ГЛАВНОЕ УПРАВЛЕНИЕ ГИДРОМЕТЕОРОЛОГИЧЕСКОЙ СЛУЖБЫ
ПРИ СОВЕТЕ МИНИСТРОВ СССР

ЦЕНТРАЛЬНЫЙ ИНСТИТУТ ПРОГНОЗОВ

Е. С. УЛАНОВА

Применение математической статистики в агрометеорологии для нахождения уравнений связей

ГИДРОМЕТЕОРОЛОГИЧЕСКОЕ ИЗДАТЕЛЬСТВО (ОТДЕЛЕНИЕ)

Москва — 1964

УДК 630 : 551.509.314

АННОТАЦИЯ

В книге в доступной форме излагаются основы математической статистики применительно к агрометеорологии. Рассмотрено применение статистических методов для нахождения уравнений прямолинейных и криволинейных связей между переменными величинами.

Даны из области агрометеорологии практические примеры расчетов уравнений прямолинейных связей двух, трех и четырех переменных величин, а также уравнений криволинейных связей.

Примеры расчетов уравнений рассматривают вопросы зависимости урожаев сельскохозяйственных культур от запасов продуктивной влаги в почве, продолжительности межфазных периодов сельскохозяйственных культур от термического фактора, изменения запасов продуктивной влаги в почве от метеорологических условий и другие вопросы.

Книга является методическим пособием для специалистов агрометеорологов. Она может быть использована работниками сельского хозяйства, а также студентами гидрометеорологических и сельскохозяйственных вузов.

ВВЕДЕНИЕ

В последние годы в развитии агрометеорологии значительно возросла роль математических наук. Особенно возросло применение математической статистики, главным образом — применение теории корреляции, где рассматриваются методы статистического изучения связей между явлениями.

Методы математической статистики и в частности теории корреляции находят все большее применение во многих науках при решении различных инженерно-технических, экономических, сельскохозяйственных и других вопросов, особенно в связи с развитием новой техники и автоматизации производства, когда приходится иметь дело с массовыми явлениями, с большим количеством наблюдений.

Как известно, основное применение, которое находит теория корреляции, относится к решению задачи обоснованного прогноза явления, т. е. к указанию пределов, в которых с наперед заданной точностью будет содержаться интересующая нас величина, если другие, связанные с ней величины, получают определенные значения.

При разработке методов агрометеорологических прогнозов и оценки условий формирования урожая различных сельскохозяйственных культур агрометеорологи для получения прогностических уравнений зависимости одних факторов от других проводят статистический анализ и обработку большого числа сопряженных наблюдений агро- и гидрометеорологических станций Советского Союза.

В настоящее время по математической статистике и, в частности, по теории корреляции накоплена большая советская и зарубежная литература. Однако большинство опубликованных книг, как монографий, так и учебных пособий, требуют специальной математической подготовки, в других же книгах изложение материала дано с учетом их специфического применения к определенной отрасли знаний. Кроме того, в большинстве изданий главное место занимает теория вопроса со сложными математическими выводами, но приводится мало примеров, в доступной форме излагающих практическое использование теории корреляции. В применении к агрометеорологии такие книги вообще отсутствовали.

В то же время было совершенно очевидно, что агрометео-

рологи особенно нуждались в методическом пособии, в котором в доступной форме излагалось бы применение статистических методов в агрометеорологии для нахождения уравнений корреляционных прямолинейных и криволинейных связей.

В предлагаемой книге, написанной автором по заданию Главного управления Гидрометеослужбы при Совете Министров СССР, излагаются основы математической статистики в области теории корреляции и применение статистических методов в агрометеорологии для нахождения уравнений линейной связи двух, трех и четырех переменных величин, а также уравнений криволинейных связей.

Автор стремился сделать книгу максимально доступной для агрометеорологов. Этим назначением книги определяется ее построение и характер изложения, где нет труднодоступной математической теории, а есть конкретные выводы из этой теории и их применение на практике на конкретных примерах, с которыми агрометеорологи встречаются в своей работе. Примеры взяты в основном из работ автора и относятся к различным агрометеорологическим вопросам: определение уравнений зависимости урожая озимой пшеницы от весенних запасов влаги в почве, зависимость межфазных периодов от температуры, зависимость изменения запасов влаги в почве от метеорологических факторов и др. Естественно, что эти примеры не претендуют на всю полноту охвата агрометеорологических вопросов, но они могут быть типичными для исследования многих агрометеорологических вопросов.

Предлагаемая книга может служить методическим пособием для специалистов агрометеорологов в их исследованиях по нахождению статистических связей между различными явлениями и определению уравнений этих связей. Она может быть использована работниками сельского хозяйства, а также студентами гидрометеорологических и сельскохозяйственных вузов.

ГЛАВА I РАЗЛИЧНЫЕ ТИПЫ ЗАВИСИМОСТЕЙ МЕЖДУ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ

§ 1. ФУНКЦИОНАЛЬНЫЕ И СТАТИСТИЧЕСКИЕ СВЯЗИ. АРГУМЕНТ И ФУНКЦИЯ. ЗАДАЧИ ТЕОРИИ КОРРЕЛЯЦИИ

Явления повседневной жизни неразрывно связаны с числами и измерениями. При наличии большого числа наблюдений и измерений, всегда появляется необходимость свести первоначальную массу данных к небольшому числу показателей. Всякий исследователь, имеющий дело с обширными наблюдениями, хочет привести эти наблюдения к определенной системе и форме, так как никакой человеческий ум не способен вместить в себя все содержание большого количества разрозненных числовых данных. Поэтому стали стремиться в относительно небольшом числе сводных показателей отразить наиболее важную и существенную закономерность, содержащуюся в данной массе наблюдений и измерений.

Для этого потребовалось создание особого рода научного математического метода, который был назван статистическим методом или математической статистикой.

Как наука, математическая статистика является одним из разделов математики и ее можно рассматривать как математику, применяемую при обработке результатов массовых наблюдений.

В математической статистике, как и во всей математике, одна и та же формула может одинаково относиться к самым различным объектам. Поэтому статистические методы применяются в самых различных областях знаний.

Однако необходимо иметь в виду, что научная статистика должна базироваться на предварительном качественном анализе и не может использоваться в отрыве от реальной природы объекта исследования.

К. Маркс и В. И. Ленин широко пользовались в своих работах статистическими методами, но предупреждали, что правильное применение статистического анализа к определенным явлениям нельзя сводить только к одним математическим приемам и расчетам. Для применения статистики прежде всего необходим предварительный качественный анализ особенностей изучаемых явлений, знания их общих закономерностей.

Одним из центральных разделов математической статистики является теория корреляции, которая изучает взаимосвязь, взаимозависимость между исследуемыми величинами. Латинское слово Corelatio означает соотношение, взаимосвязь.

Изучением зависимостей между различными явлениями занимается любая наука, так как каждое явление природы и общества не возникает само по себе, а находится в связи с другими явлениями.

Диалектический подход к изучению природы и общества требует рассмотрения явлений в их взаимосвязи и непрерывном изменении. Теория корреляции позволяет выразить эти взаимосвязи в количественной форме.

Наиболее простым видом связи между величинами является функциональная зависимость, когда каждому значению одной величины, соответствует вполне определенное значение другой величины.

К функциональным связям относится зависимость между силой тока I, напряжением E и сопротивлением R.

$$I=\frac{E}{R}.$$

Связь давления газа p, его температуры T и его объема V также является функциональной: $p=R\frac{T}{V}$ (R — постоянный коэффициент).

В качестве вида функциональной связи можно привести еще ряд примеров, это будет связь между радиусом окружности R и ее длиной C, где $C=2\pi R$. Эта формула позволяет по любому известному значению радиуса найти соответствующее ему вполне определенное значение длины окружности.

Функциональные связи между переменными величинами изучаются в специальном разделе математики — математическом анализе. Они характерны для количественных соотношений в области астрономии, механики, физики.

В природе же чаще всего наблюдаются нефункциональные связи, когда переменная величина y изменяется главным образом в зависимости от другой переменной x, но на изменение y влияют также множество дополнительных других факторов, учесть которые исследователь часто не в состоянии, и тогда каждому значению x соответствует несколько значений y.

Такие связи (зависимости), когда численному значению одной величины x соответствует не одно, а несколько значений другой величины y, т. е. целая статистическая совокупность значений y, группирующихся лишь около некоторой средней величины $\bar{y}_x$, называются статистическими или корреляционными. Таким образом, считают, что y на-

ходится в корреляционной зависимости от x, если:

1) каждому значению аргумента x соответствует ряд распределения функции y;

2) с изменением x эти ряды y закономерно изменяют свое положение.

Часто в литературе также встречается и такое определение статистических — корреляционных связей: связь между переменными величинами x и y называется статистической или корреляционной, если различным значениям одной из них (x) соответствуют определенные групповые средние другой $(\overline{y}_x)$ или наоборот.

В таких связях чаще всего одна величина рассматривается как независимая переменная, которая называется аргументом и обозначается буквой x. Другая величина является зависимой переменной, которая называется функцией, и обозначается буквой y. Например, если мы ищем связь урожая сельскохозяйственных культур с осадками, то ясно, что в данном случае независимой переменной — аргументом — будут осадки (x), а урожай будет зависимой переменной величиной от осадков, т. е. функцией (y), а не наоборот. Однако не всегда при нахождении связей так ясно бывает с определением независимой и зависимой переменной, т. е. аргумента и функции. Часто мы ищем связи между явлениями, взаимно влияющими друг на друга, взаимно зависимыми друг от друга. В данном случае мы условно одну величину принимаем за аргумент x, а другую — искомую величину — за функцию y. Например, при нахождении связей запасов влаги в различных слоях почвы 0—20 *см* и 20—50 *см* мы можем условно величину запасов влаги верхнего слоя 0—20 *см* принять за аргумент x, а величину запасов влаги слоя 20—50 *см* принять за функцию y и найти уравнение связи в отношении y (см. гл. II, § 8).

При анализе корреляционных статистических связей различных переменных величин главной задачей исследования должно было быть выяснение на основании большого числа наблюдений того, как изменяется функция (y) в связи с изменением главного своего аргумента (x), если бы ряд других ее аргументов не изменялся.

Однако в природе такого положения быть не может. Эти другие аргументы также изменяются и своей изменчивостью затушевывают и искажают интересующие нас зависимости. При этом влияние дополнительных факторов может проявиться с большей или меньшей силой. Определяя полученную зависимость одного элемента от другого главного, влияющего, мы всегда должны знать, хотя бы в общем, величину влияния других дополнительных изменяющихся, но неучтенных факторов. Если эта величина мала, то мы, зная главный аргумент

(x), можем достаточно точно определить значение функции. Если действие дополнительных факторов велико, то связь y и x получается слабой и мы с изменением x не можем достаточно точно определять изменения значения функции y.

Диалектический материализм учит, что изучение зависимости явлений состоит не в том, чтобы наблюдать за бесконечным множеством причин каждого отдельного случая, а изучать необходимо главные, решающие причины, которые определяют результат. Вся масса мелких второстепенных причин не может быть учтена исследователем, иначе он утонет в море деталей, не имеющих сколько-нибудь существенного значения.

Задача исследований взаимосвязи между явлениями состоит в выявлении главных причин изменения этих явлений.

Таким образом, первая задача теории корреляции заключается в выявлении на основе большого количества наблюдений того, как изменяется в среднем функция в связи с изменением одного или нескольких главных ее аргументов. Это изменение предполагает условие постоянства ряда других дополнительных неучтенных факторов, хотя они изменяются и искажающее их влияние на изменение функции очевидно.

Вторая задача заключается в определении степени влияния главных учитываемых и искажающих неучтенных факторов.

Первая задача решается путем определения формы связи и нахождения уравнения этой связи двух или нескольких переменных величин.

Вторая задача решается при помощи различных показателей тесноты связи, которые дают оценку степени рассеяния значений y для разных значений x.

§ 2. ОСНОВНЫЕ ВИДЫ ЛИНЕЙНЫХ И НЕЛИНЕЙНЫХ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ И ИХ УРАВНЕНИЯ

Общий вид уравнения корреляционной связи $y=f(x)$.

В наиболее точном выражении $\overline{y}_x=f(x_i)$ или $\overline{y}_i=f(x_i)$, где $f(x_i)$ представляет собой определенную однозначную функцию, дающую возможность по x_i находить приближенно соответствующие средние величины $\overline{y}_i$, где $i=1, 2, 3, \ldots n$ — знак порядкового номера наблюдений, а n — общее число наблюдений. Это уравнение называется уравнением регрессии y по x, где x — аргумент, а y — функция.

Можно также уравнение находить в отношении x, когда $\overline{x}_i$ представляет собой условную среднюю значений x, вычисляемую по условному распределению x, соответствующему y_i. Тогда $\overline{x}_i=f(y_i)$ является корреляционной связью $\overline{x}_i$ с y_i и

называется уравнением регрессии x по y. В данном случае $\bar{x}_i$ приобретает значения функции, а y_i — аргумента.

Таким образом, для каждой статистической зависимости величин можно рассматривать два вида связи y по x и x по y, которые являются различными.

По этим связям находят приближенные формулы, выражающие зависимость между значениями x_i и средними значениями $\bar{y}_i$ или наоборот. Такие формулы, полученные на основании статистического анализа экспериментальных данных, называются эмпирическими.

Существуют различные методы нахождения эмпирических формул, но при пользовании этими формулами результаты получаются в разной степени приближенного характера. Оценка точности статистических — корреляционных зависимостей и полученных эмпирических формул производится на основании корреляционного анализа.

Как уже отмечалось, первой задачей теории корреляции и корреляционного анализа является вопрос формы связи переменных величин, состоящий в определении вида функции $\bar{y}_i = f(x_i)$.

Из встречающихся форм корреляционных связей наиболее распространены линейные корреляционные связи. Они также являются и наиболее изученными. Однако часто наблюдаются и нелинейные связи между элементами. Не всегда задача выбора формы связи бывает легкой. При графическом изображении статистической связи часто точки располагаются так, что можно провести ряд линий различных типов. Например, в большей части графика могут совпадать прямая линия и гипербола или парабола. Поэтому при выборе формы связи (типа линии связи) необходимо прежде всего принимать во внимание характерные особенности линии связи, вытекающие непосредственно из самой физической сущности изучаемого явления, из знания общих закономерностей данных связей. Таким образом, выбору вида линии должен предшествовать логический анализ, обусловленный знанием общих закономерностей исследуемых явлений.

Для выбора формы статистической связи нужно хорошо знать простейшие линии и их уравнения.

Обычно в уравнениях переменные величины, связь между которыми мы ищем, обозначаются последними буквами латинского алфавита — x, y, z, u, v, а постоянные коэффициенты при этих переменных (параметры уравнения) обозначаются первыми буквами — a, b, c, d и т. д.

При определении статистических связей различных агрометеорологических элементов чаще всего могут встретиться следующие типы линий и их уравнения (рис. 1):

1) Прямая, проходящая через начало координат. Уравне-

ние этой прямой $y=ax$. Имеем зависимость прямой пропорциональности между y и x, в которой необходимо определить

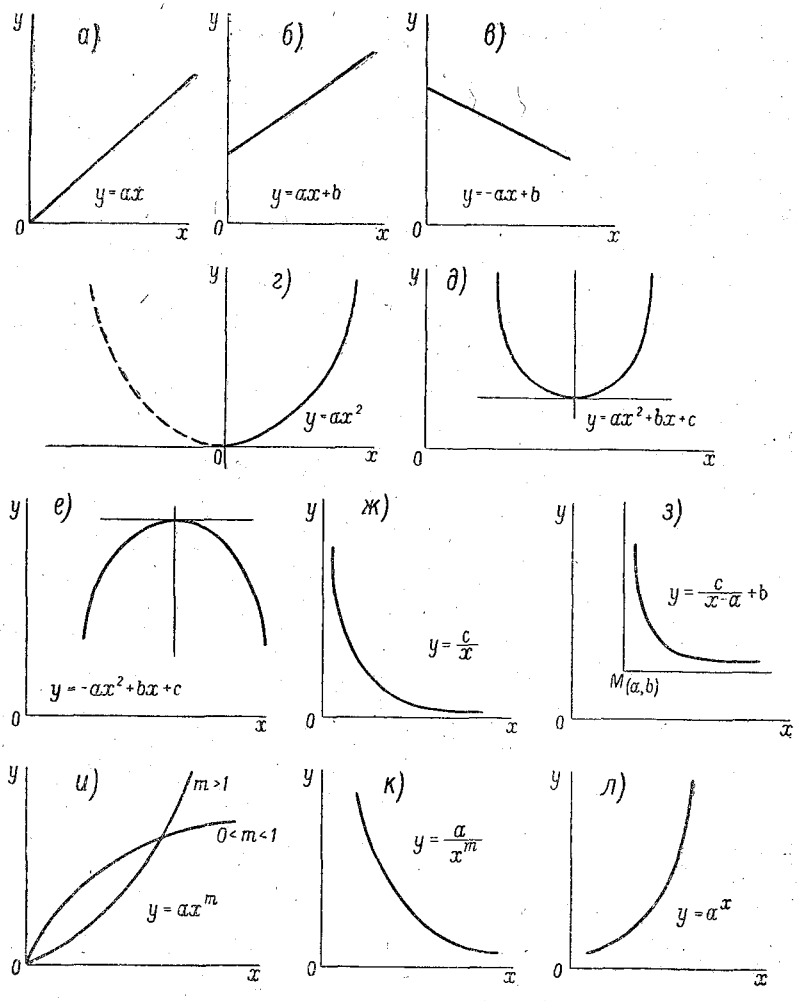

Рис. 1. Виды основных линий различных связей между переменными величинами и их уравнения.

один параметр a. Линии этого типа нужно выбирать в тех случаях, когда по смыслу при $x=0$ и $y=0$ (рис. 1 $а$).

2) Прямая, не проходящая через начало координат. Уравнение ее имеет вид $y=ax+b$ (рис. 1$б$) и $y=-ax+b$ (рис. 1 $в$). Имеем линейную прямую (рис. 1 $б$) или обратную

(рис. 1 *в*) зависимость y от x с необходимостью определения двух параметров a и b.

3) Парабола с вершиной в начале координат и симметричная одной из осей координат (рис. 1 *г*). Уравнение для $y=ax^2$. Такими параболами изображаются зависимости, где одна из величин x или y пропорциональна квадрату другой величины. Формула содержит один параметр a. По мере увеличения абсолютной величины параметра a уменьшается «раствор» параболы.

4) Парабола, симметричная прямой, параллельной оси y. Уравнение имеет вид $y=ax^2+bx+c$. Функция квадратичная. Направление выпуклости параболы зависит от знака коэффициента a. При положительном a $(a>0)$ выпуклость параболы направлена вниз, при отрицательном a $(a<0)$ — вверх. Линии этого типа выбираются при наличии одного максимума или одного минимума и кривые симметричны относительно их (рис. 1 *д* и *е*). В формуле необходимо определить три параметра a, b и c.

5) Гипербола, асимптотически приближающаяся к осям координат. Уравнение имеет вид $y=\frac{c}{x}$. Имеем зависимость обратной пропорциональности между y и x, где необходимо определение одного параметра c (рис. 1 *ж*).

6) Гипербола, асимптотически приближающаяся к прямым, параллельным осям координат. Уравнение $y=\frac{c}{x-a}+b$. Формула содержит три параметра. Параметры a и b являются координатами точки $м$. Знак параметра c зависит от расположения гиперболы в отношении асимптот (рис. 1 *з*).

7) Степенные кривые (рис. 1 *и* и *к*). Уравнение $y=ax^m$, где m может быть положительным, целым или дробным. Частным случаем степенных функций являются параболы (рис. 1 *г*) и гиперболы (рис. 1 *ж* и *з*).

8) Показательная кривая, когда с возрастанием одной величины (x) наблюдается усиленное возрастание другой величины (y). Уравнение $y=a^x$ (рис. 1 *л*).

После того, как установлена форма связи, выбран тип линии и вид общего уравнения связи, приступают к вычислению параметров уравнения данной связи и определению ее тесноты.

ГЛАВА II ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ ДВУХ ПЕРЕМЕННЫХ ВЕЛИЧИН

§ 1. КОРРЕЛЯЦИОННОЕ ПОЛЕ. КОРРЕЛЯЦИОННАЯ ТАБЛИЦА. ЭМПИРИЧЕСКИЕ ЛИНИИ РЕГРЕССИИ

При исследовании взаимосвязей различных явлений часто бывает необходимость установления зависимости между двумя переменными величинами. Наиболее распространены линейные связи между двумя величинами, которые хорошо изучены математической статистикой.

Корреляционная зависимость между случайными переменными величинами x и y называется линейной корреляцией, если обе функции регрессии $y=f(x)$ и $x=f(y)$ являются линейными. В этом случае при графическом изображении обе линии регрессии являются прямыми, они называются прямыми регрессии и выражаются следующими уравнениями регрессии:

Линейное корреляционное уравнение регрессии y по x

$$y = ax + b, \tag{1}$$

линейное корреляционное уравнение регрессии x по y

$$x = a_1 y + b_1. \tag{2}$$

Коэффициент a уравнения (1) прямой регрессии y по x называется коэффициентом регрессии y по x и и обозначается $\rho_{y/x}$.

Аналогично коэффициент a_1 уравнения (2) прямой регрессии x по y называется коэффициентом регрессии x по y и обозначается $\rho_{x/y}$.

Коэффициент регрессии a является угловым коэффициентом k прямой регрессии y по x:

$$a = k = \rho_{y/x}.$$

Коэффициент регрессии a_1 не является угловым коэффициентом соответствующей прямой регрессии x по y. Угловым

коэффициентом $k_{x/y}$ является величина, обратная коэффициенту регрессии:

$$k_{x/y} = \frac{1}{\rho_{x/y}}.$$

Расчету корреляционных уравнений, нахождению коэффициентов регрессии и показателей тесноты связи обычно предшествует первичный анализ, систематизация имеющегося материала наблюдений и его статистическая обработка.

Массовый материал данных наблюдений двух переменных величин, зависимость между которыми мы хотим определить, необходимо сначала проанализировать с точки зрения соответствия данных общим закономерностям изменения того или иного явления и его взаимосвязи с другими явлениями. После анализа и отбраковки ошибочных данных материал наблюдений представляется в виде простой таблицы-сводки, где указаны соответствующие друг другу значения x_i и y_i. Причем x_i обозначает независимое переменное (аргумент), а y_i — зависимое переменное (функцию), где i — значок любого порядкового номера x или y от 1 до n, где n — общее число пар наблюдений x и y:

№ п/п	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
.	.	.
.	.	.
.	.	.
n	x_n	y_n
n	$\overline{x_{ср}}$	$\overline{y_{ср}}$

Для выяснения линейности связи необходимо прежде всего построение графического изображения этой связи. Для этого данные каждой пары значений x и y в виде точки, должны быть нанесены на график в прямоугольной системе координат.

Графическое изображение связи, кроме установления формы связи, позволит увидеть также и тесноту связи.

Для разбора основных элементов теории корреляции приведем полученную нами связь запасов продуктивной влаги различных слоев почвы осенью под озимой пшеницей.

Эта связь была получена по данным фактических наблюдений гидрометеорологических станций под запасами влаги осенью на полях озимой пшеницы в южных районах Украины.

Как известно, в начальный период развития и роста озимой пшеницы осенью для оценки и прогноза ее влагообеспеченности агрометеорологам важно знать запасы продуктивной влаги верхних слоев почвы до полуметра, так как в южных районах при продолжительной осени корневая система озимой пшеницы к моменту прекращения вегетации осенью может достигать 30—40 *см*. Поэтому в южных засушливых районах очень важно знать распределение запасов влаги по слоям. В приведенной зависимости (рис. 2) анализировались данные наблюдений запасов продуктивной влаги (в *мм*) верхнего 0—20-сантиметрового слоя почвы и следующего слоя почвы от 20 до 50 *см* глубины.

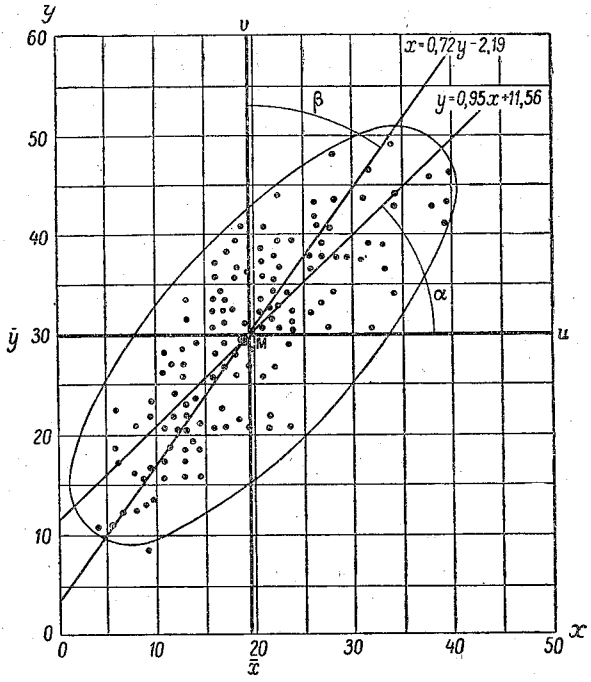

Рис. 2. Связь запасов продуктивной влаги различных слоев почвы осенью под озимой пшеницей; x — запасы влаги слоя 0—20 *см*, y — запасы влаги слоя 20—50 *см*.

За независимое переменное x в данном случае условно взяты величины запасов влаги слоя 0—20 *см*, а за зависимое переменное y — величины запасов влаги слоя 20—50 *см*. Это

сделано для того, чтобы уравнение было найдено в отношении y и давало бы нам расчет величин запасов влаги слоя 20—50 *см* в зависимости от влажности почвы верхнего 0—20-сантиметрового слоя. Данные о влажности слоя почвы 0—20 *см* можно иметь как фактические, так и прогнозируемые, рассчитанные по другим методам, в то время как для расчетов влаги слоя 20—50 *мм* методов не было.

Таким образом, при анализе материала наблюдений в первую очередь важно определить, в отношении какого элемента мы будем искать уравнение. Тот элемент, который известен при расчетах и прогнозах, мы должны считать аргументом x, а неизвестный элемент, который нужно рассчитать по искомому уравнению, мы должны обозначить функцией y и искать уравнение в отношении y.

Получив 135 случаев одновременных наблюдений за запасами влаги слоя 0—20 *см* и 20—50 *см*, записываем их сначала в виде простой таблицы-сводки, где под одним порядковым номером даются величины x и y. Затем, для выяснения линейности связи, строим график, откладывая в прямоугольной системе координат по оси x значения запасов влаги слоя почвы 0—20 *см*, а по оси y — слоя 20—50 *см*. Таким образом, мы получаем для каждого порядкового номера нашей сводной таблицы на плоскости точки с координатами x_1y_1, x_2y_2, ..., $x_{135}y_{135}$. Мы получили поле точек, которое называется полем корреляции или корреляционным полем yx (рис. 2).

Если число случаев пар наблюдений велико (больше 100), то корреляционное поле имеет вид более или менее правильного эллипса со сгущением точек в центре и сравнительно редким их расположением на периферии. Отклонение осей эллипса от координатных направлений указывает на наличие корреляции. Вытянутость же эллипса не всегда является объективным показателем, ибо она зависит от принятых масштабов по осям координат. По корреляционному полю мы качественно уже можем судить о форме связи и ее тесноте.

На основании полученного графического поля корреляции при большом числе наблюдений продолжают дальнейшую систематизацию данных, путем их группировки и построения корреляционной таблицы или корреляционной решетки.

Корреляционная таблица строится по интервалам значений x и y, выбранным самим исследователем.

Для этого на графике, где изображено поле корреляции (рис. 2), строят координатную сетку через точки, которые определяют границы выбранных интервалов значений для x и для y. Таким образом, все поле разбивается на вертикальные и горизонтальные столбцы, которые называются строями.

Вследствие пересечения строев плоскость корреляционно-

го поля разобьется на прямоугольники или клетки. Подсчитав число точек в каждом прямоугольнике, который соответствует определенным значениям интервалов x и y, и записав эти данные в виде таблицы, мы получим корреляционную таблицу или корреляционную решетку, которая нам облегчит ряд действий по нахождению линии регрессий и их уравнений.

Корреляционную таблицу или решетку можно строить непосредственно по первичной таблице-сводке, не прибегая к графическому построению корреляционного поля. В этом случае устанавливаются нужные нам интервалы для значений x и y и делается выборка данных по этим интервалам. Получим число частот (m_{xy}) сочетания значений x и y определенных интервалов.

Общий вид корреляционной таблицы или решетки представлен табл. 1.

Таблица 1

Общий вид корреляционной таблицы или решетки

x / y	Интервал Δx	Δx_1	Δx_2	. . .	Δx_s	m_y	$\bar{x}_{y_i}$
Интервал Δy	Середина интервалов	x_1	x_2	. . .	x_s		
Δy_1	y_1	m_{11}	m_{21}	. . .	m_{s1}	m_{y_1}	$\bar{x}_{y_1}$
Δy_2	y_2	m_{12}	m_{22}	. . .	m_{s2}	m_{y_2}	$\bar{x}_{y_2}$
. . .	. . .						
. . .	. . .						
Δy_k	y_k	m_{1k}	m_{2k}	. .	m_{sk}	m_{y_k}	$\bar{x}_{y_k}$
m_{x_i}		m_{x_1}	m_{x_2}	. . .	m_{x_s}	$n=s\,k$	$\bar{x}$
$\bar{y}_{x_i}$		$\bar{y}_{x_1}$	$\bar{y}_{x_2}$	. . .	$\bar{y}_{x_s}$	$\bar{y}$	

На пересечении каждого вертикального столбца и горизонтальной строки корреляционной таблицы дана частота m_{xy}, показывающая, сколько раз при данном значении x встречались указанные значения y или наоборот.

