Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет»

П.В. Медведев,
В.А. Федотов

МАТЕМАТИЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

Рекомендовано ученым советом федерального государственного бюджетного образовательного учреждения высшего образования «Оренбургский государственный университет» для обучающихся по образовательным программам высшего образования по направлению подготовки 19.04.02 Продукты питания из растительного сырья

Оренбург
2017

УДК 664.65.05 (075.8)
ББК 36.83-5я73
М 42

Рецензент – доктор технических наук, профессор В. Ю. Полищук

Медведев, П.В.
М 42 Математическая обработка результатов исследования: учебное пособие/ П.В. Медведев, В.А. Федотов; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2017. – 99 с.
ISBN978-5-7410-1772-2

В учебном пособии приведены основные теоретические вопросы курса «Математическая обработка результатов исследования» в виде лекций; задачи для проведения корреляционного и регрессионного анализа; примеры проведения статистических исследований с помощью программного комплекса StatSoft Statistica.

Учебное пособие предназначено для студентов, обучающихся по программам высшего образования по направлению подготовки 19.04.02 Продукты питания из растительного сырья

УДК 664.65.05 (075.8)
ББК 36.82-5я73

ISBN 978-5-7410-1772-2

Содержание

Введение

Одним из требований ускорения научно-технического прогресса является использование передовых технологий и методик исследования. Роль научных исследований возрастает и в связи с переходом высшей школы нашей страны на двухуровневую систему подготовки специалистов, когда окончание магистратуры завершается защитой диссертации - выпускной квалификационной работы научно-исследовательского характера. При этом на выбор темы исследования, анализ современного состояния исследуемого явления, постановку задачи и ее решение, написание статей, оформление и защиту диссертации отводится сравнительно мало времени. Получению наилучших решений за сравнительно короткое время поможет владение методами математического планирования эксперимента.

В вопросе повышения эффективности научных исследований и сокращения сроков разработки новых технологий необходимо решать задачи поиска моделей исследуемых процессов и оптимизация исследований на всех стадиях разработки, исследования и эксплуатации изучаемых процессов [1].

На практике встречаются случаи, когда теоретическое решение задачи отсутствует, особенно это относится к сложным многофакторным процессам. И тогда единственным инструментом остается проведение эксперимента и грамотная обработка результатов опытов. В настоящее время методологической основой экспериментальных исследований является математическая теория планирования эксперимента, которая базируется на приложениях теории вероятности и математической статистики. Соответствующую дисциплину изучают во многих ведущих вузах страны или же это является составной частью других дисциплин, например, основ научных исследований.

Целью данной работы является пропаганда идей математического анализа результатов исследований, способствующих улучшению качества их проведения, правильной обработке и интерпретации результатов. В работе сделана попытка «прокладки мостика» между теоретическими разработками и практическим

приложением, для чего рассмотрены практические примеры из различных отраслей техники.

1 Предварительная обработка данных исследований. Основные понятия статистики и их характеристики

Методы математической статистики широко используются при расчетах характеристик параметров, описывающих самые различные явления окружающей жизни. Например, в процессе эксплуатации сооружения и объекты испытывают воздействия, имеющие вероятностную природу. Значит, эти объекты должны быть запроектированы с высокой надежностью с учетом вероятностных законов, поскольку в случае аварии или повреждения материальный и социальный ущерб будет иметь значительные объемы [2].

Материалы, из которых возводятся объекты, характеризуются временной изменчивостью физических свойств. Определяющие функционирование объектов сейсмические, климатические, антропогенные, геологические, гидрологические и почвенно-мелиоративные условия зависят от большого множества факторов и также носят случайный характер. Методы обработки таких данных базируются на положениях математической статистики (рисунок 1).

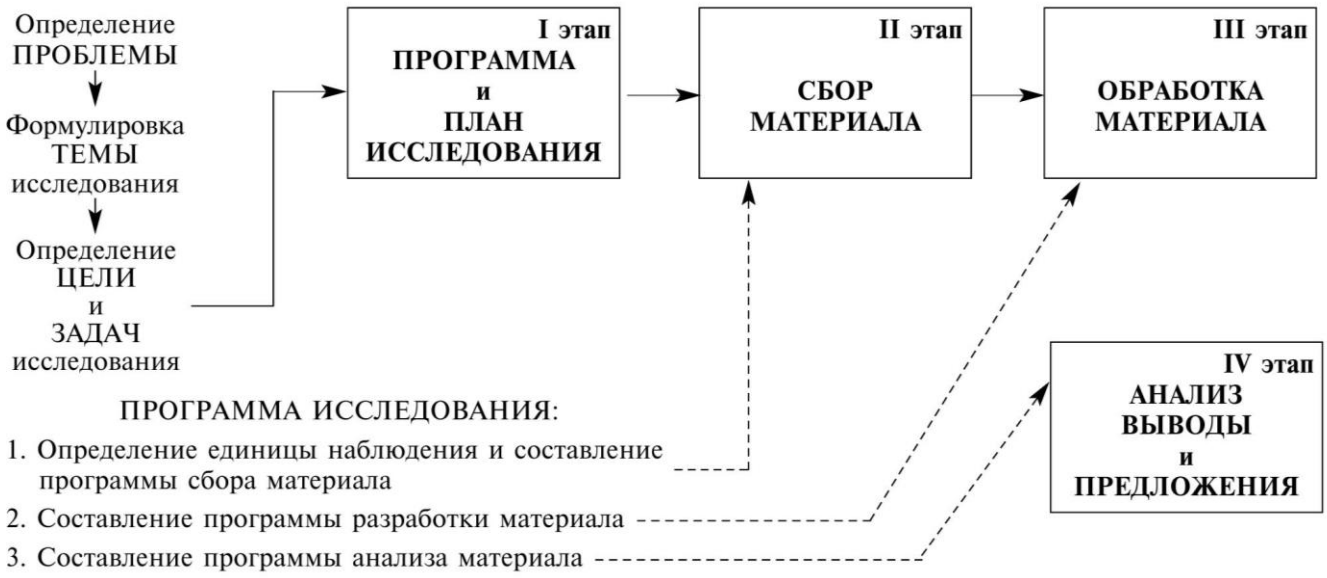


Рисунок 1 – Этапы статистического исследования

1.1 Случайные величины

Случайная величина x есть величина определенной физической размерности, принимающая в результате эксперимента то или иное числовое значение, которая в принципе нельзя предсказать, исходя из условия проведения эксперимента, то есть она может принимать случайные значения. Например, температура или влажность почвы, направление ветра, количество осадков за сезон, число аварий в системе водоснабжения и т. д.

Случайная величина называется дискретной, если она принимает последовательные различные значения и известна вероятность каждого из них, другими словами, она принимает лишь определенные значения, которые можно занумеровать или перечислить, например, число двоек на экзамене, количество одновременных потребителей воды в городе в данный момент времени, число абитуриентов в разные годы и т. д.).

Случайная величина называется непрерывной, если она может принимать все значения в заданных границах или на всей числовой оси (например, расход или скорость воды в реке, продолжительность безотказной работы затвора, температура или давление жидкости, вес студента и т. д.).

1.2 Типы данных

Значения переменных, которые регистрируются с помощью чисел, имеющих содержательный смысл, называют количественными данными. Данные, приведенные в таблице 1, в столбцах с 3 по 8, являются количественными. В то же время числа, приведенные в первом столбце этой таблицы, не являются количественными данными: они только указывают на номер государства при их алфавитном расположении. Эти числа не имеют содержательной интерпретации. С

количественными данными можно выполнять все обычные операции над числами, такие, как вычисление среднего и оценку изменчивости.

В зависимости от того, какие значения может потенциально принимать переменная, выделяют два типа количественных данных: дискретные и непрерывные.

Дискретная - это такая переменная, которая может принимать значения только из некоторого списка определенных чисел. Примерами дискретной переменной являются число детей в семье; число вызовов «скорой помощи», поступающих в больницу; число отказов изделия; число клиентов, обратившихся в фирму за определенный промежуток времени, и т. д.

Непрерывной будем считать любую переменную, не являющуюся дискретной. Она принимает значения из некоторого промежутка. Примерами непрерывной переменной является рост взрослого человека (например, от 140 до 230 см), фактическая масса буханки хлеба (например, от 750 до 830 г), дальность полета снаряда, урожайность культуры, выращенной в хозяйстве и т. п.

Есть данные, которые регистрируют определенное качество, которым обладает объект. Такие данные называют качественными. Даже если значениям этого качества можно приписать числа (например, полу человека приписать соответственно числа 0 и 1), то обрабатывать эти числа как количественные данные нельзя. Примерами качественных данных являются тип школы, где обучается ребенок (лицей, гимназия, специализированная физико-математическая школа); должность, которую занимает сотрудник на предприятии; названия газет, которые читают в определенном городе, и т. п.

Качественные данные бывают двух типов: порядковые, для которых существует имеющий содержательный смысл порядок, и номинальные, для которых нет содержательно интерпретируемого порядка.

Порядковые данные можно ранжировать и использовать это ранжирование при проведении статистического анализа. Примером порядковых данных являются ответы на вопросы анкеты, содержащей следующие варианты ответов: да; больше да, чем нет; больше нет, чем да; нет. Хоть и можно выразить эти ответы числами

(например, 4, 3, 2, 1), но предложенная шкала оценок носит субъективный характер. Нельзя считать, что разница между ответами 4 и 3 такая же, как и между ответами 2 и 1. Также нельзя считать, что ответ 3 в три раза лучше ответа 1.

Таблица 1 - Производство пшеницы по годам, в миллионах тонн

Номер государства	Страна	2010	2011	2012	2013	2014	2015	2016
1	Россия	30,1	34,5	50,6	34,1	45,4	47,7	45,0
2	Австрия	1,3	1,3	1,4	1,2	1,7	1,5	1,4
3	Албания	0,4	0,3	0,3	0,3	0,3	0,3	0,2
4	Беларусь	0,4	1,0	1,0	0,8	1,1	1,2	1,1
5	Бельгия	-	1,7	1,7	1,6	1,9	1,8	1,6
6	Болгария	3,4	2,8	4,1	2,0	4,0	3,5	3,3
7	Венгрия	4,6	3,7	3,9	2,9	6,0	5,1	4,4
8	Германия	17,8	21,6	20,8	19,3	25,4	23,7	22,4
9	Греция	2,3	2,3	2,0	1,7	2,1	2,0	1,4
10	Дания	4,6	4,7	4,1	4,7	4,8	4,9	4,8
11	Ирландия	0,6	0,7	0,9	0,8	1,0	0,8	0,8
12	Испания	3,1	7,3	6,8	6,3	7,1	4,0	5,6
13	Италия	7,9	7,5	7,5	6,2	8,6	7,7	7,1
14	Латвия	0,2	0,4	0,5	0,5	0,5	0,7	0,6
15	Литва	0,6	1,2	1,2	1,2	1,4	1,4	0,8
16	Нидерланды	1,2	1,1	1,1	1,1	1,2	1,2	1,2
17	Норвегия	0,3	0,3	0,3	0,3	0,4	0,4	0,4
18	Республика Македония	0,4	0,3	0,3	0,2	0,4	0,3	0,3
19	Республика Молдова	1,3	0,7	1,1	0,1	0,9	1,1	0,7

Продолжение таблицы 1

Номер государства	Страна	2010	2011	2012	2013	2014	2015	2016
20	Румыния	7,7	4,5	4,4	2,5	7,8	7,3	5,5
21	Словакия	1,9	1,3	1,6	0,9	1,8	1,6	1,3
22	Словения	0,2	0,2	0,2	0,1	0,1	0,1	0,1
23	Соединенное Королевство (Великобритания)	14,3	16,7	16,0	14,3	15,5	14,9	14,7
24	Украина	16,3	10,2	20,6	3,6	17,5	18,7	13,9
25	Финляндия	0,4	0,5	0,6	0,7	0,8	0,8	0,7
26	Франция	30,9	37,4	38,9	30,5	39,7	36,9	35,4
27	Чешская Республика	3,8	4,1	3,9	2,6	5,0	4,1	3,5
28	Швейцария	0,6	0,6	0,5	0,4	0,5	0,5	0,5
29	Швеция	1,6	2,4	2,1	2,3	2,4	2,2	2,0
30	Эстония	0,1	0,1	0,1	0,1	0,2	0,3	0,2

1.3 Генеральная и выборочная совокупности

Совокупность значений случайной величины, которые получены из опыта или текущего наблюдения, называют статистической совокупностью. Статистическая совокупность, содержащая все возможные значения случайной величины, называют генеральной совокупностью. Однако на практике исследователь имеет лишь ограниченное число наблюдений (опытов), которое называют выборкой. Выборка (выборочная статистическая совокупность) практически всегда является лишь небольшой частью генеральной совокупности.

В силу невозможности практической реализации большого числа опытов характеристики генеральной совокупности приходится оценивать приближенно по

выборке данных (рисунок 2). Этот метод, называемый выборочным, применим при выполнении следующих требований:

а) репрезентативность (представительность - от английского слова represent – представлять, быть представительной), то есть разные элементы генеральной совокупности должны имеет одинаковую вероятность попадания в выборку или другими словами характеристики выборки (выборочное среднее, выборочная дисперсия, коэффициент корреляции и др.) должны совпадать со значениями соответствующих характеристик генеральной совокупности с наперед заданной точностью, для чего количество опытов (объем выборки) должно быть не слишком малым;

б) независимость испытаний, когда по результатам одного опыта нельзя предсказать исход другого опыта;

в) результаты опытов считаются разными значениями (реализациями) одной и той же случайной величины, то есть опыты с разными номерами, рассматриваемые как случайные величины, имеют одну и ту же функцию распределения.

Выборка, удовлетворяющая указанным условиям, называется случайной.

Основные способы достижения этого условия (приближения к идеалу, абсолютной точности здесь достичь нельзя):

- случайная выборка;

- моделирование выборки по свойствам генеральной совокупности.

Существенным при организации выборки является вопрос о необходимом и достаточном числе испытуемых. Малое количество испытуемых не обеспечит точности результатов, большое количество приведет к увеличению трудоемкости (времени и стоимости) исследования.

В США одним из наиболее известных исторических примеров нерепрезентативной выборки считается случай, происшедший во время президентских выборов в 1936 году. Журнал «Литрери Дайджест», успешно прогнозировавший события нескольких предшествующих выборов, ошибся в своих предсказаниях, разослав десять миллионов пробных бюллетеней своим

подписчикам, а также людям, выбранным по телефонным книгам всей страны и людям из регистрационных списков автомобилей. В 25 % вернувшихся бюллетеней (почти 2,5 миллиона) голоса были распределены следующим образом: 57 % отдавали предпочтение кандидату-республиканцу Альфу Лэндону 40 % выбрали действующего в то время президента-демократа Франклина Рузвельта. На действительных же выборах, как известно, победил Рузвельт, набрав более 60 % голосов [3].

