии связи между рассматриваемыми признаками** (в генеральной совокупности) **при принятом уровне значимости не отвергается**. Воспользуемся мерой сопря-*

женности между рассматриваемыми качественными признаками (коэффициент сопряженности "Т" Чупрова А.А.), при m=k=3,

$$T = \sqrt{\frac{0{,}99^2}{35 \cdot (3-1)^{0,5} \cdot (3-1)^{0,5}}} = 0{,}12.$$

Значение величины (Т) также указывает на наличие слабой связи между рассматриваемыми признаками. Отметим, что в некоторых случаях, особенно когда непараметрические критерии для определения связей имеют отрицательные значения, их интерпретация затруднительна. Абсолютные значения критериев могут изменяться в пределах от (0) до (±1). Чем ближе абсолютные значения к единице, тем теснее связь между исследуемыми признаками. При этом, абсолютные величины коэффициентов, соответствующие условию высокой связи, в случае корреляции больше двух признаков, как правило, ниже, чем в случае корреляции двух признаков.

Верхний предел коэффициентов сопряженности зависит от размеров таблицы сопряженности. Поэтому, сравнивать можно только коэффициенты, полученные из таблиц с одинаковым числом столбцов и строк. Не принято сравнивать коэффициенты сопряженности с коэффициентами линейной корреляции, вычисленными более точно на основе параметрических критериев. Вместе с тем, ***коэффициенты ассоциации способны измерять взаимосвязи, отличающиеся от линейных с произвольным распределением частот.***

3 РЕГРЕССИОННЫЙ АНАЛИЗ И МЕТОДИКА СОСТАВЛЕНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Регрессионный анализ является логическим продолжением корреляционного анализа, он *позволяет развить* и *углубить представление о корреляционных связях*. Если корреляционный анализ представляет возможность установить лишь форму и тесноту зависимости между случайными переменными, то *регрессионный анализ позволяет математически описывать* выявленную *зависимость*, т.е. *численно оценивать одни параметры через другие*. Составив и решив уравнение регрессии, *можно* произвести выравнивание эмпирических линий регрессии, т.е. *выполнить моделирование исследуемого процесса (явления) путем подбора функции*, график которой представляет собой *теоретическую линию регрессии*. Если подобная функция отражает сущность процесса или явления, то возможно прогнозирование значений признака за пределами сделанных наблюдений.

Подобно корреляции, *регрессия может быть парной* (простой) *и множественной*, по форме связи - *линейной и нелинейной*, по зависимости - *односторонней* (изменяется лишь один признак под влиянием другого) и *двусторонней* (взаимодействуя, изменяются оба признака).

Регрессия выражается несколькими способами: путем *построения эмпирических линий*, путем *составления уравнения* и, затем, - *построения теоретических линий регрессии*, а также *с помощью коэффициента регрессии*. Уравнение наиболее точно выражает зависимость между двумя переменными (X, Y), если корреляция между ними близка к единице.

Уравнения регрессии могут быть *составлены одним из* следующих *способов*:

а) *координат точек, с использованием двух-трех точек, расположенных на эмпирической линии (желательно в ее начале, середине и конце), для тех случаев, когда не требуется большая точности расчетов;*

б) *наименьших квадратов, когда для составления уравнения регрессии привлекаются все сопряженные наблюдения и требуется повышенная точность расчетных величин..*

Рассмотрим наиболее простые способы составления уравнений регрессии.

3.1 Уравнение линейной регрессии с одним переменным фактором

Под линейной (прямолинейной) *корреляционной зависимостью* между двумя признаками (X_i) и (Y_i) *понимается* такая *зависимость*, которая носит линейный характер и *выражается уравнением прямой линии*. Это уравнение называется *уравнением регрессии* (Y_i) на (X_i), а соответствующая ему *прямая линия* - выборочной линией регрессии (Y_i) на (X_i).

В литературе (Y_i) называется - *функцией отклика, зависимой переменной, предиктором, а (X_i) - входной переменной, независимой переменой, фактором, регрессором.*

Рассмотрим метод наименьших квадратов для получения уравнения линейной регрессии.

Предположим, что линия регрессии переменой, которую обозначим (Y_i) от переменной (X_i), имеет вид ($\beta_0+\beta_1 \cdot X_i$). Тогда можно записать линейную модель

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i. \qquad (3.1)$$

Для данного (X_i) соответствующее значение (Y_i) состоит из суммы величин ($\beta_0+\beta_1 \cdot X_i$) и добавки ($\varepsilon_i$), при учете которой любое индивидуальное значение (Y_i) получает возможность не попасть на линию регрессии. Задача линейного регрессионного анализа (метода наименьших квадратов) состоит в том, чтобы, зная положение точек на плоскости, провести линию регрессии при минимальной сумме квадратов отклонений (ε_i^2) по оси (0y) (ординате). Предположим, что форма модели установлена достоверно. В уравнении (3.1) величины (β_0), (β_1) и (ε_i) неизвестны, причем, последнюю, на самом деле, будет трудно исследовать, поскольку она меняется в ходе наблюдений. Величины (β_0) и (β_1) остаются постоянными, но без точного изучения всех возможных сочетаний (Y_i) и (X_i), ограничимся лишь их оценками (b_0) и (b_1.). В данной интерпретации запишем

$$\hat{Y}_i = b_0 + b_1 \cdot X_i \, . \qquad (3.2)$$

где $\hat{Y}_i$ - обозначает предсказанное значение (Y_i) для данного "X_i" (b_0 и b_1 - определены). Если имеется множество пар (n) - наблюдений (Y_1 и X_1), (Y_2 и X_2),...,(Y_n и X_n) и сумма квадратов отклонений от

"истинной" линии определяется как

$$U = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i)^2_{min} , \qquad (3.3)$$

то при решении поставленной задачи необходимо, в каждом конкретном случае, вычислить значения коэффициентов (b_0) и (b_1), минимизирующих эту сумму. Для этого, как известно из теории математического анализа, необходимо вычислить частную производную функции (U) по (β_0), затем, - по (β_1), приравняв, в итоге, результаты к нулю. Следовательно, имеем:

$$\begin{cases} \dfrac{\partial U}{\partial \beta_0} = -2 \cdot \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i) \\ \dfrac{\partial U}{\partial \beta_1} = -2 \cdot \sum_{i=1}^{n} X_i \cdot (Y_i - \beta_0 - \beta_1 \cdot X_i); \end{cases} \qquad (3.4)$$

$$\begin{cases} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i) = 0 \\ \sum_{i=1}^{n} X_i \cdot (Y_i - \beta_0 - \beta_1 \cdot X_i) = 0. \end{cases} \qquad (3.5)$$

Подставив (b_0), (b_1), соответственно, вместо (β_0), (β_1), из (3.5) получим:

$$\begin{cases} \sum_{i=1}^{n} Y_i - n \cdot b_0 - b_1 \cdot \sum_{i=1}^{n} X_i = 0 \\ \sum_{i=1}^{n} X_i \cdot Y_i - b_0 \cdot \sum_{i=1}^{n} X_i - b_1 \cdot \sum_{i=1}^{n} X_i^2 = 0; \end{cases} \qquad (3.6)$$

или

$$\begin{cases} b_0 \cdot n + b_1 \cdot \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ b_0 \cdot \sum_{i=1}^{n} X_i + b_1 \cdot \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i \cdot Y_i. \end{cases} \qquad (3.7)$$

Эти уравнения называются нормальными. Решая систему уравнений (3.7) относительно угла наклона прямой (b_1), получим

$$b_1 = \frac{\sum_{i=1}^{n} X_i \cdot Y_i - ((\sum_{i=1}^{n} X_i) \cdot (\sum_{i=1}^{n} Y_i)) / n}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2 / n} = \frac{\sum_{i=1}^{n} (X_i - X) \cdot (Y_i - Y)}{\sum_{i=1}^{n} (X_i - X)^2}, \qquad (3.8)$$

где b_1 - правильные, но несколько различные две формы одной и той же зависимости. Решение системы (3.7), относительно свободного члена (b_0),

дает зависимость

$$b_0 = \overline{Y} - b_1 \cdot \overline{X} \ . \qquad (3.9)$$

Подставляя уравнение (3.9) в уравнение (3.2), получаем

$$\hat{Y}_i = \overline{Y} + b_1 \cdot (X_i - \overline{X}) \ . \qquad (3.10)$$

Коэффициент b_0 - свободный член уравнения регрессии, как геометрическая характеристика, представляет собой отрезок, отсекаемый на ординате линией регрессии. Коэффициент (b_1) представляет собой тангенс угла наклона линии регрессии к оси абсцисс. Существует *две модели* регрессии. Модель ($\hat{Y}_i = b_{0(yx)} + b_{1(yx)} \cdot X_i$) условно *можно назвать прямой*, а модель ($\hat{X}_i = b_{0(xy)} + b_{1(xy)} \cdot Y_i$) - *обратной регрессией*. Следовательно, уравнение ($\hat{Y}_i = b_0 + b_1 \cdot X_i$) не является алгебраическим, из которого непосредственно можно найти (X_i), так как эта модель получена минимизацией суммы квадратов отклонений вдоль оси (0y). Для статистического оценивания коэффициентов регрессии *проверяется нуль-гипотеза* (H_0:β=0), т.е. условие, - отличается ли статистически значимо оценка коэффициента регрессии от нуля? Граница значимости устанавливается на основании распределения Стьюдента:

$$t = |b| / \overline{S}_b \geq t^T_{(n-2,P)} \ ; \qquad (3.11)$$

$$\overline{S}_{b_1} = \frac{\overline{S}_{ост.}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \ ; \qquad (3.12)$$

$$\overline{S}_{b_0} = \frac{\overline{S}_{ост.}}{\sqrt{n}} \ , \qquad (3.13)$$

где $\overline{S}_{ост.} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{Y}_i)^2}{n-2}}$ - остаточная дисперсия. При использовании для построения линии регрессии значений параметров (b_0) и (b_1), включающих в себя указанные случайные ошибки, априори допускается погрешность в оценке ординат линии регрессии. *Погрешность* предсказываемого среднего значения "Y" (или "$\hat{Y}_0$" при заданном "X_0") *определяется значени-*

ем соответствующей *дисперсии*

$$\bar{S}^2_{y(x)} = \bar{S}^2_{ост.} \cdot \left(\frac{1}{n-2} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right). \quad (3.14)$$

Дисперсия ($\bar{S}^2_{y(x)}$) характеризует рассеивание ординат выборочной линии регрессии относительно линии регрессии, соответствующей генеральной совокупности. *Для проверки значимости уравнения регрессии,* в целом, с использованием F-критерия Фишера, *общая дисперсия* ($\bar{S}^2_Y$) *сравнивается с остаточной дисперсией* ($\bar{S}^2_{ост}$), которая представляет собой показатель ошибки предсказания на базе уравнения регрессии результатов опытов; во сколько раз уравнение регрессии предсказывает результаты опытов лучше, чем среднее, видно из F - критерия

$$F = \frac{\bar{S}^2_Y}{\bar{S}^2_{ост.}}. \quad (3.15)$$

Искомое уравнение, при (α) %-ном уровне значимости, описывает результаты опытов лучше среднего ($\bar{Y}$) в ($F^T_{(\nu_1,\nu_2,P\%)}$) - раз (при условии $F > F^T_{(\nu_1,\nu_2,P\%)}$).

Рассмотрим примеры, иллюстрирующие порядок применения изложенных рекомендаций:

а) по материалам 28-летних наблюдений (с 1947 года) за годовым стоком рек Ореса - с. Андреевка (X) и Птичь - с. Лучицы (Y) необходимо построить уравнение регрессии (таблица 2.3). к

Решение

1) По исходным данным определяются характеристики: $\bar{X} = 16{,}86$, *м³/с;*

$\bar{Y} = 45{,}31$, *м³/с;* $\sum_{i=1}^{28}(X_i - \bar{X})\cdot(Y_i - \bar{Y}) = 2187{,}84$; $\sum_{i=1}^{28}(X_i - \bar{X})^2 = 917{,}15$;

$\sum_{i=1}^{28}(Y_i - \bar{Y})^2 = 5737{,}61$; $\bar{S}^2_Y = 220{,}68$; $\bar{S}^2_X = 35{,}28$.

2) На основании полученных данных, по формулам (3.8) и (3.9), определяются $b_i = 2187{,}84/917{,}5 = 2{,}38$; $b_0 = 45{,}31 - 2{,}38 \cdot 16{,}86 = 5{,}18$, *при* $r = 0{,}95 \pm 0{,}006$ *(0,89...1,00).*

3) Следовательно, уравнение регрессии (Y) по (X) будет иметь вид - $Y_i = 45{,}31 + 2{,}38 \cdot (X_i - 16{,}86)$ *или* $Y_i = 5{,}18 + 2{,}38 \cdot X_i$, *м³/с, а уравнение регрессии (X) по (Y) вид -* $X_i = 0{,}38 \cdot Y_i - 0{,}42$, *м³/с.*

*4) **Остаточная дисперсия определяется как***

$\bar{S}^2_{ост.}$ *=497,69/(28-2)=19,14, а стандартные ошибки коэффициентов регрессии определяются по (3.12) и (3.13):*

$$\bar{S}_{b_0} = \frac{4,37}{\sqrt{28}} = 0,84; \qquad \bar{S}_{b_1} = \frac{4,37}{\sqrt{917,15}} = 0,14.$$

*5) **Результаты расчетов параметров уравнения регрессии следующие:***

$$b_0 \pm \bar{S}_{b_0} = 5,18 \pm 0,84 ; \qquad b_1 \pm \bar{S}_{b_1} = 2,38 \pm 0,14.$$

*6) **Стандартная ошибка ординаты уравнения регрессии определяется по (3.14)***

$$\bar{S}^2_{\bar{Y}(X)} = 4,37 \cdot \left[\frac{1}{28-2} + \frac{(X_i - 16,86)^2}{917,15} \right].$$

В частности, $\bar{S}^2_{\bar{Y}(x)}$=1,31, м^3/с (при X_i=10); $\bar{S}^2_{\bar{Y}(x)}$=0,96, м^3/с (при X_i=20); $\bar{S}^2_{\bar{Y}(x)}$=0.84=$\bar{S}_{b_0}$, (при $X_i = \bar{X}$ =16,86, м^3/с) - как это и должно быть.

*7) **Значимость коэффициентов регрессии проверяется по t-критерию Стьюдента (3.11):***

$$t_{b_1} = |2,38| / 0,14 = 17,0 ; \qquad t_{b_0} = |5,18| / 0,84 = 6,17 .$$

Табличное значение критерия (t^T) находится по Приложению (таблица П.2). При 5%-ном уровне значимости и n=28, оно составляет - $t^T_{(28,5\%)}$=1,701. Следовательно, найденные значения коэффициентов регрессии статистически значимы.

*8) Адекватность уравнения исходным данным проверяется по критерию Фишера (3.15) - F=220,68/19,14=11,53. Табличное значение (F^T) из Приложения (таблица П.4.1) - $F_{(27,26,5\%)}$=1,92 - значительно меньше фактического, следовательно, **полученное выше уравнение регрессии (Y) и (X) адекватно отражает исследуемую связь** годовых расходов воды рек Ореса - с. Андреевка (X) и Птичь - с. Лучицы (Y);*

*б) **провести корреляционный и регрессионный анализ данных по определению относительной влажности (X) и липкости (Y) чернозема обыкновенного,** которые представлены в таблице 3.1.*

Решение

*1) **Вычисляются вспомогательные величины, с использованием сумм, которые даются под расчетной таблицей 3.1 (при n=12):***

$$\bar{y} = (\Sigma Y) : n = 35,9 : 12 = 2,99 \ (г / см^2) ;$$

$$\Sigma(X - \bar{x})^2 = \Sigma X^2 - (\Sigma X)^2 : n = 25742,67 - (514,7)^2 : 12 = 3666,33;$$

$$\Sigma(Y - \bar{y})^2 = \Sigma Y^2 - (\Sigma Y)^2 : n = 171,37 - (35,9)^2 : 12 = 63,97;$$

$$\Sigma(X - \bar{x})(Y - \bar{y}) = \Sigma XY - (\Sigma X \Sigma Y) : n = = 2013,08 - (514,7 \cdot 35,9) : 12 = 473,27.$$

$$\bar{X} = \Sigma X : n = 514,7 : 12 = 42,89 .$$

*2) **Определяются коэффициенты корреляции, регрессии и составляется уравнение регрессии:***

$$r = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\sqrt{(X-\bar{x})^2\Sigma(Y-\bar{y})^2}} = \frac{473,27}{\sqrt{3666,33 \cdot 63,97}} = 0,977;$$

$$b_{yx} = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\Sigma(X-\bar{x})^2} = \frac{473,27}{3666,33} = 0,13 \quad (г/см^2);$$

$$Y - \bar{y} + b_{yx}(X-\bar{x}) = 2,99 + 0,13 \cdot (X - 42,89) = 0,13X - 2,58.$$

*3) **Вычисляются ошибки, критерий значимости и доверительные интервалы:***

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0,977^2}{12-2}} = 0,067;$$

$$s_b = s_r\sqrt{\frac{\Sigma(Y-\bar{y})^2}{\Sigma(X-\bar{x})^2}} = 0,067\sqrt{\frac{63,97}{3666,33}} = 0,009\ (г/см^2);$$

$$s_{yx} = s_r\sqrt{\Sigma(Y-\bar{y})^2} = 0,067\sqrt{63,97} = 0,54\ (г/см^2);$$

$$t_r = \frac{r}{s_r} = \frac{0,977}{0,067} = 14,58;$$

$$\nu = n-2 = 12-2 = 10;\ t_{0,5} = 2,23;$$

$$r \pm t_{05}s_r = 0,977 \pm 2,23 \cdot 0,067 = 0,977 \pm 0,149 (0,82...1,00);$$

$$b_{yx} \pm t_{05}s_r = 0,13 \pm 2,23 \cdot 0,009 = 0,13 \pm 0,02 (0,11...0,15)\ (г/см^2).$$

Таблица 3.1 Результаты расчета вспомогательных величин при вычислении корреляции и регрессии (Y) по (X)

Номер пары	Значение признаков		X^2	Y^2	XY
	X, %	Y, г/см²			
1	19,9	0,0	396,01	0,00	0,00
2	20,9	0,6	436,81	0,36	12,54
3	26,1	1,1	681,21	1,21	28,71
4	29,4	1,2	864,36	1,44	35,28
5	30,5	1,7	930,25	2,89	51,58
6	40,3	1,7	1624,09	2,89	68,51
7	44,8	2,6	2007,04	6,76	116,48
8	47,8	3,4	2284,84	11,56	162,52
9	55,6	4,2	3091,36	17,64	233,52

Продолжение таблицы 3.1

10	*58,3*	*5,8*	*3398,89*	*33,64*	*338,14*
11	*64,5*	*6,3*	*4160,25*	*39,69*	*406,35*
12	*76,6*	*7,3*	*5867,56*	*53,29*	*559,18*
Сумма	*51[illegible]7=ΣX*	*35,9=ΣY*	*25742,67=ΣX²*	*171,37=ΣY²*	*2013,08=ΣXY*

Судя по t-критерию ($t_ф > t_{05}$) и доверительным интервалам, которые не включают нулевого значения, корреляция и регрессия значимы и, следовательно, нулевая гипотеза на 5%-ном уровне отвергается.

*4) **По уравнению регрессии рассчитываются усредненные теоретические значения (Y), для экстремальных величин (X) строится теоретическая линия регрессии (Y) по (X) (рисунок 3.1):***

$Y_{x=19,9}=0,13 \cdot 19,9-2,58=0,00$ *(г/см²);*

$Y_{x=76,6}=0,13 \cdot 76,6-2,58=7,37$, *(г/см²).*

*Найденные точки (19,9; 0,00) и (76,6; 7,37) наносятся на график, соединяются прямой, характеризующей теоретическую линию регрессии (Y) по (X). Зависимость показывает, что увеличению влажности почвы на 1% соответствует увеличение липкости в среднем на 0,13 (г/см²). Судя по коэффициенту детерминации ($d_{yx}=0,977^2=0,95$), примерно 95% изменений в липкости обусловлено изменениями во влажности почвы и только 5% изменений связано с другими факторами. На графике целесообразно указать уравнение регрессии, коэффициент регрессии, коэффициент регрессии и корреляции, доверительную зону для истинной линии регрессии, в совокупности. **Чтобы ограничить доверительную зону**, необходимо вверх и вниз от теоретической линии регрессии отложить величину одной (68%-ная зона) или двух (95%-ная зона) ошибок отклонения от регрессии ($\pm s_{yx}$ или $\pm 2s_{yx}$) и соединить найденные точки пунктирными линиями. Область, заключенная между этими линиями, является доверительной зоной регрессии. На рисунке 3.1 пунктирными линиями ограничена 68%-ная доверительная зона для положения "истинной" линии регрессии, в совокупности, т. е. зона в пределах ($Y \pm s_{yx}$). Если необходимо ограничить 95%-ную доверительную зону (когда можно ожидать, что только 5% всех случаев окажутся за пределами - $Y \pm 2s_{yx}$), то значения ошибки удваиваются ($t_{05}=2$). Отметим, что общая сумма квадратов ($\Sigma(Y-y)^2$) может быть разложена на два компонента: сумму квадратов для регрессии (C_b) и сумму квадратов отклонения от регрессии (Cd_{yx}).*

Первая сумма определяется по формуле

$$C_b = \frac{[\Sigma(X-\bar{x})(Y-\bar{y})]^2}{\Sigma(X-\bar{x})^2} = \frac{473{,}27^2}{3666{,}33} = 61{,}09,$$

вторая сумма квадратов находится по разности -

$$Cd_{yx} = \Sigma(Y-\bar{y})^2 - C_b = 63{,}97 - 61{,}09 = 2{,}88.$$

Разделив полученные суммы квадратов на соответствующие степени свободы, можно найти средние квадраты, вычислить критерий (F), который и позволяет проверить нулевую гипотезу об отсутствии линейной связи (Y) с (X). Результаты расчетов представляются в виде таблицы дисперсионного анализа (таблица 3.2).

Таблица 3.2 Результаты дисперсионного анализа (Y)

Дисперсия	Сумма квадратов	Степени свободы	Средний квадрат	$F_ф$	F_{05}
Общая	63,97	11	–	–	–
Регрессия	61,09	1	61,09	212,12	4,96
Отклонения от регрессии	2,88	10	0,288	–	–

Соотношение ($F_ф > F_{05}$) указывает на то, что отклонение от линейности обусловлено случайным выборочным варьированием, и нулевая гипотеза об отсутствии линейной связи (Y) с (X) отвергается. По среднему квадрату отклонения от регрессии (S^2_{yx}=0,288) легко вычислить ошибку отклонения от регрессии (S_{yx}). Она равна - $s_{yx} = \sqrt{s^2_{yx}} = \sqrt{0{,}288} = 0{,}54 г/см^2$, т. е. ранее полученной величине;

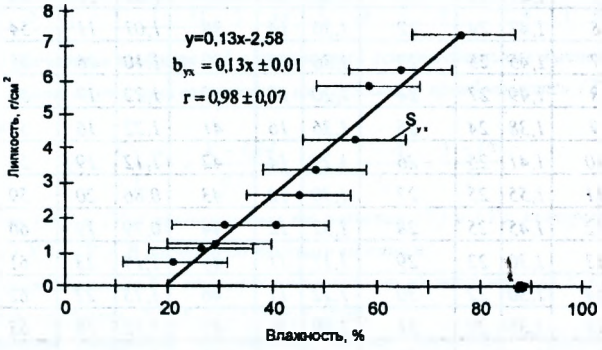


Рисунок 3.1 Точечный график и теоретическая линия регрессии при прямолинейной корреляции между липкостью и относительной влажностью чернозема обыкновенного.

в) провести корреляционный и регрессионный анализ для выборочной совокупности, в которой представлены результаты определения содержания гумуса и подвижных форм фосфатов в пахотном слое легкосуглинистой дерново-подзолистой почвы (таблица 3.3).

Решение

*1) **Данные группируются в корреляционную таблицу** (решетку), состоящую из столбцов (X) и строк (Y), количество которых соответствует числу групп ряда. При (n=64), целесообразно выделить 6...8 групп. Для рядов (X) и (Y) определяются величина интервала группировки и число групп:*

$$i_x = \frac{X_{макс} - X_{мин}}{6...8} = \frac{1{,}70 - 0{,}79}{6...8} = \frac{0{,}91}{7} = 0{,}13\%;$$

$$i_y = \frac{Y_{max} - Y_{min}}{6...8} = \frac{36 - 6}{6...8} = \frac{30}{6} = 5 \ (мг/100\ г\ почвы).$$

Таблица 3.3 Содержание гумуса (ряд X, %) и подвижного фосфора (ряд Y, мг на 100 г почвы)

Номер пары	X	Y	Номер пары	X	Y	Номер пары	X	Y	Номер пары	X	Y
1	1,57	30	17	1,35	17	33	0,96	16	49	1,42	27
2	1,58	28	18	1,31	17	34	1,08	9	50	1,36	25
3	1,21	25	19	1,29	16	35	1,16	19	51	1,55	24
4	1,21	27	20	1,38	17	36	1,12	17	52	1,36	22
5	1,41	25	21	1,38	16	37	1,01	11	53	1,46	28
6	1,47	24	22	1,36	14	38	1,07	11	54	1,39	28
7	1,45	25	23	1,36	16	39	1,10	16	55	1,63	36
8	1,49	27	24	1,20	17	40	1,22	17	56	1,57	36
9	1,38	24	25	1,36	16	41	1,22	16	57	1,37	27
10	1,41	25	26	1,29	14	42	1,12	19	58	1,48	25
11	1,55	25	27	1,30	12	43	0,86	20	59	1,61	28
12	1,45	25	28	1,32	12	44	0,79	19	60	1,61	30
13	1,30	22	29	1,17	11	45	1,19	23	61	1,70	28
14	1,30	22	30	1,22	11	46	1,15	22	62	1,62	28
15	1,39	20	31	1,09	9	47	1,13	18	63	1,04	13
16	1,46	22	32	1,13	9	48	1,34	20	64	1,22	10

***В корреляционную таблицу 3.4 последовательно переносятся исходные даты из таблицы 3.3.** Например, первая пара, имеющая X=1,57% и Y=30 (мг/100 г почвы), заносится черточкой в клетку, находящуюся на пересечении последнего столбца (против группы*

1,57–1,70, %) и второй строки (против группы 30–25, мг/100 г почвы). Так, все данные первичной таблицы переносятся в корреляционную решетку (для проверки это проделывается дважды), подсчитывается число дат в каждой ячейке и результат записывается в ней же. Эти числа представляют частоты количества вариант, имеющих одинаковые значения признаков (X) и (Y). Затем, подсчитывается частота каждой строки, столбца, общие суммы всех частот по столбцам (f_x) и по строкам (f_y), а также общее число объектов: $n=\Sigma f_x=\Sigma f_y=64$.

Таблица 3.4 Корреляционная матрица

X, % / Y, мг/100 г почвы			0,79-0,91	0,92-1,04	1,05-1,17	1,18-1,30	1,31-1,43	1,44-1,56	1,57-1,70	Суммы, fy
			середина групп							
			0,85	0,98	1,11	1,24	1,37	1,50	1,64	
36-31		33								2
30-25		28							II 2	12
25-21	сере-	23				I 1	III 3	II 2	IIIII 6	17
20-16	дина	18			I 1	IIII 4	IIIII 5	IIIIII 7		20
15-11	групп	13	II 2	I 1	IIIII 5	IIII 4	IIIIIII 8			9
10-6		8		II 2	II 2	III 3	II 2			4
					III 3	I 1				
Суммы, f_x			2	3	11	13	18	9	8	64=n

*2) **Составляется расчетная таблица 3.5 и проводятся вспомогательные вычисления.** В таблице, вместо границ групп, проставляются их середины и преобразуются (X) и (Y) по соотношениям:*

$$X_1 = \frac{X - A_x}{i_x} = \frac{X - 1{,}24}{0{,}13}; \quad Y_1 = \frac{Y - A_x}{i_y} = \frac{Y - 18}{5}.$$

За условные начала (A_x) и (A_y) принимаются те значения (X) и (Y), которые ближе всего к величинам ($\overline{X}$) и ($\overline{Y}$). В расчетную таблицу вносятся:

а) произведения отклонений в единицах интервала на их частоты - (f_xX_1) и (f_yY_1) и, соответственно, их суммы - $\Sigma(f_xX_1)=37$; $\Sigma(f_yY_1)=30$;

б) произведения квадратов отклонений на их частоты - $\Sigma(f_xX_1^2)=167$; $\Sigma(f_yY_1)=108$;

в) суммы произведений отклонений в интервалах на их частоту (fX_1Y_1) и общая сумма - $\Sigma(fX_1Y_1)=96$. Для этого частота (f), указанная в каждой клетке таблицы, умножается на соответствующие значения (X_1) и (Y_1) и суммируются по каждой колон-

ке полученные цифры. Так, для первой колонки - $f(X_1 Y_1)= 2\cdot(-3)\cdot 0 = 0$; для второй колонки - $fX_1Y_1=1\cdot(-2)\cdot 0+2\cdot(-2)\cdot(-1)=4$; и т. д. Затем, находится общая сумма произведений - $fX_1Y_1=(0+4+...+54)=96$;

г) групповые или частные средние ($\bar{y}_x$) для каждого значения (X), определенные по формуле

$$\bar{y}_x = A_y + i_y\left(\frac{\Sigma f Y_1}{f}\right),$$

где $A_y=18$; $i_y=5$;

д) значения ($\bar{x}$), ($\bar{y}$), $\Sigma(X-\bar{x})$, $\Sigma(Y-\bar{y})$ и $\Sigma(X-\bar{x})(Y-\bar{y})$ в исходных единицах, ицах, которые записываются под таблицей 3.5. Следует иметь в виду, что если в процессе кодирования производилось деление или умножение на (i_x) и (i_y), то суммы квадратов, в одном случае, надо умножить, а в другом - разделить на (i_x^2) или (i_y^2); сумму произведений отклонений, в первом случае, надо умножить, а во втором - разделить на ($i_x i_y$).

Таблица 3.5 Результаты расчета вспомогательных величин при вычислении корреляции и регрессии (Y) по (X)

$Y_1=\frac{Y-18}{5}$	$X_1=\frac{X-1,24}{0,13}$	–3	–2	–1	0	1	2	3	f_y	f_yY_1	$f_yY_1^2$
	X / Y	0,85	0,98	1,11	A_x= =1,24	1,37	1,50	1,63			
3	33							2	2	6	18
2	28				1	3	2	6	12	24	48
1	23			1	4	5	7		17	17	17
0	A_y=18	2	1	5	4	8			20	0	0
–1	13		2	2	3	2			9	–9	9
–2	8			3	1				4	–8	16
f_x		2	3	11	13	18	9	8	64=n	30= $=\Sigma f_yY_1$	108= $\Sigma f_yY_1^2$
f_xX_1		–6	–6	–11	0	18	18	24	$37=\Sigma f_xX_1$		
$f_xX_1^2$		18	12	11	0	18	36	72	$167=\Sigma f_xX_1^2$		
$f_xX_1Y_1$		0	4	7	0	9	22	54	$96=\Sigma fX_1Y_1$		
y_x		18,0	14,7	15,7	18,0	19,7	24,1	29,2			

$n=64$; $\bar{x} = A_x + i_x\Sigma\left(f_xX_1\right)/n = 1,24 + 0,13\cdot 37/64 = 1,32\%$;

$\bar{y} = A_y + i_y\left(\Sigma f_yY_1\right)/n = 18 + 5\cdot 30/64 = 20,3$ мг / 100 г;

$\Sigma(X-\bar{x})^2 = i_x^2(\Sigma f_xX_1^2 - \Sigma\left(f_xX_1\right)^2/n) = 0,13^2(167 - 37^2/64) = 2,46$;

$\Sigma(Y-\bar{y})^2 = i^2{}_y(\Sigma(f_yY_1^2) - (\Sigma f_yY_1)^2/n) = 5^2(108 - 30^2/64) = 2348,5$;

$\Sigma(X-\bar{x})\cdot(Y-\bar{y}) = i_{xy}(\Sigma\left(fX_1Y_1\right) - \Sigma(f_xX_1)\cdot\Sigma(f_yY_1)/n) = 0,13(96 - 37\cdot 30/64) = 51,13$.

3) Вычисляется выборочный коэффициент корреляции, регрессии и составляется уравнение регрессии (Y) по (X):

$$r = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\sqrt{\Sigma(X-\bar{x})^2\Sigma(Y-\bar{y})^2}} = \frac{51{,}13}{\sqrt{2{,}46\cdot 2348{,}5}} = 0{,}67;$$

$$b_{yx} = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{\Sigma(X-\bar{x})^2} = \frac{51{,}13}{2{,}46} = 20{,}8 \text{ мг/100 г почвы};$$

$$Y = \bar{y} + b_{yx}(X-\bar{x}) = 20{,}3 + 20{,}8(X-1{,}32) = 20{,}8X - 7{,}2.$$

4) Определяются ошибки, критерий значимости, доверительные интервалы для (r) и (b_{yx}) и проверяется (H_0):

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0{,}67^2}{64-2}} = 0{,}094;$$

$$s_b = s_r\sqrt{\frac{\Sigma(Y-\bar{y})^2}{\Sigma(X-\bar{x})^2}} = 0{,}094\sqrt{\frac{2348{,}5}{2{,}46}} = 2{,}9 \text{ мг/100 г почвы};$$

$$s_{yx} = s_r\sqrt{\Sigma(Y-\bar{y})^2} = 0{,}094\sqrt{2348{,}5} = 4{,}61 \text{мг/100 г почвы};$$

$$t_r = \frac{r}{s_r} = \frac{0{,}67}{0{,}094} = 7{,}13;$$

$$\nu = n = 64-2 = 62; \quad t_{05} = 2{,}00;$$

$$r \pm t_{05}s_r = 0{,}67 \pm 2{,}00\cdot 0{,}094 = 0{,}67 \pm 0{,}19 \ (0{,}49 \ldots 0{,}86);$$

$$b_{yx} \pm t_{05}s_r = 20{,}8 \pm 2{,}00\cdot 2{,}9 = 20{,}8 \pm 5{,}8 \ (15{,}0 \ldots 26{,}6);$$

(H_0) отвергается, ибо ($t_\phi > t_{05}$).

5) По принятому уравнению регрессии, рассчитываются средние теоретические кие *значения ($\bar{y}_X$) для экстремальных групповых значений (X) и строится теоретическая линия регрессии (Y) по (X) (рисунок 3.2):*

$$\bar{y}_{X=0{,}85} = 20{,}8\cdot 0{,}85 - 7{,}2 = 10{,}5 \text{ мг/100 г почвы};$$

$$\bar{y}_{X=1{,}64} = 20{,}8\cdot 1{,}64 - 7{,}2 = 26{,}9 \text{ мг/100 г почвы}.$$

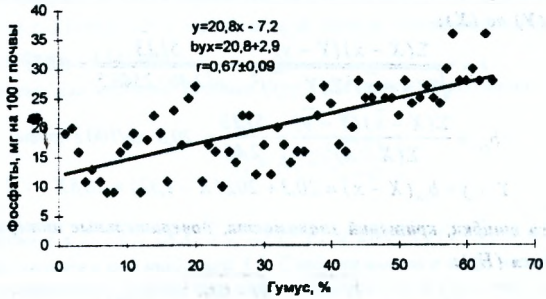


Рисунок 3.2 Точечный график (◆ исходный ряд) и теоретическая линия регрессии (- линейная зависимость) при прямолинейной корреляции между содержанием гумуса (%) и подвижного фосфора (мг на 100 г легкосуглинистой дерново-подзолистой почвы).

Для проверки гипотезы о линейности связи (Y) с (X), вычисляются суммы квадратов для регрессии (C_b) и отклонения от регрессии (Cd_{yx}):

$$C_b = \frac{(\Sigma(X-\bar{x})(Y-\bar{y}))^2}{\Sigma(X-\bar{x})^2} = \frac{51{,}13^2}{2{,}46} = 1062{,}7;$$

$$Cd_{yx} = \Sigma(Y-\bar{y})^2 - C_b = 2348{,}5 - 1062{,}7 = 1285{,}8.$$

Результаты дисперсионного анализа приведены в таблице 3.6.

Таблица 3.6 Результаты дисперсионного анализа (Y)

Дисперсия	Сумма квадратов	Степени свободы	Средний квадрат	F_ϕ	F_{05}
Общая	2348,5	63	–	–	–
Регрессия	1062,7	1	1062,70	51,3	4,00
Отклонения от регрессии	1285,8	62	20,73	–	–

Нулевая гипотеза об отсутствии линейной связи (Y) с (X) отвергается ($F_\phi > F_{05}$) и, следовательно, отклонения от линейности обусловлены случайным выборочным варьированием данных.

***На основании полученных данных,** можно считать, что между содержанием гумуса и подвижными фосфатами легкосуглинистой дерново-подзолистой почвы имеет место средняя взаимосвязь, когда и (r_a) всей совокупности лежит в пределах от 0,49 до 0,85. Нулевая гипотеза (H_0: r=0) на 5%-ном уровне значимости отвергается ($t_r = t_b < t_{05}$).*

Судя по коэффициенту детерминации ($d_{yx} = r^2 = 0{,}67^2 = 0{,}45$), примерно 45% изменений в содержании фосфора обусловлено изменениями в содержании почвенного гумуса. На 1 % изменения содержания гумуса приходится 20,8 мг/100 г почвы изменения содержания подвижных фосфатов. По уравнению вида (Y=20,8X–7,2) для любых значений (X), включая те, которых нет в исходных данных, можно рассчитать значение (Y). Однако, нельзя использовать полученное уравнение регрессии для интерполяции данных за пределы таблицы.

3.2 Нелинейная парная регрессия

В случаях, *когда* по корреляционным критериям *гипотеза линейности может быть отброшена* или при графическом представлении точек явно прослеживается нелинейность парной зависимости, *есть смысл получить по экспериментальным данным ее нелинейную форму*. При этом, можно ожидать, что нелинейная форма даст меньшую остаточную дисперсию ($\bar{S}^2_{ост.}$), т.е. лучше предскажет результаты опытов. Но следует помнить, что речь идет о зависимости, нелинейной по фактору (X); по параметрам зависимость остается линейной.

Нелинейная парная регрессия сводится к получению заданной нелинейной зависимости (Y(X)), приближающей совокупность чисел (X_i) и (Y_i) с наименьшей среднеквадратической погрешностью. Сведение нелинейной регрессии к линейной выполняется с помощью *линеаризирующих преобразований**) (Y_i') и (X_i'). *Используя метод наименьших квадратов и линеаризирующие преобразования, можно построить* практически *любые формы нелинейной парной связи.*

В таблице 3.6 приведены наиболее часто встречающиеся парные зависимости и линеаризирующие преобразования переменных. Качество предсказания результатов проверяется с помощью уравнения ($\hat{Y}_i' = b_0' + b_1' \cdot X_i'$). После вычисления коэффициентов (b_0') и (b_1') по методу наименьших квадратов (как для парной линейной зависимости),

*) **линеаризирующие преобразования** - методы, позволяющие свести решения нелинейных задач к последовательному решению поставленных линейных задач.

выполняются обратные преобразования [по (b_0') и (b_1') определяются (b_0) и (b_1)], в соответствии с содержанием таблицы 3.6. Очевидно, что парная зависимость может иметь разнообразную форму.

Таблица 3.6 Функции и линеаризующие преобразования

N п/п	Функция	Линеаризующие преобразования			
		Преобразование переменных		Выражения для величин (b_0) и (b_1)	
		Y'	X'	b_0	b_1
1	2	3	4	5	6
1	$Y=b_0+b_1 \cdot X$	Y	X	b_0'	b_1'
2	$Y=1/(b_0+b_1 \cdot X)$	1/Y	X	b_0'	b_1'
3	$Y=b_0+b_1/X$	Y	1/X	b_0'	b_1'
4	$Y=X/(b_0+b_1 \cdot X)$	X/Y	X	b_0'	b_1'
5	$Y=b_0 \cdot b_1^X$	lgY	X	$10^{b_0'}$	$10^{b_1'}$
6	$Y=b_0+\exp(b_1/X)$	lnY	1/X	$\exp(b_0')$	b_1'
7	$Y=b_0 \cdot 10^{b_1 X}$	lgY	X	$10^{b_0'}$	b_1'
8	$Y=1/(b_0+b_1 \cdot \exp(-X))$	1/Y	exp(-X)	b_0'	b_1'
9	$Y=b_0 \cdot X^{b_1}$	lgY	lgX	$10^{b_0'}$	b_1'
10	$Y=b_0+b_1 \cdot \lg X$	Y	lgX	b_0'	b_1'
11	$Y=b_0+b_1 \cdot \ln X$	Y	lnX	b_0'	b_1'
12	$Y=b_1/(b_1+X)$	1/Y	X	$1/b_1'$	$b_0' \cdot b_0$
13	$Y=b_0 \cdot X/(b_1+X)$	1/Y	1/X	$1/b_0'$	$b_0 \cdot b_1'$
14	$Y=b_0 \cdot \exp(b_1/X)$	lnY	1/X	$\exp(b_0')$	b_1'
15	$Y=b_0 \cdot 10^{b_1 X}$	lgY	1/X	$10^{b_0'}$	b_1'
16	$Y=b_0+b_1 \cdot X^n$	Y	X^n	b_0'	b_1'

До тех пор, пока не проверены все известные формы связи, исследователь не может быть уверен, что выбрана лучшая из них (с точки зрения точности обобщения результатов опытов); исключение - случай, когда "облако" точек имеет определенную и интерпретируемую форму. Кроме приведенных в таблице 3.6 уравнений, можно использовать квадратическую парную регрессию, которая имеет вид

$$Y_i = b_0 + b_1 \cdot X_i + b_2 \cdot X_i^2 . \qquad (3.16)$$

Коэффициенты регрессии квадратического уравнения (b_0), (b_1), (b_2) находятся аналогично с уравнением (3.2), при решении системы нормальных уравнений с тремя неизвестными:

$$\begin{cases} b_0 \cdot n + b_1 \cdot \sum_{i=1}^{n} X_i + b_2 \cdot \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i; \\ b_0 \cdot \sum_{i=1}^{n} X_i + b_1 \cdot \sum_{i=1}^{n} X_i^2 + b_2 \cdot \sum_{i=1}^{n} X_i^3 = \sum_{i=1}^{n} X_i Y_i; \\ b_0 \cdot \sum_{i=1}^{n} X_i^2 + b_1 \cdot \sum_{i=1}^{n} X_i^3 + b_2 \cdot \sum_{i=1}^{n} X_i^4 = \sum_{i=1}^{n} X_i^2 Y_i. \end{cases} \quad (3.17)$$

Нетрудно заметить, что аналогичная система уравнений может иметь место при получении уравнения парной зависимости любого порядка. Статистическая оценка коэффициентов регрессии и проверка значимости уравнения регрессии, в целом, осуществляются как для линейной регрессии (раздел 3.1).

Для примера, рассмотрим следующие задачи:

а) по данным таблицы 2.4 найти вид эмпирического уравнения регрессии, выражающего зависимость расхода воды реки от размера водосборной площади.