В предпоследнюю строку и предпоследний столбец вписывают суммы частот по столбцам и строкам $\sum m_x = \sum^{y} m_{xy}$; $\sum m_y = \sum^{x} m_{xy}$. $\sum m_x = \sum m_y = n$, где n — общее

число наблюдений. Значки x и y над знаками сумм обозначают суммирование вдоль столбца или вдоль строки, т. е. $\sum^{y}$ означает суммирование частот y по интервалам при неизменном x, а $\sum^{x}$ — суммирование частот x по интервалам при неизменном y.

В последнюю строку и в последний столбец вписывают условные средние взвешенные по частотам $\bar{y}_x$ и $\bar{x}_y$ значений y и x по столбцам и строкам

$$\bar{y}_x = \frac{\sum^{y} m_{xy}\, y}{m_x}; \quad \bar{x}_y = \frac{\sum^{x} m_{xy}\, x}{m_y}. \tag{3}$$

Суммы величин, стоящих в предпоследней строке и предпоследнем столбце должны быть равны общему числу наблюдений n:

$$\sum^{x} m_x = n; \qquad \sum^{y} m_y = n.$$

В предпоследние клетки последней строки и последнего столбца вписывают подсчитанные общие средние взвешенные значения всех y и x, т. е. $\bar{y}$ и $\bar{x}$. Они могут быть вычислены, как средние всех y или x, взвешенные по частотам m_{xy}, или как средние из $\bar{y}_x$ и $\bar{x}_y$, взвешенные по частотам m_x или m_y;

$$\bar{y} = \frac{\sum^{xy} m_{xy}\, y}{n} \text{ или } \bar{y} = \frac{\sum^{x} m_x \bar{y}_x}{n}. \tag{4}$$

$$\bar{x} = \frac{\sum^{xy} m_{xy}\, x}{n} \text{ или } \bar{x} = \frac{\sum^{y} m_y \bar{x}_y}{n}. \tag{5}$$

По корреляционной таблице мы так же, как и по полю корреляции, можем судить о форме связи и ее тесноте, так как мы по существу получаем корреляционную зависимость y от x в виде таблицы значений $\bar{y}_x$ для каждого значения x с указанием частот, а также корреляционную зависимость x от y в виде таблицы значений $\bar{x}_y$ для каждого значения y с указанием частот. Если частоты расположены по диагонали вниз направо, то связь между величинами прямая, т. е. при

увеличении x увеличивается и y. Если же частоты расположены по диагонали вверх направо, то связь обратная, т. е. с увеличением x уменьшается y.

По корреляционной таблице легко можно построить графическое поле корреляции, накладывая на координатную плоскость сетку применительно к интервалам таблицы и изображая частоту каждой клетки в виде соответствующего числа точек, равномерно распределенных внутри клетки. Подобное поле корреляции, составленное на основании корреляционной таблицы по частотам интервалов называется вторичным корреляционным полем (рис. 3).

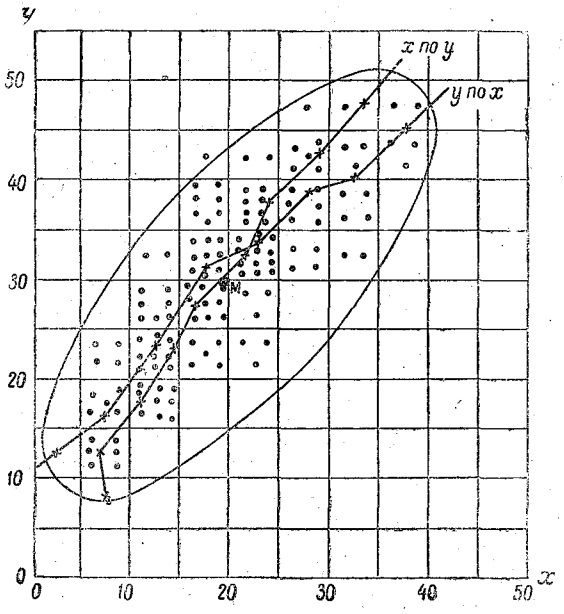

Рис. 3. Связь запасов влаги различных слоев почвы осенью под озимой пшеницей (вторичное корреляционное поле); x — запасы влаги слоя 0—20 *см*, y — запасы влаги слоя 20—50 *см*.

Корреляционная таблица облегчает построение эмпирической и теоретической линий регрессии и расчет уравнений регрессии методом сгруппированных данных, о чем будет изложено ниже (§ 8). Однако следует помнить, что корреляционная таблица строится обычно при большом числе пар наблюдений (больше 100).

При небольшом числе данных составление корреляционных таблиц по интервалам не рекомендуется. При обработке корреляционной таблицы считают, что число случаев в каж-

дой клетке относится к серединам интервалов, а это при малом числе наблюдений может дать заметные ошибки.

Кроме того, интервалы не должны быть большими, иначе мы будем получать искаженные данные и будем терять часть наблюдений, так как вместо фактических данных наблюдений при группировке мы берем условные, относящиеся к серединам интервалов. Доказано, что потеря информации, обусловленная группировкой, составляет менее 1%, если интервал группировки не превосходит четвертую часть среднего квадратического отклонения σ (§ 3). При надлежащем подборе интервала группировки ущерб в отношении точности не велик, но значительно сокращаются затраты труда на обработку данных.

В нашем примере, где мы ищем связь между величинами запасов влаги различных слоев почвы, число пар наблюдений x и y составляет 135 ($n=135$), поэтому мы можем построить корреляционную таблицу (табл. 2). Это легко сделать по построенному уже полю корреляции (рис. 2).

Берем по оси x и y интервалы значений в 5 *мм* и строим по этим интервалам сетку, разбивая все поле на клетки. Такие же интервалы берем для корреляционной таблицы, делая в ней такую же сетку, как и на графике, но с указанием значений середин интервалов по x и y. Затем подсчитываем число точек на графике в каждой клетке и вписываем это число в соответствующую клетку таблицы с теми же интервалами значений. Например, на графике в клетке при x от 10 до 15 *мм* и при y от 20 до 25 *мм* мы имеем 9 точек. В клетку таблицы для этих же интервалов x и y мы ставим число 9 и т. д. Получив частоты по столбцам и строкам, складываем их по вертикалям и горизонталям. Например, сумма частот y при $x=7{,}5$ будет равна 16, а при $x=12{,}5$ сумма частот y равна 27. ($\overset{y}{\Sigma} m_{7,5}=16$; $\overset{y}{\Sigma} m_{12,5}=27$ и т. д.). Затем подсчитываем суммы частот x при различных серединах интервалов y. Например $\overset{x}{\Sigma} m_{12,5}=7$; $\overset{x}{\Sigma} m_{17,5}=14$ и т. д.

Получив частные суммы частот x для середин различных интервалов y, суммируем их и получаем общую сумму частот:

$$\Sigma m_y = 135.$$

То же самое суммируя все частоты y по серединам интервалов x, получим

$$\Sigma m_x = 135.$$

После этого приступаем к расчету средних взвешенных величин $\overline{y}_x$ и $\overline{x}_y$ по столбцам и строкам.

В табл. 2 мы имеем восемь столбцов с различной суммой частот y.

Таблица 2

Пример составления корреляционной таблицы для связи запасов влаги различных слоев почвы 0—20 *см* и 20—50 *см*

y \ x	Интервал Δx	0—5	5—10	10—15	15—20	20—25	25—30	30—35	35—40	m_y	$\bar{x}_y$
Интервал Δy	Середина	2,5	7,5	12,5	17,5	22,5	27,5	32,5	37,5		
0—5	2,5										
5—10	7,5		1							1	7,5
10—15	12,5	1	6							7	6,8
15—20	17,5		5	9						14	10,7
20—25	22,5		4	9	5	3				21	14,1
25—30	27,5			7	9	3				19	16,5
30—35	32,5			2	9	12	4	2		29	21,6
35—40	37,5				7	8	6	4		25	23,9
40—45	42,5				1	2	5	3	3	14	29,3
45—50	47,5						1	2	2	5	33,5
m_x		1	16	27	31	28	16	11	5	135	$\bar{x}=19$
$\bar{y}_x$		12,5	16,2	22,9	30,9	33,0	38,4	39,8	44,5	$\bar{y}=30$	

Следовательно нам необходимо найти восемь условных средних значений y_x. Это достигается путем суммирования по вертикали произведений числа частот каждой клетки на соответствующее значение середины интервала y:

1) $\overline{y}_{2,5} = 12{,}5,$

2) $\overline{y}_{7,5} = \frac{7{,}5 + 6 \cdot 12.5 + 5 \cdot 17{,}5 + 4 \cdot 22{,}5}{16} = \frac{260{,}0}{16} = 16{,}2,$

3) $\overline{y}_{12,5} = \frac{9 \cdot 17{,}5 + 9 \cdot 22{,}5 + 7 \cdot 27{,}5 + 2 \cdot 32{,}5}{27} = \frac{617{,}5}{27} = 22{,}9,$

4) $\overline{y}_{17,5} = \frac{5 \cdot 22{,}5 + 9 \cdot 27{,}5 + 9 \cdot 32.5 + 7 \cdot 37{,}5 + 42{,}5}{31} = \frac{957{,}5}{31} = 30{,}9,$

5) $\overline{y}_{22,5} = \frac{3 \cdot 22{,}5 + 3 \cdot 27{,}5 + 12 \cdot 32{,}5 + 8 \cdot 37{,}5 + 2 \cdot 42{,}5}{28} = \frac{925{,}0}{28} = 33{,}0,$

6) $\overline{y}_{27,5} = \frac{4 \cdot 32{,}5 + 6 \cdot 37{,}5 + 5 \cdot 42{,}5 + 47{,}5}{16} = \frac{615{,}0}{16} = 38{,}4,$

7) $\overline{y}_{32,5} = \frac{2 \cdot 32{,}5 + 4 \cdot 37{,}5 + 3 \cdot 42{,}5 + 2 \cdot 47{,}5}{11} = \frac{437{,}5}{11} = 39{,}8,$

8) $\overline{y}_{37,5} = \frac{3 \cdot 42{,}5 + 2 \cdot 47{,}5}{5} = \frac{222{,}5}{5} = 44{,}5,$

По горизонтали в табл. 2 нам нужно рассчитать 10 условных средних значений $\overline{x}_y$, которые находим путем суммирования по горизонтали произведений числа частот каждой клетки на соответствующее значение середины интервала x.

Начинаем со второй строки, так как в первой строке значений x не было:

2) $\overline{x}_{7,5} = 7{,}5,$

3) $\overline{x}_{12,5} = \frac{2{,}5 + 67{,}5}{7} = \frac{47{,}5}{7} = 6{,}8,$

4) $\overline{x}_{17,5} = \frac{5 \cdot 7{,}5 + 9 \cdot 12{.}5}{14} = \frac{150{,}0}{14} = 10{,}7,$

5) $\overline{x}_{22,5} = \frac{4 \cdot 7{,}5 + 9 \cdot 12{,}5 + 5 \cdot 17{.}5 + 3 \cdot 22{,}5}{21} = \frac{297{,}5}{21} = 14{,}1,$

6) $\overline{x}_{27,5} = \frac{7 \cdot 12{.}5 + 9 \cdot 17{,}5 + 3 \cdot 22{,}5}{19} = \frac{312{,}5}{19} = 16{,}5,$

7) $\overline{x}_{32,5} = \frac{2 \cdot 12{,}5 + 9 \cdot 17{.}5 + 12 \cdot 22{.}5 + 4 \cdot 27{,}5 + 2 \cdot 32{,}5}{29} = \frac{627{,}5}{29} = 21{,}6,$

8) $\overline{x}_{37,5} = \frac{7 \cdot 17{,}5 + 8 \cdot 22{,}5 + 6 \cdot 27{,}5 + 4 \cdot 32{,}5}{25} = \frac{597{,}5}{25} = 23{,}9,$

9) $\bar{x}_{42,5} = \frac{17,5+2\cdot 22,5+5\cdot 27,5+3\cdot 32,5+3\cdot 37,5}{14} = \frac{410,0}{14} = 29,3,$

10) $\bar{x}_{47,5} = \frac{27,5+2\cdot 32,5+2\cdot 37,5}{5} = \frac{167,5}{5} = 33,5.$

Получив условные средние взвешенные значения $\bar{y}_x$ и $\bar{x}_y$ по строям, находим общее среднее взвешенное значение всех y:

$$\bar{y} = \frac{\sum^{x} m_x \bar{y}_x}{n} = \frac{12,5+16\cdot 16,2+27\cdot 22,9+\ldots+5\cdot 44.5}{135} = \frac{4046,5}{135} = 30.$$

Также находим общее среднее взвешенное значение всех x:

$$\bar{x} = \frac{\sum^{y} m_y \bar{x}_y}{n} = \frac{7,5+7\cdot 6,8+14\cdot 10,7+\ldots+5\cdot 33,5}{135} = \frac{2616,1}{135} = 19.$$

Записываем эти числа $\bar{y}$ и $\bar{x}$ в корреляционную таблицу. Рассчитав таким образом условные средние значения $\bar{y}_x$ и $\bar{x}_y$, а также общие средние $\bar{y}$ и $\bar{x}$, мы можем приступить к построению эмпирических линий регрессии y по x и x по y.

По данным полученной корреляционной таблицы построим вторичное поле корреляции, нанося в каждую клетку на графике рис. 3 число точек соответственно числу частот в таблице и распределяя их равномерно по клетке, ограниченной данными интервалами. После этого наносим на график (рис. 3) данные восьми условных средних значений $\bar{y}_x$ для середин интервалов x и общее среднее значение $\bar{y}$. Соединяя точки средних значений, получаем ломаную линию, которая называется эмпирической линией регрессии y по x.

Наносим условные средние значения $\bar{x}_y$ для середин интервалов y, а также общую среднюю величину $\bar{x}$. Соединяя эти значения всех средних получаем эмпирическую линию регрессии x по y. Точка M на графике с координатами $\bar{y}\bar{x}$, называется центром распределения (рис. 3).

Вторичное корреляционное поле строить не обязательно. Эмпирические линии регрессии можно построить и на первичном корреляционном поле, нанеся на него условные и общие средние величины $\bar{y}_x$, $\bar{x}_y$, $\bar{y}$, $\bar{x}$.

Если число случаев наблюдений мало, корреляционную таблицу не составляют, а эмпирические линии регрессии получают следующим образом. По таблице-сводке наблюдений строят график корреляционного поля. Оси X и Y разбивают на

нужные интервалы и получают строи по x и строи по y. Находят среднее значение $\bar{y}_x$ для каждого интервала оси X и общее среднее значение всех y $(\bar{y})$. Наносят значения $\bar{y}_x$ на график для середин интервалов x и соединяют данные этих средних значений ломаной линией, получая эмпирическую линию регрессии y по x. Определив по строям оси Y средние значения $\bar{x}_y$ и нанеся их на график для середин интервалов y, строят эмпирическую линию регрессии x по y.

Эмпирические линии регрессии получаются ломаными, поэтому проводят их сглаживание или выравнивание и получают плавные прямые линии регрессии, которые называются теоретическими линиями регрессии. Они для лучшей наглядности нанесены на рис. 2.

Теоретическую линию регрессии следует проводить после нахождения уравнения данной функции. Выравнивание по уравнению называется аналитическим выравниванием эмпирической линии регрессии.

Построенный аппарат теории корреляции двух переменных величин для определения тесноты связи и нахождения уравнений связи дает в качестве суммарных характеристик пять основных показателей:

а) средние арифметические значения каждой из величин $\bar{x}$ и $\bar{y}$;

б) средние квадратические отклонения каждой величины σ_x и σ_y;

в) коэффициент корреляции r.

§ 2. СРЕДНЯЯ АРИФМЕТИЧЕСКАЯ И ЕЕ СВОЙСТВА

Средняя арифметическая величина является простейшей и в то же время очень важной величиной, так как она является первой сводной статистической характеристикой.

Средней арифметической величиной называется сумма значений признака (элемента), разделенная на число этих значений:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots, x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (6)$$

Такая средняя арифметическая называется простой.

Если значениям x соответствуют различные частоты (m), то величина средней арифметической зависит не только от

значений x, но и от их частот (m):

$$\bar{x} = \frac{m_1 x_1 + m_2 x_2 + \ldots + m_k x_k}{m_1 + m_2 + \ldots + m_k} = \frac{\sum_{i=1}^{k} m_i x_i}{\sum_{i=1}^{k} m_i}. \qquad (7)$$

Такая средняя называется средней взвешенной величиной, а частоты m_1, m_2, ..., m_k — весами.

Значение признака (элемента), соответствующее каждой отдельной группе или интервалу, называется вариантом. Если x_1, x_2, x_3, ..., x_n — варианты, а m_1, m_2, ..., m_k — их частоты, то определение взвешенной средней арифметической может быть следующим. Взвешенная средняя арифметическая равна сумме произведений вариантов на их веса (или частоты), разделенной на сумму весов.

Средние взвешенные величины мы находили при определении средних в корреляционной табл. 2. Если число наблюдений очень большое, то расчеты средней арифметической получаются очень громоздкими, и тогда применяют упрощенные методы вычисления средней, учитывая ряд свойств средней арифметической величины.

1-е свойство: Если все значения x уменьшить на одно и то же число, то и средняя арифметическая уменьшится на это же число.

$$\overline{x - a} = \frac{\Sigma m_i x_i}{\Sigma m_i} - a = \bar{x} - a.$$

Число a может быть каким угодно, но лучше его брать из середины ряда значений x.

2-е свойство. Если все значения x разделить на одно и то же число, то средняя арифметическая тоже разделится на это же число.

$$\frac{x}{l} = \frac{1}{l} \frac{\Sigma m_i x_i}{\Sigma m_i} = \frac{\bar{x}}{l}.$$

3-е свойство. Средняя арифметическая суммы равна сумме средних арифметических, а средняя арифметическая разности равна разности средних арифметических:

$$\bar{x} = \frac{\Sigma (y + z)}{n} = \bar{y} + \bar{z}; \quad \bar{x} = \overline{y - z} = \bar{y} - \bar{z}.$$

4-е свойство. Сумма отклонений от средней арифметической равна нулю:

$$\Sigma (x - \bar{x}) = \overline{nx} - n\bar{x} = 0.$$

5-е свойство. Сумма квадратов отклонений от средней арифметической меньше, чем сумма квадратов отклонений от любого другого числа.

§ 3. ДИСПЕРСИЯ И СРЕДНЕЕ КВАДРАТИЧЕСКОЕ ОТКЛОНЕНИЕ. ИХ СВОЙСТВА

При различном числе наблюдений важно знать не только средние величины, но и отклонения отдельных значений от средней. Отклонением называется разность между отдельным значением x_i и средним значением $\bar{x}$. Отклонение положительно, когда x_i больше $\bar{x}$ и отрицательно, когда x_i меньше $\bar{x}$. Сумма всех положительных и отрицательных отклонений от средней арифметической, согласно 4-му свойству, будет равна нулю. Поэтому среднее отклонение нельзя использовать, как характеристику рассеяния. Для этого вводят другой показатель — дисперсию (D или σ^2), которая является средней арифметической квадратов отклонений, устраняющей влияние знаков на результат.

Дисперсия характеризует рассеяние значений переменной величины около средней арифметической $\bar{x}$.

$$D = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \quad \text{или} \quad D = \sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 m_i}{\sum_{i=1}^{k} m_i}. \qquad (8)$$

Таким образом, для вычисления дисперсии сначала все отклонения возводятся в квадрат, а потом вычисляется средняя арифметическая квадратов отклонений.

Очень важной статистической характеристикой является среднее квадратическое отклонение. Средним квадратическим отклонением называется абсолютное значение корня квадратного из дисперсии $\sigma = \sqrt{D}$.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}, \qquad (9)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 m_i}{\sum_{i=1}^{k} m_i}}. \qquad (10)$$

В математической статистике среднее квадратическое отклонение часто называют стандартным отклонением или просто стандартом.

Распределение величин по мере увеличения объема выборки (числа наблюдений n) приближается к нормальному распределению. Нормальное распределение, как известно, в графическом изображении симметрично, с наибольшей частотой в центре.

При нормальном распределении около 2/3 всех отклонений значений величины от ее среднего арифметического значения не превышают по абсолютной величине среднее квадратическое отклонение, т. е входят в эту величину. Среднее квадратическое отклонение σ_x выборочной средней x называют средней квадратической ошибкой или просто средней ошибкой (то же и для y).

Дисперсия и среднее квадратическое отклонение имеют следующие свойства.

1) Если все значения признака увеличить или уменьшить на одну и ту же величину a, то на ту же величину увеличится или уменьшится их средняя, отклонения же останутся без изменения. Следовательно останутся без изменения и среднее квадратическое отклонение и дисперсия.

2) Если все значения признака умножить на одно и то же число k, то в k раз увеличится и их средняя, следовательно в k раз увеличатся и отклонения. Квадраты же отклонений увеличатся в k^2 раз. Таким образом, дисперсия будет увеличена в k^2 раз, а среднее квадратическое отклонение окажется увеличенным в k раз.

3) Если все значения признака одинаковы, то они совпадают со своей средней и отклонения равны нулю. Вследствие этого и дисперсия и квадратическое отклонение равны нулю.

4) Средняя величина квадратов отклонений значений признака от любой величины a больше дисперсии на квадрат отклонения этой величины a от средней величины признака $\bar{x}$:

$$\frac{\Sigma (x-a)^2}{n} = \sigma^2 + (a-\bar{x})^2.$$

§ 4. КОЭФФИЦИЕНТ ЛИНЕЙНОЙ КОРРЕЛЯЦИИ ДВУХ ПЕРЕМЕННЫХ ВЕЛИЧИН

Кроме средних величин $\bar{x}$ и $\bar{y}$ и средних квадратических отклонений σ_x и σ_y, для нахождения уравнений регрессии и характеристики тесноты связи нам необходима еще одна величина, называемая коэффициентом линейной корреляции r, который является одним из наиболее совершенных методов измерения тесноты связи.

На рис. 2 было проведено две прямых линии регрессии y по x и x по y.

Направление этих прямых определяется коэффициентами регрессии. Первый из них — это тангенс угла, образованного прямой регрессии y по x с осью u, второй — тангенс угла между прямой регрессии x по y с осью v. Обозначим эти углы α и β. Следовательно коэффициенты регрессии — это $\operatorname{tg}\alpha$ и $\operatorname{tg}\beta$. В уравнениях $y=ax+b$ и $x=a_1y+b_1$ $a=\operatorname{tg}\alpha$, $a_1=\operatorname{tg}\beta$.

Коэффициенты регрессии могут быть оба положительными или оба отрицательными. В общем случае корреляционной связи эти две прямые регрессии не совпадают. Они совпадут, если зависимость между y и x будет функциональной, так как в этом случае не будет совокупности одних величин при определенном значении других величин, т. е. угол φ между прямыми будет равен нулю; не будет $\overline{y_x}$ и $\overline{x_y}$, а каждому значению одной величины будет соответствовать только одно значение другой.

С помощью угла φ между прямыми регрессии можно судить о тесноте связи между y и x. Чем больше угол φ между прямыми, тем слабее связь, и чем ближе угол φ к нулю, тем связь ближе к функциональной.

Для совпадающих прямых при $\varphi=0$ мы имеем $\alpha=90-\beta$ и поэтому $\operatorname{tg}\alpha=\operatorname{tg}(90-\beta)=\operatorname{ctg}\beta=\frac{1}{\operatorname{tg}\beta}$. Отсюда $\operatorname{tg}\alpha\operatorname{tg}\beta=1$.

Если связи между величинами нет, то y мало изменяется при изменении x и наоборот. В этом случае α и β близки к нулю и в пределе $\operatorname{tg}\alpha\operatorname{tg}\beta=0$.

Корень квадратный из числа $\operatorname{tg}\alpha\operatorname{tg}\beta$ принимают за критерий степени близости корреляционной связи к линейной функциональной зависимости и называют коэффициентом корреляции двух переменных величин x и y, обозначая его r:

$$r=\sqrt{\operatorname{tg}\alpha\operatorname{tg}\beta}. \qquad (11)$$

В уравнениях прямых регрессии $y=ax+b$ и $x=a_1y+b_1$ коэффициенты регрессии a и a_1 равны $a=\operatorname{tg}\alpha$, $a=\operatorname{tg}\beta$; коэффициент корреляции

$$r=\sqrt{\operatorname{tg}\alpha\operatorname{tg}\beta}=\sqrt{aa_1}. \qquad (12)$$

Таким образом, коэффициентом линейной корреляции называется средняя геометрическая величина из коэффициентов регрессий.

В полученных формулах и определении проявляется геометрический смысл коэффициента корреляции, состоящий в том, что он представляет собой среднюю геометрическую ко-

эффициентов регрессий прямых, образованных выравниванием каждого признака по другому признаку.

Если $r>0$, то обе прямые регрессии y по x и x по y проходят через центр распределения $M(\bar{x}, \bar{y})$ и образуют острые углы с положительными направлениями осей X и Y. В этом случае корреляция называется положительной, так как с возрастанием одной величины, возрастают соответственно условные средние другой. При $r=0$, в случае независимости величин x и y, прямая регрессии y по x параллельна оси X, а прямая регрессии x по y параллельна оси Y.

Следовательно угол между этими прямыми равен 90°. С возрастанием r наклон каждой из прямых к соответствующей оси координат возрастает, а острый угол между прямыми убывает. При $r=1$ обе прямые сливаются в одну, в этом случае зависимость будет функциональной. В случае $r<0$ при отрицательной корреляции, прямые регрессии проходят через точку $M(\bar{x}, \bar{y})$ и образуют тупые углы с положительным направлением осей координат. Угол φ между прямыми острый и всегда убывает по мере приближения r к -1. В том случае, когда $r=-1$, обе прямые сливаются в одну и мы имеем случай обратной линейной функциональной зависимости одной величины от другой.

Таким образом, коэффициент корреляции является безразмерной величиной и изменяется в пределах $-1 \leqslant r \leqslant 1$.

Кроме формулы 12 предложен еще ряд формул коэффициента корреляции r, выраженных через средние величины $\bar{x}$, $\bar{y}$ и средние квадратические отклонения σ_x, σ_y.

Из метода наименьших квадратов известно следующее выражение коэффициентов прямых регрессии:

$$a_{y/x}=\frac{\overline{xy}-\bar{x}\,\bar{y}}{\overline{x^2}-(\bar{x})^2} \quad \text{и} \quad a_{1x/y}=\frac{\overline{xy}-\bar{x}\,\bar{y}}{\overline{y^2}-(\bar{y})^2}. \tag{13}$$

Знаменатели обоих выражений обозначают соответствующие дисперсии:

$$\overline{x^2}-(\bar{x})^2=D(x)=\sigma_x^2, \quad \overline{y^2}-(\bar{y})^2=D(y)=\sigma_y^2,$$

откуда имеем простую формулу для r, выраженную через $\bar{x}$, $\bar{y}$, σ_x, σ_y:

$$r=\sqrt{aa_1}=\sqrt{\frac{\overline{(xy-\bar{x}\,\bar{y})^2}}{\sigma_x^2\sigma_y^2}}=\frac{\overline{xy}-\bar{x}\,\bar{y}}{\sigma_x\sigma_y}, \tag{14}$$

где r — линейный коэффициент корреляции между x и y; $\overline{xy}$ — среднее значение произведения x на y; $\overline{xy}=\frac{\Sigma xy}{n}$;

$\bar{x}$ и $\bar{y}$ — средние арифметические величины соответствующих признаков x и y;

$$\bar{x} = \frac{\Sigma x_i}{n}; \quad \bar{y} = \frac{\Sigma y_i}{n};$$

σ_x и σ_y — средние квадратические отклонения, найденные по признаку x и по признаку y;

$$\sigma_x = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}; \quad \sigma_y = \sqrt{\frac{\Sigma (y - \bar{y})^2}{n}}.$$

Для практического пользования в формулах делаем замену:

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}; \quad \sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}.$$

Тогда

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{[\overline{x^2} - (\bar{x})^2]\,[\overline{y^2} - (\bar{y})^2]}}. \tag{15}$$

Эти формулы сразу показывают, что между независимыми величинами корреляции не существует, так как для таких величин выполняется равенство $\overline{xy} = \bar{x}\,\bar{y}$.

Указанная выше формула r позволяет выразить каждый коэффициент регрессии через коэффициент корреляции.

В случае регрессии y по x

$$a_{y/x} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x} \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{\sigma_y}{\sigma_x} r, \tag{16}$$

в случае регрессии x по y

$$a_{1x/y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_y^2} = \frac{\sigma_x}{\sigma_y} \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_y \sigma_x} = \frac{\sigma_x}{\sigma_y} r. \tag{17}$$

Таким образом, нахождение уравнений регрессии будет облегчено, если мы найдем значения r, σ_x и σ_y.

Для этого необходимо найти

$$\overline{xy} - \bar{x}\,\bar{y} = \frac{\Sigma xy}{n} - \bar{x}\,\bar{y}, \quad \sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}$$

и

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{\frac{\Sigma y^2}{n} - (\bar{y})^2}.$$

Тогда

$$r = \frac{\frac{\Sigma xy}{n} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} \tag{18}$$

или

$$r = \frac{\frac{\Sigma xy}{n} - \bar{x}\,\bar{y}}{\sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}\ \sqrt{\frac{\Sigma y^2}{n} - (\bar{y})^2}}. \tag{19}$$

Для расчета r можно также применить следующую видоизмененную формулу (20), умножив числитель и знаменатель формулы (19) на n^2:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}. \tag{20}$$

Несмотря на свой громоздкий вид формула (20) удобна для расчетов, особенно при пользовании счетной машиной.