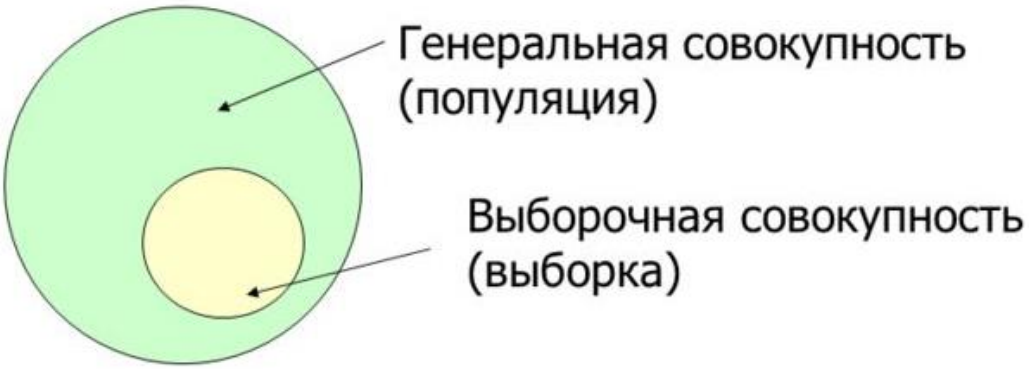


Рисунок 2 - Генеральная и выборочная совокупности

Ошибка «Литрери Дайджест» заключалась в следующем: желая увеличить репрезентативность выборки, так как им было известно, что большинство их подписчиков считают себя республиканцами - они расширили выборку за счёт людей, выбранных из телефонных книг и регистрационных списков. Однако они не учли современных им реалий и в действительности набрали ещё больше республиканцев: во время Великой депрессии обладать телефонами и автомобилями могли себе позволить в основном представители среднего и высшего класса (то есть большинство республиканцев, а не демократов).

1.4 Числовые характеристики случайных величин

Среднее арифметическое ряда чисел – это сумма данных чисел, поделенная на количество слагаемых. Среднее арифметическое называют средним значением числового ряда.

Пример: Найдем среднее арифметическое чисел 2, 6, 9, 15.

Решение. У нас четыре числа. Значит, надо их сумму разделить на 4. Это и будет среднее арифметическое данных чисел

$$\frac{2+6+9+15}{4}=8.$$

Размах ряда чисел – это разность между наибольшим и наименьшим из этих чисел.

Пример: Найти размах чисел 2, 5, 8, 12, 33.

Решение: Наибольшее число здесь 33, наименьшее 2. Значит, размах составляет

$$33 – 2 = 31.$$

Мода ряда чисел – это число, которое встречается в данном ряду чаще других.

Пример: Найти моду ряда чисел 1, 7, 3, 8, 7, 12, 22, 7, 11, 22, 8.

Решение: Чаще всего в этом ряде чисел встречается число 7 (3 раза). Оно и является модой данного ряда чисел.

Медиана. В упорядоченном ряде чисел: медиана нечетного количества чисел – это число, записанное посередине.

Пример: В ряде чисел 2, 5, 9, 15, 21 медианой является число 9, находящееся посередине.

Медиана четного количества чисел – это среднее арифметическое двух чисел, находящихся посередине.

Пример: Найти медиану чисел 4, 5, 7, 11, 13, 19.

Решение: Здесь четное количество чисел (6). Поэтому ищем не одно, а два числа, записанных посередине. Это числа 7 и 11. Находим среднее арифметическое этих чисел

$$\frac{7+11}{2}=9.$$

Число 9 и является медианой данного ряда чисел.

Медианой произвольного ряда чисел называется медиана соответствующего упорядоченного ряда.

Пример: Найдем медиану произвольного ряда чисел 5, 1, 3, 25, 19, 17, 21.

Решение: Располагаем числа в порядке возрастания: 1, 3, 5, 17, 19, 21, 25.

Посередине оказывается число 17. Оно и является медианой данного ряда чисел.

Пример: Добавим к нашему произвольному ряду чисел еще одно число, чтобы ряд стал четным, и найдем медиану: 5, 1, 3, 25, 19, 17, 21, 19.

Решение: Снова выстраиваем упорядоченный ряд: 1, 3, 5, 17, 19, 19, 21, 25.

Посередине оказались числа 17 и 19. Находим их среднее значение

$$\frac{17+19}{2}=18.$$

Число 18 и является медианой данного ряда чисел [4].

1.5 Задания для самостоятельной работы по теме «Основные понятия статистики и их характеристики»

1.5.1 Найти среднее арифметическое чисел 8, 9, 10, 15, 2, 89, 54, 25, 21, 47.

1.5.2 Найти медиану чисел 5, 9, 10, 14, 21, 89, 54, 25, 9, 10, 14, 47, 74, 21.

1.5.3 Найти размах чисел 89, 54, 25, 9, 1, 47, 97, 74, 9, 10, 14, 32, 47, 20, 15, 14.

1.5.4 Найти размах чисел 9, 1, 89, 54, 25, 47, 97, 74, 9, 10, 14, 21, 12, 78, 23.

1.5.5 Найти среднее арифметическое чисел 89, 54, 25, 9, 1, 47, 97, 74, 9, 10, 14, 32, 47, 20, 15, 14.

1.5.6 Найти среднее арифметическое чисел 5, 9, 10, 14, 21, 89, 54, 25, 9, 10, 14, 47, 74, 21.

1.6 Нормальное распределение. Критерии нормальности

Встречаемые в обыденной жизни большое число наблюдений, измеряемые значения параметров, погрешности измерений и ошибки подчиняются нормальному закону. Этот закон можно считать основой математической статистики. Нормальным называется закон распределения случайной величины x, если плотность распределения вероятности определяется по формуле

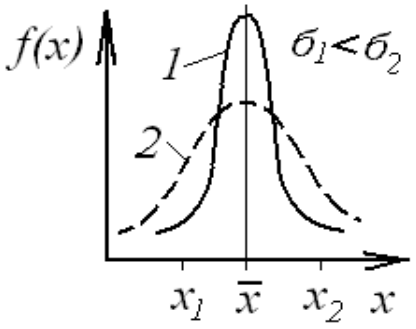
$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma}(x-\bar{x})^2}, \quad (-\infty < x < +\infty,\ \sigma > 0), \qquad (1)$$

где σ – среднее квадратическое отклонение,

$\bar{x}$ - среднее значение (математическое ожидание),

π = 3,141593, e = 2,718282 – математические константы.

Из формулы следует, что нормальное распределение полностью определяется параметрами $\bar{x}$ и σ. Среднее квадратическое отклонение определяет форму кривой: чем больше σ (разброс данных), тем кривая становится пологой (рисунок 3). Математическое ожидание $\bar{x}$ определяет положение кривой на оси абсцисс, кривая симметрична относительно этого значения [5, 6].



1 – с меньшим разбросом данных, 2 – с большим разбросом данных

Рисунок 3 – Виды графиков нормального распределения случайной величины

Нормальное распределение часто встречается в природе. Например, следующие случайные величины хорошо моделируются нормальным распределением:

- отклонение при стрельбе;
- погрешности измерений (однако погрешности некоторых измерительных приборов имеют не нормальные распределения);
- рост человека;
- давление крови в течение дня;
- экзаменационные оценки;
- некоторые характеристики живых организмов в популяции [7].

Многие непрерывные случайные величины не являются ни точно, ни приближенно нормальными. Свойства таких величин довольно сильно отличаются от свойств нормального распределения, перечисленных выше (рисунок 4).

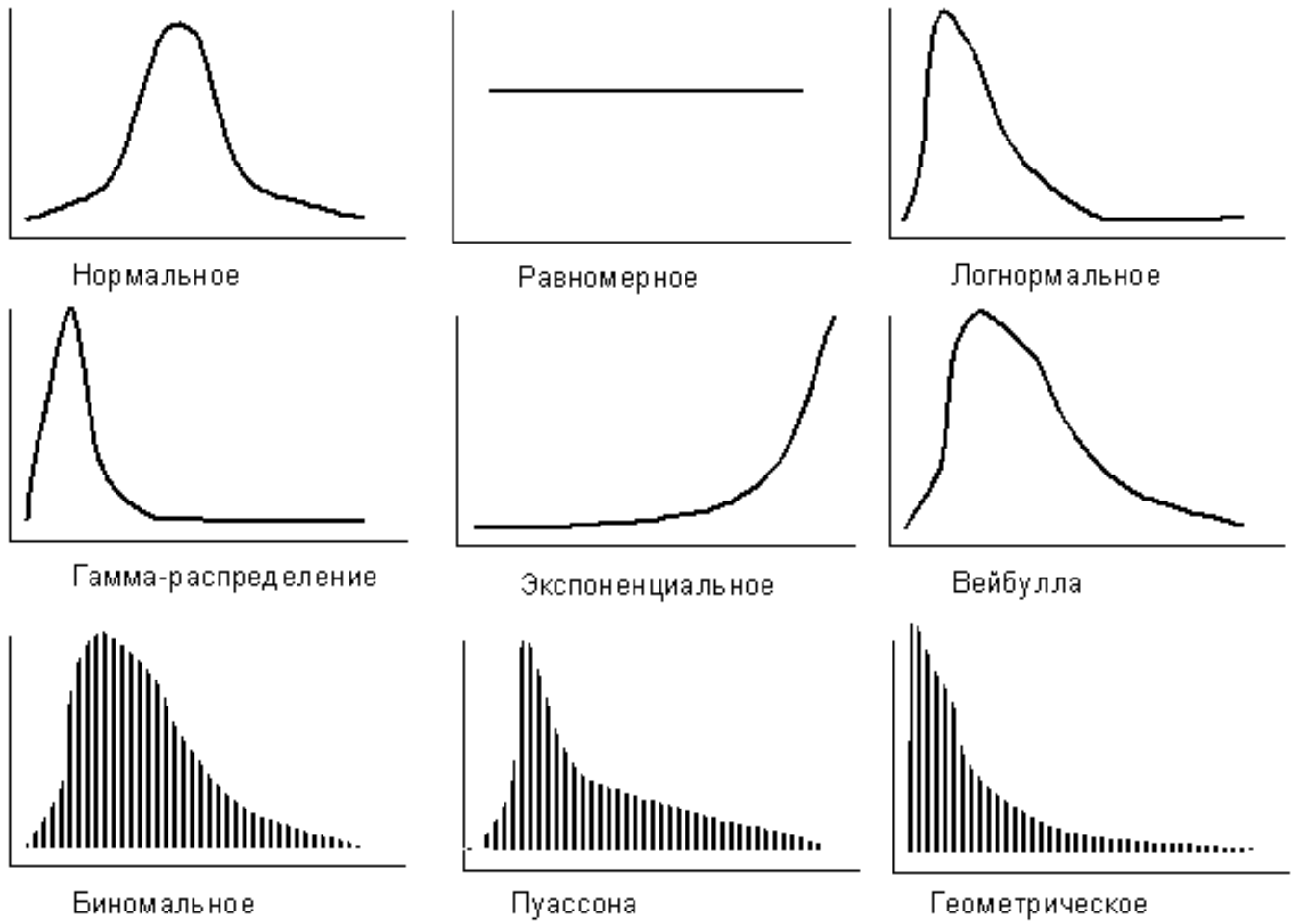


Рисунок 4 – Виды распределений величин

1.7 Равномерный закон распределения

На практике встречаются случайные величины, о которых заранее известно, что они могут принять какое-либо значение в строго определенных границах, причем в этих границах все значения случайной величины имеют одинаковую вероятность (обладают одной и той же плотностью вероятностей).

Например, при поломке часов остановившаяся минутная стрелка будет с одинаковой вероятностью (плотностью вероятности) показывать время, прошедшее от начала данного часа до поломки часов. Это время является случайной величиной, принимающей с одинаковой плотностью вероятности значения, которые не выходят

за границы, определенные продолжительностью одного часа. К подобным случайным величинам относится также и погрешность округления. Про такие величины говорят, что они распределены равномерно, т. е. имеют равномерное распределение [8].

Одним из наиболее часто встречающихся распределений является нормальное распределение. Оно играет большую роль в теории вероятностей и занимает среди других распределений особое положение. Нормальный закон распределения является предельным законом, к которому приближаются другие законы распределения при часто встречающихся аналогичных условиях.

Если предоставляется возможность рассматривать некоторую случайную величину как сумму достаточно большого числа других случайных величин, то данная случайная величина обычно подчиняется нормальному закону распределения. Суммируемые случайные величины могут подчиняться каким угодно распределениям, но при этом должно выполняться условие их независимости (или слабой зависимости). При соблюдении некоторых не очень жестких условий указанная сумма случайных величин подчиняется приближенно нормальному закону распределения и тем точнее, чем большее количество величин суммируется [9].

Ни одна из суммируемых случайных величин не должна резко отличаться от других, т. е. каждая из них должна играть в общей сумме примерно одинаковую роль и не иметь исключительно большую по сравнению с другими величинами дисперсию.

Для примера рассмотрим изготовление некоторой детали на станке-автомате. Размеры изготовленных деталей несколько отличаются от требуемых. Это отклонение размеров от стандарта вызывается различными причинами, которые более или менее независимы друг от друга. К ним могут относиться: неравномерный режим обработки детали; неоднородность обрабатываемого материала; неточность установки заготовки в станке; износ режущего инструмента и деталей станков; упругие деформаций узлов станка; состояние микроклимата в цехе; колебание напряжения в электросети и т. д. Каждая из перечисленных и подобных

им причин влияет на отклонение размера изготовляемой детали от стандарта. Таким образом, общее отклонение размера, фиксируемое измерительным прибором, является суммой большего числа отклонений, обусловленных различными причинами.

Если ни одна из этих причин не является доминирующей, то суммарное отклонение является случайной величиной, имеющей нормальный закон распределения.

1.8 Отсев грубых погрешностей

Предварительная обработка результатов измерений или наблюдений необходима для того, чтобы отсеять так называемые грубые промахи, убрать из рассмотрения статистически незначимые факторы и в дальнейшем использовать оставшиеся данные для построения адекватных моделей исследуемого процесса.

Грубые промахи могут быть вызваны ошибками измерительных приборов, субъективными ошибками исследователя, методом обработки данных, влиянием неучтенных случайных факторов, округлением при вычислениях, другими причинами.

Другим важным моментом предварительной обработки данных является проверка соответствия распределения результатов измерения закону нормального распределения. Если эта гипотеза неприемлема, то следует определить, какому закону распределения подчиняются опытные данные, и если это возможно, преобразовать данное распределение к нормальному.

Только после предварительного анализа и исключения грубых промахов можно использовать оставшиеся данные для получения правильных решений. Существуют различные рекомендации для проведения отсева грубых погрешностей наблюдения. Чащу всего грубые погрешности можно заметить на визуальных отображениях - графиках величин (рисунок 5).