Решение

1) Из теоретических соображений, в качестве зависимости между площадью водосбора (X) и расходом воды реки (Y) по таблице 3.6 принимаем зависимость (9): $\hat{Y}_i = b_0 \cdot X_i^{b_1}$, где b_0 и b_1 - некоторые постоянные, подлежащие определению.

2) Выполнив преобразование ($X' = \ln X$) и ($Y' = \ln Y$); ($b_0' = \ln b_0$) и ($b_1' = b_1$), получим систему нормальных уравнений для определения параметров (b_0) и (b_1):

$$\begin{cases} b_0' \cdot n + b_1' \cdot \sum_{i=1}^{n} X_i' = \sum_{i=1}^{n} Y_i'; \\ b_0' \cdot \sum_{i=1}^{n} X_i + b_1' \cdot \sum_{i=1}^{n} X_i'^2 = \sum_{i=1}^{n} (Y_i' \cdot X_i'). \end{cases}$$

Неизвестные (b_0') и (b_1'), при этом, можно найти по формулам:

$$\begin{cases} b_0' = \dfrac{\sum_{i=1}^{n}(\ln Y_i)\cdot\sum_{i=1}^{n}(\ln X_i)^2 - \sum_{i=1}^{n}(\ln Y_i\cdot\ln X_i)\cdot\sum_{i=1}^{n}(\ln X_i)\cdot}{n\sum_{i=1}^{n}(\ln X_i^2) - (\sum_{i=1}^{n}(\ln X_i)^2}; \\ b_1' = \dfrac{n\cdot\sum_{i=1}^{n}(\ln Y_i\cdot\ln X_i) - \sum_{i=1}^{n}(\ln X_i)\cdot\sum_{i=1}^{n}(\ln Y_i)}{n\cdot\sum_{i=1}^{n}(\ln X_i^2) - (\sum_{i=1}^{n}(\ln X_i)^2}. \end{cases}$$

В таблице 3.7 показана последовательность определения констант (b_0') и (b_1') решаемого уравнения. Подставляя найденные (таблица 3.7) промежуточные значения в формулы, получаем параметры (b_0' и b_1'):

$$\begin{cases} b_0' = \dfrac{16{,}6174\cdot 558{,}6992 - 99{,}6817\cdot 98{,}3439}{18\cdot 558{,}6992 - (98{,}3439)^2} = -1{,}3477\ ; \\ b_1' = \dfrac{18\cdot 99{,}6817 - 98{,}3439\cdot 16{,}6174}{18\cdot 558{,}6992 - (98{,}3439)^2} = 0{,}4156. \end{cases}$$

3) Используя преобразование (-1,3477=lnb_0) или (exp(-1,3477)), т.е. b_0=0,2598, получим уравнение регрессии -

$$\hat{Y}_i = 0{,}2598\cdot X_i^{0{,}4156}.$$

4) Подставляя в уравнение значения (X), т.е. площадь водосбора (км^2), находим расчетные величины расхода воды в реке (Y_i, м^3/с) (таблица 3.7). Сравнение теоретических ($\hat{Y}_i$) и фактических (Y_i) расходов воды для одних и тех же водосборов свидетельствует о хорошем их согласии. Средняя квадратическая погрешность вычисления расхода воды в реке по теоретической зависимости (при n=18; i=1,2,...,n) составляет

$$\delta = \sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_i)^2} = \sqrt{\frac{1}{18}\cdot 1{,}947} = \pm 0{,}329,\ м^3/с.$$

Ошибка вычисления (δ=±0,329, м^3/с) значительно меньше инструментальной погрешности измерения расхода воды в реке, что свидетельствует о хорошем согласии определяемого теоретически и замеряемого фактического расходов воды реки. Адекватность полученного уравнения проверим по критерию Фишера (формула 3.15): F=20,398/1,947=12,02; табличное значение $F^T_{(17;16,5\%)}$ = 2,398 (Приложение, табли-ца П.4.1) значительно меньше фактического, следовательно, ***искомое уравнение адекватно отражает исследуемую связь;***

Таблица 3.7 Результаты расчет параметров b_0' и b_1'

N n/n	Y_i	X_i	lnY_i	lnX_i	$(lnX_i)^2$	$lnY_i lnX_i$	$\hat{Y}_i$	$\hat{Y}_i - Y_i$
1	2	3	4	5	6	7	8	9
1	0,65	50	-0,4308	3,9120	15,3039	-1,6852	1,320	0,670
2	1,15	25	0,1398	3,2189	10,3612	0,4499	0,990	-0,160
3	1,25	75	0,2231	4,3175	18,6407	0,9634	1,563	0,313
4	1,75	50	0,5596	3,9120	15,3039	2,1892	1,320	-0,430
5	2,25	100	0,8109	4,6052	21,2076	3,7345	1,761	-0,489
6	2,10	150	0,7419	5,0106	25,1065	3,7176	2,085	-0,015
7	2,55	150	0,9361	5,0106	25,1065	4,6904	2,085	-0,465
8	2,50	200	0,9163	5,2983	28,0722	4,8548	2,349	-0,151
9	3,05	250	1,1151	5,5215	30,4865	6,1572	2,578	-0,472
10	2,75	300	1,0116	5,7038	32,5331	5,7700	2,781	0,031
11	3,15	400	1,1474	5,9915	35,8977	6,8746	3,134	-0,016
12	3,60	450	1,2809	6,1092	37,3229	7,8255	3,291	-0,309
13	3,50	550	1,2528	6,3099	39,8151	7,9048	3,577	0,077
14	4,00	650	1,3863	6,4770	41,9512	8,9790	3,834	-0,166
15	3,75	700	1,3218	6,5511	42,9167	8,6589	3,954	0,204
16	3,85	800	1,3481	6,6846	44,6840	9,0113	4,180	0,330
17	4,00	900	1,3863	6,8024	46,2726	9,4301	4,390	0,390
18	4,35	1000	1,4702	6,9078	47,7171	10,1556	4,586	0,236
Σ	50,2 $\bar{Y}$=2,79	6800 $\bar{X}$=378	16,6174	98,3439	558,6992	99,6817	--	--

*б) **по небольшой выборке (n=12) определить корреляционное отношение (η_{yx}) и построить линию регрессии**, характеризующую потери аммиака от испарения (Y%), в зависимости от его концентрации (X, кг/100 м³ воды) в поливной воде (таблица 3.8). **Нулевая гипотеза (H_0:η_{yx}=0).***

Таблица 3.8 Исходные данные: потери аммиака от испарения (ряд Y, %) в зависимости от его концентрации в поливной воде (ряд X, кг/100 м³ воды)

Номер пары	X, кг на 100 м³ воды	Y, %	Группа
1	3	25	1
2	4	23	
3	5	15	2
4	6	14	
5	8	12	3
6	8	11	
7	17	5	4
8	18	7	
9	18	7	
10	25	5	5
11	27	3	
12	45	2	

Решение

*1) **Ранжируются по возрастающей концентрации аммиака** - X (как в таблице 3.8) и **разбивается ряд** на 4...7 групп так, чтобы в каждой группе независимого признака (X) было не менее двух наблюдений. При этом, интервалы групп могут быть различны по величине. Данные таблицы 3.8 целесообразно разбить **на пять групп** (последняя колонка).*

*2) **Составляется вспомогательная таблица 3.9** и вычисляются необходимые суммы квадратов отклонений и средние. После подстановки итоговых данных таблицы 3.9 в формулы, определяются корреляционное отношение, ошибка, критерий существенности, доверительный интервал корреляционного отношения и проверяется нулевая гипотеза ($H_0: \eta_{yx} = 0$).*

Таблица 3.9. Результаты расчета вспомогательных величин для вычисления корреляционного отношения

Номер пары	X	$\bar{x}_y$	n_x	Y	$\bar{y}_x$	$Y-\bar{y}_x$	$(Y-\bar{y}_x)^2$	$Y-\bar{y}$	$(Y-\bar{y})^2$
1	3	3,50	2	25	24,00	1,00	1,00	14,25	203,06
2	4			23		-1,50	1,00	12,25	150,06
3	5	5,50	2	14	14,50	0,50	0,25	4,25	18,06
4	6			15		-0,50	0,25	3,25	10,56

Продолжение таблицы 3.9

5	8	8,00	2	12	11,50	0,50	0,25	1,25	1,56
6	8			11		-0,50	0,25	0,25	0,06
7	17			5	6,33	-1,33	1,77	-5,75	33,06
8	18	17,67	3	7		0,67	0,45	-3,75	14,06
9	18			7		0,67	0,45	-3,75	14,06
10	25			5	3,33	1,67	2,79	-5,75	33,06
11	27	32,33	3	3		-0,33	0,11	-7,75	60,06
12	45			2		-1,33	1,77	-8,75	76,56
	184= $=\sum X$	15,33= $=\bar{x}$	12= $=\bar{n}$	129= $=\sum Y$	10,75= $=\bar{y}$	0,02	10,34= $=\sum(Y-\bar{y}_x)^2$	0,00= $=\sum(Y-\bar{y})$	614,22= $=\sum(Y-\bar{y})^2$

*3) **Точки** с координатами, **соответствующими групповым средним** ($\bar{x}_y$) и ($\bar{y}_x$): 3,50 и 24,0; 5,50 и 14,50 и т. д., **наносятся на график и соединяются плавной линией** (на рисунке 3.3 эти точки обозначены крестиками), которая является линией регрессии (Y) по (X). Она показывает, что потери аммиака из поливной воды особенно резко возрастают, когда концентрация его в поливной воде составляет меньше 15 кг на 100 м³ воды. **Далее, на этот график** последовательно **переносятся все данные фактических наблюдений** (таблица 3.8) и указывается значение корреляционного отношения* $\eta_{y,x} = (0{,}99 \pm 0{,}04)$.

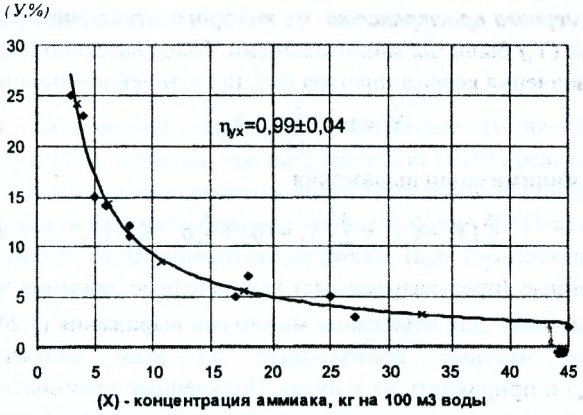


Рисунок 3.3 Зависимость потерь аммиака из аммиачных солей в связи с испарением (Y, %) от его концентрации в поливной воде (X, кг на 100 м³ воды).

3.3 Линейная множественная регрессия

При практическом анализе результатов научных исследований нередко количественное изменение изучаемого явления *(функции отклика)* зависит не от одной, а от нескольких причин *(факторов)*. При проведении экспериментов в такой множественной ситуации, *исследователь учитывает состояние функции отклика* и всех *факторов*, от которых она зависит (X_j). *Результатами* наблюдений *являются* уже *не два вектор - столбца* (Y) и (X), как при проведении парного регрессионного анализа, *а матрица результатов наблюдений:*

$$\begin{Vmatrix} Y_1 & X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1k} \\ Y_2 & X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ Y_i & X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{ik} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ Y_n & X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{nk} \end{Vmatrix}, \qquad (3.18)$$

где n - количество опытов; k - число факторов; X_{ij} - значение j-го фактора для i-го опыта; Y_i - значение функции отклика для i-го опыта. *В задачу* множественного регрессионного анализа *входит построение уравнения плоскости* (+1) - *мерного пространства*, от которого отклонения результатов наблюдений (Y_i) были бы минимальными. Следовательно, необходимо вычислить значения коэффициентов (b_0), (b_j) в линейном полиноме

$$\hat{Y}_i = b_0 + \sum_{j=1}^{k} b_j \cdot X_{ij}\ , \qquad (3.19)$$

что равносильно минимизации выражения

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\left[Y_i - (b_0 + \sum_{j=1}^{k} b_j \cdot X_{ij}) \right]^2 , \qquad (3.20)$$

где $\hat{Y}_i$ - вычисленные (предсказываемые) выровненные значения исследуемой характеристики. Для отыскания минимума выражения (3.20), необходимо найти частные производные по всем неизвестным ($b_0, b_1, b_2, \dots, b_j, \dots, b_k$) и приравнять их к нулю. Полученные *зависимости образуют систему нормальных уравнений*, при решении которых рассчитываются неизвестные коэффициенты регрессии. Данный способ отыскания неизвестных коэффициентов регрессии (*способ наименьших квадратов*) детально *рассмотрен в разделе* 3.1.

Рассмотрим другой *способ отыскания коэффициентов регрессии с помощью детерминантов или определителей*, широко используемый в гидролого-гидрогеологических исследованиях. В этом случае, общее выражение для коэффициентов регрессии (b_j) можно записать в форме

$$b_j = (-1)^{j+1} \cdot \frac{\sigma_y}{\sigma_{x_j}} \cdot \frac{D_{1j}}{D_{11}}, \qquad (3.21)$$

где σ_y - среднее квадратическое отклонение зависимой переменной (функции); σ_{xj} - среднее квадратическое отклонение независимой переменной; D_{1j} и D_{11} - миноры определителя (D) (2.22). В результате проведенных операций полином первой степени (3.19) получается с известными коэффициентами (b_0), (b_j). Проверка значимости (качества предсказания) множественного уравнения регрессии, в принципе, мало отличается от соответствующей проверки парной зависимости. Остаточная дисперсия вычисляется по формуле

$$\bar{S}^2_{ост.} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}, \qquad (3.22)$$

и, затем, сравнивается с дисперсией среднего $\left(\bar{S}^2_y\right)$ с помощью F-критерия Фишера -

$$F = \frac{\bar{S}^2_Y}{\bar{S}^2_{ост.}}, \qquad (3.23)$$

с числом степеней свободы в числителе ($\nu_1=(n-1)$), в знаменателе - ($\nu_2=(n-k-1)$). Считается, что по уравнению (3.19) предсказываются результаты опытов лучше среднего, если (F) превышает или достигает границы значимости при выбранном ее уровне: обычно (P=(1-q)=5%).

Значимость коэффициентов регрессии (b_0), (b_j) проверяется по t-критерию Стьюдента

$$t = \frac{|b_j|}{S_{b_j}}, \qquad (3.24)$$

в котором $\bar{S}_{b_j}$ - погрешность коэффициента регрессии, определяемая как

$$\bar{S}_{b_j} = \frac{\sigma_Y}{\sigma_{X_j}} \cdot \sqrt{\frac{1-R^2}{n-k} \cdot \frac{\Delta_{jj}}{D_{11}}}, \qquad (3.25)$$

где Δ_{jj} - минор определителя (D), который получается из (D) вычеркиванием (j+1) строки столбца. Вычисленное значение (t) сравнивается с (t^T), при числе степеней свободы (v=(n-k-1)). Доверительный интервал для коэффициентов регрессии устанавливается по соотношению

$$b_j - t^T \cdot \bar{S}_{b_j} \le \beta_j \le b_j + t^T \cdot \bar{S}_{b_j}, \qquad (3.26)$$

где β_j - соответствующее значение для коэффициентов регрессии в генеральной совокупности.

На примере приведения к многолетнему периоду величин годового стока реки Птичь - с. Лучицы, рассмотрим порядок применения линейной множественной регрессии).

Для этого необходимо построить уравнение множественной регрессии зависимости расхода реки Птичь-с.Лучицы от расходов рек - аналогов. В качестве аналогов, как и в других примерах, принимаем реки Ясельду - с. Сенин, Орессу - с. Верхутино, Орессу - с. Андреевка. Парные коэффициенты корреляции приведены в таблице 2.7.

Решение

1) Определитель*) (D) и его миноры будут равны:

$$D = \begin{vmatrix} 1 & 0{,}8230 & 0{,}6464 & 0{,}9537 \\ 0{,}8230 & 1 & 0{,}6510 & 0{,}8255 \\ 0{,}6464 & 0{,}6510 & 1 & 0{,}6635 \\ 0{,}9537 & 0{,}8255 & 0{,}6635 & 1 \end{vmatrix} = 0{,}010 ;$$

$$D_{yy} = \begin{vmatrix} 1 & 0{,}6510 & 0{,}8255 \\ 0{,}6510 & 1 & 0{,}6635 \\ 0{,}8255 & 0{,}6635 & 1 \end{vmatrix} = 0{,}1676 ;$$

*) - **определитель** - детерминант (квадратной матрицы порядка "n" - многочлен от элементов матрицы), каждый член которого снабжен определенным знаком и является произведением "n" - элементов, взятых по одному из каждой строки и каждого столбца.

$$D_{yx_1} = \begin{vmatrix} 0,8230 & 0,6510 & 0,8255 \\ 0,6464 & 1 & 0,6635 \\ 0,9537 & 0,6635 & 1 \end{vmatrix} = 0,0186 ;$$

$$D_{yx_2} = \begin{vmatrix} 0,8230 & 1 & 0,8255 \\ 0,6464 & 0,6510 & 0,6635 \\ 0,9537 & 0,8255 & 1 \end{vmatrix} = -0,0006 ;$$

$$D_{yx_3} = \begin{vmatrix} 0,8230 & 1 & 0,6510 \\ 0,6464 & 0,6510 & 1 \\ 0,9537 & 0,8255 & 0,6635 \end{vmatrix} = 0,1441$$

2) Подсчет средних значений и средних квадратических отклонений исходных рядов (см. таблицу 2.3) дает следующие оценки:

$$\overline{Y} = 45,31, \quad \overline{X}_1 = 18,55, \quad \overline{X}_2 = 2,74, \quad \overline{X}_3 = 16,86,$$
$$\sigma_y = 14,31, \quad \sigma_{x_1} = 8,14, \quad \sigma_{x_2} = 0,83, \quad \sigma_{x_3} = 0,1441.$$

3) Коэффициенты регрессии уравнения связи определяются по формуле (3.21):

$$b_1 = (-1)^{1+1} \cdot \frac{14,31 \cdot 0,0186}{8,14 \cdot 0,1676} \approx 0,1951 ; b_2 = (-1)^{2+1} \cdot \frac{14,31 \cdot (-0,0006)}{0,83 \cdot 0,1676} = 0,0668 ;$$

$$b_3 = (-1)^{3+1} \cdot \frac{14,31 \cdot 0,1441}{5,72 \cdot 0,1676} = 2,1511.$$

4) Уравнение регрессии, согласно исходным данным, получает вид -

$$Y - \overline{Y} = b_1 \cdot (X_1 - \overline{X}) + b_2 \cdot (X_2 - \overline{X}) + b_3 \cdot (X_3 - \overline{X}),$$

или-

$$Y - 45,31 = 0,1951 \cdot (X_1 - 18,55) + 0,0668 \cdot (X_2 - 2,74) + 2,1511 \cdot (X_3 - 16,86),$$

т.е.-

$$Y = 0,1951 \cdot X_1 + 0,0668 \cdot X_2 + 2,1511 \cdot X_3 + 5,24.$$

5) Погрешность полученной связи распределяется как

$$\sigma_{\overline{Y}} = \sigma_Y \cdot \sqrt{1 - R^2} = 14,31 \cdot \sqrt{1 - 0,969^2} = 3,54 .$$

Погрешности коэффициентов регрессии (b_1), (b_2), (b_3), применительно к случаю четырех переменных, согласно формулы (3.25), определяются как:

$$\overline{S}_{b_1} = \frac{\sigma_{\overline{Y}}}{\sigma_{X_1}} \cdot \sqrt{\frac{1 - r^2_{X_1X_2}}{(n-3) \cdot (1 - r^2_{X_1X_2} - r^2_{X_1X_3} - r^2_{X_2X_3} + 2 \cdot r_{X_1X_2} r_{X_1X_3} r_{X_2X_3})}} =$$

$$= \frac{3,54}{8,14} \cdot \sqrt{\frac{1 - 0,6464^2}{(28-3) \cdot (1 - 0,6510^2 - 0,8255^2 - 0,6635^2 + 2 \cdot 0,6510 \cdot 0,8255 \cdot 0,6635)}} = 0,1589 ;$$

$$\overline{S}_{b_2} = \frac{\sigma_{\overline{Y}}}{\sigma_{X_2}} \cdot \sqrt{\frac{1 - r^2_{X_1X_2}}{(n-3)\cdot(1 - r^2_{X_1X_2} - r^2_{X_1X_3} - r^2_{X_2X_3} + 2 \cdot r_{X_1X_2} r_{X_1X_3} r_{X_2X_3})}} =$$

$$= \frac{3{,}54}{0{,}83} \cdot \sqrt{\frac{1 - 0{,}8255^2}{(28-3)\cdot(1 - 0{,}6510^2 - 0{,}8255^2 - 0{,}6635^2 + 2 \cdot 0{,}6510 \cdot 0{,}8255 \cdot 0{,}6635)}} = 1{,}1758;$$

$$\overline{S}_{b_1} = \frac{\sigma_{\overline{Y}}}{\sigma_{X_1}} \cdot \sqrt{\frac{1 - r^2_{X_1X_2}}{(n-3)\cdot(1 - r^2_{X_1X_2} - r^2_{X_1X_3} - r^2_{X_2X_3} + 2 \cdot r_{X_1X_2} r_{X_1X_3} r_{X_2X_3})}} =$$

$$= \frac{3{,}54}{5{,}72} \cdot \sqrt{\frac{1 - 0{,}6510^2}{(28-3)\cdot(1 - 0{,}6510^2 - 0{,}8255^2 - 0{,}6635^2 + 2 \cdot 0{,}6510 \cdot 0{,}8255 \cdot 0{,}6635)}} = 0{,}2295.$$

6) Доверительный интервал для коэффициентов регрессии определяется по формуле (3.26) *(при значении $t^T = 1{,}711$, числе степеней свободы - ν=28-3-1=24, согласно Приложению, таблица П.2):*

$$0{,}1951 - 1{,}711 \cdot 0{,}1589 \le \beta_1 \le 0{,}1951 + 1{,}711 \cdot 0{,}1589;$$
$$-0{,}0768 \le \beta_1 \le 0{,}4670;$$
$$0{,}0668 - 1{,}711 \cdot 1{,}1758 \le \beta_2 \le 0{,}0668 + 1{,}711 \cdot 1{,}1758;$$
$$-1{,}9450 \le \beta_2 \le 2{,}0786;$$
$$2{,}1511 - 1{,}711 \cdot 0{,}2295 \le \beta_3 \le 2{,}1511 + 1{,}711 \cdot 0{,}2295;$$
$$1{,}7584 \le \beta_3 \le 2{,}5438.$$

7) Значимость коэффициентов регрессии (b_1), (b_2), (b_3) определяется с использованием критерия Стьюдента по формуле (3.24) :

$$t_1 = \frac{|b_1|}{\overline{S}_{b_1}} = \frac{0{,}1951}{0{,}1589} = 1{,}23 < t^T = 1{,}711;$$

$$t_2 = \frac{|b_2|}{\overline{S}_{b_2}} = \frac{0{,}0668}{1{,}1758} = 0{,}06 < t^T = 1{,}711;$$

$$t_3 = \frac{|b_3|}{\overline{S}_{b_3}} = \frac{2{,}1511}{0{,}2295} = 9{,}37 > t^T = 1{,}711.$$

Судя по критерию Стьюдента, коэффициенты регрессии (b_1) и (b_2) статистически незначимы, следовательно, они могут быть исключены из полученного уравнения (имеет место достаточно высокая корреляция между расходами рек - аналогов).

8) Для проверки качества предсказания результатов по полученному уравнению, по

формуле (3.22) вычисляется остаточная дисперсия

$$\bar{S}^2_{ост.} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1} = \frac{2285,82}{28-3-1} = 95,24,$$

которая сравнивается с дисперсией среднего ($\bar{S}_Y = 204,78$) и с помощью F"-критерия, определенного по Приложению (таблица П.4.1) при числе степеней свободы - (ν_1=28-1=27), (ν_2=28-3-1=24) и 5%-ном уровне значимости, устанавливается соотношение (F) и (F^m)

$$F = \frac{\bar{S}_y}{\bar{S}^2_{ост}} = \frac{204,78}{95,84} = 2,15 > F^m = 1,98.$$

Вывод

Полученное уравнение предсказывает годовые расходы воды реки Птичь-с. Лучицы в 2,15-раза лучше, чем их средняя величина.

3.4 Нелинейная множественная регрессия

В тех случаях, *когда методами линейной множественной регрессии не удается получить* приемлемую *математическую модель, прибегают к моделям нелинейной множественной регрессии.*

Первый этап нелинейного множественного анализа - получение, так называемой, квадратичной формы или *модели второго порядка*, имеющей вид

$$\begin{aligned} Y_j = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + ... + \beta_j \cdot X_{ji} + ... + \beta_k \cdot X_{ki} + \\ + \beta_{22} \cdot X^2_{2i} + ... + \beta_{jj} \cdot X^2_{ji} + ... + \beta_{kk} \cdot X^2_{ki} + + \beta_{12} \cdot X_{1i} \cdot X_{2i} + \\ + \beta_{13} \cdot X_{1i} \cdot X_{3i} + ... + \varepsilon_i. \end{aligned} \qquad (3.27)$$

Аналогичным способом *можно получить* и *модель более высоких порядков*, повышая степень уравнения до тех пор, пока уменьшается остаточная дисперсия $\left(\bar{S}^2_{ост.}\right)$. Для определения оценок коэффициентов регрессии (b_0), (b_j), (b_{jm}), (b_{jj}), *модель* типа (3.27) *приводится к линейному виду* путем замены переменных: $X'_{1i}=X_{1i}$; $X'_{2i}=X_{2i}$, ...$X'_{k+1,i}=X_{1i}^2$; $X'_{k+2,i}=X_{2,i}^2$ ...; затем, *параметры* нового (*расширенного*) линейного *полинома определяются по методике, приведенной в разделе* 3.3. *В случае неадекватности модели второго порядка, используется модель третьего порядка.* Однако, вряд ли стоит механически добавлять в модель члены более высоких порядков. Часто оказывается продуктивным исследование возможностей

каких-то иных преобразований предиктов, откликов или и тех, и других, одновременно. Таким образом, *другой формой* проведения *нелинейного регрессионного анализа является использование,* так называемых, *"внутренне линейных" форм* уравнений, т.е. форм, которые легко линеаризируются логарифмированием или другим преобразованием.

Рассмотрим **некоторые из приемов**, используемых в практике **нелинейного регрессионного анализа:**

а) мультипликативная модель -

$$Y_i = \beta_0 \cdot X_{1i}^{\beta_1} \cdot X_{2i}^{\beta_2} \cdot \ldots \cdot X_{ji}^{\beta_j} \cdot \ldots \cdot X_{ki}^{\beta_k} \cdot \varepsilon_i \,, \qquad (3.28)$$

где β_0, β_1, β_2,..., β_j, ...β_k - неизвестные параметры; ε_i - мультипликативная случайная ошибка. Логарифмирование уравнения (3.28) по основанию позволяет перевести модель в линейную форму

$$\begin{aligned} \ln Y_i = \ln\beta_0 + \beta_1 \cdot \ln X_{1i} + \beta_2 \cdot \ln X_{2i} + \ldots \\ \ldots + \beta_j \cdot \ln X_{ji} + \ldots + \beta_k \cdot \ln X_{ki} + \ln\varepsilon_i \,. \end{aligned} \qquad (3.29)$$

Далее, производится замена переменных : $Y'_i = \ln Y_i$; $\beta'_0 = \ln\beta_0$; $X'_{1i} = \ln X_{1i}$; $X'_{2i} = \ln X_{2i}$,,,;и т.д. Преобразованная модель (3.29), с заменой переменных, имеет форму уравнения (3.19) и, поэтому, возможно применение стандартных методов исследования линейной регрессии, описанных в разделе 3.3. В конце расчетов производится только одно обратное преобразование для получения величины (β_0);

б) "обратная" модель -

$$Y_i = \frac{1}{\beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \ldots + \beta_j \cdot X_{ji} + \ldots + \beta_k \cdot X_{ki} + \varepsilon_i} \,. \qquad (3.30)$$

Обращая обе части, получим

$$\frac{1}{Y_i} = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \ldots + \beta_j \cdot X_{ji} + \ldots + \beta_k \cdot X_{ki} + \varepsilon_i \,; \qquad (3.31)$$

в) экспоненциальная модель -

$$Y_i = \exp(\beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \ldots + \beta_j \cdot X_{ji} + \ldots + \beta_k \cdot X_{ki}) \cdot \varepsilon_i \,, \qquad (3.32)$$

логарифмируя обе части по натуральному основанию, получим

$$\ln Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \ldots + \beta_j \cdot X_{ji} + \ldots + \beta_k \cdot X_{ki} + \ln\varepsilon_i \,; \qquad (3.33)$$

г) комбинированная модель -

$$Y_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \ldots + \beta_j \cdot X_{ji} + \ldots + \beta_k \cdot X_{ki} + \varepsilon_i)} \,, \qquad (3.34)$$

применяя обращение и вычитая единицу, затем, логарифмируя по натуральному основанию обе части уравнения, получаем -

$$\ln(\frac{1}{Y_i} - 1) = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + ... + \beta_j \cdot X_{ji} + ... + \beta_k \cdot X_{ki} + \varepsilon_i \ . \quad (3.35)$$

Данный пример (последовательного преобразования зависимой переменной) ***используется для приведения сложной нелинейной модели к линейному виду. Более простые модели преобразований, приводятся ниже;***

д) "обратное" преобразование -

$$Y_i = \beta_0 + \beta_1 \cdot \frac{1}{X_{1i}} + \beta_2 \cdot \frac{1}{X_{2i}} + ... + \beta_j \cdot \frac{1}{X_{ji}} + ... + \beta_k \cdot \frac{1}{X_{ki}} + \varepsilon_i \ ; \quad (3.36)$$

е) логарифмическое преобразование -

$$Y_i = \beta_0 + \beta_1 \cdot \ln X_{1i} + \beta_2 \cdot \ln \cdot X_{2i} + ... + \beta_j \cdot \ln X_{ji} + ... + \beta_k \cdot \ln X_{ki} + \varepsilon_i \ ; \quad (3.37)$$

ж) преобразование типа квадратного корня -

$$Y_i = \beta_0 + \beta_1 \cdot \sqrt{X_{1i}} + \beta_2 \cdot \sqrt{X_{2i}} + ... + \beta_j \cdot \sqrt{X_{ji}} + ... + \beta_k \cdot \sqrt{X_{ki}} + \varepsilon_i \ . \quad (3.38)$$

Мерой тесноты связи в нелинейной зависимости ***служит множественное корреляционное отношение***, которое вычисляется по формуле (2.26), при использовании для вычисления ($\hat{Y}_i$) нелинейной формы уравнения. *Сравнение множественного корреляционного отношения с коэффициентом множественной корреляции*, вычисленным по линейной форме, *дает* некоторое *представление о "кривизне"* изучаемой *зависимости.* Ясно, что *существует множество преобразований* и можно постулировать модели, содержащие меньшее или большее количество факторов. В одной и той же модели может содержаться несколько различных преобразований. Нередко трудно решить, что предпочесть, если можно сделать любое преобразование. Часто выбор осуществляется на основе предыдущих знаний о входящих в модель факторах. Цель преобразований такого рода состоит в том, чтобы получить для преобразованных переменных более простую регрессионную модель, чем для исходных. Процедура множественного регрессионного анализа громоздка, трудоемка и изначально ориентируется на применение ЭВМ.

Множественный нелинейный регрессионный анализ нами использовался для получения регрессионной модели нормы годового расхода малых рек Беларуси

$$Q = 2{,}3 \cdot 10^{-6} \cdot X^{0.668} \cdot (\varphi + 300)^{0.386} \cdot (J_P + 1)^{0.193} \cdot (f_{сл} + f_{зл} + 1)^{0.107} \cdot F^{1.122} \ ,$$

где $\overline{Q}$ - норма годового стока, м^3/с; X - годовая норма атмосферных осадков, мм; (φ+300) - географический параметр, в котором φ - широта центра тяжести водосбора в километровой сетке с началом координат в Минске, км; $J_р$ - уклон реки, ‰; $f_{сл}$ - площадь водосбора, занятая сухим лесом, %; $f_{зл}$ - площадь водосбора, занятая лесом на заболоченных землях, %; F - общая площадь водосбора, км^2. Линейное корреляционное отношение предлагаемой нами модели - R=0,987±0,0003.

О трудоемкости построения и испытания данной модели можно судить по тому, что нами использовались 12 факторов, в той или иной степени влияющих на формирование стока более чем 100 рек Беларуси, в различном их сочетании, с соответствующей отбраковкой несущественных из них, апробацией и оптимизацией модели на независимых экспериментальных данных.

3.5 Выбор оптимальной модели

Теоретически, точность аппроксимации *можно повысить* одновременно с повышением степени полинома, однако, *практически, для полиномов высоких степеней*, при проведении матричных операций на ЭВМ *накапливаются* столь *значительные погрешности округления*, что решение становится невозможным. Поэтому, *на практике* обычно *ограничиваются построением полинома второго порядка и проведением шагового регрессионного анализа,* с включением или исключением переменных. *Для выбора оптимальной формы регрессии было бы лучше* всего вычислить коэффициенты всех возможных уравнений регрессии, а затем *выбрать* "лучшее" *уравнение по минимальной остаточной дисперсии или значению коэффициента корреляции.* Однако, такое решение задачи не всегда реально. *Например*, общее число уравнений при 10 определяющих факторах составляет 2^{10}=1024. Если сюда добавить и нелинейные формы уравнений, то *чисто технически решение* такой *задачи проблематично. Возникают трудности, также, при выборке из нескольких уравнений с одинаковыми статистически значимыми критериями*, когда *необходимо принять во внимание следующие моменты*: 1) могут быть основательные теоретические причины, которые заставляют нас поверить в то, что регрессионное уравнение имеет определенную математическую форму; 2) регрессионное уравнение должно обеспечить наибольшее приближение к наблюденным значениям (тогда уравнение может быть с достаточной уверенностью использовано в прогнозных целях); 3) регрессионное уравнение должно быть как можно более простым, но, при этом, адекватно описывать ис-

следуемые явления; исходя из требований пунктов 2,3, очевидна необходимость компромисса, и выбор "наилучшего" регрессионного уравнения становится субъективным.

При формальных методах отбора факторов для сравнения их важности *имеются пять показателей*: 1) коэффициенты парной корреляции между функцией отклика и изучаемым фактором; 2) частные коэффициенты корреляции; 3) t-критерии Стьюдента; 4) остаточные дисперсии ($\bar{S}^2_{ост.}$); 5) частные F-критерии Фишера.

Решающее значение при выборе того или иного фактора имеют природа явления или даже интуитивные представления о важности этих факторов. В литературе встречается *множество подходов к построению регрессионных моделей*, но, в принципе, это *методы: включения переменных; исключения переменных*.

Алгоритм множественного регрессионного анализа методом включения факторов (переменных) следующий:

1) вычислить частные коэффициенты корреляции;
2) ранжировать факторы по значению коэффициентов корреляции ($r_{yx_j} > r_{yx_{j+1}}$);
3) построить линейную модель с одним фактором, у которого частный коэффициент корреляции наибольший;
4) вычислить (R), (R^2) и $\left(\bar{S}^2_{ост.}\right)$;
5) добавить следующий фактор, у которого коэффициент корреляции наибольший из оставшихся и построить регрессионную модель типа (3.19);
6) вычислить критерии для оценки качества предсказания результатов опыта на базе модели (R, R^2 и $\bar{S}^2_{ост.}$);
7) вычислить выражения -

$$\left|\frac{S_{j+1}}{\sigma_y} - \frac{S_j}{\sigma_y}\right| > \varepsilon \ ; \qquad (3.39)$$

и

$$\left|R_{j+1} - R_j\right| > \beta \ , \qquad (3.40)$$

в которых S_j и S_{j+1}- относительные ошибки при использовании (j) и (j+1)-факторов; σ_y - среднеквадратическое отклонение исследуемо-

го ряда; R_j и R_{j+1} - соответственно, коэффициенты множественной корреляции; ε и β - заданные пороги чувствительности модели;

8) если условия (3.39) и (3.40) выполняются, то привлекается очередной фактор и расчет продолжается с пункта 5;

8.1) если условия (3.39) и (3.40) не выполняются и первые из (j) - факторов значимы, модель считается построенной.

Алгоритм множественного регрессионного анализа методом исключения факторов (переменных) основан на принципах, изложенных выше (для метода включения переменных):

1) уравнение регрессии расширяется сразу до полной квадратичной, при возможности, - до полной кубической формы;

2) исключение начинается с фактора, имеющего наименьший t-критерий Стьюдента;

3) после исключения очередного фактора, для нового уравнения регрессии, вычисляются: множественный коэффициент корреляции, остаточная дисперсия, и т.д.;

4) для прекращения исключения факторов используются подходы-

4.1) исключение прекращается, начиная с момента увеличения остаточной дисперсии,

4.2) то же, исходя из условия назначения уровня значимости (α, %) при вычислении t-критерия Стьюдента для последнего оставляемого фактора.

4 МЕТОДИКА АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

В предыдущих разделах, *определяя* расчетным путем, *как случайные величины*, некоторые *гидролого-климатические характеристики*, мы *не касались методик учета характера их появления во времени. Случайная функция* представляет собой *обобщенное понятие случайной величины при ее распределении*, как правило, *относительно координат пространства и времени*. Причем, случайная величина, получаемая опытным путем, имеет неизвестные заранее значения. *Конкретный вид* случайной функции *называется реализацией случайной функции.*

При исследовании природных явлений целесообразно, а *иногда* и просто *необходимо, считать их случайными процессами. Случайным временным или стохастическим процессом принято называть некоторое множество функций времени*, которое *можно описать некоторыми статистическими закономерностями*. Каждая *отдельная функция времени* называется *выборочной функцией* или, *при конечном интервале времени, - реализацией случайного процесса.*

Временной ряд - есть перечень значений случайной переменной в зависимости от времени. Обычно, *интервал времени* между наблюдениями случайной переменной (временной интервал) *постоянен.*

Класс случайных функций, или случайных *процессов*, включает в себя широкий набор процессов (от случайных, не зависящих от предшествующих значений, до строго периодических, легко предсказываемых в последующее время).

Классификация случайных *процессов* обычно выполняется при их группировке по следующим признакам:

1) *по числу реализаций* с разделением их на *многомерные* и *одномерные* (*одномерные* - частный случай многомерных; *считаются удовлетворяющими*, как правило, *условиям стационарности и эргодичности*);

2) *по зависимости* их *характеристик от времени* (*стационарные* - в которых средняя амплитуда и характер колебаний "X(t)" существенно не изменяются во времени; *нестационарные* - имеющие определенную тенденцию изменяться во времени). С известным приближением, на некоторых временных отрезках, нестационарные процессы могут считаться стационарными. Кроме того, большое число процессов, которые не могут по определению рассматриваться как стационарные, можно с допущениями описать как стационарные. Например, если нестационарность функции

вызвана изменяющимся во времени математическим ожиданием - $m_X(t)$, то процесс легко может быть приравнен к стационарному преобразованием типа:

$$x(t) - m_x(t) = \Delta_x(t) \; ; \qquad (4.1)$$

$$x(t) / m_x(t) = k(t) \; , \qquad (4.2)$$

приводящим к математическому ожиданию, соответственно, равному нулю или единице. Случайный процесс называется стационарным, в широком смысле, если его математическое ожидание не зависит от времени (t), а корреляционная функция - $r_X(t)$ является только функцией интервала времени

$$m_x(t) = m_x = \text{const} \; ; \quad r_x(t, t+\tau) = r_x \; ; \qquad (4.3)$$

3) *по множеству реализаций наличия (отсутствия) связей между характеристиками (эргодические и неэргодические). Эргодическое свойство стационарного процесса* состоит в том, что *каждая отдельная выборка может характеризовать* всю генеральную *совокупность*. Среднее (по времени) и другие характеристики случайного процесса можно приближенно определять по одной, достаточно длинной выборке. В качестве формального признака используется оценка затухания корреляции до нуля (если автокорреляционная функция (r(t)), при ($t \to \infty$), стремится к нулю, - *процесс является эргодическим)*. Если *автокорреляции при увеличении интервалов* между сечениями (t_2), (t_1) *стремятся к* некоторой (*отличной от нуля*) величине, то такая *функция не является эргодичной. Эргодичность* и *стационарность* - это *два различных свойства случайных процессов*. Каждый *эргодический процесс является стационарным*, тогда как *стационарный процесс не обязательно* должен *быть эргодическим*. Благодаря свойству эргодичности стационарных процессов, отпадает необходимость исследовать большое число выборок, данные о которых, как правило, отсутствуют, но достаточно исследовать одну выборку в течение длительного периода. Это свойство некоторых случайных процессов, как и свойство стационарности, существенно облегчает решение как гидрометеорологических задач, так и любых других задач рационального природопользования. Поскольку практические наблюдения гидрометеорологических, гидролого-климатических и других природных явлений все же ограничены во времени, можно признать справедливым

равенство во времени среднего значения множества выборок и среднего значения лишь с некоторой степенью достоверности;

4) *по типу закона распределения случайной функции в каждом сечении* (*нормальные* - гауссовские, если любая система сечений $x(t_1)$, $x(t_2)$,...,$x(t_n)$ подчиняется нормальному закону; *пуассоновские*; *биноминальные*, и *другие*);

5) *по типу корреляционных функций* (в эколого - метереологических исследованиях - *величина максимального сдвига* "τ_{max}" между связанными сечениями случайной функции; при $\tau_{max} = 1$, имеют место *процессы без последствия - непереходные*, при $\tau_{max} > 1$, *процессы с последствием - переходные*). *Случайные процессы*, в которых *корреляционные связи* $(r(t,t')\neq 0)$, *при* $(t=t')$, называются *цепями Маркова*. При этом, *если* для каждого момента времени *вероятность* любого состояния системы *в будущем зависит* только *от состояния системы в настоящий момент* (t_0) и *не зависит от предшествующих* значений, математическая *модель процесса называется простой цепью Маркова*. В таком процессе автокорреляция имеет место только между смежными членами ряда

$$r(t, t+l) = r(\tau = l) \neq 0 \ . \qquad (4.4)$$

Случайная последовательность - x(t), в которой для каждого момента времени (t) значения и вероятность любого состояния системы в будущем зависят от ее состояния в предшествующие моменты времени- t_i(i=1,2,3,...,n), называется *сложной цепью Маркова. Процессы*, связанные по типу цепей Маркова с оценками вероятностей перехода из одного состояния в другое, *часто называются переходными*;

6) *по характеру аналитического описания* (*с наличием* или *отсутствием тренда*); в общем виде запись процесса (Z(t)) осуществляется -

как *при наличии трендовой составляющей*

$$Z(t) = X_0(t) + X_T(t) + X_u(t) + \xi(t) \ , \qquad (4.5)$$

так *при отсутствии тренда*

$$Z(t) = X_u(t) + \xi(t) \ , \qquad (4.6)$$

где $X_0(t)$ - постоянная (средняя); $X_T(t)$ - трендовая; $X_Ц(e)$ - циклическая; $\xi(t)$ - случайная (шумовая) независимая составляющие. В зависимости от класса стационарного эргодического процесса, более развернутое аналитическое представление для одномерных стационарных временных про-

цессов (рядов) может записываться следующими соотношениями:

$$X(t) = A \cdot \cos(\omega \cdot t) + \xi(t) \; ; \qquad (4.7)$$

$$X(t) = \xi(t) + \alpha \cdot \xi(t-1) \; ; \qquad (4.8)$$

$$X(t) + \beta \cdot X(t-1) = \xi(t) \; ; \qquad (4.9)$$

$$X(t) = P(t) + q(t) \cdot \xi(t) \; , \qquad (4.10)$$

в которых P(t) и q(t) - полиномы по (t). Уравнение (4.7) описывает линейный циклический процесс, у которого A - амплитуда, ω - основная угловая частота. С его помощью, в частности, можно было бы описать годовой ход температуры воздуха, сглаженных сумм атмосферных осадков или суммарного испарения с поверхности почвы. Уравнение (4.8) описывает процессы скользящего осреднения, где α - весовой коэффициент, учитывающий предысторию. Уравнение (4.9) характеризует простой авторегрессионный процесс, называемый также марковским процессом 1-го порядка, или простой инерционностью. Здесь, настоящее значение (X(t)) определяется предыдущим (X(t-1)). Независимо от ожидаемого типа колебаний, в целях физической интерпретации, ряд обычно раскладывается на сезонную, циклическую компоненты, компоненту тренда и случайный остаток. В связи с этим, *анализ временных рядов* целесообразно *проводить по-этапно*: 1) выделение периодических, регулярных и сезонных циклов (годовой, сезонный, суточный ход); 2) выделение нерегулярных циклов (тренд, непериодические, квази-периодические составляющие); 3) сглаживание и фильтрация отдельных частот; 4) проверка на случайность колебаний; 5) анализ однородности колебаний во времени и в пространстве; 6) прогноз колебаний.