По приведенным формулам для r вычисляют коэффициент корреляции для несгруппированных данных. Для нахождения коэффициента корреляции по формулам (19) и (20) строят дополнительную таблицу, по которой удобно проводить расчеты (табл. 3).

Таблица 3

n_i	x	y	xy	x^2	y^2
1	2	3	4	5	6
$\Sigma n_i = n$	Σx	Σy	Σxy	Σx^2	Σy^2

Указанная табл. 3, как и последующие таблицы, дает подробное расположение величин, их произведений и квадратов для каждого порядкового номера. Приведенные таблицы удобны при ручном счете без помощи счетных машин. Одна-

ко необходимо стремиться к тому, чтобы расчеты статистических величин и нахождения параметров уравнений проводить с помощью вычислительной техники. Тогда не нужно будет заполнять графы 4, 5 и 6, а можно сразу рассчитывать произведение и суммировать величины.

При обработке с помощью вычислительной машины таблица для расчетов примет вид табл. 3*а*.

Таблица 3*а*

№ п/п	x	y
1	2	3
n	Σx	Σy
	Σx^2	Σy^2
	Σxy	

По такому же типу при расчете на счетной машине перестраиваются и остальные таблицы, когда можно сразу получить различные необходимые степени величин и их суммы, а также произведения величин и сумму произведений.

Когда x и y являются большими величинами, указанные формулы r приводят к громоздким расчетам, так как нужно брать произведение xy и их квадраты. Поэтому более удобно при несгруппированных данных использовать другую формулу r, где учитываются не сами величины, а их отклонения от средней, что позволяет проводить действия с меньшими числами, чем сами величины.

Проведем некоторые преобразования в формуле:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\frac{\Sigma (x - \bar{x})^2}{n} \cdot \frac{\Sigma (y - \bar{y})^2}{n}}} =$$

$$= \frac{\overline{xy} - \bar{x}\,\bar{y}}{\frac{1}{n} \sqrt{\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2}}. \qquad (21)$$

Возьмем выражение $\frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}$ и, раскрывая скобки под знаком суммы, преобразуем его. Тогда получим

$$\frac{\Sigma (xy - \bar{x}y - y\bar{x} + \bar{x}\bar{y})}{n} = \frac{\Sigma xy - \bar{y}\Sigma x - \bar{x}\Sigma y + \bar{x}\,\bar{y}\,n}{n} =$$

$$= \frac{\Sigma xy}{n} - \frac{\bar{y}\Sigma x}{n} - \frac{\bar{x}\Sigma y}{n} + \frac{\bar{x}\,\bar{y}\,n}{n} = \overline{xy} - \bar{y}\bar{x} - \bar{x}\,\bar{y} +$$

$$+ \bar{x}\,\bar{y} = \overline{xy} - \bar{x}\,\bar{y}.$$

Заменяя в формуле для r выражение $\overline{xy} - \bar{x}\,\bar{y}$ на $\frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}$, получим формулу r, где даны не сами величины, а их отклонения от средних:

$$r = \frac{\frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}}{\frac{1}{n} \sqrt{\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2}} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2}}.$$

Обозначая отклонения $x - \bar{x}$ через Δx, а $y - \bar{y}$ через Δy, получим

$$r = \frac{\Sigma \Delta x \Delta y}{\sqrt{\Sigma \Delta x^2 \Sigma \Delta y^2}}. \qquad (22)$$

Эта формула для практического использования несгруппированных данных является наиболее удобной. Для расчетов по этой формуле данные располагают в следующей табл. 4.

Таблица 4

n	x	y	$\Delta x = x - \bar{x}$	$\Delta y = y - \bar{y}$	$\Delta x\ \Delta y$	Δx^2	Δy^2
n	Σx	Σy			$\Sigma \Delta x\ \Delta y$	$\Sigma \Delta x^2$	$\Sigma \Delta y^2$

Для расчетов r можно применить также формулу Пирсона:

$$r = \frac{\Sigma \Delta x \Delta y}{n \sigma_x \sigma_y}. \qquad (23)$$

Выражение $\frac{\Sigma \Delta x \Delta y}{n} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}$ называется ковариацией или первым моментом произведений (μ_{11}).

Понятие момента в математической статистике занимает важное место. Центральным моментом статистической величины называется сумма произведений тех или иных степеней отклонений x от $\bar{x}$ или y от $\bar{y}$ на соответствующую частоту, деленная на сумму всех частот:

$$\mu = \frac{1}{n} \Sigma m_i (x_i - \bar{x})^h.$$

Второй центральный момент (μ_2) является дисперсией, а первый центральный момент (μ_1) — просто средним отклонением от средней арифметической величины.

Отсюда формула коэффициента корреляции, выраженная через моменты, будет иметь следующий вид:

$$r = \frac{\mu_{11}}{\sigma_x \sigma_y}. \qquad (24)$$

Если данные сгруппированы в корреляционную таблицу по частотам m_{xy}, то для расчетов применяются следующие формулы r и система расчетов:

$$r = \frac{\frac{1}{n}\Sigma m_{xy}\,(x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y}. \qquad (25)$$

В формуле (25) можно провести замену:

$$\frac{1}{n}\Sigma m_{xy}\,(x-\bar{x})\,(y-\bar{y}) = \frac{1}{n}\Sigma m_{xy}\,(xy-\bar{x}y-\bar{y}x+\bar{x}\,\bar{y}) =$$

$$= \frac{1}{n}[\Sigma m_{xy}\,xy - \bar{x}\Sigma m_{xy}\,(y-\bar{y})\Sigma m_{xy}\,x + \bar{x}\,\bar{y}\Sigma m_{xy}] =$$

$$= \frac{1}{n}[n\overline{xy} - n\bar{x}\,\bar{y} - n\bar{x}\,\bar{y} + n\bar{x}\,\bar{y}] = \overline{xy} - \bar{x}\,\bar{y}.$$

Получаем более удобную для расчетов формулу:

$$r = \frac{\frac{\Sigma m_x xy}{n} - \bar{x}\,\bar{y}}{\sqrt{\frac{\Sigma m_x x^2}{n} - (\bar{x})^2}\sqrt{\frac{\Sigma m_y y^2}{n} - (\bar{y})^2}}. \qquad (26)$$

Для сгруппированных данных при большом числе наблюдений расчеты упрощает введение новых условных значений x' и y':

$$x' = \frac{x - c_x}{i_x}, \quad y' = \frac{y - c_y}{i_y},$$

где c_x и c_y — новые начала отсчета, i_x и i_y — интервалы по x и по y.

Тогда формула для вычисления r будет иметь вид

$$r_{y'|x'} = \frac{n\Sigma mx'y' - \Sigma m\,x'y'}{\sqrt{n\Sigma mx'^2 - (\Sigma mx')^2}\quad\sqrt{n\Sigma my'^2 - (\Sigma my')^2}} \qquad (27)$$

При этом $r_{y|x} = r_{y'|x'}$.

Для сгруппированных данных можно пользоваться также следующими формулами:

$$r = \frac{\Sigma m_{xy}\,xy - n\,\bar{x}\,\bar{y}}{n\,\sigma_x \sigma_y} \qquad (28)$$

или

$$r = \frac{\frac{\Sigma m_{xy}\, xy}{n} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}, \qquad (29)$$

где m_{xy} — частоты в каждой клетке корреляционной таблицы, n — общее число случаев.

Мы привели ряд формул для вычисления коэффициента корреляции для несгруппированных и сгруппированных данных, встречающихся в литературе. Выбор той или иной формулы для расчетов зависит от материала наблюдений и числа случаев. Поэтому, прежде чем приступить к расчетам, необходимо, учитывая техническую сторону расчетов, выбрать ту формулу для определения r, которая даст менее громоздкие расчеты.

§ 5. СВОЙСТВА КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

1-е свойство. Величина коэффициента корреляции не изменяется, если из всех значений x и y вычтем какие-нибудь постоянные a и b и полученные результаты разделим на какие-то постоянные k и l (это свойство основано на свойствах среднего арифметического и дисперсии):

$$r = \frac{\Sigma m_{xy}\left[\left(\frac{x-a}{k} - \frac{\bar{x}-a}{k}\right)\left(\frac{y-b}{l} - \frac{\bar{y}-в}{l}\right)\right]}{n\,\frac{\sigma_x}{k}\,\frac{\sigma_y}{l}} =$$

$$= \frac{\Sigma m_{xy}\,\frac{x-\bar{x}}{k}\,\frac{y-\bar{y}}{l}}{n\,\frac{\sigma_x}{k}\,\frac{\sigma_y}{l}} = \frac{\Sigma m_{xy}\,(x-\bar{x})(y-\bar{y})}{n\,\sigma_x\,\sigma_y}.$$

Иначе говоря, коэффициент корреляции не изменится, если от первоначальных значений x и y перейти к новым условным значениям x' и y':

$$x' = \frac{x-a}{k}; \quad y' = \frac{y-b}{l}; \quad x = kx' + a; \quad y = ly' + b;$$

$$\bar{x} = k\bar{x}' + a; \quad \bar{y} = l\bar{y}' + b; \quad \sigma_x = k\sigma_{x'}; \quad \sigma_y = l\sigma_{y'};$$

отсюда

$$r_{y|x} = \frac{\Sigma m_{xy}(x-\bar{x})(y-\bar{y})}{n\,\sigma_x\,\sigma_y} = \frac{\Sigma m_{xy}\,k(x'-\bar{x}')\,l(y'-\bar{y}')}{n\,k\,\sigma_{x'}\,l\,\sigma_{y'}};$$

так как $m_{xy} = m_{x'y'}$, то

$$r_{y/x} = \frac{\Sigma m_{x'y'}(x' - \bar{x'})(y' - \bar{y'})}{n\sigma_{x'}\sigma_{y'}} = r_{y'/x'},$$

т. е. $r_{y/x} = r_{y'/x'}$.

На этом свойстве основан упрощенный метод вычислений коэффициента корреляции. Это метод расчета коэффициента корреляции по сгруппированным данным по условным величинам переменных, который приведен ниже (гл. II, § 8).

Указанное свойство можно также сформулировать следующим образом: линейные преобразования, сводящиеся к изменению масштаба или начала отсчета переменных величин не изменяют коэффициента корреляции между ними.

2-е свойство. Коэффициент корреляции равен отношению разности средней арифметической произведений всевозможных значений x и y ($\overline{xy}$) и произведения средних $\bar{x}\bar{y}$ к произведению стандартных отклонений σ_x и σ_y:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y},$$

где $\overline{xy} = \frac{\Sigma mxy}{n}$ — среднее арифметическое произведения x на y.

3-е свойство. Величина коэффициента корреляции не превосходит единицы:

$$-1 \leqslant r \leqslant 1.$$

4-е свойство. При наличии линейной функциональной связи между величинами $r = \pm 1$. При этом прямые регрессии y на x и x на y совпадают.

5-е свойство. Чем ближе r к единице, тем теснее прямолинейная корреляция между величинами x и y.

6-е свойство. Если регрессия y на x линейная и коэффициент корреляции равен нулю, то все групповые средние $\bar{y}_x$ совпадают и равны общей средней $\bar{y}$ переменной y, т. е. между y и x в этом случае линейной корреляционной связи нет. В этом случае прямые регрессии будут параллельны осям координат.

Однако, когда $r = 0$, невозможна лишь линейная корреляционная связь. Нелинейные корреляционные связи при этом возможны, не исключаются и нелинейные функциональные связи.

Таким образом, к оценке связи только по одному коэффициенту корреляции r нужно относиться очень осторожно. Если $r=0$, то это еще не означает, что связи нет, связь может быть нелинейной, которая не учитывается величиной коэффициента корреляции r.

На практике принято считать, что величины достаточно связаны, если $r>0{,}6$. Однако можно говорить о связи и при $r<0{,}6$, если эту связь можно объяснить физическими причинами.

При физически обоснованной связи приведенные свойства коэффициента корреляции позволяют считать его доброкачественной мерой тесноты связи в условиях линейной корреляции. Правильное же истолкование его при криволинейной корреляции представляет нелегкую задачу. Для измерения тесноты криволинейной связи пользуются корреляционным отношением. В то время, как значения корреляционного отношения не зависят от формы кривых регрессии, значения коэффициента корреляции r существенно зависят от того, в какой мере линии регрессии отличаются от прямых.

§ 6. УРАВНЕНИЕ ЛИНЕЙНОЙ КОРРЕЛЯЦИОННОЙ СВЯЗИ МЕЖДУ ДВУМЯ ПЕРЕМЕННЫМИ

После определения коэффициента корреляции r и установления линейной корреляционной связи приступают к нахождению параметров уравнений этой связи. Общий вид этих уравнений $y=ax+b$ и $x=a_1y+b_1$.

Из аналитической геометрии известно, что если прямые проходят через некоторую точку с координатами $\bar{x}$ и $\bar{y}$, то уравнения этих прямых регрессии имеют вид

$$y-\bar{y}=a_{y/x}(x-\bar{x}) \text{ и } x-\bar{x}=a_{1y/x}(y-\bar{y}). \quad (30)$$

Коэффициенты $a_{y/x}$ и $a_{1x/y}$ называются коэффициентами линейной регрессии, а уравнения — уравнениями регрессии.

В предыдущем разделе мы рассмотрели связь между коэффициентом корреляции r и коэффициентами регрессии a и a_1.

Коэффициенты регрессии в уравнениях линейной связи двух переменных равны

$$a=r\frac{\sigma_y}{\sigma_x}; \quad a_1=r\frac{\sigma_x}{\sigma_y},$$

Величину параметра b для уравнения $y=ax+b$ получаем из уравнения $y-\bar{y}=a(x-\bar{x})$ путем нахождения величины выражения $\bar{y}-a\bar{x}$.

Таким образом, вычислив средние арифметические значения $\bar{x}$ и $\bar{y}$, средние квадратические отклонения σ_x и σ_y и коэффициент корреляции r, мы легко путем решения уравнений связи вида

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \quad \text{и} \quad x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

можем вычислить параметры a, a_1, b, b_1 уравнений

$$y=ax+b \text{ и } x=a_1y+b_1,$$

по которым можно будет в зависимости от изменений одной величины получать наиболее вероятное значение другой величины.

Обычно при определении зависимости одного элемента от другого находят только одно уравнение $y=ax+b$, но если необходимо также установить как зависит x от y, то ищут и второе уравнение $x=a_1y+b_1$.

§ 7. СРЕДНЯЯ И ВЕРОЯТНАЯ ОШИБКИ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ. СРЕДНЯЯ ОШИБКА УРАВНЕНИЯ РЕГРЕССИИ

При исследовании корреляционной связи переменных величин даже при большом числе наблюдений мы имеем дело с выборочной совокупностью величин из генеральной совокупности p.

Генеральной совокупностью p можно считать все те наблюдения, которые теоретически можно было бы сделать, изучая взаимосвязь двух или нескольких явлений. Практически получение генеральной совокупности при бесконечном числе членов часто бывает неосуществимо или слишком громоздко и обычно имеют дело с выборочной совокупностью, т. е. когда число наблюдений n значительно меньше, чем могло быть при генеральной совокупности.

Поэтому важно знать различие выборочных коэффициентов корреляции и выборочных коэффициентов уравнений регрессии от тех же величин генеральной совокупности. При очень большой выборке эти величины могут мало различаться.

В математической статистике имеется положение, что с вероятностью, близкой к единице, можно утверждать, что при достаточно большом объеме выборки n выборочный коэффи-

циент линейной корреляции $r_в$ будет мало отличаться от такого же генерального коэффициента корреляции $r_г$.

Средняя квадратическая ошибка коэффициента корреляции σ_r при замене генерального коэффициента корреляции $r_г$ выборочным $r_в$ равна

$$\sigma_r = \pm \frac{1 - r^2}{\sqrt{n}} \qquad (31)$$

и

$$\sigma_r = \pm \frac{1 - r^2}{\sqrt{n-1}} \text{ (для малых и средних } r\text{)}. \qquad (32)$$

С вероятностью 0,954 считают, что случайная ошибка не будет превышать $2\sigma_r$ т. е.

$$r_г = r_в \pm 2\sigma_r.$$

После расчета σ_r находят отношение $\frac{r}{\sigma_r}$. Если величина этого отношения превышает три при числе наблюдений больше 50, то можно считать, что полученный выборочный коэффициент корреляции надежен и отображает искомую связь $\frac{r}{\sigma_r} > 3$.

Величина $r - 3\sigma_r$ является гарантийным минимумом, а величина $r + 3\sigma_r$ — гарантийным максимумом коэффициента корреляции, т. е.

$$r \pm 3\sigma_r.$$

Кроме средней ошибки коэффициента корреляции σ_r, можно вычислять вероятную ошибку коэффициента корреляции (E_r), которая составляет 0,67 σ_r:

$$E_r = \pm 0{,}67 \frac{1 - r^2}{\sqrt{n}}. \qquad (33)$$

Вероятное значение коэффициента корреляции заключено в пределах $r \pm E_r$, а предельная величина близка к $r \pm 4E_r$. Если $r > 4E_r$, то связь доказана.

Если выборка очень мала (число случаев меньше 50), то вычисление средних ошибок по указанным формулам нежелательно, в таких случаях следует оценивать корреляцию по критерию Фишера.

$$z = \frac{1}{2} \ln \frac{1 + r}{1 - r}, \qquad (34)$$

выражая через десятичные логарифмы, получим

$$z = 1{,}151 \lg \frac{1+r}{1-r}. \qquad (35)$$

Функция z подчиняется закону нормального распределения. Вычислив z, находят ее ошибку по формуле

$$\sigma_z = \frac{1}{\sqrt{n-3}}, \qquad (36)$$

где n — объем выборки.

Кроме σ_z, вычисляют еще вероятное отклонение

$$p_z = 0{,}675 \frac{1}{\sqrt{n-3}} = 0{,}675\,\sigma_z, \qquad (37)$$

где n — число пар значений величин, вошедших в определение r и z.

Допустим, что мы нашли коэффициент корреляции нашей связи равным 0,62. Находим z из уравнения $z = 1{,}151 \lg \frac{1+r}{1-r}$;

$$z = 0{,}725;\ \sigma_z = 0{,}123;\ p_z = 0{,}083.$$

В табл. 5 дано $r = f(z)$. Находим интервал изменения r, соответственно интервалу изменения z:

$$z \pm \sigma_z = 0{,}725 \pm 0{,}123 = \begin{cases} 0{,}85 — \text{верхняя граница} \\ 0{,}60 — \text{нижняя граница}. \end{cases}$$

По верхней и нижней границе z находим границы изменения r по табл. 5, откуда имеем $r = 0{,}69$ — верхняя граница и $r = 0{,}54$ — нижняя граница, т. е. $r = \begin{cases} 0{,}69 \\ 0{,}54 \end{cases}$.

При корреляционной связи, где каждому значению x_i соответствует ряд значений y_i, вычисленное по уравнению $\bar{y}_i$ в зависимости от x_i будет, естественно, отличаться от каждого значения y_i этого ряда, которые соответствовали значению x_i при наблюдениях.

Зная значение x и подставляя его в полученное уравнение для расчета y, получаем y, как бы с ошибкой или с отклонением от наблюдений; получаем, как говорилось ранее, среднюю величину $\bar{y}_i$ при заданном значении x_i. Определение величины этой ошибки позволит судить о том, насколько рассеяны точки корреляционного поля относительно линии прямой регрессии.

Значения коэффициента корреляции r в зависимости от значений z функции Фишера

Таблица 5

z	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10
0,0	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0699	0,0798	0,0898	0,0997
0,1	0,1096	0,1194	0,1293	0,1391	0,1489	0,1586	0,1684	0,1781	0,1877	0,1974
0,2	0,2070	0,2165	0,2260	0,2355	0,2449	0,2543	0,2636	0,2729	0,2821	0,2913
0,3	0,3004	0,3095	0,3185	0,3275	0,3364	0,3452	0,3540	0,3627	0,3714	0,3800
0,4	0,3885	0,3969	0,4053	0,4136	0,4219	0,4301	0,4382	0,4462	0,4542	0,4621
0,5	0,4699	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299	0,5370
0,6	0,5441	0,5511	0,5580	0,5649	0,5717	0,5784	0,5850	0,5915	0,5980	0,6044
0,7	0,6107	0,6169	0,6231	0,6291	0,6351	0,6411	0,6469	0,6527	0,6584	0,6640
0,8	0,6696	0,6751	0,6805	0,6858	0,6911	0,6963	0,7014	0,7064	0,7114	0,7163
0,9	0,7211	0,7259	0,7306	0,7352	0,7398	0,7443	0,7487	0,7531	0,7574	0,7616
1,0	0,7658	0,7699	0,7739	0,7779	0,7818	0,7857	0,7895	0,7932	0,7969	0,8005
1,1	0,8041	0,8076	0,8110	0,8144	0,8178	0,8210	0,8243	0,8275	0,8306	0,8337
1,2	0,8367	0,8397	0,8426	0,8455	0,8483	0,8511	0,8538	0,8565	0,8591	0,8617
1,3	0,8643	0,8668	0,8692	0,8717	0,8741	0,8764	0,8787	0,8810	0,8832	0,8854
1,4	0,8875	0,8896	0,8917	0,8937	0,8957	0,8977	0,8996	0,9015	0,9033	0,9051
1,5	0,9069	0,9087	0,9104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9201	0,9217
1,6	0,9232	0,9246	0,9261	0,9275	0,9289	0,9302	0,9316	0,9329	0,9341	0,9354
1,7	0,9366	0,9379	0,9391	0,9402	0,9414	0,9425	0,9436	0,9447	0,9458	0,94681
1,8	0,94783	0,94884	0,94983	0,95080	0,95175	0,95268	0,95359	0,95449	0,95537	0,95624
1,9	0,95709	0,95792	0,95873	0,95953	0,96032	0,96102	0,96185	0,96259	0,96331	0,96403
2,0	0,96473	0,96541	0,96609	0,96675	0,96739	0,96803	0,96865	0,96926	0,96986	0,97045
2,1	0,97103	0,97159	0,97215	0,97269	0,97323	0,97375	0,97426	0,97477	0,97526	0,97574
2,2	0,97622	0,97668	0,97714	0,97759	0,97803	0,97846	0,97888	0,97929	0,97970	0,98010
2,3	0,98049	0,98087	0,98124	0,98161	0,98197	0,98233	0,98267	0,98301	0,98335	0,98367
2,4	0,98399	0,98431	0,98462	0,98492	0,98522	0,98551	0,98579	0,98607	0,98635	0,98661
2,5	0,98688	0,98714	0,98739	0,98764	0,98788	0,98812	0,98835	0,98858	0,98881	0,98903
2,6	0,98924	0,98945	0,98966	0,98987	0,99007	0,99026	0,99045	0,99064	0,99083	0,99101
2,7	0,99118	0,99136	0,99153	0,99170	0,99186	0,99202	0,99218	0,99233	0,99248	0,99263
2,8	0,99278	0,99292	0,99306	0,99320	0,99333	0,99346	0,99359	0,99372	0,99384	0,99396
2,9	0,99408	0,99420	0,99431	0,99443	0,99454	0,99464	0,99475	0,99485	0,99495	0,99505

Средняя ошибка уравнения регрессии y по x выражается следующей формулой:

$$s_y = \pm \sigma_y \sqrt{1 - r^2}. \quad (38)$$

Средняя ошибка уравнения регрессии x по y аналогично равна

$$s_x = \pm \sigma_x \sqrt{1 - r^2}, \quad (39)$$

где σ_y и σ_x — средние квадратические отклонения от средних арифметических величин;

$$\sigma_y = \sqrt{\frac{\Sigma (y - \overline{y})^2}{n}}; \quad \sigma_x = \sqrt{\frac{\Sigma (x - \overline{x})^2}{n}},$$

а r — коэффициент корреляции связи x и y.

При $r = \pm 1$ значения s_y и s_x равны нулю. В этом случае уравнения регрессии дают точные значения y по x и наоборот, т. е. мы имеем линейную функциональную связь.

Максимальная величина ошибки уравнения регресии в три раза больше средней ошибки:

$$s_{\text{макс}} = 3\, s_y. \quad (40)$$

Средние ошибки выборочных коэффициентов регрессии a и a_1 в уравнениях могут быть выражены следующими формулами: в уравнении $y = ax + b$ формулой

$$\sigma_a = \pm \frac{\sigma_y}{\sigma_x} \sqrt{\frac{1 - r^2}{n}}, \quad (41)$$

в уравнении $x = a_1 y + b_1$ формулой

$$\sigma_{a_1} = \pm \frac{\sigma_x}{\sigma_y} \sqrt{\frac{1 - r^2}{n}}. \quad (42)$$

Таким образом, чем больше число наблюдений n и коэффициент корреляции r, тем меньше ошибки коэффициента корреляции, коэффициентов регрессии и уравнения регрессии.

§ 8. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЛИНЕЙНОЙ СВЯЗИ ДВУХ ПЕРЕМЕННЫХ ВЕЛИЧИН ПО СГРУППИРОВАННЫМ ДАННЫМ

(связь запасов продуктивной влаги в различных слоях почвы)

Когда число случаев пар наблюдений велико (больше 100), нахождение уравнений связи и вычисление коэффициентов корреляции для уменьшения громоздкости расчетов проводят способом группировки данных. При этом учитывают веса или частоты данных величин по корреляционной таблице или корреляционной решетке.

В гл. II (§ 1) приведен пример построения корреляционного поля и корреляционной таблицы (решетки) связи запасов продуктивной влаги различных слоев почвы осенью под озимой пшеницей (рис. 2 и 3, табл. 2).

Табл. 2 нам необходима была для построения эмпирических линий регрессии (рис. 3). Для нахождения уравнения данной связи и теоретической линии регрессии $y=ax+b$ способом группировки данных по частотам возьмем этот же пример, но несколько видоизменим корреляционную таблицу с дополнением граф, необходимых для расчетов коэффициента корреляции и уравнения регрессии (табл. 6).

Расчеты граф табл. 6 проводят следующим образом.

В данном примере мы имеем 135 случаев парных наблюдений величин запасов влаги слоев почвы 0—20 и 20—50 *см*. Обозначаем через x запасы влаги слоя почвы 0—20 *см*, а через y — запасы влаги слоя 20—50 *см*. Разбиваем значения x и y на интервалы по 5 *мм* (интервалы берутся произвольно и могут быть разными для x и для y) и записываем средние значения интервалов в графах таблицы. Затем ведем подсчет числа случаев парных наблюдений запасов влаги, попадающих в соответствующие интервалы, т. е. находим частоты повторения y по данному интервалу x и наоборот. Делать это проще с графика корреляционного поля, как было изложено в § 1, гл. II.

Вертикальные столбцы дадут нам число частот (m_x) различных y по каждому данному интервалу x, а горизонтальные столбцы дадут число частот (m_y) различных x по каждому данному интервалу y. Более подробно построение этой части корреляционной таблицы дано в гл. II, § 1.

На основании 1-го свойства коэффициент линейной корреляции r остается неизменным, если от первоначальных значений x и y перейти к условным значениям x' и y'. Введением условных значений x' и y' значительно упрощаются расчеты.

Обозначим средние горизонтальные и вертикальные столбцы через 0, т. е. примем их за начальные, а следующие от них столбцы вправо и влево (для вертикальных) и вверх и вниз (для горизонтальных) по мере возрастания значений x или y будем обозначать $+1$, $+2$ и т. д., а по мере убывания значений x или y будем обозначать -1, -2 и т. д.

Таким образом, за условную единицу мы принимаем один интервал значений x или y в 5 действительных единиц.

Возьмем за начальные столбцы для x и для y средние столбцы с интервалами 20—25 *мм* и обозначим их через 0. Соседние столбцы с возрастанием значений x и y обозначим через $+1$, а соседние столбцы с убыванием значений обозначим через -1 и т. д.

Таким образом, за начало отсчета мы взяли $x_0=22{,}5$ *мм* и $y_0=22{,}5$ *мм*, обозначив их через $x'=0$ и $y'=0$.

Таблица 6

Корреляционная таблица для расчета уразнения связи запасов влаги различных слоев почвы по сгруппированным данным

y \ x Интервал Δy	Интервал Δx Середина Δx и Δy	Условные единицы x' / y'	0—5	5—10	10—15	15—20	20—25	25—30	30—35	35—40	m_y	$\Sigma m_y x'_y$	$m_y x'_y y'$	$m_y y'$	y'^2	$m_y y'^2$
			2,5	7,5	12,5	17,5	22,5	27,5	32,5	37,5						
			—4	—3	—2	—1	0	+1	+2	+3						
0—5	2,5	—4														
5—10	7,5	—3		1							1	—3	9	—3	9	9
10—15	12,5	—2	1	6							7	—22	44	—14	4	28
15—20	17,5	—1		5	9						14	—33	33	—14	1	14
20—25	22,5			4	9	5	3				21	—35	0	0	0	0
25—30	27,5	+1			7	9	3				19	—23	—23	19	1	19
30—35	32,5	+2			2	9	12	4	2		29	—5	—10	58	4	116
35—40	37,5	+3				7	8	6	4		25	7	21	75	9	225
40—45	42,5	+4				1	2	5	3	3	14	19	76	56	16	224
45—50	47,5	+5						1	2	2	5	11	55	25	25	125
m_x			1	16	27	31	28	16	11	5	$n=135$	$\Sigma=-84$	$\Sigma=205$	$\Sigma=202$	$\Sigma=69$	$\Sigma=760$
$\Sigma m_x y'_x$			—2	—20	2	52	59	51	38	22	$\Sigma=202$					
$m_x y'_x x'$			8	60	—4	—52	0	51	76	66	$\Sigma=205$					
$m_x x'$			—4	—48	—54	—31	0	16	22	15	$\Sigma=-84$					
x'^2			16	9	4	1	0	1	4	9	$\Sigma=44$					
$m_x x'^2$			16	144	108	31	0	16	44	45	$\Sigma=404$					

Обозначив горизонтальные и вертикальные столбцы условными единицами, переходим к дальнейшим расчетам корреляционной таблицы, необходимым для нахождения r и уравнения связи $y=ax+b$.