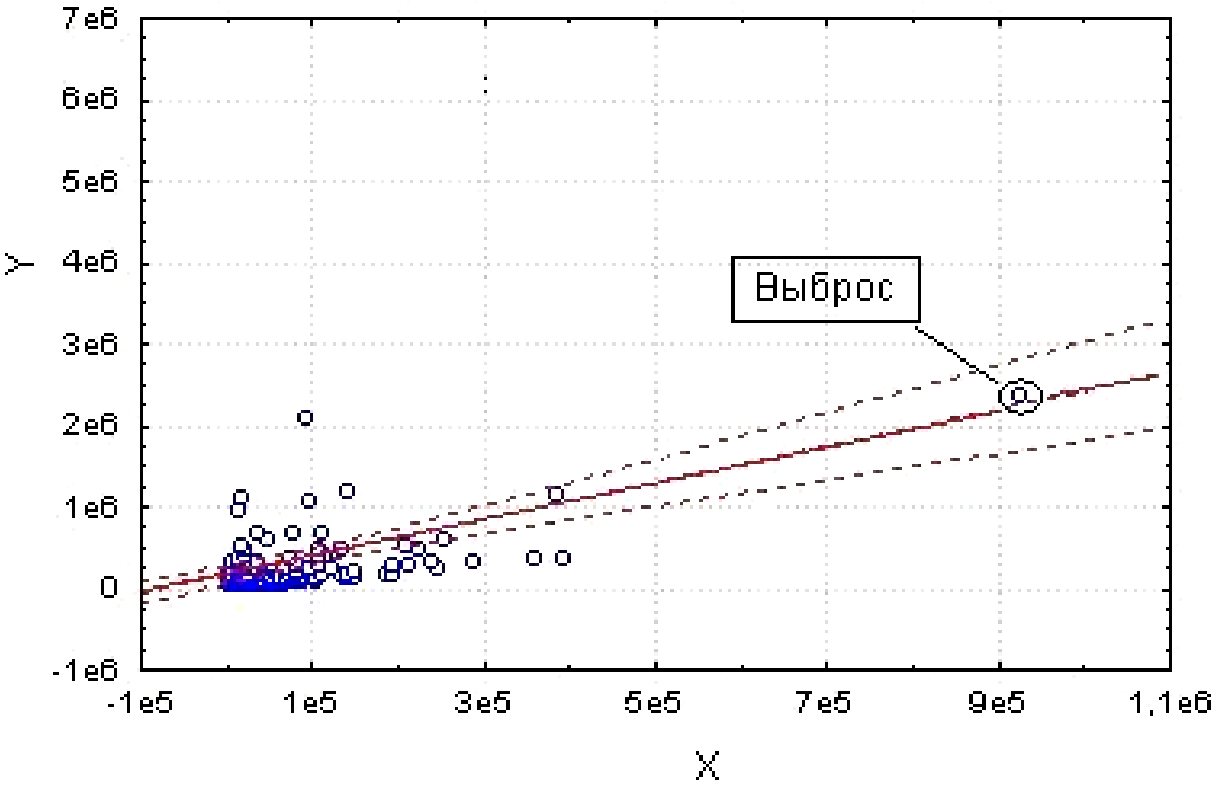


Рисунок 5 – Распределение случайной величины

1.9 Ошибки параллельных опытов

Эксперимент является основным и наиболее совершенным методом познания. Он может быть активным и пассивным. Осуществление пассивного эксперимента не зависит от экспериментатора, и ему приходится довольствоваться лишь ролью наблюдателя. Основной вид эксперимента – активный, проводится в контролируемых и управляемых условиях.

Все факторы, влияющие на исследуемые параметры объекта, предусмотреть, как правило, не удается. Так, в сложных системах, зависящих от множества факторов, некоторые воздействия не могут контролироваться или управляться. Воздействие этих факторов рассматриваются как белый шум, наложенный на истинные результаты эксперимента. Чтобы отделить факторы,

интересующие экспериментатора, от шумового фона, применяются специальные методы, называемые рандомизацией эксперимента.

Проведение активного эксперимента зачастую требует больших материальных затрат. Поэтому важной задачей является получение необходимых сведений при минимальном числе опытов. Решением этой проблемы занимается теория планирования эксперимента, представляющая собой раздел математической статистики.

Каждый эксперимент содержит элемент неопределенности вследствие ограниченности экспериментального материала. Постановка повторных (или параллельных) опытов не дает полностью совпадающих результатов, потому что всегда существует ошибка опыта (ошибка воспроизводимости). Эту ошибку и нужно оценить по параллельным опытам. Для этого опыт воспроизводится по возможности в одинаковых условиях несколько раз и затем берется среднее арифметическое всех результатов. Среднее арифметическое $\bar{y}$ равно сумме всех n отдельных результатов, деленной на количество параллельных опытов n

$$\bar{y} = \frac{y_1 + y_2 + ... + y_n}{n} = \frac{1}{n} \cdot \sum_{q-1}^{n} y_q \,. \qquad (2)$$

Отклонение результата любого опыта от среднего арифметического можно представить как разность $y_q - \bar{y}$, где y_q – результат отдельного опыта. Наличие отклонения свидетельствует об изменчивости, вариации значений повторных опытов. Для измерения этой изменчивости чаще всего используют дисперсию. Дисперсией называется среднее значение квадрата отклонений величины от ее среднего значения. Дисперсия обозначается s^2 и выражается формулой

$$s^2 = \frac{1}{n-1} \cdot \sum_{q-1}^{n} (\mathrm{y}_q - \bar{y})^2 \,, \qquad (3)$$

где $(n-1)$ – число степеней свободы, равное количеству опытов минус единица. Одна степень свободы использована для вычисления среднего.

Корень квадратный из дисперсии, взятый с положительным знаком, называется средним квадратическим отклонением, стандартом или квадратичной ошибкой

Стандарт имеет размерность той величины, для которой он вычислен. Дисперсия и стандарт – это меры рассеяния, изменчивости. Чем больше дисперсия и стандарт, тем больше рассеяны значения параллельных опытов около среднего значения.

Ошибка опыта являемся суммарной величиной, результатом многих ошибок: ошибок измерений факторов, ошибок измерений параметра оптимизации и др. Каждую из этих ошибок можно, в свою очередь, разделить на составляющие.

Вопрос о классификации ошибок довольно сложный и вызывает много дискуссий. В качестве примера одной из возможных схем классификации приведем схему (рисунок 6).

Все ошибки принято разделять на два класса: систематические и случайные.

Систематические ошибки порождаются причинами, действующими регулярно, в определенном направлении. Чаще всего эти ошибки можно изучить и определить количественно.

Систематические ошибки обусловливаются причинами, действующими вполне определённым образом. Примером систематической ошибки при взвешивании может являться смещение стрелки ненагруженных весов относительно нулевой отметки на некоторую постоянную величину. Зная это смещение (например, взвесив гирю, масса которой точно известна), можно, всякий раз измеряя массу на этих весах, вычитать из показаний прибора. Таким образом, систематические ошибки могут быть устранены или достаточно точно учтены.

Систематические ошибки находят, калибруя измерительные приборы и сопоставляя опытные данные с изменяющимися внешними условиями (например, при градуировке термопары по реперным точкам, при сравнении с эталонным прибором).

Если систематические ошибки вызываются внешними условиями (переменной температуры, сырья и т. д.), следует компенсировать их влияние. Как это делать, будет показано ниже.

Случайными ошибками называются те, которые появляются нерегулярно, причины возникновения которых неизвестны и которые невозможно учесть заранее.

Систематические и случайные ошибки состоят из множества элементарных ошибок. Для того, чтобы исключать инструментальные ошибки, следует проверять приборы перед опытом, иногда в течение опыта и обязательно после опыта. Ошибки при проведении самого опыта возникают вследствие неравномерного нагрева реакционной среды, разного способа перемешивания и т.п. При повторении опытов такие ошибки могут вызвать большой разброс экспериментальных результатов [10].

Очень важно исключить из экспериментальных данных грубые ошибки, так называемый брак при повторных опытах. Для отброса ошибочных опытов существуют правила. Для определения брака используют, например, критерий Стьюдента (таблица А.1, рисунок А.1)

$$\frac{y - \bar{y}}{s} \geq t . \qquad (4)$$

Значение t берут из таблицы t-распределения Стьюдента. Опыт считается бракованным, если экспериментальное значение критерия t по модулю больше табличного значения.

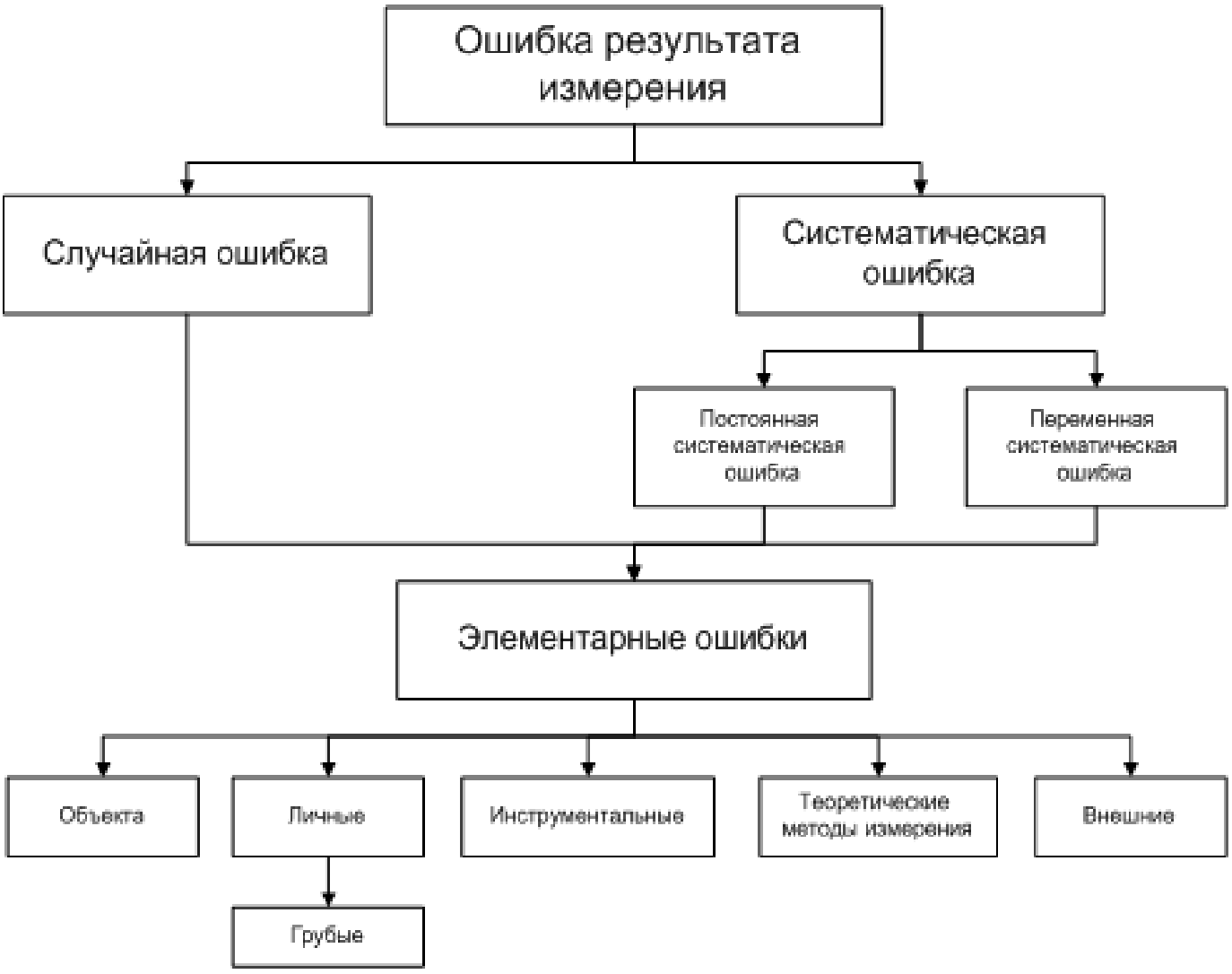


Рисунок 6 – Классификация ошибок измерений

1.10 Проверка гипотезы нормального распределения

Проверку нормальности распределения случайной величины можно производить разными способами, рассмотрим простую рекомендацию. Надо вычислить среднее абсолютное по обычной формуле

$$e = \frac{\sum_{1}^{n} |x_i - \bar{x}|}{n}. \qquad (5)$$

Если выполняется условие

$$\left| \frac{e}{S} - 0{,}7979 \right| < \frac{0{,}4}{\sqrt{n}}, \qquad (6)$$

то гипотеза нормальности распределения выборки данных принимается. В противном случае надо постараться преобразовать распределение к нормальному или применить другое распределение.

1.11 Преобразование распределений к нормальному

Если гипотеза нормальности распределения не может быть принята, то возможно, что с помощью существующих методов можно так преобразовать исходные данные, что их распределение будет подчиняться нормальному закону.

В самом начале преобразования данных большую помощь может оказать вид гистограммы. При крутой левой части и пологой правой, то есть при явной асимметрии путем, например, логарифмирования можно добиться симметричного распределения. При логарифмировании исходных данных левая ветвь кривой распределения сильно растягивается и распределение принимает приближенно нормальный характер.

Асимметричное распределение с одной вершиной часто приводится к нормальному преобразованием $x' = \ln(x \pm a)$. В некоторых случаях можно применять и другие преобразования: $x' = 1/x$, $x' = 1/\sqrt{x}$, $x' = ax^b$ и др. (a, b - некоторые постоянные).

Для нормализации смещенного право распределения служит тригонометрические преобразования. В этом и во многих других случаях универсальным является степенные преобразования $x' = ax^b$, причем b будет тем больше, чем больше выражено правое смещение.

1.12 Алгоритм предварительной обработки данных

Резюмируя вышесказанное, алгоритм предварительной обработки наблюдений сводится к следующему:

1) вычисление выборочных характеристик;
2) отсев грубых погрешностей;
3) проверка нормального закона;
4) преобразование распределения к нормальному в случае необходимости.

1.13 Задания для самостоятельной работы по теме «Нормальное распределение. Критерии нормальности»

1.13.1 Проверить на нормальное распределение ряд чисел 21, 89, 54, 25, 9, 10, 14, 47, 74, 21, 89, 54, 25, 9, 1, 47, 97.

1.13.2 Установить, соответствие критериям нормальности численного ряда 89, 54, 25, 9, 1, 47, 97, 74, 9, 10, 14, 32, 47, 20, 15, 14.

1.13.3 Построить диаграмму нормальности для ряда чисел 97, 74, 9, 10, 14, 32, 47, 54, 78, 12, 45, 56, 54.

2 Анализ результатов исследования

2.1 Характеристика видов связей между наблюдениями

На практике сама необходимость измерений большинства величин вызывается тем, что они не остаются постоянными, а изменяются в функции от изменения других величин. В этом случае целью проведения эксперимента является установление вида функциональной зависимости y = f(X). Для этого должны одновременно определяться как значения X, так и соответствующие им значения y, а задачей эксперимента является установление математической модели исследуемой зависимости. Фактически речь идет об установлении связи между двумя рядами наблюдений (измерений).

Связи в общем случае являются достаточно многообразными и сложными. Обычно выделяют следующие виды связей: функциональная (рисунок 7) и корреляционная (рисунок 8).

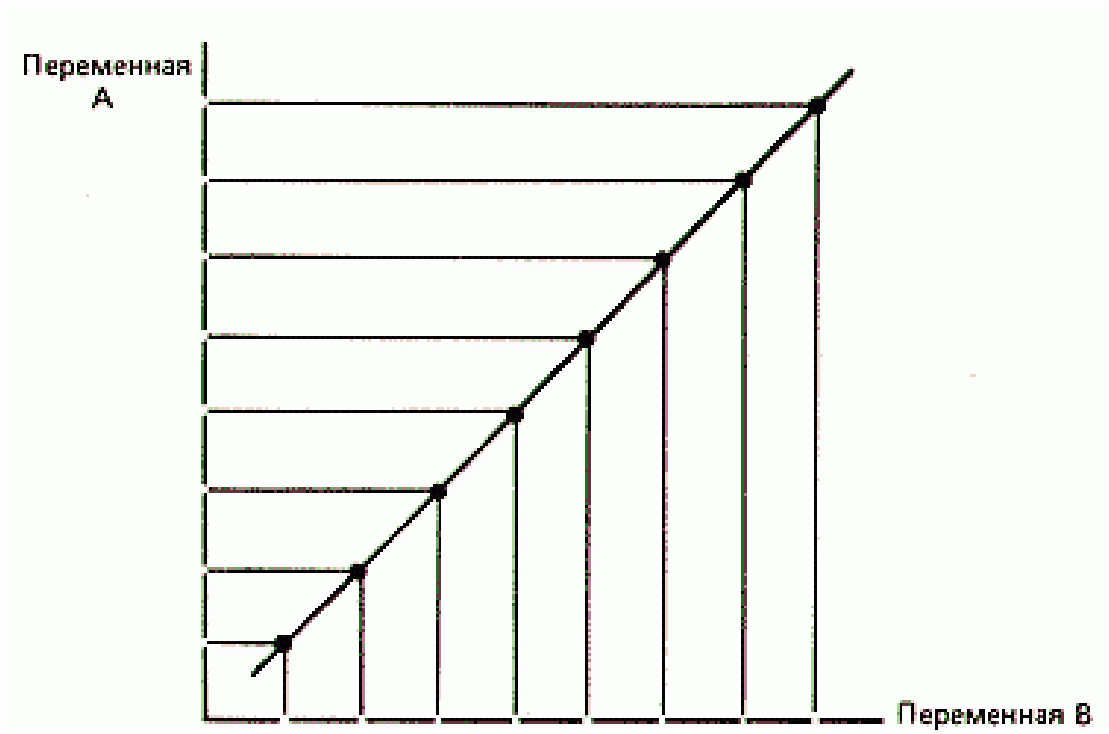


Рисунок 7 – Пример функциональной связи

Рисунок 8 – Пример корреляционной связи

2.2 Понятие корреляционного и регрессионного анализа

Основным аппаратом математической обработки результатов исследований является раздел математической статистики – корреляционно-регрессионный анализ.