Практика анализа временных рядов показывает, что общий ход или колебание гидрометеорологической случайной величины во времени представляет собой сумму нескольких колебаний. Некоторые из них носят периодический характер, другие - непериодический. Анализ любых временных рядов проводится, во-первых, с целью разделения их на периодические и непериодические компоненты и, во-вторых, - для изучения каждой из компонент, в отдельности.

4.1 Анализ периодических колебаний

Выделение регулярных циклов

К периодическим колебаниям в гидрометеорологических временных рядах *относятся годовой и суточный ход* гидролого-климатических, тепловодобалансовых и других характеристик. Годовой ход гидрометеорологической величины лучше всего можно проанализировать, вычислив средние значения для каждого месяца или сезона и представив их в графическом виде в зависимости от времени. Также, можно представить и средние суточные величины, но подобный ряд безусловно даст случайные колебания, обусловленные короткопериодическими нерегулярными изменениями. Суточный ход обычно анализируется либо по результатам наблюдений в течение только одного месяца, либо по средним величинам за весь год, годовой ход анализируется - по результатам наблюдений в один и тот же час суток, - по средним суточным данным.

Гармонический анализ

Гармонический анализ получил наибольшее распространение при исследовании периодического хода гидрометеорологических параметров. *В задачу гармонического анализа входит разложение функций на простейшие периодические*. Такой анализ позволяет понять физическую сущность периодических колебаний. Исходя из основных принципов математического анализа, *любую функцию*, заданную в каждой точке интервала, *можно представить бесконечным рядом* синусоидальных и косинусоидальных функций (рядом Фурье). *Метод* подобного нахождения функций *называется анализом Фурье*. Если в анализируемом ряду имеется конечное число точек, то все результаты наблюдений могут быть выражены конечным числом синусов и косинусов. *Представление рядов конечной суммой членов с синусами и косинусами называется гармоническим анализом. Первая* (основная) *гармоника**) имеет период, равный длине всего исследуемого периода. *Вторая гармоника* имеет период, равный половине основного периода, *третья* - период, равный третьей части основного и т.д. Вообще, если число наблюдений (n), то число гармоник (n/2). *Различные гармоники* выделяются таким образом, что каждую из них можно рассматривать как независимый объект и объяснять разными физическими причинами. Часто ход гидрометеорологической характеристики не может

*) - **гармоника** - простейшая периодическая функция, например, $y=a\sin(\omega x+\varphi_0)$

быть объяснен полно, в то время как отдельные гармоники этому объяснению поддаются. Однако, в каждую гармонику, в отдельности, не обязательно вкладывать отчетливый физический смысл. Так будет всякий раз, когда периодическая функция не носит синусоидального характера. Не всегда определяются все (n/2) - гармоник. Обычно, изменение периодической функции достаточно хорошо описывается первыми двумя или, в крайнем случае, - тремя гармониками. Но, в случае периодических функций, дело обстоит совершенно иначе. *Полную сумму, задающую случайную переменную* (X(t)), *можно записать в виде*

$$X(t) = \overline{X} + \sum_{i=1}^{n/2}(A_i \cdot \sin(\frac{2 \cdot \pi}{P} \cdot i \cdot t) + B_i \cdot \cos(\frac{2 \cdot \pi}{P} \cdot i \cdot t)) \ . \qquad (4.11)$$

То есть, временная сумма есть среднее плюс сумма всех (n/2) - гармоник. Здесь (P) - основной период, или полный период периодической функции; (i) - номер гармоники. Следует отметить, что (P) не всегда равно (n). Если наблюдения производить каждые два часа одного дня, то (n=12), а (P=24). Первые два члена суммы (4.11) проходят полный цикл за один основной период, а быстрее всего меняется последний член, имеющий период (2·P/n), и если существуют какие-либо более короткие периоды, то их можно обнаружить только на основании более частых наблюдений. *Гармонический анализ* начинается с нахождения коэффициентов приведенного ряда, каждый из которых может быть вычислен независимо от другого по следующим формулам:

$$A_i = \frac{2}{n} \cdot \sum_{t=1}^{n} X(t) \cdot \sin(\frac{2 \cdot \pi}{P} \cdot i \cdot t) \ ; \qquad (4.12)$$

$$B_i = \frac{2}{n} \cdot \sum_{t=1}^{n} X(t) \cdot \cos(\frac{2 \cdot \pi}{P} \cdot i \cdot t) \ , \qquad (4.13)$$

где i - может иметь любое целое значение от 1 до (n/2-1). Для последней гармоники (An=0), (Bn) - равно величине, получаемой из (4.13), но деленной на 2, т.е. $(\frac{Bn}{2})$. Далее рассчитываются значения амплитуды i-ой гармоники и фазы (времени, при котором i-ая гармоника имеет

максимум - φ_i)

$$C_i = \sqrt{A_i^2 + B_i^2} \;\; ; \;\; \varphi_i = \frac{P}{2 \cdot \pi \cdot i} \cdot \operatorname{arctg}\left(\frac{A_i}{B_i}\right) . \qquad (4.14)$$

В связи с тем, что $\operatorname{arctg}\left(\frac{A_i}{B_i}\right)$ имеет два значения (в интервале от 0 до "2·π"), для выбора правильного решения используется дополнительное условие

$$\varphi_i = \frac{P}{2 \cdot \pi \cdot i} \cdot \arcsin\left(\frac{A_i}{C_i}\right) . \qquad (4.15)$$

Каждая гармоника учитывает некоторую часть полной дисперсии ($\sigma(t)$). Если несколько первых гармоник учитывают значительную часть полной дисперсии ($\sigma(t)$), то дальнейшие расчеты нецелесообразны. Дисперсия за счет i-ой гармоники равна ($C_i^2/2$) для всех гармоник, за исключением последней, для которой она равна (C_i^2). Часть дисперсии, учитываемая i-ой гармоникой, может быть представлена в виде отношения величины ($C_i^2/2$) или (C_i^2) к полной дисперсии (σ_X^2). Поскольку, никакие две гармоники не будут учитывать одну и ту же часть дисперсии ($\sigma(t)$), то дисперсии, учитываемые различными гармониками, можно складывать.

Исключение регулярных циклов

С целью анализа нерегулярных колебаний временного ряда, необходимо, после выделения и анализа регулярных циклов, исключить последние, т.е. вычесть их из исходных данных так, чтобы можно было проанализировать оставшийся временной ряд. *Если период регулярного цикла короче, чем* предполагаемые *периоды нерегулярных колебаний*, то *для исключения регулярного цикла можно использовать только результат наблюдения в одной и той же точке цикла* или *использовать средние из всех результатов наблюдений за полный регулярный цикл*. Так, *для исключения суточного хода температуры* воздуха, *можно использовать только результат наблюдения в конкретный час каждых суток* (например, в полдень) *или только средние суточные температуры. Если период* регулярного *цикла длиннее, чем период нерегулярных колебаний, то результат* каждого наблюдения *может быть выражен как отклонение от среднего* (нормы). Например, если временной ряд состоит из средних месячных температур воздуха, то каждое значение можно заменить разностью между средней месячной температурой и климатической нормой температуры для того же месяца.

В качестве примера, выполним гармонический анализ временного ряда средних месячных температур воздуха в городе Бресте. *Исходный ряд представлен в таблице 4.1.*

Решение

Таблица 4.1 Средние месячные температуры воздуха по метеостанции Брест

Месяц	*t°C*	*Месяц*	*t°C*	*Месяц*	*t°C*
I	*-4,7*	*V*	*13,6*	*IX*	*13,1*
II	*-3,8*	*VI*	*16,9*	*X*	*7,7*
III	*0,4*	*VII*	*18,4*	*XI*	*2,6*
IV	*7,3*	*VIII*	*17,4*	*XII*	*-2,0*

1) В зависимости от (n) и (P) (для всех значений "i" и "t") составляется таблица множителей (A_i ; B_i) (таблица 4.2): $(\frac{2}{n}\cdot(\sin\frac{2\cdot\pi}{P}\cdot i\cdot t))$ и $(\frac{2}{n}\cdot(\cos\frac{2\cdot\pi}{P}\cdot i\cdot t))$. Для последней гармоники (i=n/2) множители определяются как :$(\frac{1}{n}\cdot(\cos\frac{2\cdot\pi}{P}\cdot i\cdot t))$.

2) Значения анализируемого ряда (таблица 4.1) умножаются на соответствующие множители (таблица 4.2) и произведения (в столбцах) суммируются. Суммы (по столбцам) являются коэффициентами Фурье. Далее, по (4.14), определяются амплитуда (C_i) и дисперсия гармонического колебания (i), равная ($\sigma_i = C_i^2 / 2$); результаты расчетов вносятся в таблицу 4.3; под таблицей приведен ход расчета величин (A_i); (B_i); (C_i); (φ_i);($C_i^2 / 2$).

Таблица 4.2 Множители (A_i;B_i) гармонического анализа результатов 12 наблюдений

$$(\frac{2}{n}\cdot\sin(\frac{360^0}{P}\cdot i\cdot t)\ и\ \frac{2}{n}\cdot\cos(\frac{360^0}{P}\cdot i\cdot t))$$

Номер наблюдения	A_1	A_2	A_3	A_4	A_5
1	*2*	*3*	*4*	*5*	*6*
1	*0,0833*	*0,1443*	*0,1667*	*0,1443*	*0,0833*
2	*0,1443*	*0,1443*	*0,0000*	*-0,1443*	*-0,1443*
3	*0,1667*	*0,0000*	*-0,1667*	*0,0000*	*0,1667*

Продолжение таблицы 4.2

4	0,1443	-0,1443	0,0000	0,1443	-0,1443
5	0,0833	-0,1443	0,1667	-0,1443	0,0833
6	0,0000	0,0000	0,0000	0,0000	0,0000
7	-0,0833	0,1443	-0,1667	0,1443	-0,0833
8	-0,1443	0,1443	0,0000	-0,1443	0,1443
9	-0,1667	0,0000	0,1667	0,0000	-0,1667
10	-0,1443	-0,1443	0,0000	0,1443	0,1443
11	-0,0833	-0,1443	-0,1667	-0,1443	-0,0833
12	0,0000	0,0000	0,0000	0,0000	0,0000

→ *Продолжение таблицы 4.2*

Номер наблюдений	B_1	B_2	B_3	B_4	B_5	B_6
1	7	8	9	10	11	12
1	0,1443	0,0833	0,0000	-0,0833	-0,1443	-0,0833
2	-0,0833	-0,0833	-0,1667	-0,0833	0,0833	0,0833
3	0,0000	-0,1667	0,0000	0,1667	0,0000	-0,0833
4	-0,0833	-0,0833	0,1667	-0,0833	-0,0833	0,0833
5	0,1443	0,0833	0,0000	-0,0833	0,1443	-0,0833
6	-0,1667	0,1667	-0,1667	0,1667	-0,1667	0,0833
7	-0,1443	0,0833	0,0000	-0,0833	0,1443	-0,0833
8	-0,0833	-0,0833	0,1667	-0,0833	-0,0833	0,0833
9	0,0000	-0,1667	0,0000	0,1667	0,0000	-0,0833
10	0,0833	-0,0833	-0,1667	0,0833	0,0833	0,0833
11	0,1443	0,0833	0,0000	-0,0833	-0,1443	-0,0833
12	0,1667	0,1667	0,1667	0,1667	0,1667	0,0833

$$A_1=0,0833\cdot(-4,7)+0,1443\cdot(-3,8)+0,1667\cdot 0,4+0,1443\cdot 7,3+0,0833\cdot 13,6+$$
$$+0,0000\cdot 16,9+(-0,0833)\cdot 18,4+(-0,1443)\cdot 17,4+(-0,1667)\cdot 13,1+$$
$$+(-0,1443)\cdot 7,7+(-0,0833)\cdot 2,6+0,0000\cdot(-2,0)=-6,2419;$$

$$B_1=0,1443\cdot(-4,7)+(-0,0833)\cdot(-3,8)+0,0000\cdot 0,4+(-0,0833)\cdot 7,3+$$
$$+(-0,1443)\cdot 13,6+(-0,1667)\cdot 16,9+(-0,1443)\cdot 18,4+(-0,0833)\cdot 17,4+$$
$$+0,0000\cdot 13,1+0,0833\cdot 7,7+0,1443\cdot 2,6+0,1667\cdot(-2,0)=-9,1708;$$

$$C_1=\sqrt{A_1^2+B_1^2}=\sqrt{(-6,2419)^2+(-9,1708)^2}=11,0935\ ;$$

$$\varphi_1 = \frac{12}{2 \cdot 3{,}14 \cdot 1} \cdot \arcsin\left(\frac{-6{,}2419}{11{,}0935}\right) = -1{,}1413 \ ;$$

$$C_1^2 / 2 = 61{,}5335 .$$

Таблица 4.3 Результаты гармонического анализа средних месячных температур воздуха по метеостанции Брест

i	A_i	B_i	C_i	φ_i	$C_i^2/2$
1	-6,2419	-9,1708	11,0936	-1,11413	61,5335
2	-0,5628	0,3417	0,6584	-0,9789	0,2167
3	0,1033	0,3167	0,3332	0,2007	0,0555
4	-0,1587	4,1441	1,1550	-0,0658	0,6670
5	-0,0081	0,0374	0,0382	-0,0820	0,0007
6	--	0,0083	0,0083	0,0000	0,0000

3) Поскольку (S^2=67,85), доля дисперсии, учитываемая первой гармоникой, определяется как

$$\frac{C_1^2}{2 \cdot S^2} = \frac{61{,}5335}{67{,}85} = 0{,}9069 \approx 91\%,$$

второй гармоникой как -

$$\frac{C_2^2}{2 \cdot S^2} = \frac{0{,}2117}{67{,}85} = 0{,}00319 \approx 0{,}3\%,$$

Выводы

Как и следовало ожидать, первая гармоника, описывающая годовой ход температуры воздуха в пункте Брест, учитывает большую часть общей дисперсии. Взятые вместе две первые гармоники описывают 95% суммарного изменения температуры. Вычислять дальнейшие гармоники нет необходимости.

4.2 Выделение и анализ нерегулярных циклов

Остаточные члены временного ряда, после исключения из исходного ряда периодов регулярных колебаний, по-видимому, не объединены какой-либо особой периодичностью, т.е. не представляют собой явно выраженных циклов. Для такого временного ряда обычно характерно несколько типов колебаний:

1) короткопериодные флуктуации настолько малого временного масштаба, что они проходят за половину (меньше половины) периода между, следующими друг за другом, наблюдениями. Такие циклы не мо-

гут быть изучены из-за недостаточной частоты наблюдений; их влияние может быть в значительной степени исключено с помощью сглаживающих, низкочастотных фильтров;

2) медленное, постепенное изменение случайной переменной в течение всего анализируемого периода, называемое *трендом;* тренд никогда не длится бесконечно, а скорее является частью колебаний с периодами, длительность которых сравнима с периодом наблюдений (тренды эффективно исключаются с помощью высокочастотной фильтрации);

3) нерегулярные колебания, характеризующиеся промежуточным временным масштабом.

Ниже *приводятся основные методы выделения и анализа* выявленных *нерегулярных циклов.*

Скользящее сглаживание

Одной из причин получения противоречивых результатов исследований колебаний и связей является пренебрежение статистическими критериями оценки реальности полученных результатов. Часто, вообще, недооцениваются методические аспекты исследования гелиофизических связей и изменений климата. В климатологии широко используются *различные методы сглаживания рядов*, в частности, *метод скользящих перекрывающихся средних*. Суть его заключается в следующем. Если имеется ряд последовательных значений элемента ($x_1,x_2,...,x_n$), то при осреднении по (m) членов, где (m<n), получим ряд

$$\frac{1}{m}\cdot\sum_{i=1}^{m} x_i,\ \frac{1}{m}\cdot\sum_{i=2}^{m+1} x_i,\ \ldots,\frac{1}{m}\cdot\sum_{i=n+1-m}^{n} x_i.$$

Метод скользящих средних представляет едставляет собой фильтр, позволяющий гасить волны коротких колебаний и выделять колебания с большей длиной волны. Достоинством метода является то обстоятельство, что для волн синусоидального характера (если отнести осредненные члены к середине интервала) осредненная фаза не меняется. Вместе с тем, необходимо учитывать, что не все колебания, обнаруженные в таких рядах, реальны. Еще в 1927 году Слуцким Е.Е. было доказано, что сложение случайных причин может порождать волнообразные ряды гармонического характера. Возникновение таких волн связывается с тем, что при скользящем осреднении влияние каждого члена ряда распространяется на (m) - членов скользящего ряда (n), и между ними может возникать корреляция. По этим же соображениям, следует избегать многократного сглаживания. Кроме того, этот метод растягивает и

сглаживает резкие скачки в ряду и, тем самым, может создать представление о циклах с длительностью, равной тройному периоду осреднения, аналогичному при ложном заключении о наличии трехлетнего периода в случайном ряду по числу экстремумов, число которых теоретически равно (2/3).

Другим, часто применяющимся методом исследований колебаний климатических характеристик, является *метод интегральных или интегрально-разностных кривых*. Если обозначить отклонение каждого члена ряда (x_i) через ($d_i = x_i - \bar{x}$), то интегральным называется ряд, члены которого равны -

$$(x_1),\ (x_1 + x_2),\ (x_1 + x_2 + x_3),\ldots,(x_1 + x_2 + \ldots + x_n),$$

а интегрально-разностным - ряд, члены которого равны-

$$(d_1),\ (d_1 + d_2),\ (d_1 + d_2 + d_3),\ldots,(d_1 + d_2 + \ldots + d_n).$$

Интегрирование (суммирование) периодических или циклических колебаний увеличивает амплитуды подобных колебаний пропорционально длине соответствующего периода, облегчает выделение низкочастотных колебаний. Метод интегрально-разностных кривых может быть рекомендован, когда важны кумулятивные свойства ряда, например, при интерпретации рядов осадков, стока, суммарного испарения и т.д. Однако, как и в случае скользящего осреднения, при этих методах, вследствие увеличения внутрирядной связности суммируемых последовательностей (x_i), (d_i), может возникать ложная цикличность. Кроме того, даже при исходном бессвязном ряде в середине интегрально-разностного ряда, дисперсия его членов будет наибольшей и наложение случайных ошибок может увеличивать амплитуду случайных колебаний в этой части ряда.

Выделение тренда

Поскольку тренды используются для оценки тенденции будущих значений ряда, исключение тренда - одна из основных задач анализа нерегулярных колебаний. Обычно, тренд выделяется по методу наименьших квадратов способом скользящей средней или по определенной, характерной для данного ряда аналитической формуле, т.е. находятся средние ($\bar{x}_i$) и соответствующие отклонения аномалий (d) от среднего ($\bar{x}$) или уровня аналитической кривой (x(t)). Алгебраически можно показать, что при переходе от уровней к их разностям (d) исключается влияние общей тенденции на колеблемость. Если тренд ряда может быть представлен пря-

мой ($\bar{y}_t = \alpha + b \cdot t$), то, обозначая последовательные моменты времени через ($t_1, t_2, \ldots, t_n$), получим для:

$$\left.\begin{array}{ll} t = 1 & - \quad y_1 = \alpha + b + d_{y1}; \\ t = 2 & - \quad y_2 = \alpha + 2 \cdot b + d_{y2}; \\ t = 3 & - \quad y_3 = \alpha + 3 \cdot b + d_{y3}; \\ t = 4 & - \quad y_4 = \alpha + 4 \cdot b + d_{y4} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \end{array}\right\}. \qquad (4.16)$$

Найдем первые разности -

$$\left.\begin{array}{l} \Delta'_1 = y_2 - y_1 = b + (d_{y2} - d_{y1}); \\ \Delta'_2 = y_3 - y_2 = b + (d_{y3} - d_{y2}); \\ \Delta'_3 = y_4 - y_3 = b + (d_{y4} - d_{y3}) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array}\right\}. \qquad (4.17)$$

Так как во всех разностях присутствует одна и та же константа (b), то очевидно, что колебания рассчитанных разностей (Δ) зависят только от (d_y), т.е. влияние общей тенденции (тренда) механически исключается. Для случая, когда исходные данные изменяются по параболе 2-го порядка ($\bar{y}_t = \alpha + b + c \cdot t^2$), получим, при ($t_1 = y_1$), ($t_1 = \alpha + b + c + d_{y1}$), при ($t_2 = y_2$), ($t_2 = \alpha + 2 \cdot b + 4 \cdot c + d_{y2}$) и т.д. Находим первые разности

$$\left.\begin{array}{l} \Delta'_1 = y_2 - y_1 = b + 3 \cdot c + (d_{y2} - d_{y1}); \\ \Delta'_2 = y_3 - y_2 = b + 5 \cdot c + (d_{y3} - d_{y2}); \\ \Delta'_3 = y_4 - y_3 = b + 7 \cdot c + (d_{y4} - d_{y3}) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array}\right\}. \qquad (4.18)$$

Как видно, первые разности содержат, кроме постоянного (b), еще и переменные слагаемые (3c, 5c, 7c,...). Чтобы добиться устранения влияния общей тенденции, на основе первых разностей (Δ'), рассчитаем вторые разности (Δ'')

$$\left.\begin{array}{l} \Delta''_1 = \Delta'_2 - \Delta'_1 = 2 \cdot c + (d_{y3} - 2 \cdot d_{y2} + d_{y1}); \\ \Delta''_2 = \Delta'_3 - \Delta'_2 = 2 \cdot c + (d_{y4} - 2 \cdot d_{y3} + d_{y2}) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array}\right\}. \qquad (4.19)$$

Как видно, колебания (Δ') определяются только величинами (d_y), так как (2c) - величина постоянная во всех вторых разностях. Таким образом, для дальнейших исследований следует оперировать не с исходными рядами, а с их разностями различного порядка : при линейном тренде - с первыми разностями, при параболическом - с вторыми разностями, при аппроксимации зависимости параболой (n)-го порядка или полиномами - с разностями (n)-го порядка. Простейший прием экстраполяции динамики рядов или их тренда следующий:

1) проводится обобщающая линия, отражающая тенденцию ряда на основе зрительного впечатления о расположении фактических точек;

2) определяется постоянство и рассчитывается средний абсолютный прирост за последние годы - $\left(\Delta = \frac{y_n - y_1}{n-1}\right)$. Далее, он последовательно прибавляется к оценке тренда на последний срок столько раз, на сколько периодов экстраполируется ряд. Можно также уменьшить последний член тренда на коэффициент роста в степени , пропорциональной периоду экстраполяции;

3) устанавливаются соотношения изменений показателей динамики сравниваемых рядов на основе корреляции между ними.

На рисунке 4.1 показано *выделение тренда урожайности яровой пшеницы.*

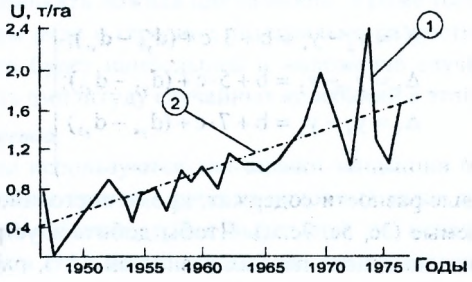


Рисунок 4.1 Колебания урожайности яровой пшеницы (1) в Поволжье и линия тренда (2).

Автокорреляционный анализ

Автокорреляция означает корреляцию параметра с самим собой. Коэффициенты автокорреляции являются коэффициентами линейной корре-

ляции между значениями временного ряда в данный момент времени и его же значениями в последующий момент времени. Для практических целей коэффициенты автокорреляции определяются по формуле

$$r_x(\tau) = \frac{\sum_{i=0}^{n-\tau-1} (x_i - \bar{x}) \cdot (x_{i+\tau} - \bar{x})}{(n - \tau - 1) \cdot \sigma_x^2}, \qquad (4.20)$$

где τ - запаздывание; $\bar{x}$ - среднее значение ряда; σ_x^2 - его дисперсия. Если запаздывание мало, то в природных процессах коэффициенты автокорреляции обычно положительны, поскольку для природных процессов характерна устойчивость. По мере увеличения запаздывания, коэффициент автокорреляции уменьшается и может стать отрицательным. *Зависимость* между коэффициентом автокорреляции и периодом запаздывания *носит название автокорреляционной функции* и графически представляется автокоррелограммой (рисунок 4.2).

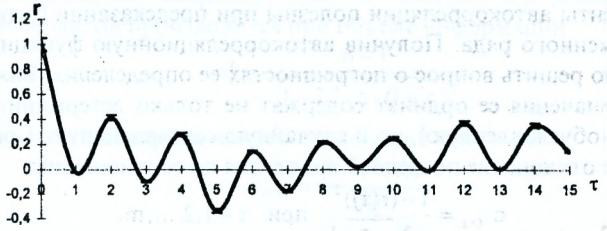


Рисунок 4.2 Автокорреляционная функция годовых величин испаряемости - максимально возможного испарения (по данным метеостанции Брест).

В зависимости от динамики развития исследуемого процесса, корреляционные функции имеют различную степень внутрирядной связности и, как следствие, - различный вид и форму аппроксимации. Например, корреляционная функция чисто случайного ряда имеет один всплеск при сдвиге (τ=0), при остальных сдвигах (r(τ)) - она равняется нулю. Достаточно простой вид имеет корреляционная функция процессов по типу простой цепи Маркова

$$r(\tau) = [r(\tau)]^{\tau} \approx r^{-\alpha(\tau)}. \qquad (4.21)$$

Корреляционная функция (r(τ)) случайного ряда (периодом "m"-лет), сглаженного скользящим образом, определяется равенством

$$r(\tau) = \frac{m - \tau}{m}, \qquad (4.22)$$

а ряда из суммы синусоид ($x(t) = \sum \alpha_k \cdot \sin(\omega \cdot t)$) - соотношением

$$r(\tau) = \sum b_k \cdot \cos(\omega_k \cdot \tau) \ . \qquad (4.23)$$

Корреляционные функции, несмотря на их универсальность, часто не поддаются интерпретации. Кроме того, непосредственно по ним трудно установить сопряженность колебаний с тем или иным периодом. В связи с этим, при анализе структуры случайных процессов, используются спектральные характеристики, позволяющие оценить распределение амплитуд (энергии) колебаний в частотной области. То обстоятельство, что коэффициенты автокорреляции отличны от нуля, означает, что многие из статистических оценок неприменимы, поскольку при их выводе предполагается, что данные временного ряда независимы. С другой стороны, коэффициенты автокорреляции полезны при предсказании будущих значений временного ряда. Получив автокорреляционную функцию, затем, необходимо решить вопрос о погрешностях ее определения. Важно отметить, что значения ее ординат содержат не только детерминированную (причинно обусловленную), но и случайную составляющую. Среднеквадратическое отклонение последней находится по соотношению

$$\sigma_{r(x)} = \frac{1 - (r(\tau))^2}{\sqrt{N - \tau - 1}}, \text{ при } \tau = 1,2,\ldots,m. \qquad (4.24)$$

В случае, когда величины (r(τ)) малы, отклонения коэффициентов автокорреляции от их истинных значений распределены по нормальному закону. Тогда, доверительные границы (Н-нижняя, В-верхняя) для оценки автокорреляционной функции, с доверительной вероятностью 95%, могут быть найдены как:

$$r_н(\tau) = r(\tau) - \frac{2}{\sqrt{N - \tau}} \cdot (1 - r^2(\tau)) \ ; \qquad (4.25)$$

$$r_в(\tau) = r(\tau) + \frac{2}{\sqrt{N - \tau}} \cdot (1 - r^2(\tau)) \ . \qquad (4.26)$$

Необходимо отметить, что данные зависимости получены в предположении нормальности статистического распределения параметров природных процессов. Тщательное изучение вопроса показывает, что отличия "от нормальности" невелики и, как правило, в целом не нарушают стационарности рядов. Вместе с тем, необходимо отметить, что оценки

ошибок по соотношению (4.24) свидетельствуют о наличии достоверных связей только между смежными членами ряда (сдвиг τ=1). Оценки более дальних связей в значительной степени отражают лишь случайные флуктуации выборочных данных, связанные с ограниченностью длительности имеющихся рядов, т.е. малодостоверны. Это исключает возможность использовать их в целях экстраполяции (прогноза на будущее). Вместе с тем, эмпирические автокорреляционные функции, при τ>1, могут рассматриваться как показатели цикличности многолетних колебаний, особенно при использовании массового материала. Здесь следует быть осторожным, особенно при интерпретации периодов, являющихся совпадением по фазе двух циклов (например, увеличенные ординаты, при τ=25, могут быть следствием совпадения циклов меньшей продолжительности: τ=8 и τ=17 лет). Если проверка на достоверность производится относительно простой цепи Маркова, т.е. проверяется гипотеза об отсутствии внутрирядных связей, например, при (τ>3 лет), то построение доверительных интервалов производится при объеме информации

$$n' = \frac{n-1}{\sqrt{1-2{,}2\cdot r_1\cdot(1+r_1)^2}}, \qquad (4.27)$$

а оценка среднего квадратического отклонения как

$$\sigma_{r(\tau)} = \frac{(1-r^2(\tau))}{\sqrt{n'}}. \qquad (4.28)$$

В общем случае, независимо от вида последовательности $(x_1,x_2,\ldots,x_n)$, построение доверительных интервалов (r(τ)) осуществляется по формуле

$$r_{\alpha\tau} = t_\alpha \cdot \sigma_{r(\tau)} = 0, \qquad (4.29)$$

где $r_{\alpha\tau}$ - доверительные границы (r(τ)) для данного (τ); t_α - критерий Стьюдента с (n-τ-1) - степенями свободы при данном уровне значимости (α); $\sigma_{r(\tau)}=0$ - среднее квадратическое отклонение выборочных значений (r(τ)), определяемое по системе зависимостей (4.27) - (4.28).

Взаимная корреляционная функция

Аналогично автокорреляционной функции *понятие взаимной корреляционной функции.* Эта функция характеризует зависимость значений одного временного ряда в данный момент от значений другого временного ряда в другой момент времени. Если имеются аномалии двух рядов-

$$\left.\begin{matrix}\Delta x_1,\Delta x_2,\ldots,\Delta x_\tau,\ldots,\Delta x_N\\ \Delta y_1,\Delta y_2,\ldots,\Delta y_\tau,\ldots,\Delta y_N\end{matrix}\right\}, \qquad (4.30)$$

то оценку функции взаимной корреляции можно представить в виде

$$r_{xy} = \begin{cases} \dfrac{1}{(N-\tau)\cdot\sigma_x\cdot\sigma_y}\cdot\sum\limits_{i=1}^{N-\tau}\Delta x_i\cdot\Delta y_{i+\tau}, & \text{при } \tau \geq 0, \\ \dfrac{1}{(N-|\tau|)\cdot\sigma_x\cdot\sigma_y}\cdot\sum\limits_{i=1}^{N+\tau}\Delta x_i\cdot\Delta y_{i+|\tau|}, & \text{при } \tau \leq 0. \end{cases} \qquad (4.31)$$

Первая из формул описывает зависимость ($y_{i+\tau}$) от (x_i), вторая, наоборот, ($x_{i+|\tau|}$) от (y_i). В противоположность автокорреляционной функции, взаимная корреляционная функция необязательно имеет максимум при (τ=0). Максимальное значение ($r_{xy}(\tau)$) указывает на сдвиг между анализируемыми рядами, при котором эти ряды наиболее тесно связаны. И автокорреляционная функция, и функция взаимной корреляции могут быть использованы в прогнозировании значения данной величины в зависимости от значения этой или другой величины в предшествующий момент времени -

$$x(t+\tau) = r(\tau)\cdot x(\tau) + \omega(\tau)\cdot\sqrt{1-r^2(\tau)}\ , \qquad (4.32)$$

здесь r(τ) - коэффициент корреляции между (x(t+τ)) и (x(t)); $\omega(\tau)$ - некоторая некоррелированная с (x(τ)) функция, которая обычно определяет ошибку прогноза. В результате, по (4.32) получаем, так называемый, статистический метод прогноза по одной точке предыстории. Прогнозируемая величина (x(t+τ)), при этом, будет всегда близка к последнему значению (x(t)). Отметим, что включение в прогностическое соотношение переменных за более ранние сроки не всегда приводит к положительным результатам, так как чаще всего оказывается, что прогностические уравнения типа (4.32) нарушаются при проверке на других периодах. Многие свойства взаимно корреляционных функций идентичны свойствам автокорреляционных функций, но имеется и различие. Так, свойство четности у взаимных корреляционных функций заменяется симметричностью нулевого сдвига, т.е. ($r_{xy}(\tau) \neq r_{xy}(-\tau)$). Другое отличие в свойствах определяется тем обстоятельством, что фазы колебаний взаимодействующих процессов далеко не всегда совпадают. Из этого следует, что гармоники с одинаковыми фазами и периодами колебаний на графике взаимной корреляции будут образовывать всплески, соответствующие максимальным амплитудам циклов при нулевой разности фаз. Различие частот, наоборот, приводит к взаимной корреляции, близкой к нулевым значениям.

Автоспектральный анализ

С помощью автокорреляционной функции можно описать внутреннюю структуру процесса, определяемую доминирующими компонентами, во временной области. В тех случаях, когда процесс складывается из составляющих разных временных масштабов (явление, широко распространенное в гидрометеорологии), знание структуры процесса во временной области часто оказывается недостаточным. При решении многих природоведческих (природопользовательских) задач необходимо также знать распределение интенсивности процесса между составляющими различных временных масштабов, т.е. необходимо выполнить описание случайного процесса в частной области. Для этой цели используется спектральное разложение процесса. Спектр временного ряда аналогичен оптическому спектру, который показывает вклад различных длин волн или частот в энергию заданного источника света. Физический смысл спектра временного ряда состоит в том, что он показывает вклад колебаний с различными частотами в полную дисперсию временного ряда и, следовательно, может быть получен с помощью метода гармонического анализа. Если вычислить все (n/2) - гармоник и построить график полуквадрата их амплитуды, как функцию частоты, то разброс точек окажется очень большим. При этом, если построить спектры для двух отдельных частей одного стационарного временного ряда, то отдельные точки спектра будут занимать совершенно различное положение, поскольку отдельные резкие пики в нерегулярных колебаниях временного ряда носят чисто случайный характер. Поэтому, с помощью спектрального анализа обычно не пытаются определить амплитуды отдельных гармоник. Его целью является нахождение сглаженного спектра, который остается одинаковым для различных частей одного и того же временного ряда. При этом, рассматривается не спектр данного короткого временного ряда, а спектр бесконечного длинного временного ряда, из которого данный временной ряд представляет собой короткую случайную выборку. Для длинного ряда можно определить сглаженный спектр, использовав соответствующую методику предельного перехода. Задачей практического спектрального анализа является оценка этого сглаженного спектра на основании данного короткого ряда. Спектральное разложение отображает стационарную случайную функцию в виде разложения на периодические колебания раз-

личных частот (ωi). Если имеется какой-либо квазигармонический процесс, представляемый набором гармоник

$$x(t) = \bar{x} + \sum_{i=1}^{n} (A_i \cdot \cos\omega t_i + B_i \cdot \sin\omega t_i) \ , \qquad (4.33)$$

то его спектром называется функция

$$C_i = \sqrt{A_i^2 + B_i^2}, \qquad (4.34)$$

описывающая распределение амплитуд гармоник по различным частотам. В отличие от простых колебательных движений, стационарный случайный процесс описывается каноническими разложениями случайных процессов типа

$$x(t) = \bar{x}(t) + \sum \vartheta_i \cdot \varphi_i(t) \ , \qquad (4.35)$$

где $\bar{x}(t)$ - математическое ожидание случайной функции; $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$ - -некоррелированные случайные величины с математическим ожиданием, равным нулю. Плотность распределения дисперсий по частотам непрерывного спектра называется спектральной плотностью дисперсии или спектральной плотностью стационарной случайной функции (S(x)). Таким образом, дисперсия будет представлена в виде

$$D_x = \int_0^{\infty} S_x(\omega) \cdot d\omega \ . \qquad (4.36)$$

Иногда, при решении практических задач, вместо спектральной плотности (S(x)) пользуются нормированной спектральной плотностью ($\Psi_x(\omega)$), связанной преобразованиями Фурье:

$$r_x(\tau) = \int \Psi_x(\omega) \cdot \cos(\omega \cdot \tau) \cdot d\omega \ ; \qquad (4.37)$$

$$\Psi_x(\omega) = \frac{1}{\pi} \cdot \int r_x(\tau) \cdot \cos(\omega \cdot \tau) \cdot d\tau \ . \qquad (4.38)$$

Если положить (τ=0) и учесть, что ($r_x(\tau=0)$=1), получим $(\int_0^{\infty} \Psi_x(\omega) \cdot d\omega = 1)$, т.е. полную площадь, ограниченную графиком нормированной плотности, равной единице. Имея аналитическое выражение корреляционной функции, можно легко определить вид спектральной функции. Так спектральная плотность процесса (по типу простой цепи Маркова), корреля-

ционная функция которой описывается формулой $\left[r(\tau) = r^{-\alpha(\tau)}\right]$, с учетом выражения (4.38), будет определяться как

$$\Psi(\omega) = \frac{a}{\pi \cdot \omega^2} + d^2 . \qquad (4.39)$$

Если аналитическое выражение корреляционной функции неизвестно или аналитическое описание носит сложный характер, то аппроксимация подинтегральной функции осуществляется кривой с соответствующим переходом от точного значения ее площади к приближенному значению, как суммы площадей конечного числа трапеций

$$\Psi(\omega) \approx \frac{2}{\pi} \cdot \sum_{i=3}^{m-1} r_i \cdot t_i \cdot \frac{\sin(\omega \cdot t_i)}{\omega \cdot t_i} \cdot \frac{\sin(\omega \cdot \Delta_i)}{\omega \cdot \Delta_i}, \qquad (4.40)$$

где $r_i = r_\tau - r_{i-1}$ - разности ординат соседних переломных точек ломаной кривой функции; $\Delta_i = (\tau_i - \tau_{i-1})/2$ - полуразности абсцисс соседних точек; $t_i = \tau_i + \Delta_i$ - абсциссы середин между этими точками; $\omega = 2\pi/T$ - круговая частота; m - максимальный сдвиг (t_{max}). Расчет спектральной функции ведется по выражению

$$\Psi_i = \frac{r}{m} + \frac{2}{m} \cdot \sum_{\tau=1}^{m-1} \left(r_\tau \cdot \cos\left(\frac{360}{2 \cdot m} \cdot i \cdot \tau\right)\right) + (-1)^n \cdot \frac{r_m}{m}, \qquad (4.41)$$

где i - номер гармоники (i=1,2,..., n); m - максимальный сдвиг; r_0 и r_m - -значения (r(τ)), при (τ=0) и (τ=m). Значения (Ψ_0) и (Ψ_m) нужно уменьшить в 2 раза. Период колебания и номер гармоники связаны соотношением

$$T = \frac{2 \cdot m}{i} . \qquad (4.42)$$

Отсюда период первой гармоники принимается равным (2·m) , второй - (m), третьей (2·m/3), и т. д. Приближенная оценка доверительной границы ($I_в$) когерентности, при уровне вероятности (P), производится по формуле Гудмена

$$I_в(H(\omega)) = \sqrt{1 - P^{1/\ell - 1}}, \qquad (4.43)$$

в которой $\ell = \frac{2 \cdot N - m/2}{m}$ - число степеней свободы (N) данных реализаций; m - число запаздываний. Значения когерентности при 1% и 5%-ном уровнях значимости для различного числа степеней свободы приведены в таблице 4.4.

Таблица 4.4 Доверительные границы когерентности (H(ω))

Уровень значимости, %	Степени свободы			
	4	10	20	40
1	0,89	0,63	0,46	0,33
5	0,80	0,53	0,38	0,27

Из табличных данных следует, что, при числе степеней свободы $\ell = 20$, когерентность, равная 0,38 и больше, будет недостоверна в одном случае из 20. При использовании спектрального анализа для решения каждой конкретной задачи необходимо исследовать соотношения между временным шагом выборки (Δt), длиной выборки (N), максимальным сдвигом корреляционной функции (m_{max}) для расчета спектра, числом степеней свободы ($\ell = 2 \cdot N \neq m$) и нормированной стандартной ошибкой ($\sigma_{S(\omega)}$) оценок функции спектральной плотности. С одной стороны, число сдвигов (m) должно быть мало по сравнению с длиной выборки, число степеней свободы (ℓ) - возможно большим. Это сохранит определенную степень статистической надежности. С другой стороны, число сдвигов должно быть достаточно большим, чтобы получить большее разрешение по полосе частот; при этом, надежность статистических оценок в пределах частотной полосы уменьшается. Обычно, выбор максимального временного сдвига (τ_{max}=m) производится, исходя из возможной точности расчетов спектральной плотности, в частности, по формулам

$$\tau_{2\%} = \frac{2 \cdot \pi \cdot n}{50}; \quad \tau_{5\%} = \frac{2 \cdot \pi \cdot n}{120}; \quad \tau_{10\%} = \frac{2 \cdot \pi \cdot n}{10}, \qquad (4.44)$$

где $\tau_{2\%}$, $\tau_{5\%}$, $\tau_{10\%}$ - значения максимального сдвига при допустимой погрешности расчета - (S(ω)), соответственно, 2; 5; 10%. Большинство авторов, при 50...100-летних рядах наблюдений, принимают - τ_{max}=10...30 лет. Спектральные функции, рассчитанные по выборочным данным, будут отличаться от спектра генеральной совокупности. Как и в случае оценок одномерных распределений, при оценке значимости спектра используется нуль-гипотеза. Нуль-гипотеза заключается в предположении отсутствия гармонических колебаний в спектре исследуемого ряда на фоне спектра реализации "белого" (горизонтальная линия - нулевой континуум) или "красного" шума (убывающая экспоненциальная линия). Предполагается, что исходная выборка не случайна, а ее значения распределены по нормальному закону. Гипотеза проверяется в результате сравнения ($S_x(\omega)$) со значениями (S(ω)) заданной обеспеченности, принимаемыми за

границы доверительного интервала ($I_\beta(S\omega)$). В этом случае, спектральные оценки полагаются приблизительно распределенными по значению (χ^2) и нормированными на число степеней свободы (ℓ)

$$S(\omega) = \frac{\chi^2}{\ell}, \tag{4.45}$$

где ℓ - число степеней свободы, определяемое как

$$\ell = \frac{2 \cdot N - 0{,}5 \cdot \tau_m}{\tau_m}; \tag{4.46}$$

N - объем выборки, используемый для оценки спектра; τ_m - максимальный сдвиг на коррелограмме. Тогда для отыскания доверительного интервала ($I_\beta(S_\omega)$) справедливо равенство

$$I_\beta(S_\omega) = \frac{\bar{S}_x(\omega) \cdot \chi^2}{\ell}, \tag{4.47}$$

в котором $S_x(\omega)$ - средний уровень спектральной плотности (приравниваемый к "белому шуму"), вычисляемый в интервале значений автокорреляционных функций в пределах сдвигов (запаздываний от 1 до "m"). Выход пиков на спектрограмме за границы заданных ($I_\beta(S_\omega)$) будет свидетельствовать о достоверности выявляемых частот колебаний. Вероятностные точки распределения (χ^2 / ℓ) даны в таблице 4.5.