Находим суммы частот m_y, записанных в горизонтальных столбцах, и суммы частот m_x, записанных в вертикальных столбцах; m_y обозначает сумму частот различных x для данного постоянного интервала y; m_x обозначает сумму частот различных y для данного постоянного интервала x. Сумма всех m_x равна общему числу случаев n, и сумма всех m_y также равна общему числу всех случаев n.

В нашем примере $\Sigma m_x = \Sigma m_y = n = 135$.

Далее находим значения столбцов $\Sigma m_x y_x'$ и $\Sigma m_y x_y'$, беря y' и x' в условных единицах. Значение $\Sigma m_y x_y'$ для каждого горизонтального столбца получается как алгебраическая сумма из произведений каждого числа частот m на соответствующее ему значение x' в условных единицах.

Например, в первом горизонтальном столбце с частотами имеется одно значение частоты, равное 1. Умножаем его на $x'=-3$ и ставим значение -3 в первую горизонтальную графу $\Sigma m_y x'_y$. Во втором горизонтальном столбце с частотами, значение частоты 1 умножаем на $x'=-4$, получаем -4. Затем второе значение частоты в этом же горизонтальном столбце, равное 6, умножаем на $x'=-3$, получаем -18. Складываем -4 и -18, сумма равна -22, ее записываем во вторую горизонтальную графу $\Sigma m_y x'_y$. В третьем горизонтальном столбце частоту $m=5$ умножаем на $x'=-3$, получаем -15, затем $m=9$ умножаем на $x'=-2$, получаем -18. Складываем -15 и -18, получаем сумму -33, которую записываем в третью графу $\Sigma m_y x'_y$ и т. д.

Величина $\Sigma m_x y'_x$ для каждого вертикального столбца получается как алгебраическая сумма (с учетом знака) из произведений каждого числа частот m на соответствующее ему значение y' в условных единицах. Например, в первом вертикальном столбце частоту $m=1$ умножаем на $y'=-2$, получаем -2 и записываем в графу $\Sigma m_x y'_x$ данного столбца.

Во втором вертикальном столбце умножаем $1\cdot(-3)=-3$; $6\cdot(-2)=-12$; $5\cdot(-1)=-5$ и $4\cdot 0=0$. Складывая эти произведения, получаем сумму, равную -20, и записываем ее в графу $\Sigma m_x y'_x$ второго вертикального столбца и т. д. Затем находим общую алгебраическую сумму всего вертикального столбца $\Sigma m\ x'_y=-84$ и общую сумму всего горизонтального столбца $\Sigma m_x y'_x=202$.

Подсчитываем далее вертикальные и горизонтальные столбцы значений $m_y x_y' y'$ и $m_x y_x' x'$. Они получаются умножением соответствующих значений $m_y x'_y$, получен-

ных в предыдущем столбце на соответствующие им условные y', или $m_x y_x'$ на x'. Например, в первой строке $m_y x_y'$ равно -3. Умножаем -3 на y' также равное -3, получаем 9 и заносим эту цифру в графу $m_y x_y' y'$.

Во второй графе -22 умножаем на -2, получаем 44 и т. д. В конце всего вертикального столбца подсчитываем алгебраическую сумму

$$\Sigma m_y x_y' y' = 205.$$

Рассчитываем графу $m_x y_x' x'$, умножая значения $m_x y'_x$ предшествующей графы на соответствующие им условные значения x'. Получаем числа $-2 \cdot (-4) = 8$; $-20 \cdot (-3) = 60$; $2 \cdot (-2) = -4$ и т. д. В конце этой горизонтальной графы подсчитываем алгебраическую сумму

$$\Sigma m_x y_x' x' = 205.$$

Сумма $\Sigma m_y x_y' y'$ должна быть равна $\Sigma m_x y_x' x'$. В нашем примере $\Sigma m_y x_y' y' = \Sigma m_x y_x' x' = 205$. В этом состоит проверка правильности расчетов данной корреляционной таблицы.

Коэффициент корреляции r данной связи мы рассчитаем по формуле

$$r = \frac{\frac{\Sigma m x' y'}{n} - \bar{x}' \bar{y}'}{\sigma_{x'} \sigma_{y'}}.$$

Значение $\Sigma m x' y'$ мы нашли, оно равно 205. Находим выражение

$$\frac{\Sigma m x' y'}{n} = \frac{205}{135} = 1{,}52.$$

Рассчитываем другие члены $\bar{x}'$, $\bar{y}'$, $\sigma_{x'}$ и $\sigma_{y'}$, входящие в формулу коэффициента корреляции.

Находим $\bar{x}'$ и $\bar{y}'$ — средние арифметические величины:

$$\bar{x}' = \frac{\Sigma m_x x'}{n} = \frac{-84}{135} = -0{,}622,$$

$$\bar{y}' = \frac{\Sigma m_y y'}{n} = \frac{202}{135} = 1{,}496.$$

Средние квадратические отклонения $\sigma_{x'}$ и $\sigma_{y'}$ находим по формулам

$$\sigma_{x'} = \sqrt{\frac{\Sigma m_x x'^2}{n} - (\bar{x}')^2}; \quad \sigma_{y'} = \sqrt{\frac{\Sigma m_y y'^2}{n} - (\bar{y}')^2}.$$

Для нахождения $\Sigma m_x x'^2$ берем квадраты значений x' каждого столбца и умножаем на m_x, затем берем сумму этих значений. В нашем примере

$$\Sigma m_x x'^2 = 404, \text{ а } \frac{\Sigma m_x x'^2}{n} = \frac{404}{135} = 2{,}993;$$

$\bar{x}'$ у нас было вычислено: $\bar{x}' = -0{,}622$. Возводим $\bar{x}'$ в квадрат: $\bar{x}'^2 = 0{,}387$. Находим

$$\sigma_{x'} = \sqrt{\frac{\Sigma m_x x'^2}{n} - (\bar{x}')^2} = \sqrt{2{,}993 - 0{,}387} = \sqrt{2{,}606} = 1{,}61.$$

Подобным образом находим $\sigma_{y'}$. Берем квадраты значений y' каждого горизонтального столбца, умножаем на число равных между собой y' этого столбца, т. е. на m_y, и получаем величины $m_y y'^2$ каждого столбца. Складывая их, получаем $\Sigma m_y y'^2 = 760$. Отсюда находим

$$\frac{\Sigma m_y y'^2}{n} = \frac{760}{135} = 5{,}630.$$

По предыдущим вычислениям

$$\bar{y}' = 1{,}496; \quad \bar{y}'^2 = 2{,}238.$$

Находим $\sigma_{y'}$:

$$\sigma_{y'} = \sqrt{\frac{\Sigma m_y y'^2}{n} - (\bar{y}')^2} = \sqrt{5{,}630 - 2{,}238} = \sqrt{3{,}392} = 1{,}84.$$

Таким образом, мы нашли все величины, необходимые для вычисления коэффициента корреляции r,

$$r_{y'/x'} = \frac{\frac{\Sigma m x' y'}{n} - \bar{x}' \bar{y}'}{\sigma_{x'} \sigma_{y'}} = \frac{1{,}52 - (-0{,}622)(1{,}496)}{1{,}61 \cdot 1{,}84} = \frac{2{,}450}{2{,}962} = 0{,}83.$$

Находим среднюю ошибку коэффициента корреляции

$$\sigma_r = \pm \frac{1 - r^2}{\sqrt{n}} = \pm \frac{1 - (0{,}83)^2}{\sqrt{135}} = \pm 0{,}03,$$

$$r \pm \sigma_r = 0{,}83 \pm 0{,}03 = \begin{cases} 0{,}86 \\ 0{,}80. \end{cases}$$

Максимальное и минимальное значение r лежит в пределах $r \pm 3\sigma_r$; $3\sigma_r = 0{,}09$;

$$r \pm 3\sigma_r = 0{,}83 \pm 0{,}09 = \begin{cases} 0{,}92 \\ 0{,}74. \end{cases}$$

Вероятная ошибка коэффициента корреляции r

$$E_r = \pm 0{,}67\,\sigma_r = \pm 0{,}67 \cdot 0{,}03 = \pm 0{,}02.$$

Предельные значения для r могут быть также выражены через вероятную ошибку. Они равны $r \pm 4E_r$:

$$r \pm 4E_r = 0{,}83 \pm 0{,}08 = \begin{cases} 0{,}91 \\ 0{,}75 \end{cases}.$$

Обычно находится одна из ошибок σ_r или E_r, так как предельные значения коэффициента корреляции r, вычисленные по σ_r $(r \pm 3\sigma_r)$ и по E_r $(r \pm 4E_r)$, близки между собой.

Мы получили величину коэффициента корреляции $r = 0{,}83$ для условных единиц x' и y'.

Исходя из 1-го свойства коэффициента корреляции (§ 5, гл. II), можно его принять и для наших действительных значений x и y.

Вычисляем уравнение линейной регрессии связи двух переменных величин по формуле

$y - \bar{y} = R(x - \bar{x})$, где $R = r\frac{\sigma_y}{\sigma_x}$, или по формуле $y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$.

Средние квадратические отклонения σ_x и σ_y у нас выражены в условных единицах. Переводим их в действительные. За условную единицу мы принимали 5 действительных единиц, следовательно, для получения σ_x и σ_y в действительных единицах их значения в условных умножаем на 5. В результате этого получаем

$$\sigma_y = \sigma_{y'} \cdot 5 = 1{,}84 \cdot 5 = 9{,}2; \quad \sigma_x = \sigma_{x'} \cdot 5 = 1{,}61 \cdot 5 = 8{,}05.$$

Среднее арифметическое значение $\bar{x}$ в действительных единицах вычисляем по формуле

$$\bar{x} = x_0 + \bar{x}'\,d,$$

где x_0 — величина начального отсчета $\bar{x}$ при $x' = 0$; d — количество действительных единиц, взятых за одну условную.

В нашем примере $x_0 = 22{,}5$, а $d = 5$. Отсюда $\bar{x} = 22{,}5 + (-0{,}622) \cdot 5 = 22{,}5 - 3{,}11 = 19{,}39$.

Подобным образом вычисляем $\bar{y}$ в действительных единицах; y_0 — также было равно 22,5,

$$\bar{y} = y_0 + y' d;\ d = 5.\ \text{Отсюда}$$

$$\bar{y} = 22{,}5 + 1{,}496 \cdot 5 = 22{,}5 + 7{,}48 = 29{,}98.$$

Находим уравнение регрессии

$$y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) = 29{,}98 + 0{,}83 \cdot \frac{9{,}20}{8{,}05}(x - 19{,}39) =$$

$$= 29{,}98 + 0{,}95\,(x - 19{,}39).$$

Таким образом, $y = 0{,}95x + 11{,}56$.

Мы получили уравнение связи запасов влаги различных слоев почвы.

Зная запасы влаги верхнего (0—20 *см*) слоя почвы (x), мы можем рассчитать запасы влаги слоя почвы 20—50 *см* (y) и таким образом иметь представление о запасах влаги полуметрового слоя почвы.

Найдем среднюю квадратическую ошибку уравнения регрессии:

$$s_y = \pm \sigma_y \sqrt{1 - r^2} = \pm 9{,}2 \sqrt{1 - (0{,}83)^2} = \pm 5{,}15\ \textit{мм}.$$

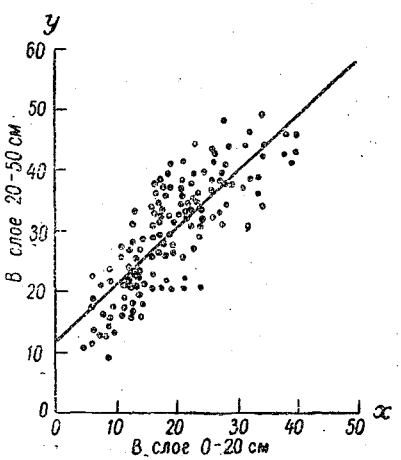

Рис. 4. Связь запасов продуктивной влаги различных слоев почвы осенью под озимой пшеницей.

По найденному уравнению построим теоретическую линию регрессии y по x на корреляционном поле (рис. 4). При

$x=0$, $y=11{,}56$. Откладываем точку с такими координатами на оси Y. Зададим еще ряд значений x:

при $x=20$; $y=30{,}56$,

при $x=40$; $y=49{,}56$.

Откладываем на графике точки с данными координатами и через них проводим прямую линию. Это и есть искомая теоретическая линия регрессии, уравнение которой мы нашли: $y=0{,}95x+11{,}56$. Можно было провести линию по двум точкам, задав значения x начала и конца расположения точек корреляционного поля.

Метод сгруппированных данных удобен в расчетах, менее громоздок, но и менее точен. Им можно пользоваться только при большом числе наблюдений (более 100). При малом числе наблюдений метод группировки данных применять не следует, так как это может повлечь к ошибкам, о которых говорилось в гл. II, § 1.

§ 9. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЛИНЕЙНОЙ СВЯЗИ ДВУХ ПЕРЕМЕННЫХ ВЕЛИЧИН ПО НЕСГРУППИРОВАННЫМ ДАННЫМ (зависимость урожая озимой пшеницы от весенних запасов влаги в почве)

Уравнение регрессии и коэффициент корреляции двух переменных величин часто находят не группируя данные и не прибегая к условным единицам. Когда число случаев пар наблюдений не слишком велико (не больше 100), то расчеты и по несгруппированным данным могут быть не слишком громоздкими, если применить для вычисления r формулу (23), где величины выражены отклонениями от средних. В этом способе есть свои преимущества в более быстром прямом пути расчета, при котором меньше возможностей допустить ошибки.

Приведем пример расчета коэффициента корреляции и уравнения регрессии по указанному способу.

Нами была найдена зависимость урожая озимой пшеницы от весенних запасов влаги в почве (рис. 5). Эта зависимость более четко проявляется в зоне недостаточного летнего увлажнения почвы на Украине и на Северном Кавказе. В пределах запасов продуктивной влаги весной в метровом слое почвы от 100 до 200 *мм*, при благоприятных осенних и зимних условиях, когда весной число стеблей у озимой пшеницы на 1 $м^2$ составляет 1000—1900, при высокой агротехнике для одного и того же сорта эта зависимость урожая озимой пшеницы от весенних запасов влаги является прямолинейной (рис. 5).

Проведя анализ данных об урожаях озимой пшеницы при высокой агротехнике на полях опытных станций, сельскохо-

зяйственных институтов, передовых колхозов и совхозов Украины и Северного Кавказа и наблюдений над запасами продуктивной влаги, проводимыми на тех же полях агро- и гидрометеостанциями Гидрометеослужбы, мы получили большой материал сопряженных наблюдений двух величин для каждого года — урожаев озимой пшеницы и весенних запасов продуктивной влаги в метровом слое почвы.

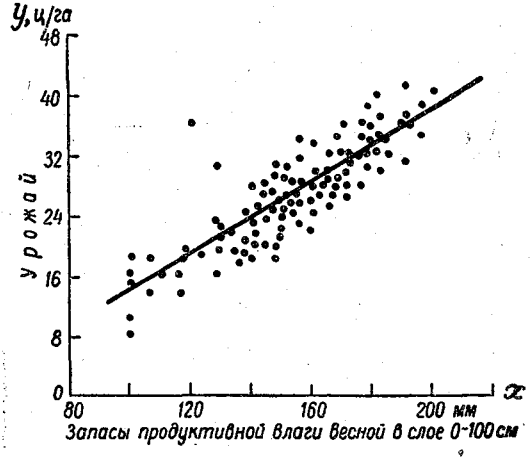

Рис. 5. Зависимость урожая озимой пшеницы от весенних запасов влаги в почве на Украине и на Северном Кавказе при числе стеблей пшеницы весной 1000—2000 на 1 $м^2$; y — урожай озимой пшеницы (в *ц/га*), x — запасы влаги (в *мм*) в метровом слое почвы в декаду последнего перехода температуры воздуха через +5° весной.

Урожай — зависимая переменная величина от запасов влаги—является функцией (y). Запасы влаги весной — независимая от урожая переменная величина является аргументом (x).

Данные каждого года x и y под одним порядковым номером поместили в таблицу и получили таблицу-сводку, где было 100 сопряженных значений x и y, т. е. n — общее число случаев, равное 100.

После этого, отложив на вертикальной оси значения y — урожаев озимой пшеницы, а на горизонтальной оси значения x — запасов влаги весной, мы для каждой пары значений x и y (откладывая их одновременно) получили точки на плоскости с координатами x и y. Таким образом, мы получили корреляционное поле из 100 случаев пар двух переменных величин x и y.

По расположению точек на графике ясно видно, что между этими величинами существует тесная линейная корреляционная связь, для которой необходимо найти уравнение прямой линии регрессии и коэффициент корреляции, определяющий степень тесноты этой связи.

Так как значения запасов влаги метрового слоя представляют собой большие величины, то для нахождения коэффициента корреляции и уравнения прямой регрессии удобнее всего использовать формулу r, где все расчеты ведутся не через сами величины, а через их отклонения (Δ) по сравнению со средними величинами, тем самым исключая громоздкие большие числа самих величин: $r=\frac{\Sigma(x-\overline{x})(y-\overline{y})}{\sqrt{\Sigma(x-\overline{x})^2\,\Sigma(y-\overline{y})^2}}$.

Для облегчения расчетов по указанной формуле необходимо составить табл. 7 и определить указанные в ней величины.

В первой графе таблицы ставится порядковый номер пары наблюдений x и y. Во второй и третьей графах даны значения каждой пары x и y, относящиеся к одному году и к одному полю.

В четвертой и пятой графах рассчитываются отклонения каждого x_i и y_i от их средних арифметических величин ($\overline{x}$ и $\overline{y}$), где берется алгебраическая разность с учетом знака.

В шестой и седьмой графах берутся квадраты отклонений, а в восьмой графе произведение отклонений (произведение берется с учетом знака).

Девятая и десятая графы рассчитываются для контроля.

Таким образом, для расчета указанных граф табл. 7 мы в первую очередь должны найти средние арифметические значения $\overline{x}$ и $\overline{y}$.

В нашем примере

$$\overline{x}=\frac{\Sigma x}{n}=\frac{15\,300}{100}=153;\ \overline{y}=\frac{\Sigma y}{n}=\frac{2650}{100}=26{,}5.$$

Найдя отклонения $\Delta x=x_i-\overline{x}$ и $\Delta y=y_i-\overline{y}$, их произведение $\Delta x\,\Delta y$, их квадраты Δx^2 и Δy^2, а также суммы этих величин, мы должны провести контроль наших расчетов по формуле

$$\Sigma(x-\overline{x})^2+\Sigma(y-\overline{y})^2+2\Sigma(x-\overline{x})(y-\overline{y})=$$
$$=\Sigma[(x-\overline{x})+(y-\overline{y})]^2.$$

Если наши расчеты в таблице верны, то числа левой и правой части формулы будут одинаковы, в противном случае необходимо провести пересчеты, пока не обнаружится ошибка и не будет соблюдено указанное равенство. Только после этого можно приступать к дальнейшим расчетам коэффициента корреляции и уравнения регрессии.

В нашем примере контроль показал правильность расчетов:

$$\Sigma (x-\overline{x})^2+\Sigma (y-\overline{y})^2+2\Sigma (x-\overline{x})(y-\overline{y})=$$
$$=60049+4777+2\cdot 14705=94236,$$
$$\Sigma [(x-\overline{x})+(y-\overline{y})]^2=94\,236.$$

Коэффициент корреляции r находим по следующей формуле:

$$r=\frac{\Sigma (x-\overline{x})(y-\overline{y})}{\sqrt{\Sigma (x-\overline{x})^2\Sigma (y-\overline{y})^2}}=\frac{\Sigma \Delta x\Delta y}{\sqrt{\Sigma \Delta x^2\Sigma \Delta y^2}}=$$
$$=\frac{14705}{\sqrt{60049\cdot 4777}}=0{,}86.$$

Средняя ошибка коэффициента корреляции

$$\sigma_r=\pm\frac{1-r^2}{\sqrt{n}}=\frac{1-(0{,}86)^2}{\sqrt{100}}=\pm 0{,}026.$$

Найдем вероятную ошибку коэффициента корреляции r

$$E_r=\pm 0{,}67\ \sigma_r=\pm 0{,}67\cdot 0{,}026=\pm 0{,}017.$$

Отсюда вероятное значение коэффициента корреляции заключено в пределах

$$r\pm E_r=0{,}86\pm 0{,}017=\begin{cases}0{,}88.\\0{,}84.\end{cases}$$

Предельная величина r близка к $r\pm 4E_r$ или к $r\pm 3\sigma_r$

$$r\pm 4E_r=0{,}86\pm 0{,}07=\begin{cases}0{,}93\\0{,}79\end{cases}$$

$$\text{или } r\pm 3\sigma_r=0{,}86\pm 0{,}08=\begin{cases}0{,}94\\0{,}78.\end{cases}$$

Как следует из этих расчетов, на Украине и Северном Кавказе наблюдается очень тесная связь урожая озимой пшеницы с весенними запасами влаги, коэффициент корреляции которой очень высокий: $r=0{,}86$, а его ошибки небольшие. Даже с учетом этих ошибок предельная величина коэффициента корреляции не становится меньше 0,78, что говорит о хорошей связи между указанными величинами.

Перейдем теперь к расчету уравнения линии прямой регрессии по формуле $y-\overline{y}=R(x-\overline{x})$, где R — коэффициент

Пример расчета уравнения зависимости урожая озимой
(корреляция двух переменных величин

№ п/п	x	y	$\Delta x = x - \bar{x}$	$\Delta y = y - \bar{y}$	$\Delta x^2 = (x - \bar{x})^2$
1	2	3	4	5	6
1	100	8	—53	—18,5	2809
2	100	10	—53	—16,5	2809
3	100	15	—53	—11,5	2809
4	100	16	—53	—10,5	2809
5	100	18	—53	—8,5	2809
6	105	14	—48	—12,5	2304
7	105	18	—48	—8,5	2304
8	110	16	—43	—10,5	1849
9	115	14	—38	—12,5	1444
10	115	16	—38	—10,5	1444
11	118	18	—35	—8,5	1225
12	118	19	—35	—7,5	1225
13	120	36	—33	—9,5	1089
14	122	19	—31	—7,5	961
15	128	16	—25	—10,5	625
.					
.					
.					
81	175	32	22	5,5	484
82	175	34	22	7,5	484
83	175	36	22	9,5	484
84	178	32	25	5,5	625
85	178	36	25	9,5	625
86	178	38	25	11,5	625
87	180	32	27	5,5	729
88	180	37	27	10,5	729
89	180	40	27	13,5	729
90	182	30	29	3,5	841
91	182	32	29	5,5	841
92	182	34	29	7,5	841
93	182	34	29	7,5	841
94	190	31	37	4,5	1369
95	190	36	37	9,5	1369
96	190	36	37	9,5	1369
97	190	36	37	9,5	1369
98	195	34	42	7,5	1764
99	195	38	42	11,5	1764
100	200	40	47	13,5	2209
$\Sigma = 100$	15300	2650			60 049

Таблица 7

пшеницы (y) от запасов продуктивной влаги весной (x)
для несгруппированных данных)

$\Delta y^2=(y-\overline{y})^2$	$\Delta x\,\Delta y=$ $=(x-\overline{x})\,(y-\overline{y})$	$\Delta x+\Delta y=$ $=(x-\overline{x})+(y-\overline{y})$	$(\Delta x+\Delta y)^2=$ $=[(x-\overline{x})+(y-\overline{y})]^2$
7	8	9	10
342,25	980,5	—71,5	5112,25
272,25	874,5	—69,5	4830,25
132,25	609,5	—64,5	4160,25
110,25	556,5	—63,5	4032,25
72,25	450,5	—61,5	3782,25
156,25	600,0	—60,5	3660,25
72,25	408,0	—56,5	3192,25
110,25	451,5	—53,5	2862,25
156,25	475,0	—50,5	2550,25
110,25	399,0	—48,5	2352,25
72,25	297,5	—43,5	1892,25
56,25	262,5	—42,5	1806,25
90,25	313,5	—23,5	552,25
56,25	232,5	—38,5	1482,25
110,5	262,5	—35,5	1260,25
30,25	121,0	27,5	756,25
56,25	165,0	29,5	870,25
90,25	209,0	31,5	992,25
30,25	137,5	30,5	930,25
90,25	237,5	34,5	1190,25
132,25	287,5	36,5	1332,25
30,25	148,5	32,5	1056,25
110,25	283,5	37,5	1406,25
182,25	364,5	40,5	1640,25
12,25	101,5	32,5	1056,25
30,25	159,5	34,5	1190,25
56,25	217,5	36,5	1332,25
56,25	217,5	36,5	1332,25
20,25	166,5	41,5	1722,25
90,25	351,5	46,5	2162,25
90,25	351,5	46,5	2162,25
90,25	351,5	46,5	2162,25
56,25	315,0	49,5	2450,25
132,25	483,0	53,5	2862,25
182,25	634,5	60,5	3660,25
4777	14 705		94 236

уравнения регрессии,

$$R_{\frac{y}{x}} = r\frac{\sigma_y}{\sigma_x};$$

σ_y и σ_x — средние квадратические отклонения,

$$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{60\,049}{100}} = 24{,}5;$$

$$\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \sqrt{\frac{4777}{100}} = 6{,}9.$$

Откуда

$$R = r\frac{\sigma_y}{\sigma_x} = 0{,}86\ \frac{6{,}9}{24{,}5} = 0{,}24.$$

Подставляя значения $\bar{x}$, $\bar{y}$ и R в уравнение прямой линии, получаем

$$y - 26{,}5 = 0{,}24\,(x - 153);\ y = 0{,}24\,x - 36{,}72 + 26{,}50,$$

отсюда получаем уравнение прямой линии окончательного вида, характеризующее найденную нами зависимость:

$$y = 0{,}24\,x - 10{,}22,$$

где y — урожай озимой пшеницы (в *ц/га*); x — запасы продуктивной влаги (в *мм*) в метровом слое почвы весной при переходе средней декадной температуры воздуха через +5°.

Определим среднюю ошибку найденного уравнения регрессии:

$$s_y = \pm\,\sigma_y\sqrt{1-r^2} = \pm\,6{,}9\sqrt{1-(0{,}86)^2} = \pm\,3{,}5\ \textit{ц/га}.$$

Таким образом, по найденному уравнению мы можем по весенним запасам влаги примерно с трехмесячной заблаговременностью, независимо от будущей погоды, определять виды на урожай озимой пшеницы с ошибкой ±3,5 *ц/га*.

При нахождении уравнений корреляционных связей следует указывать пределы их действия.

Найденное нами уравнение, как было указано выше, действует в пределах значений весенних запасов влаги от 100 до 200 *мм*.

По указанному уравнению, задавая различные значения x, находим значения y и строим теоретическую линию регрессии y по x (рис. 5). Например, задаем значение $x=100$ *мм*, тогда $y=13{,}8$ *ц/га*. Отмечаем эту точку на графике. При $x=160$ *мм*, $y=28{,}2$ *ц/га* получаем вторую точку на графике. При $x=200$ *мм*, $y=37{,}8$ *ц/га* получаем третью точку на графике.

Через указанные три точки проводим прямую линию. Это и будет искомая теоретическая линия регрессии, уравнение которой $y=0{,}24x-10{,}22$.

При нахождении теоретической линии регрессии достаточно задать только два значения x и, рассчитав два значения y, получить на графике две точки. Как известно, через две точки можно провести только одну прямую. Поэтому, имея две точки, мы также правильно проведем искомую теоретическую линию прямой регрессии, уравнение которой мы нашли.

ГЛАВА III МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ ТРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН

§ 1. ЧАСТНЫЕ И ОБЩИЙ КОЭФФИЦИЕНТЫ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ. УРАВНЕНИЕ СВЯЗИ ТРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН

При исследованиях связей между различными явлениями часто можно встретиться с тем, что на одну переменную величину оказывают влияние сразу несколько переменных величин. В этой главе рассмотрим вопрос о связи трех переменных величин, т. е. когда одна зависимая переменная величина z зависит главным образом от двух других величин, независимых переменных x и y. Наиболее простой формой такой связи является линейная.

В случае зависимости между тремя величинами, уравнение линейной множественной корреляции будет иметь вид

$$z = ax + by + c, \tag{43}$$

где a, b, c — постоянные величины, параметры данного уравнения, которые необходимо определить.

Параметры уравнения можно определить двумя путями: 1) через коэффициенты корреляции данных переменных величин, 2) методом наименьших квадратов.

Рассмотрим первый способ расчета параметров уравнения через коэффициенты корреляции.

Линейное уравнение связи между тремя переменными величинами можно представить также в следующем виде:

$$z - \bar{z} = a(x - \bar{x}) + b(y - \bar{y}), \tag{44}$$

где $\bar{z}$, $\bar{x}$ и $\bar{y}$ — средние арифметические значения величин z, x и y.