Задача корреляционного анализа – выявление характера и степени взаимосвязи между показателями, являющимися случайными величинами.

Задача регрессионного анализа – выявление того, насколько изменение одной переменной (фактора) в среднем влияет на изменение другой переменной (результативного признака). В корреляционном анализе определяется один показатель, характеризующий степень тесноты взаимосвязи технологических показателей [11].

В регрессионном анализе строится модель регрессии в виде математической функции, которая показывает влияние факторов на некоторый технологический показатель.

Теоретически корреляция и регрессия связаны между собой. Рассмотрим виды зависимости между случайными величинами. Пусть имеется двумерная (многомерная) случайная величина, например X, У.

Зависимость между случайными величинами может быть следующих видов:

- функциональная – если значению случайной величины Х по определенному закону ставится в соответствие значение случайной величины У;

- статистическая (вероятностная) – если значению случайной величины Х ставится в соответствие определенное распределение случайной величины У.

Математическое ожидание У, определенное для каждого значения Х в вероятностной зависимости, называется условным математическим ожиданием.

Статистическая зависимость, в которой при изменении случайной величины Х изменяется условное математическое ожидание случайной величины У, называется корреляционной зависимостью. При этом, если условное математическое ожидание меняется по линейному закону, корреляционная зависимость называется линейной, если по нелинейному закону – нелинейной. Если условное математическое ожидание случайной величины У не изменяется при изменении значений Х, то корреляционной зависимости нет (т.е. любая корреляционная зависимость является статистической, но не всякая статистическая зависимость является корреляционной).

Функция, которая описывает закон изменения условного математического ожидания случайной величины У при изменении другой случайной величины Х, называется функцией регрессии Y на X. Если двумерная случайная величина (Х, У) распределена по нормальному закону, то функция регрессии линейная, корреляционная зависимость тоже линейная. Из вышесказанного понятна практическая интерпретация корреляционного и регрессионного анализа – выводы исследования можно интерпретировать лишь как некоторые усредненные свойства, обобщенные характеристики технологических процессов. Это необходимо иметь в виду тем, кто применяет корреляционный и регрессионный анализ.

Корреляционный и регрессионный анализ являются методами математической статистики – дисциплины, целью которой является получение научных и практических выводов относительно некоторых явлений в целом на

основе исследования корреляционной и регрессионной зависимости в выборке или временном ряду.

Любое конкретное исследование имеет дело с некоторыми выборочными данными из некоторой генеральной совокупности. Задача исследования состоит в том, чтобы на основе анализа выборки сделать выводы о свойствах всей совокупности. Эта задача решается с помощью специальных приемов математической статистики – проверки статистических гипотез. Принимается с определенной вероятностью гипотеза о наличии (отсутствии) данного свойства в генеральной совокупности. Эта гипотеза проверяется с помощью специального показателя (статистического критерия), который позволяет с заданной вероятностью ошибки сделать вывод о том, значимы ли значения выборочных показателей для выводов о наличии определенных свойств в генеральной совокупности или их значения случайно отличны от нуля и они не значимы.

Для проверки статистической значимости различных видов выборочных статистических показателей существуют тесты, использующие свои критерии, каждый из которых является случайной величиной, имеющей то или иное известное вероятностное распределение, зависящее от степеней свободы и выбранного уровня значимости (доверительной вероятности).

2.3 Корреляционный анализ

Корреляционный анализ – метод математической статистики, используемый для изучения, исследования взаимосвязи между (генеральными) экономическими показателями на основе их наблюдаемых статистических (выборочных) аналогов.

Парный корреляционный анализ – изучение взаимосвязи между двумя показателями, описывающими свойства однотипных объектов из некоторой совокупности.

Шаги корреляционного анализа:

1) постановка задачи.

2) сбор и анализ данных; определение формы корреляционной связи (линейная, криволинейная).

3) вычисление показателя тесноты корреляционной связи.

4) оценка статистической значимости показателя тесноты корреляционной связи.

Сбор данных. Сбор данных осуществляется методом случайной выборки некоторого количества наблюдаемых объектов из некоторой однородной совокупности, фиксации для каждого выбранного объекта пары признаков (свойств), взаимосвязь которых будет предметом исследования (рисунок 9).

Визуальный анализ данных. Визуальная оценка осуществляется на основе графического анализа. Данные на графике можно представить в виде точечной диаграммы в MS Excel. В результате такой оценки может быть сделана гипотеза о наличии линейной корреляционной связи, о нелинейной корреляционной связи или об отсутствии корреляционной связи.

	A	B	C	D
1	№	Y	X	
2	1	1,5	6,5	
3	2	3,5	6	
4	3	0,5	6,5	
5	4	1,5	4,5	
6	5	2	8	
7	6	1,5	5,5	
8	7	4,5	6,5	
9	8	3	6	
10	9	3,5	8	
11	10	3	4,5	
12				

Рисунок 9 – Таблица данных

Вычисление показателей тесноты корреляционной связи. Если визуальный анализ позволяет принять гипотезу о линейной форме связи между показателями – для оценки степени тесноты связи применяется линейный коэффициент корреляции r.

Границы измерения: $-1 \leq r \leq 1$.

Если взаимосвязь между показателями обратная (отрицательная) то корреляционная связь отрицательная: $-1 < r < 0$.

Если взаимосвязь между показателями прямая, то корреляционная зависимость положительная: $0 < r < 1$.

Если $r = 0$, линейная корреляционная зависимость отсутствует (рисунок 10).

В крайних случаях $|r| = 1$ имеется функциональная линейная зависимость между показателями x и y. [12]

Если визуальный анализ не позволяет принять гипотезу о линейной форме связи, коэффициент корреляции в этом случае применять неправомерно (рисунок 11).

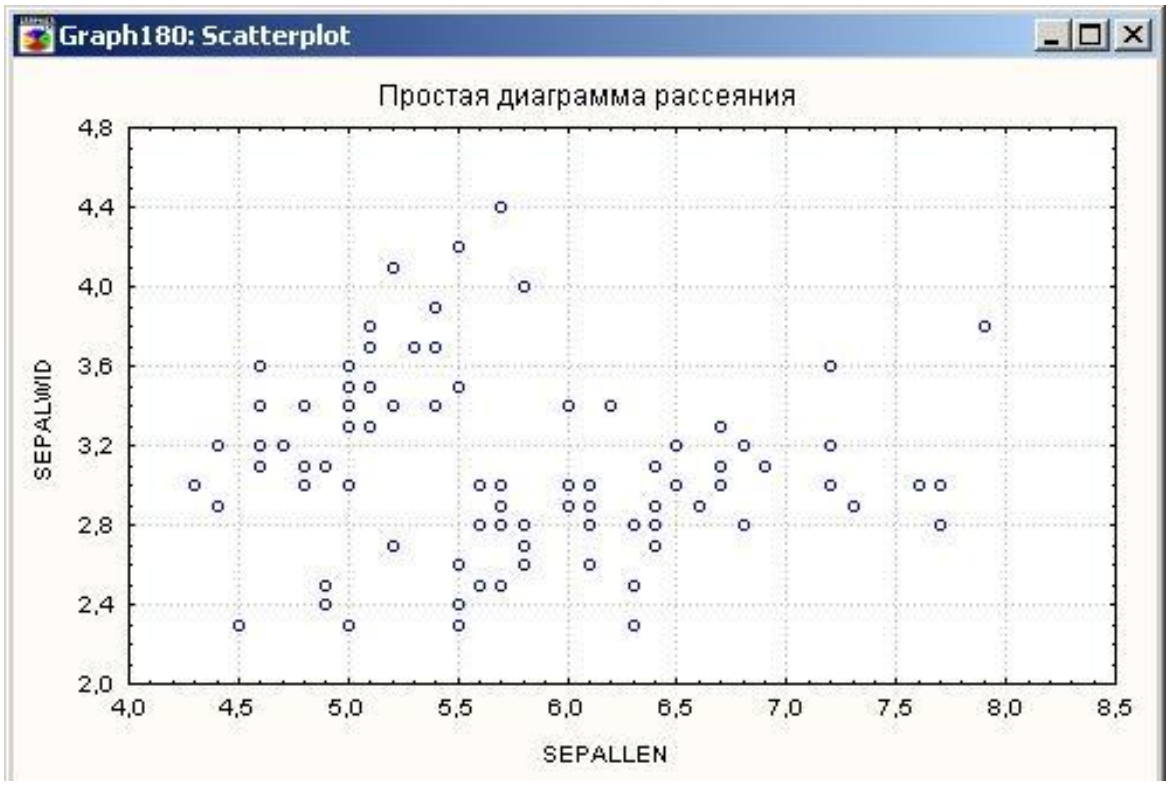


Рисунок 10 – Пример отсутствия зависимости

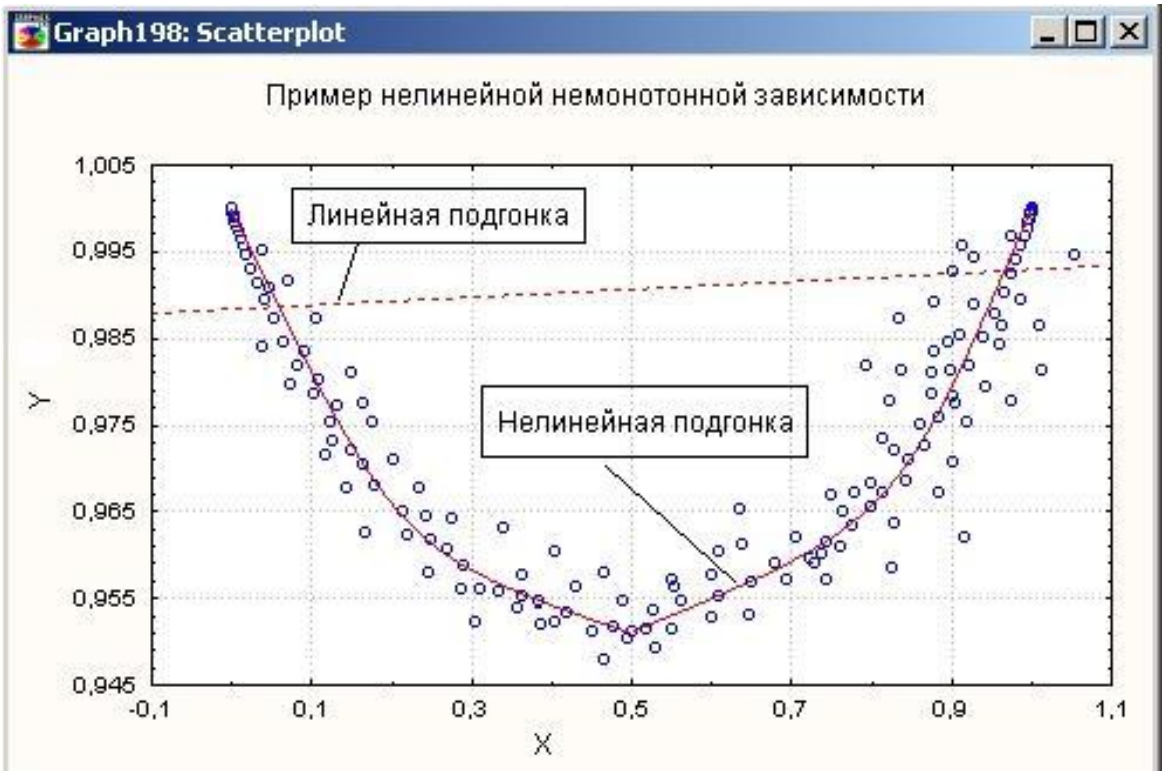


Рисунок 11 – Пример нелинейной корреляции

В некоторых случаях, несмотря на высокие значения коэффициента корреляции обнаруженная связь оказывается ложной (рисунок 12).

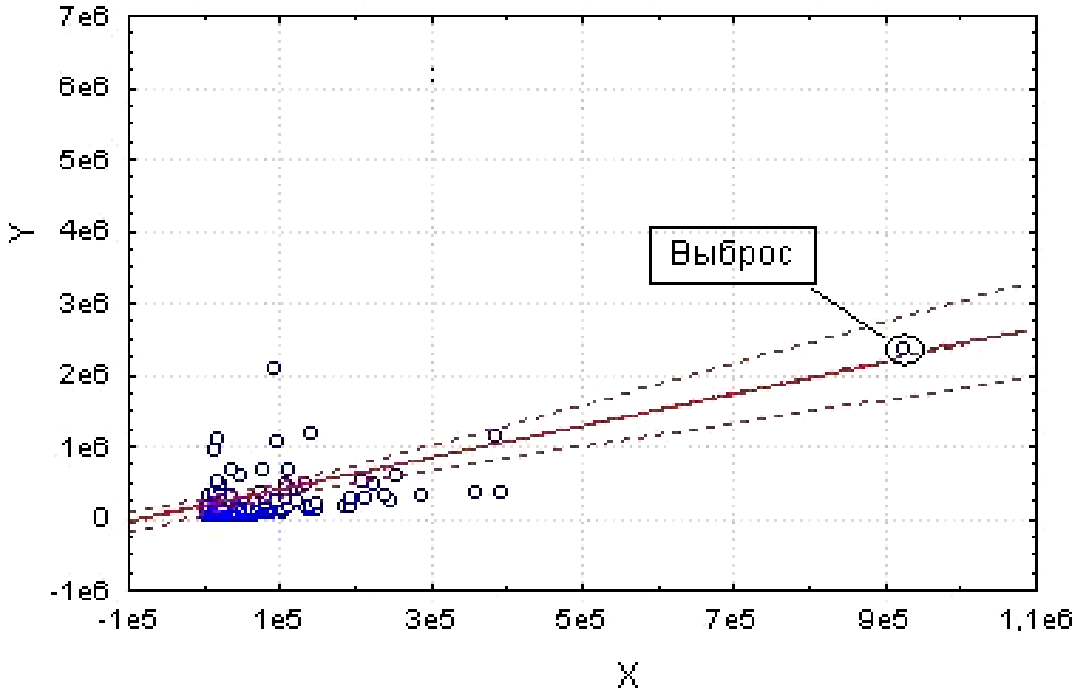


Рисунок 12 – Диаграмма сильной корреляции, обусловленной единственным выбросом (r = 0,8284)

В некоторых случаях, низкий коэффициент корреляции изначально ошибочен, и исследуемую группу образцов или явления следует разделить на две-три разные группы. Внутри каждой группы корреляция окажется существенной (рисунок 13).