Если имеются только таблицы (χ^2), а не (χ^2 / ℓ), то при исследуемом уровне значимости (α) получаются отысканием значения (χ^2_α), которые соответствуют значениям (ℓ), с последующим делением на (ℓ). При значительном вкладе в исследуемый процесс "красного шума" марковского процесса первого порядка, средний уровень спектральной плотности (линейный в силу стационарности случайного процесса) превращается в нелинейный, убывающий к низким частотам. Оценка континуума "красного шума" осуществляется с использованием выражения

$$S_k(\omega) = \bar{S}_x(\omega) \cdot \frac{1 - r_x^2(\tau_1)}{1 + r_x^2(\tau_1) - 2 \cdot r_x(\tau_1) \cdot \cos\frac{\pi \cdot k}{\tau_m}}, \tag{4.48}$$

которое содержит значения автокорреляционной функции единичного (τ_1) и максимального (τ_m) сдвигов, превышающих, при (τ_i), значение ($r(\tau)$), равное 0,4.

Таблица 4.5 Вероятностные (по Халду) точки распределения (χ^2 / ℓ)

Степени свободы, ℓ	Вероятность в процентах				
	1	*5**	*95**	*99*	*99,9*
2	0,010	0,052	3,000	4,605	6,908
5	0,111	0,229	2,214	3,017	4,103
10	0,256	0,394	1,831	2,321	2,959
20	0,413	0,543	1,570	1,877	2,266
30	0,498	0,616	1,459	1,696	1,990
50	0,594	0,695	1,350	1,523	1,733
60	0,625	0,720	1,318	1,473	1,660
80	0,669	0,755	1,274	1,404	1,560
100	0,701	0,699	1,243	1,358	1,494
200	0,782	0,841	1,170	1,247	1,338
400	0,843	0,887	1,119	1,172	1,238
1000	0,899	0,928	1,075	1,107	1,144

Примечание: *) - вероятность, эквивалентная 95%-ой точке значимости для одностороннего критерия спектрального минимума.

Сглаженный спектр, как указывалось, может быть получен методом гармонического анализа с последующим сглаживанием амплитуд всех отдельных гармоник, посредством какой-либо алгебраической процедуры осреднения. Этот метод громоздок и в настоящее время реализуется на ЭВМ с использованием, так называемого, алгоритма быстрого преобразования Фурье. Винером и Хинчином отмечалась возможность вычисления спектра через корреляционные функции; Блэкман и Тьюки реализовали эту идею практически. Сегодня этот метод, практически, вытеснен другими методами, но процедура Блэкмана-Тьюки для небольших запаздываний сохраняет свою актуальность. Ниже *дается алгоритм вычисления спектра* с использованием процедуры Блэкмана-Тьюки:

1) если среднее временного ряда (x(i)), состоящего из (n) - членов, не равно нулю, то его вычисляют и вычитают из всех значений ряда;

2) для (m+1) - значений индекса (τ) вычисляются значения автокорреляционной функции -

$$r_x(\tau) = \frac{\sum_{i=0}^{n-\tau-1} x_i \cdot x_{i+\tau}}{(n-\tau-1)\cdot \sigma_x^2}, \text{ при } \tau = 0,1,\ldots,m; \qquad (4.49)$$

3) выбирается корреляционное окно -

а) Хеннинга

$$V_m^{(1)}(\tau) = \frac{1}{2} \cdot (1 + \cos\frac{\pi \cdot \tau}{m}) \; ; \qquad (4.50)$$

б) Хемминга

$$V_m^{(2)}(\tau) = 0{,}54 + 0{,}46 \cdot (1 + \cos\frac{\pi \cdot \tau}{m}) \; ; \qquad (4.51)$$

в) Парзена

$$V_m^{(3)}(\tau) = \begin{cases} (1 - 6 \cdot \frac{\tau}{m})^2 \cdot (1 - \frac{\tau}{m}), & \text{при } \tau < \frac{m}{2} \; ; \\ 2 \cdot (1 - \frac{\tau}{m})^3, & \text{при } \tau > \frac{m}{2} \; , \end{cases} \qquad (4.52)$$

и с помощью одного из них (окон) сглаживается автокорреляционная функция

$$\bar{r}_x(\tau) = r_x(\tau) \cdot V_m^{(\cdot)}(\tau) \; ; \qquad (4.53)$$

4) посредством интегрирования (методом трапеции) для различных частот вычисляются спектральные оценки

$$S_x(k) = 2 \cdot \left(1 + 2 \cdot \sum_{\tau=1}^{m-1} \bar{r}_x(\tau) \cdot \cos\frac{\pi \cdot \tau \cdot k}{m} \right), \text{ при } 0 \le k \le m \; ; \qquad (4.54)$$

график выборочных спектральных оценок целесообразно строить в логарифмическом масштабе, поскольку построение доверительного интервала для спектра сводится к откладыванию около выборочной спектральной оценки одного и того же интервала для всех частот; доверительный интервал для логарифма спектра рассчитывается по зависимостям

$$\lg S_x(k) + \lg\frac{\vartheta}{x_\vartheta(1-\alpha/2)}; \qquad \lg S_x(k) + \lg\frac{\vartheta}{x_\vartheta(\alpha/2)}, \qquad (4.55)$$

в которых ϑ - число степеней свободы (для окон: Хеннинга - $\vartheta = 2{,}667 \cdot \frac{n}{m}$; Парзена - $\vartheta = 3{,}71 \cdot \frac{n}{m}$).

Фильтрация временных рядов

Спектральный состав временного ряда можно описать с помощью его статистических характеристик. Если исходный временной ряд содержит

некоторые частоты или периоды, которые в данный момент не представляют интереса для исследователя, амплитуда этих волн может быть уменьшена с помощью статистической фильтрации. При этом, изменяется спектр исходного временного ряда. Формой фильтрации, создающей временной ряд, в котором спектральные компоненты с высокой частотой уменьшены, является сглаживание. Такой тип фильтра называют фильтром пропускания низких частот, так как сглаживание слабо влияет на волны с низкой частотой (длиннопериодические волны). Величина рассматриваемой характеристики в сглаженном временном ряду является оценкой величины во временном ряду, в котором нежелательные высокие частоты отсутствовали бы. Можно отфильтровать низкие частоты, оставив в ряду только высокочастотные колебания. Этот тип фильтра называется фильтром пропускания высоких частот. Можно отфильтровать как низкие, так и высокие частоты, оставив в получающемся ряду только средние. Такой тип фильтра носит название фильтра пропускания средних частот. Операция "фильтрации" в гидрометеорологии и природопользовании выполняется обычно численными методами, при следующем математическом описании

$$\tilde{x}_i = \sum_{m=-M}^{M} h_m \cdot x_{i+m}, \qquad (4.56)$$

где x - исходный ряд; $\tilde{x}$ - отфильтрованный ряд; h_m - весовые коэффициенты фильтра; M - параметр, определяющий число весов, которое для симметричных фильтров равно (2·M+1). Изменение спектра исходного ряда в результате фильтрации может быть оценено с помощью частотной характеристики фильтра, которая представляет собой отношение амплитуды колебания данной частоты после фильтрации к амплитуде до фильтрации. Это отношение меняется вместе с частотой. Частотная характеристика любой дискретной симметричной сглаживания или фильтрующей функции выражается следующим равенством

$$R(f) = h_0 + 2 \cdot \sum_{m=1}^{n} h_m \cdot \cos(2 \cdot \pi \cdot m \cdot f \cdot \Delta t), \qquad (4.57)$$

где R(f) - частотная характеристика; f - частота; h_m - (m)-ый вес, причем число (m) отличается от главного весового коэффициента (h_0); Δt - интервал времени между последовательными наблюдениями во временном ряде. Весовые функции и частотные характеристики наиболее употребительных низкочастотных фильтров приведены ниже:

1) *скользящее равновесное среднее*; весовая функция этого фильтра

$$h_m = \begin{cases} \dfrac{1}{2 \cdot M + 1}, & \text{при } |m| \le M; \\ 0, & \text{при } |m| > M, \end{cases} \tag{4.58}$$

где - (2·M+1) - интервал сглаживания. Частотная характеристика фильтра

$$R(f) = \frac{\sin(\pi \cdot f \cdot (2 \cdot M + 1))}{\pi \cdot f \cdot (2 \cdot M + 1)}. \tag{4.59}$$

Простое осреднение значений величины в пределах соседних отрезков времени есть частный случай скользящего равновесного осреднения. Скольжение осуществляется здесь "скачками", длина которых равна длине интервала осреднения. Можно сказать, что, несмотря на широкое распространение, простое осреднение является едва ли не худшим видом сглаживания. Каждое значение исходного ряда, в отличие от других фильтров, используется лишь однажды. Кроме того, фазы ряда гармоник меняются на 180^0, при сравнительно плохой фильтрации этих гармоник. Поэтому, простое осреднение дает искаженное представление о процессе;

2) *фильтр Барплетта* (треугольный)

$$h_m = \begin{cases} \dfrac{2}{2 \cdot M + 1} - \dfrac{4 \cdot |m|}{(2 \cdot M + 1)^2}, & \text{при } |m| \le M; \\ 0, & \text{при } |m| > M; \end{cases} \tag{4.60}$$

$$R(f) = \left[\frac{2 \cdot \sin\left(\dfrac{\pi \cdot f \cdot (2 \cdot M + 1)}{2}\right)}{\pi \cdot f \cdot (2 \cdot M + 1)} \right]^2; \tag{4.61}$$

характеристика (4.61), в отличие от (4.59), есть неотрицательная функция;

3) *фильтр Тьюки*

$$h_m = \begin{cases} \dfrac{1 + \cos\left(\dfrac{2 \cdot \pi \cdot m}{2 \cdot M + 1}\right)}{2 \cdot M + 1}, & \text{при } |m| \le M; \\ 0, & \text{при } |m| > M; \end{cases} \tag{4.62}$$

$$R(f) = \frac{\sin\left(\frac{\pi \cdot f \cdot (2 \cdot M + 1)}{2}\right)}{\pi \cdot f \cdot (2 \cdot M + 1)} \cdot \frac{1}{1 - (f \cdot (2 \cdot M + 1))^2}. \qquad (4.63)$$

Этот фильтр характерен тем, что боковые лепестки его частотной характеристики весьма малы. К недостаткам фильтра Тьюки, так же как и фильтра Бартлетта, следует отнести сравнительно медленное затухание его частотной характеристики. Располагая характеристиками низкочастотных фильтров, легко осуществить выделение гармоник с высокими частотами. Частотная характеристика высокочастотного фильтра ($R_H(f)$) просто выражается через частотную характеристику сглаживающего фильтра

$$R_н(f) = 1 + R_L(f). \qquad (4.64)$$

Значения весовой функции высокочастотного фильтра равны по величине и обратные по знаку значениям весовой функции низкочастотного фильтра, за исключением центрального веса (h_0), который дополняет до единицы величину центрального веса низкочастотного фильтра. На рисунке 4.3 приведена нормированная разностная интегральная кривая годового стока реки Неман в створе Гродно.

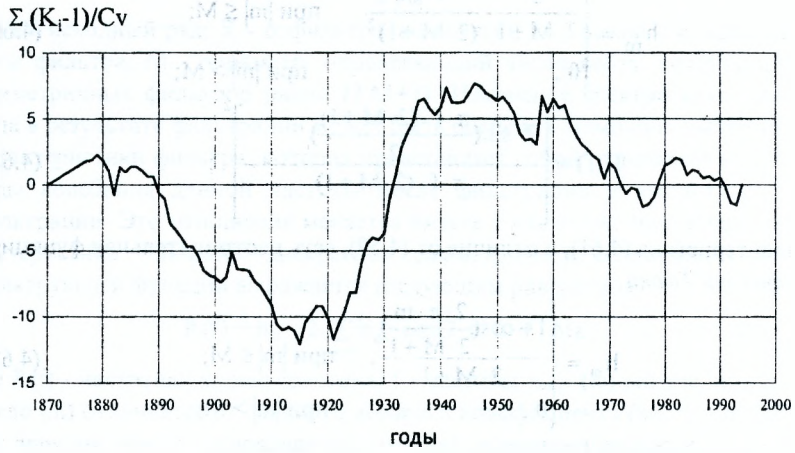


Рисунок 4.3 Нормированная разностная интегральная кривая годового стока реки Неман в створе Гродно.

4.3 Понятие о статистических методах предсказания природных процессов

Учет структурных особенностей, отмеченных выше, при диагнозе в долгосрочном прогнозе, возможен *при следующих предпосылках. Во-первых*, чтобы статистические характеристики рядов были устойчивы; *во-вторых*, необходимо наличие определенной регулярности в проявлении цикличности, под которой обычно понимаются колебания связанных величин различной степени регулярности при условии устойчивости математических ожиданий параметров этих колебаний. Первое условие в природных процессах выполняется весьма приближенно, что касается второго, то оно накладывает следующие ограничения - наличие в квазипериодических колебаниях закономерного чередования высоких и низких значений, а в квазислучайных (стахостических процессах) - хотя бы группировки высоких и низких значений, отличной от случайной. Чтобы выявить наличие случайности (неслучайности) колебаний, необходимо использовать критерии систематического ряда, например, критерии экстремумов, повышений или понижений ряда (критерии Б.П. Вайнберга, М.А. Омшанского, О.А. Дроздова и др.). Цель использования таких критериев - выявить свойства в колебаниях анализируемых рядов, отличных от колебаний, имеющих место в случайных бессвязных рядах. Отметим, что в последних (рядах) числа повышений и понижений приблизительно равны, а число экстремумов распределено асимптотически нормально со средним ($m_э \approx 2 \cdot n/3$) и дисперсией ($D_э \approx (16 \cdot n - 29)/90$). Для проверки гипотезы случайности ряда (x_i) достаточно рассчитать фактическое нормированное число экстремумов

$$t_э = n^*_э - \frac{m_э}{\sqrt{D_э}}, \qquad (4.65)$$

и сравнить его с нормированным значением (t_α) нормального закона распределения при заданном уровне значимости (α). При ($t_э > t_\alpha$), гипотеза о случайности ряда будет, по-видимому, неверна. Сущность критерия проверки случайности по числу повышений или понижений ряда состоит в следующем. Пусть имеется выборка ($x_1, x_2, \ldots, x_n$); к повышениям (+) в ряду относится ситуация, когда ($x_{i-1} < x_i$), к понижениям (-) - когда ($x_{i-1} > x_i$). Общее число ситуаций ((+)+(-)=N) распределено асимптотически нормально с математическим ожиданием

$$m_+ = m_- = \frac{n}{2} \tag{4.66}$$

и дисперсией

$$D_+ = D_- = n + \frac{1}{12}. \tag{4.67}$$

Зная (m) и (D), можно рассчитать нормированные значения повышений или понижений

$$t^*_+ = \frac{n^*_+ - m_+}{\sqrt{D_+}}; \quad t^*_- = \frac{n^*_- - m_-}{\sqrt{D_-}}, \tag{4.68}$$

где n^*_+, n^*_- - соответственно, число повышений (понижений) в ряду. Затем, проводится сравнение (t^*_+) и (t^*_-) со значениями нормированных ординат таблицы закона нормального распределения. Если (t^*_+) и (t^*_-) по таблице окажется меньше уровня значимости, то гипотеза о случайности исследуемого ряда отвергается и принимается гипотеза об устойчивости тенденции к повышению или понижению. Критерий Б.П. Вайнберга позволяет обнаружить изменение уровня ряда, а критерий М.А. Омшанского - оценить длительность разных циклов. Критерий случайности О.А. Дроздова базируется на вычислении разностей-($d_1=x_2-x_1$), ($d_2=x_3-x_2$), ($d_{n-1}=x_n-x_{n-1}$), между членами исходной последовательности. Суммирование разностей (d_i) приводит к выражению

$$D_k = \sum_{i=1}^{k} d_i = x_{k+1} - x_1, \tag{4.69}$$

т.е. при суммировании разностей вновь получается разностный ряд, но с интервалом в (k) - членов. Относя эти накопления разности к среднему квадратическому отклонению разностного ряда (σ_{Dk}), получаем искомый критерий случайности (C_k) по сравнению с естественной изменчивостью

$$C_k = \frac{D_k}{\sigma_D}. \tag{4.70}$$

При наличии систематических тенденций эволюций уровня, ряд - - C_k (k=1,2,3,...,N), при одном из значений (k), наконец, выйдет за пределы нескольких единиц, что будет характеризовать надежность установления тенденции. Мощность критерия (C_k) возрастает при вычислении

разностей не между отдельными членами ряда, а между исследуемыми (n)-летиями

$$d_k^{(n)} = \frac{!}{n} \cdot \left(\sum_{i=k+1}^{k+n} x_i - \sum_{i=k}^{k+n+1} x_i \right). \qquad (4.71)$$

Суммирование ($d_k^{(n)}$) дает величины

$$D_k^{(n)} = \frac{!}{n} \cdot \left(\sum_{i=k+1}^{k+n} x_i - \sum_{i=1}^{n} x_i \right). \qquad (4.72)$$

Для случайного бессвязного ряда дисперсия определяется как

$$\sigma_D^{2(n)} = 2 \cdot \frac{\sigma_x^2}{n}, \qquad (4.73)$$

откуда -

$$C_k^n = \frac{D_k^{(n)}}{\sigma_D^{(n)}}. \qquad (4.74)$$

Сравнивая ряд ($D_k^{(n)}$) с ($\sigma_D^{(n)}$), можно более точно оценить существенность отличия изменений в ряду (x_i) от колебаний в случайном бессвязном ряду. Математически постановку задачи статистического прогноза можно сформулировать следующим образом. Пусть (x(t)) - стационарный процесс, наблюдавшийся до момента (t_0). После (t_0) сведений о значениях процесса нет. Требуется предсказать ($\hat{x}(t_0 + \Theta)$) - значение процесса в момент ($t_0+\Theta$), причем, - с наилучшей точностью. Истинное значение процесса ($x(t_0+\Theta)$), как правило, не совпадает с предсказанным ($\hat{x}(t_0 + \Theta)$). Их разность -

$$e(t + \Theta) = x(t_0 + \Theta) - \hat{x}(t_0 + \Theta) - \qquad (4.75)$$

представляет ошибку прогноза на время (Θ), произведенного в момент (t_0). Располагая прошлыми и текущими значениями прогноза (предысторией), можно получить характеристику процесса, определяющую связь между его значениями, разделенными временным промежутком (Θ), а именно - корреляционную функцию. В гидрометеорологических исследованиях понятие статистического прогноза обычно связывается с задачей экстраполяции (интерполяции) и сглаживания случайного процесса. По известной реализации

$$z(t) = x(t) + y(t), \qquad (4.76)$$

в которой x(t) - детерминированная составляющая; y(t) - случайная составляющая; в случае y(t)=0 (процесс без ошибок) прогноз сводится к чистой экстраполяции. В случае наличия ошибок (y(t)), прежде чем опре-

делить истинное значение реализации x(t) в некоторый момент (t+Θ), необходимо отделить его от ошибки наблюдения. Это задача о сглаживании (фильтрации) случайного процесса. Задача об экстраполяции тесно связана со сглаживанием, так как реализация случайного процесса включает в себя ошибки измерения. При этом, задача экстраполяции сглаживанием состоит в том, чтобы по имеющейся реализации (4.76) на промежутке (t_0+Θ) дать прогноз реализации (x(t)) в момент (t+Θ), при (Θ>0). При (Θ<0) - имеет место задача интерполяции со сглаживанием. В математической постановке задачи предполагается, что математические ожидания процессов - ($m_x(t)$), ($m_y(t)$), их автокорреляционные и взаимные корреляционные функции заданы. При этом, обычно считают ($m_x(t)$) и ($m_y(t)$) равными 0. В противном случае, вместо (x(t)) и (y(t)) рассматриваются их центрированные случайные функции (аномалии) - (Δx(t)), (Δy(t)). Математическое решение задачи статистического прогноза сводится к получению наилучшего результата по всему множеству реализаций, т.е. к нахождению такого оператора (L), который в применении к множеству реализаций (z(t)) давал бы наилучшее, в некотором смысле, значение реализации ($x(t_0+\Theta)$)

$$x(t_0+\Theta) = L(x(t)+y(t)). \qquad (4.77)$$

Для оценки качества прогнозирования вводится критерий качества прогнозирования, так или иначе связанный с ошибкой прогноза, - средним квадратом ошибки

$$\bar{e}^2(t_0+\Theta) = M((x(t_0+\Theta)-\hat{x}(t_0+\Theta))^2) \Rightarrow \min. \qquad (4.78)$$

Чтобы вычислить предсказанное значение, нужно уметь выбрать правило вычисления ожидаемой оценки ($\hat{x}(t_0+\Theta)$) - алгоритм прогноза. Алгоритм, предсказания ($\hat{x}$) должен связать с ним предысторию процесса и его вероятностные характеристики. С другой стороны, качество алгоритма определится дисперсией ошибки прогноза. *Простейшие алгоритмы прогноза рассмотрены ниже.*

Вероятностное прогнозирование значений случайного процесса

Оценка значения в реализации случайного процесса (в силу случайности физических явлений) в будущем (в момент времени - "t") не может быть вычислена по точной формуле, но может быть описана в вероятностном виде. В случае стационарного и эргодического процессов, (P(x,t)) не зави-

сит от времени и может быть определена по единственной реализации (x(t)) как

$$P(x) = P(x(t) \le \xi) = \lim_{T \to \infty} \left(\frac{T(x(t) \le \xi)}{T}\right), \qquad (4.79)$$

где T[x(t)≤ ξ] - общее время, в течение которого реализация (x(t)) находится не выше уровня (ξ) . В этом случае, значение (x(t)) в произвольный момент времени не превышает данного значения (ξ). При (x→-∞), интегральная функция распределения (P(x)=P(x,t)) стремится к 0, при (x→∞), - стремится к 1. По характеру изменения функции распределения от 0 до 1 различаются случайные процессы с разной вероятностной структурой. Можно сказать, что плотность вероятности определяет скорость изменения функции распределения, поскольку вероятность различных событий можно находить интегрированием плотности вероятности на кривой

$$P(x_1 \le x(t) \le x_2) = \int_{x_1}^{x_2} P(x) \cdot dx = P(x_2) - P(x_1). \qquad (4.80)$$

Функция распределения существует как для непрерывных, так и для прерывных случайных величин и является универсальной характеристикой случайных величин, так как плотность характеризует их с вероятностной точки зрения. Зная функцию распределения случайной величины, можно найти вероятность ее попадания на заданный участок, которая равна приращению функции распределения на этом участке. При ($x_1 \to 0$), получим (x_2),

$$P(0 \le x(t) \le x_2) = \int_{-\infty}^{x_2} P(x) \cdot dx = P(x_2), \qquad (4.81)$$

т.е. площадь под графиком плотности вероятности левее точки (x_2) равна значению дифференциальной функции распределения в точке (x_2). Таким образом, прогнозирование вероятности того или иного элемента может быть осуществлено при знании или прогнозировании функции распределения. Задача прогнозирования при использовании вероятностных моделей заключается в определении по кривой распределения вероятностей величины параметра (x), такого, когда вероятность (P(x)) равна заданному значению (P). Следует помнить, что точность прогноза, с вероятностной точки зрения, в этом случае, будет зависеть от точности прогноза функции распределения.

Прогноз по последнему значению

Прогнозирование по последнему значению реализации, называемое "экстраполяцией нулевого порядка" (инерционный прогноз), заключается в том, что в качестве предсказанного значения ($\hat{x}(t_0+\Theta)$) принимается значение ($x(t_0)$)

$$\hat{x}(t_0+\Theta) = x(t_0). \qquad (4.82)$$

Предсказанное значение здесь не зависит от предыстории прогноза (предыстория представлена лишь одной точкой - последним значением "$x(t_0)$"), а вероятностные характеристики не учитываются совсем. Алгоритм прогноза заключается в умножении значения последнего наблюдения ($x(t_0)$) на 1, т.е. не требует выполнения никаких вычислительных операций. Прогноз, таким образом, можно выполнить, ничего не зная о процессе, не производя никаких вычислений. Однако, точность прогноза очень низкая. Возможная ошибка прогноза по алгоритму (4.75) здесь определяется как

$$e(t_0+\Theta) = x(t_0+\Theta) - x(t_0), \qquad (4.83)$$

а ее средний квадрат ($\bar{e}^2$), если ($m_x=0$),

$$\bar{e}^2(\Theta) = M((x(t_0+\Theta) - x(t))^2) = \sigma_x^2 - 2\cdot r_x(\Theta) + \sigma_x^2 = 2\cdot(\sigma_x^2 - r_x(\Theta)) \ . \qquad (4.84)$$

Средний квадрат ошибки прогноза растет от 0, при $\Theta=0$, когда $r_x(0) = \sigma_x^2$, до ($2\cdot\sigma_x^2$), при $\Theta=\infty$, когда $r_x(\infty) = 0$. Об истинном качестве этого способа прогноза можно говорить после сравнения полученной ошибки с ошибками других алгоритмов и способов прогноза. Простота этого способа обеспечила ему широкое распространение.

Прогноз по математическому ожиданию

Прогнозирование по математическому ожиданию заключается в том, что в качестве предсказанного значения ($\hat{x}(t_0+\Theta)$) принимается математическое ожидание (m_x). Как и в предыдущем способе, предсказанное значение здесь не зависит от времени прогноза (Θ). Различие заключается в том, что, хотя не требуется никакой информации о предыстории, необходимы сведения о свойствах процесса - о его математическом ожидании. Алгоритм прогноза не требует никаких вычислительных операций. Ошибка прогноза вычисляется по зависимости -

$$e(\Theta) = x(t+\Theta) - m_x \ - \qquad (4.85)$$

и представляет собой отклонение процесса от среднего в момент ($t_0+\Theta$).

Средний квадрат ошибки не зависит от времени прогноза и равен дисперсии прогноза

$$\bar{e}^2(\Theta) = M((x(t_0 + \Theta) - m_x))^2) = \sigma_x^2. \quad (4.86)$$

При малой заблаговременности (Θ) прогноз по последнему значению явно предпочтителен, однако, после получения некоторой критической величины заблаговременности прогноза (Θ*), когда ($\bar{e}^2(\Theta^*) = \sigma_x^2$), метод прогноза по математическому ожиданию дает большую точность. Наконец, при (Θ→∞), квадрат ошибки прогноза по математическому ожиданию (норме) вдвое меньше, чем по последнему отсчету. Так, предсказывая расход воды в реке на несколько дней, мы, руководствуясь инерцией, ориентируемся на ее текущее состояние, совершенно игнорируя средние многолетние величины. Наоборот, пытаясь предвидеть летом весеннее половодье, мы, напротив, прежде всего ориентируемся на "норму" половодья.

Статистический прогноз по одной точке

Стационарный эргодический процесс может быть как ансамблем реализаций, так и одной реализацией неограниченной длительности. Сечения ансамбля представляют собой случайные величины, функция распределения которых отождествляется с одномерной функцией распределения процесса. Обозначим случайную величину ($x(t_0)$) как сечение процесса в момент (t_0) - через (x), а сечение ($x(t_0+\Theta)$) - через (y) и будем рассматривать систему двух случайных величин (x, y) - последнего значения предыстории и предсказанного значения. Компоненты системы (x) и (y) подчинены одномерным нормальным законам - ($N_1(m_x,\sigma_x)$), ($N_2(m_y,\sigma_y)$).

Алгоритм прогноза в рассматриваемом способе формулируется так, что в качестве предсказанного значения ($\hat{x}(t_0 + \Theta)$) выступает условное математическое ожидание ($m_{y/x}$) величины (y), при условии, что ($x=x(t_0)$) -

$$\hat{x}(t_0 + \Theta) = m_{y/x}, \quad (4.87)$$

где аналогично уравнению регрессии:

$$m_{y/x} = m_y + r\frac{\sigma_y}{\sigma_x} \cdot (x - m_x) \; ; \quad (4.88)$$

$$\sigma_{y/x} = \sigma_y \cdot \sqrt{1 - r^2}. \quad (4.89)$$

Известно, что закон, постулированный при условии, что первая компонента (x) приняла определенное значение, называется условным законом распределения и имеет вид

$$f_{(y/x)} = \frac{f(x,y)}{f(x)} = \frac{1}{\sigma_y \cdot \sqrt{2\cdot\pi}\cdot\sqrt{1-r^2}} \cdot \exp\left(-\frac{1}{2(1-r^2)}\right) \cdot \left(\frac{y-m_y}{\sigma_y} - r\cdot\frac{x-m_x}{\sigma_x}\right) \quad (4.90)$$

или

$$f_{(y/x)} = \frac{1}{\sigma_y \cdot \sqrt{2\cdot\pi}\cdot\sqrt{1-r^2}} \cdot \exp\left(-\frac{1}{2(1-r^2)}\right) \cdot \left(y - m_y - r\cdot\frac{\sigma_y}{\sigma_x}\cdot(x-m_x)\right). \quad (4.91)$$

С учетом (4.88) и (4.89), получим плотность нормально распределенной условной случайной величины

$$f_{(y/x)} = \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_{y/x}} \cdot \exp\left(-\frac{y-m_{y/x}}{2\cdot\sigma^2_{y/x}}\right). \quad (4.92)$$

Из (4.92) следует, что, при изменении одной из компонент, вид закона распределения второй компоненты не меняется, а меняется лишь его параметр ($m_{y/x}$) - условное математическое ожидание (4.88). Условная дисперсия (4.89) от значения (x) также не зависит. Зависимость ($m_{y/x}$) от (x) линейна и называется регрессией (y) на (x). Ошибка прогноза по соотношению (4.88) определяется как -

$$e(t_0+\Theta) = x(t_0+\Theta) - \hat{x}(t_0+\Theta) = y - m_{y/x} \quad (4.93)$$

и представляет собой отклонение случайной величины (y) от своего условного математического ожидания, а средний квадрат ошибки- $\bar{e}^2(\Theta) = M((y-m_{y/x})^2)$ равен условной дисперсии ($\sigma^2_{y/x}$). Учитывая (4.89) и (4.90), установим, что

$$\hat{x}(t_0+\Theta) = m_{x/y} = m_y + r_{x/y}\cdot\frac{\sigma_y}{\sigma_x}\cdot(x-m_x) \quad (4.94)$$

и

$$\bar{e}^2(\Theta) = \sigma_{y/x} = \sigma^2_y\cdot(1-r^2). \quad (4.95)$$

Поскольку процесс (x(t)) стационарен, математические ожидания и дисперсии сечений одинаковы: $\sigma_y = \sigma_x = \sigma$; $m_y = m_x = m$. Коэффициент корреляции (r_{xy}) равен значению нормированной корреляционной

(автокорреляционной) функции

$$r_{xy} = r(\Theta). \qquad (4.96)$$

Теперь алгоритм прогноза прост

$$\hat{x}(t_0 + \Theta) = m + r(\Theta) \cdot (x(t_0) - m), \qquad (4.97)$$

а средний квадрат ошибки оказывается зависящим от Θ

$$e^{-2} = \sigma^2 \cdot (1 - r(\Theta)). \qquad (4.98)$$

Алгоритм предполагает знание отклонения процесса от среднего в момент (t_0), т.е. одной точки предыстории, знание нормированной корреляционной функции (r(Θ)) и математического ожидания (m).

Динамико-стохастический метод сверхдолгосрочного прогноза

Математический аппарат прогнозирования динамико-стохастическим методом содержит в своей основе предположение о стационарности прогнозируемых процессов во времени и наличии в процессе внутрирядных связей (даже при сдвиге степени τ>1). Задача линейного экстраполирования (прогноза) стационарной случайной последовательности, удовлетворяющей условиям: m_x=const, r_x=r(τ)≠0, заключается в необходимости подбора таких действительных коэффициентов (κ_τ), при заданных (m>0) и (Θ>0), при которых линейная комбинация -

$$\hat{x}_0(t_0 + \Theta) = \bar{x} + k_1 \cdot \Delta x_{t-1} + k_2 \cdot \Delta x_{t-2} + \ldots + k_m \cdot \Delta x_{t-m} = \bar{x} + \sum_{\tau=1}^{m} k_\tau \cdot \Delta x_{t-\tau} \quad (4.99)$$

является наиболее точным приближением к случайной величине ($\hat{x}(t)$). В соотношении (4.99): ($\bar{x}_0 \cdot (t_0 + \Theta)$) - прогнозируемое на момент ($t=t_0+\Theta$) значение исследуемой величины (x(t)); $\bar{x}$ - среднее значение (x(t)); Δx_{t-1}, Δx_{t-2} ,..., Δx_{t-m} - последовательность предшествующих моменту (t) значений величины (x(t)) в отклонениях от ($\bar{x}$) ; $k_1, k_2, \ldots, k_m$ - коэффициенты обратной связи, определяемые путем решения системы уравнений

$$\left.\begin{aligned} k_1 + k_2 \cdot r_2 + \ldots + k_m \cdot r_{m-1} &= r_1, \\ k_1 \cdot r_1 + k_2 + \ldots + k_m \cdot r_{m-2} &= r_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ k_1 \cdot r_{m-1} + k_2 \cdot r_{m-2} + \ldots + k_m &= r_m \end{aligned}\right\}, \qquad (4.100)$$

где $r_1, r_2, \ldots, r_m$ - последовательность значений корреляционной функции (r(τ)); m - оптимальный период обратной связи. Для определения периода

обратной связи ($m_{опт.}$) рекомендуется выполнять проверочное прогнозирование, при значениях (m), последовательно увеличивающихся от 1 до 30 лет. В качестве оптимального выбирается то значение, при котором ошибка прогноза ($\bar{e}^2$) становится минимальной. Оценка точности прогнозов, как правило, производится по последовательности эмпирических коэффициентов связи между фактическими ($\Delta x_{i\phi}$) и прогностическими (Δx_{ip}) аномалиями ряда. При этом, прогностическое значение ($\Delta x(t+\theta)$) отыскивается по уравнению авторегрессии вида

$$\Delta x(t+\Theta) = \sum_{k=0}^{m} \alpha_k \cdot \Delta x(t-k). \qquad (4.101)$$

Коэффициенты (α_k) для каждого заданного значения (θ), определяются, исходя из условия минимума ошибки экстраполяции, при решении системы уравнений

$$r_{\Delta x}(\Theta + j) = \sum_{k=1}^{m} \alpha_k \cdot r_{\Delta x}(k-j); \text{ при } j=1,2,...,m, \qquad (4.102)$$

где $r_{\Delta x}(\tau)$ - корреляционная функция отклонений. Число слагаемых (m) в сумме ($\sum_{k=1}^{m}$) следует выбирать таким, чтобы корреляционные моменты ($r_{\Delta x}(k-j)$) определялись по данным наблюдений в (m) - точках с требуемой надежностью. На рисунке 4.4 приведены результаты прогноза годового стока реки Волга в створе Самары с заблаговременностью 1 год.

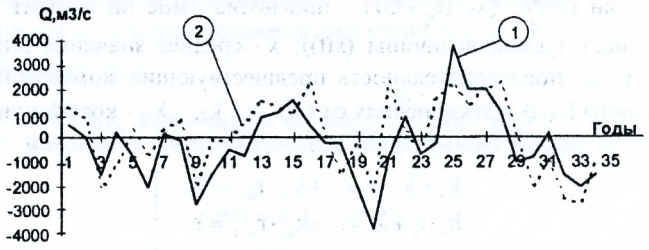


Рисунок 4.4 Результаты прогноза годового стока (Q, м³/с) реки Волга у Самары с заблаговременностью один год: 1-фактический сток, 2- прогнозные величины стока.

5 ДИСПЕРСИОННЫЙ АНАЛИЗ И СПОСОБЫ ЕГО ИСПОЛЬЗОВАНИЯ

В природопользовании, при постановке прикладных комплексных исследований, экспериментальные данные часто требуется разбить на группы, отличающиеся между собой в количественном отношении, и установить сходство (различие) между ними. Например, определить степень влияния географических условий на ход тех или иных природных процессов и явлений. Лучше всего этим требованиям отвечает дисперсионный анализ, который широко используется при решении именно практических задач. Дисперсионный анализ позволяет установить с определенной долей уверенности влияние на изучаемый объект каждого из исследуемых факторов, в отдельности, или в определенных их сочетаниях. Необходимым условием использования дисперсионного анализа является разбивка каждого учитываемого фактора не менее чем на две группы. Если влияние факторов нельзя выразить количественными показателями, то они могут быть представлены качественными показателями, выраженными в виде баллов. Дисперсионный анализ разработан и введен в практику английским ученым Р. Фишером, который открыл закон распределения отношения средних квадратов (дисперсий)

$$\frac{S_1^2}{S_2^2} = F \,, \qquad (5.1)$$

где S_1^2 - среднеквадратическая ошибка средних отдельного опыта; S_2^2 - суммарная среднеквадратическая ошибка средних всех опытов. Использование методов дисперсионного анализа позволяет дать ответ на следующие вопросы: 1) *значимо ли влияет изучаемый фактор* на воспроизводимость и в целом *на результат*? 2) если *установлено значимое влияние* какого-либо фактора на результат, в целом, то начиная *с какого уровня* фактора это влияние действует; значимо ли различаются выборочные средние между собой? 3) *какой* количественной *мерой* можно *оценить степень* установленного *влияния*? Сущностью дисперсионного анализа является расчленение общей суммы квадратов отклонений и общего числа степеней свободы на части - компоненты, соответствующие структуре эксперимента, и оценка значимости действия и взаимодействия изучаемых факторов по F-критерию. С этой целью, дисперсия разделяется на независимые слагаемые, которые, затем, сравниваются между собой. До-

пустим, в результате измерения величины (M) получено значение (X) и пусть на процесс измерения влияют случайные независимые факторы (A) и (B). Тогда, отклонение - (M-X=α+β+γ), где α - отклонение под влиянием фактора (A) , β - то же под влиянием фактора (B), а γ - под влиянием остальных, неучтенных факторов, причем, (α), (β) и (γ)-независимы. В этом случае, дисперсия- D(M-X)=D(α+β+γ), а DX=(Dα+Dβ+Dγ), где Dα - характеризует влияние фактора (A), Dβ - влияние фактора (B), Dγ - влияние остальных, неучтенных факторов. Дисперсия (Dγ) называется *остаточной дисперсией.* Для оценки значимости факторов (A) и (B) сравниваются соответствующие дисперсии (Dα) и (Dβ) с остаточной дисперсией (Dγ). Если исследуется влияние одного фактора, то говорят об однофакторном дисперсионном анализе, при исследовании влияния двух факторов - о двухфакторном дисперсионном анализе и т.п.

Рассмотрим *простейшие способы применения дисперсионного анализа*, техника проведения которого довольно разнообразна.

5.1 Однофакторный дисперсионный анализ

При решении практических задач в природопользовании наиболее часто используется однофакторный дисперсионный анализ. При этом, в опыте должно быть предусмотрено не менее трех повторностей. Исследуемые данные разбиваются на группы с целью выявления оптимального значения фактора, влияющего на результативный признак. При однофакторном дисперсионном анализе обычно изучается действие одного фактора на (m) - уровнях (k>2), при равном числе определений (измерений) на каждом уровне (n). Пусть фактор имеет (m) - уровней. Из каждого уровня делается выборка из (n) - элементов. Общее количество выбранных элементов обозначается как - (N=m·n). Вся выборка представляет собой матрицу

$$X = \begin{Vmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{Vmatrix}. \qquad (5.2)$$

Полагая, что данная выборка сделана из нормально распределенной генеральной совокупности, и задавая уровень значимости (α), необходимо проверить гипотезу (H_0) о равенстве средних значений на всех уровнях

фактора - ($\mu_1=\mu_2=...=\mu_m$). При альтернативной гипотезе (H_1) не все средние значения (μ_i) должны быть равными. В качестве статистики используется величина (F), определяемая по аналогичной с (5.1) зависимости

$$F = \frac{S_1^2}{S_2^2}, \qquad (5.3)$$

в которой S_1^2 - дисперсия, характеризующая влияние исследуемого фактора (факторная дисперсия); S_2^2 - дисперсия, характеризующая влияние остальных факторов (остаточная дисперсия). Если гипотеза (H_0) верна, то случайная величина (F) имеет F-распределение со степенями свободы (m-1) и (N-m=m·(n-1)). При проверке гипотезы (H_0) используется правосторонняя критическая область, исходя из условия

$$P(F > f_a) = \alpha . \qquad (5.4)$$

Если значение статистики входит в критическую область, то гипотеза (H_0) о равенстве средних значений на всех уровнях фактора отвергается, т.е. считается значимым влияние исследуемого фактора. В противном случае, принимается гипотеза (H_0), т.е. считается, что значимость влияния фактора не установлена. При определении (F) находится сумма квадратов отклонений элементов выборки относительно общего среднего арифметического

$$Q = \sum_{j=1}^{m} \cdot \sum_{i=1}^{n} (X_{ij} - \overline{X})^2 , \qquad (5.5)$$

где $\overline{X} = \frac{1}{N} \cdot \sum_{j=1}^{m} \cdot \sum_{i=1}^{n} X_{ij}$, которая, в свою очередь, может быть разделена на два независимых слагаемых -

$$Q_1 = n \cdot \sum_{j=1}^{m} (\overline{X}_j - \overline{X})^2 , \qquad (5.6)$$

$$Q_2 = \sum_{j=1}^{m} \cdot \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2 - \qquad (5.7)$$

так, чтобы выполнялось равенство ($Q=Q_1+Q_2$); здесь ($\overline{X}_j = \frac{1}{n} \cdot \sum_{j=1}^{n} X_{ij}$), при (j=1,2,...,m) - групповое среднее. Суммой квадратов межгрупповых отклонений, характеризующих влияние исследуемого фактора, является сумма квадратов отклонений групповых средних относительно общей средней (сумма Q_1). С другой стороны, (Q_2) представляет собой сумму квадратов отклонений значений выборки относительно групповых средних, так называемую сумму квадратов внутри групповых отклонений.