После определения вида уравнения встает задача определения тесноты связи между тремя переменными величинами, т. е. определения общего коэффициента корреляции R.

Общий коэффициент корреляции вычисляется по формуле

$$R = \sqrt{\frac{r^2_{zx} + r^2_{zy} - 2\, r_{zx}\, r_{zy}\, r_{xy}}{1 - r^2_{xy}}}, \tag{45}$$

где r_{zx}, r_{zy}, r_{xy} — есть частные коэффициенты парной корреляции.

Общий коэффициент корреляции (R) обладает следующими свойствами:

1) Значение R всегда положительно и изменяется от 0 до 1:

$$0 \leqslant R \leqslant 1,$$

2) если R равно 0, то z не может быть линейно связано с x и y. Однако при этом возможна нелинейная корреляционная и даже функциональная связь z с x и y,

3) если R равно единице, то z связано с x и y линейной функциональной связью,

4) если R отлично от своих крайних значений (0 и 1), то при приближении R к единице теснота линейной связи z с x и y увеличивается.

Для того чтобы выделить влияние на полученный результат каждого фактора в отдельности и для определения общего коэффициента корреляции (R) вычисляются частные коэффициенты корреляции по следующим формулам:

$$1)\ r_{zx} = \frac{\Sigma \Delta x \Delta z}{\sqrt{\Sigma \Delta x^2 \Sigma \Delta z^2}};\quad 2)\ r_{zy} = \frac{\Sigma \Delta y \Delta z}{\sqrt{\Sigma \Delta y^2 \Sigma \Delta z^2}};$$

$$3)\ r_{xy} = \frac{\Sigma \Delta x \Delta y}{\sqrt{\Sigma \Delta x^2 \Sigma \Delta y^2}},$$

где $\Delta x = x_i - \bar{x};\ \Delta y = y_i - \bar{y};\ \Delta z = z_i - \bar{z};$

Частные коэффициенты множественной корреляции совершенно аналогичны простым коэффициентам корреляции двух переменных величин r и имеют те же свойства. Каждый из них изменяется от -1 до $+1$ ($-1 \leqslant r \leqslant 1$). Когда $r=0$, линейная связь двух величин исключена. Когда $r = \pm 1$, имеет место функциональная связь между двумя величинами.

Каждый из указанных частных коэффициентов корреляции, при наличии связи трех переменных величин, определяет тесноту линейной связи между двумя величинами, когда третья величина остается (условно) постоянной. Частные коэффициенты корреляции r могут быть определены и по другим формулам, указанным в гл. II, § 4.

Определив частные коэффициенты корреляции r_{zx}, r_{zy}, r_{xy}, общий коэффициент корреляции R и убедившись в достаточно надежной тесноте связи между исследуемыми величинами, переходим к определению параметров a, b и c уравнения ли-

нейной регрессии $z=ax+by+c$. Формулы для вычисления параметров a и b имеют следующий вид:

$$a=\frac{\sigma_z}{\sigma_x}\frac{r_{zx}-r_{zy}r_{xy}}{1-r^2_{xy}}; \tag{46}$$

$$b=\frac{\sigma_z}{\sigma_y}\frac{r_{zy}-r_{zx}r_{xy}}{1-r^2_{xy}}; \tag{47}$$

где σ_z, σ_x и σ_y — средние квадратические отклонения для рядов z, x и y.

$$\sigma_z=\sqrt{\frac{\Sigma\Delta z^2}{n}};\quad \sigma_x=\sqrt{\frac{\Sigma\Delta x^2}{n}};\quad \sigma_y=\sqrt{\frac{\Sigma\Delta y^2}{n}}$$

(здесь n — общее число наблюдений).

Подставляя значения параметров a и b в уравнение $z-\bar{z}=a(x-\bar{x})+b(y-\bar{y})$, мы получим общий вид уравнения регрессии трех переменных величин:

$$z-\bar{z}=\frac{\sigma_z}{\sigma_x}\frac{r_{zx}-r_{zy}r_{xy}}{1-r^2_{xy}}(x-\bar{x})+\frac{\sigma_z}{\sigma_y}\frac{r_{zy}-r_{zx}r_{xy}}{1-r^2_{xy}}(y-\bar{y}). \tag{48}$$

Решая уравнение (48), мы получим окончательное уравнение множественной регрессии $z=ax+by+c$, которое будет характеризовать найденную нами связь между тремя переменными величинами.

Средняя квадратическая ошибка уравнения регрессии трех переменных величин вычисляется по формуле

$$s_z=\pm\sigma_z\sqrt{\frac{1-r^2_{zx}-r^2_{zy}-r^2_{xy}+2r_{zx}r_{zy}r_{xy}}{1-r^2_{xy}}}. \tag{49}$$

Средняя квадратическая ошибка общего коэффициента множественной корреляции вычисляется по формуле

$$\sigma_R=\pm\frac{1-R^2}{\sqrt{n}}. \tag{50}$$

Величины, необходимые для определения общего коэффициента множественной корреляции R и уравнения регрессии трех переменных величин рассчитываются по табл. 8.

§ 2. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЛИНЕЙНОЙ СВЯЗИ ТРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН (зависимость запасов влаги в почве от осадков в разные периоды)

Для характеристики условий роста и развития озимых культур осенью в Западной Сибири, нами была установлена

Таблица 8

№ п/п	z	x	y	Δz	Δx	Δy	Δz^2	Δx^2	Δy^2	$\Delta z\,\Delta x$	$\Delta z\,\Delta y$	$\Delta x\,\Delta y$	$\Delta z+\Delta x+\Delta y$	$(\Delta z+\Delta x+\Delta y)^2$
1														
2														
3														
.														
.														
.														
n	Σ	Σ	Σ				Σ	Σ	Σ	Σ	Σ	Σ		Σ

зависимость запасов продуктивной влаги пахотного слоя почвы от осадков текущего и предшествующего месяца. Рассмотрим в качестве примера линейной связи трех переменных величин эту зависимость запасов влаги пахотного слоя почвы в августе, когда проходит сев озимых культур в Западной Сибири, от осадков августа и июля.

Нахождение данных связей было небходимо для того, чтобы имея большой ряд лет наблюдений над осадками и небольшие ряды наблюдений над влажностью почвы, можно было по осадкам рассчитывать запасы влаги в почве и давать агроклиматическую оценку условий увлажнения почвы в период сева озимых.

Таким образом, нам необходимо было найти уравнение указанной связи трех переменных величин вида

$$z = ax + by + c,$$

z — средние за месяц запасы продуктивной влаги (в *мм*) в августе в слое почвы 0—20 *см*; x — сумма осадков (в *мм*) в августе; y — сумма осадков (в *мм*) в июле.

Для нахождения неизвестных параметров уравнения a, b, c и коэффициента множественной корреляции R, указывающего на тесноту связи, данные наблюдений необходимо расположить в табл. 9 и провести расчеты остальных граф.

Две последние графы (14 и 15-я) в табл. 9 рассчитываются для контроля.

После анализа материала наблюдений и установления его пригодности для обработки, записываем данные по запасам

Пример расчета уравнения зависимости средних запасов продуктивной (множественная корреляция между

№ п/п	z	x	y	Δz	Δx	Δy	Δz^2
1	2	3	4	5	6	7	8
1	5	5	30	—21	—43	—27	441
2	5	15	15	—21	—33	—42	441
3	7	20	15	—19	—28	—42	361
4	8	20	40	—18	—28	—17	324
5	8	25	20	—18	—23	—37	324
6	10	35	30	—16	—13	—27	256
7	10	20	25	—16	—28	—32	256
8	11	30	25	—15	—18	—32	225
9	12	25	20	—14	—23	—37	196
10	12	34	5	—14	—14	—52	196
.							
.							
.							
43	40	68	70	14	20	13	196
44	40	70	120	14	22	63	196
45	40	80	68	14	32	11	196
46	44	85	70	18	37	13	324
47	45	65	85	19	17	28	361
48	45	80	65	19	32	8	361
49	50	90	80	24	42	23	576
50	50	100	90	24	52	33	576
51	50	80	90	24	32	33	576
52	55	100	110	29	52	53	841
Σ 52	1352	2496	2964	—	—	—	8790

Таблица 9

влаги в слое почвы 0—20 *см* за август от осадков июля и августа тремя переменными величинами)

Δx^2	Δy^2	$\Delta z\ \Delta x$	$\Delta z\ \Delta y$	$\Delta x\ \Delta y$	$\Delta z+\Delta x+\Delta y$	$(\Delta z+\Delta x+\Delta y)^2$
9	10	11	12	13	14	15
1849	729	903	567	1161	—91	8281
1089	1764	693	882	1386	—96	9216
784	1764	532	798	1176	—89	7921
784	289	504	306	476	—63	3969
529	1369	414	666	851	—78	6084
169	729	208	432	351	—56	3136
784	1024	448	512	896	—76	5776
324	1024	270	480	576	—65	4225
529	1369	322	518	851	—74	5476
196	2704	196	728	728	—80	6400
400	169	280	182	260	47	2209
484	3969	308	882	1386	99	9801
1024	121	448	154	352	57	3249
1369	169	666	234	481	68	4624
289	784	323	532	476	64	4096
1024	64	608	152	256	59	3481
1764	529	1008	552	966	89	7921
2704	1089	1248	792	1716	109	11 881
1024	1089	768	792	1056	89	7921
2704	2809	1508	1537	2756	134	17 956
30 298	51 274	13 695	15 208	20 023	—	188 214

влаги (z), осадкам августа (x) и осадкам июля (y) в табл. 9 и находим их средние величины:

$$\bar{z} = \frac{\Sigma z}{n} = \frac{1352}{52} = 26,$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{2496}{52} = 48,$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{2964}{52} = 57.$$

После этого находим разности (Δ) значений каждой величины z_i и $\bar{z}$; x_i и $\bar{x}$; y_i и $\bar{y}$:

$$\Delta z = z_i - \bar{z}; \ \Delta x = x_i - \bar{x}; \ \Delta y = y_i - \bar{y}.$$

Найдя разности указанных значений, их квадраты, их произведения и суммы этих значений, делаем контроль наших расчетов по формуле

$$\Sigma\Delta z^2 + \Sigma\Delta x^2 + \Sigma\Delta y^2 + 2\Sigma\Delta z\Delta x + 2\Sigma\Delta z\Delta y + 2\Sigma\Delta x\Delta y =$$

$$= \Sigma(\Delta x + \Delta y + \Delta z)^2;$$

$$\Sigma(\Delta x + \Delta y + \Delta z)^2 = 188\,214$$

$$8790 + 30\,298 + 51\,274 + 2 \cdot 13\,695 + 2 \cdot 15\,208 + 2 \cdot 20\,023 = 188\,214.$$

Контроль показал правильность наших расчетов, поэтому можно перейти к нахождению частных коэффициентов корреляции между коррелируемыми величинами r_{zx}; r_{zy}; r_{xy}:

$$r_{zx} = \frac{\Sigma\Delta x\Delta z}{\sqrt{\Sigma\Delta x^2 \Sigma\Delta z^2}} = \frac{13\,695}{\sqrt{30\,298 \cdot 8790}} = 0{,}84,$$

$$r_{zy} = \frac{\Sigma\Delta y\Delta z}{\sqrt{\Sigma\Delta y^2 \Sigma\Delta z^2}} = \frac{15\,208}{\sqrt{51\,274 \cdot 8790}} = 0{,}71,$$

$$r_{xy} = \frac{\Sigma\Delta x\Delta y}{\sqrt{\Sigma\Delta x^2 \Sigma\Delta y^2}} = \frac{20\,023}{\sqrt{30\,298 \cdot 51\,274}} = 0{,}51.$$

После нахождения частных коэффициентов корреляции r находим общий коэффициент множественной корреляции R и

его вероятную ошибку E_R :

$$R = \sqrt{\frac{r^2_{zx} + r^2_{zy} - 2\, r_{zx}\, r_{zy}\, r_{xy}}{1 - r^2_{xy}}} =$$

$$= \sqrt{\frac{(0,84)^2 + (0,71)^2 - 2\,(0,84 \cdot 0,71 \cdot 0,51)}{1 - (0,51)^2}} = \sqrt{\frac{0,60}{0,74}} = 0,90;$$

$$E_R = \pm\, 0,67\, \frac{1 - R^2}{\sqrt{n}} = \pm\, 0,67\, \frac{1 - (0,90)^2}{\sqrt{52}} = \pm\, 0,02.$$

Следовательно, R вероятно в пределах

$$R \pm E_R = 0,90 \pm 0,02 = \begin{cases} 0,92 \\ 0,88. \end{cases}$$

Предельные значения коэффициента корреляции

$$R \pm 4\, E_R = 0,90 \pm 0,08 = \begin{cases} 0,98 \\ 0,82. \end{cases}$$

Как видим, значения R получились очень высокие, связь запасов влаги в слое почвы 0—20 *см* с осадками августа и июля очень тесная.

Переходим к нахождению уравнения регрессии этой связи. Рассчитываем средние квадратические отклонения:

$$\sigma_z = \sqrt{\frac{\Sigma \Delta z^2}{n}} = \sqrt{\frac{8790}{52}} = 13,0; \quad \sigma_x = \sqrt{\frac{\Sigma \Delta x^2}{n}} =$$

$$= \sqrt{\frac{30\,298}{52}} = 24,14;$$

$$\sigma_y = \sqrt{\frac{\Sigma \Delta y^2}{n}} = \sqrt{\frac{51\,274}{52}} = 31,4.$$

Подставляя в уравнение найденные значения $\bar{z}$, $\bar{x}$, $\bar{y}$, σ_z, σ_x, σ_y, r_{zx}, r_{zy}, r_{xy}, находим искомые параметры a, b и c.

$$z - \bar{z} = \frac{\sigma_z}{\sigma_x}\, \frac{r_{zx} - r_{zy}\, r_{xy}}{1 - r^2_{xy}}\, (x - \bar{x}) + \frac{\sigma_z}{\sigma_y}\, \frac{r_{zy} - r_{zx}\, r_{xy}}{1 - r^2_{xy}}\, (y - \bar{y}),$$

$$z - 26,0 = \frac{13,0}{24,14}\, \frac{0,84 - 0,71 \cdot 0,51}{0,74}\, (x - 48) +$$

$$+ \frac{13,0}{31,4}\, \frac{0,71 - 0,84 \cdot 0,51}{0,74}\, (y - 57),$$

$$z = 26,0 + 0,54 \cdot 0,65\,(x - 48) + 0,41 \cdot 0,38\,(y - 57),$$

$$z = 0,35\, x + 0,16\, y + 0,1. \qquad (51)$$

Таким образом, уравнение зависимости запасов влаги от осадков августа и июля имеет вид $z = 0{,}35x + 0{,}16y + 0{,}1$, где z — средние за август запасы продуктивной влаги (в *мм*) в слое почвы 0—20 *см*, x — сумма осадков (в *мм*) за август, y — сумма осадков (в *мм*) за июль.

Находим среднюю квадратическую ошибку данного уравнения регрессии

$$s_z = \pm \sigma_z \sqrt{1 - R^2} = 13{,}0 \sqrt{1 - (0{,}90)^2} = \pm 5{,}7 \text{ мм}.$$

Для удобства расчетов по найденному уравнению построим график, где по оси абсцисс (x) отложим суммы осадков августа, а по оси ординат (y) — суммы осадков июля (рис. 6).

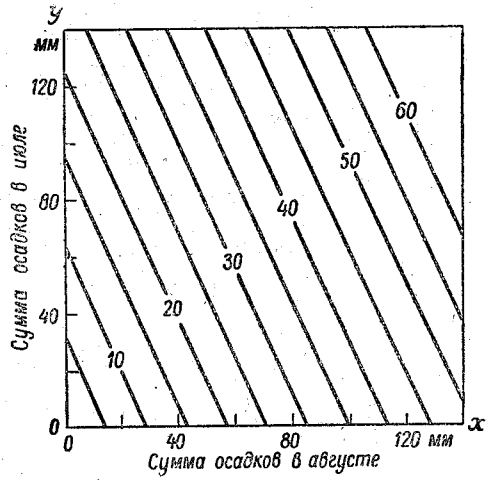

Рис. 6. Зависимость запасов продуктивной влаги слоя почвы 0—20 *см* в августе от осадков августа и июля в Западной Сибири.

Задавая различные значения x и y, получим в поле графика значения запасов влаги (z). Эти значения будут самые различные, по ним трудно будет провести линии равных значений (z) через определенные интервалы, которые мы хотим получить на графике. Поэтому график лучше строить следующим образом.

Допустим, мы хотим получить на графике значения z через интервал 5 *мм*. Задаем значения z через каждые 5 *мм* и решаем уравнение сначала относительно x при $y=0$; а затем при тех же значениях z решаем уравнение относительно y при $x=0$. Например, нам необходимо на графике получить линию $z=5$ *мм*. Решаем уравнение для данного $z=5$ *мм* при $y=0$,

в отношении x, получаем $5=0{,}35x+0{,}1$; откуда $x=14{,}0$. Откладывая 14 на оси абсцисс, получаем точку с координатами $z=5$, $y=0$, $x=14$.

Теперь решаем уравнение в отношении y для этого же $z=5$ *мм*, но при $x=0$, получаем $5=0{,}16y+0{,}1$, откуда $y=31$. Откладывая 31 на оси y, получаем точку с координатами $z=5$, $x=0$, $y=31$. Таким образом, мы получили две точки на осях, где $z=5$ *мм*. Через две точки, как известно, можно провести только одну прямую. Проводим эту прямую для $z=5$ *мм*.

Если мы хотим построить график для значений z через интервал 5 *мм*, то дальнейшие расчеты проводим точно таким же образом для $z=10$, $z=15$, $z=20$ *мм* и т. д.

Получив точку на оси x при $z=10$, $y=0$, $x=28$ и точку на оси y при $z=10$, $x=0$, $y=62$, проводим через две эти точки линию равных значений $z=10$ *мм* и так далее для $z=15$ *мм*, $z=20$ *мм* и т. д.

Таким образом, мы получаем графическое изображение нашей зависимости, по которому легко производить расчеты запасов влаги в зависимости от осадков (рис. 6).

Расчеты уравнения связи трех переменных величин можно проводить также методом сгруппированных данных с введением условных единиц. Тогда для подсчета каждого частного коэффициента корреляции r_{zx}, r_{zy}, r_{xy} составляется отдельная корреляционная таблица, подобно табл. 4 (§ 8 гл. II). Таким образом, составляется три таких таблицы, по данным которых находят r_{zx}, r_{zy}, r_{xy}, $\bar{z}$, $\bar{x}$, $\bar{y}$, σ_z, σ_x и σ_y точно таким способом, как было подробно изложено в § 8 гл. II. Затем подставляя эти величины в формулы для R и для z, приведенные в настоящем разделе, мы получаем множественный коэффициент корреляции R и уравнение регрессии трех переменных величин.

ГЛАВА IV МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ ЧЕТЫРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН

§ 1. УРАВНЕНИЕ СВЯЗИ ЧЕТЫРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН. ЧАСТНЫЕ И ОБЩИЙ КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ

Часто бывает, что связь между двумя или тремя величинами недостаточно тесная и необходимо учитывать еще ряд факторов. Тогда ищут связь между четырьмя величинами или точнее ищут зависимость одной переменной величины от трех других переменных величин. Уравнение этой зависимости будет иметь вид

$$u = ax + by + cz + d. \qquad (52)$$

Для удобства расчетов параметров уравнения a, b, c, d и коэффициента множественной корреляции R рассчитывается следующая табл. 10, где $\Delta u = u_i - \bar{u}$; $\Delta x = x_i - \bar{x}$; $\Delta y = y_i - \bar{y}$; $\Delta z = z_i - \bar{z}$, а $\bar{u}$, $\bar{x}$, $\bar{y}$ и $\bar{z}$ — средние арифметические величины.

При корреляции четырех переменных величин необходимо найти шесть частных коэффициентов корреляции, которые вычисляются по способу, изложенному для двух переменных величин:

Таблица 10

№ п/п	u	x	y	z	Δu	Δx	Δy	Δz	Δu^2	Δx^2	Δy^2	Δz^2	$\Delta u\ \Delta x$	$\Delta u\ \Delta y$	$\Delta u\ \Delta z$	$\Delta x\ \Delta y$	$\Delta x\ \Delta z$	$\Delta y\ \Delta z$	$(\Delta x+\Delta y+\Delta z+\Delta u)$	$(\Delta x+\Delta y+\Delta z+\Delta u)^2$
n	Σ	Σ	Σ	Σ					Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ		Σ

$$1)\ r_{ux}=\frac{\Sigma\Delta u\,\Delta x}{\sqrt{\Sigma\Delta u^2\,\Sigma\Delta x^2}};\quad 2)\ r_{uy}=\frac{\Sigma\Delta u\,\Delta y}{\sqrt{\Sigma\Delta u^2\,\Sigma\Delta y^2}};$$

$$3)\ r_{uz}=\frac{\Sigma\Delta u\,\Delta z}{\sqrt{\Sigma\Delta u^2\,\Sigma\Delta z^2}};\quad 4)\ r_{xy}=\frac{\Sigma\Delta x\,\Delta y}{\sqrt{\Sigma\Delta x^2\,\Sigma\Delta y^2}};$$

$$5)\ r_{xz}=\frac{\Sigma\Delta x\,\Delta z}{\sqrt{\Sigma\Delta x^2\,\Sigma\Delta z^2}};\quad 6)\ r_{yz}=\frac{\Sigma\Delta y\,\Delta z}{\sqrt{\Sigma\Delta y^2\,\Sigma\Delta z^2}}.$$

После этого находят общий коэффициент множественной корреляции четырех величин R:

$$R=\sqrt{1-\frac{D}{D_0}}, \tag{53}$$

$$D=1-r^2_{ux}-r^3_{uy}-r^3_{uz}-r^2_{xy}-r^2_{xz}-r^2_{yz}+$$
$$+r^2_{ux}r^2_{yz}+r^2_{xy}\,r^2_{uz}+r^2_{uy}\,r^2_{xz}+2\,(r_{xy}\,r_{xz}\,r_{yz}+$$
$$+r_{yz}r_{uz}r_{uy}+r_{ux}r_{uz}\,r_{xz}+r_{ux}r_{uy}r_{xy})-$$
$$-2\,(r_{ux}r_{uz}r_{xy}r_{yz}+r_{uz}r_{uy}r_{xy}r_{xz}+r_{ux}r_{uy}r_{xz}r_{yz}) \tag{54}$$

$$D_0=1-r^2_{xy}-r^2_{yz}-r_{xz}{}^2+2\,r_{xy}\,r_{xz}\,r_{yz}. \tag{55}$$

Уравнение регрессии четырех переменных величин, выраженное через r и σ имеет общий вид

$$u-\bar{u}=\frac{\sigma_u}{\sigma_x}\left[\frac{r_{ux}(1-r^2_{yz})-r_{uy}r_{xy}-r_{uz}r_{xz}+r_{yz}(r_{uy}r_{xz}+r_{uz}r_{xy})}{1-r^2_{yz}-r^2_{xy}-r^2_{xz}+2\,r_{yz}r_{xy}r_{xz}}\right]\cdot$$

$$\cdot(x-\bar{x})+\frac{\sigma_u}{\sigma_y}\left[\frac{r_{uy}(1-r^2_{xz})-r_{ux}r_{xy}-r_{uz}r_{yz}+r_{xz}(r_{ux}r_{yz}+r_{uz}r_{xy})}{1-r^2_{yz}-r^2_{xy}-r^2_{xz}+2\,r_{yz}r_{xy}r_{xz}}\right]\cdot$$

$$\cdot(y-\bar{y})+\frac{\sigma_u}{\sigma_z}\left[\frac{r_{uz}(1-r^2_{xy})-r_{ux}r_{xz}-r_{uy}r_{yz}+r_{xy}(r_{ux}r_{yz}+r_{uy}r_{xz})}{1-r^2_{yz}-r^2_{xy}-r^2_{xz}+2\,r_{yz}r_{xy}r_{xz}}\right]\cdot$$

$$\cdot(z-\bar{z}) \tag{56}$$

Определив средние квадратические отклонения $\sigma_u=\sqrt{\frac{\Sigma(u-\bar{u})^2}{n}}$;

$$\sigma_x=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}};\quad \sigma_y=\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}};\quad \sigma_z=\sqrt{\frac{\Sigma(z-\bar{z})^2}{n}}$$

и решив общее уравнение в отношении u, получим уравнение окончательного вида для связи четырех переменных $u=ax+by+cz+d$.

Средняя квадратическая ошибка уравнения регрессии рассчитывается по формуле

$$s_u = \pm \sigma_u \sqrt{1 - R^2}. \qquad (57)$$

Как видим, расчеты уравнения связи четырех переменных не сложны, но очень громоздки. При введении еще большего числа переменных расчеты становятся еще более громоздкими. В этом случае подсчитывают частные коэффициенты корреляции искомой величины с каждой из величин, связь с которыми мы ищем. Частные коэффициенты корреляции покажут, какие величины следует учесть с более высокими коэффициентами корреляции, а какие можно не брать в расчет.

§ 2. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЛИНЕЙНОЙ КОРРЕЛЯЦИИ ЧЕТЫРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН (зависимость запасов влаги в почве от осадков, температуры и исходных запасов влаги)

В агрометеорологии, как известно, прогнозирование запасов продуктивной влаги под различными культурами проводится на основе расчетов по уравнениям с четырьмя переменными. Это уравнения зависимости запасов влаги к концу декады от суммы осадков и средней температуры за декаду и исходных запасов влаги к началу декады.

Нами были найдены уравнения зависимости величины запасов влаги в почве под кукурузой к концу декады от суммы осадков, средней температуры воздуха за декаду и исходных запасов влаги к началу декады. Эти зависимости были найдены для метрового и пахотного слоя почвы по различным отрезкам вегетационного периода кукурузы.

Приведем пример нахождения указанного уравнения зависимости величины запасов влаги (в *мм*) в метровом слое почвы к концу декады (u) от исходных запасов влаги к началу декады (x), от суммы осадков за декаду (y) и от средней температуры за декаду (z) в период от выбрасывания султана до молочной спелости кукурузы (табл. 11).

После анализа большого материала наблюдений данные заносим в табл. 11 и рассчитываем средние величины $\bar{u}$, $\bar{x}$, $\bar{y}$ и $\bar{z}$.

$$\bar{u} = \frac{\Sigma u}{n} = \frac{14\,848}{232} = 64; \quad \bar{x} = \frac{\Sigma x}{n} = \frac{19\,024}{232} = 82;$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{3944}{232} = 17; \quad z = \frac{\Sigma z}{n} \frac{5034}{232} = 21{,}7$$

Затем находим разности значений каждой величины со средними ($\Delta u = u_i - \bar{u}$; $\Delta x = x_i - \bar{x}$ и т. д.), квадраты этих разностей, произведения разностей и суммы этих величин.