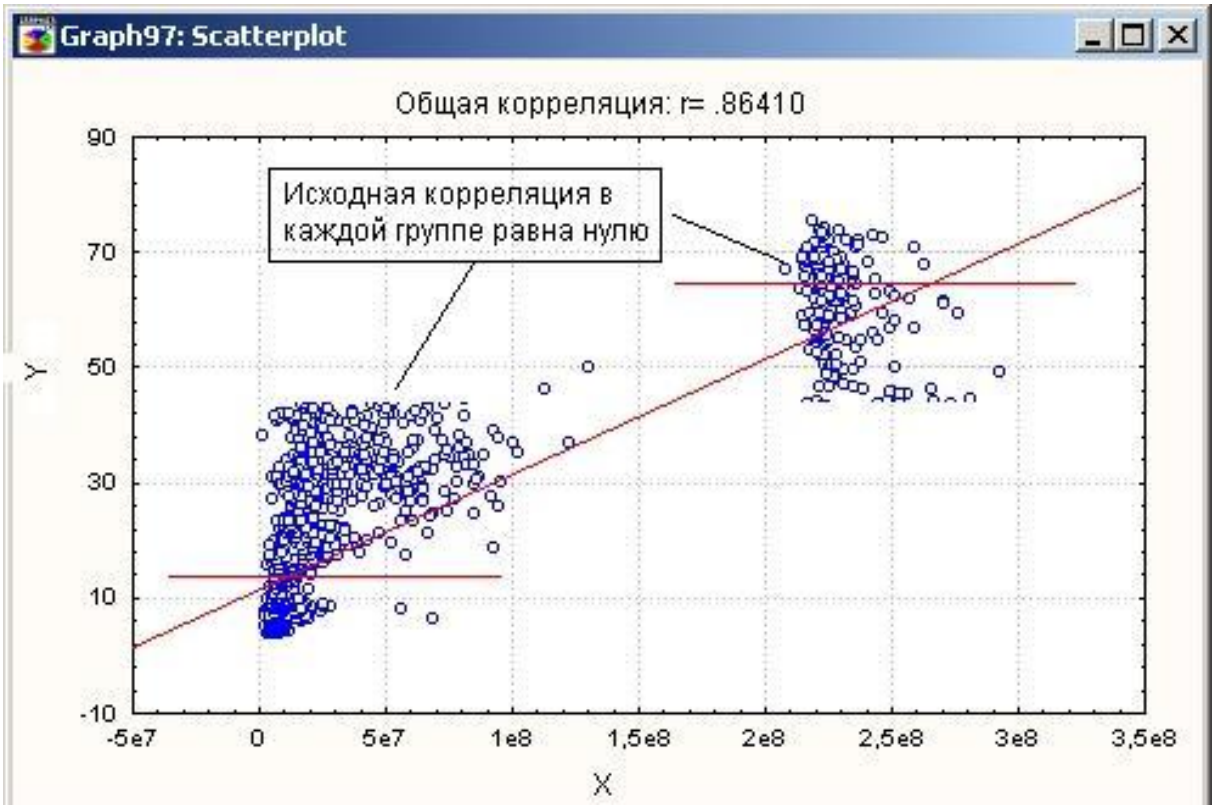


Рисунок 13 – Диаграмма данных, нуждающихся в группировании для оценки их корреляции

2.4 Интерпретация коэффициента корреляции

Как интерпретировать значение коэффициента корреляции Пирсона? Для оценки тесноты, или силы, корреляционной связи обычно используют общепринятые критерии, согласно которым абсолютные значения r < 0,3 свидетельствуют о слабой связи, значения r от 0,3 до 0,7 – о связи средней тесноты,

значения r > 0,7 – о сильной связи. Более точную оценку силы корреляционной связи можно получить, если воспользоваться таблицей Чеддока (таблица 2).

Таблица 2 – Таблица Чеддока для интерпретации силы корреляционной связи

Абсолютное значение r	Теснота (сила) корреляционной связи
менее 0,3	слабая
от 0,3 до 0,5	умеренная
от 0,5 до 0,7	заметная
от 0,7 до 0,9	высокая
более 0,9	весьма высокая

Другим способом оценки силы связи по коэффициенту корреляции является его сравнение с некоторым критическим, который в свою очередь зависит от количества измерений, опытов, наблюдений (таблица А.2).

Допустим, проводится независимое измерение различных параметров у одного типа объектов. Из этих данных можно получить качественно новую информацию - о взаимосвязи этих параметров. [13]

Например, измеряем рост и вес человека, каждое измерение представлено точкой в двумерном пространстве (рисунок 14). Несмотря на то, что величины носят случайный характер, в общем наблюдается некоторая зависимость - величины коррелируют.

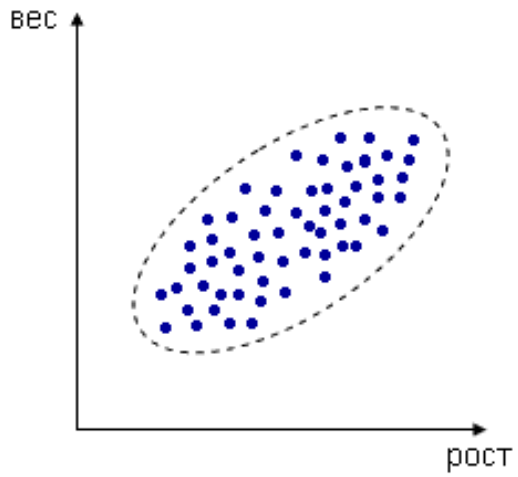


Рисунок 14 – Положительная корреляция

В данном случае это положительная корреляция (при увеличении одного параметра второй тоже увеличивается). Возможны также такие случаи (рисунки 15, 16).

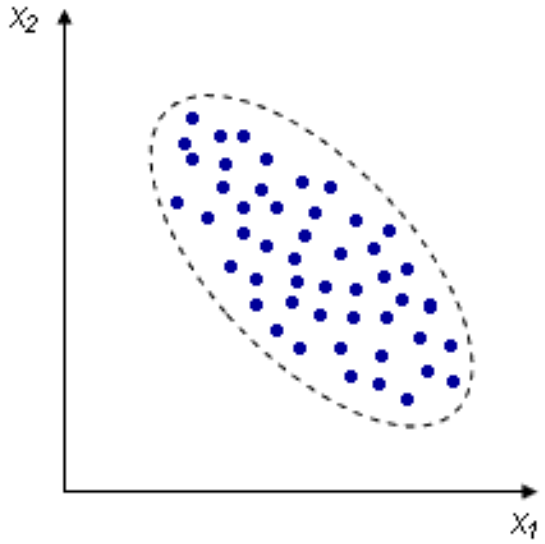


Рисунок 15 – Отрицательная корреляция

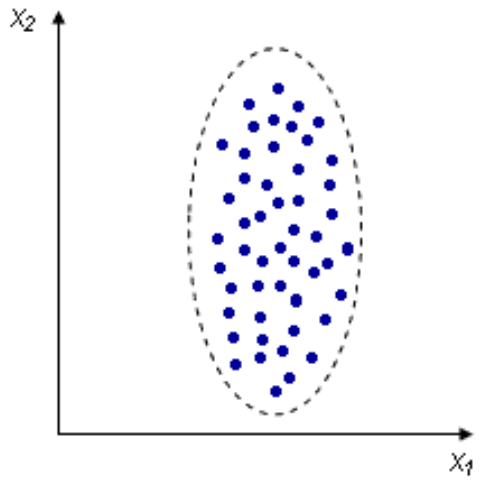


Рисунок 16 – Отсутствие корреляции

Взаимосвязь между переменными необходимо охарактеризовать численно, чтобы, например, различать такие случаи (рисунки 17, 18).

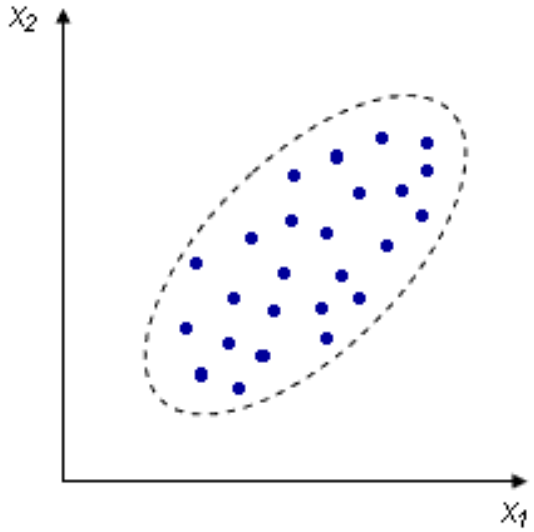


Рисунок 17 – Слабая связь

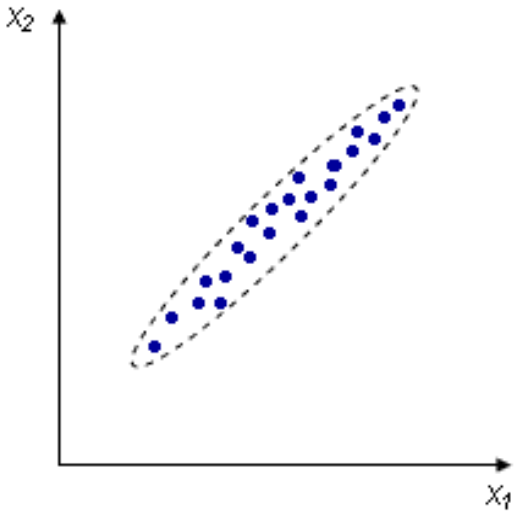


Рисунок 18 – Сильная связь

2.5 Оценка статистической значимости показателя корреляционной связи

Оценка статистической значимости линейного коэффициента корреляции проводится с помощью теста Стьюдента по t-статистике

$$t_r = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}. \quad (7)$$

Если значение коэффициента t будет больше или равно табличного значения (таблица А.1), то коэффициент корреляции статистически значим.

Таким образом, при определенном уровне значимости проверяется гипотеза о том, что в генеральной совокупности нет корреляционной зависимости между анализируемыми показателями.

2.6 Условия и ограничения применения критерия Пирсона

У критерия Пирсона существует несколько допустимых к использованию условий.

1. Сопоставляемые показатели должны быть измерены в количественной шкале (например, стекловидность, содержание белка).

2. Посредством критерия корреляции Пирсона можно определить лишь наличие и силу линейной взаимосвязи между величинами.

Прочие характеристики связи, в том числе направление (прямая или обратная), характер изменений (прямолинейный или криволинейный), а также наличие зависимости одной переменной от другой - определяются при помощи регрессионного анализа.

3. Количество сопоставляемых величин должно быть равно двум. В случае анализ взаимосвязи трех и более параметров следует воспользоваться методом факторного анализа.

4. Критерий корреляции Пирсона является параметрическим, в связи с чем условием его применения служит нормальное распределение сопоставляемых переменных. В случае необходимости корреляционного анализа показателей, распределение которых отличается от нормального, в том числе измеренных в порядковой шкале, следует использовать коэффициент ранговой корреляции Спирмена.

5. Следует четко различать понятия зависимости и корреляции. Зависимость величин обуславливает наличие корреляционной связи между ними, но не наоборот.

2.7 Пример расчета коэффициента корреляции Пирсона ручным способом

Одним из критериев, позволяющих определить, какова теснота (или сила) корреляционной связи между двумя показателями, измеренными в количественной шкале является критерий корреляции Пирсона.

Целью исследования явилось выявление, определение тесноты и статистической значимости корреляционной связи между двумя количественными показателями: газообразующей способностью муки ГОСМ (X) и показателя амилолитической активности муки («число падения» или ЧП) - (Y). Исходные данные для выборки, состоящей из 5 опытов (n = 5), сведены в таблице 3.

Таблица 3 – Исходные данные по результатам экспериментов

Газообразующая способность муки, мл CO2 на 100 г (X)	Число падения, % (Y)
951	83
874	76
957	84
1084	89
903	79

Вычислим суммы анализируемых значений X и Y

$$\Sigma(X) = 951 + 874 + 957 + 1084 + 903 = 4769,$$

$$\Sigma(Y) = 83 + 76 + 84 + 89 + 79 = 441.$$

Найдем средние арифметические для X и Y

$$Mx = \Sigma(X) / n = 4769 / 5 = 953,8,$$

$$My = \Sigma(Y) / n = 441 / 5 = 82,2.$$

Рассчитаем для каждого значения сопоставляемых показателей величину отклонения от среднего арифметического dx и dy (таблица 4).

Таблица 4 – Расчет отклонений значений экспериментов от среднего арифметического

Газообразующая способность муки, мл CO2 на 100 г (X)	Число падения, % (Y)	Отклонение ГОСМ от среднего значения (d_x)	Отклонение % ЧП от среднего значения (d_y)
951	83	-2,8	0,8
874	76	-79,8	-6,2
957	84	3,2	1,8
1084	89	130,2	6,8
903	79	-50,8	-3,2

Возведем в квадрат каждое значение отклонения d_x и d_y – таблица 5.

Таблица 5 – Расчеты промежуточных статистических данных

Газообразующая способность муки, мл CO2 на 100 г (X)	Число падения, % (Y)	Отклонение ГОСМ от среднего значения (d_x)	Отклонение % ЧП от среднего значения (d_y)	d_x^2	d_y^2
951	83	-2,8	0,8	7,84	0,64
874	76	-79,8	-6,2	6368,04	38,44
957	84	3,2	1,8	10,24	3,24
1084	89	130,2	6,8	16952,04	46,24
903	79	-50,8	-3,2	2580,64	10,24

Рассчитаем для каждой пары анализируемых значений произведение отклонений d_x и d_y (таблица 6).

Определим значения суммы квадратов отклонений $\sum d_x^2$ и $\sum d_y^2$

$$\sum d_x^2 = 25918{,}8,$$

$$\sum d_y^2 = 98{,}8.$$

Таблица 6 – Значения промежуточных статистических данных

Газообразующая способность муки, мл CO2 на 100 г (X)	Число падения, % (Y)	Отклонение ГОСМ от среднего значения (d_x)	Отклонение % ЧП от среднего значения (d_y)	d_x^2	d_y^2	$d_x \cdot d_y$
951	83	-2,8	0,8	7,84	0,64	-2,24
874	76	-79,8	-6,2	6368,04	38,44	494,76
957	84	3,2	1,8	10,24	3,24	5,76
1084	89	130,2	6,8	16952,04	46,24	885,36
903	79	-50,8	-3,2	2580,64	10,24	162,56

Найдем значение суммы произведений отклонений $\sum(\mathrm{d}_x \cdot \mathrm{d}_y)$

$$\sum(\mathrm{d}_x \cdot \mathrm{d}_y) = 1546,2$$

Рассчитаем значение коэффициента корреляции Пирсона r по приведенной выше формуле

$$r = \frac{\sum(\mathrm{d}_x \cdot \mathrm{d}_y)}{\sqrt{(\sum d_x^2 \cdot \sum d_y^2)}}, \quad (8)$$

$$r = \frac{1546,2}{\sqrt{25918,8 \cdot 98,8}} = 0,966.$$

Найдем значение t-критерия для оценки статистической значимости корреляционной связи

$$t_r = \frac{0,97 \cdot \sqrt{5-2}}{\sqrt{1-0,97^2}} = 7,0.$$

Критическое значение t-критерия найдем по таблице, где при числе степеней свободы f = n - 2 = 3 и уровне значимости p = 0,01 значение $t_{table} = 5,84$. Рассчитанное значение t_r (7,0) больше t_{table} (5,84), следовательно связь является статистически значимой.