Эта сумма характеризует влияние остальных, неучтенных факторов. Имея суммы (Q), (Q_1, Q_2), можно вычислить соответствующие дисперсии

$$S^2 = \frac{Q}{m \cdot n - 1}; \quad S_1^2 = \frac{Q_1}{m - 1}; \quad S_2^2 = \frac{Q_2}{m \cdot (n - 1)}. \qquad (5.8)$$

Две последние дисперсии используются при вычислении (F). При практических вычислениях, величины (Q) и (Q_1) находятся, обычно, по выборке, а (Q_2) - определяется как разность

$$Q_2 = Q - Q_1. \qquad (5.9)$$

В простейших случаях, на каждом уровне фактора, выбирается одинаковое количество объектов исследований, но нахождение сумм по (5.5)...(5.7) достаточно сложно, и более приемлемые формулы можно получить, преобразуя выражения (5.5) и (5.6), когда:

$$Q = \sum_{j=1}^{m}(\sum_{i=1}^{n} X_{ij}^2) - \frac{1}{m \cdot n} \cdot (\sum_{j=1}^{m}(\sum_{i=1}^{n} X_{ij}))^2; \qquad (5.10)$$

$$Q_1 = \frac{1}{n} \cdot \sum_{j=1}^{m}(\sum_{i=1}^{n} X_{ij})^2 - \frac{1}{m \cdot n} \cdot (\sum_{j=1}^{m}(\sum_{i=1}^{n} X_{ij}))^2. \qquad (5.11)$$

Тогда, обозначив ($R_j = \sum_{i=1}^{n} X_{ij}^2$) и ($L_j = \sum_{i=1}^{n} X_{ij}$), окончательно можно получить:

$$Q = \sum_{j=1}^{m} R_j - \frac{1}{m \cdot n} \cdot (\sum_{j=1}^{m} L_j)^2; \qquad (5.12)$$

$$Q_1 = \frac{1}{n} \cdot \sum_{j=1}^{m} L_j^2 - \frac{1}{m \cdot n} \cdot (\sum_{j=1}^{m} L_j)^2. \qquad (5.13)$$

По исходной матрице (выборке) вычисляются суммы элементов и их квадратов по столбцам (L_j) и (R_j), при (j=1,2,...,m). Далее производится замена переменных ($Y_{ij}=X_{ij}-C$). Целесообразно их выбрать близкими к общему среднему. В результате замены (для Y_{ij}) получаются следующие формулы:

$$Q = \sum_{j=1}^{n} P_j - \frac{1}{N} \cdot (\sum_{j=1}^{m} T_j)^2; \qquad (5.14)$$

$$Q_1 = \frac{1}{n} \cdot \sum_{j=1}^{n} T_j^2 - \frac{1}{N} \cdot (\sum_{j=1}^{m} T_j)^2, \qquad (5.15)$$

где $N=m \cdot n$; $P_j = \sum_{i=1}^{n} Y_{ij}^2$ и $T_j = \sum_{i=1}^{n} Y_{ij}$, при (j=1,2,...,m). На практике не всегда удается гарантировать одинаковое количество элементов на каждом

уровне фактора. Если количество элементов на j-м уровне, обозначенное через (n_j), при (j=1,2,...,m), то объем выборки составит

$$N = \sum_{j=1}^{m} n_j \ . \qquad (5.16)$$

Формула (5.15) может быть записана в виде

$$Q_1 = \sum_{j=1}^{m} \frac{1}{n_j} \cdot T_j^2 - \frac{1}{N} \cdot (\sum_{j=1}^{m} T_j)^2 \ , \qquad (5.17)$$

а суммы (P_j) и (Tj) как - ($P_j = \sum_{i=1}^{n_j} Y_{ij}^2$) и ($T_j = \sum_{i=1}^{n_j} Y_{ij}$). Тогда дисперсии определяются по зависимостям

$$\sigma^2 = \frac{Q}{N-1} \ ; \quad \sigma_1^2 = \frac{Q_1}{m-1} \ ; \quad \sigma_2^2 = \frac{Q_2}{N-m} \ . \qquad (5.18)$$

Остальные формулы, используемые при решении задач, не изменяются.

На примере материалов полевого опыта, *в котором сравнивается урожайность озимой пшеницы при пяти вариантах технологий внесения удобрений и обработки почвы (таблица 5.1),* ***можно проследить порядок выполнения дисперсионного анализа экспериментальных данных.***

Таблица 5.1 Средняя урожайность озимой пшеницы по вариантам опыта, ц/га

Вариант	Урожайность по повторениям				Суммы по вариантам	Средняя урожайность
	I	II	III	IV		
A (контроль)	47,8	46,9	45,4	44,1	184,2	46,0
B	53,7	50,3	50,6	48,0	202,6	50,6
C	46,7	42,0	43,4	40,7	172,8	43,2
D	48,0	47,0	45,9	45,7	186,6	46,6
E	41,8	40,0	43,0	41,6	166,4	41,6
Суммы по повторениям	238,0	226,2	228,3	220,1	912,6=ΣXi= =ΣVi=ΣPi	45,6= $\overline{X}$

Алгоритм обработки опытных данных следующий:

*1) **в исходной таблице подсчитываются суммы урожаев** - по вариантам (строки) - (Vi), по повторениям (столбцы) - (Pi) и определяются средние урожайности по вариантам (последний столбец);*

*2) **проверяется правильность вычислений**, для чего подсчитывается сумма сумм урожаев по вариантам и повторениям -*

$\sum Vi=184,2+202,6+172,8+186,6+166,4=912,6$ (ц/га),

$\sum Pi=238,0+226,2+228,3+220,1=912,6$ (ц/га),

и проверяется равенство - $\sum Xi=\sum Vi=\sum Pi=912,6$ (ц/га);

*3) **вычисляются средние значения урожаев по вариантам** путем деления сумм по вариантам (Vi) на число повторений (в рассматриваемом примере, n=4);*

*4) **определяется средняя урожайность озимой пшеницы**, в целом по опыту, делением общей суммы урожаев $\sum Xi$ на общее число делянок (Ni) в опыте (20 дел.);*

*5) **преобразуются исходные данные** по соотношению ($X'_{ij} = X_{ij} - C$), когда приняв за условное среднее число 45, близкое к среднему урожаю по опыту, можно облегчить вычисления сумм квадратов. Для варианта А (повторение I), при урожае 47,8 (ц/га) значение X'_{11} =47,8-45=2,8 и т.д. Преобразования значительно упрощают все последующие вычисления и не оказывают влияния на величину сумм квадратов отклонений. Преобразованные данные записываются в таблицу 5.2, определяются суммы по повторениям (графам) и вариантам (строкам) и проверяется правильность расчетов по равенству - ($\sum P'_{ij} = \sum V'_{ij} = \sum X'_{ij}$) = 12,6 (ц/га);*

Таблица 5.2 Преобразованные данные

Вариант	$X'_{ij} = X_{ij} - C = X_{ij} - 45$ (ц/га)				$\sum V'_{ij}$
	I	II	III	IV	V
А (контроль)	2,8	1,9	0,4	-0,9	4,2
В	8,7	5,3	5,6	3,0	22,6

Продолжение таблицы 5.2

C	*1,7*	*-3,0*	*-1,6*	*-4,3*	*-7,2*
D	*3,0*	*2,0*	*0,9*	*0,7*	*6,6*
E	*-3,2*	*-5,0*	*-2,0*	*-3,4*	*-13,6*
$\sum P_{ij}^{'}$	*13,0*	*1,2*	*3,3*	*-4,9*	$\sum X_{ij}^{'}$ *=12,6*

*6) **вычисляются суммы квадратов отклонений** для различных компонентов варьирования в следующей последовательности -*

а) общее число наблюдений - $(N=m\cdot n)=5\cdot 4=20$;

б) корректирующий фактор - $(S=(\sum X_{ij}^{'})^2/N)=(12,6)^2/20=7,94$;

в) суммы квадратов -

общая - $(S_Y=\sum(X_{ij}^{'})^2-S)=(2,8^2+1,9^2+...+3,4^2)-7,94=246,67$,

повторений - $(S_P=\sum(P_{ij}^{'})^2/m-S)=(13,0^2+1,2^2+2,3^2+4,9^2)/5-7,94=33,13$,

вариантов - $(S_V=\sum(V_{ij}^{'})^2/n-S)=(4,2^2+22,6^2+...+13,6^2)/4-7,94=194,25$,

ошибки: $(S_z=S_Y-S_P-S_V)=246,67-33,13-194,25=19,29$;

*7) **заполняется таблица дисперсионного анализа** (таблица 5.3).*

Таблица 5.3 Результаты дисперсионного анализа

Варианты	*Сумма квадратов отклонений*	*Степень свободы*	*Дисперсия*	*Критерий Фишера*	
				F_Φ	$F_{0.05}^T$
Общая	*246,67*	*19*	--	--	--
Повторений	*33,13*	*3*	--	--	--
Вариантов	*194,25*	*4*	*478,56*	*30,25*	*3,26*
Остаток (ошибки)	*19,29*	*12*	*1,6*	--	--

Значение критерия $F_{0.05}^T$ находится по Приложению (таблица П.4.1) для 4-х степеней свободы дисперсии вариантов (числитель) и для 12 степеней - дисперсии ошибки (знаменатель). В опыте есть существенные различия между вариантами и (H_0:d=0) отвергается ($F_\Phi > F_{0.05}^T$);

*8) для оценки существенности частных различий и группировки вариантов **вычисляются - ошибка опыта, ошибка разности средних и наименьшая существенная***

разность ($НСР_{0,05}$) в абсолютных и относительных величинах -

$$S_{\bar{X}} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{1,60}{4}} \approx 0,64 \text{ (ц / га)},$$

$$S_d = \sqrt{\frac{2 \cdot S^2}{n}} = \sqrt{\frac{2 \cdot 1,60}{4}} \approx 0,90 \text{ (ц / га)},$$

$$НСР_{0,05} = t_{0,05} \cdot S_d = 2,18 \cdot 0,90 = 1,96 \approx 2,0 \text{ (ц / га)},$$

$$НСР_{0,05} = \frac{t_{0,05} \cdot S_d}{\bar{X}} \cdot 100\% = \frac{2,18 \cdot 0,90}{45,6} \cdot 100\% = 4,3 \text{ (\%)}.$$

Для 12 степеней свободы ошибки, которые находятся как $(m-1)(n-1)=(5-1)(4-1)=12$, по Приложению (таблица П.4.1) отыскивается $F^T_{0,05}=2,18$;

*9) **результаты эксперимента и статистической обработки записываются в итоговую таблицу**; на основе ($НСР_{0,05}$) распределяются варианты по группам и делаются выводы.*

Таблица 5.4 Сравнение урожайности озимой пшеницы с контролем

Вариант	Средняя урожайность, (ц/га)	Разность с контролем		Группа	Заключение о существенности разности
		ц/га	%		
A (контроль)	46,0	--	--	--	--
B	50,6	4,6	10,0	I	Существенна
C	43,2	-2,8	-6,1	III	Существенна
D	46,6	0,6	1,3	II	Не существенна
E	41,6	-4,4	-9,6	II	Существенна
$НСР_{0,05}$	--	2,0	4,3		

При распределении вариантов по величине ($НСР_{0,05}$) на три группы, руководствуются следующими положениями:

***I группа** - отклонения средних урожаев от контроля с положительным знаком больше ($НСР_{0,05}$) (существенное повышение урожая);*

***II группа** - отклонение не выходит за пределы ($\pm НСР_{0,05}$) (разность несущественная);*

***III группа** - отклонения с отрицательным знаком больше по абсолютной величине ($НСР_{0,05}$) (существенное снижение урожая).*

***Исходя из подобной группировки**, вариант **(B)** (группа I) существенно превышает по урожаю контрольный вариант, а варианты **(C)** и **(E)** (III группа) существенно уступают ему. Вариант **(D)** (II группа) несущественно отличается от стандарта (контроля). **Следовательно**, на основе статистической обработки данных по урожай-*

*ности озимой ржи, полученной в результате полевого опыта с пятью вариантами технологий внесения удобрений и обработки почвы, **можно считать**, что **в варианте (В)** в данных условиях **получены более высокие урожаи**, чем при традиционной технологии в контрольном варианте (А).*

Рассмотрим другой пример, ***когда используется критерий Фишера для установления различия влияния фактора на конечный результат анализа.***

*Известно, что оптимальным условиям питания культурных растений соответствует среда достаточно увлажненной дерновой легкосуглинистой гумусированной нейтральной почвы. Ее можно создать путем внесения в пахотный горизонт добавок минерального грунта с определенным механическим составом и торфа. Формирование антропогенного почвенного слоя требует предварительных полевых экспериментов. В связи с этим, **поставлена задача**: определить влияние на урожай зерна ячменя различных доз торфа (200, 300, 400, тонн абсолютно сухого вещества на гектар) при внесении его на фоне органоминеральных, органических удобрений и доломитовой муки. **Исходная почва** - дерново-подзолистая, глеевая, связносупесчаная, мелиорированная. **Сведения об урожайности** зерна ячменя в названных условиях приведены в таблице 5.5, куда занесена исходная информация по группам влияющего фактора (вариантам опыта) и некоторые результаты расчетов (для удобства сделано округление по урожайности до целых чисел).*

*Производятся расчеты по вариантам опыта (строкам), с разноской данных по столбцам: **суммарный по повторностям урожай** ячменя ($\sum X_i$) по каждому варианту опыта **вносится в столбец 6** (числитель); затем, **в столбце 7, приводятся квадраты суммарного урожая** ячменя **по повторностям** ($(\sum X_i)^2$; среднее арифметическое ($\overline{X}_i$) по общему варианту опыта, **заносится в столбец 8; определяется общее среднее** ($\overline{X}_{общ}$).*

*После получения данных по вариантам опыта **производятся расчеты** необходимых показателей **по повторностям** (X_k). Сначала **суммируются данные** по урожайностям ячменя и приводятся **в строке** ($\sum X_k$) по повторностям. **Суммы** урожайности ячменя **по вариантам опыта и повторностям должны совпасть и дать сумму** всех вариант ($\Sigma\Sigma X_{i,k}=495$). Аналогично суммируются квадраты этих показателей (знаменатель) по повторностям ($\sum X_k^2$). **Суммы сумм квадратов по вариантам и повторностям опыта должны совпадать и дать сумму квадратов всех вариант** ($\sum X_i^2 = \sum X_k^2 = 15920$). Ниже **записываются результаты возведения в квадрат сумм вариант** по каждой повторности ($\sum X_k)^2$ и их сумма - $\Sigma(\sum X_k)^2=61269$. **Вычисляется среднее арифметическое по каждой повторности опыта** ($\overline{X}_k$). **Общее среднее арифметическое всех вариант опыта** составляет*

$$\overline{X}_{общ.} = (\Sigma X_{i,k})/N = 495/16 = 30{,}93.$$

Таблица 5.5 Однофакторный дисперсионный комплекс

Варианты опыта	Урожай ячменя по повторностям, ц/га*)				$\frac{\sum X_i}{\sum X_i^2}$	$(\sum X_i)^2$	M_i
	I	II	III	IV			
1	2	3	4	5	6	7	8
Контроль (фон)	20 400	21 441	22 484	20 400	83 1725	6889	20,75
Фон+200 т/га торфа	30 900	32 1024	32 1032	31 961	125 3909	15625	31,25
Фон+300 т/га торфа	35 1225	36 1296	35 1225	36 1296	142 5032	20164	35,50
Фон+400 т/га торфа	36 1296	35 1225	37 1369	37 1369	145 5254	21025	36,25
$\sum X_k$	121	124	126	124	$\sum\sum X_{i,k}$ = =495	$\sum(\sum X_i)^2$ = =63703	
$\sum X_k^2$	3816	3981	4102	4021	$\sum\sum X_{ik}^2$ =15920		
$(\sum X_k)^2$	14641	15376	15876	15376	$\sum(\sum X_k)^2$=61269		
$\bar{X}_k$	30,25	31,00	31,50	31,00		$\bar{X}_{общ}$ =30,93	

) **В числителе - опытные данные, **в знаменателе** - квадраты этих показателей.*

***Следующий этап расчетов** - нахождение сумм квадратов отклонений, т.е. расчленение общего варьирования признака на составные части, исходя из равенства ($Q=Q_1+Q_2+Q_3$), где Q - сумма квадратов отклонений по общему варьированию данных; Q_1 - то же по группам фактора (варианты опыта); Q_2 - то же по повторностям опыта; Q_3 - то же по остаточному варьированию.*

Общая сумма квадратов отклонений вычисляется следующим образом: $(Q=(\sum\sum X_{ik}^2-(\sum\sum X_{i,k})^2/N))=(15920-(495^2)/16)=620,94$.

Затем, находится сумма квадратов отклонений по группам фактора опыта по формуле: $(Q_1=(\sum(X_i)^2-(\sum\sum X_{i,k})^2/k)/i)=63703-495^2/4/4=611,69$, где $k=4$ - число групп фактора; $i=4$ - число слагаемых в сумме по вариантам опыта (равное количеству повторностей). В данном случае, должно соблюдаться равенство $(N=k\cdot i)=4\cdot4=16$.

Сумма квадратов отклонений по повторностям опыта находится по формуле $(Q_2=\sum(X_k)^2-(\sum\sum X_{i,k})^2/i)/k)=(61269-495^2/4/4=3,19$.

Сумма квадратов отклонений по остаточному варьированию определяется из равенства (Q_3=Q-Q_1-Q_2)=620,94-611,68-3,18=6,08.

Результаты дисперсионного анализа *данных урожая ячменя приведены* ***в таблице*** *5.6.*

Таблица 5.6 Результаты однофакторного дисперсионного анализа

Варьирование данных	*Сумма квадратов отклонений, (Q)*	*Степень свободы, (ν)*	*Дисперсия, (σ^2)*	*Критерий Фишера, (F)*	
				$F_ф$	$F_т$
Общее по опыту	620,94	15	41,39	—	—
По вариантам опыта	611,68	3	203,89	304,31	8,81
По повторностям	3,18	3	1,05	1,56	8,81
Случайное (остаточное)	6,08	9	0,67	—	—

В таблицу вносятся рассчитанные суммы квадратов отклонений (Q, Q_1, Q_2, Q_3). Число степеней свободы определяется следующим образом: по общей сумме квадратов отклонений (ν=N-1)=16-1=15; *по вариантам опыта* (ν_1=n_1-1)=4-1=3; *по повторностям -* (ν_2=n_2-1)=4-1=3; *по остаточной сумме -* (ν_3=ν-ν_1-ν_2)=15-3-3=9.

Дисперсия определяется путем деления сумм квадратов отклонений (Q, Q_1, Q_2, Q_3) на соответствующие им числа степеней свободы (ν, ν_1, ν_2, ν_3), что можно выразить в общем виде формулой: (σ^2=Q/ν). *Фактический критерий Фишера* ($F_ф$) определяется путем деления каждой из величин дисперсий на значение остаточной дисперсии. *Критическое (табличное) значение критерия Фишера* ($F_ф$) находится по Приложению (таблица П.4.1) на пересечении значений большей и меньшей степеней свободы, которые устанавливаются по величине сравниваемых дисперсий.

Так как $F_ф$>$F_т$, то ***внесение добавок минерального грунта и торфа положительно влияет на величину урожая ячменя*** в исследуемых условиях.

5.2 Двухфакторный дисперсионный анализ

Если в дисперсионный анализ включить несколько факторов, влияющих на результативный признак, то они должны быть независимыми друг от друга. При обработке данных исходной информации, алгоритм расчетов аналогичен однофакторному дисперсионному анализу.

В качестве примера, *требуется определить влияние метеорологических условий (фактор I) и мелиоративных мероприятий (фактор II) на урожай биомассы трав в различных агроландшафтах.*

Решение

Здесь имеет место обработка данных с двумя факторами, каждый из которых делится на две группы. *Для этого составляется комбинационный (двухфакторный) дисперсионный комплекс (таблица 5.7).* ***Каждый фактор характеризуется тремя наблюдениями*** *(повторностями). Аналогичную схему можно использовать для двухфакторного анализа с большим числом групп и повторностей в каждом факторе.*

Таблица 5.7 Двухфакторный дисперсионный комплекс

Повторяемость опыта по фактору (II)	*Биомасса, кг/м²*		$\frac{\sum Y_i}{\sum Y_i^2}$	$(\sum Y_i)^2$	$\bar{Y}$
	Группы по фактору (I)				
	1982 год (сухой)	*1984 год (влажный)*			
1	*2*	*3*	*4*	*5*	*6*
Группа фактора (II) (неосушенный агроландшафт)					
Первая	*5/25*	*4/16*	*9/41*		
Вторая	*6/36*	*5/25*	*11/61*		
Третья	*5/25*	*6/36*	*11/61*		
∑/∑	*16/86*	*15/77*	*31/163*	*961*	*5,16*

Продолжение таблицы 5.7

Группа фактора (II) (осушенный агроландшафт)					
Первая	3/9	5/25	8/34		
Вторая	4/16	6/36	10/52		
Третья	4/16	6/36	10/52		
Σ/Σ	11/41	17/97	28/138	784	4,66
$\frac{\sum X_i}{\sum X_i^2}$	27/127	32/174	59/301	$\Sigma(\Sigma Y_i)^2=1745$	
$(\sum X_i)^2$	729	1024		$\Sigma(\Sigma X_i)^2=1753$	
$\overline{M}$	4,50	5,33		$\overline{M}_{общ.}=4,90$	

Примечание *: X_i - варианты опыта по фактору (I), Y_i - то же по фактору (II). В числителе - урожай биомассы, в знаменателе - квадрат чисел.*

Суть двухфакторного дисперсионного анализа *можно представить равенством*

$$Q=Q_1+Q_2+Q_3+Q_4+Q_5, \qquad (5.19)$$

где Q - общая сумма квадратов; Q_1 и Q_2 - соответственно, сумма квадратов отклонений для факторов (I) и (II); Q_3 - сумма квадратов отклонений, имеющих место при взаимодействии факторов (I) и (II); Q_4 - сумма квадратов отклонений по повторностям; Q_5 - остаточная сумма квадратов отклонений неучтенных факторов. ***Общая сумма квадратов отклонений*** *находится как -*

$$Q=(\Sigma\Sigma X_i^2, Y_i^2-(\Sigma\Sigma X_i, Y_i^2/N))=(301-(59^2/12))=10,92,$$

где N=12 - общий объем выборки ; ***сумма квадратов отклонений по фактору (I)*** *как -*

$$Q_1=(\Sigma(\Sigma X_i)^2-(\Sigma\Sigma X_i, Y_i)/n_X)/k_X=(1753-59^2/2)/6=2,08,$$

где $n_X=2$ - число групп фактора (1); $k_X=6$ - число вариант в каждой отдельной сумме; ***сумма квадратов отклонений по фактору (II)*** *вычисляется аналогично*

$$Q_2=(\Sigma(\Sigma Y_i)^2-(\Sigma\Sigma X_i, Yi/n_y))/k_y=(1745-59^2/2)/6=0,75;$$

сумма квадратов отклонений, вызываемых взаимодействием факторов (I) и (II), *определяется следующим образом*

$$Q_3=(\Sigma(\Sigma Z_i^2)-(\Sigma\Sigma X_i, Y_i)^2/nZ)/k_Z-Q_1-Q_2=(891-59^2/4)/3-2,08-0,75=4,08,$$

где $\Sigma(\Sigma Z_i^2)=(16^2+15^2+11^2+17^2)=891$ - сумма квадратов сумм значений вариант по группам выборки комбинационной таблицы; $n_Z=4$ - число сумм вариант по группам; $k_Z=3$ - число слагаемых вариант в каждой группе выборки; ***сумма квадратов отклонений по повторностям*** *(Q_4) определяется как-*

$$Q_4=(\Sigma(\Sigma X_i)^2-(\Sigma\Sigma X_i, Y_i)/n_{x,y})/k_{x,y}=(1171-59^2/3)4=2,67,$$

где $n_{x,y}=3$ - число сумм по повторностям; $k_{x,y}=4$ - число слагаемых в каждой сумме; сумма квадратов сумм (X_i), вычисленная как -

$$\Sigma(\Sigma X_i)^2=((5+4)+(3+5))^2+((6+5)+(4+6))^2+((5+6)+(4+6))^2=1171;$$

***сумма квадратов отклонений по остаточному варьированию** составляет -*

$$Q_5=Q-Q_1-Q_2-Q_3-Q_4=10,92-2,08-0,75-4,08-2,67=1,14;$$

число степеней свободы для (Q) будет - ($\nu = (N-1)$)=11; для (Q_1) и (Q_2) - соответственно, равно числу градаций фактора минус единица - ($\nu_1=(n_1-1)$)=(2-1)=1, ($\nu_2=(n_2-1)$)=(2-1)=1; для (Q_3) определится как - ($\nu_3=\nu_1.\nu_2$)=1·1=1; для (Q_4) - равно числу повторностей минус единица - ($\nu_4=(3-1)$)=2; для (Q_5) этот показатель определяется следующим образом - ($\nu_5=(\nu-\nu_2-\nu_3-\nu_4)$)=(11-1-1-1-2)=6.

Полученные расчетным путем характеристики сведены в таблицу 5.8.

Таблица 5.8 Результаты двухфакторного дисперсионного анализа.

Варьирование данных	Сумма квадратов отклонений (Q)	Степень свободы (ν)	Дисперсия (σ^2)	Критерий Фишера	
				F_ϕ	$F^T_{0.05}$
1	2	3	4	5	6
Общие по опыту	10,92	11	0,99	5,21	4,03
По фактору (I)	2,08	1	2,08	10,94	5,99
По фактору (II)	0,75	1	0,75	3,94	5,99
По взаимодействию факторов (I) и (II)	4,08	1	4,08	21,47	5,99
По повторностям	2,67	2	1,34	7,05	5,14
Остаточное	1,14	6	0,19	1,00	—

*Показатели дисперсии (таблица 5.8) вычисляются путем деления значений сумм квадратов отклонений на соответствующие значения степеней свободы. Фактический критерий Фишера определяется путем деления каждой из величин дисперсий на значение остаточной дисперсии. Критическое значение критерия Фишера найдено по Приложению (таблица П.4.1). **Анализируя критерии Фишера, можно заключить, что влияние исследуемых параметров на биомассу существенно (в целом по опыту, по фактору - I, по взаимодействию факторов и по повторностям, т.е. во всех рассмотренных случаях)** при вариации опытных данных $V=\frac{\sqrt{\sigma^2_{общ.}}}{M_{общ.}}\cdot 100\% = =\frac{\sqrt{0,99}}{4,9}\cdot 100\% = 20,0\ \%$ и $F_\phi > F^T_{0.05}$; **действие группы факторов (II) на исследованных агроландшафтах не доказано** ($F_\phi < F^T_{0.05}$).*

Примечание

Оценку результатов эксперимента можно сделать по критериям наименьшей существенной разности (НСР) и Стьюдента. Для вычисления (НСР) и (t) находится ошибка среднего арифметического ($m_м$) всего опыта и ошибка разности средних по формулам:

$$m_м = \sqrt{\sigma^2_{ост.} / N} = \sqrt{0{,}19 / 12} = 0{,}1258;$$

$$m_d = \sqrt{2 \cdot \sigma^2_{ост.} / n} = \sqrt{2 \cdot 0{,}19 / 6} = 0{,}25;$$

$$НСР = m_d \cdot t^T_{0.05} = 0{,}25 \cdot 2{,}45 = 0{,}61,$$

в которых n - численность меньшей из сравниваемых частных групп. По критерию Стьюдента сравниваются средние арифметические данные по осушенному и неосушенному агроландшафтам

$$t = (M_{у.1} - M_{у.2}) / m_{d_y} = (5{,}16 - 4{,}66) / 0{,}25 = 2{,}00.$$

По Приложению (таблица П.2) критерий Стьюдента - ($t^T_{0.05}$=2,45), при (P=0,95) для (ν=6). Таким образом, на биомассу трав в агроландшафтах не влияет мелиорация (т.е. фактор II), так как ($t_Ф$=2,0<$t^T_{0.05}$=2,45), при (P=0,95); метеорологические условия (фактор I) достоверно влияют на биомассу трав, при (P=0,95). ***Выводы, сделанные с использованием критериев Фишера и Стьюдента, совпадают. В заключение*** *обычно определяется* ***точность опыта****, которая составляет*

$$P = (m_м / M_{общ.}) \cdot 100\% = 0{,}1258/4{,}9 \cdot 100\% = 2{,}56\%.$$

Точность опыта ***признается достаточно высокой****, поскольку (P<3%).* ***Коэффициент варьирования опытных данных -***

$$V = \frac{\sqrt{\sigma^2_{общ.}}}{M_{общ}} \cdot 100\% = \frac{\sqrt{0{,}99}}{4{,}9} \cdot 100\% = 20{,}0\%,$$

также ***незначителен****, что удовлетворяет требованиям опыта.*

Рассмотрим второй пример *двухфакторного дисперсионного анализа, в котором* ***требуется установить, значимо ли различие в действии форм азотных удобрений на урожай овсяницы луговой*** *(таблица 5.9). Нулевая гипотеза (H_0: d=0).*

Решение

Особенностью по-вариантной обработки данных вегетационного опыта с разной повторностью является необходимость вычисления нескольких значений наименьшей существенной разности, так как не все средние равноточны. В примере варианты (1–2) имеют четыре, а варианты (3–4) – шесть наблюдений. В установленном порядке, выполняются следующие вычислительные операции:

1) ***определяются суммы урожаев и средние по вариантам, общая сумма и средний урожай по опыту*** *(таблица 5.9);*

Таблица 5.9 Урожай овсяницы (г на сосуд)

Варианты (формы азота)	Урожай (X)						Число наблюдений (n)	Суммы (V)	Средние
1	16,0	17,2	14,4	15,8	–	–	4	63,4	15,85
2	29,4	30,4	30,3	28,1	–	–	4	118,2	29,55
3	26,0	29,2	26,7	27,1	26,0	28,1	6	163,1	27,18
4	25,3	24,8	26,1	23,2	25,7	24,0	6	149,1	24,85
Общая сумма							20=Σn= =N	493,8= =ΣX	24,69=$\bar{x}$

*2) **преобразуются исходные даты по соотношению** ($X_1=X-A$), при условной средней (A), принимаемой равной числу 25, - близкому к среднему урожаю по опыту ($\bar{X}$ =24,69) (таблица 5.10), с последующим вычислением сумм квадратов отклонений.*

Таблица 5.10 Преобразованные данные

Варианты	$X_1=X-25$						Суммы (V)
1	–9,0	–7,8	–10,6	–9,2	–	–	–36,6
2	4,4	5,4	5,3	3,1	–	–	18,2
3	1,0	4,2	1,7	2,1	1,0	3,1	13,1
4	0,3	–0,2	1,1	–1,8	0,7	–1,0	–0,9
Общая сумма							–6,2=ΣX_1

При по-вариантном вычислении сумм квадратов отклонений необходимо иметь ввиду, что в суммы (V) входит разное число наблюдений (n). Далее, последовательно определяются:

а) общее число наблюдений ($N=\Sigma n=20$;

б) корректирующий фактор ($C=(\Sigma X_1)^2 : N=(-6,2)^2 : 20=0,07$;

в) суммы квадратов отклонений

$$(C_y = \Sigma X_i^2 - C) = (9,0^2 + 7,8^2 + \ldots + 1,0^2 - 1,92 = 465,70;$$

$$C_v = \Sigma\left(\frac{V_1^2}{n_1} + \frac{V_2^2}{n_2} + \ldots + \frac{V_l}{n_l}\right) - C = \left(\frac{36,6^2}{4} + \frac{18,2^2}{4} + \frac{13,1^2}{6} + \frac{0,9^2}{6}\right) - 1,92 = 444,51;$$

$$C_z = C_y - C_v = 465,70 - 444,51 = 21,19.$$

Вычисленные суммы квадратов отклонений вносятся в таблицу 5.11, наряду с другими составляющими дисперсионного анализа.

Таблица 5.11 Результаты дисперсионного анализа

Дисперсия	*Сумма квадратов*	*Степени свободы*	*Средний квадрат*	$F_ф$	F_{05}
Общая	*465,70*	*19*	–	–	–
Вариантов	*444,51*	*3*	*148,80*	*112,7*	*3,24*
Остаток (ошибки)	*21,19*	*16*	*1,32*	–	–

Значение (F_{05}) принимается по Приложению (таблица П.4.1) для 3-х степеней свободы дисперсии вариантов (числитель) и 16-ти степеней свободы остатка (знаменатель). Так как ($F_ф > F_{05}$), то между вариантами опыта имеются существенные различия на 5%-ном уровне значимости и гипотеза (H_0) отвергается;

*3) при оценке существенности частных **различий в опыте** с разной повторностью **необходимо учесть неравноточность сравнения средних**. Ошибки средних первых двух вариантов ($\bar{x}_1$ и $\bar{x}_2$) предопределяются числом наблюдений ($n_1=n_2$)=4, а двух последних ($\bar{x}_3$ и $\bar{x}_4$), соответственно, ($n_3=n_4$)=6 наблюдений. Поэтому, ошибку разности между средними нужно определять по формуле, учитывающей разную повторность по (1-2) и (3-4) вариантам*

$$s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{s^2 \frac{n_1 + n_2}{n_1 n_2}}.$$

Тогда:

а) ошибка разности средних, при сравнении $\bar{x}_1$ с $\bar{x}_2$ ($n_1=n_2=4$), будет -

$$s_d' = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{2 \cdot 1,32}{4}} = 0,81 \text{ г / сосуд},$$

при сравнении $\bar{x}_1$ и $\bar{x}_2$ с $\bar{x}_3$ и $\bar{x}_4$ ($n_1=4$ и $n_2=6$), будет -

$$s_d'' = \sqrt{s^2 \frac{n_1 + n_2}{n_1 n_2}} = \sqrt{1,32 \cdot \frac{4+6}{4 \cdot 6}} = 0,74 \text{ г / сосуд},$$

при сравнении $\bar{x}_3$ с $\bar{x}_4$ ($n_3=n_4=6$), будет -

$$s_d''' = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{2 \cdot 1,32}{6}} = 0,66 \text{ г / сосуд};$$

б) наименьшая существенная разность (НСР) для 5%-ного (или 1%-ного) уровня значимости определяется как -

$$НСР'_{05} = t_{05}s'_d = 2,12 \cdot 0,81 = 1,72 \text{ г / сосуд;}$$

$$НСР''_{05} = t_{05}s''_d = 2,12 \cdot 0,74 = 1,57 \text{ г / сосуд;}$$

$$НСР'''_{05} = t_{05}s'''_d = 2,12 \cdot 0,66 = 1,40 \text{ г / сосуд.}$$

Значения критерия (t_{05}=2,12) принимаются из Приложений (таблица П.2) для 16-ти степеней свободы дисперсии ошибки (остатка). Результаты опыта и статистической обработки заносятся в таблицу 5.12.

.Таблица 5.12 Различия в действии форм азотных удобрений на урожай овсяницы луговой (г на сосуд)

Варианты	Урожай	Сравнение с контролем		Сравнение с аммиачной селитрой	
		разность	$НСР_{05}$	разность	$НСР_{05}$
Без удобрений (контроль)	15	–	–	–11,4	1,57
Сульфат аммония	29,6	13,8	1,72	2,4	1,57
Аммиачная селитра	27,2	11,4	1,57	–	–
Мочевина	24,8	9,0	1,57	–2,4	1,40

***Результаты анализа показывают**, что все формы азотных удобрений существенно повышают урожай овсяницы луговой. Аммиачная селитра и мочевина примерно равноценны по эффективности; сульфат аммония обеспечивает статистически значимый на 5%-ном уровне эффект в сравнении с аммиачной селитрой.*

Рассмотрим еще один пример ***двухфакторного дисперсионного анализа, в котором по данным учета числа зерен в колосе у гибридов ячменя (таблица 5.13) необходимо вычислить коэффициент наследуемости.***

Решение

*1) **проводится дисперсионный анализ** для двухфакторного комплекса:*

$$N = l_A l_B n = 3*2*4 = 24;$$

$$C = (\Sigma X)^2 : N = 575^2 : 24 = 13776,04;$$

$$C_1 = \Sigma X^2 - C = (20^2 + 20^2 + ... + 32^2) - 13776,04 = 428,96;$$

$$C_P = \Sigma P^2 : l_A l_B - C = (141^2 + 143^2 + 145^2 + 146^2) : 3*2 - 13776,04 = 2,46;$$

$$C_V = \Sigma V^2 : n - C = (83^2 + 76^2 + ... + 123^2) : 4 - 13776,04 = 406,71;$$

$$C_Z = C_1 - C_P - C_V = 428,96 - 2,46 - 406,71 = 19,79;$$

*2) **для выполнения оценки существенности действия материнских и отцовских форм и их взаимодействия на результативный признак гибридов составляется таблица 5.14**, в*

Таблица 5.13 Число зерен в колосе у гибридов ячменя

Материнская форма (A)	Отцовская форма (B)	Повторения (X)				Суммы (V)	Средние
		1	2	3	4		
1	2	3	4	5	6	7	8
a_1	b_1	20	20	22	21	83	20,8
	b_2	18	19	20	19	76	19,0
a_2	b_1	23	23	20	22	88	22,0
	b_2	22	23	24	24	93	23,2
a_3	b_1	29	28	27	28	112	28,0
	b_2	29	30	32	32	123	30,8
Суммы (P)		141	143	145	146	575=ΣX	

которую вписываются суммы (V) для каждого гибрида и находятся суммы по факторам (A) и (B).

Таблица 5.14 Вычисление сумм для определения эффектов (A), (B) и взаимодействия (AB)

Фактор (A)	Фактор (B)		Суммы (A)
	b_1	b_2	
a_1	83	76	159
a_2	88	93	181
a_3	112	123	235
Суммы (B)	283	292	575=ΣX

*3) **по результатам дисперсионного анализа (таблица 5.15) имеем**: общее варьирование (C_{A+B+AB}) – внутренняя часть таблицы (численно $C_{A+B+AB}=C_v=406,71$ и вычислено нами ранее), варьирование факторов (A) и (B). Взаимодействие (AB) находится по разности: ($C_A=\Sigma A^2 : l_B \cdot n - C$)=($159^2+181^2+235^2$):2*4–13776,04=382,34, при (l_A–l)=(3–1)=2 - -степенях свободы; ($C_B=\Sigma B^2:l_A \cdot n-C$)=($283^2+292^2$):3*4–13776,04=3,38, при (l_B–l)= =(2–1)=1 - степени свободы; ($C_{AB}=C_{A+B+AB}-C_A-C_B$)=406,71–382,34–3,38=20,99, при (l_A–l) (l_B–l)=(3-1) (2–1)=2 - степенях свободы; существенность действия и взаимодействия факторов оценивается по критерию (F) (Приложение, таблица П.4.1).*

В рассматриваемом примере существенным оказалось действие (А) (материнских форм) и взаимодействие (АВ) (материнская формахотцовская форма). Следовательно, имеет смысл вычислить коэффициенты наследуемости, характеризирующие силу генетического влияния материнских форм и взаимодействия; вариабельность числа зерен в колосе у гибридов ячменя не зависит существенно от отцовских форм ($F_\phi < F_{05}$);

Таблица 5.15 Результаты дисперсионного анализа

Дисперсия	*Сумма квадратов*	*Степени свободы*	*Средний квадрат*	F_ϕ	F_{05}
Общая	*428,96*	*23*	-	-	-
Повторений	*2,46*	*3*	-	-	-
Материнских форм (А)	*382,34*	*2*	*191,17*	*144,83*	*3,60*
Отцовских форм (В)	*3,38*	*1*	*3,38*	*2,56*	*4,54*
Взаимодействия (АВ)	*20,99*	*2*	*10,50*	*7,95*	*3,60*
Остаток	*19,79*	*15*	*1,32*	-	-

*4) **в двухфакторном комплексе дисперсия групповых средних имеет более сложную природу**, чем в однофакторном, и определяется как генетической изменчивостью, обусловленной генотипами материнских и отцовских форм и их взаимодействием, так случайной изменчивостью (остаток). **Общий коэффициент наследуемости в этом случае равен***

$$h^2 = h^2_A + h^2_B + h^2_{AB}.$$

*Для уяснения сущности вычислительных операций при определении дисперсий, характеризующих влияние на фенотипическую изменчивость генотипов материнских форм (s^2_A), отцовских форм (s^2_B) и их взаимодействия (s^2_{AB}), **целесообразно дополнительно рассмотреть схему компонентного двухфакторного анализа результатов эксперимента**..*

В настоящем примере, существенным оказалось влияние материнских форм (А) и взаимодействия (АВ).

Однако, при подборе пар для скрещивания, необходимо иметь ввиду, что, как показал проведенный нами численный эксперимент, проявление результативного признака в гибридах зависит в основном (на 86 %) от материнского растения.

6 СТАТИСТИЧЕСКИЕ МЕТОДЫ ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТА

Эксперименты, поставленные таким образом, что в каждом опыте варьируются одновременно все независимые переменные (факторы) по специальному плану, *называются активными. В отличие от* обычных, традиционных, *пассивных* экспериментов, в активных экспериментах в каждом отдельном опыте *варьируется только один фактор.*

Статистические методы планирования активного эксперимента являются одним из эмпирических способов получения математического описания статики сложных объектов исследования, т.е. уравнения связи отклика объекта (Y) и независимых управляемых нормированных входных переменных (факторов) - $(\bar{Z}^T = (z_1; z_2; \ldots; z_n))$. При этом, математическое описание представляется в виде некоторого полинома- отрезка ряда Тейлора, в который разлагается неизвестная зависимость основной точки $(\bar{Z}_0)$

$$M\{Y\} = \varphi(z_1, z_2, \ldots, z_n) = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot z_i + + \sum_{\substack{i,l=1 \\ i<l}}^{n} \beta_{il} \cdot z_i \cdot z_l + \sum_{i=1}^{n} \beta_{ii} \cdot z_i^2 + \ldots, \qquad (6.1)$$

где $\beta_i = \left.\frac{\partial \varphi}{\partial z_i}\right|_{\bar{Z} = \bar{Z}_0}$; $\beta_{il} = \left.\frac{\partial^2 \varphi}{\partial z_i \cdot \partial z_l}\right|_{\bar{Z} = \bar{Z}_0}$; $\beta_{ii} = \frac{1}{2} \cdot \left.\frac{\partial^2 \varphi}{\partial z_i^2}\right|_{\bar{Z} = \bar{Z}_0}$ -

теоретические коэффициенты. Из-за наличия неуправляемых и даже неконтролируемых факторов, изменение величины (Y) носит случайный характер, поэтому функциональная зависимость $(\varphi(\bar{Z}))$ не дает точной связи между управляемыми факторами и откликом (Y_g) объекта в каждом g-ом опыте, а лишь между управляемыми факторами и математическим ожиданием случайной величины (Y)

$$M\{Y_g\} = \varphi(\bar{Z}_g^T) \; . \qquad (6.2)$$

Здесь $\bar{Z}_g^T = (z_{1g}; z_{2g}; \ldots; z_{ng})$ - g-ая точка пространства независимых управляемых факторов (факторного пространства). В таком случае, по результатам эксперимента можно отыскать уравнение регрессии в форме некоторого полинома

$$\hat{Y} = b_0 + \sum_{i=1}^{n} b_i \cdot z_i + \sum_{\substack{i,l=1 \\ i<l}}^{n} b_{il} \cdot z_{il} + \sum b_{ii} \cdot z_i^2 + \ldots \; , \qquad (6.3)$$

где b_0, b_i, b_{ii}, b_{ii}, - выборочные коэффициенты регрессии, которые являются лишь оценками для теоретических коэффициентов, соответственно, (β_0, β_i, β_{ii}, β_{ii},...); $\hat{Y}$ - оценка для $M\{Y\}$; $Z_g^T (g = 1,2,\ldots, N)$ - точки факторного пространства, в которых проводится эксперимент. Задача отыскания оценок коэффициентов уравнения регрессии (6.3) по результатам опытов в (N) - точках факторного пространства является типичной задачей множественного регрессионного анализа в том случае, если имеются следующие предпосылки:

1) результаты наблюдений (y_1, y_2,...,y_N) отклика в (N) - точках факторного пространства представляют собой независимые нормально распределенные случайные величины, т.е. на них воздействуют нормально распределенные случайные помехи (ε) с нулевым математическим ожиданием - ($M\{\varepsilon\}=0$);

2) дисперсии ($\sigma^2\{Y_g\}$), при (g=1,2,...,N) равны; это значит, что, получаемые при проведении многократных повторных наблюдений над величиной (Y_g) в точках (Z_g^T), выборочные оценки ($S_g^2\{Y\}$) однородны, дисперсия же ($\sigma^2\{Y_g\}$) не зависит от математического ожидания ($M\{Y_g\}$), т.е. не отличается от дисперсии ($\sigma^2\{Y_g\}$), полученной при повторных наблюдениях в любой другой точке (Z_g^T) факторного пространства (воспроизводимость с равной точностью);

3) независимые управляемые факторы (z_1, z_1,...,z_N) измеряются с пренебрежимо малыми ошибками по сравнению с ошибкой в определении (Y) (имеется в виду влияние их ошибок на величину "Y" по сравнению с влиянием неуправляемых и неконтролируемых факторов - "ε".