Таблица 11

Пример расчета уравнения зависимости запасов продуктивной влаги к концу декады (u) от запасов влаги в начале декады (x), суммы осадков (y) и температуры (z) за декаду (множественная корреляция четырех переменных величин)

№ п/п	u	x	y	z	Δu	Δx	Δy	Δz	Δu^2	Δx^2	Δy^2	Δz^2	$\Delta u\,\Delta x$	$\Delta u \Delta y$	$\Delta u \Delta z$	$\Delta x \Delta y$	$\Delta x \Delta z$	$\Delta y \Delta z$
1	109	140	3	21,8	45	58	—14	0,1	2025	3364	196	0,0	2610	—630	4,5	—812	5,8	—1,4
2	131	109	62	23,2	67	27	45	1,5	4489	729	2025	2,2	1809	3015	100,5	1215	40,5	67,5
3	102	131	10	19,7	38	49	—7	—2,0	1444	2401	49	4,0	1862	—266	—76,0	—343	—98,0	14,0
4	101	154	0	20,8	37	72	—17	—0,9	1369	5184	289	0,8	2664	—629	—33,3	—1224	—64,8	15,3
5	129	101	61	18,2	65	19	44	—3,5	4225	361	1936	12,2	1235	2860	—227,5	836	—66,5	—154,0
6	123	129	32	23,0	59	47	15	1,3	3481	2209	225	1,7	2773	885	76,7	705	61,1	19,5
7	100	129	0	26,0	36	47	—17	4,3	1296	2209	289	18,5	1692	—612	—154,8	—799	202,1	—73,1
8	75	100	0	25,3	11	18	—17	3,6	121	324	289	13,0	198	—187	39,6	—306	64,8	—61,2
9	63	75	23	21,8	—1	—7	6	0,1	1	49	36	0,0	7	—6	—0,6	—42	—0,7	0,6
10	55	75	14	20,9	—9	—7	—3	—0,8	81	49	9	0,6	63	27	7,2	21	5,6	2,4
.																		
.																		
.																		
223	81	116	16	21,4	17	34	—1	—0,3	289	1156	1	0,1	578	—17	—5,1	—34	—10,2	0,3
224	118	125	39	21,0	54	43	22	—0,7	2916	1849	484	0,5	2279	1166	—37,8	946	—30,1	—15,4
225	84	118	4	20,9	20	36	—13	—0,8	400	1296	169	0,6	720	—260	—16,0	—468	—28,8	10,4
226	65	92	7	21,1	1	10	—10	—0,6	1	100	100	0,4	10	—10	—0,6	—100	—6,0	6,0
227	93	65	62	18,1	29	—17	45	—3,6	841	289	2025	13,0	—493	1305	104,4	—765	61,2	—162,0
228	67	93	11	19,0	3	11	—6	—2,7	9	121	36	7,3	33	—18	—8,1	—66	—29,7	16,2
229	40	62	8	21,6	—24	—20	—9	—0,1	576	400	81	0,0	480	216	2,4	180	2,0	0,9
230	45	65	7	23,0	—19	—17	—10	1,3	361	289	100	1,7	323	190	24,7	170	—22,1	—13,0
231	12	45	0	22,8	—52	—37	—17	1,1	2704	1369	289	1,2	1924	884	—57,2	629	—40,7	—18,7
232	4	12	45	24,4	—60	—70	28	2,7	3600	4900	784	7,3	4200	—1680	—162,0	—1960	—189,0	75,6
n=232	14848	19024	3944	5034	—	—	—	—	239838	308717	95252	1062	222633	48989	—4648	—10144	—1465	—2255

После расчета всех граф табл. 11 и получения сумм указанных величин по графам, приступаем к нахождению шести частных коэффициентов корреляции между различными величинами:

$$1)\ r_{ux} = \frac{\Sigma \Delta u \Delta x}{\sqrt{\Sigma \Delta u^2 \Sigma \Delta x^2}} = \frac{222633}{\sqrt{239\ 838 \cdot 308\ 717}} = 0{,}82,$$

$$2)\ r_{uy} = \frac{\Sigma \Delta u \Delta y}{\sqrt{\Sigma \Delta u^2\ \Sigma \Delta y^2}} = \frac{48\ 989}{\sqrt{239\ 838 \cdot 95\ 252}} = 0{,}32,$$

$$3)\ r_{uz} = \frac{\Sigma \Delta u \Delta z}{\sqrt{\Sigma \Delta u^2 \Sigma \Delta z^2}} = \frac{-4648}{\sqrt{239\ 838 \cdot 1062}} = -0{,}29,$$

$$4)\ r_{xy} = \frac{\Sigma \Delta x \Delta y}{\sqrt{\Sigma \Delta x^2 \Sigma \Delta y^2}} = \frac{-10\ 144}{\sqrt{308\ 717 \cdot 95\ 252}} = -0{,}06,$$

$$5)\ r_{xz} = \frac{\Sigma \Delta x \Delta z}{\sqrt{\Sigma \Delta\ x^2\ \Sigma \Delta z^2}} = \frac{-1465}{\sqrt{308\ 717 \cdot 1062}} = -0{,}08,$$

$$6)\ r_{yz} = \frac{\Sigma \Delta y\ \Delta z}{\sqrt{\Sigma \Delta y^2\ \Sigma \Delta z^2}} = \frac{-2255}{\sqrt{95\ 252 \cdot 1062}} = -0{,}22.$$

Находим общий коэффициент множественной корреляции $R = \sqrt{1 - \frac{D}{D_0}}$.

Рассчитываем D по формуле (54):

$$D = 1 - 0{,}82^2 - 0{,}32^2 - (-0{,}29)^2 - (-0{,}06)^2 - (-0{,}08)^2 - (0{,}22)^2 + 0{,}82^2 (-0{,}22)^2 + (-0{,}06)^2 (-0{,}29)^2 + 0{,}32^2 (-(0{,}08)^2 + 2[(-0{,}06)(-0{,}08)(-0{,}22) + (-0{,}22)(-0{,}29)\,0{,}32 + 0{,}82(-0{,}29)(-0{,}08) + 0{,}82\,0{,}32(-0{,}06)] - 2[0{,}82(-0{,}29)(-0{,}06)(-0{,}22) + (-0{,}29)\,0{,}32(-0{,}06)(-0{,}08) + 0{,}82 \cdot 0{,}32(-0{,}08)(-0{,}22)] = 0{,}1593;$$

$$D_0 = 1 - r^2_{xy} - r^2_{yz} - r^2_{xz} + 2 r_{xy} r_{xz} r_{yz} = 1 - (-0{,}06)^2 - (-0{,}22)^2 - (-0{,}08)^2 + 2[(-0{,}06)(-0{,}08)(-0{,}22)] = 0{,}9396;$$

откуда

$$R = \sqrt{1 - \frac{D}{D_0}} = \sqrt{1 - \frac{0,1593}{0,9396}} = \sqrt{1 - 0,1695} =$$

$$= \sqrt{0,8305} = 0,91.$$

Вероятная ошибка коэффициента корреляции

$$E_R = \pm 0,67 \frac{1 - R^2}{\sqrt{n}} = \pm 0,67 \frac{1 - 0,91^2}{\sqrt{232}} = \pm 0,007.$$

Предельное значение R равно

$$R \pm 4 E_R = 0,91 \pm 0,028 = \begin{cases} 0,94 \\ 0,88. \end{cases}$$

Связь, судя по коэффициенту корреляции, очень хорошая.

Находим средние квадратические отклонения:

$$\sigma_u = \sqrt{\frac{\Sigma \Delta u^2}{n}} = \sqrt{\frac{239\,838}{232}} = 32,1;$$

$$\sigma_x = \sqrt{\frac{\Sigma \Delta x^2}{n}} = \sqrt{\frac{308\,717}{232}} = 36,5;$$

$$\sigma_y = \sqrt{\frac{\Sigma \Delta y^2}{n}} = \sqrt{\frac{95\,252}{232}} = 20,3; \quad \sigma_z = \sqrt{\frac{\Sigma \Delta z^2}{n}} = \sqrt{\frac{1062}{232}} = 2,1$$

После этого приступаем к расчету уравнения регрессии по формуле (56):

$$u - 64 = \frac{32,1}{36,5} \frac{0,82\,[1 - (-0,22)^2] - [0,32\,(-0,06)] - (-0,29)\,(-0,08) +}{1 - (-0,22)^2 - (-0,06)^2 - (-0,08)^2 +}$$

$$\frac{+ (-0,22)\,[0,32\,(-0,08) + (-0,29)\,(-0,06)]}{+ 2\,(-0,22)\,(-0,06)\,(-0,08)} (x - 82) +$$

$$+ \frac{32,1}{20,3} \frac{0,32\,[1 - (-0,08)^2] - 0,82\,(-0,06) - (-0,29)\,(-0,22) +}{1 - (-0,22)^2 - (-0,06)^2 - (-0,08)^2 + 2(-0,22)(-0,06)(-0,08)}$$

$$\frac{+ (-0,08)\,[0,82\,(-0,22) + (-0,29)\,(-0,06)]}{} (y - 17) +$$

$$+ \frac{32,1}{2,1} \frac{(-0,29)\,[1 - (-0,06)^2] - 0,82\,(-0,08) -}{}$$

$$\frac{- 0,32\,(-0,22) + (-0,06)\,[0,82\,(-0,22) + 0,32\,(-0,08)]}{1 - (-0,22)^2 - (-0,06)^2 - (-0,08)^2 + 2\,(-0,22)\,(-0,06)\,(-0,08)} (z - 21,7) =$$

$$= 0,73\,(x - 82) + 0,54\,(y - 17) - 2,29\,(z - 21,7) \text{ или}$$

$$u - 64 = 0,73\,x - 59,86 + 0,54\,y - 9,18 - 2,29\,z + 49,69$$

Окончательно уравнение искомой зависимости имеет вид

$$u = 0{,}73\,x + 0{,}54\,y - 2{,}29\,z + 44{,}65,$$

где u — запасы влаги (в *мм*) в метровом слое почвы к концу декады под кукурузой в период от выбрасывания султана до молочной спелости; x — исходные запасы влаги (в *мм*) в метровом слое почвы к началу декады; y — сумма осадков за декаду; z — средняя температура воздуха за декаду.

По найденному уравнению можно построить график, по которому значительно быстрее производятся расчеты величины u (рис. 7).

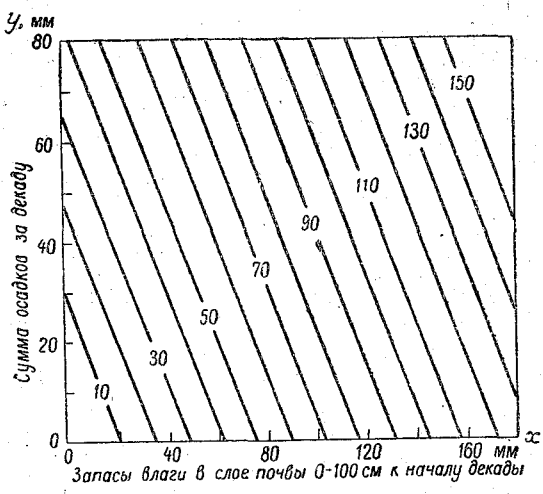

Рис. 7- Зависимость запасов влаги (в *мм*) в метровом слое почвы под кукурузой к концу декады от исходных запасов влаги к началу декады, суммы осадков за декаду и средней температуры воздуха за декаду (для периода от выбрасывания султана до молочной спелости кукурузы).

Средняя температура воздуха за декаду	18	19	20	21	22	23	24	25	26
Поправка на величину запасов влаги (мм)	9	7	5	2	0	-2	-5	-7	-9

Однако на плоскости можно изобразить связь только трех переменных величин, поэтому графическую зависимость строят для трех величин (u, x, y, при постоянном z), а на четвертую величину z рассчитывают по уравнению поправки, которые учитывают при окончательном расчете u.

Обычно график рассчитывается для u, x и y при постоянном z, равном $\bar{z}$, т. е. при среднем арифметическом значении z. В нашем примере $\bar{z}=22{,}0$.

Таким образом, взяв $\bar{z}=22{,}0$, рассчитываем график связи трех переменных величин u, x и y, как это было подробно из-

ложено в гл. III § 2, с той разницей, что берем здесь значения и интервалы запасов влаги для метрового слоя почвы (рис. 7).

Судя по наблюдениям, использованным для расчета уравнения, запасы влаги метрового слоя в этот период под кукурузой могут изменяться в пределах от 0 до 160 *мм*.

Возьмем интервалы для u через 10 *мм* и найдем на графике линии, для которых $u=10$ *мм*; $u=20$ *мм*; $u=30$ *мм* и т. д. По оси абсцисс откладываем исходные запасы влаги к началу декады (x), по оси ординат — сумму осадков за декаду (y), в поле графика будем строить линии равных значений u с интервалом 10 *мм* при $\bar{z}=22{,}0$.

Задавая определенные значения u и находя при $y=0$ и $z=22{,}0$ значения x, а также при том же значении u, но при $x=0$ и $z=22{,}0$ — значения y, мы получим на осях x и y точки, соответствующие данному значению u. Соединив эти точки, получим линию равных значений u. Таким же образом строим линии равных значений для различных u с интервалом 10 *мм*. Получаем график связи u, x и y при $z=22{,}0$. Следовательно, все расчеты запасов влаги по этому графику будут соответствовать уровню температуры воздуха в 22,0°. Для учета различных значений температуры находим поправки. Для $z=22{,}0$ поправка для графика равна 0, так как график был построен при таких значениях температуры.

Проводим по уравнению расчета u для различных значений температуры z через градус, при одинаковых значениях x и y и таким образом получаем разные значения u только в зависимости от z.

Разности u при $\bar{z}=22°$ и различных z дадут нам величины поправок на температуру воздуха, которые сводятся по градациям в табличку под графиком. Алгебраическая сумма величины u, снятой с графика, и величины поправки даст нам окончательную величину рассчитываемых запасов влаги u.

Расчеты частных коэффициентов корреляции r_{ux}, r_{uy}, r_{uz}, r_{xy}, r_{xz} и r_{yz}, а также и средних квадратических отклонений σ_u, σ_x, σ_y и σ_z можно проводить также методом сгруппированных данных по частотам с введением условных единиц. Для нахождения каждого частного r рассчитываем таблицу, подобно табл. 6, подробное изложение расчетов которой дано в гл. II, § 8. Следовательно, в общей сложности рассчитываем шесть таблиц и находим величины шести частных r и σ и средние арифметические величины $\bar{u}$, $\bar{x}$, $\bar{z}$. Затем подставляя их в формулы для R и для u, приведенные в этом разделе, рассчитываем множественный коэффициент корреляции и уравнение регрессии четырех переменных.

ГЛАВА V НАХОЖДЕНИЕ УРАВНЕНИЙ ЛИНЕЙНЫХ СВЯЗЕЙ ПЕРЕМЕННЫХ ВЕЛИЧИН ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ

§ 1. НАХОЖДЕНИЕ УРАВНЕНИЙ ЛИНЕЙНОЙ СВЯЗИ ДВУХ ПЕРЕМЕННЫХ ВЕЛИЧИН ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ

Если найдена линейная зависимость двух переменных величин, общий вид уравнения которой $y=ax+b$, то коэффициенты (параметры) уравнения, кроме определения через r, σ_x и σ_y, могут быть определены по методу наименьших квадратов.

Допустим, что характер расположения точек на корреляционном поле нам показывает прямолинейную связь, теоретическая линия регрессии которой выразится уравнением $y=ax+b$. Параметры a и b, характерные для нашей линии регрессии, нам неизвестны.

Из бесчисленного множества прямых линий, которые можно провести на плоскости по точкам корреляционного поля, нам следует выбрать одну, наилучшим образом соответствующую нашим экспериментальным данным.

При корреляционной связи при одном и том же значении x мы будем иметь несколько значений y. Чтобы прямая регрессии ближе всего подходила к точкам, необходимы наименьшие отклонения ординат различных точек от данной прямой. Но отклонения ординат точек от прямой могут быть положительными и отрицательными, в зависимости от того, где расположены точки — выше или ниже прямой. Возможен и такой случай, когда сумма отклонений $\Sigma(y_i-\bar{y}_x)$ окажется очень малой из-за различия знаков, а точки будут располагаться далеко от прямой. Чтобы избежать этого и исключить влияние знаков отклонений, достаточно искать наименьшее значение не суммы отклонений $\Sigma(y_i-\bar{y}_x)$, а суммы квадратов отклонений $\Sigma(y-\bar{y}_x)^2$.

Таким образом, для отыскания лучшей прямой регрессии данного корреляционного поля необходимо, чтобы

$$\Sigma(y_i-\bar{y}_x)^2=\min, \qquad (58)$$

т. е. сумма квадратов отклонений фактических ординат (y_i) от ординат, вычисленных по уравнению прямой $(\bar{y}_x)$, должна быть наименьшей.

Формула (58) носит название основного условия наименьших квадратов, а метод отыскания параметров уравнений, основанный на этом условии, называется методом наименьших квадратов.

Обозначим сумму квадратов отклонений $\Sigma (y_i - \bar{y}_x)^2$ через f и заменяя $\bar{y}_x$ через $ax + b$ получим

$$f = \sum_{i=1}^{n} (y_i - \bar{y}_x)^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Так как величина f зависит от a и b, то ее можно рассматривать как функцию двух неизвестных параметров a и b.

Для отыскания ее минимума можно применить известный прием дифференциального исчисления, заключающийся в отыскании двух частных производных первого порядка от функции f по a и b, приравнивании их к нулю и определении критических значений a и b из полученных двух уравнений.

На основании этого будем иметь

$$\frac{\partial f}{\partial a} = 0; \quad \frac{\partial f}{\partial b} = 0;$$

$$\left.\begin{aligned} \frac{\partial f}{\partial a} &= 2 \sum_{i=1}^{n} (ax_i + b - y_i) x_i = 0 \\ \frac{\partial f}{\partial b} &= 2 \sum_{i=1}^{n} (ax_i + b - y_i) = 0 \end{aligned}\right\}$$

Раскрывая скобки и разбивая на почленные суммы, а затем вынося общие множители a и b за знаки сумм и заменяя Σb на nb, уравнения можно свести к следующему виду:

$$\left.\begin{aligned} a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i &= 0 \\ a \sum_{i=1}^{n} x_i + nb - \sum_{i=1}^{n} y_i &= 0 \end{aligned}\right\} \qquad (59)$$

Таким образом, мы получили систему двух нормальных уравнений первой степени относительно неизвестных параметров a и b. Их можно записать также в следующем виде:

$$\left.\begin{aligned} a\sum_{i=1}^{n} x_i + nb &= \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i \end{aligned}\right\} \qquad (60)$$

где n — общее число случаев или число точек корреляционного поля. Значения y_i и x_i нам известны из наблюдений. Таким образом, параметры уравнения прямой регрессии a и b можем определить на основании данных уравнений по следующим формулам:

$$a = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \qquad (61)$$

— формула для коэффициента регрессии уравнения;

$$b = \frac{\Sigma y_i \Sigma x^2_i - \Sigma x_i \Sigma x_i y_i}{n\Sigma x^2_i - (\Sigma x_i)^2} \qquad (62)$$

— формула для свободного члена уравнения.

Для вычисления Σx_i; Σy_i; Σx_i^2 и $\Sigma x_i y_i$ составляется табл. 12:

Таблица 12

№ п/п	x_i	y_i	x^2_i	$x_i y_i$
n	Σx_i	Σy_i	Σx_i^2	$\Sigma x_i y_i$

Для определения параметров уравнения a и b в литерату-

ре встречаются также формулы другого вида. Разделив числитель и знаменатель формулы (61) на n^2, получим

$$a = \frac{\frac{\Sigma x_i y_i}{n} - \frac{\Sigma x_i}{n}\frac{\Sigma y_i}{n}}{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2} \text{ или, что то же самое,}$$

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2}, \tag{63}$$

где $\overline{x^2} - \overline{x}^2 = \sigma_x^2$, тогда

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x^2}. \tag{64}$$

Для взвешенных данных, если расчеты ведутся с учетом частот (m_{xy}) мы будем иметь следующие уравнения и формулы для определения a и b:

$$\left.\begin{aligned} a\sum_{i=1}^{k} m_{x_i} x_i^2 + b\sum_{i=1}^{k} m_{x_i} x_i &= \sum_{i=1}^{k} m_{x_i} x_i y_i \\ a\sum_{i=1}^{k} m_{x_i} x_i + bn &= \sum_{i=1}^{k} m_{x_i} y_i \end{aligned}\right\} \tag{65}$$

откуда

$$a = \frac{n\sum_{i=1}^{k} m_{x_i} x_i y_i - \sum_{i=1}^{k} m_{x_i} x_i \sum_{i=1}^{k} m_{x_i} y_i}{n\sum_{i=1}^{k} m_{x_i} x^2{}_i - \left(\sum_{i=1}^{k} m_{x_i} x_i\right)^2} \tag{66}$$

— формула для коэффициента регрессии уравнения $y = ax + b$;

$$b = \frac{\sum_{i=1}^{k} m_{x_i} y_i \sum_{i=1}^{k} m_{x_i} x^2{}_i - \sum_{i=1}^{k} m_{x_i} x_i \sum_{i=1}^{k} m_{x_i} x_i y_i}{n\sum_{i=1}^{k} m_{x_i} x^2{}_i - \left(\sum_{i=1}^{k} m_{x_i} x_i\right)^2} \tag{67}$$

— формула для свободного члена того же уравнения.

Для расчетов параметров уравнений с учетом весов или частот составляется табл. 13.

Определив параметры a и b уравнения $y=ax+b$, строят по этому уравнению теоретическую линию регрессии, задавая различные значения x и получая рассчитанные значения y.

Таблица 13

№ п/п	x_i	y_i	m_{x_i}	$m_{x_i} x_i$	$m_{x_i} x^2{}_i$	$m_{x_i} y_i$	$m_{x_i} x_i y_i$
n	—	—	—	$\Sigma m_{x_i} x_i$	$\Sigma m_{x_i} x^2{}_i$	$\Sigma m_{x_i} y_i$	$\Sigma m_{x_i} x_i y_i$

При составлении уравнений способом наименьших квадратов можно отсчитывать значения x и y от произвольного начала, а затем учесть это изменение отсчета в окончательной формуле. Это уменьшает большие числа и громоздкость расчетов.

Значения x и y можно отсчитывать от средних $\bar{x}$ и $\bar{y}$, если они не дробные числа, тогда расчеты значительно упрощаются.

В этом случае уравнение прямой регрессии будут иметь вид

$$y-\bar{y}=a'(x-\bar{x})+b'. \tag{68}$$

Нормальные уравнения для расчетов будут иметь вид

$$\left.\begin{aligned} a'\Sigma(x-\bar{x})+nb' &= \Sigma(y-\bar{y}) \\ a'\Sigma(x-\bar{x})^2+b'\Sigma(x-\bar{x}) &= \Sigma(x-\bar{x})(y-\bar{y}) \end{aligned}\right\} \tag{69}$$

откуда

$$a'=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2},$$

$$b'=\frac{\Sigma(y-\bar{y})}{a'\Sigma(x-\bar{x})}.$$

§ 2. НАХОЖДЕНИЕ ЛИНЕЙНЫХ УРАВНЕНИЙ СВЯЗИ ТРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ

Линейную зависимость одной переменной величины (z) от двух других переменных величин (x и y) можно выразить уравнением

$$z = ax + by + c,$$

где a, b, c — неизвестные параметры уравнения, которые можно определить по методу наименьших квадратов.

Для этого необходимо решить следующую систему трех нормальных уравнений

$$\left.\begin{aligned} a\Sigma x + b\Sigma y + cn &= \Sigma z \\ a\Sigma xy + b\Sigma y^2 + c\Sigma y &= \Sigma zy \\ a\Sigma x^2 + b\Sigma xy + c\Sigma x &= \Sigma zx \end{aligned}\right\} \qquad (70)$$

где n — общее число случаев наблюдений сочетания трех переменных величин; Σx, Σy и Σz — суммы соответствующих значений каждой из переменных величин.

Таблица 14

№ п/п	z	x	y	zx	zy	xy	y^2	x^2
n	Σz	Σx	Σy	Σzx	Σzy	Σxy	Σy^2	Σx^2

Для решения этих уравнений необходимо провести расчеты по табл. 14.

Получив из табл. 14 необходимые данные, составляем систему уравнений (70) и, решая ее, находим параметры a, b и c уравнения связи $z = ax + by + c$.

§ 3. НАХОЖДЕНИЕ ЛИНЕЙНЫХ УРАВНЕНИЙ СВЯЗИ ЧЕТЫРЕХ ПЕРЕМЕННЫХ ВЕЛИЧИН ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ЗАВИСИМОСТИ УРОЖАЯ ЯРОВОЙ ПШЕНИЦЫ ОТ ОСАДКОВ В РАЗНЫЕ ПЕРИОДЫ И ИСПАРЕНИЯ

Линейное уравнение связи четырех переменных величин выразим формулой

$$y = a_0 + a_1 x + a_2 z + a_3 u,$$

где неизвестные параметры уравнения a_0, a_1, a_2, a_3 находятся путем решения системы четырех нормальных уравнений:

$$\left.\begin{aligned} na_0 + a_1 \Sigma x + a_2 \Sigma z + a_3 \Sigma u &= \Sigma y \\ a_0 \Sigma x + a_1 \Sigma x^2 + a_2 \Sigma xz + a_3 \Sigma xu &= \Sigma xy \\ a_0 \Sigma z + a_1 \Sigma xz + a_2 \Sigma z^2 + a_3 \Sigma zu &= \Sigma yz \\ a_0 \Sigma u + a_1 \Sigma xu + a_2 \Sigma zu + a_3 \Sigma u^2 &= \Sigma yu \end{aligned}\right\} \qquad (71)$$

Для решения указанных уравнений удобно пользоваться табл. 15. Взяв необходимые данные из наблюдений и рассчитав графы табл. 15, составляем систему указанных уравнений (71) и рассчитываем параметры уравнения.

Таблица 15

№ п/п	x	z	u	y	xz	xu	zu	x^2	z^2	u^2	y^2	xy	zy	uy
n	Σx	Σz	Σu	Σy	Σxz	Σxu	Σzu	Σx^2	Σz^2	Σu^2	Σy^2	Σxy	Σzy	Σuy

При увеличении числа переменных увеличивается число членов уравнения связи, а следовательно увеличивается и число нормальных уравнений, решение которых необходимо. Расчеты при большом числе членов (больше четырех) становятся очень громоздкими и трудоемкими, поэтому их следует проводить на электронно-вычислительных машинах.

При нахождении уравнения связи многих переменных следует брать только те величины или факторы, которые оказывают существенное влияние на искомую переменную величину, иначе учет несущественных факторов сильно загромождает расчетную работу при нахождении уравнения связи, но мало приближает к более полному изучению связи.

Приведем пример расчета параметров уравнения связи четырех переменных по методу наименьших квадратов. Пример взят из книги В. С. Немчинова «Сельскохозяйственная статистика с основами общей теории» (Сельхозгиз, М., 1946 г.).

Найдем уравнение зависимости урожая пшеницы (y) от суммы осадков (в *мм*) за период от сева до начала кущения (x), от суммы осадков (в *мм*) от начала кущения до начала

цветения (z) и от испарения (в *мм*) в период от начала кущения до начала цветения пшеницы (u).

Для нахождения параметров уравнения указанной зависимости по методу наименьших квадратов нам необходимо решить систему уравнений, где нужные суммы переменных величин находятся по табл. 16.

Составляем табл. 16 и вычисляем необходимые графы. После вычисления сумм подставляем их величины в систему уравнений (71), решаем эту систему:

$$\left.\begin{aligned} 25\,a_0 + 714\,a_1 + 1454\,a_2 + 4857\,a_3 &= 231{,}0 \\ 714\,a_0 + 32\,116\,a_1 + 46\,766\,a_2 + 121\,212\,a_3 &= 8224{,}1 \\ 1459\,a_0 + 46\,766\,a_1 + 106\,111\,a_2 + 251\,798\,a_3 &= 15\,610{,}0 \\ 4857\,a_0 + 121\,212\,a_1 + 251\,798\,a_2 + 1\,034\,455\,a_3 &= 40\,229{,}7 \end{aligned}\right\}$$

Далее, решая эту систему уравнений, производим следующие действия:

1) Делим все члены уравнений на коэффициенты при a_0

$$\left.\begin{aligned} a_0 + 28{,}56\,a_1 + 58{,}16\,a_2 + 194{,}28\,a_3 &= 9{,}24; \\ a_0 + 44\,980{,}4\,a_1 + 65{,}4986\,a_2 + 169{,}7647\,a_3 &= 11{,}5183; \\ a_0 + 32{,}0535\,a_1 + 72{,}7286\,a_2 + 172{,}5826\,a_3 &= 10{,}6991; \\ a_0 + 24{,}9561\,a_1 + 51{,}8423\,a_2 + 212{,}9822\,a_3 &= 8{,}2828. \end{aligned}\right\}$$

2) Вычитаем из второго уравнения уравнение первое, из второго — третье уравнение и из третьего — четвертое. Получим три уравнения:

$$\left.\begin{aligned} \text{II} - \text{I} \quad 16{,}4204\,a_1 + 7{,}3385\,a_2 - 24{,}5153\,a_3 &= 2{,}2783 \\ \text{II} - \text{III} \quad 12{,}9269\,a_1 + 7{,}2300\,a_2 - 2{,}8179\,a_3 &= 0{,}8192 \\ \text{III} - \text{IV} \quad 7{,}0974\,a_1 + 20{,}8863\,a_2 - 40{,}3996\,a_3 &= 2{,}4163. \end{aligned}\right\}$$

3) Делим все члены полученных уравнений на коэффициенты при a_1:

$$\begin{aligned} a_1 + 0{,}4469\,a_2 - 1{,}4930\,a_3 &= 0{,}1387; \\ a_1 - 0{,}5637\,a_2 - 0{,}2180\,a_3 &= 0{,}0634; \\ a_1 + 2{,}9428\,a_2 - 5{,}6924\,a_3 &= 0{,}340. \end{aligned}$$

4) Вычитаем из первого уравнения второе и из третьего — второе. Получим систему двух уравнений:

$$\begin{aligned} \text{I}-\text{II} \quad & 1{,}0106\,a_2 - 1{,}2750\,a_3 = 0{,}0753; \\ \text{III}-\text{II} \quad & 3{,}5065\,a_2 - 5{,}9104\,a_3 = 0{,}2770. \end{aligned}$$

Таблица расчета сумм величин, необходимых для

Годы	x	z	u	y	xz	xu	zu
1905	21	76	211	10,5	1596	4431	16 036
1906	1	49	268	6,6	49	268	13 132
1907	44	42	163	10,5	1848	7172	6846
1908	25	36	222	6,1	900	5550	7992
1909	37	58	126	11,1	2146	4662	7308
1910	53	58	145	14,7	3074	7685	8410
1911	1	18	310	2,4	18	310	5580
1912	24	89	175	12,1	2136	4200	15 575
1913	27	84	154	10,0	2268	4158	12 936
1914	36	52	171	10,8	1872	6156	8892
1915	88	96	159	17,7	8448	13 992	15 264
1916	14	93	141	8,2	1302	1974	13 113
1917	9	20	211	5,4	180	1899	4220
1918	14	89	124	11,8	1246	1736	11 036
1919	17	109	149	10,2	1853	2533	16 241
1920	32	57	214	9,3	1824	6848	12 198
1921	1	5	364	1,3	5	364	1320
1922	19	49	171	9,6	931	3249	8879
1923	40	55	178	8,7	2200	7120	9790
1924	8	18	309	3,0	144	2172	5562
1925	12	103	167	11,4	1236	2009	17 201
1926	66	74	180	13,3	4884	11 880	13 320
1927	20	18	229	5,2	360	4580	4122
1928	66	71	135	14,1	4686	8910	9585
1929	39	40	181	7,0	1560	7059	7240
$\Sigma n=25$	714	1454	4857	231,0	46 766	121 212	251 798

y — урожай яровой пшеницы (в *ц/га*);
x — осадки от начала сева до начала кущения (в *мм*);
z — осадки от начала кущения до начала цветения (в *мм*);
u — испарение от начала кущения до начала цветения (в *мм*).