Сделаем статистический вывод. Значение коэффициента корреляции Пирсона составило 0,97, что соответствует весьма высокой тесноте связи между газообразующей способностью муки и амилолитической активностью муки. Данная корреляционная связь является статистически значимой ($p < 0,01$).

2.8 Задания для самостоятельной работы по теме «Корреляционный анализ»

2.8.1 Определите отсутствие или наличие линейной корреляционной взаимосвязи между основными характеристиками зернового анализа зерна Оренбургской области. Для анализа используйте данные Оренбургского государственного центра агрохимической службы (таблица 7).

Этапы выполнения задания:

1) постройте точечную диаграмму и выдвинете гипотезу о характере связи между рассматриваемыми переменными;

2) рассчитайте коэффициент корреляции;

3) проверьте значимость коэффициента корреляции;

4) сделайте выводы.

Таблица 7 – Основные характеристики зернового анализа пшеницы

Сорт	Зона произрастания	Объем зерновки, куб.мм	Натура, г/л	Масса 1000 зерен, г	Стекловидность, %
Твердые сорта					
Харьковская 3	восток	22,60	810	31,1	88
	центр	22,63	807	31,5	85
	запад	22,57	825	30,2	85
Оренбургская 10	запад	22,50	820	30	98
	восток	22,13	811	29,4	96
	центр	22,57	775	29,6	94
Оренбургская 21	восток	21,62	830	29,9	95
	центр	22,83	802	29,4	98
	запад	21,55	784	29	93
Безенчукская Янтарь	восток	22,45	789	30,1	90
	центр	23,82	803	30,2	94
	запад	23,32	798	30	94
Безенчукская 200	восток	23,45	830	31,7	95
	центр	21,4	830	30,7	96
	запад	24,53	825	28,2	95
Степь 3	восток	21,65	755	29,3	96
	центр	20,5	750	29,0	95
	запад	20,1	723	28,4	90

Продолжение таблицы 7

Сорт	Зона произрастания	Объем зерновки, куб.мм	Натура, г/л	Масса 1000 зерен, г	Стекловидность, %
Мягкие сорта					
Юго-Восточная 3	восток	22,65	781	30,2	98
	центр	23,54	780	30,7	95
	запад	24,32	772	31,6	96
Учитель	восток	23,85	745	30,5	93
	запад	22,94	768	29,3	90
	центр	21,56	798	29	95
Варяг	восток	21,26	803	31,3	98
	центр	20,54	800	30,7	95
	запад	21,7	811	31,4	97
Оренбургская 13	восток	21,87	820	31,6	94
	центр	21,64	812	30,5	95
	запад	21,49	815	31,4	95
Прохор	восток	23,8	820	30,4	94
	центр	23,58	804	29,5	93
	запад	23,7	800	30,1	93
Л 503	восток	25,7	790	32,5	95
	центр	25,4	786	32	92
	запад	24,53	775	32,2	91
Саратовская 42	восток	24,21	710	30,8	85
	центр	23,82	745	30,3	87
	запад	24,58	750	29,9	88

2.8.2 Определите отсутствие или наличие линейной корреляционной взаимосвязи между зараженностью зерна пшеницы спорами «картофельной» палочки (показатель выражается в единицах КОЕ/г – колониеобразующих единиц бактерий в 1 грамме зерна) и тепло- и влагообеспеченностью района произрастания (показатели суммы температур и коэффициента атмосферного увлажнения), стекловидности зерна (выражена в %). Для анализа используйте данные Оренбургского гос. центра агрохимической службы (таблица 8).

Таблица 8 – Входные данные для корреляционного анализа

№ образца зерна	Сумма температур, ºС	Коэффициент атмосферного увлажнения	Обсемененность спорами «картофельной» палочки, КОЕ/г	Стекловидность, %
1	2270	0,22	400	98
2	2360	0,15	900	90
3	2350	0,25	500	67
4	2400	0,15	1100	69
5	2490	0,25	800	68
6	2610	0,15	1600	74
7	2670	0,22	1300	75
8	2665	0,12	2000	79
9	2730	0,16	1400	95
10	2770	0,12	2100	96
11	2790	0,25	1800	85
12	2770	0,15	2200	80
13	2900	0,25	1600	87
14	2970	0,15	2600	84
15	2240	0,25	380	89
16	2290	0,12	850	90
17	2450	0,25	510	95
18	2470	0,15	1090	96
19	2520	0,25	790	97
20	2540	0,15	1580	98
21	2580	0,22	1250	92
22	2640	0,12	1990	91
23	2630	0,15	1360	83
24	2675	0,12	2080	87
25	2850	0,22	1770	76
26	2900	0,12	2160	74

2.8.3 Провести статистическую обработку результатов антропометрических данных на примере группы людей.

2.8.4 Провести корреляционный анализ, используя в качестве исходных данных – значения церебрального индекса животных.

2.8.5 Провести корреляционный анализ, используя в качестве исходных данных – значения индекса массы группы людей.

2.8.6 Провести статистический анализ данных (вычислить описательные статистики, корреляции, экстраполяция - прогноз) на примере динамики курса валют за продолжительный период времени в связи с основными экономическими характеристиками.

2.9 Основы работы со статистическим пакетом StatSoft Statistica

Работа с популярным и очень мощным статистическим пакетом Statistica компании StatSoft Inc существенно облегчает и ускоряет проведение статистических анализов. В частности, с его помощью легко определить существование и силу связей между различными величинами [14].

Для того, чтобы рассчитывать корреляционные взаимосвязи методом попарной корреляции Пирсона, нужно проделать следующее.

1. Внести данные в программу «Statistica» (ручным способом или импортировать): пункт меню File – Open.

Для примера, нас интересует существуют ли связи между зараженностью зерна пшеницы спорами «картофельной» палочки (показатель выражается в единицах КОЕ/г – колониеобразующих единиц бактерий в 1 грамме зерна) и тепло- и влагообеспеченностью района произрастания этой пшеницы (показатели суммы температур и коэффициента атмосферного увлажнения). Кроме того добавлен такой априори незначительный показатель как день фазы Луны в момент посева зерна.

2. Выбрать пункт меню Statistics – Basic Statistics/Tables – Correlation matrices. Появится панель как на рисунке 19.

3. Нажимаем кнопку One variable list (один список переменных) и выбираем все переменные, нажав кнопочку Select all, а потом Ok. Таким образом мы включаем в анализ все наши переменные. При желании можно выбрать только некоторые из них, либо сгруппировать два списка переменных (кнопка Two lists...) для перекрестного определения корреляций только между этими переменными (рисунок 20).

4. На вкладке Options, в секции Display format for correlation matrices, отмечаем пункт выставленный по умолчанию Display simple matrix, с подсветкой достоверных уровней значимости p. В секции управления отсутствующими значениями «MD deletion» выбираем пункт попарной обработки Pairwise и жмем Summary. Попарная обработка означает, что программа, формируя выборку для анализа, не будет учитывать отсутствующие значения.

5. В результате мы получили корреляционную матрицу, представляющую собой таблицу в первом столбце и в первой строке которой перечислены все наши переменные, а на пересечении любых двух переменных указан соответствующий их взаимосвязи коэффициент корреляции. По диагонали матрицы идут коэффициенты корреляции переменной «сама с собой» равные 1. Например, можно отметить, что величина обсеменности спорами «картофельной» палочки зерна прямо коррелирует с суммой температур в месте произрастания пшеницы, с коэффициентом корреляции $r=0,87$ и $p<0,05$ (коэффициенты корреляции с уровнем значимости $p<0,05$ выделены в таблице красным цветом) – рисунок 21. Корреляция обсеменности спорами «картофельной» палочки зерна с атмосферным увлажнением – сильная отрицательная (-0,57). Связь обсемененности с фазами Луны незначительная, т.е. ее нет (-0,24).

6. Мы можем сохранить полученную таблицу в экселевский файл. Для этого кликнем правой кнопкой мыши по пункту Correlations (test) в левом древовидном списке Рабочей книги программы (Workbook1), выберем в появившемся меню пункт Extract as stand-alone windows – Original (рисунок 21), означающий, что мы желаем

экспортировать оригинальное окно расчета, а не его копию. Программа экстрагирует выходную таблицу, которую уже можно сохранить обычным способом, через главное меню File – Save as и выбрав тип файлов xls (рисунок 22). Далее будет предложено сохранить идентификаторы полей, т.е. вычленить строки и столбцы, которые в экселевской таблице будут заголовками полей. Отмечаем флажками оба пункта Put case… и Put variable… и жмём Ok.

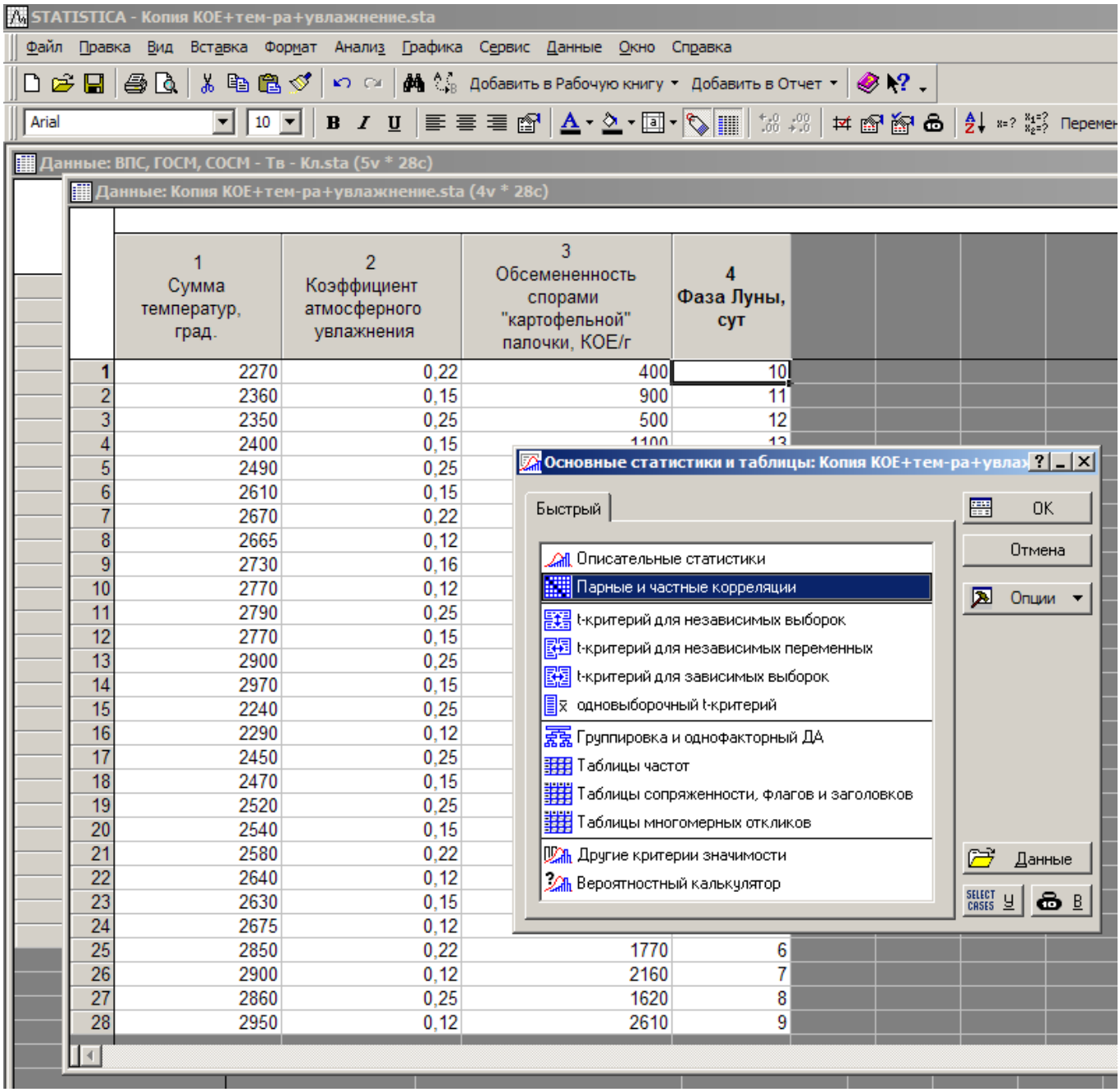


Рисунок 19 – Панель выбора описательных статистик в программе Statistica

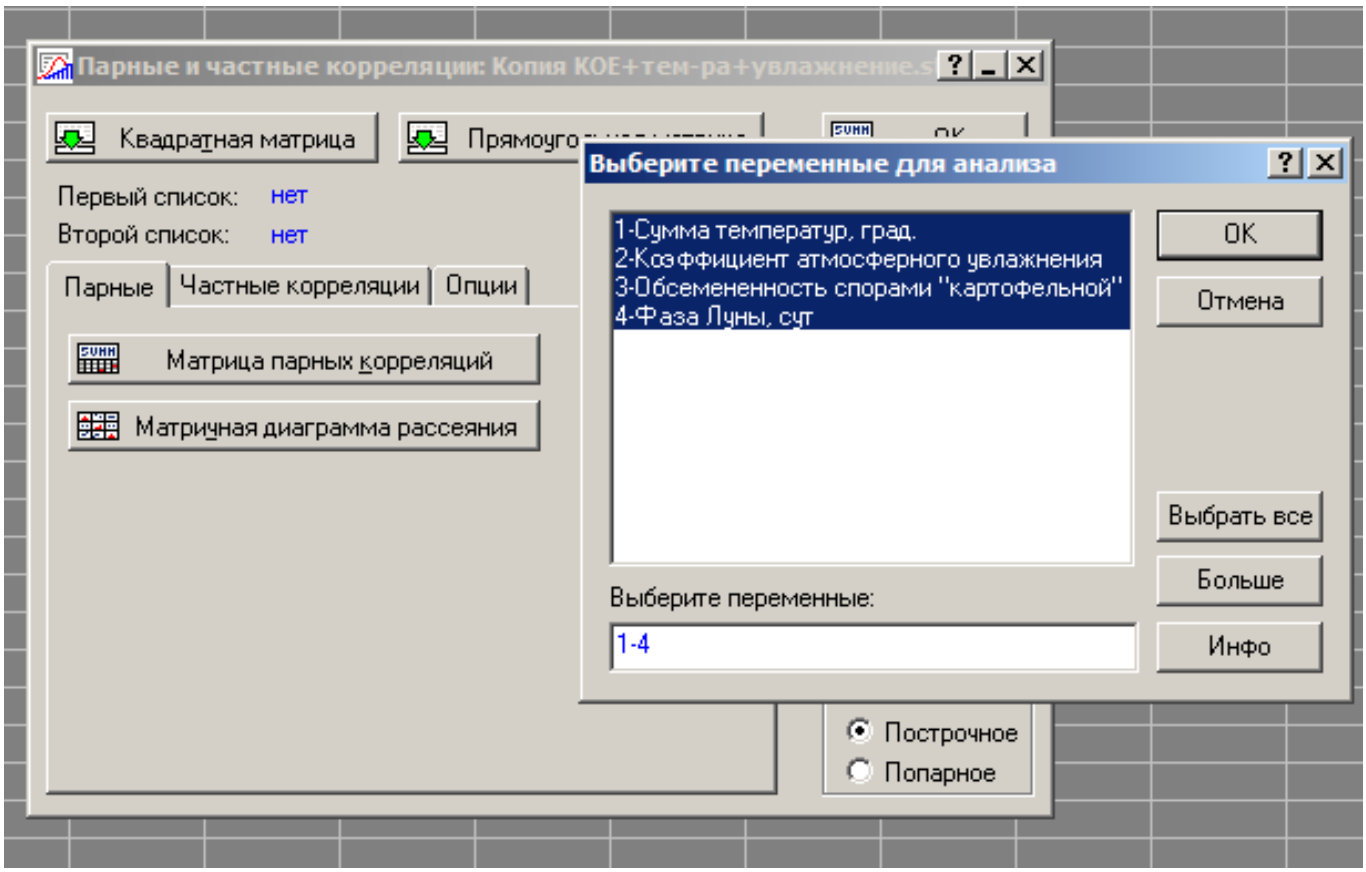
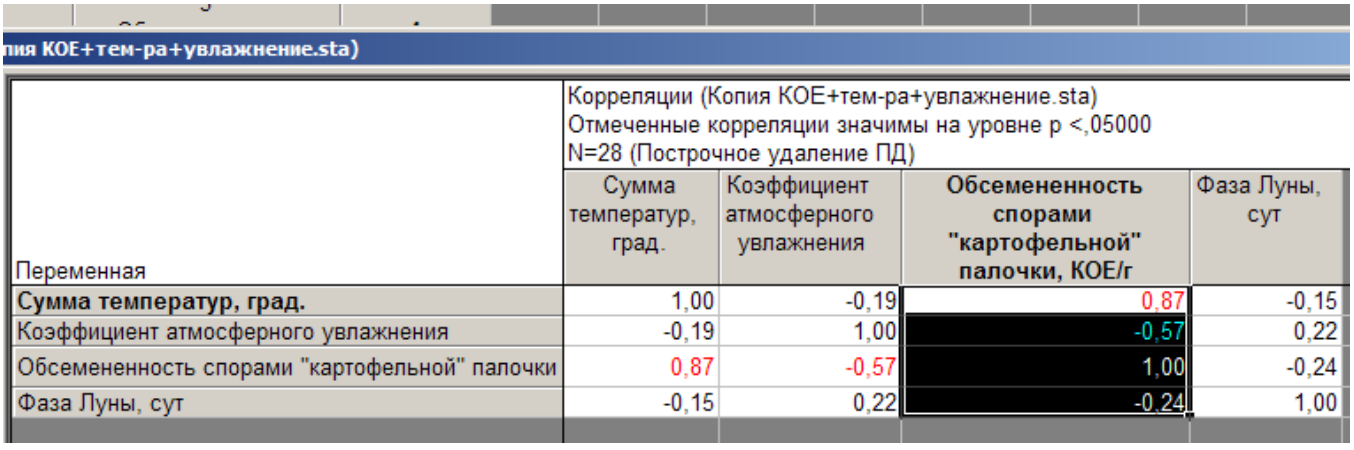