Активные эксперименты обладают следующими **преимуществами**:

1) поскольку план экспериментов составляется заранее, перед началом опытов, то он может максимально способствовать упрощению последующей обработки материалов экспериментов при построении регрессионных моделей;

2) оптимальное использование факторного пространства при активном экспериментировании позволяет при минимальных затратах (минимуме экспериментов) получать максимум информации об изучаемых явлениях;

3) при планировании экстремальных экспериментов, кроме аппроксимации функции отклика, можно попутно решить более важные для практи-

ки задачи - поиска экстремума в (P)-мерном факторном пространстве или / и оптимального управления процессами;
4) методы планирования экспериментов позволяют опытным путем проранжировать факторы по степени их влияния на функцию отклика;
5) планируемые эксперименты дают возможность математически описать более сложные процессы, формализовать (методами дисперсионного анализа) изучаемые явления, в том числе с использованием косвенных факторов;
6) планирование эксперимента позволяет изучать и математически описывать процессы и явления при неполном знании их внутреннего механизма.

Рассмотрим основные понятия и определения теории планирования эксперимента.

Объект исследования (ОИ) - носитель некоторых неизвестных и подлежащих изучению свойств и качеств. Как правило, **любой объект исследования можно представить в виде "черного ящика" с определенным количеством входов и выходов**. При этом, **традиционно выделяются**:

1) входные контролируемые и управляемые переменные, которые варьируются исследователем по своему усмотрению - вектор ($X = \|x_1, x_2, \ldots, x_n\|$);

2) входные контролируемые, но неуправляемые переменные - вектор ($Z = \|z_1, z_2, \ldots, z_k\|$);

3) неуправляемые и неконтролируемые переменные - вектор ($E = \|e_1, e_2, \ldots, e_j\|$);

4) выходные показатели - вектор ($Y = \|y_1, y_2, \ldots, y_j\|$).

Переменные (X) и (Z) принято называть *факторами*, а пространство контролируемых переменных образует *факторное пространство*. Выходную переменную (Y) - зависимую переменную объекта - часто называют *откликом*; зависимость отклика от рассматриваемых факторов - *функцией отклика*, а геометрическое представление функции отклика - *поверхность отклика*. *Эксперимент* - система операций, воздействий и наблюдений, направленных на получение информации об объекте при исследовательских работах. *Опыт* - воспроизведение исследуемого явления в определенных условиях проведения эксперимента при возможности регистрации его результатов. *План эксперимента* - совокупность данных, определяющих число, условия и порядок реализации опытов. *Планирование экс-*

перимента - выбор плана эксперимента, удовлетворяющего заданным требованиям. *Область действия* - область возможных значений факторов (X) при экспериментировании. *Область планирования* - область значений факторов (X), в которой находятся точки, отвечающие условиям проведения опытов согласно плану эксперимента. *Точка плана* - упорядоченная совокупность численных значений факторов, соответствующая условиям проведения опыта, точке факторного пространства, в которой проводится эксперимент. Точке плана с номером (Y) отвечает вектор ($X_g^T = \|x_{1g}, x_{2g}, \ldots, x_{nj}\|$). Общая совокупность таких векторов (X_g=..., g=1,2,...,L) образует *план эксперимента,* а совокупность различных векторов - *спектр плана*. Фиксированное значение фактора называется *уровнем фактора*. Факторы могут различаться по числу уровней, на которых возможна их фиксация в данной задаче. Понятие уровня фактора часто используется при описании характерных точек из области действия фактора (x_i): минимальный ($X_{i\,min}$) и максимальный ($X_{i\,max}$) уровни, основной уровень фактора (x_i^0), при (i=1,2,...,n). Обычно вектор ($X^{0T} = \|x_1^0, x_2^0, \ldots, x_n^0\|$) задает в факторном пространстве точку, являющуюся, в каком-то смысле, центром области планирования, центром эксперимента. В ее окрестностях и располагаются все точки плана. Часто координаты (x_i^0) выбираются с помощью соотношения ($x_i^0=(X_{i\,max}+X_{i\,min})/2$). В ряде случаев, полезно понятие *интервала (шага) варьирования* фактора (x_i): ($\Delta x_i=(X_{i\,max}-X_{i\,min})/2$), так что ($X_{i\,max}=x_i^0+\Delta x_i$); ($X_{i\,min}=x_i^0-\Delta x_i$). Зная ($x_i^0$), ($\Delta x_i$), при (i=1,2,...,n), можно реализовать операцию нормализации факторов. Операция нормализации сводится к изменению начала отсчета координатных осей и масштаба, в соответствии с соотношением ($x_i=(x_i-x_i^0)/\Delta x_i$, при (i=1,2,...,N)). Таким образом, для переменной (x_i), в стандартизированном масштабе, начало координат совмещено с центром эксперимента, а в качестве единицы измерения используется шаг варьирования фактора. Применение безразмерных переменных (факторов) - (x_i) часто существенно облегчает математические выкладки и запись конечных результатов. *Принципы*, положенные в основу *теории планирования эксперимента*, направлены на повышение эффективности экспериментирования, т.е. на *получение максимума информации при минимуме опытов*:

1) отказ от полного перебора возможных входных состояний

Для получения исчерпывающей информации о свойствах функции отклика, вообще, необходимо проведение бесконечного числа опытов во всех точках области планирования эксперимента. В противном случае, всегда существует теоретическая возможность пропустить некоторую особенность поверхности отклика. Ясно, что это носит чисто умозрительный, гипотетический характер, и не реализуемо на практике. Экспериментатор просто вынужден задаться дискретной сеткой значений факторов, выбрать какое-то фиксированное число уравнений каждого фактора. Коль это так, то экспериментатор, желая того или нет , фактически задается некоторыми свойствами поверхности отклика и, тем самым, постулирует определенную степень гладкости этой поверхности. В теории планирования эксперимента сознательно отказываются от полного перебора входных состояний или от эксперимента, близкого к нему по своей конструкции. Выбор числа уровней варьирования по каждому фактору непосредственно связывается с выбором вида функции отклика. В свою очередь, сам этот выбор можно осуществить, используя принцип постепенного усложнения математической модели;

2) постепенное усложнение математической модели

Этот принцип очень прост: в отсутствие априорной информации о свойствах функции отклика, нет смысла сразу строить сложную модель объекта. Получение сложной модели требует, естественно, большего числа опытов, чем модели простой, и может оказаться, что в сложной модели нет необходимости, так как она вырождается в простую модель (поскольку таковы свойства объекта). Поэтому, теория планирования эксперимента рекомендует, как правило, начинать с простейшей модели, соответствующей имеющейся априорной информации. Согласно этой концепции, при проведении эксперимента необходимо использовать последовательную, шаговую стратегию. После каждого шага, производится анализ результатов, затем, принимается решение о дальнейшей деятельности. Исследователь отказывается от попытки заранее задать строго фиксированную схему проведения эксперимента. Такая стратегия предусматривает возможность принятия решений в зависимости от результатов, полученных на отдельных этапах исследования. Не менее важным является и вопрос о проверке качества результатов на каждом этапе исследования. В теории планирования эксперимента для проверки исполь-

зуются различные, математически строго обоснованные, статистические процедуры, вытекающие из некоторых вероятностных свойств шума;

3) сопоставление с шумом

Точность полученной модели обязательно должна быть сопоставлена с интенсивностью случайной помехи, воздействующей на результат измерения отклика (Y). При прочих равных условиях, чем меньше уровень помехи, тем более точной должна быть модель; чем выше уровень помехи, тем в большей степени можно ожидать, что более простая модель окажется работоспособной. Поскольку многие реальные объекты характеризуются высоким уровнем помех, при их описании получили наибольшее распространение полиномиальные регрессионные модели, причем в подавляющем числе случаев порядок такой модели равен (1) или (2). Подобные модели широко используются при создании различного рода методик инженерных расчетов тех или иных устройств, схем и агрегатов, так как необходимая точность расчетов обычно весьма невелика (порядка 5...15%). Конечно, с теоретической точки зрения, для выявления физической сущности механизма явления подобные полиномиальные модели менее содержательны, чем теоретические модели типа дифференциальных уравнений. Однако, с практических позиций, полученные полиномиальные модели являются весьма эффективными, порой, единственными средствами изучения сложных объектов;

4) рандомизация (приведение к случайности)

Этот принцип состоит в такой организации эксперимента, которая позволяет сделать случайными (рандомизировать) систематически действующие переменные, не поддающиеся (поддающиеся с трудом) учету и контролю, для того чтобы можно было рассматривать их как случайные величины и, следовательно, учитывать статистически. Иными словами, не в силах учесть действие неслучайных переменных, исследователь искусственно создает в эксперименте случайную ситуацию, т.е. переводит такие переменные в разряд случайных, избавляясь от возможных систематических ошибок в конечных результатах. При этом, уровень шумового поля увеличивается, что, однако, особой роли не играет. При различных статистических исследованиях принцип рандомизации предусматривает чисто случайный выбор элементов для последующего анализа из общей совокупности, подлежащей изучению. Тем самым, обеспечивается представительность полученной выборки, т.е. гарантируется возмож-

ность, с помощью измерения свойств конечного набора элементов из совокупности, высказать обоснованное суждение о свойствах совокупности, в целом. При проведении различного рода экспериментов принцип рандомизации предусматривает случайный порядок реализации опытов, т.е. случайный порядок реализации строк матрицы плана. Использование такого простейшего приема, как рандомизация опытов, позволяет устранить в математических моделях те или иные смещения, вызванные действием неконтролируемых систематических переменных;

5) оптимальность планирования эксперимента

Этот принцип является центральным в теории планирования эксперимента. В соответствии с ним, план эксперимента должен обладать некоторыми оптимальными свойствами, с точки зрения определенного, заранее выбранного, критерия оптимальности плана или совокупности подобных критериев. Критерии оптимальности планов могут формулироваться по-разному. Фактически, в таких критериях, в строгой математической форме, представлены (формализованы) те или иные интуитивные соображения специалистов - экспериментаторов о "хорошем", качественном эксперименте. При этом, общая направленность теории планирования - "меньше опытов - больше информации, выше качество результатов" - сохраняется. Конкретная форма критерия зависит, прежде всего, от типа решаемой задачи, назначения плана, хотя, даже в рамках одного типа задач, могут быть использованы различные критерии.

6.1 Полный факторный эксперимент

Для построения линейных и неполных степенных математических моделей применяется полный факторный эксперимент, обладающий ортогональной матрицей планирования. Математическое описание поверхности отклика объекта в окрестности точки базового режима-($\vec{X}_0^T = (x_{10}; x_{20}; \ldots; x_{n0})$) можно получить варьированием каждого из факторов (x_i) на двух уровнях, отличающихся от базового уровня (x_{i0}) на величину интервала варьирования (Δx_i). Интервал варьирования по каждому управляемому фактору выбирается так, чтобы приращение величины отклика (Y) к базовому значению (Y_0), при реализации ($x_{i0} \pm \Delta x_i$), можно было выделить на фоне "шума" при небольшом числе параллельных опытов.

Полным факторным экспериментом (ПФЭ) называется эксперимент, реализующий все возможные неповторяющиеся комбинации уравнений (n) - независимых управляемых факторов, каждый из которых варьируется на двух уровнях. Числом этих комбинаций ($N=2^n$) определяется тип ПФЭ. Для упрощения, дальнейшее изложение строится на примере планирования типа ($N=2^3$), т.е. на примере объекта с тремя (n=3) - независимыми управляемыми факторами (x_1,x_2,x_3). При планировании эксперимента проводится преобразование размерных управляемых независимых факторов (x_i) в безразмерные, нормированные

$$z_i = (x_i - x_{i0}) / \Delta x_i \,. \qquad (6.4)$$

Это дает возможность легко построить ортогональную матрицу планирования и значительно облегчить дальнейшие расчеты, так как, в этом случае, верхние и нижние уровни варьирования ($z_{iв}$) и ($z_{iн}$) в относительных единицах, равны соответственно, (+1) и (-1) независимо от физической природы факторов, значений основных уровней и интервалов варьирования факторов (Δx_i). Если для трехфакторной задачи теоретическое уравнение регрессии относительно нормированных факторов имеет вид -

$$M\{Y\} = \beta_0 + \sum_{i=1}^{3} \beta_i \cdot z_i + \sum_{\substack{i;l=1 \\ i<l}}^{3} \beta_{il} \cdot z_i \cdot z_l + \beta_{123} \cdot z_1 \cdot z_2 \cdot z_3 \,, \qquad (6.5)$$

когда степенями факторов выше первой можно пренебречь, то ПФЭ дает возможность найти раздельные оценки коэффициентов (β_i). Так как изменение выходной величины (Y) носит случайный характер, то имеется возможность определить лишь выборочные коэффициенты регрессии (b_i), (b_{il}) для оценивания теоретических коэффициентов (β_i), (β_{il}). Процесс нахождения модели (идентификации) методом ПФЭ состоит из : 1) планирования эксперимента, 2) проведения эксперимента на объекте исследования, 3) проверки воспроизводимости (однородности выборочных дисперсий "S_g^2") эксперимента, 4) получения математической модели объекта с проверкой статистической значимости выборочных коэффициентов регрессии, 5) проверки адекватности математического описания. *В таблице 6.1 даются*: запись всех априорных сведений об уравнении регрессии, запись базовых уровней, шагов варьирования, верхних и нижних уровней управляемых факторов, матрицы планирования, результатов наблюдений эксперимента, промежуточных и конечных результатов рас-

чета для проверки воспроизводимости эксперимента, проверки значимости коэффициентов, проверки адекватности математического описания, а *в таблице 6.2 - приведена матрица планирования полного факторного эксперимента* (ПФЭ):

1) планирование полного факторного эксперимента

Матрица планирования ПФЭ представляется в виде таблицы, составляемой по следующим правилам (таблица 6.2) -

а) каждая (g)-ая строка матрицы содержит набор координат (z_{ig}) точки, в которой проводится (g)-ый опыт (i=1,2,...,n ; g=1,2,...,N);

б) вводится фиктивная переменная (Z_0=+1);

в) поскольку переменные (z_i) принимают лишь значения (+1) и (-1), то все взаимодействия - ($z_i \cdot z_l$(i; l=1,2,3;)) могут принимать только такие же значения;

г) в первой строке (g=1) все управляемые факторы выбираются на нижнем уровне, т.е. (z_i=-1); последующие (g)-ые варианты варьирования при составлении матрицы планирования выбираются так: при построчном переборе всех вариантов, частота смены знака факторов для каждого последующего фактора (z_{i+1}) вдвое меньше, чем для предыдущего (z_i); три столбца управляемых факторов образуют собственно план эксперимента (обведено жирной чертой), а остальные столбцы матрицы планирования получаются перемножением соответствующих значений управляемых факторов и необходимы для расчета соответствующих коэффициентов при взаимодействиях.

Таблица 6.1 Этапы планирования и характеристики полного факторного эксперимента

Факторы	X_1	X_1	X_2	X_3	Априорные сведения	Оценка коэффициентов регрессии	
Базовый уровень	X_{10}				$\beta_{12}=0$	$b_0 \to \beta_0$	$b_{12} \to \beta_{12}$
Интервал варьирования	ΔX_1				$\beta_{13}=0$	$b_1 \to \beta_1$	$b_{13} \to \beta_{13}$
Верхний уровень	$X_{1в}$					$b_2 \to \beta_2$	$b_{23} \to \beta_{23}$
Нижний уровень	$X_{1н}$				$\beta_{23}=0$	$b_3 \to \beta_3$	$b_{123} \to \beta_{123}$

Матрица планирования и результаты расчета оценок коэффициентов регрессии											Результаты наблюдений и проверки воспроизводимости эксперимента					Проверка адекватности модели	
q	K_1	K_2	K_3	Z_0	Z_1	Z_2	Z_3	Z_1Z_2	Z_1Z_3	Z_2Z_3	$Z_1Z_2Z_3$	Y_{gk_1}	Y_{gk_2}	Y_{gk_3}	$\overline{Y}_g$	S_g^2	$\hat{Y}_g$	$(\overline{Y}_g - \hat{Y}_g)^2$
1	21	12	3	+1	-1	-1	-1	+1	+1	+1	-1							
2	23	14	7	+1	+1	-1	-1	-1	-1	+1	+1							
3	8	13	4	+1	-1	+1	-1	-1	+1	-1	+1							
4	5	16	19	+1	+1	+1	-1	+1	-1	-1	-1							
5	10	24	17	+1	-1	-1	+1	+1	-1	-1	+1							
6	15	22	6	+1	+1	-1	+1	-1	+1	-1	-1							
7	2	1	11	+1	-1	+1	+1	-1	-1	+1	-1							
8	9	20	18	+1	+1	+1	+1	+1	+1	+1	+1							
Оценки коэффициентов регрессии				b_0	b_1	b_2	b_3	b_{12}	b_{13}	b_{23}	b_{123}	$\frac{\sum_{g=1}^{g} S_g^2}{m \; a \; x \; \{S_g^2\}}$					$\frac{\sum \overline{Y}_g - \hat{Y}_g)^2}{S_{ad}^2}$	
												G					F	
Проверка значимости оценок коэффициентов регрессии												$q_{табл}[\%]$			5%		$q_{адд}[\%]$	5%
$q_{табл}[\%]$	5%	$S^2\{b\}$										$\nu_{1табл}$			7		$\nu_{1адд}$	
$t_{табл}$	2.119	$S\{b\}$										$\nu_{2табл}$			8		$\nu_{2адд}$	
S_{bi}^2		Вывод										G			0.5157		$F_{табл}$	
Уравнение регрессии (неполная степенная модель)											d	Вывод					Вывод	
$\hat{Y} = b_0 + b_1Z_1 + b_2Z_2 + b_3Z_3 + b_{12}Z_1Z_2 + b_{13}Z_1Z_3 + b_{23}Z_2Z_3 + b_{123}Z_1Z_2Z_3$																		

Таблица 6.2 Матрица планирования полного факторного эксперимента

g	Z_0	Z_1	Z_2	Z_3	$Z_1 \cdot Z_2$	$Z_1 \cdot Z_3$	$Z_2 \cdot Z_3$	$Z_1 \cdot Z_2 \cdot Z_3$
1	+1	-1	-1	-1	+1	+1	+1	-1
2	+1	+1	-1	-1	-1	-1	+1	+1
3	+1	-1	+1	-1	-1	+1	-1	+1
4	+1	+1	+1	-1	+1	-1	-1	-1
5	+1	-1	-1	+1	+1	-1	-1	+1
6	+1	+1	-1	+1	-1	+1	-1	-1
7	+1	-1	+1	+1	-1	-1	+1	-1
8	+1	+1	+1	+1	+1	+1	+1	+1

План ПФЭ типа 2^4 (n=4) можно построить либо указанным выше способом, либо на базе плана ПФЭ типа 2^3, повторив его дважды: один раз - при величине (z_4=-1), второй раз - при (z_4=+1). Аналогично могут быть получены планы для сколь угодно большого числа (n) - независимых управляемых факторов;

2) проведение эксперимента на объекте исследования

Так как изменение отклика (Y) носит случайный характер, то в каждой точке (X_g) приходится проводить (m) - параллельных опытов и результаты наблюдений ($y_{g1}, y_{g2}, \ldots, y_{gm}$) осреднять

$$\overline{Y}_g = \frac{1}{m} \cdot \sum_{k=1}^{m} y_{gk} \quad . \qquad (6.6)$$

Пусть в рассматриваемом случае - (m=3). Перед реализацией плана на объекте необходимо рандомизировать варианты варьирования факторов, т.е. с помощью таблицы равномерно распределенных случайных чисел определить последовательность реализации вариантов варьирования плана в (N·m) - опытах. Рандомизация проводится следующим образом. Из Приложения (таблица П.1) случайных чисел выбирается любой столбец, из которого в порядке следования берутся числа от (1) до (N·m) и записываются (таблица 6.1) последовательно в (m) - столбцов ($k_1, k_2, \ldots, k_m$). Пусть, например, (k_1=8), при (g=3); это значит, что третий вариант варьирования реализуется в эксперименте восьмым по порядку. Результаты наблюдений, в соответствии с вариантами варьирования плана, записываются в столбцы ($y_{gk_1}, y_{gk_2}, y_{gk_3}$);

3) проверка воспроизводимости эксперимента

Она состоит в проверке выполнения второй предпосылки регрессионного анализа об однородности выборочных дисперсий (S_g^2). Задача сводится к проверке гипотезы о равенстве генеральных дисперсий - $(\sigma^2\{y_1\}=\sigma^2\{y_2\}=\ldots=\sigma^2\{y_N\})$, при опытах, соответственно, в точках $(x_1, x_2, \ldots, x_N)$. Оценки дисперсий находятся по формуле

$$S_g^2 = \frac{1}{m-1}\cdot\sum(y_{gk}-\overline{Y}_g)^2 \ . \qquad (6.7)$$

Так как все оценки дисперсий получены по выборкам одинакового объема (m=3), то число степеней свободы для всех из них одинаково и составляет

$$\nu_{1\cdot вос.} = m-1 \ . \qquad (6.8)$$

В этом случае, для проверки гипотезы об однородности оценок (S_g^2) дисперсий следует пользоваться критерием Кохрэна, который основан на законе распределения отношения максимальной оценки дисперсии к сумме всех сравниваемых оценок дисперсий, т.е. -

$$G = \frac{\max\{S_g^2\}}{\sum_{g=1}^{N} S_g^2\{Y\}} \ . \qquad (6.9)$$

Если вычисленное по данным эксперимента значение критерия (G) окажется меньше критического значения ($G_{кр.}$), найденного по Приложению (таблица П.7.1), для ($\nu_{1\cdot вос.}$=m-1) и ($\nu_{2\cdot вос.}$=N) (в данном случае [$\nu_{1\cdot вос.}$=2] и [$\nu_{2\cdot вос.}$=8]) при выбранном уровне значимости [$g_{вос.}$, %, обычно 5%]), то гипотеза об однородности выборочных дисперсий отвечает результатам наблюдений. При этом, всю группу выборочных дисперсий (S_g^2) можно считать оценками для одной и той же генеральной дисперсии ($\sigma^2\{Y\}$) воспроизводимости эксперимента, откуда наилучшая ее оценка имеет вид

$$S_{вос.}^2\{Y\} = \frac{1}{N}\cdot\sum_{g=1}^{N} S_g^2\{Y\} \ , \qquad (6.10)$$

с числом степеней свободы -

$$\nu_{зн.} = N\cdot(m-1) \ . \qquad (6.11)$$

Если проверка воспроизводимости эксперимента дала отрицательный результат, то остается признать его невоспроизводимость относительно управляемых фактов, вследствие наличия неблагоприятных флуктуаций неуправляемых и неконтролируемых факторов. При этом, следует либо увеличить число параллельных опытов для вариантов варьирования с большими значениями выборочных дисперсий (S_g^2), либо использовать в дальнейшем модификацию метода наименьших квадратов, пригодную при невыполнении предпосылки и воспроизводимости эксперимента;

4) получение математической модели

На основе метода ПФЭ, получаются независимые оценки (b_0, b_i, b_{il}) соответствующих коэффициентов (β_0, β_i, β_{il}), т.е. ($b_0 \to \beta_0$), ($b_i \to \beta_i$), ($b_{il} \to \beta_{il}$) с использованием формул -

$$b_0 = \frac{1}{N} \cdot \sum_{g=1}^{N} z_{0g} \cdot \bar{Y}_g \;;\; b_i = \frac{1}{N} \cdot \sum_{g=1}^{N} z_{ig} \cdot \bar{Y}_g \cdot (i = 1,2,\ldots,n) \;; \tag{6.12}$$

$$b_{il} = \frac{1}{N} \cdot \sum_{g=1}^{N} z_{ig} \cdot z_{lg} \cdot \bar{Y}_g \cdot (i;l = 1,2,\ldots,n; i \neq l) \;.$$

После определения оценок (b) коэффициентов регрессии, необходимо проверить гипотезы об их значимости, т.е. проверить соответствующие нуль-гипотезы (β=0). Проверка таких гипотез производится с помощью критерия Стьюдента, эмпирическое значение которого

$$t = \frac{|b|}{S^2\{b\}}, \tag{6.13}$$

где $S^2\{b\} = \frac{1}{N \cdot m} \cdot S^2_{вос.}\{Y\}$ - (6.14)

- дисперсия оценки (b) коэффициента регрессии; N - число точек факторного пространства, в которых проводится эксперимент; m - число параллельных опытов в этих точках. Если найденная величина критерия (t) превышает значение ($t_{кр}$) для числа степеней свободы - ($\nu_{эн.} = N \cdot (m-1)$), при заданном уровне значимости ($g_{зн}$), обычно 5%, т.е. ($sign(t-t_{кр})=+1$), то проверяемую нуль-гипотезу (H_0:β=0) отвергают и соответствующую оценку (b) коэффициента регрессии признают значимой. В противном случае, т.е. при $sign(t-t_{кр})=-1$, нуль-гипотезу не отвергают и оценку (b) считают статистически незначимой, т.е. (β=0). Статистическая незначимость оценки (b_i) коэффициента регрессии может быть обусловлена следующими причинами: а) данный i-й фактор не имеет функциональной

связи с откликом (Y), т.е. $\beta_i=0$; б) уровень (x_{i0}) базового режима (X_0) находится в точке частного экстремума функции отклика по фактору (x_i) и тогда ($\beta_i = \frac{\partial y}{\partial z_i} = 0$); в) интервал варьирования (Δx_i) выбран малым; г) вследствие влияния неуправляемых факторов, велика ошибка воспроизводимости эксперимента.

Ортогональное планирование позволяет определить доверительные границы независимо для каждого из коэффициентов регрессии. Поэтому, если какая-либо из оценок коэффициентов окажется незначимой, то ее можно отбросить без пересчета всех остальных. После этого, математическая модель объекта составляется в виде уравнения связи отклика (Y) и факторов (z_i), включающего только значимые оценки коэффициентов;

5) проверка адекватности математического описания

Чтобы проверить гипотезу об адекватности математического описания процесса опытным данным, достаточно оценить отклонение предсказанной, по полученному уравнению регрессии, величины отклика ($\hat{Y}_g$) от результатов наблюдений ($\overline{Y}_g$) в одних и тех же (g)-х точках факторного пространства. Рассеяние результатов наблюдений вблизи уравнения регрессии, оценивающего истинную функцию отклика, можно охарактеризовать с помощью дисперсии адекватности

$$S^2_{ад.} = \frac{m}{N-2} \cdot \sum_{g=1}^{N} (\overline{Y}_g - \hat{Y}_g)^2 , \qquad (6.15)$$

где d - число членов аппроксимирующего полинома. Дисперсия адекватности определяется с числом степеней свободы

$$\nu_{ад.} = N - d . \qquad (6.16)$$

Проверка гипотезы об адекватности состоит в выяснении соотношения между дисперсией адекватности ($S^2_{ад.}$) и оценкой дисперсии воспроизводимости отклика ($S^2_{вос.}$). Если эти оценки дисперсии однородны, то математическое описание адекватно представляет результаты опыта; если нет, то описание считается неадекватным. Проверка гипотезы об адекватности производится с использование F-критерия Фишера. Критерий Фишера позволяет проверить гипотезу об однородности двух выборочных дисперсий ($S^2_{ад.}$) и ($S^2_{вос.}\{Y\}$). В том случае, если ($S^2_{ад.} > S^2_{вос.}\{Y\}$), F-критерий характеризуется отношением

$$F = \frac{S^2_{ад.}}{S^2_{вос.}\{Y\}} . \qquad (6.17)$$

Если вычисленное по результатам наблюдений эмпирическое значение F-критерия меньше критического ($F_{кр}$), найденного из Приложения (таблица П.4.1) для соответствующих степеней свободы -

$$\nu_{1\cdot ад.} = N - d \ , \quad \nu_{2 ад.} = \nu_{3\cdot н.} = N \cdot (m - 1) \ , \qquad (6.18)$$

при заданном уровне значимости ($g_{ад}$), то гипотеза об адекватности не отвергается. В противном случае, гипотеза отвергается и математическое описание признается неадекватным. Проверка адекватности возможна, при ($\nu_{1\cdot ад.}>0$). Если число (T) - вариантов варьирования плана ПФЭ равно числу всех значимых оценок коэффициентов регрессии (N=d), то для проверки гипотезы об адекватности математического описания степеней свободы не остается ($\nu_{1\cdot ад.}=0$). Если некоторые оценки коэффициентов регрессии оказались незначительными, то число (d) - членов проверяемого уравнения, в этом случае, меньше числа (N) - вариантов варьирования (N>d) и для проверки гипотезы об адекватности остается одна или несколько степеней свободы ($\nu_{1ад} > 0$). В случае, когда гипотеза об адекватности отвергается, необходимо переходить к более сложной форме математического описания, либо, по возможности, проводить эксперимент с меньшим интервалом варьирования (ΔX_i). Следует отметить, что максимальная величина интервала варьирования определяется условием адекватного описания объекта в области варьирования. Если при больших интервалах варьирования математическая модель неадекватна, то возникают систематические ошибки в определении коэффициентов, для уменьшения которых следует сузить область выравнивания. Однако, с уменьшением интервала варьирования, появляется целый ряд новых трудностей: растет отношение помехи к полезному сигналу, что приводит к необходимости увеличения числа параллельных опытов для выделения полезного сигнала на фоне шума, т.е. уменьшаются абсолютные значения оценок (b_i) коэффициентов регрессии, величины которых непосредственно зависят от Δx_i (для уравнения с нормированными факторами "z_i"), и оценки коэффициентов могут стать статистически незначимыми.

6.2 Дробный факторный эксперимент

Во многих практических задачах идентификации влияние взаимодействий второго и высших порядков отсутствует или пренебрежимо ма-

ло. Кроме того, на первый этапах исследования часто необходимо получить в первом приближении лишь линейную аппроксимацию изучаемого уравнения связи при минимальном количестве опытов. Поэтому, *неэффективно использовать полный факторный эксперимент (ПФЭ)* для оценивания коэффициентов регрессии лишь при линейных членах и некоторых парных произведениях, из-за реализации большого числа вариантов варьирования (2^n), в особенности при большом числе факторов (n). При линейном росте числа независимых факторов, число вариантов варьирования для ПФЭ растет по показательному закону, в результате чего на проверку гипотезы об адекватности остается излишне много степеней свободы. *Дробным факторным экспериментом (ДФЭ) называется эксперимент, реализующий часть (дробную реплику) полного факторного эксперимента (ПФЭ).* Он позволяет получить, например, линейное приближение искомой функциональной зависимости ($M\{Y\} = \varphi(\bar{X})$) в некоторой небольшой окрестности точки базового режима при минимуме опытов:

1) планирование эксперимента

Для решения трехфакторной (n=3) задачи регрессии в линейном приближении можно ограничиться четырьмя вариантами варьирования, если в планировании ПФЭ типа (2^2) произведение ($z_1 \cdot z_2$) приравнять к третьему независимому фактору (z_3). Такое планирование, представленное матрицей (таблица 6.3), позволяет найти свободный член (b_0) и три оценки коэффициентов регрессии при линейных членах (b_1), (b_2), (b_3) (из четырех опытов нельзя получить более четырех оценок коэффициентов регрессии).

Таблица 6.3 Матрица дробного факторного эксперимента

g	z_0	z_1	z_2	z_3	$z_1 \cdot z_2$	$z_1 \cdot z_3$	$z_2 \cdot z_3$	$z_1 \cdot z_2 z_3$
1	+1	-1	-1	+1	+1	-1	-1	+1
2	+1	+1	-1	-1	-1	-1	+1	+1
3	+1	-1	+1	-1	-1	+1	-1	+1
4	+1	+1	+1	+1	+1	+1	+1	+1

Применение ДФЭ всегда связано со смешиванием, т.е. с совместным оцениванием нескольких теоретических коэффициентов математической модели. В рассматриваемом случае, если коэффициенты регрессии (β_{ij}), при парных произведениях, отличны от нуля, каждый из найденных коэффициентов (β_i) служит оценкой двух теоретических коэффициентов

регрессии

$$b_0 \rightarrow \beta_0 + \beta_{123}\,;\ b_1 \rightarrow \beta_1 + \beta_{23}\quad b_2 \rightarrow \beta_2 + \beta_{13}\,;\ b_3 \rightarrow \beta_3 + \beta_{12}\,. \qquad (6.19)$$

Действительно, указанные теоретические коэффициенты в таком планировании не могут быть оценены раздельно, поскольку столбцы матрицы планирования для линейных членов и парных произведений совпадают. Рассмотренный план ДФЭ представляет половину плана ПФЭ типа (2^3) и называется *полурепликой* от ПФЭ типа (2^3) или планированием типа ($N=2^{3-1}$) (таблица 6.2). Для правильного планирования ДФЭ необходимо использовать все полученные ранее сведения теоретического и интуитивного характера об объекте и выделить те факторы и произведения факторов, влияние которых на отклик существенно. При этом, смешивание нужно производить так, чтобы линейные коэффициенты (β_0, β_1,...,β_n) были смешаны с коэффициентами при взаимодействиях самого высокого порядка (так как обычно они в модели отсутствуют), или при тех взаимодействиях, о которых априори известно, что они не оказывают влияния на отклик. Следовательно, недопустима произвольная разбивка плана ПФЭ типа (2^3) на две части для выделения полуреплики типа (2^{3-1}). При большом числе (n) - факторов, для получения линейного приближения, можно построить дробные реплики высокой степени дробности. Так, при (n=7), можно составить дробную реплику на основе ПФЭ типа (2^3), приравняв четыре из семи факторов к взаимодействиям трех других факторов: парным и тройному. Обозначим тип дробной реплики записью (2^{n-p}), если (n-p) - факторов приравнены к произведениям остальных факторов. План ДФЭ можно построить, приравнивая факторы различным взаимодействиям; разумеется, при этом, меняется система совместных оценок теоретических коэффициентов. Для получения системы совместных оценок и анализа разрешающей способности дробных реплик *удобно пользоваться понятиями генерирующего* и *определяющего соотношений. Генерирующее соотношение* служит для построения дробной реплики. Так, в рассматриваемом планировании задается полуреплика плана ПФЭ типа (2^3) с помощью генерирующего соотношения ($z_3=z_1 \cdot z_2$). *Определяющим соотношением* называется соотношение, задающее элементы первого столбца матрицы планирования для фиктивной переменной (все они всегда равны "1"). Выражение определяющего соотношения в рассматриваемом случае получается умножением левой и правой частей приведенного генерирующего соотношения на (z_3), т.е. ($1=z_1 \cdot z_2 \cdot z_3$), так

как всегда ($z_i^2=1$). Знание определяющего соотношения позволяет найти всю систему совместных оценок без изучения матрицы планирования ДФЭ. Соотношения, задающие эти оценки, можно найти, последовательно перемножив независимые факторы на определяющее соотношение

$$Z_0=z_1 \cdot z_2 \cdot z_3,\ z_1=z_2 \cdot z_3,\ z_2=z_1 \cdot z_3,\ z_3=z_1 \cdot z_2, \qquad (6.20)$$

откуда легко находятся смешиваемые теоретические коэффициенты регрессии и их оценки

$$b_0 \rightarrow \beta_0+\beta_{123};\quad b_1 \rightarrow \beta_1+\beta_{23};\quad b_2 \rightarrow \beta_2+\beta_{13};\quad b_3 \rightarrow \beta_3+\beta_{12}. \qquad (6.21)$$

Если априори можно принять, что коэффициенты при всех парных и тройном взаимодействиях равны нулю, то реализация этой полуреплики позволит получить раздельные оценки для всех четырех линейных коэффициентов регрессии. Разрешающая способность полуреплик определяется их генерирующими соотношениями. Разрешающая способность тем выше, чем более высок порядок взаимодействий, с коэффициентами которых смешаны линейные коэффициенты. Она увеличивается для главных полуреплик с ростом числа независимых факторов. Для четвертьреплики в пятифакторном планировании типа (2^{5-2}) должны быть заданы два генерирующих соотношения, например, ($z_4 = z_1 \cdot z_2 \cdot z_3$; $z_5=z_1 \cdot z_2$), причем, полагаем ($\beta_{123}=0$), т.е. (x_1, x_2, x_3) - все вместе не взаимодействуют, и ($\beta_{12}=0$), т.е. (x_1), и (x_2) также не взаимодействуют. Определяющие соотношения для этой реплики, согласно приведенным выше правилам, имеют вид ($1=z_1 \cdot z_2 \cdot z_3 \cdot z_4$); ($1=z_1 \cdot z_2 \cdot z_5$). Если для дробной реплики имеют место два (или более) определяющих соотношения, то их перемножают между собой, используя все возможные комбинации. В рассматриваемом случае, имеется одна новая комбинация: ($1=z_3 \cdot z_4 \cdot z_5$). *Обобщающее определяющее соотношение*, построенное на основе всех полученных определяющих соотношений, полностью характеризует разрешающую способность реплик высокой степени дробности. Так, в данном случае ($1=z_1 \cdot z_2 \cdot z_3 \cdot z_4=z_1 \cdot z_2 \cdot z_5=z_3 \cdot z_4 \cdot z_5$). Совместные оценки здесь определяются вспомогательными соотношениями

$$
\begin{aligned}
&Z_0 = z_1 \cdot z_2 \cdot z_3 \cdot z_4 = z_1 \cdot z_2 \cdot z_5 = z_3 \cdot z_4 \cdot z_5;\\
&z_1 = z_2 \cdot z_3 \cdot z_4 = z_2 \cdot z_5 = z_1 \cdot z_3 \cdot z_4 \cdot z_5;\\
&z_2 = z_1 \cdot z_3 \cdot z_4 = z_1 \cdot z_5 = z_2 \cdot z_3 \cdot z_4 \cdot z_5;\\
&z_3 = z_1 \cdot z_2 \cdot z_4 = z_1 \cdot z_2 \cdot z_3 \cdot z_5 = z_4 \cdot z_5;\\
&z_4 = z_1 \cdot z_2 \cdot z_3 = z_1 \cdot z_2 \cdot z_4 \cdot z_5 = z_3 \cdot z_5;\\
&z_5 = z_1 \cdot z_2 \cdot z_3 \cdot z_4 \cdot z_5 = z_1 \cdot z_2 = z_3 \cdot z_4;\\
&z_1 \cdot z_3 = z_2 \cdot z_4 = z_2 \cdot z_3 \cdot z_5 = z_1 \cdot z_4 \cdot z_5;\\
&z_1 \cdot z_4 = z_2 \cdot z_3 = z_2 \cdot z_4 \cdot z_5 = z_1 \cdot z_3 \cdot z_5.
\end{aligned}
\qquad (6.22)
$$

Данные вспомогательные соотношения позволяют установить, какие столбцы матрицы планирования окажутся линейно зависимыми и, следовательно, совместной оценкой каких теоретических коэффициентов является тот или иной выборочный коэффициент регрессии

$$
\begin{aligned}
&b_0 \rightarrow \beta_0+\beta_{1234}+\beta_{125}+\beta_{345}; \quad && b_4 \rightarrow \beta_4+\beta_{123}+\beta_{1245}+\beta_{35};\\
&b_1 \rightarrow \beta_1+\beta_{234}+\beta_{25}+\beta_{1345}; && b_5 \rightarrow \beta_5+\beta_{12345}+\beta_{12}+\beta_{34};\\
&b_2 \rightarrow \beta_2+\beta_{134}+\beta_{15}+\beta_{2345}; && b_{13} \rightarrow \beta_{13}+\beta_{24}+\beta_{235}+\beta_{145};\\
&b_3 \rightarrow \beta_3+\beta_{124}+\beta_{1235}+\beta_{45}; && b_{14} \rightarrow \beta_{14}+\beta_{23}+\beta_{245}+\beta_{135}.
\end{aligned}
\qquad (6.23)
$$

Разрешающая способность этой четвертьреплики невысокая, так как все теоретические линейные коэффициенты регрессии смешаны с коэффициентами при парных взаимодействиях. Следует иметь в виду, что план ДФЭ всегда можно расширить до плана ПФЭ недостающими дробными репликами. В данном примере, для остальных трех четвертьреплик генерирующие соотношения запишутся в виде

$$
\begin{cases} z_4 = z_1 \cdot z_2 \cdot z_3; \\ z_5 = -z_1 \cdot z_2; \end{cases}
\quad
\begin{cases} z_4 = -z_1 \cdot z_2 \cdot z_3; \\ z_5 = z_1 \cdot z_2; \end{cases}
\quad
\begin{cases} z_4 = -z_1 \cdot z_2 \cdot z_3; \\ z_5 = -z_1 \cdot z_2, \end{cases}
\qquad (6.24)
$$

а обобщающие определяющие соотношения - в виде:

$$
\begin{aligned}
&1 = z_1 \cdot z_2 \cdot z_3 \cdot z_4 = -z_1 \cdot z_2 \cdot z_5 = -z_3 \cdot z_4 \cdot z_5;\\
&1 = -z_1 \cdot z_2 \cdot z_3 \cdot z_4 = z_1 \cdot z_2 \cdot z_5 = -z_3 \cdot z_4 \cdot z_5;\\
&1 = -z_1 \cdot z_2 \cdot z_3 \cdot z_4 = z_1 \cdot z_2 \cdot z_5 = z_3 \cdot z_4 \cdot z_5.
\end{aligned}
\qquad (6.25)
$$

Осуществление этих дополняющих четвертьреплик означает реализацию ПФЭ в целом и, следовательно, раздельное оценивание всех теоретических коэффициентов регрессии (если априори известно, что $\beta_{ii}=0$, $\beta_{iii}=0$);

2) планирование эксперимента на объекте исследования

Реализация плана ДФЭ ничем не отличается от реализации плана ПФЭ;

3) проверка воспроизводимости эксперимента

Проверка однородности оценок дисперсии отклика в различных точках факторного пространства проводится в полном соответствии с методикой, изложенной для ПФЭ, различие состоит лишь в числе точек плана;

4) получение математической модели объекта

Процедуры определения оценок коэффициентов регрессии и проверки их значимости полностью совпадают с процедурой, применяемой при исследовании объекта методом ПФЭ;

5) проверка адекватности математического описания

Адекватность математического описания функции отклика проверяется теми же методами, что и для ПФЭ.