Таблица 16

решения системы уравнений связи четырех переменных

x^2	z^2	u^2	y^2	xy	zy	uy
441	5776	44 521	110,25	220,5	798,0	2215,5
1	2401	71 824	43,56	6,6	323,4	1768,8
1936	1764	26 569	110,25	462,0	441,0	1711,5
625	1296	49 284	37,21	152,5	219,6	1354,2
1369	3364	15 876	123,21	410,7	643,8	1398,6
2809	3364	21 025	216,09	779,1	852,6	2131,5
1	324	96 100	5,76	2,4	43,2	744,0
576	7921	30 625	146,41	290,4	1076,9	2117,5
729	7056	23 716	100,00	270,0	840,0	1540,0
1296	2704	29 241	116,64	388,8	561,6	1846,8
7744	9216	25 281	313,29	1557,6	1699,2	2814,3
196	8649	19 881	67,24	114,8	762,6	1156,2
81	400	44 521	29,16	48,6	108,0	1139,4
196	7921	15 376	139,24	165,2	1050,2	1463,2
289	11 881	22 201	104,04	173,4	1111,8	1519,8
1024	3249	45 796	86,49	297,6	530,2	1990,2
1	25	132 496	1,69	1,3	6,5	473,2
361	2401	29 241	92,16	182,4	470,4	1641,6
1600	3025	31 684	75,69	348,0	478,5	1548,6
64	324	95 481	9,00	24,0	54,0	927,0
144	10 609	27 889	129,96	136,8	1174,2	1903,8
4356	5476	32 400	176,89	877,8	984,2	2394,0
400	324	52 441	30,25	110,0	99,0	1259,5
4356	5041	18 225	198,81	930,6	1001,1	1908,5
1521	1600	32 761	49,00	273,0	280,0	1267,0
32 116	106 111	1 034 455	2512,29	8224,1	15 610,0	40 229,7

5) Делим члены уравнений на коэффициенты при a_2 и, вычитая из второго уравнения первое или наоборот, находим параметр a_3:

$$\begin{array}{l} a_2 - 1{,}2616\ a_3 = 0{,}0745 \\ a_2 - 1{,}6855\ a_3 = 0{,}0790 \\ \hline -0{,}4239\ a_3 = 0{,}0045 \\ a_3 = -0{,}0106. \end{array}$$

6) Находим параметр a_2 подстановкой величины a_3:

$$a_2 = 0{,}0745 - 1{,}2616 \cdot 0{,}0106 = 0{,}0611.$$

7) Находим параметр a_1:

$$a_1 + 0{,}0611 \cdot 0{,}4469 + 1{,}4930 \cdot 0{,}0106 = 0{,}1389;$$
$$a_1 = 0{,}0273 + 0{,}0158 = 0{,}1389;$$
$$a_1 = 0{,}1389 - 0{,}0431 = 0{,}0958.$$

8) Находим параметр a_0:

$$a_0 + 28{,}56 \cdot 0{,}0958 + 58{,}16 \cdot 0{,}0611 - 194{,}28 \cdot 0{,}0106 = 9{,}24;$$
$$a_0 = 5{,}0098.$$

Подставляя полученные величины параметров в общий вид уравнения связи, получаем искомое корреляционное уравнение связи четырех переменных величин:

$$\bar{y}_{x,\ z,\ u} = 5{,}0098 + 0{,}0958\ x + 0{,}0611\ z - 0{,}0106\ u.$$

Указанный пример расчета параметров уравнения линейной связи четырех величин включил в себя также расчеты по методу наименьших квадратов связи трех величин (решение системы трех уравнений) и двух величин (решение системы двух уравнений).

ГЛАВА VI КРИВОЛИНЕЙНЫЕ КОРРЕЛЯЦИОННЫЕ СВЯЗИ МЕЖДУ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ

§ 1. НАХОЖДЕНИЕ ПАРАМЕТРОВ УРАВНЕНИЙ ПАРАБОЛИЧЕСКИХ СВЯЗЕЙ МЕЖДУ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ

Прямолинейные зависимости между x и y являются наиболее простой формой связи. Однако часто связи между величинами являются более сложными, представляющие в графическом изображении кривую линию. Криволинейная, четко выраженная зависимость, бывает видна и в корреляционной таблице по расположению частот.

Нередки случаи, когда при исследовании корреляционных связей точки $A_1(x_1 y_1)$, $A_2(x_2 y_2)$ ... $A_n(x_n y_n)$ располагаются на графике вблизи некоторой параболы. Следовательно, эмпирическую формулу необходимо искать следующего вида:

$$y = ax^2 + bx + c. \tag{72}$$

Это уравнение параболы второго порядка. Оно выражает параболическую зависимость, когда имеет место ускоренное возрастание или убывание одной величины (y) при равномерном возрастании другой (x). Парабола второго порядка — кривая с одним возможным максимумом или минимумом.

Метод наименьших квадратов дает возможность найти параметры указанного параболического уравнения a, b, c путем решения системы следующих нормальных уравнений:

$$\left.\begin{aligned} cn + b\Sigma x_i + a\Sigma x_i^2 &= \Sigma y_i \\ c\Sigma x_i + b\Sigma x_i^2 + a\Sigma x_i^3 &= \Sigma x_i y_i \\ c\Sigma x_i^2 + b\Sigma x_i^3 + a\Sigma x_i^4 &= \Sigma x_i^2 y_i \end{aligned}\right\} \tag{73}$$

Это система уравнений для несгруппированных данных простой перечневой таблицы, где для каждого значения x_i указано одно значение y_i (частота $m_{x_i}=1$).

Определив необходимые суммы значений x и y по табл. 17 и решив указанную систему уравнений, мы найдем парамет-

Таблица 17

№ п/п	x_i	y_i	$x^2{}_i$	$x^3{}_i$	$x^4{}_i$	$x_i y_i$	$x^2{}_i\, y_i$
n	Σx_i	Σy_i	$\Sigma x^2{}_i$	$\Sigma x^3{}_i$	$\Sigma x^4{}_i$	$\Sigma x_i y_i$	$\Sigma x_i{}^2\, y_i$

ры уравнения искомой параболы, которая будет ближе всего располагаться около точек A_i.

Согласно принципу наименьших квадратов, искомой параболой будет та, для которой сумма квадратов отклонений по ординате точек A_1, A_2, A_3 ... A_n от точек, лежащих на параболе, при тех же абсциссах будет наименьшей.

В случае сгруппированных данных по частотам m_{xy} система уравнений для нахождения параметров a, b и c имеет следующий вид:

$$\left.\begin{aligned} c\,\Sigma\, m_{x_i} + b\,\Sigma\, m_{x_i} x_i + a\,\Sigma\, m_{x_i} x_i{}^2 &= \Sigma\, m_{x_i} \bar{y}_{x_i} \\ c\,\Sigma\, m_{x_i} x_i + b\,\Sigma\, m_{x_i} x_i{}^2 + a\,\Sigma\, m_{x_i} x_i{}^3 &= \Sigma\, m_{x_i} x_i \bar{y}_{x_i} \\ c\,\Sigma\, m_{x_i} x_i{}^2 + b\,\Sigma\, m_{x_i} x_i{}^3 + a\,\Sigma\, m_{x_i} x_i{}^4 &= \Sigma\, m_{x_i} x_i{}^2 \bar{y}_{x_i} \end{aligned}\right\} \quad (74)$$

Для расчета сумм данной системы уравнений пользуются табл. 18. Вычисления параметров параболического уравнения громоздки, связаны с большими числами, так как переменная x берется в квадрате, в третьей и четвертой степенях. Эти вычисления значительно упрощаются, если провести преобра-

Таблица 18

x_i	m_{x_i}	$m_{x_i} x_i$	$m_{x_i} x_i{}^2$	$m_{x_i} x_i{}^3$	$m_{x_i} x_i{}^4$	$m_{x_i} \bar{y}_{x_i}$	$m_{x_i} \bar{y}_{x_i} x_i$	$m_{x_i} \bar{y}_{x_i} x_i{}^2$
Σx_i	Σm_{x_i}	$\Sigma m_{x_i} x_i$	$\Sigma m_{x_i} x^2{}_i$	$\Sigma m_{x_i} x^3{}_i$	$\Sigma m_{x_i} x^4{}_i$	$\Sigma m_{x_i} \bar{y}_{x_i}$	$\Sigma m_{x_i} \bar{y}_{x_i} x_i$	$\Sigma m_{x_i} \bar{y}_{x_i} x_i{}^2$

зования и перенести начало координат в точку, наиболее близкую к вершине; тогда значения x_i и y_i значительно уменьшаются, а вычисления упрощаются:

$$x' = x - \alpha, \; y' = y - \beta,$$

где α и β — координаты нового начала отсчета.

Делают и другие упрощения. Часто значения x уменьшают в несколько раз (в 10, 20 раз и т. д), что значительно уменьшает величины и громоздкость расчетов.

Кроме параболы второго порядка, при изучении связи между величинами применяются параболы более высоких порядков. Чем выше порядок параболы, тем точнее он воспроизводит опытные данные.

Уравнение параболы третьего порядка имеет вид

$$y = ax^3 + bx^2 + cx + d.$$

Для нахождения параметров уравнения a, b, c, d необходимо решить следующую систему нормальных уравнений:

$$\left.\begin{aligned} nd + c\Sigma x + b\Sigma x^2 + a\Sigma x^3 &= \Sigma y \\ d\Sigma x + c\Sigma x^2 + b\Sigma x^3 + a\Sigma x^4 &= \Sigma yx \\ d\Sigma x^2 + c\Sigma x^3 + b\Sigma x^4 + a\Sigma x^5 &= \Sigma yx^2 \\ d\Sigma x^3 + c\Sigma x^4 + b\Sigma x^5 + a\Sigma x^6 &= \Sigma yx^3 \end{aligned}\right\} \qquad (75)$$

Расчеты необходимых сумм в уравнениях проводятся с помощью табл. 19.

Таблица 19

n_i	x	y	x^2	x^3	x^4	x^5	x^6	yx	yx^2	yx^3
n	Σx	Σy	Σx^2	Σx^3	Σx^4	Σx^5	Σx^6	Σyx	Σyx^2	Σyx^3

При увеличении порядка параболы расчеты параметров уравнений становятся очень сложными и громоздкими.

Если мы имеем уравнение параболы некоторой степени e, то ее уравнение имеет вид $y = a_0 + a_1x + a_2x^2 + \ldots + a_ex^e$.

Параметры данного уравнения a_1, a_2 ... a_e находятся путем решения системы $e+1$ нормальных уравнений при груп-

пировке данных:

$$\left.\begin{aligned} a_0 \Sigma m_{x_i} + a_1 \Sigma m_{x_i} x_i + \ldots + a_e \Sigma m_{x_i} x_i^e &= \Sigma m_{x_i} \bar{y}_{x_i} \\ a_0 \Sigma m_{x_i} x_i + a_1 \Sigma m_{x_i} x_i^2 + \ldots + a_e \Sigma m_{x_i} x_i^{e+1} &= \Sigma m_{x_i} x_i \bar{y}_{x_i} \\ a_0 \Sigma m_{x_i} x_i^e + a_1 \Sigma m_{x_i} x_i^{e+1} + \ldots + a_e \Sigma m_{x_i} x_i^{2e} &= \Sigma m_{x_i} x_i^e \bar{y}_{x_i} \end{aligned}\right\} \quad (76)$$

§ 2. КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ—МЕРА ТЕСНОТЫ СВЯЗИ ДЛЯ КРИВОЛИНЕЙНЫХ ЗАВИСИМОСТЕЙ

Вопрос тесноты связи переменных величин y и x в криволинейных зависимостях не может быть решен нахождением коэффициента корреляции r. Для этого находят корреляционное отношение (η).

Корреляционное отношение η является общим показателем тесноты связи y с x любой формы, в этом его преимущество перед коэффициентом корреляции r. При прямолинейной связи корреляционное отношение равно коэффициенту корреляции:

$$\eta = r.$$

Корреляционное отношение η называют также индексом корреляции.

Корреляционным отношением y по x называется отношение среднего квадратического отклонения условных средних $\bar{y}_x$ относительно общей средней $\bar{y}$ (межгруппового $\sigma(\bar{y}_x)$) к среднему квадратическому отклонению всех значений y относительно $\bar{y}$ (общему σ_y).

Для связи y по x

$$\eta_{y/x} = \frac{\sigma(\bar{y}_x)}{\sigma_y}; \quad (77)$$

$$\sigma(\bar{y}_x) = \sqrt{\frac{\sum^x m_x (\bar{y}_x - \bar{y})^2}{n}}; \quad \sigma_y = \sqrt{\frac{\sum^y m_y (y - \bar{y})^2}{n}}.$$

Для связи x по y

$$\eta_{x/y} = \frac{\sigma(\bar{x}_y)}{\sigma_x}, \quad (78)$$

$$\sigma(\bar{x}_y) = \sqrt{\frac{\sum^y m_y (\bar{x}_y - \bar{x})^2}{n}}; \quad \sigma_x = \sqrt{\frac{\sum^x m_x (x - \bar{x})^2}{n}}.$$

Таким образом, корреляционное отношение будет разное для связи y по x и x по y:

$$\eta_{y/x}=\frac{\sigma(\bar{y}_x)}{\sigma_y}=\frac{1}{\sigma_y}\sqrt{\frac{1}{n}\Sigma m_x(\bar{y}_x-\bar{y})^2}=\frac{\sqrt{\Sigma m_x(\bar{y}_x-\bar{y})^2}}{\sqrt{\Sigma m_y(y-\bar{y})^2}}, \quad (79)$$

$$\eta_{x/y}=\frac{\sigma(\bar{x}_y)}{\sigma_x}=\frac{1}{\sigma_x}\sqrt{\frac{1}{n}\Sigma m_y(\bar{x}_y-\bar{x})^2}=\frac{\sqrt{\Sigma m_y(\bar{x}_y-\bar{x})^2}}{\sqrt{\Sigma m_x(x-\bar{x})^2}}. \quad (80)$$

Следовательно, для расчетов корреляционного отношения связи y по x необходимо определение квадратического отклонения его средних величин в строях $\sigma(\bar{y}_x)$ и определение общего квадратического отклонения σ_y величин y.

Корреляционное отношение выражают также через дисперсии. В этом случае η дается следующее определение. Корреляционным отношением называется корень квадратный из отношения межгрупповой дисперсии $\sigma^2(\bar{y}_x)$ к общей дисперсии $\sigma_y{}^2$:

$$\eta_{y/x}=\sqrt{\frac{\sigma^2(y_x)}{\sigma_y{}^2}}, \quad (81)$$

где $\sigma^2(\bar{y}_x)=\dfrac{\Sigma(\bar{y}_x-\bar{y})^2 m_x}{n}$ — межгрупповая дисперсия,

$\sigma_y{}^2=\dfrac{\Sigma(y-\bar{y})^2 m_y}{n}$ — общая дисперсия.

В двухмерной статистической совокупности могут быть три вида дисперсий:

1) Внутригрупповая дисперсия $\sigma_{xi}\ (y)$ — дисперсия значений y в каждой группе по столбцам или строям корреляционной таблицы около условного среднего $\bar{y}_x$ для определенного значения x. При этом значения y меняются при неизменном x. Значит это внутригрупповое рассеяние y не зависит от x, а зависит от других факторов. Значок x_i показывает, к какому условному распределению относится данное рассеяние y.

2) Межгрупповая дисперсия $\sigma^2(\bar{y}_x)$ — дисперсия условных средних $\bar{y}_x$ около общего среднего $\bar{y}_x$. В этом случае с изменением значений x условные средние y_x также меняют свои значения при переходе от одного условного распределения к другому.

3) Общая дисперсия $\sigma^2{}_y$ — дисперсия всех значений y (по всей таблице) около общего среднего $\bar{y}$. Таким образом, общая дисперсия $\sigma^2{}_y$ состоит из двух видов рассеяния y, двух дисперсий внутригрупповой $\sigma_x{}^2{}_i\ (y)$ и межгрупповой $\sigma^2(\bar{y}_x)$,

первая не зависит от x, а вторая дисперсия — межгрупповая — целиком зависит от x.

Корреляционное отношение η, выраженное через дисперсии, показывает, какую долю в общей мере рассеяния (дисперсии) занимает дисперсия данной величины y, возникающая вследствие влияния другой величины x.

Корреляционное отношение имеет следующие свойства:

1) Корреляционное отношение η всегда имеет значение между нулем и единицей

$$0 \leqslant \eta_{y/x} \leqslant 1.$$

2) Если корреляционное отношение равно нулю ($\eta_{y/x} = 0$ или $\eta_{x/y} = 0$), то между x и y не может быть никакой связи.

3) Если корреляционное отношение равно единице $\eta_{y/x} = 1$ или $\eta_{x/y} = 1$, то между x и y существует точная функциональная связь.

4) С возрастанием корреляционного отношения от нуля до единицы увеличивается теснота связи x и y, переходя при $\eta = 1$ в функциональную.

5) Корреляционное отношение всегда больше или равно коэффициенту корреляции $\eta \geqslant r$.

Средняя квадратическая ошибка корреляционного отношения равна

$$\sigma_\eta = \frac{1 - \eta^2}{\sqrt{n}}. \qquad (82)$$

§ 3. ПРИМЕР РАСЧЕТА УРАВНЕНИЯ ПАРАБОЛИЧЕСКОЙ СВЯЗИ И КОРРЕЛЯЦИОННОГО ОТНОШЕНИЯ ЗАВИСИМОСТИ УРОЖАЯ ОЗИМОЙ ПШЕНИЦЫ ОТ ВЕСЕННИХ ЗАПАСОВ ВЛАГИ ПРИ ЗАГУЩЕНИИ ПОСЕВОВ

Приведем пример определения параметров уравнения криволинейной параболической связи.

Нами был проведен анализ данных по урожайности озимой пшеницы сортов Одесская 3 и Одесская 16 и по весенним запасам влаги на Украине и Северном Кавказе. Мы выделили годы, когда озимая пшеница имела весной очень большое число стеблей (2000—2600) на 1 $м^2$ и построили корреляционное поле связи урожаев озимой пшеницы в эти годы с весенними запасами влаги (рис. 8).

На графике (рис. 8) видно, что связь явно криволинейная. Урожаи растут при увеличении запасов влаги весной до 170—180 *мм*, в годы же, когда было сочетание сильной загущенности посевов и больших запасов влаги весной, урожаи снижались.

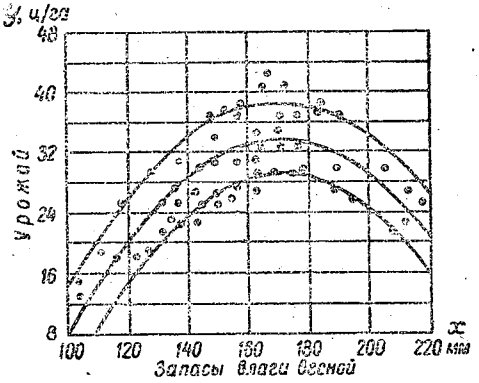

Рис. 8. Зависимость урожая (в *ц/га*) озимой пшеницы Одесская 3 и Одесская 16 от весенних запасов влаги (в *мм*) в метровом слое почвы в декаду перехода температуры воздуха через +5° весной на Украине при сильном загущении посевов (число стеблей пшеницы на 1 $м^2$ весной больше 2000).

Таким образом, при графическом изображении видно, что мы имеем расположение точек в виде параболической кривой.

Рассчитаем параметры уравнения для параболы второго порядка, общий вид уравнения которой

$$y = ax^2 + bx + c.$$

Будем вести расчеты методом группировки данных по частотам m_{xy}, который будет нам также необходим и для расчета корреляционного отношения. Для определения параметров уравнения нам необходимо рассчитать систему уравнений вида

$$\left.\begin{aligned} c\Sigma m_{x_i} + b\Sigma m_{x_i} x_i + a\Sigma m_{x_i} x^2{}_i &= \Sigma m_{x_i} \overline{y}_{x_i} \\ c\Sigma m_{x_i} x_i + b\Sigma m_{x_i} x^2{}_i + a\Sigma m_{x_i} x^3{}_i &= \Sigma m_{x_i} x_i \overline{y}_{x_i} \\ c\Sigma m_{x_i} x^2{}_i + b\Sigma m_{x_i} x^3{}_i + a\Sigma m_{x_i} x^4{}_i &= \Sigma m_{x_i} x^2{}_i \overline{y}_{x_i} \end{aligned}\right\}$$

Следовательно, нам нужно определить частоты m_{x_i} для y при

определенных значениях x и рассчитать условные средние $\bar{y}_{x_i}$ по каждому значению середины интервала x (гл. II, см. § 8). Для этого составляем корреляционную таблицу (табл. 20). Разбив на интервалы ось x и ось y (рис. 8), проводим вертикальные и горизонтальные линии по этим интервалам, которые дадут нам столбцы или строи табл. 20.

Таблица 20

Корреляционная таблица зависимости урожая озимой пшеницы (y) от весенних запасов влаги (x) при числе стеблей пшеницы 2000—2600 на 1 м² весной
(связь параболическая)

Интервалы	x	100—120	120—140	140—160	160—180	180—200	200—220	
y	середина интервала / x \ y	110	130	150	170	190	210	m_{y_i}
12—16	14	2	0					2
16—20	18	2	2					4
20—24	22	0	3	1			2	6
24—28	26	1	3	6	1	2	2	15
28—32	30		2	3	5	1	1	12
32—36	34			1	5	0		6
36—40	38			4	2	3		9
40—44	42				3			3
	Σm_{x_i}	5	10	15	16	6	5	57
	$\bar{y}_{x_i}$	18,0	24,0	30,3	34,2	32,7	25,2	29

Подсчитываем на рис. 8 число точек в каждой клетке, соответствующей определенному интервалу x и y, и заносим это число в корреляционную таблицу — в клетку с такими же интервалами. Получаем частоты m_{x_i} значений y для определенной величины середины интервала x_i.

Подсчитываем сумму частот m_{x_i} или m_{xy} по вертикальным столбцам и находим для каждого столбца условное среднее $\bar{y}_{x_i}$ по формуле $\bar{y}_{x_i} = \frac{\sum^{y} m_{x_y} y}{\Sigma m_{x_i}}$:

$$1)\ \bar{y}_{x=110} = \frac{2 \cdot 14 + 2 \cdot 18 + 0 \cdot 22 + 1 \cdot 26}{5} = 18;$$

2) $\overline{y}_{x=130} = \frac{0 \cdot 14 + 2 \cdot 18 + 3 \cdot 22 + 3 \cdot 26 + 2 \cdot 30}{10} = 24;$

3) $\overline{y}_{x=150} = \frac{1 \cdot 22 + 6 \cdot 26 + 3 \cdot 30 + 1 \cdot 34 + 4 \cdot 38}{15} = 30{,}3;$

4) $\overline{y}_{x=170} = \frac{1 \cdot 26 + 5 \cdot 30 + 5 \cdot 34 + 2 \cdot 38 + 3 \cdot 42}{16} = 34{,}2;$

5) $\overline{y}_{x=190} = \frac{2 \cdot 26 + 1 \cdot 30 + 0 \cdot 34 + 3 \cdot 38}{6} = 32{,}7;$

6) $\overline{y}_{x=210} = \frac{2 \cdot 22 + 2 \cdot 26 + 1 \cdot 30}{5} = 25{,}2.$

Затем находим общее среднее значение $\overline{y}$.

Затем по формуле $\overline{y} = \frac{\Sigma m_{x_i} \overline{y}_{x_i}}{n}$ находим среднее значение

$$\overline{y}_{\text{общ.}} = \frac{5 \cdot 18 + 10 \cdot 24 + 15 \cdot 30{,}3 + 16 \cdot 34{,}2 + 6 \cdot 32{,}7 + 5 \cdot 25{,}2}{57} = \frac{1654}{57} = 29{,}02.$$

Затем рассчитываем величины сумм, указанные в системе уравнений по табл. 21.

Полученные величины сумм подставляем в систему уравнений и получаем

$$\left.\begin{array}{l} 57\,c + 9010\,b + 1\,466\,500\,a = 1654 \\ 9010\,c + 1\,466\,500\,b + 245\,317\,000\,a = 266\,037 \\ 1\,466\,500\,c + 245\,317\,000\,b + 42\,088\,570\,000\,a = 43\,824750 \end{array}\right\}$$

Решаем указанную систему уравнений.

1) Делим каждое уравнение на коэффициенты при c, получаем

$$\begin{array}{l} c + 158{,}07\ b + 25\,728{,}07\ a = 29{,}02; \\ c + 162{,}76\ b + 27\,227{,}19\ a = 29{,}53; \\ c + 167{,}28\ b + 28\,700{,}01\ a = 29{,}88. \end{array}$$

2) Вычитаем из второго и третьего уравнений первое и получаем систему из двух уравнений с двумя неизвестными:

$$\begin{array}{l} 4{,}69\ b + 1499{,}12\ a = 0{,}51; \\ 9{,}21\ b + 2971{,}94\ a = 0{,}86. \end{array}$$

Расчет величины сумм, необходимых для определения

x_i	m_{x_i}	$m_{x_i}x_i$	$x^2{}_i$	$m_{x_i}x^2{}_i$	$x^3{}_i$	$m_{x_i}x^3{}_i$
110	5	550	12 100	60 500	1 331 000	6 655 000
130	10	1300	16 900	169 000	2 197 000	21 970 000
150	15	2250	22 500	337 500	3 375 000	50 625 000
170	16	2720	28 900	462 400	4 913 000	78 608 000
190	6	1140	36 100	216 600	6 859 000	41 154 000
210	5	1050	44 100	220 500	9 261 000	46 305 000
Σ	57	9010		1 466 500		245 317 000

3) Делим уравнения на коэффициенты при b и получаем

$$b + 319{,}642\ a = 0{,}109;$$

$$b + 322{,}686\ a = 0{,}093.$$

4) Вычитая из второго уравнение первое, находим значение параметра a:

$$3{,}044\ a = -\ 0{,}016;$$

$$a = -\frac{0{,}016}{3{,}044} = -\ 0{,}0052.$$

5) Подставляя значение a в одно из предыдущих уравнений, находим значения параметра b:

$$b = 0{,}109 - 319{,}642\,(-\ 0{,}0052) = 1{,}77.$$

6) Подставляем значения $\overline{a}$ и $\overline{b}$ в одно из уравнений с параметром c, находим значение c:

$$c = 29{,}02 - 158{,}07 \cdot (1{,}77) - 25728{,}07 - (-\ 0{,}0052) = -116{,}978.$$

7) Вносим значения параметров a, b и c в общий вид уравнения параболы второй степени и получаем искомое уравнение

$$y = -\ 0{,}0052x^2 + 1{,}77\ x - 116{,}98.$$

8) Находим теоретическую линию регрессии по данному уравнению. Задавая различные значения x, вычисляем по ним y:

При $x_1 = 110$ *мм*, y_1 по уравнению равен 14,8 *ц/га*.
При $x_2 = 140$; $y_2 = 28{,}9$.
При $x_3 = 160$; $y_3 = 33{,}1$.

Таблица 21

параметров уравнения параболы второй степени

x_i^4	$m_{x_i} x_i^4$	$\bar{y}_{x_i}$	$m_{x_i} \bar{y}_{x_i}$	$m_{x_i} x_i \bar{y}_{x_i}$	$m_{x_i} x_i^2 \bar{y}_{x_i}$
146 410 000	732 050 000	18	90	9 900	1 089 000
285 610 000	2 856 100 000	24	240	31 200	4 056 000
506 250 000	7 593 750 000	30,3	454,5	68 175	10 226 250
835 210 000	13 363 360 000	34,2	547,2	93 024	15 814 080
1 303 210 000	7 819 260 000	32,7	196,2	37 278	7 082 820
1 944 810 000	9 724 050 000	25,2	126	26 460	5 556 600
	42 088 570 000		1654	266 037	43 824 750

При $x_4=170$; $y_4=33{,}6$.
При $x_5=190$; $y_5=31{,}6$.
При $x_6=200$; $y_6=29{,}02$.
При $x_7=210$ $y_7=25{,}4$.
При $x_8=220$; $y_8=20{,}7$.
При $x_9=230$; $y_9=15{,}04$.

Наносим полученные точки с координатами $x_1y_1, x_2y_2 \ldots x_9y_9$ на график и проводим по ним теоретическую кривую — параболу второго порядка (рис. 8).

Часто на графиках проводят также линии значений y, рассчитанных по уравнению с учетом ошибки уравнения $(y \pm s_y)$. Между этими ограничивающими линиями бывает заключено более двух третей всех данных, вошедших в расчет уравнения связи.

Для криволинейной связи ошибка уравнения (s_y) вычисляется по формуле

$$S_y = \pm \sigma_y \sqrt{1-\eta^2},$$

σ_y — среднее квадратическое отклонение; η — корреляционное отношение.