Рисунок 20 – Панель выбора переменных для проведения корреляционного анализа в программе Statistica



пия КОЕ+тем-ра+увлажнение.sta)

Корреляции (Копия КОЕ+тем-ра+увлажнение.sta)
Отмеченные корреляции значимы на уровне p <,05000
N=28 (Построчное удаление ПД)

Переменная	Сумма температур, град.	Коэффициент атмосферного увлажнения	**Обсемененность спорами "картофельной" палочки, КОЕ/г**	Фаза Луны, сут
Сумма температур, град.	1,00	-0,19	0,87	-0,15
Коэффициент атмосферного увлажнения	-0,19	1,00	-0,57	0,22
Обсемененность спорами "картофельной" палочки	0,87	-0,57	1,00	-0,24
Фаза Луны, сут	-0,15	0,22	-0,24	1,00

Рисунок 21 – Результат корреляционного анализа

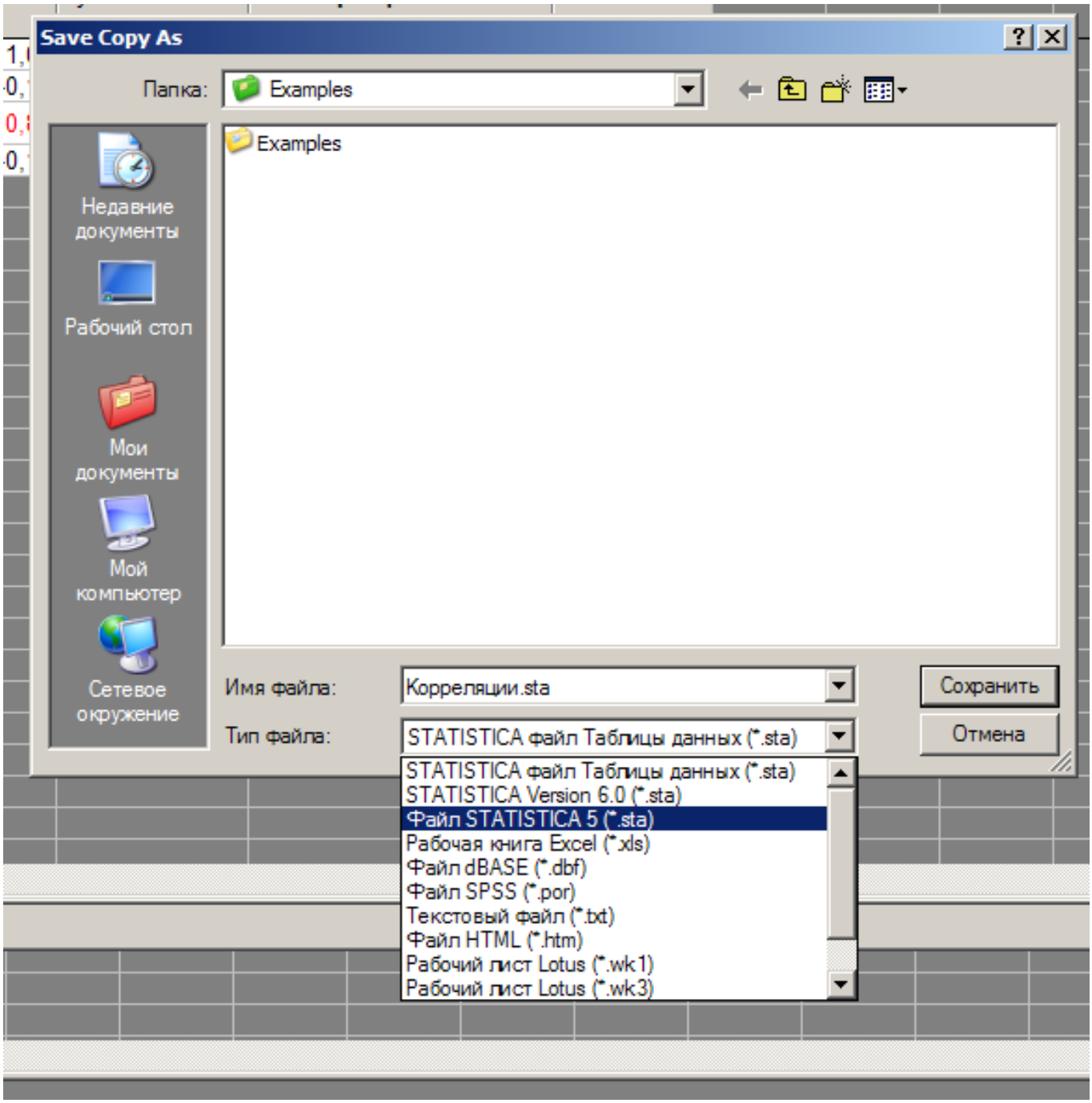


Рисунок 22 – Сохранение результата анализа в файл Excel

В результате получаем корреляционную матрицу, пример оформления такой матрицы в таблице 9.

Таблица 9 – Связь белково-протеиназного комплекса с показателями качества зерна яровой пшеницы (критическое значение коэффициента корреляции r = 0,49, число образцов 12)

Показатели	Альбумин	Глобулин	Глиадин	Глютенин	Отношение глиадина к глютенину	Протеиназа	Твердозерность	ГОСМ	ВПС
Альбумин	-	0,32	0,29	0,42	-0,04	0,72*	0,49*	-0,17	-0,22
Глобулин	0,32	-	0,61*	0,16	0,47	0,36	0,48	-0,29	-0,65*
Глиадин	0,29	0,61*	-	0,27	0,73*	0,29	0,75*	0,48	-0,80*
Глютенин	0,42	0,16	0,27	-	-0,46	-0,13	-0,20	0,33	-0,38
Отношение глиадина к глютенину	-0,04	0,47	0,73*	-0,46	-	0,35	0,82*	-0,68*	-0,46
Протеиназа	0,72*	0,36	0,29	-0,13	0,35	-	0,67*	-0,42	-0,07
ЧП	0,24	-0,05	0,17	0,05	0,10	0,49*	0,59*	-0,44	0,18
Твердозерность	0,49*	0,28	0,75*	-0,20	0,82*	0,67*	-	0,39	0,58*
ГОСМ	-0,17	-0,29	-0,48	0,33	-0,68*	-0,42	0,39	-	0,29
ВПС	-0,22	-0,65*	-0,80*	-0,38	-0,46	-0,07	0,58*	0,29	-
Объемный выход хлеба	-0,52*	-0,24	0,58*	-0,25	0,89*	-0,51*	-0,24	0,62*	0,37
* - коэффициент корреляции является существенным при уровне значимости α = 0,05									

2.10 Закон больших чисел

Поскольку корреляционная связь является статистической, первым условием возможности ее изучения является общее условие всякого статистического исследования: наличие данных по достаточно большой совокупности явлений. По

отдельным явлениям можно получить совершенно превратное представление о связи признаков, ибо в каждом отдельном явлении значения признаков кроме закономерной составляющей имеют случайное отклонение (вариацию). Например, сравнивая два хозяйства, одно из которых имеет лучшее качество почв, по уровню урожайности, можно обнаружить, что урожайность выше в хозяйстве с худшими почвами. Ведь урожайность зависит от сотен факторов и при том же самом качестве почв может быть и выше, и ниже. Но если сравнивать большое число хозяйств с лучшими почвами и большое число - с худшими, то средняя урожайность в первой группе окажется выше и станет возможным измерить достаточно точно параметры корреляционной связи.

Какое именно число явлений достаточно для анализа корреляционной и вообще статистической связи, зависит от цели анализа, требуемой точности и надежности параметров связи, от числа факторов, корреляция с которыми изучается. Обычно считают, что число наблюдений должно быть не менее чем в 5, а лучше - не менее чем в 10 раз больше числа факторов. Еще лучше, если число наблюдений в несколько десятков или в сотни раз больше числа факторов, тогда закон больших чисел, действуя в полную силу, обеспечивает эффективное взаимопогашение случайных отклонений от закономерного характера связи признаков [15].

2.11 Регрессионный анализ

Общее назначение множественной регрессии (этот термин был впервые использован в работе Пирсона) состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной.

В общественных и естественных науках процедуры множественной регрессии чрезвычайно широко используются в исследованиях. Множественная регрессия позволяет исследователю задать вопрос (и, вероятно, получить ответ) о том, «что

является лучшим предиктором для...». Например, исследователь в области образования мог бы пожелать узнать, какие факторы являются лучшими предикторами успешной учебы в университете. А психолога мог быть заинтересовать вопрос, какие индивидуальные качества позволяют лучше предсказать степень социальной адаптации индивида. Заметим, что термин «множественная» указывает на наличие нескольких предикторов или регрессоров, которые используются в модели.

На диаграмме рассеяния имеется независимая переменная или переменная X и зависимая переменная Y. Эти переменные могут, например, представлять возраст детей и показатели их роста, соответственно. Каждая точка на диаграмме представляет данные одного ребенка, т.е. его соответствующие показатели Возраст (лет) и Рост (см). Целью процедур линейной регрессии является подгонка прямой линии по точкам. А именно, алгоритм строит линию регрессии так, чтобы минимизировать квадраты отклонений этой линии от наблюдаемых точек. Поэтому на эту общую процедуру иногда ссылаются как на оценивание по методу наименьших квадратов.

В итоге, прямая линия на плоскости (в пространстве двух измерений) задается уравнением

$$Y = a + b \cdot X \ . \qquad (9)$$

Переменная Y (Рост) может быть выражена через некоторую константу a и угловой коэффициент b, умноженный на переменную X (Возраст).

Константу иногда называют также свободным членом, а угловой коэффициент – коэффициентом регрессии или B-коэффициентом (рисунок 23).

Коэффициент регрессии - абсолютная величина, на которую в среднем изменяется величина одного признака при изменении другого связанного с ним признака на единицу.

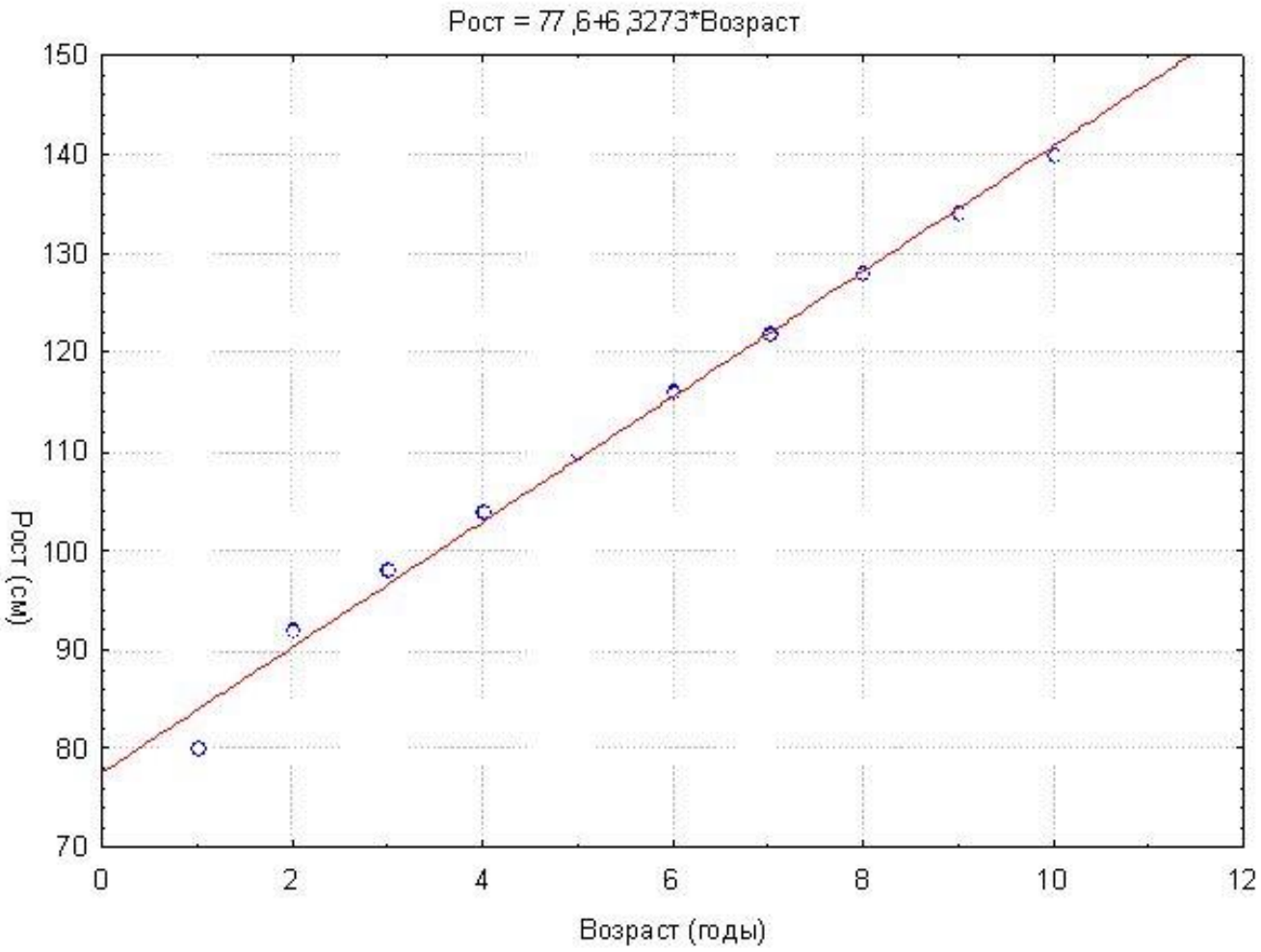


Рисунок 23 – График связи роста X и возраста Y

Формула расчета коэффициента регрессии следующая

$$R_{xy} = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}\ , \qquad (10)$$

где R_{xy} - коэффициент регрессии;

r_{xy} - коэффициент корреляции между признаками X и Y;

σ_x и σ_y - среднеквадратические отклонения признаков X и Y.