Пример

*Необходимо **дать рекомендации по постановке полевого эксперимента с целью установления оптимальных параметров** горизонтального закрытого **дренажа** для осушения участка в сложных гидрогеологических условиях. В результате критического анализа литературных источников, а также предварительных расчетов по формулам различных авторов определено, что расстояние между дренами колеблется от 10 до 20 м, а глубина заложения - от 1,0 до 1,2 м. Эти данные необходимо проверить и уточнить на основе данных полевых исследований на экспериментальном участке.*

Решение

Назначаются три градации расстояний между дренами 10, 15, 20 (м) и три градации глубины заложения дрены - 1,0, 1,1, 1,2 (м). Как видно из условия задачи, необходимо запланировать полный факторный эксперимент для оценки влияния двух факторов с тремя градациями каждого (3^2):

1) *обозначаем расстояние между дренами фактором (А), глубину заложения дрен - фактором (В);*

2) *градации расстояний между дренами закодируем следующим образом: a_1 - 10 (м); a_2 - 15 (м); a_3 - 20 (м), а градации глубин заложения как - b_1 - 1,0 (м); b_2 - 1,1 (м); b_3 - 1,2 м;*

3) составим матрицу планирования ПФЭ (3^2)

Номер варианта	*Фактор*		*Обозначение варианта*
	A	*B*	
1 - й	*1*	*1*	*0 (контроль)*
2 - й	*2*	*1*	
3 - й	*3*	*1*	
4 - й	*1*	*2*	
5 - й	*2*	*2*	
6 - й	*3*	*2*	
7 - й	*1*	*3*	
8 - й	*2*	*3*	
9 - й	*3*	*3*	

Поскольку в планируемом опыте нет нулевых градаций, то градации факторов (А) и (В) в таблице обозначаем числами 1, 2 и 3;

4) на опытном участке необходимо заложить опыт по следующей схеме

Номер варианта	*1 - й*	*2 - й*	*3 - й*	*4 - й*	*5 - й*	*6 - й*	*7 - й*	*8 - й*	*9 - й*
Расстояние между дренами (м)	*10*	*15*	*20*	*10*	*15*	*20*	*10*	*15*	*20*
Глубина заложения дрен (м)	*1,0*	*1,0*	*1,0*	*1,1*	*1,1*	*1,1*	*1,2*	*1,2*	*1,2*

После проведения эксперимента, представится возможность обработки опытных данных, и статистическими методами, рассмотренными выше, обосновать оптимальные расстояния между дренами и глубины их заложения.

7 МЕТОДЫ ПРОСТРАНСТВЕННОГО ОБОБЩЕНИЯ ГИДРОМЕТЕОРОЛОГИЧЕСКОЙ И ЭКОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Пространственное обобщение гидрометеорологической и экологической информации является заключительным этапом статистической обработки результатов наблюдений с их представлением в виде информационных полей. Основное содержание пространственного обобщения информации: оценка статистических показателей пространственной структуры полей, картирование и районирование метеорологических, климатических, гидрологических и эколого-ландшафтных характеристик. Исследование элементов пространственной структуры приобретает важное значение в связи с необходимостью представления рассматриваемой информации непосредственно в границах природно-экономических (административных) районов, осуществление программы природно - климатического мониторинга и экологического аудита. Поля гидрометеорологических и экологических элементов обычно задаются данными в отдельных точках пространства (опытными по отдельным пунктам или проинтерполированными в узлы регулярной сетки с помощью объективного анализа). Поскольку, каждое значение элемента в координатах поля представлено случайным элементом выборки, то такие информационные поля элементов являются случайными полями. Для описания совокупности рядов в пространстве, используются корреляционные, ковариационные и спектральные функции. Данные функции однозначно связаны между собой и равно пригодны для описания статистической структуры полей элементов, но на практике чаще используются пространственные корреляционные функции (ПКФ) - как более точные, универсальные, менее зависимые от колебаний сезонного и географического характера. Пространственные корреляционные функции случайного поля ($M(\varphi_j,\lambda_j,t_i)$), в общем случае, характеризуются зависимостью

$$r_{jk} = \frac{\sum_{i=1}^{n_{jk}} (M_{ij} - \overline{M}_j) \cdot (M_{ik} - \overline{M}_k)}{\sqrt{\sum_{i=1}^{n_{jk}} (M_{ij} - \overline{M}_j)^2 \cdot \sum_{i=1}^{n_{jk}} (M_{ik} - \overline{M}_k)^2}}, \qquad (7.1)$$

в которой M_{ij} - значение элемента гидрометеорологического поля в (j)-ой точке с координатами (φ_j) и (λ_j) в (i)-ый интервал времени; $\overline{M}_j$ - -норма гидрометеорологической величины в (j)-ой точке; n_{jk} - количество

интервалов совместных наблюдений за элементом поля. Практический интерес представляют *однородные* и *изотропные поля. Однородными* являются поля с одинаковыми законами распределения вероятностей той или иной величины во всех координатах поля (равенство одноточечных характеристик средних и дисперсий). Если пространственная корреляционная функция поля зависит только от расстояния между рассматриваемыми точками и не зависит от направления между ними, поле считается *изотропным.* В случае невыполнения этих условий, говорят об однородности и изотропности поля только относительно пространственных корреляционных функций. Для однородных и изотропных полей ПКФ зависит только от расстояния между наблюдаемыми точками, т.е. (R=R(ρ)). Следует иметь в виду, что для реальных информационных полей, из-за сложности их структуры, более точным является использование введенных А.Н. Колмогоровым терминов "*локальная однородность*" и "*локальная изотропность*", когда поля принимаются однородными и изотропными в пределах пространственных и временных моментов, характерных для исследуемых процессов. В зависимости от этого, рассматриваются *микро-*, *мезооднородность* - изотропность в пределах от нескольких километров до сотен километров и *макрооднородность* - изотропность от сотен километров и далее. Это связано с тем, что в отличие от микро- и мезоизменчивости, макроизменчивость вызывается, по существу, принципиально иными причинами, связанными с глобальными процессами циркуляции атмосферы, физическая природа которых отличается от природы микро- и мезомасштабных процессов. В случае мезо-, макромасштабных явлений, условия локальной однородности выполняются лишь в горизонтальном направлении, как правило до расстояний не более 2000 км, хотя для ряда метеорологических полей (осадки, запасы воды в снеге и др.) однородность (изотропность) нарушается на гораздо меньших расстояниях. В любом случае, это происходит из-за существования неоднородности полей гидрометеорологических элементов в горизонтальном направлении.

7.1 Оценка статистической структуры поля

Кроме пространственных корреляционных функций, к числу наиболее важных статистических характеристик, описывающих пространственную структуру гидрометеорологических и экологических полей, относятся кривая распределения в пространстве (ее параметры - средняя

арифметическая - "$\overline{X}_F$", и дисперсия - "σ_F"по площади - "F"), коэффициент аномальности поля , пространственная структурная функция, а также показатели сходства и различия полей. Осреднение используется как способ получения результатов для сравнивания фоновых характеристик рассматриваемых элементов в отдельных районах или их изменений на больших территориях. Осреднение - необходимый этап обобщения исходных данных при анализе полей гидрометеорологических элементов в силу их большой пространственной изменчивости. Существует множество способов площадного осреднения, но наиболее часто используются методы среднего арифметического взвешивания по площадям - квадратам, изолиний и оптимального осреднения. Обобщенное выражение для оценки средних по площади записывается в виде

$$\overline{X}_F = \sum_{i=1}^{n} \alpha_i \cdot x_i(\varphi_i, \lambda_i) , \qquad (7.2)$$

где $\overline{X}_F$ - среднее значение рассматриваемого элемента по площади; $X_i(\varphi_i, \lambda_i)$ - значение элемента в точке с координатами (φ_i, λ_i); α_i - коэффициент, зависящий от веса и способа осреднения; n - число метеостанций, используемых при осреднении. При равномерном расположении метео-, мониторинговых станций в условиях равнинного рельефа (когда веса "$\alpha_i \approx 1$"), чаще всего используется метод арифметического осреднения

$$\overline{X}_F = \frac{\sum_{i=1}^{n} x_i(\varphi_i, \lambda_i)}{n} . \qquad (7.3)$$

Арифметическая сумма по методу взвешивания по площадям (метод полигонов) вычисляется как

$$\overline{X}_F = \frac{\sum_{i=1}^{n} X_i(\varphi_i, \lambda_i) \cdot F_i}{\sum_{i=1}^{n} F_i} , \qquad (7.4)$$

где F_i - площадь полигона, относящаяся к метео-, мониторовой станции $(X_i(\varphi_i, \lambda_i))$. При неравномерном распределении пунктов по площади целесообразно применять метод квадратов, который при использовании ЭВМ называется еще методом оптимальной интерполяции. По методу квадратов исследуемая площадь разбивается на регулярную сеть квадратов или точек, данные в которых определяются путем осреднения наблюденных значений непосредственно в квадратах или снимаются с карт

изолиний. В последующем, для определения ($\overline{X}_F$), производится осреднение характеристики по формуле (7.2). В методе оптимальной интерполяции весовые множители (P_j) находятся, при наличии сведений о статистической структуре поля, путем решения системы уравнений

$$\sum_{j=1}^{n} P_j \cdot R_{ij} = R_{0i}\ ,\ \text{при}\ (i=1,2,...,n), \qquad (7.5)$$

или в случае, когда учитывается ошибка наблюдений (Δf (ρ_i=0)) в точке, при (ρ_i=0), по соотношению

$$\sum_{j=1}^{n} P_j \cdot R_{ij} + P_j \cdot \eta_j^2 = R_{0i}\ ,\ \text{при}\ (i=1,2,...,n), \qquad (7.6)$$

в котором $\eta_j^2 = \frac{\sigma_{\Delta_i}^2}{\sigma_i^2}$ - мера ошибок наблюдений в (j)-ой точке; R_{0i}, R_{ji} -

-значения коэффициентов корреляции истинных значений исследуемой характеристики в точках (j) и (0). Оценка пространственной дисперсии (σ_F^2), как правило, рассчитывается по формуле, аналогичной обычной формуле оценки дисперсии для случайной выборки

$$\sigma_F^2 = \frac{\sum_{i=1}^{n} (x_i(\varphi_i, \lambda_i) - \overline{X}_F)^2}{n}\ . \qquad (7.7)$$

При обработке исходных данных возникает необходимость оценить степень аномальности поля. Оценку аномальности поля можно вести, исходя из отклонений значений элемента от средней в отдельных точках поля (аномалий) по повторяемости аномалий, по весу площади, занимаемой аномалией и т.д. Одной из распространенных, хотя и несколько формальных, характеристик является коэффициент аномальности поля, предложенный А.Н. Багровым

$$A = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(x_i(\varphi_i, \lambda_i) - \overline{X})^2}{\sigma_i^2}\ , \qquad (7.8)$$

где $(X_i(\varphi_i, \lambda_i) - \overline{X})$ - аномалия; σ_i^2 - дисперсия поля в (i)-той точке. Предполагается, что число точек (n) достаточно велико и они расположены относительно равномерно по территории. В каждой отдельной реализации случайного поля величина (X_i) в координатах (φ_i) и (λ_i) принимает то или иное значение, в связи с чем возникает вопрос, насколько близки к нему значения (x_i) в окружающих точках. Полное статистическое описание одновременного поведения случайного поля требует задания много-

мерных функций распределения, что осложняет объективное решение задачи описания и интерполяции. На практике ограничиваются более простыми характеристиками в виде пространственных корреляционных функций (ПКФ) - R=f(ρ). ПКФ принимает максимальное значение при (ρ =0), с увеличением расстояния (R=f(ρ)) - убывает. При очень больших расстояниях между метео-, мониторинговыми станциями, связь между элементами практически отсутствует, приближаясь к нулю.

Алгоритм расчета ПКФ предусматривает проведение следующих операций:

1) составление матрицы гидрологической характеристики на (k)- станциях за (N) - лет

$$\|M_{ij}\|, \text{ при } (i = \overline{1, N_j}; j = \overline{1, k}); \tag{7.9}$$

2) расчет одноточечных моментов по всем рядам наблюдений-

а) среднее для каждого пункта

$$\overline{M}_j = \frac{1}{N_j} \cdot \sum_{i=1}^{N_j} M_{ij}\ ; \tag{7.10}$$

б) среднеквадратическое отклонение

$$\sigma_j = \sqrt{\frac{1}{N_j} \cdot \sum_{i=1}^{N_j} (M_{ij} - \overline{M}_j)^2}\ ; \tag{7.11}$$

в) среднеквадратическое отклонение выборочной средней арифметической

$$\overline{\sigma}_j = \frac{\sigma_j}{\sqrt{N_j}}\ ; \tag{7.12}$$

г) коэффициент вариации

$$C_{V_j} = \frac{\sigma_j}{\overline{M}_j}\ ; \tag{7.13}$$

д) среднеквадратическая ошибка коэффициента вариации

$$\sigma_{C_{Vj}} = \sqrt{\frac{1 + C_{Vj}^2}{2 \cdot N_j}}\ ; \tag{7.14}$$

3) оценка коэффициентов парной корреляции за совместный период наблюдений

$$r_{jk} = \frac{\sum_{i=1}^{n_{jk}} (x_{ik} - \overline{X}_k) \cdot (x_{ij} - \overline{X}_j)}{\sigma_k \cdot \sigma_j \cdot n_{jk}}; \qquad (7.15)$$

4) осреднение парных коэффициентов корреляции по градациям расстояний (Δρ) (с учетом числа попаданий - "k_j" в "j"-градацию)

$$\bar{r}_j = \frac{\sum_{i=1}^{k_j} r_{ji} \cdot n_{ji}}{\sum_{i=1}^{k_j} n_{ji}}; \qquad (7.16)$$

5) аппроксимация ПКФ, по всей совокупности коэффициентов корреляции или по средневзвешенным значениям, зависимостями различного типа, выбор которых производится по критерию минимальной остаточной дисперсии и критерию Фишера

$$R(\rho) = \begin{cases} R(0) - \alpha \cdot \rho, \\ \exp(-\alpha \cdot \rho^{\beta}), \\ (1 + \alpha \cdot \rho) \cdot (\exp - \alpha \cdot \rho), \\ R(0) \cdot \exp(-\rho / \rho_0)^n, \\ \dfrac{\sin(\beta \cdot \rho)}{\beta \cdot \rho} \cdot \exp(-\alpha \cdot \rho), \\ \exp(-\alpha \cdot \rho) \cdot \cos(\beta \cdot \rho); \end{cases} \qquad (7.17)$$

6) оценка ошибки взвешенных средних коэффициентов корреляции по соотношению

$$\sigma_{r_{jk}} = \frac{1 - \bar{r}_j^2}{\sqrt{n_j - 1}}; \qquad (7.18)$$

а для распределения величины

$$Z_j = \frac{1}{2} \cdot \ln \frac{1 + \bar{r}_j}{1 - \bar{r}_j} + \frac{\bar{r}_j}{2 \cdot (n - 1)}. \qquad (7.19)$$

оценка ошибки

$$\sigma_{z_{jk}} = \frac{1}{\sqrt{n_j - 3}}; \qquad (7.20)$$

7) построение соответствующих доверительных интервалов при соответствующих уровнях доверительной вероятности

$$\bar{Z}(\rho_{jk}) - t_{1-P} \cdot \sigma_{Z_{jk}} < Z_{jk} < \bar{Z}(\rho_{jk}) + t_{1-P} \cdot \sigma_{Z_{jk}} , \qquad (7.21)$$

где $t_{1-P} = (Z_{jk} - \bar{Z}(\rho_{jk})) / \sigma_{Z_{jk}}$ - квантиль нормального распределения при заданной доверительной вероятности (1-P);

8) оценка пространственной однородности ПКФ выполняется на основе критериев согласия Колмогорова (χ^2), (ω^2), по распределению Фишера или каким-либо другим способом.

7.2 Оценка точности характеристик статистической структуры поля

Расчет статистических характеристик гидрометеорологических и эколого-ландшафтных полей по ограниченному объему данных не позволяет получить генеральное значение, а дает лишь их оценки. Для практических целей важно знать точность, с которой получены оценки статистической структуры поля. Кроме того, *вследствие неоднородности исходных данных и нестационарности материалов наблюдений*, связанных с особенностями измерений, условиями макро- и мезопроцессов (масштабы которых меньше пространственно-временного разрешения системы наблюдений) *имеют место* определенные *неточности* (систематические погрешности). *Разделить ошибки* за счет указанных факторов зачастую *невозможно*, и *приходится* их *объединять под* общим *названием случайные ошибки*. При измерении или вычислении величины в точке получается не истинное ее значение, а некоторая величина

$$M'_k = M_k \pm \Delta_{M_k} . \qquad (7.22)$$

Ошибка (Δ_{M_k}) включает в себя систематическую погрешность ($\bar{\Delta}_{M_k}$), одинаковую для всех измерений или вычислений в аналогичных условиях, и случайную величину (δ_{M_k}), которая может принимать различные значения

$$\Delta_{M_k} = \bar{\Delta}_{M_k} + \delta_{M_k} . \qquad (7.23)$$

Погрешность в определении гидрометеорологических элементов составляет

$$\bar{\sigma}^2_{M_k} = (M'_k - \bar{M}_k)^2 = \sigma^2_{M_k} + \Delta^2_k = \sigma^2_{M_k} \cdot (1 + \eta^2_{M_k}) , \qquad (7.24)$$

где Δ^2_k - дисперсия ошибок наблюдений; $\eta_{M_k} = \frac{\Delta^2_k}{\sigma_{M_k}}$ - мера случайных погрешностей в исходных данных. Средняя величина, вычисленная по реальным данным, изменяется лишь при наличии систематической погрешности ($\bar{\Delta}_{M_k}$). На дисперсию наличие систематических погрешностей не оказывает влияния, но она завышается на величину, равную дисперсии ошибок измерения. С коэффициентами корреляции -

$$\bar{r}_{jk} = r_{jk} \cdot \sqrt{\frac{1 + \eta^2_{M_j}}{1 + \eta^2_{M_k}}} , \qquad (7.25)$$

дело обстоит иначе. Здесь наличие ошибок в исходных данных, как правило, приводит к их занижению. В случае однородных и изотропных полей, естественно считать одинаковыми в разных точках и меры ошибок в исходных данных, следовательно,

$$R(\rho) = \frac{R(0)}{1 + \eta^2_M} , \qquad (7.26)$$

где R(0) - экстраполированное значение эмпирической ПКФ до значения (ρ=0). Отметим также, что при уменьшении расстояния (ρ), корреляционная функция (R(ρ)) стремится не к (1), а к некоторой положительной величине

$$\bar{R}(0) = \frac{1}{1 + \eta^2_M} . \qquad (7.27)$$

Исходя из этого, значения эмпирических функций на малых расстояниях могут быть использованы для оценки точности исходных данных. Величина (R(0)) получается путем экстраполяции (R(ρ)) при малых значениях аргумента. После этого можно получить

$$\eta^2_M = \frac{(1 - R(0))}{R(0)} . \qquad (7.28)$$

В итоге, приведение эмпирической корреляционной функции к теоретической осуществляется по формуле

$$R(\rho) = \frac{\tilde{R}(\rho)}{R(0)} . \qquad (7.29)$$

Наряду с систематическими погрешностями за счет ошибок наблюдений, возможны искажения характеристик статистической структуры полей элементов из-за недостаточных объемов выборок экспериментальных данных, а также из-за различной длины их статистических моментов. Поэтому, определенные на эмпирическом материале значения ПКФ всегда имеют разброс и перед их использованием приходится проводить сглаживание. Алгоритм сглаживания содержит в себе осреднение коэффициентов корреляции для всех пар точек, попавших в некоторую градацию по расстоянию. Полученные значения корреляционной функции сглаживаются для того, чтобы согласовать значения для различных градаций, что позволяет уточнить характеристики для некоторых, недостаточно статистически обеспеченных градаций. В большинстве случаев, довольно надежные результаты получаются при простом арифметическом осреднении коэффициентов корреляции. Но при достаточно высокой связности ($r>0,8$) необходимо учитывать несимметричность распределения выборочных коэффициентов корреляции, которое приводит к тому, что более высокие их значения являются более точными и должны учитываться с большим весом. В этом случае, предлагается осуществлять осреднение не самих коэффициентов корреляции, а величин (Z), получаемых путем преобразования Z-Фишера (7.19). Распределение величин (Z_j), как правило, существенно ближе к нормальному закону, и их осреднение является более оправданным. После осреднения величин (Z_j) по градациям расстояний, соответствующие значения коэффициента корреляции получаются путем обратного преобразования: ($r = th(\bar{Z})$). Расчеты характеристик пространственной структуры связаны с учетом однородности и изотропности, особенно, при использовании ПКФ для решения ряда практических задач. Если однородность и / или изотропность поля нарушаются, точность определения коэффициентов корреляции по ПКФ снижается, что, естественно, сказывается, в целом, на результатах расчетов. Наибольшее практическое распространение получил следующий подход к решению задачи оценки однородности и изотропности. С помощью ка-

ких-либо статистических критериев оценивается степень разброса точек на ПКФ. Если разброс точек невелик и удовлетворяет выбранному статистическому критерию, поле считается однородным и изотропным. Необходимым и достаточным условием однородности корреляционной функции в пределах рассматриваемого района по критерию Г.А. Алексеева является выполнение неравенства

$$\left|z_{jk} - \bar{Z}(\rho_{jk})\right| \geq \sigma_{z_{jk}} \approx 31{,}7\% \text{ или } \geq 2 \cdot \sigma_{z_{jk}} \approx 4{,}6\% , \qquad (7.30)$$

где 31,7(%) и 4,6(%), соответственно, число случаев от общего числа $(C^2(p) = p \cdot (p-1)/2)$ эмпирических значений (z_{jk}). Другими словами, в пределах квантилей нормального распределения (t=1,t=2), общее эмпирическое число превышений $(K_Э)$ должно быть теоретически (по нормальному закону распределения) равно числу превышений, т.е.

$$K_Э(1) \approx 0{,}317 \cdot C_\rho^2 = 0{,}317 \cdot n \cdot (n-1)/2 ; \qquad (7.31)$$

$$K_Э(2) \approx 0{,}046 \cdot C_\rho^2 = 0{,}046 \cdot n \cdot (n-1)/2 , \qquad (7.32)$$

где n - число пунктов наблюдений по территории. При больших расхождениях между теоретическими и эмпирическими вероятностями "нулевая" гипотеза отвергается и признается альтернативная гипотеза "неоднородности" эмпирической пространственной корреляционной функции. В этом случае, поле рассматриваемого элемента должно быть уменьшено. Проверка однородности и изотропности, при этом, повторяется. Кроме того, в некоторых случаях исключаются пункты, вызывающие неоднородность статистической структуры полей элементов. Однако, следует отметить, что достижение однородности и изотропности поля путем исключения части информации, если не проводится специальных исследований по поводу достоверности данных, не всегда правомерно, так как исключение данных приводит к утере ценной, возможно, реальной информации. Альтернативой может служить подход, который основан на выяснении генетических причин разброса точек и решении задач на случай неоднородного и анизотропного поля. ПКФ анизотропного поля зависят не только от расстояния между точками, но и от направления, их связывающего. На использовании этого условия разработана методика оценки изотропности путем анализа схем изокоррелят, построенных относительно различных пунктов - центров корреляции. Этот метод,

хотя и наглядный, но трудоемкий, поскольку требует обработки большого объема информации вручную. В связи с этим, следует отметить более наглядный *способ представления пространственной связности поля по развернутым пространственным корреляционным функциям* (РПКФ). В соответствии с этим методом, осреднение коэффициентов корреляции осуществляется не только по градациям расстояний, но и в зависимости от направления между метеостанциями, которое отсчитывается от параллели или меридиана. Для получаемой, таким образом, трехмерной поверхности строятся линии равных уровней (изокорреляты). Если линии уровня РПКФ близки к окружностям, принимается гипотеза об изотропности поля; в общем случае, по форме изокоррелят можно наглядно проследить направления большей и меньшей связности поля, т.е. проанализировать характер его анизотропности. В отличие от изокоррелят, построенных относительно отдельных станций, линии уровня РПКФ получаются путем осреднения всех индивидуальных изокоррелят, в силу чего, несущественные индивидуальные детали сглаживаются и четко проявляются основные закономерности анизотропности поля.

7.3 Примеры комплексного анализа статистической структуры гидрометеорологических полей и экологических ареалов

Рассмотрим сущность анализа статистической структуры поля на примере речного стока и атмосферных осадков Беларуси. Если случайное поле ($\xi(\rho)$) представлено в виде независимых составляющих ($\xi(\rho)=\eta(\rho)+\delta(\rho)$), где $\eta(\rho)$ - мелкомасштабная, $\delta(\rho)$ - крупномасштабная составляющие, то его корреляционная функция, обладающая свойством аддитивности по отношению к независимым составляющим поля, может быть представлена как ($R_\xi(\rho)=R_\eta(\rho)+R_\delta(\rho)$). Для выявления соотношений между мелкомасштабной и крупномасштабной составляющими строятся эмпирические ПКФ по эмпирическим коэффициентам корреляции (r_{jk}) и соответствующим им расстояниям (ρ) между метео - , мониторинговыми станциями (центрами тяжести водосборов), которые аппроксимируются линейными зависимостями типа

$$R(\rho) = R(0) - \alpha_r \cdot \rho, \qquad (7.33)$$

характеризующими закономерность убывания эмпирических коэффициентов корреляции с увеличением расстояния между пунктами наблюде-

ний. Величина (R(0)), которой определяются значения ПКФ, при (ρ=0), как правило, меньше единицы. Это обусловлено наличием в данных наблюдений случайных ошибок, а также микроклиматических - (η(ρ)) различий в расположении станций (бассейнов). Хотя эти различия в каждом пункте вызывают систематическое расхождение, при рассмотрении гидрометеорологического поля на большой территории они выступают как случайные. Именно этими различиями, в основном, и определяется имеющий место значительный разброс коэффициентов корреляции относительно средних величин. При отсутствии ошибок измерений и микроклиматических различий имело бы место - (R(0)=1). В действительности, выполняется соотношение (7.26). Таким образом, при (R(0)<1) можно оценить, какая доля изменчивости поля определяется естественной изменчивостью рассматриваемых элементов на территории, а какая возникает за счет случайного размещения станций, погрешностей наблюдений; $\left(\alpha_r = \frac{dR(\rho)}{d\rho}\right)$ - по физическому смыслу представляет градиент поля, т.е. показывает величину изменения ПКФ на единицу расстояния. Градиент поля (α_r) служит характеристикой при совместном анализе и сопоставлении различных гидрометеорологических и экологических полей. Приведение эмпирических ПКФ к теоретическим или, точнее, откорректированных функций осуществляется путем деления каждого члена уравнения (7.33) на (R(0)). В результате чего ПКФ имеет вид

$$\hat{R}(\rho) = 1 - \hat{\alpha}_r \cdot \rho \,, \qquad (7.34)$$

где $\hat{\alpha}_r$ - приведенный градиент ПКФ поля. Величины корреляционных функций (R(0)), а также мер ошибок случайных погрешностей в исходных данных (η_M), приведенного градиента ПКФ поля ($\hat{\alpha}_r$), коэффициентов корреляции функций (7.33) стока и осадков Беларуси (r) представлены в таблице 7.1.

Годовой цикл эмпирических ПКФ месячных значений атмосферных осадков и речного стока представлен на рисунке 7.1. Чтобы заведомо не упрощать картину принятием каких-либо гипотез о виде ПКФ, годовой ход представлен не по аппроксимирующим функциям, а изокоррелятами, полученными путем интерполяции эмпирических коэффициентов корреляции. Шаг интерполяции ПКФ - (0,1). В мезомасштабной области, для различных элементов водного баланса, вклад крупномасштабной состав-

ляющей неодинаков. В метеорологических полях он значительно выше, чем в гидрологических. Поскольку коррелированность поля крупномасштабной составляющей выше, чем мелкомасштабной - значения ($R_\xi(\rho)$) завышаются для величин (ρ), не превышающих среднего масштаба крупных флуктуаций поля. Поэтому, ПКФ элементов водного баланса для большинства интервалов временной дискретизации полей являются выгнутыми, подчиняющимися экспоненциально-степенной зависимости типа

$$R(\rho) = \exp(-\alpha \cdot \rho^{\beta}) \ . \qquad (7.35)$$

Таблица 7.1 Экстраполированные значения основных характеристик ПКФ атмосферных осадков и речного стока на территории Беларуси для различных временных интервалов (R_0; η_M, ($\hat{\alpha}_r$); p)

Интервал осреднения	Атмосферные осадки				Речной сток			
	R(0)	$\eta_м$	$(\hat{\alpha}_r)10^{-3}$	r	R(0)	$\eta_м$	$(\hat{\alpha}_r)10^{-3}$	r
1	2	3	4	5	6	7	8	9
Январь	0,84	0,43	0,76	0,39	0,90	0,33	0,67	0,50
Февраль	0,76	0,56	0,70	0,36	0,83	0,44	0,54	0,42
Март	0,79	0,51	0,65	0,30	0,88	0,37	1,07	0,66
Апрель	0,80	0,50	1,40	0,69	0,82	0,46	1,03	0,56
Май	0,59	0,83	0,99	0,37	0,85	0,42	1,35	0,69
Июнь	0,59	0,83	0,90	0,43	0,73	0,60	1,56	0,62
Июль	0,66	0,71	1,32	0,50	0,80	0,50	1,59	0,70
Август	0,64	0,46	1,33	0,52	0,82	0,46	1,84	0,75
Сентябрь	0,82	0,40	1,01	0,52	0,83	0,44	1,85	0,79
Октябрь	0,86	0,42	0,91	0,42	0,90	0,33	1,82	0,77
Ноябрь	0,85	0,42	0,73	0,40	0,88	0,37	1,40	0,75
Декабрь	0,83	0,44	0,77	0,42	0,87	0,38	1,28	0,75
Год	0,71	0,64	1,08	0,48	0,84	0,43	1,61	0,79

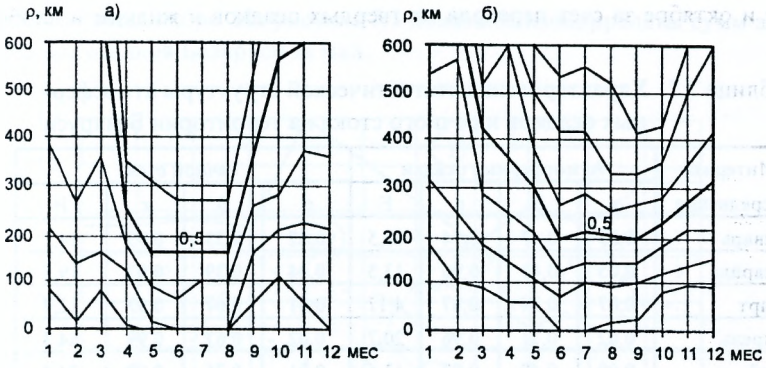


Рисунок 7.1 Годовой цикл пространственных корреляционных функций: а) атмосферных осадков; б) речного стока (при шаге интерполяции - 0,1).

Приведенная зависимость одинаково хорошо описывает поля всех элементов водного баланса, в том числе атмосферных осадков и речного стока. Недостатком зависимостей типа (7.35) является то, что они неопределимы при отрицательных коэффициентах корреляции. Значения параметров, характеризующих статистическую структуру полей атмосферных осадков и речного стока Беларуси для различных периодов осреднения, а также коэффициентов корреляции и критериев Фишера представлены в таблице 7.2. Линии регрессии (R(ρ)) атмосферных осадков для всех месяцев года вогнуты, так что с увеличением расстояния между метеостанциями убывание связности полей осадков замедляется. Для всех месяцев имеет место неравенство (R(0)<1), которое увеличивается в летние месяцы, когда роль локальной неоднородности растет за счет конвективных осадков. Наименьший разброс точек относительно эмпирической линии регрессии (R(ρ)) отличается в зимние месяцы. В пространственной коррелированности месячных сумм атмосферных осадков четко выражен годовой ход связности их полей в холодный период, которая существенно выше, чем в теплый период. Минимальная коррелированность наблюдается в летние месяцы, что вызвано несколько большей масштабностью конвективных процессов. Отметим и такую особенность, как некоторое увеличение разнородности осадков по территории в фев-

рале и октябре за счет перехода от твердых осадков к жидким и наоборот.

Таблица 7.2 Характеристика статистической структуры атмосферных осадков и речного стока на территории Беларуси

Интервал осреднения	Атмосферные осадки				Речной сток			
	α	β	r	F	α	β	r	F
Январь	0,03	0,47	0,94	10,5	0,02	0,53	0,95	14,1
Февраль	0,05	0,41	0,96	13,5	0,04	0,39	0,97	19,5
Март	0,07	0,34	0,87	4,17	0,01	0,67	0,97	12,7
Апрель	0,02	0,71	0,96	20,7	0,02	0,62	0,99	64,3
Май	0,09	0,40	0,97	13,5	0,01	0,76	0,98	24,4
Июнь	0,11	0,36	0,99	14,9	0,02	0,67	0,97	20,1
Июль	0.07	0,42	0,92	7,01	0,01	0,73	0,96	22,0
Август	0,07	0,48	0,95	11,7	0,01	0,91	0,96	20,7
Сентябрь	0,02	0,61	0,97	26,8	0,01	0,93	0,98	39,8
Октябрь	0,02	0,58	0,92	10,9	0,01	1,04	0,97	33,7
Ноябрь	0.02	0,50	0,94	10,7	0,01	0,83	0,98	45,3
Декабрь	0,03	0,49	0,96	17,0	0,01	0,70	0,96	18,4
Год	0,07	0,43	0,93	8,40	0,01	0,91	0,98	36,4

Проверка показала, что при доверительных вероятностях 68,3(%) и 95,4(%) пространственные корреляционные функции атмосферных осадков неоднородны. Разделение территории Беларуси на Черноморский и Балтийский склоны позволяет для ряда временных интервалов получить однородные поля осадков. Анизотропность поля атмосферных осадков можно оценить с помощью коэффициентов анизотропности поля (χ), которые определяются путем деления градиента поля на градиент ориентированного поля. Годовой ход коэффициентов анизотропности поля атмосферных осадков представлен на рисунке 7.2. Поля изокоррелят для каждого месяца имеют вид эллипсов, большая ось которых ориентирована в направлении преобладающего переноса воздушных масс. Так, в январе-феврале анизотропию поля сумм атмосферных осадков определяет западный перенос воздушных масс, в мае, июне, июле - северо-восточный. Наименьшая анизотропия полей осадков наблюдается в ап-

реле, августе, ноябре. На рисунке 7.3 показаны изокорреляты сумм атмосферных осадков Беларуси за май.

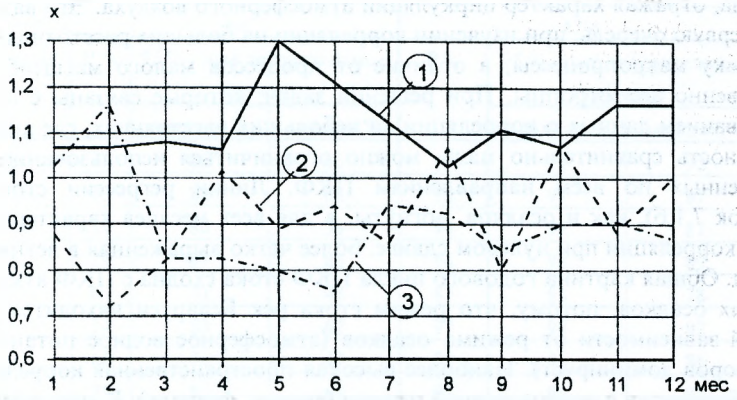


Рисунок 7.2 Годовой ход анизотропности пространственной корреляции среднемесячных сумм атмосферных осадков Беларуси по направлениям: 1 - северо-восток - юго-запад; 2 - запад - -восток; 3 - северо-запад - юго-восток.

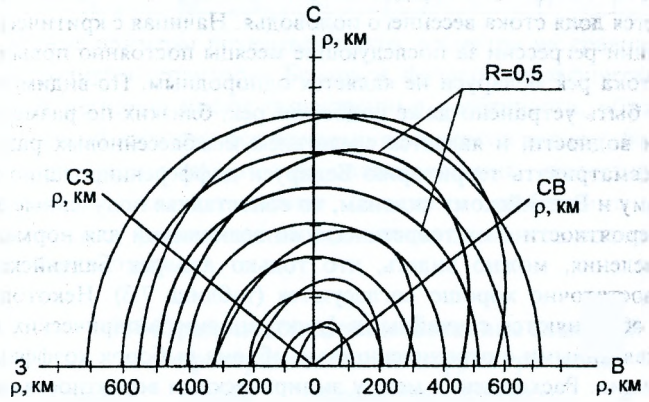


Рисунок 7.3 Поле изокоррелят атмосферных осадков Беларуси в мае (при шаге интерполяции - 0,1).

Необходимо углубленное изучение особенностей анизотропности полей элементов водного баланса, так как она меняется по территории в течение года, отражая характер циркуляции атмосферного воздуха. Это важно, в первую очередь, при изучении корреляции на больших расстояниях, поскольку макропроцессы, в отличие от процессов малого масштаба, существенно анизотропны. При решении задач, которые связаны с использованием данных о корреляции на небольших расстояниях, где анизотропность сравнительно мала, можно ограничиться использованием осредненных по всем направлениям ПКФ. Линии регрессии стока (рисунок 7.1,б), как и осадков, вогнуты, а для всех месяцев характерна срезка корреляции при нулевом сдвиге, более четко выраженная в летние месяцы. Общая картина годового цикла ПКФ стока сходна с ПКФ атмосферных осадков, потому, что режим стока рек Беларуси находится в прямой зависимости от режима осадков (атмосферное водное питание водосборов доминирует). Наиболее высокая пространственная корреляция наблюдается в период зимней межени (январь, февраль). В этот период реки имеют, преимущественно, грунтовое питание, нарушаемое отдельными оттепелями, охватывающими большие территории. Несколько меньшая, но достаточно высокая, пространственная корреляция речного стока наблюдается в весенние месяцы, в период половодья. Причем, линии регрессии за эти месяцы последовательно понижаются в той степени, как уменьшается доля стока весеннего половодья. Начиная с критического месяца, линии регрессии за последующие месяцы постоянно повышаются. Поле стока рек Беларуси не является однородным. По-видимому, это не может быть устранено даже подбором рек, близких по размерам водосборов и водности, и является следствием межбассейновых различий. Если рассматривать территорию Беларуси дифференцированно по Черноморскому и Балтийскому склонам, то сопоставляя полученные эмпирические вероятности с их теоретическими величинами для нормального распределения, можно видеть, что только для рек Балтийского склона они достаточно хорошо согласуются (таблица 7.3). Некоторые расхождения объясняются случайными флуктуациями эмпирических вероятностей, связанными с ограниченностью объема выборок коэффициентов корреляции. Расхождения между эмпирическими вероятностями и теоретическими их значениями для нормального закона распределения в большинстве случаев находятся в 95%-ной доверительной области.

Таблица 7.3 Оценка однородности пространственных корреляционных функций стока рек Балтийского склона Беларуси

Период осреднения	±σ			±2·σ		
	P,%	95 % - ные доверительные границы		P,%	95 % - ные доверительные границы	
		верхняя,%	нижняя,%		верхняя,%	нижняя,%
Январь	77	85	65	97	99	92
Февраль	73	82	62	95	97	87
Март	71	80	58	94	97	87
Апрель	79	87	68	96	98	89
Май	60	71	47	87	93	78
Июнь	74	83	62	97	99	92
Июль	60	71	47	92	96	86
Август	77	85	65	97	99	92
Сентябрь	83	90	72	96	98	83
Октябрь	67	77	55	93	97	87
Ноябрь	76	85	65	95	97	87
Декабрь	82	90	72	97	99	92
Год	72	82	61	95	97	87

Приведенные примеры показывают, что, несмотря на сравнительно небольшие размеры территории Беларуси, физико-географические условия отдельных регионов имеют существенные особенности, обуславливающие различный характер формирования элементов водного баланса. Поэтому, при планировании природопользования, проектировании и управлении водохозяйственными системами необходимо учитывать асинхронность в формировании и динамике элементов водного баланса по территории страны.

7.4 Практическое использование сведений о пространственной структуре поля

Обрабатывая данные гидрометеорологических и эколого-ландшафтных наблюдений, исследователь, *во-первых*, должен правильно осмыслить, к какому временному интервалу и к какой пространственной области целесообразно относить это наблюдение в пределах необходимой точности; *во-вторых*, уметь восстанавливать временное пространст-

венное распределение, т.е. строить поля анализируемых элементов как непосредственно измеренных, так и тесно связанных с измеряемыми величинами. Круг этих и смежных вопросов принято объединять термином *"интерпретация данных наблюдений"*. Важнейшие *элементы* и *задачи интерпретации*: *оценка точности* показаний гидрометеорологических приборов, *интерполяция, экстраполяция, согласование, объективный анализ, оценка дифференциальных характеристик* исследуемых полей, *методы контроля данных* наблюдений, *планирование сети станций* и т.д.

Каждый из этих вопросов представляет самостоятельную проблему и требует обстоятельного изложения. Остановимся на использовании элементов пространственной структуры исследуемых полей элементов для целей интерполяции (на примере стока рек Беларуси).