Найдем корреляционное отношение, показывающее тесноту найденной параболической связи:

$$\eta_{y/x} = \sqrt{\frac{\sigma^2(\bar{y}_x)}{\sigma^2 y}}.$$

Рассчитываем межгрупповую дисперсию, пользуясь данными корреляционной табл. 20, по формуле

$$\sigma^2(\bar{y}_x) = \frac{\Sigma(\bar{y}_x - \bar{y})^2 m_{x_i}}{n}:$$

$$\sigma^2(\overline{y_x}) = \frac{(18-29)^2 \cdot 5 + (24-29)^2 \cdot 10 + (30{,}3-29)^2 \cdot 15}{57} +$$

$$+ \frac{(34{,}2-29)^2 \cdot 16 + (32{,}7-29)^2 \cdot 6 + (25{,}2-29)^2 \cdot 5}{57} = \frac{1467{,}33}{57} = 25{,}74.$$

Находим величину общей дисперсии по формуле

$$\sigma_y^2 = \frac{\Sigma (y-\overline{y})^2 m_{y_i}}{n}:$$

$$\sigma^2_y = \frac{(14-29)^2 \cdot 2 + (18-29)^2 \cdot 4 + (22-29)^2 \cdot 6 + (26-29)^2 \cdot 15}{57} +$$

$$+ \frac{(30-29)^2 \cdot 12 + (34-29)^2 \cdot 6 + (38-29)^2 \cdot 9 + (42-29)^2 \cdot 3}{57} =$$

$$= \frac{2761}{57} = 48{,}44.$$

Рассчитываем корреляционное отношение

$$\eta = \sqrt{\frac{\sigma^2(\overline{y_x})}{\sigma_y^2}} = \sqrt{\frac{25{,}74}{48{,}44}} = \sqrt{0{,}53} = 0{,}73.$$

Средняя ошибка корреляционного отношения

$$\sigma_\eta = \frac{1-\eta^2}{\sqrt{n}} = \frac{1-(0{,}73)^2}{\sqrt{57}} = \frac{0{,}47}{7{,}55} = \pm 0{,}06$$

Таким образом η находится в пределах $\eta \pm \sigma_\eta = \begin{cases} 0{,}79 \\ 0{,}67. \end{cases}$

Находим ошибку уравнения криволинейной регрессии по формуле

$$s_y = \pm \sigma_y \sqrt{1-\eta^2}.$$

По нашим расчетам $\sigma^2_y = 48{,}44$, следовательно, $\sigma_y = 6{,}96$.

$$s_y = \pm 6{,}96 \sqrt{1-(0{,}73)^2} = \pm 4{,}8 \ \text{ц/га}.$$

Связь урожая с весенними запасами влаги при сильном загущении посевов, как видим, несколько хуже по сравнению с очень тесной связью при меньшем травостое (1000—1900 стеблей на 1 $м^2$ весной, рис. 5).

§ 4. НАХОЖДЕНИЕ ПАРАМЕТРОВ УРАВНЕНИЙ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ МЕЖДУ ПЕРЕМЕННЫМИ ВЕЛИЧИНАМИ ГИПЕРБОЛИЧЕСКИХ, СТЕПЕННЫХ И ПОКАЗАТЕЛЬНЫХ КРИВЫХ

Часто связь между двумя переменными величинами выражена гиперболической кривой, уравнение которой в случае регрессии y на x имеет общий вид

$$y = \frac{a}{x} + b. \qquad (83)$$

Как было изложено выше, метод наименьших квадратов применим для линейных и параболических связей, для других видов связей метод наименьших квадратов можно применить к нахождению параметров уравнения только после некоторых преобразований.

При нахождении параметров уравнения гиперболы обычно делают замену переменных и приводят уравнение гиперболической зависимости к линейному виду. После этого находят параметры линейного уравнения методом наименьших квадратов или через r, σ_x, σ_y, $\bar{x}$ и $\bar{y}$.

Зависимость $y=\frac{a}{x}+b$ можно свести к линейной, заменив $\frac{1}{x}$ на x^1. Тогда уравнение будет иметь вид $y=ax^1+b$. Это, как известно, уравнение прямой линии.

Для нахождения параметров a и b уравнения гиперболы по методу наименьших квадратов необходимо значения x заменить значениями $x^1=\frac{1}{x}$ и составить систему необходимых уравнений.

Система нормальных уравнений для нахождения параметров a и b в этом случае будет иметь вид

$$\left.\begin{aligned} nb + a\sum\frac{1}{x} &= \sum y \\ b\sum\frac{1}{x} + a\sum\frac{1}{x^2} &= \sum\frac{y}{x} \end{aligned}\right\} \text{ или } \left.\begin{aligned} nb + a\sum x' &= \sum y \\ b\sum x' + a\sum x'^2 &= \sum x'y \end{aligned}\right\} \qquad (84)$$

Расчеты, необходимые для решения уравнений величин, проводятся по табл. 22. Подставляем полученные по таблице необходимые величины сумм в систему уравнений и решаем ее обычным образом:

1) Делим каждый член обоих уравнений на коэффициенты при b и получим два новых уравнения.

2) Вычитаем из второго уравнения первое и находим параметр a.

3) Подставив в одно из уравнений полученное значение a, находим значение параметра b.

Таблица 22

n	x	y	$x'=\frac{1}{x}$	$(x')^2=\frac{1}{x^2}$	$x'y=\frac{y}{x}$
			$\Sigma x'$	$\Sigma x'^2$	$\Sigma x'y'$

4) Подставляя полученные величины a и b в уравнение гиперболы, получаем окончательное искомое уравнение гиперболической связи.

5) Задавая значения x в найденном уравнении гиперболы, получаем значения y и строим теоретическую кривую линию регрессии.

Параметры уравнения a и b при данной замене можно определить с помощью определителей

$$a=\frac{\begin{vmatrix}\sum y & \sum\frac{1}{x}\\ \sum\frac{y}{x} & \sum\frac{1}{x^2}\end{vmatrix}}{\begin{vmatrix}n & \sum\frac{1}{x}\\ \sum\frac{1}{x} & \sum\frac{1}{x^2}\end{vmatrix}};\quad b=\frac{\begin{vmatrix}n & \sum y\\ \sum\frac{1}{x} & \sum\frac{y}{x}\end{vmatrix}}{\begin{vmatrix}n & \sum\frac{1}{x}\\ \sum\frac{1}{x} & \sum\frac{1}{x^2}\end{vmatrix}}.$$

Для гиперболы в случае сгруппированных данных по корреляционной таблице система уравнений имеет вид

$$\left.\begin{aligned} nb+a\sum m_{x_i}\frac{1}{x_i}&=\sum m_{x_i}\bar{y}_{x_i}\\ b\sum m_{x_i}\frac{1}{x_i}+a\sum m_{x_i}\frac{1}{x^2{}_i}&=\sum m_{x_i}\frac{1}{x_i}\bar{y}_{x_i}\end{aligned}\right\}\qquad(85)$$

При гиперболической зависимости можно сделать другую замену. Приведя уравнение гиперболы $y=\frac{a}{x}+b$ к общему знаменателю, имеем $xy=a+bx$. Обозначаем xy через z и получаем уравнение прямой линии $z=a+bx$, параметры которого определяются системой уравнений

$$\left.\begin{aligned} na+b\Sigma x&=\Sigma z\\ a\Sigma x+b\Sigma x^2&=\Sigma xz\end{aligned}\right\}\quad\text{или}\quad\left.\begin{aligned} na+b\Sigma x&=\Sigma xy\\ a\Sigma x+b\Sigma x^2&=\Sigma x^2y\end{aligned}\right\}\qquad(86)$$

Необходимые для решения уравнений суммы вычисляются по табл. 23.

Таблица 23

№ п/п	x	y	xy	x^2	x^2y
n	Σx	Σy	Σxy	Σx^2	Σx^2y

Решая указанную систему уравнений, находим параметры a и b и, подставляя их в уравнение гиперболы, получаем искомое уравнение.

Значение параметров или коэффициентов уравнения a и b при данной замене можно вычислить также с помощью определителей

$$a = \frac{\begin{vmatrix} \Sigma xy & \Sigma x \\ \Sigma x^2 y & \Sigma x^2 \end{vmatrix}}{\begin{vmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{vmatrix}} = \frac{\Sigma xy \Sigma x^2 - \Sigma x \Sigma x^2 y}{n \Sigma x^2 - \Sigma x \Sigma x}, \qquad (87)$$

$$b = \frac{\begin{vmatrix} n & \Sigma xy \\ \Sigma x & \Sigma x^2 y \end{vmatrix}}{\begin{vmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{vmatrix}} = \frac{n \Sigma x^2 y - \Sigma xy \Sigma x}{n \Sigma x^2 - \Sigma x \Sigma x}. \qquad (88)$$

Корреляционная связь между двумя переменными величинами может быть выражена также степенными (89) и показательными (90) (экспоненциальными) функциями:

$$y = bx^a, \qquad (89)$$

$$y = ab^x. \qquad (90)$$

Приведение степенных и показательных функций к линейным делают путем логарифмирования формулы. Если мы имеем степенную функцию $y = bx^a$ то, логарифмируя ее уравнение, получим $\lg y = a \lg x + \lg b$.

Обозначая $y' = \lg y$; $x' = \lg x$; $\lg b = B$ получаем уравнение прямой линии $y' = ax' + B$. Таким образом, соотношение между $\lg x$ и $\lg y$ является линейным и мы можем применять уже известные нам методы.

При расчете параметров уравнения методом наименьших квадратов или через r и σ таблицы с величинами наблюдений x и y пересчитывают на x' и y', т. е. берутся не сами величины, а их логарифмы.

При $y=\frac{b}{x^a}$ $\lg y=-a\lg x+\lg b$. Обозначаем $y'=\lg y$, $x'=\lg x$, $B=\lg b$. Отсюда будем иметь $y'=-ax'+B$.

Если расчеты параметров a и B проводят по методу наименьших квадратов, то составляют следующую систему уравнений:

$$\left.\begin{aligned} n\lg b+a\Sigma\lg x&=\Sigma\lg y\\ \lg b\,\Sigma\lg x+a\Sigma\lg x^2&=\Sigma\lg x\lg y\end{aligned}\right\} \qquad (91)$$

или

$$\left.\begin{aligned} nB+a\Sigma x'&=\Sigma y'\\ B\Sigma x'+a\Sigma x'^2&=\Sigma x'y'\end{aligned}\right\} \qquad (92)$$

Необходимые суммы для решения системы уравнений (91) рассчитывают по табл. 24.

Таблица 24

№ п/п	x	y	$x'=\lg x$	$y'=\lg y$	x'^2	$x'y'$
n			$\Sigma x'$	$\Sigma y'$	$\Sigma x'^2$	$\Sigma x'y'$

Если расчеты параметров уравнения ведут через коэффициент корреляции r и средние квадратические отклонения σ_x, σ_y, то пользуются табл. 25.

Определив параметр B, по нему определяют параметр b из соотношения $\lg b=B$.

Связь между двумя переменными может быть также выражена показательной (экспоненциальной) кривой, уравнение которой $y=ab^x$, где a и b — параметры уравнения, постоянные коэффициенты.

Подобными уравнениями выражаются связи между двумя величинами, когда увеличение функции (y) происходит значительно быстрее, чем увеличение аргумента (x).

Таблица 25

№ п/п	x	y	$x'=\lg x$	$y'=\lg y$	$\Delta x'=x'-\bar{x}'$	$\Delta y'=y'-\bar{y}'$	$\Delta x'^2$	$\Delta y'^2$	$\Delta x'\Delta y'$
n			$\Sigma x'$	$\Sigma y'$			$\Sigma\Delta x'^2$	$\Sigma\Delta y'^2$	$\Sigma\Delta x'\Delta y'$

Путем логарифмирования уравнение кривой приводят к уравнению прямой линии:

$$\lg y = \lg a + x \lg b. \tag{93}$$

Обозначив $\lg y = y'$; $\lg a = A$; $\lg b = B$, получим уравнение прямой $y' = Bx + A$. Параметры данного уравнения можно определить способом наименьших квадратов, составив систему уравнений и табл. 26:

Таблица 26

№ п/п	x	y	$y'=\lg y$	x^2	$xy'=x\lg y$
n	Σx		$\Sigma y'$	Σx^2	$\Sigma xy'$

$$\left.\begin{aligned} nA + B\Sigma x &= \Sigma y' \\ A\Sigma x + B\Sigma x^2 &= \Sigma xy' \end{aligned}\right\} \tag{94}$$

или

$$\left.\begin{aligned} n\lg a + \lg b\,\Sigma x &= \Sigma \lg y \\ \lg a\,\Sigma x + \lg b\,\Sigma x^2 &= \Sigma x \lg y \end{aligned}\right\} \tag{95}$$

Найдя по табл. 26 суммы и решив систему указанных двух уравнений, находим параметры A и B. Из выражений $A=\lg a$ и $B=\lg b$ находим значения a и b. Подставляя их в уравнение $y=ab^x$, находим искомое уравнение для данной криволинейной связи.

§ 5. ПРИМЕР РАСЧЕТА ПАРАМЕТРОВ УРАВНЕНИЯ СТЕПЕННЫХ КРИВЫХ (зависимость продолжительности периода всходы — кущение озимой ржи от температуры)

В агрометеорологии важным вопросом является вопрос нахождения зависимости межфазных периодов сельскохозяйственных культур от температуры воздуха. Связь длины межфазного периода многих культур, выраженная в днях, со средней температурой воздуха за период является обратной связью и представляет чаще всего гиперболическую или степенную кривую.

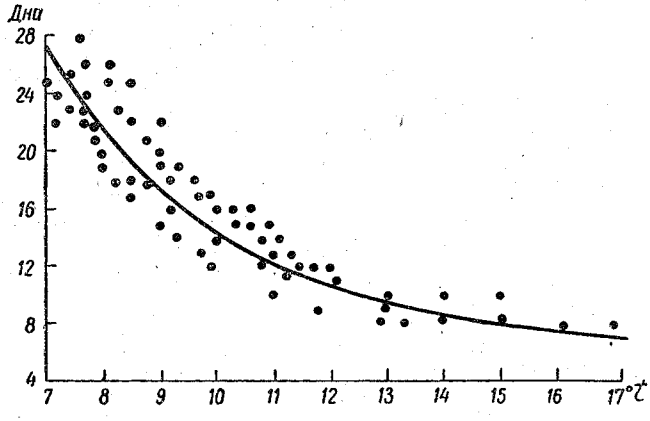

Рис. 9. Зависимость продолжительности периода всходы—кущение озимой ржи (y) от средней температуры за период (x) при хорошем увлажнении почвы (30—60 *мм*) в слое 0—22 *см*.

Рассмотрим найденную нами связь продолжительности периода от всходов до начала кущения озимой ржи с температурой воздуха за этот период осенью. Прежде всего, проводим анализ данных наблюдений. Так как продолжительность указанного межфазного периода зависит не только от температуры, но и от влажности почвы, то для нахождения зависимости от температуры исключим влияние влажности. Возьмем только те случаи, когда влажность почвы была оптимальной, т. е. когда влияние температуры было главным.

Наносим данные наблюдений за продолжительностью периода (y) и температурой воздуха (x) на график, т. е. строим корреляционное поле (рис. 9).

По расположению точек на графике мы видим, что связь криволинейная, обратная, скорее всего степенная (гиперболическая связь является также частным случаем степенной связи).

Уравнение обратной связи степенной кривой имеет вид

$$y = \frac{b}{x^a}. \qquad (96)$$

Для нахождения параметров данного уравнения необходимо привести его к линейному виду. Прологарифмируем это уравнение и получим $\lg y = -a \lg x + \lg b$. Обозначаем $\lg y = y_1$; $\lg x = x_1$; $\lg b = B$ и получаем $y_1 = -ax_1 + B$ — уравнение прямой линии для обратной связи.

Если мы будем брать не сами значения x и y, а их логарифмы и наносить значения этих логарифмов в логарифмической шкале на график, то точки будут на графике располагаться в виде прямой линии, для которой мы должны найти уравнение связи.

Параметры искомого уравнения прямой регрессии можно найти двумя путями: методом наименьших квадратов по системе уравнений (91) и по табл. 24 (см. гл. VI, § 4) или через коэффициент корреляции r и средние квадратические отклонения σ_x и σ_y по уравнению $y_1 - \bar{y}_1 = R(x_1 - \bar{x}_1)$, где $R = r\frac{\sigma_{y_1}}{\sigma_{x_1}}$.

Проведем расчет параметров уравнения регрессии вторым путем через r и σ с помощью табл. 27.

В табл. 27 вносим под одним порядковым номером среднюю температуру воздуха за период от всходов до кущения озимой ржи (x) и продолжительность этого периода в днях (y). Затем по таблицам логарифмов вычисляем логарифмы значений x и y. После этого находим суммы значений $x_1 = \lg x$ и $y_1 = \lg y$ и рассчитываем средние арифметические значения $\bar{x}_1$ и $\bar{y}_1$:

$$\bar{x}_1 = \frac{65{,}5351}{66} = 0{,}9930; \quad \bar{y}_1 = \frac{78{,}6721}{66} = 1{,}1920.$$

Рассчитываем последующие графы табл. 27, находим отклонения каждого x_1 и y_1 от средней величины, квадраты отклонений, произведение отклонений и подсчитываем суммы этих величин (указанные действия были изложены в § 9, гл. II).

Получив необходимые суммы, рассчитываем коэффициент корреляции связи:

$$r = \frac{\Sigma \Delta x_1 \Delta y_1}{\sqrt{\Sigma \Delta x_1^2 \Sigma \Delta y_1^2}} = \frac{-0{,}9001}{\sqrt{0{,}5661 \cdot 1{,}6443}} = -0{,}93.$$

Таблица для расчета связи продолжительности периода всходы—

№ п/п	x	y	$\lg x = x_1$	$\lg y = y_1$	$\Delta x_1 = x_1 - \bar{x}_1$	$\Delta y_1 = y_1 - \bar{y}_1$
1	7,6	28	0,8808	1,4472	—0,1122	0,2552
2	7,7	26	0,8865	1,4150	—0,1065	0,2230
3	7,0	25	0,8451	1,3979	—0,1479	0,2059
4	7,4	25	0,8692	1,3979	—0,1238	0,2059
5	7,2	24	0,8573	1,3802	—0,1357	0,1882
6	7,7	24	0,8865	1,3802	—0,1065	0,1882
7	7,3	23	0,8633	1,3617	—0,1297	0,1697
8	7,7	23	0,8865	1,3617	—0,1065	0,1697
9	7,7	22	0,8865	1,3424	—0,1065	0,1504
10	7,2	22	0,8573	1,3424	—0,1357	0,1504
.						
.						
.						
56	12,1	11	1,0828	1,0414	0,0898	—0,1506
57	13,0	10	1,1139	1,0000	0,1209	—0,1920
58	13,0	9	1,1139	0,9542	0,1209	—0,2378
59	12,9	8	1,1106	0,9031	0,1176	—0,2889
60	13,3	8	1,1239	0,9031	0,1309	—0,2889
61	14,0	10	1,1461	1,0000	0,1531	—0,1920
62	13,9	7	1,1430	0,8451	0,1500	—0,3469
63	15,0	10	1,1761	1,0000	0,1831	—0,1920
64	15,0	8	1,1761	0,9031	0,1831	—0,2889
65	16,1	8	1,2068	0,9031	0,2138	—0,2889
66	17,0	8	1,2304	0,9031	0,2374	—0,2889
$\Sigma 66$	—	—	65,5351	78,6721	—	—

Таблица 27

кущение озимой ржи от средней температуры за период

$\Delta x_1^2 = (x_1 - \bar{x}_1)^2$	$\Delta y_1^2 = (y_1 - \bar{y}_1)^2$	$(x_1 - \bar{x}_1)(y_1 - \bar{y}_1)$	$(x_1 - \bar{x}_1) + (y_1 - \bar{y}_1)$	$[(x_1 - \bar{x}_1) + (y_1 - \bar{y}_1)]^2$
0,0126	0,0651	—0,0286	0,1430	0,0204
0,0113	0,0497	—0,0237	0,1165	0,0136
0,0219	0,0424	—0,0304	0,0580	0,0034
0,0153	0,0424	—0,0255	0,0821	0,0067
0,0184	0,0354	—0,0255	0,0525	0,0027
0,0113	0,0354	—0,0200	0,0817	0,0067
0,0168	0,0288	—0,0220	0,0400	0,0016
0,0113	0,0288	—0,0181	0,0632	0,0040
0,0113	0,0226	—0,0160	0,0439	0,0019
0,0184	0,0226	—0,0204	0,0147	0,0002
0,0081	0,0227	—0,0135	—0,0608	0,0037
0,0146	0,0369	—0,0232	—0,0711	0,0050
0,0146	0,0565	—0,0288	—0,1169	0,0136
0,0138	0,0835	—0,0340	—0,1713	0,0293
0,0171	0,0835	—0,0378	—0,1580	0,0250
0,0234	0,0369	—0,0294	—0,0389	0,0015
0,0225	0,1203	—0,0520	—0,1969	0,0388
0,0335	0,0369	—0,0352	—0,0089	0,0000
0,0335	0,0835	—0,0529	—0,1058	0,0112
0,0457	0,0835	—0,0618	—0,0751	0,0056
0,0564	0,0835	—0,0686	—0,0515	0,0026
0,5661	1,6443	—0,9001	—	0,4102

Ошибка коэффициента корреляции

$$\sigma_r = \pm \frac{1 - r^2}{\sqrt{n}} = \frac{1 - (-0{,}93)^2}{\sqrt{66}} \approx \pm 0{,}017,$$

$$r \pm \sigma_r = -0{,}93 \pm 0{,}017 = \begin{cases} -0{,}95 \\ -0{,}91. \end{cases}$$

Коэффициент уравнения регрессии

$$R = r \frac{\sigma_{y_1}}{\sigma_{x_1}}; \quad \sigma_{y_1} = \sqrt{\frac{\Sigma \Delta y_1^2}{n}};$$

$$\sigma_{x_1} = \sqrt{\frac{\Sigma \Delta x_1^2}{n}}; \quad \frac{\sigma_{y_1}}{\sigma_{x_1}} = \frac{\sqrt{\Sigma \Delta y_1^2}}{\sqrt{\Sigma \Delta x_1^2}} = 1{,}70.$$

Определяем коэффициент уравнения регрессии

$$R = r \frac{\sigma_{y_1}}{\sigma_{x_1}} = -0{,}93 \cdot 1{,}70 = -1{,}58.$$

Подставляем в уравнение $y_1 - \bar{y}_1 = R(x_1 - \bar{x})$ величины R, $\bar{y}_1$, $\bar{x}_1$:

$$y_1 - 1{,}1920 = -1{,}581\ (x_1 - 0{,}9930);$$
$$y_1 = 1{,}1920 - 1{,}581\ x_1 + 1{,}5699.$$

Получаем искомое уравнение связи:

$$y_1 = 2{,}76 - 1{,}58\ x_1 \text{ или } \lg y = 2{,}76 - 1{,}58\ \lg x.$$

По таблице антилогарифмов любого математического справочника находим b: $\lg b = 2{,}76$; $b = 575$.

Таким образом, мы можем написать искомое уравнение нашей криволинейной связи в виде $y = \frac{575}{x^{1{,}58}}$. Задавая различные значения x, рассчитываем ряд значений y и строим теоретическую кривую связи. Эти расчеты ведутся по формуле $\lg y = 2{,}76 - 1{,}58 \lg x$, так как по формуле $y = \frac{575}{x^{1{,}58}}$, где участвует степень 1,58, рассчитывать без логарифмирования значения y нельзя.

Задаем произвольно пять значений x, получаем пять значений y:

1) При $x_1 = 8$ $\lg y = 2{,}76 - 1{,}58 \lg 8 =$
2) При $x_2 = 10$ $\lg y = 2{,}76 - 1{,}58 \lg 10 =$
3) При $x_3 = 12$ $\lg y = 2{,}76 - 1{,}58 \lg 12 =$
4) При $x_4 = 14$ $\lg y = 2{,}76 - 1{,}58 \lg 14 =$
5) При $x_5 = 16$ $\lg y = 2{,}76 - 1{,}58 \lg 16 =$

$= 2{,}76 - 1{,}58 \cdot 0{,}90 = 1{,}34$; $y = 21{,}9$.
$= 2{,}76 - 1{,}58 \cdot 1{,}00 = 1{,}18$; $y = 15{,}1$.
$= 2{,}76 - 1{,}58 \cdot 1{,}08 = 1{,}05$; $y = 11{,}2$.
$= 2{,}76 - 1{,}58 \cdot 1{,}15 = 0{,}94$; $y = 8{,}7$.
$= 2{,}76 - 1{,}58 \cdot 1{,}20 = 0{,}86$; $y = 7{,}2$.

Наносим точки со значениями x_1y_1; x_2y_2 ... x_5y_5 на корреляционное поле рис. 9 и проводим по ним теоретическую линию кривой регрессии, по которой впоследствии без расчета по уравнению можно снимать значения y по заданным значениям x.

Для определения величин логарифмов, различных степеней переменных, величин x и y и других значений следует пользоваться математическими таблицами, где указанные величины уже рассчитаны.

Кроме этого, вычисления следует проводить при помощи счетных машин, не заполняя в таблицах графы степеней и произведений по каждому порядковому номеру, а получая на машине и записывая сразу значения сумм степеней и произведений. Это значительно уменьшит объем работы по расчетам параметров уравнений корреляционных связей.

ЛИТЕРАТУРА

1. Берфин П. Составление эмпирических формул зависимости по экспериментальным данным. Брянск, 1957.
2. Ван дер. Варден. Математическая статистика. Изд-во иностр. лит. М., 1960.
3. Венецкий И. Г., Венецкая В. И. Основы математической статистики для заочников. Профиздат. М., 1960.
4. Длин А. М. Математическая статистика в технике. Изд-во. Сов. наука. М., 1958.
5. Дунин И. В., Барковский и Смирнов Н. В. Теория вероятностей и математическая статистика в технике. Гостехиздат. М., 1955.
6. Кондратьева Е. Прямолинейная корреляция. Ред.-изд. отд. ЦУЕГМС СССР, М., 1936.

7. Крамер Г. Математические методы статистики. Изд-во иностр. лит. М., 1948.
8. Леонтьев Н. Л. Статистическая обработка результатов наблюдений. М. Л., 1952.
9. Линник Ю. В. Метод наименьших квадратов и основы математико-статистической обработки наблюдений. М., 1958.
10. Лукомский Я. И. Теория корреляции и ее применение к анализу производства. Госстатиздат. М., 1961.
11. Маслов П. Корреляция. М., 1955.
12. Маркович Э. С. Элементы теории корреляции. М., 1958.
13. Мевзос Л. М. Методическое пособие по математической статистике. Харьков, 1963.
14. Митропольский А. К. Статистическое исчисление. Л., 1952.
15. Митропольский А. К. Техника статистических вычислений. Физматгиз М., 1961.
16. Романовский В. И. Математическая статистика, кн. 1 и 2. Изд. АН УзССР, 1960.
17. Румшицкий Л. З. Элементы теории вероятностей. Физматгиз М., 1963.
18. Рокицкий П. Ф. Основы вариационной статистики для биологов. Минск, 1961.
19. Снедекор Дж. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. Изд-во иностр. лит. М., 1961.
20. Уорсинг А. и Геффнер Дж. Методы обработки экспериментальных данных. Изд. иностр. литературы. М., 1949.
21. Урбах В. Ю. Математическая статистика для биологов и медиков. М., 1963.
22. Успенский А. К. Выбор вида и нахождение параметров эмпирической формулы. М., 1960.
23. Фишер Р. А. Статистические методы для исследователей. Госстатиздат. М., 1958.
24. Фокин Г. К., Мелентьев Е. К. Методическое пособие к решению задач по математической статистике. Куйбышев, 1963.
25. Хальд А. Математическая статистика с техническими приложениями. Изд-во иностр. лит. М., 1956.
26. Чупров А. А. Основные проблемы теории корреляции. О статистическом исследовании связи между явлениями. Госстатиздат. М.
27. Юл Д. Э. и Кэндел М. Д. Теория статистики. Госстатиздат. М., 1960.
28. Ястремский Б. С. Некоторые вопросы математической статистики. Госстатиздат М., 1961.

Примеры корреляционных связей и расчеты уравнений взяты из работ:

1. Немчинов В. С. Сельскохозяйственная статистика с основами общей теории. Сельхозгиз. М., 1946.
2. Уланова Е. С. Методы агрометеорологических прогнозов. Гидрометеоиздат. Л., 1959.
3. Уланова Е. С. Метод долгосрочного прогноза агрометеорологических условий формирования урожая озимой пшеницы. Метеорология и гидрология, № 11, 1963.
4. Уланова Е. С. Агроклиматические условия осеннего периода развития и роста озимых культур в Западной Сибири. Труды ЦИП, вып. 47 (74), 1956.
5. Уланова Е. С., Рымар А. Л. О связи запасов продуктивной влаги в различных слоях почвы под озимой пшеницей в осенний период. Труды ЦИП, вып. 131, 1963.
6. Уланова Е. С., Цао Ин. Зависимость запасов продуктивной влаги под кукурузой от метеорологических факторов на Украине. Труды ЦИП, вып. 131, 1963.

СОДЕРЖАНИЕ

ЕВГЕНИЯ СТАНИСЛАВОВНА УЛАНОВА

ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В АГРОМЕТЕОРОЛОГИИ ДЛЯ НАХОЖДЕНИЯ УРАВНЕНИЙ СВЯЗИ

Отв. редактор *Г. И. Морской*
Редактор *В. В. Рощина* Техн. ред. *И. М. Зарх* Корректор *Т. Д. Сурикова*
Московское отделение Гидрометеоиздата. Москва, ул. Горького, д. 18-а

Т-09830 Сдано в набор 27/III 1964 г. Подписано к печати 26/VI 1964 г.
Изд. № 190 Индекс М-М-190 Бумага 60×90¹/₁₆ Печ. л. 7 Уч. изд. л. 6,21
Заказ № 415 Цена 31 коп. Тираж 2790

1-я типолитография Гимиза. Москва, Измайловское шоссе, 42