Восстановление стоковых характеристик методом пространственной интерполяции

Разработка водохозяйственных мероприятий, как правило, включает в себя определение расчетных характеристик речного стока. Нередко эти расчеты выполняются при отсутствии данных гидрометрических наблюдений или при наличии в них искажений антропогенного характера. Под восстановлением стока имеется ввиду пространственная интерполяция гидрологических величин, наблюденных в реперных пунктах, с целью их косвенного получения при отсутствии данных наблюдений в искомом пункте, а также корректировка в сторону уточнения измеренных данных при наличии в них антропогенных составляющих. Оптимальная пространственная интерполяция (ОПИ) сводится к следующему. Значение гидрологической величины в любой точке поля вычисляется по формуле

$$M_{i0} = \sum_{j=1}^{k} P_j \cdot M_{ij} \ , \qquad (7.36)$$

где M_{ij} - значение элемента в (i)-ый срок на (j)-ый влияющий аналог; P_j - весовые коэффициенты; k - число влияющих рек-аналогов. Чаще уравнение (7.36) решается для отклонений от среднего, т.е. в виде

$$M_{i0} = \overline{M}_0 + \sum_{j=1}^{k} P_j \cdot \Delta M_{ij} \ , \qquad (7.37)$$

где $\overline{M}_0$ - норма гидрологической характеристики в точке интерполяции; ΔM_{ij} - отклонения от нормы на (j)-ом аналоге в (i)-ый срок. После определения состава влияющих аналогов составляется система линейных

уравнений для определения интерполяционных весов (P_j) по формуле (7.5). Количество влияющих рек-аналогов для каждого отдельного случая выбирается в зависимости от наличия данных по стоку за конкретный срок наблюдений, а также от расстояния между гидропостами, поэтому, совокупность влияющих рек-аналогов не остается постоянной, а система (7.5) также индивидуальна для различных периодов. Реальное число гидропостов, привлекаемых к процедуре интерполяции, может меняться от двух до шести. Для получения результатов с требуемой для практики точностью обычно достаточно уже трех аналогов. После решения системы (7.5) интерполяцию по точкам поля можно проводить по уравнению (7.36). Для интерполяции по уравнению (7.37) необходимо определить норму гидрологической характеристики в заданной точке интерполяции одним из способов: по соответствующим картам; оптимальной пространственной интерполяцией по значениям норм. Выполненная интерполяция имеет среднюю квадратическую ошибку

$$\varepsilon^2 = 1 - \sum_{j=1}^{k} r_{0j} \cdot P_j \ . \qquad (7.38)$$

По величине (ε) просто оценить возможную точность интерполяции, использовав ее для поиска грубых ошибок в данных наблюдений. При анализе результатов интерполяции стока, можно обнаружить ошибки двух типов: ошибки в измерениях; ошибки при пространственной интерполяции, которые, в свою очередь, зависят от различий в условиях формирования стока и подбора аналогов. В большинстве случаев, близкое расположение гидропостов обеспечивает сходность физико-географических условий, высокие значения парных коэффициентов корреляции и хорошее совпадение вычисленных и измеренных величин. Среднеквадратическая ошибка оптимальной пространственной интерполяции (ОПИ) месячных величин стока, в среднем, составляет 10...20%, хотя в отдельные месяцы могут наблюдаться и большие отклонения. Годовые (сезонные) значения гидрологических характеристик определяются с большей достоверностью.

В качестве примера, *на рисунках 7.4 и 7.5 приведены результаты интерполяции модулей стока за теплый период (IV...X - месяцы) реки Ницца - створ Соколище (реки-аналоги: Дрисса - створ Демехи; Дрисса - створ Дерновичи; Свольня - створ Пользино) и за год реки Котра - створ Котра (реки-аналоги: Скиделька - створ Скидель; Невища - створ Половня; Свислочь - створ Сухая Долина).*

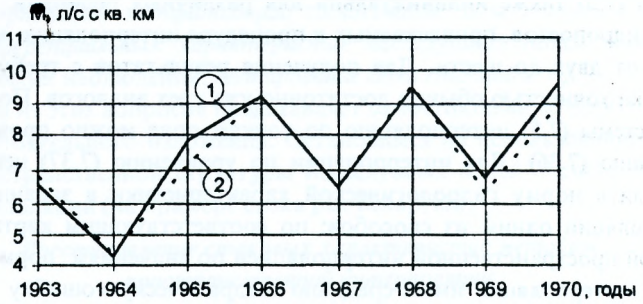


Рисунок 7.4 Наблюденные (1) и вычисленные (2) значения стока реки Ницца - створ Соколище (теплый период).

Суммарная относительная ошибка интерполяции и исходных данных составляет 5...10 процентов от средних величин модуля стока. При восстановлении месячных величин стока контроль осуществляется путем сопоставления суммы месячных величин (за год) с годовыми значениями. Если невязка получается меньше допустимой, ее распределяют помесячно пропорционально абсолютным величинам стока. В противном случае, требуется дополнительный анализ как исходной информации, так и репрезентативности аналогов. В таблице 7.4 представлены результаты интерполяции модулей норм годового стока реки Цны - створ Дятловичи при благоприятном выборе влияющих рек-аналогов (Ясельда - створ Сенин; Бобрик - створ Парохонск; Птичь - створ Лучицы).

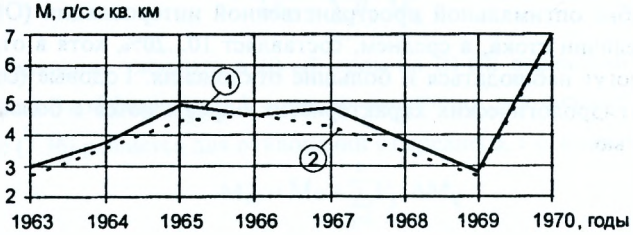


Рисунок 7.5 Наблюденные (1) и вычисленные (2) значения годового стока реки Котра - створ Котра.

Таблица 7.4 Наблюденные и вычисленные значения модулей годового стока реки Цна- -створ Дятловичи

Характеристики	Месяцы					
	январь	февраль	март	апрель	май	июнь
Наблюденный сток	1,77	2,10	6,18	16,1	6,29	2,49
Вычисленный сток	2,09	2,18	6,01	14,0	6,98	2,83
Среднеквадратическая ошибка	0,18	0,02	0,03	0,13	0,11	0,14

→ *Продолжение таблицы 7.4*

Характеристики	Месяцы					
	июль	август	сентябрь	октябрь	ноябрь	декабрь
Наблюденный сток	1,19	0,75	0,75	1,05	2,32	2,55
Вычисленный сток	1,73	1,08	1,05	1,00	2,53	3,04
Среднеквадратическая ошибка	0,45	0,24	0,17	0,02	0,09	0,19

Аналоги практически равномерно расположены вокруг пункта приведения, поэтому, достигаются хорошие совпадения наблюденных и вычисленных величин. *При использовании рек-аналогов, находящихся на значительном удалении, расхождения между измеренными и вычисленными значениями стока значительны, хотя в общих чертах* ***восстановленные величины отражают естественный характер колебаний модулей годового стока.***

ЗАКЛЮЧЕНИЕ

Специалисты отрасли "Мелиорация и водное хозяйство", при решении практических инженерных задач, используют закономерности процессов тепловлаго-массообмена деятельной поверхности (почвогрунтов), глубина и устойчивость связей в которых зависят от комплекса региональных, зональных и локальных характеристик мелиорируемых земель. Оптимизация природопользования предполагает синтез разрозненных знаний и информации по конкретным направлениям решаемых проблем, прогнозную оценку состояния природной Среды, в целом, и ее компонентов, в частности, аналитические проработки, расчеты, практические решения, обеспечивающие равновесие экосистем, экологическую безопасность мелиорируемых территорий. В процессе предварительной обработки экспериментальных данных, как показывает опыт подготовки специалистов высшей квалификации, широко используются знания и элементы общей теории ошибок, числовые характеристики генеральной совокупности и выборок, эмпирические и теоретические распределения, положения теории оценок и статистических гипотез. Используя теорию корреляции, студенты установят не только форму той или иной корреляционной зависимости, но и определят вид функции регрессии одной переменной (случайной) величины по другой, оценят тесноту корреляционных зависимостей. Логическим продолжением корреляционного анализа явится регрессионный анализ, который позволяет развивать и углублять полученные представления в корреляционных связях, выбирать оптимальные расчетные зависимости и модели. Исходя из практики анализа временных рядов, устанавливается, что общий ход или колебание во времени гидрометеорологической, гидрологической, тепловоднобалансовой или другой случайной характеристики представляет собой сумму нескольких колебаний. Суть этого анализа заключается, во-первых, в разделении временных рядов на периодические и непериодические компоненты и, во-вторых, - в изучении каждой из компонент, в отдельности. При решении практических задач природопользования и постановке прикладных комплексных исследований найдут свое применение дисперсионный анализ, статистические методы планирования эксперимента, методы пространственного обобщения гидрометеорологической и экологической информации.

ЛИТЕРАТУРА

1 Алехин Ю.М. Статистические прогнозы в геофизике. - Л.: Изд-во ЛГУ, 1963.-59 с.

2 Бондаренко Н.Ф. Физические основы мелиорации почв. - Л.: Колос, 1975. - 258 с.

3 Боровиков В.П., Боровиков И.П. Статистический анализ и обработка данных в среде Windows. - М.: Информационно-издательский дом "Филинъ", 1997. - 608 с.

4 Бочаров М.К. Методы математической статистики в географии. -М.: Мысль, 1971. - 371 с.

5 Брукс К., Карузерс Н. Применение статистических методов в метеорологии: Пер. с англ. Е.Ф. Ивановой, Л.Л. Френкеля; Под ред. Н.А. Багрова. - Л.: Гидрометеоиздат, 1963. - 416 с.

6 Вальвачев Н.И., Римша М.И. Статистический метод в медицинской практике с применением микро - ЭВМ и персональных компьютеров. - Минск: Беларусь, 1989. - 112 с.

7 Верещагин М.А., Наумов Э.П., Шанталинский К.М. Статистические методы в метеорологии. - Казань: Изд-во Казанского университета, 1990. - 110 с.

8 Доспехов Б.А. Методика полевого опыта (с основами статистической обработки результатов исследований). - 5-е изд., доп. и перераб. - М.: Агропромиздат, 1985. - 351 с.

9 Дрейпер Н., Смит Г. Прикладной регрессионный анализ: В 2-х кн. Кн. 1 / Пер. с англ. - 2-е изд., перераб. и доп. - М.: Финансы и статистика, 1986. - 366 с.

10 Дэвис Дж. С. Статистический анализ данных в геологии: Пер. с англ. в 2 кн. / Пер. В.А. Голубевой; Под ред. Д.А. Родионова. Кн. 1 - М.: Недра, 1990. - 319 с.

11 Дэвис Дж. С. Статистический анализ данных в геологии: Пер. с англ. в 2 кн. / Пер. В.А. Голубевой; Под ред. Д.А. Родионова. Кн.2 - М.: Недра, 1990. - 427 с.

12 Герасимович А.И. Математическая статистика. - Минск: Высшая школа, 1983. - 279 с.

13 Гильдерман Ю.И. Закон и случай. - Новосибирск: Наука, 1991. - 200 с.

14 Исаев А.А. Статистика в метеорологии и климатологии. - М.: Изд-во МГУ, 1988, - 248 с.
15 Каждан А.Б., Гуськов О.И. Математические методы в геологии. -М.: Высшая школа, 1983. - 251 с.
16 Климат Беларуси / Под ред. В.Ф. Логинова. - Минск: Институт геологических наук АН Беларуси, 1996. - 234 с.
17 Колде Я.К. Практикум по теории вероятностей и математической статистике. - М.: Высшая школа, 1991. - 157 с.
18 Красовский Г.И., Филаретов Г.Ф. Планирование эксперимента. -Минск: Изд-во БГУ, 1982. - 302 с.
19 Литтл Р. Дж. А., Рубин Д.Б. Статистический анализ данных с пропусками / Пер. с англ. - М.: Финансы и статистика, 1990. - 336 с.
20 Львовский Е.Н. Статистические методы построения эмпирических формул. - М.: Высшая школа, 1988. - 239 с.
21 Мелиорация: Энцикл. справочник / [Редкол.: И.П. Шамякин (гл. ред.) и др.; Под общ. ред. А.И. Мурашко]. - Минск: Белорус. Сов. Энцикл., 1984. - 567 с.
22 Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия: В 2-х вып. Вып.1 / Пер. с англ. Ю.Н. Благовещенского; Под ред. и с предисл. Ю.П. Адлера. - М.: Финансы и статистика, 1982. - 317 с.
23 Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия: В 2-х вып. Вып.2 / Пер. с англ. Ю.Н. Благовещенского; Под ред. и с предисл. Ю.П. Адлера. - М.: Финансы и статистика, 1982. - 239 с.
24 Научно-прикладной справочник по климату СССР / Многолетние данные. Вып. 7. - Л.: Гидрометеоиздат, 1987. - 301 с.
25 Основы научных исследований. Гидромелиорация / Вознюк Т.С., Гочаров С.М., Ковалев С.В. - Киев: Вища школа, 1985. - 192 с.
26 Пановский Г.А., Брайер Г.В. Статистические методы в метеорологии: Пер. с англ. / Пер. И.П. Гейбера, В.А. Шнайдмана; Под ред. Л.С. Гандина, Р.Л. Кагана. - Л.: Гидрометеоиздат, 1972. - 210 с.
27 Поллард Дж. Справочник по вычислительным методам статистики / Пер. с англ. В.С. Занадворова: Под ред. и с предисл. Е.М. Четыркина. - М.: Финансы и статистика, 1982. - 344 с.
28 Проектирование водохозяйственных систем: Пер. с чеш. Г.В. Шевелева; Под ред. В.Х. Отмана. - М.: Стройиздат, 1984. - 368 с.

29 Репин С.В., Шеин С.А. Математические методы обработки статистической информации с помощью ЭВМ. - Минск: Изд-во "Университетское", 1990. - 128 с.
30 Рождественский А.В., Чеботарев А.И. Статистические методы в гидрологии. - Л.: Гидрометеоиздат, 1974. - 424 с.
31 Статистические методы в гидрологии. - Л.: Гидрометеоиздат, 1970. - 270 с.
32 Статистические методы в инженерных исследованиях (лабораторный практикум) / Бородюк В.П., Вощинин А.П., Ива-нов А.З. и др.; Под ред Г.К. Круга. - М.: Высшая школа, 1983. - 216 с.
33 Тейлор Дж. Введение в теорию ошибок: Пер. с англ. - М.: Мир, 1985. - 272 с.
34 Уланова Е.С., Сиротенко О.Д. Методы статистического анализа в агрометеорологии. - Л.: Гидрометеоиздат, 1968. - 198 с.
35 Чарыков А.К. Математическая обработка результатов химического анализа. - Л.: Химия, 1984. - 168 с.
36 Чертко Н.К. Математические методы в физической географии. - Минск.: Изд-во "Университетское", 1987. - 151 с.
37 Чини Р.Ф. Статистические методы в геологии: Пер. с англ. - М.: Мир, 1986. - 189 с.
38 Эколого-социальные аспекты освоения водно-земельных ресурсов и технологий управления режимами гидромелиораций / П.В. Шведовский, В.Е. Валуев, А.А. Волчек, В.Г. Федоров. - Минск: Ураджай, 1998. - 363 с.

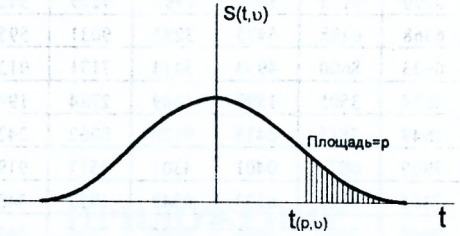
ПРИЛОЖЕНИЕ

Таблица П.1 Случайные числа

3393	6270	4228	6069	9407	1865	8549	3217	2351	8410
9108	2330	2157	7416	0398	6173	1703	8132	9065	6717
7891	3590	2502	5945	3402	0491	4328	2365	6175	7695
9085	6307	6910	9174	1753	1797	9229	3422	9861	8357
2638	2908	6368	0398	5495	3283	0031	5955	6544	3883
1313	8338	0623	8600	4950	5414	7131	0134	7241	0651
3897	4202	3814	3505	1599	1649	2784	1994	5775	1406
4380	9543	1649	2815	8415	9120	8062	2421	6161	4634
1618	6309	7909	0874	0401	4301	4517	9197	3350	0434
4858	4676	7363	9141	6133	0549	1972	3461	7116	1496
5354	9142	0847	5393	5416	6505	7156	5634	9703	6221
0905	6986	9396	3975	9255	0537	2479	4589	0562	5345
1420	0470	8679	2328	3939	1292	0406	5428	3789	2882
3218	9080	6604	1813	8209	7039	2086	3369	4437	3798
9697	8431	4387	0622	6893	8788	2320	9358	5904	9539
0912	4964	0502	9683	4636	2861	2876	1273	7870	2030
4636	7072	4868	0601	3894	7182	8417	2367	7032	1003
2515	4734	9897	6761	5636	2949	3979	8650	3430	0635
5964	0412	5012	2369	6461	0678	3693	2928	3740	8047
7848	1523	7904	1521	1455	7089	8094	9872	0898	7174
5182	2571	3643	0707	3434	6818	5729	8615	4298	4129
8438	8325	9886	1805	0226	2310	3675	5058	2515	2388
8166	6349	0319	5436	6838	2460	6433	0644	7428	8556
9158	8263	6504	2562	1160	1526	1816	9690	1215	9590
6061	3525	4048	0382	4224	7148	8256	6526	5340	4064

Таблица П.2 Критические точки распределения Стьюдента
(t-распределение)

В таблице приведены значения квантилей ($t_{p,\nu}$) в зависимости от числа степеней свободы (ν) и вероятности (p).



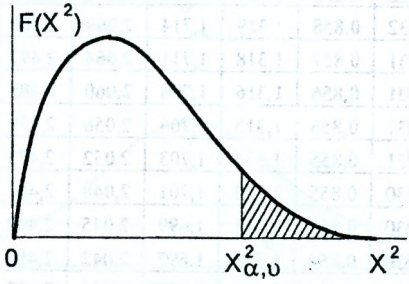
	Вероятность, (p)									
ν	0.40	0.30	0.20	0.10	0.050	0.025	0.010	0.005	0.001	0.0005
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1	0.325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	5,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	3,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850

Продолжение таблицы П.2

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,262	3,495
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Таблица П.3 Критические точки распределения Пирсона (χ^2-распределение)

В таблице приведены *значения квантилей* ($\chi^2_{\alpha,\nu}$) *в зависимости от* числа *степеней свободы* (ν) *и уровней значимости* (α).



Число степеней свободы (ν)	Уровень значимости (α)							
	0,99	**0,95**	**0,75**	**0,50**	**0,25**	**0,10**	**0,05**	**0,01**
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	...	...	0,10	0,45	1,32	2,71	3,84	6,63
2	0,02	0,10	0,58	1,39	2,77	4,61	5,99	9,21
3	0,11	0,35	1,21	2,37	4,11	6,25	7,81	11,34
4	0,30	0,71	1,92	3,36	5,39	7,78	9,49	13,28
5	0,55	1,15	2,67	4,35	6,63	9,24	11,07	15,09
6	0,87	1,64	3,45	5,35	7,84	10,64	12,59	16,81
7	1,24	2,17	4,25	6,35	9,04	12,02	14,07	18,48
8	1,65	2,73	5,07	7,34	10,22	13,36	15,51	20,09
9	2,09	3,33	5,90	8,34	11,39	14,68	16,92	21,67
10	2,56	3,94	6,74	9,34	12,55	15,99	18,31	23,21
11	3,05	4,57	7,58	10,34	13,70	17,28	19,68	24,72
12	3,57	5,23	8,44	11,34	14,85	18,55	21,03	26,22
13	4,11	5,89	9,30	12,34	15,98	19,81	22,36	27,69
14	4,66	6,57	10,17	13,34	17,12	21,06	23,68	29,14
15	5,23	7,26	11,04	14,34	18,25	22,31	25,00	30,58
16	5,81	7,96	11,91	15,34	19,37	23,54	26,30	32,00
17	6,41	8,67	12,79	16,34	20,49	24,77	27,59	33,41

Продолжение таблицы П.3

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
18	7,01	9,39	13,68	17,34	21,60	25,99	28,87	34,81
19	7,63	10,12	14,56	18,34	22,72	27,20	30,14	36,19
20	8,26	10,85	15,45	19,34	23,83	28,41	31,41	37,57
21	8,90	11,59	16,34	20,34	24,93	29,62	32,67	38,93
22	9,54	12,34	17,24	21,34	26,04	30,81	33,92	40,29
23	10,20	13,09	18,14	22,34	27,14	32,01	35,17	41,64
24	10,86	13,85	19,04	23,34	28,24	33,20	36,42	42,98
25	11,52	14,61	19,94	24,34	29,34	34,38	37,65	44,31
26	12,20	15,38	20,84	25,34	30,43	35,56	38,89	45,64
27	12,88	16,15	21,75	26,34	31,53	36,74	40,11	46,93
28	13,56	16,93	22,66	27,34	32,62	37,92	41,34	48,28
29	14,26	17,71	23,57	28,34	33,71	39,09	42,56	49,59
30	14,95	18,49	24,48	29,34	34,80	40,26	43,77	50,89
40	22,16	26,51	33,66	39,34	45,62	51,80	55,76	63,69
50	29,71	34,76	42,94	49,33	56,33	63,17	67,50	76,15
60	37,48	43,19	52,29	59,33	66,98	74,40	79,08	88,38
70	45,44	51,74	61,70	69,33	77,58	85,53	90,53	100,42
80	53,54	60,39	71,14	79,33	88,13	96,58	101,88	112,33
90	61,75	69,13	80,62	89,33	98,64	107,56	113,14	124,12
100	70,06	77,93	90,13	99,33	109,14	118,50	124,34	135,81

Таблица П.4.1 Критические точки F-распределения Фишера на 5 % - ном уровне значимости (значения отношений дисперсий "F" при различных степенях свободы "ν_1" и "ν_2" выборок)

При степенях свободы меньшей дисперсии (знаменатель)	**При степенях свободы большей дисперсии (числитель)**													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**24**	**50**	**100**
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1	161	200	216	225	230	234	237	239	241	242	244	249	252	253
2	18,51	19,00	19,16	19,33	19,30	19,33	19,36	19,37	19,38	19,39	19,41	19,45	19,47	19,49
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,74	8,64	8,58	8,56
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,77	5,70	5,66
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,68	4,53	4,44	4,40
6	5,99	5,14	4,76	4,53	4,39	4,27	4,21	4,15	4,10	4,06	4,00	3,84	3,75	3,71
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,57	3,41	3,32	3,28
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,28	3,12	3,03	2,98
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,07	2,90	2,80	2,76
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,91	2,74	2,64	2,59
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,79	2,61	2,50	2,45
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,69	2,50	2,40	2,35
13	4,64	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,60	2,42	2,32	2,26
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,53	2,35	2,24	2,19
15	4,54	3,60	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,48	2,29	2,18	2,12

Продолжение таблицы П.4.1

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,24	2,13	2,07
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,38	2,19	2,08	2,02
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,15	2,04	1,98
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,31	2,11	2,00	1,94
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,28	2,08	1,96	1,90
21	4,32	3,47	3,07	2,84	2,68	2,54	2,49	2,42	2,37	2,32	2,25	2,05	1,93	1,87
22	4,30	3,44	3,05	2,84	2,66	2,55	2,41	2,40	2,35	2,30	2,23	2,03	1,91	1,84
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,38	2,32	2,28	2,20	2,00	1,88	1,82
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,18	1,98	1,86	1,80
25	4,24	3,38	2,99	2,76	2,60	2,49	2,41	2,34	2,25	2,24	2,16	1,96	1,84	1,77
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	1,95	1,82	1,76
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19	2,12	1,91	1,78	1,72
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,12	2,09	1,89	1,76	1,69
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,00	1,79	1,66	1,59
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,95	1,74	1,60	1,52

Таблица П.4.2 Критические точки F-распределения Фишера на 1 % - ном уровне значимости (значения отношений дисперсий "F" при различных степенях свободы "ν_1" и "ν_2" выборок)

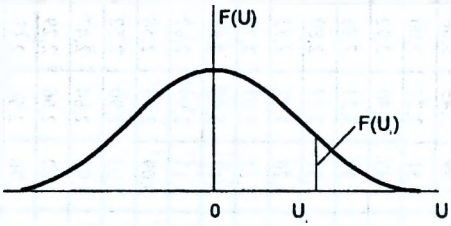
При степенях свободы меньшей дисперсии (знаменатель)	**При степенях свободы большей дисперсии (числитель)**													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**24**	**50**	**100**
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056	6106	6234	6302	6334
2	98,49	99,01	99,17	99,25	99,30	99,33	99,34	99,36	99,38	99,40	99,42	99,46	99,48	99,49
3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,60	26,35	26,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	11,66	14,54	14,37	13,93	13,69	13,57
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,89	9,47	9,24	9,13
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,31	7,09	6,99
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,47	6,07	5,85	5,75
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,67	5,28	5,06	4,96
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,11	4,73	4,51	4,41
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,71	4,33	4,12	4,01
11	9,85	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,40	4,02	3,80	3,70
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,16	3,78	3,56	3,46
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,59	3,37	3,27
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,43	3,21	3,11
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,29	3,07	2,97

Продолжение таблицы П.4.2

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
16	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,78	3,69	3,61	3,45	3,18	2,96	2,86
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,45	3,08	2,86	2,76
18	8,28	6,01	5,09	5,58	4,25	4,01	3,85	3,71	3,60	3,51	3,37	3,00	2,78	2,68
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,68	3,52	2,43	3,30	2,92	2,70	2,63
20	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37	3,23	2,86	2,63	2,53
21	8,02	5,78	4,87	4,37	4,04	3,81	3,65	3,51	3,40	3,31	3,17	2,80	2,58	2,47
22	7,94	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,75	2,53	2,42
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,70	2,48	2,37
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,25	3,17	3,03	2,66	2,44	2,33
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,21	3,13	2,99	2,62	2,40	2,29
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,17	3,09	2,96	2,58	2,36	2,25
28	7,64	5,45	4,57	4,07	3,76	3,53	3,36	3,23	3,11	3,03	2,90	2,52	2,30	2,18
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06	2,98	2,84	2,47	2,24	2,13
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,88	2,80	2,66	2,29	2,05	1,94
50	7,17	5,06	4,20	3,72	3,41	3,18	3,02	2,88	2,78	2,70	2,56	2,18	1,94	1,81
100	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59	2,51	2,36	1,98	1,73	1,59

Таблица П.5 Нормальное распределение. Плотность вероятностей нормированного нормального распределения

$$u \to N(0,1) \qquad f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$



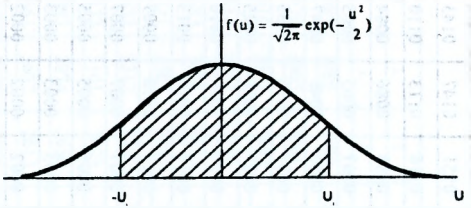
u	0	1	2	3	4	5	6	7	8	9
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804

Продолжение таблицы П.5

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Таблица П.6 Нормальное распределение. *Значение функции.*

$$\Phi(u_i) = \frac{2}{\sqrt{2\pi}} \int_0^{u_i} \exp(-\frac{x^2}{2}) dx = P(|u| < u_i).$$



Целые и десятые доли	Сотые доли									
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0,0	0,0000	0,0080	0,0160	0,0239	0,0319	0,3999	0,0478	0,0558	0,0638	0,0717
0,1	0797	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2960	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778

Продолжение таблицы П.6

Целые и десятые доли	Сотые доли									
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1,0	**0,6827**	**0,6875**	**0,6923**	**0,6970**	**0,7017**	**0,7063**	**0,7109**	**0,7154**	**0,7199**	**0,7243**
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7994	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	**0,9545**	**0,9556**	**0,9566**	**0,9576**	**0,9586**	**0,9596**	**0,9606**	**0,9616**	**0,9625**	**0,9634**
2,1	9643	9651	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9841	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9910	9912	9915	9917	9920	9922	9924	9926	9928
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947

Продолжение таблицы П.6

Целые и десятые доли	Сотые доли									
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
2,8	9949	9951	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	**0,9973**	**0,9974**	**0,9975**	**0,9976**	**0,9976**	**0,9977**	**0,9978**	**0,9979**	**0,9979**	**0,9980**
3,1	9981	9981	9982	9983	9983	9984	9984	9985	9985	9986
3,2	9986	9987	9987	9988	9988	9989	9989	9989	9990	9990
3,3	9990	9991	9991	9991	9992	9992	9992	9992	9993	9993
3,4	9993	9994	9994	9994	9994	9994	9995	9995	9995	9995
3,5	9995	9996	9996	9996	9996	9996	9996	9996	9997	9997
3,6	9997	9997	9997	9997	9997	9997	9997	9998	9998	9998
3,7	9998	9998	9998	9998	9998	9998	9998	9998	9998	9998
3,8	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
3,9	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,0	**0,999936**	**9999**	**9999**	**9999**	**9999**	**9999**	**9999**	**9999**	**9999**	**9999**
4,5	0,999994	–	–	–	–	–	–	–	–	–
5,0	0,99999994	–	–	–	–	–	–	–	–	–

Таблица П.7.1 G-распределение Кохрена (α=0,05)

Значения ($G_{\nu,k,\alpha}$) в зависимости от числа степеней свободы (ν), (k) и фиксированной вероятности (α):

$$P\{G>G_{\nu,k,\alpha}\}=\alpha$$

k\ν	1	2	3	4	5	6	7	8	9	10	16	36	144	∞
2	0,9985	0,9750	0,9392	0,9057	0,8772	0,8534	0,8332	0,8159	0,8010	0,7880	0,7341	0,6602	0,5813	0,5000
3	9669	8709	7977	7457	7071	6771	6530	6333	6167	6025	5466	4748	4031	3333
4	9065	7679	6841	6287	5895	5598	5365	5175	5017	4884	4366	3720	3093	2500
5	0,8412	0,6838	0,5981	0,5440	0,5063	0,4783	0,4564	0,4387	0,4241	0,4118	0,3645	0,3066	0,2513	0,2000
6	7808	6161	5321	4803	4447	4184	3980	3817	3682	3568	3135	2612	2119	1667
7	7271	5612	4800	4307	3974	3726	3535	3384	3259	3154	2756	2278	1833	1429
8	0,6798	0,5157	0,4377	0,3910	0,3595	0,3362	0,3185	0,3043	0,2926	0,2829	0,2462	0,2022	0,1616	0,1250
9	6385	4785	4027	3584	3286	3067	2901	2768	2659	2568	2226	1820	1446	1111
10	6020	4450	3733	3311	3029	2823	2666	2541	2439	2353	2032	1655	1308	1000
12	0,5410	0,3924	0,3264	0,2880	0,2624	0,2439	0,2299	0,2187	0,2098	0,2020	0,1737	0,1403	0,1100	0,0833
15	4709	3346	2758	2419	2195	2034	1911	1815	1736	1671	1429	1144	0889	0667
20	3894	2705	2205	1921	1735	1602	1501	1422	1357	1303	1108	0879	0675	0500
24	0,3434	0,2354	0,1907	0,1656	0,1493	0,1374	0,1286	0,1216	0,1160	0,1113	0,0942	0,0743	0,0567	0,0417
30	2929	1980	1593	1377	1237	1137	1061	1002	0958	0921	0771	0604	0457	0333
40	2370	1576	1259	1082	0968	0887	0827	0780	0745	0713	0595	0462	0347	0250
60	0,1737	0,1131	0,0895	0,0766	0,0682	0,0623	0,0583	0,0552	0,0520	0,0497	0,0411	0,0316	0,0234	0,0167
120	0998	0632	0495	0419	0371	0337	0312	0292	0279	0266	0218	0165	0120	0083

Таблица П.7.2 G-распределение Кохрена (α=0,01)

Значения ($G_{\nu,k,\alpha}$) в зависимости от числа степеней свободы (ν), (k) и фиксированной вероятности (α):

$$P\{G>G_{\nu,k,\alpha}\}=\alpha$$

k\v	1	2	3	4	5	6	7	88	9	10	16	36	144	∞
2	0,9999	0,9950	0,9794	0,9586	0,9373	0,9172	0,8988	0,8823	0,8674	0,8539	0,7949	0,7067	0,6062	0,5000
3	9933	9423	8831	8355	7933	7606	7335	7107	6912	6743	6059	5153	4230	3333
4	9676	8643	7814	7212	6761	6410	6129	5897	5702	5536	4884	4057	3251	2500
5	0,9279	0,7885	0,6957	0,6329	0,5875	0,5531	0,5259	0,5037	0,4854	0,4697	0,4094	0,3351	0,2644	0,2000
6	8828	7218	6258	5635	5195	4866	4608	4401	4229	4084	3529	2858	2229	1667
7	8376	6644	5685	5080	4659	4347	4105	3911	3751	3616	3105	2494	1929	1429
8	0,7945	0,6162	0,5209	0,4627	0,4226	0,3932	0,3704	0,3522	0,3373	0,3248	0,2779	0,2214	0,1700	0,1250
9	7544	5727	4810	4251	3870	3592	3378	3207	3067	2950	2514	1992	1521	1111
10	7175	5358	4469	3934	3572	3308	3106	2945	2813	2704	2297	1811	1376	1000
12	0,6528	0,4751	0,3919	0,3428	0,3099	0,2861	0,2680	0,2535	0,2419	0,2320	0,1961	0,1535	0,1157	0,0833
15	5747	4069	3317	2882	2593	2386	2228	2104	2002	1918	1612	1251	0934	0667
20	4799	3297	2654	2288	2048	1877	1748	1646	1567	1501	1248	0960	0709	0500
24	0,4247	0,2871	0,2295	0,1970	0,1759	0,1608	0,1495	0,1406	0,1338	0,1283	0,1060	0,0810	0,0595	0,0417
30	3632	2412	1913	1635	1454	1327	1232	1157	1100	1054	0867	0658	0480	0333
40	2940	1915	1508	1281	1135	1033	0957	0898	0853	0816	0668	0503	0363	0250
60	0,02151	0,1371	0,1069	0,0902	0,0796	0,0722	0,0668	0,0625	0,0594	0,0567	0,0461	0,0344	0,0245	0,0167
120	1252	0759	0585	0489	0429	0387	0357	0334	0316	0302	0242	0178	0125	0083

Таблица П.8 Критические значения коэффициентов корреляции при различных уровнях значимости и числах степеней свободы (ν=n-2)

Степеней свободы (n-2)	0,05	0,01	Степеней свободы (n-2)	0,05	0,01	Степеней свободы (n-2)	0,05	0,01
1	0,997	1,000	16	0,468	0,590	35	0,325	0,418
2	0,950	0,990	17	0,456	0,575	40	0,304	0,393
3	0,878	0,959	18	0,444	0,561	45	0,288	0,372
4	0,811	0,917	19	0,433	0,549	50	0,273	0,354
5	0,754	0,874	20	0,423	0,537	60	0,250	0,325
6	0,707	0,834	21	0,413	0,526	70	0,232	0,302
7	0,666	0,798	22	0,404	0,515	80	0,217	0,283
8	0,632	0,765	23	0,396	0,505	90	0,205	0,267
9	0,602	0,735	24	0,388	0,496	100	0,195	0,254
10	0,576	0,708	25	0,381	0,487	150	0,159	0,208
11	0,553	0,684	26	0,374	0,478	200	0,138	0,181
12	0,532	0,661	27	0,367	0,470	300	0,113	0,148
13	0,514	0,641	28	0,361	0,463	400	0,098	0,128
14	0,497	0,623	29	0,355	0,456	500	0,088	0,115
15	0,482	0,606	30	0,349	0,449	1000	0,062	0,081

Таблица П.9 Критические значения К - С - критерия

n	$D_{0,10}$	$D_{0,05}$	n	$D_{0,10}$	$D_{0,05}$
3	0,636	0,708	23	0,247	0,275
4	0,565	0,624	24	0,242	0,269
5	0,509	0,563	25	0,238	0,264
6	0,468	0,519	26	0,233	0,259
7	0,436	0,483	27	0,229	0,254
8	0,410	0,454	28	0,225	0,250
9	0,387	0,430	29	0,221	0,246
10	0,369	0,409	30	0,218	0,242
11	0,352	0,391	31	0,214	0,238
12	0,338	0,375	32	0,211	0,234
13	0,325	0,361	33	0,208	0,231
14	0,314	0,349	34	0,205	0,227
15	0,304	0,338	35	0,202	0,224
16	0,295	0,327	36	0,199	0,221
17	0,286	0,318	37	0,196	0,218
18	0,278	0,309	38	0,194	0,215
19	0,271	0,301	39	0,191	0,213
20	0,265	0,294	40	0,189	0,210
21	0,259	0,287	50	0,170	0,177
22	0,253	0,281	100	0,121	0,134

ТаблицаП.10 Сводная таблица распределений
(дискретные случайные величины)

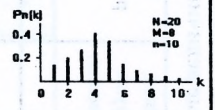
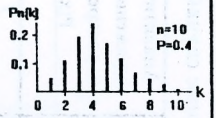
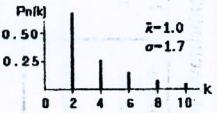
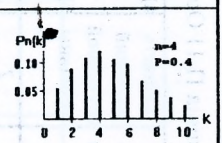
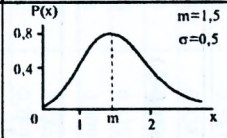
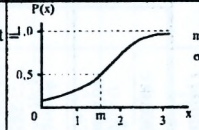
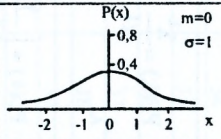
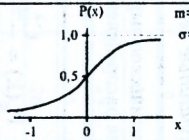
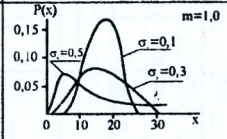
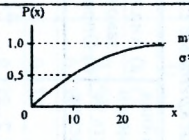
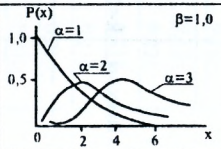
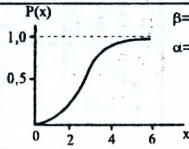
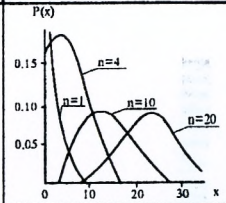
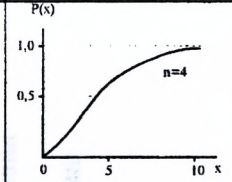
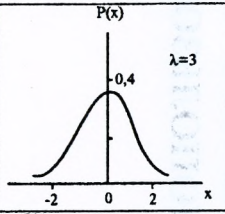
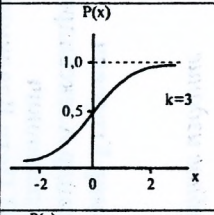
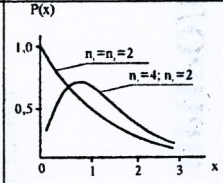
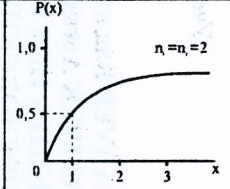
Закон распределения	Случайная величина и область ее изменения	а) Аналитическое выражение закона распределения б) Определяющие параметры	График закона распределения
1 Гипергеометрический	k=0,1,2,..., min(M,n)	а) $P_n(k)=\dfrac{C_M^k \cdot C_{N-M}^{n-k}}{C_M^n}$ б) N, M, n	
2 Биномиальный (Бернулли)	k=0,1,2,...,n	а) $P_n(k)=C_n^k \cdot P^k \cdot (1-P)^{n-k}$ б) n, P	
3 Пуассона	k=0,1,2,...,n	а) $P_n(k)=\dfrac{\lambda^k}{k!}\cdot\exp(-\lambda)$ б) λ	
4 Сложное распределение Пуассона	k=0,1,2,...,n	а) $P_n(k)=\left(\dfrac{\gamma}{\gamma+1}\right)^{\alpha}\cdot \dfrac{\alpha\cdot(\alpha+1)\cdot(\alpha+k-1)}{k!}\cdot \left(\dfrac{1}{\gamma+1}\right)^{k}$ $\alpha=\dfrac{\bar{k}^2}{\sigma^2\cdot\bar{k}};\quad \gamma=\dfrac{\bar{k}}{\sigma^2-\bar{k}}$ б) $\bar{k},\sigma$	
5 Отрицательный биномиальный закон	k=n,n+1,... или k=0,1,2,...	а) $P_n(k)=C_{k-1}^{n-1}\cdot P^n\cdot(1-P)^{k-n}$ или $P_n(k)=C_{k+n-1}^{n-1}\cdot P^n\cdot(1-P)^k$ б) n, P	

Таблица П.11 Основные законы распределений (непрерывные случайные величины)

Закон распределения	Область значений случайной величины	а) аналитическое выражение плотности вероятности, P(x) б) определяющие параметры	График плотности вероятности, P(x)	Аналитическое выражение функции распределения, P(x)	График функции распределения, P(x)
1	*2*	*3*	*4*	*5*	*6*
1 Нормальный (Гаусса)	$-\infty<x<\infty$	а) $\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2}\right)$ б) m,σ $\quad m=\bar{x}$		$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{(t-m)^2}{2\sigma^2}\right)dt=$ $=\Phi\left(\frac{x-m}{\sigma}\right)$	
2 Нормальный стандартный	$-\infty<x<\infty$	а) $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ б) $m=0,\quad \sigma=1$		$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-t^2/2\right)dt=$ $=\Phi(x)$	
3 Логарифмически нормальный	$0<x<\infty$	а) $\frac{\log e}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log x-m)^2}{2\sigma^2}\right)$ $m=M(\log x)$ $\sigma^2=D(\log x)$ б) m,σ		$\begin{cases}0 \text{ при } x<0\\ \Phi\left(\frac{\log x-m}{\sigma}\right) \text{ при } x>0\end{cases}$	
4 Гамма- распределение	$0<x<\infty$	а) $\frac{1}{\beta^{\alpha+1}\Gamma(\alpha+1)}x^{\alpha}\exp\left(-\frac{x}{\beta}\right)$ б) $\alpha,\beta \quad \alpha>-1,\beta>0$		$\begin{cases}0 \text{ при } x\le 0\\ \frac{\Gamma(\alpha+1;x/\beta)}{\Gamma(\alpha+1)}\\ \text{при } x>0\end{cases}$	

Продолжение таблицы П.11

1	2	3	4	5	6
5 χ^2 - распределение	$0<x<\infty$	а) $\frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} n^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right)$ б) $n = \ell$		$\begin{cases} 0 & \text{при } x < 0 \\ \frac{\Gamma\left(\frac{n}{2}; \frac{x}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} & \text{при } x > 0 \end{cases}$	
6 Стьюдента (t-распределения)	$-\infty<x<\infty$	а) $\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$ б) $k = \ell$		$\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot \int_{-\infty}^{\infty} \left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}} dt$	
7 Фишера (F-распределения)	$0<x<\infty$	а) $\frac{\Gamma\frac{n_1+n_2}{2}}{\Gamma\left(\frac{n_1}{2}\right)\cdot\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{n_2}} \cdot x^{\frac{n_1}{2}-1} \left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}$ б) n_1, n_2			

Учебное издание

ВАЛУЕВ Владимир Егорович
ВОЛЧЕК Александр Александрович
ПОЙТА Петр Степанович
ШВЕДОВСКИЙ Петр Владимирович

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПРИРОДОПОЛЬЗОВАНИИ

Редактор Т.В. Строкач
Ответственный за выпуск В.Е. Валуев
Художник (компьютерная графика) А.А. Волчек
Набор и компьютерная верстка Т.Е. Мозоль

Лицензия № 382 от 30.04.1999 г.

Сдано в набор 3.03.99. Подписано в печать 7.06.99. Формат 60×84 1/16. Бумага Uni Paper. Гарнитура Times New Cyr. Усл. печ. л. 14,9. Уч. изд. л. 16. Заказ № 490. Тираж 200 экз.

Отпечатано на ризографе
Брестского политехнического института.
224017, г. Брест, ул. Московская, 267.

ISBN 985-6584-02-7

9 789856 584025