Характеристику меры разнообразия результативного признака Y показывает сигма регрессии

$$\sigma R_{xy} = \sigma_y \cdot \sqrt{1 - r_{xy}^2} \ , \qquad (11)$$

где σR_{xy} - сигма (среднеквадратическое отклонение) регрессии;

σ_y - среднеквадратическое отклонение признака Y;

r_{xy} - коэффициент корреляции между признаками X и Y.

В приведенном на рисунке примере коэффициент регрессии равен 6,3272, это означает, что при изменении Возраста на 1 год, Рост в среднем увеличивается на 6,3272 см.

В многомерном случае, когда имеется более одной независимой переменной, линия регрессии не может быть отображена в двумерном пространстве, однако она также может быть легко оценена в виде множественной регрессии. Например, если в дополнение к Возрасту вы имеете другие предикторы Роста (Совокупный рост родителей, Качество питания и т.д.), вы можете построить линейное уравнение, содержащее все эти переменные. Тогда, в общем случае, процедуры множественной регрессии будут оценивать параметры линейного уравнения вида

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_p \cdot X_p . \qquad (12)$$

Регрессионные коэффициенты (или B-коэффициенты) представляют независимые вклады каждой независимой переменной в предсказание зависимой переменной. Другими словами, переменная X_1, к примеру, коррелирует с переменной Y после учета влияния всех других независимых переменных. Этот тип корреляции упоминается также под названием частной корреляции.

Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем, очевидно, лучше прогноз. Например, если связь между переменными X и Y отсутствует, то отношение остаточной изменчивости переменной Y к исходной дисперсии равно 1. Если X и Y жестко связаны, то остаточная изменчивость отсутствует, и отношение дисперсий будет равно 0. В большинстве случаев отношение будет лежать где-то между этими экстремальными

значениями, т.е. между 0 и 1. Единица минус это отношение называется R-квадратом или коэффициентом детерминации. Это значение непосредственно интерпретируется следующим образом. Если имеется R-квадрат равный 0,4, то изменчивость значений переменной Y около линии регрессии составляет (1-0,4) от исходной дисперсии; другими словами, 40 % от исходной изменчивости могут быть объяснены, а 60 % остаточной изменчивости остаются необъясненными. В идеале желательно иметь объяснение если не для всей, то хотя бы для большей части исходной изменчивости. Значение R-квадрата является индикатором степени подгонки модели к данным (значение R-квадрата близкое к 1 показывает, что модель объясняет почти всю изменчивость соответствующих переменных).

Обычно, степень зависимости двух или более предикторов (независимых переменных или переменных X) с зависимой переменной (Y) выражается с помощью коэффициента множественной корреляции R. По определению он равен корню квадратному из коэффициента детерминации.

Это неотрицательная величина, принимающая значения между 0 и 1. Для интерпретации направления связи между переменными смотрят на знаки (плюс или минус) регрессионных коэффициентов или B-коэффициентов.

Если B-коэффициент положителен, то связь этой переменной с зависимой переменной положительна; если B-коэффициент отрицателен, то и связь носит отрицательный характер. Конечно, если B-коэффициент равен 0, связь между переменными отсутствует.

2.12 Описание данных и постановка задачи для построения регрессионной модели

В качестве исходных данных будем использовать данные таблицы 17 (Входные данные для корреляционного анализа). Исходная таблица содержит 4 переменных и 28 наблюдения. Список переменных приведен в таблице 10.

Таблица 10 – Список переменных для мат. обработки

Порядковый номер переменной	Переменные
1	Сумма температур, град. С
2	Коэффициент атмосферного увлажнения
3	Обсемененность зерна пшеницы спорами «картофельной» палочки, КОЕ/г
4	Стекловидность, %

Наша цель - построить регрессионную модель для переменной №3 «Обсемененность зерна».

Этапы решения:

1) Сначала проведем разведочный анализ имеющихся данных на предмет выбросов и незначимых данных (построение линейных графиков и диаграмм рассеяния).

2) Проверим наличие возможных зависимостей между наблюдениями и между переменными (построение корреляционных матриц).

3) Если наблюдения будут образовывать группы, то для каждой группы построим регрессионную модель для переменной «Обсемененность зерна» (множественная регрессия).

Перенумеруем переменные по порядку в таблице. Зависимой переменной (отклик) будем называть переменную «Обсемененность зерна». Независимыми (предикторами) назовем все остальные переменные.

2.13 Пошаговое построение регрессионной модели с помощью программных средств

Шаг 1. Диаграммы рассеяния (рисунок 24, 25) явных выбросов не выявили. В то же время, на многих графиках явно просматривается линейная зависимость.

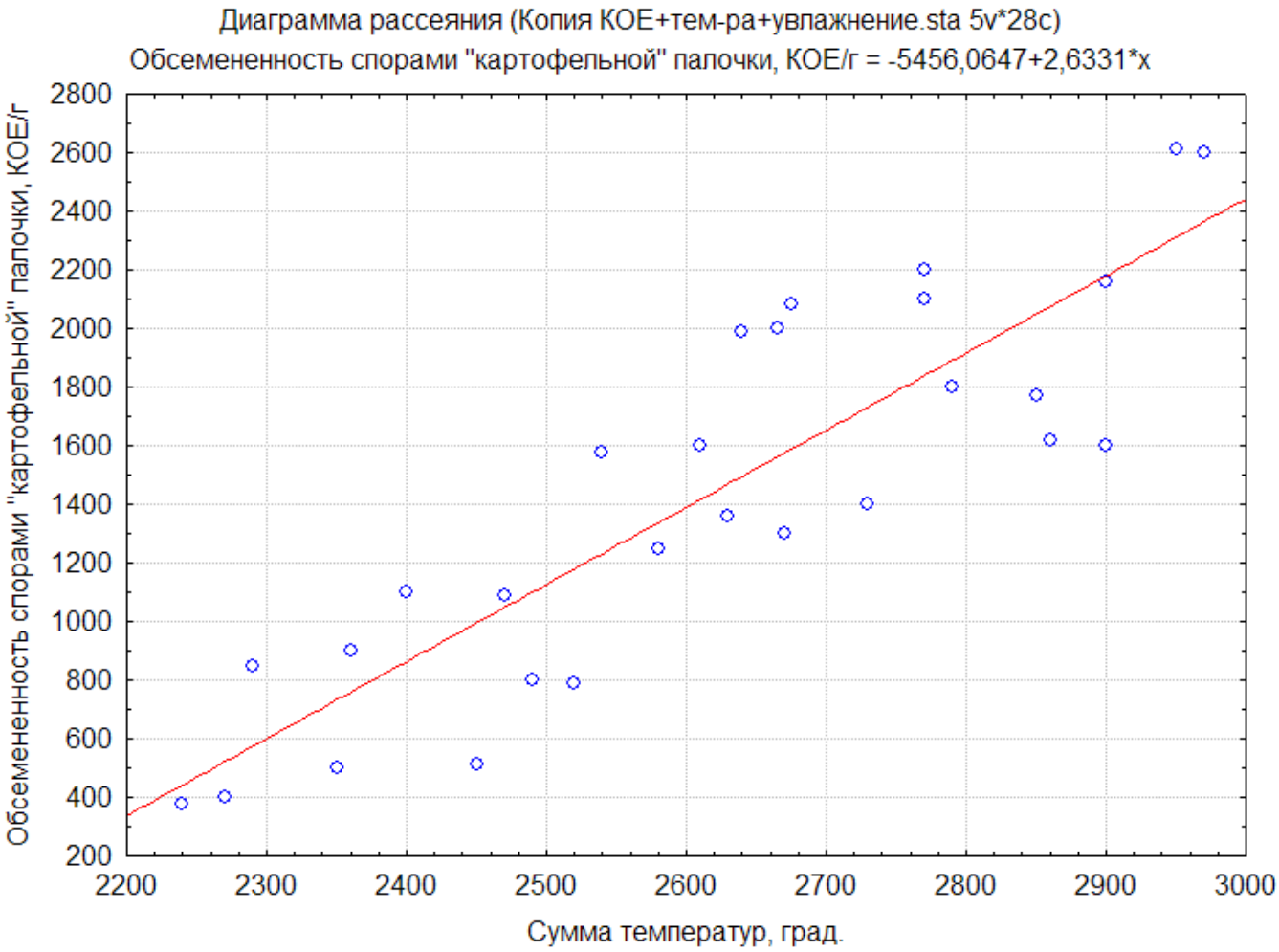


Рисунок 24 - Диаграммы рассеяния - зависимость показателя № 3 от показателя № 1

Понятно, что переменная № 2, скорее всего, войдет в регрессионную модель, т.к. нашей задачей является выявление именно линейной зависимости между предикторами и откликом.

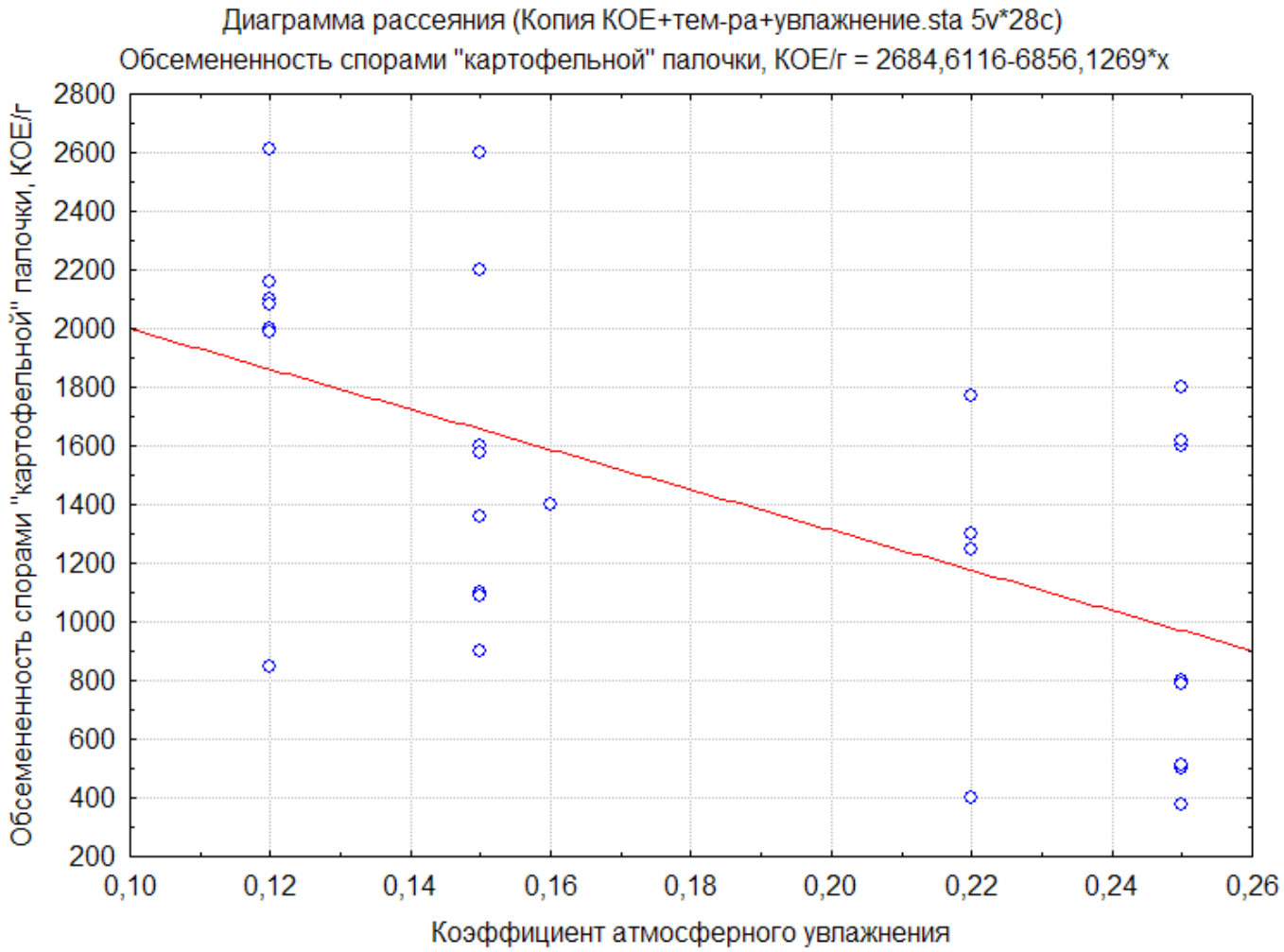


Рисунок 25 - Диаграммы рассеяния - зависимость показателя № 3 от показателя № 2

Шаг 2. Построим матрицу корреляций для всех переменных, чтобы проверить предположение относительно линейной зависимости и учесть возможные сильные корреляции между переменными при построении регрессионной модели. Так как есть пропущенные данные, корреляционная матрица была построена с опцией попарного удаления пропущенных данных (рисунок 26).

Из корреляционной матрицы в частности понятно, некоторые переменные очень сильно коррелируют друг с другом.

Стоит отметить, что достоверность больших значений корреляции возможна только при отсутствии выбросов в исходной таблице. Поэтому диаграммы рассеяния для зависимой переменной и всех остальных переменных обязательно должны учитываться при корреляционном анализе.

Переменная	Корреляции (Копия КОЕ+тем-ра+увлажнение.sta) Отмеченные корреляции значимы на уровне p <,05000 N=28 (Построчное удаление ПД) Сумма температур, град.	Коэффициент атмосферного увлажнения	Обсемененность спорами "картофельной" палочки, КОЕ/г	**Стекловидность, %**
Сумма температур, град.	1,00	-0,19	0,87	-0,09
Коэффициент атмосферного увлажнения	-0,19	1,00	-0,57	-0,15
Обсемененность спорами "картофельной"	0,87	-0,57	1,00	-0,03
Стекловидность, %	-0,09	-0,15	-0,03	1,00

Рисунок 26 – Корреляционная матрица переменных

Например, переменная № 1 и № 2 (Инвестиции в нефтяную и газовую промышленность соответственно), например, рисунок 27.

Высокий коэффициент корреляции между переменными (мультиколлиниарность) нужно учитывать при построении регрессионной модели. Здесь могут возникнуть большие ошибки при вычислении коэффициентов регрессии (плохообусловленная матрица при вычислении оценки через МНК).

В нашем случае (рисунок 28) связь между показателями № 1 и № 2 незначительна, устранение мультиколлиниарности не требуется.

Приведем наиболее распространенные способы устранения мультиколлиниарности:

1. Гребневая регрессия. Данная опция задается при построении множественной регрессии.

2. Исключение одной из объясняющих переменных. В этом случае из анализа исключается одна объясняющая переменная имеющая высокий парный коэффициент корреляции ($r > 0,8$) с другим предиктором.

3. Использование пошаговых процедур с включением/исключением предикторов.

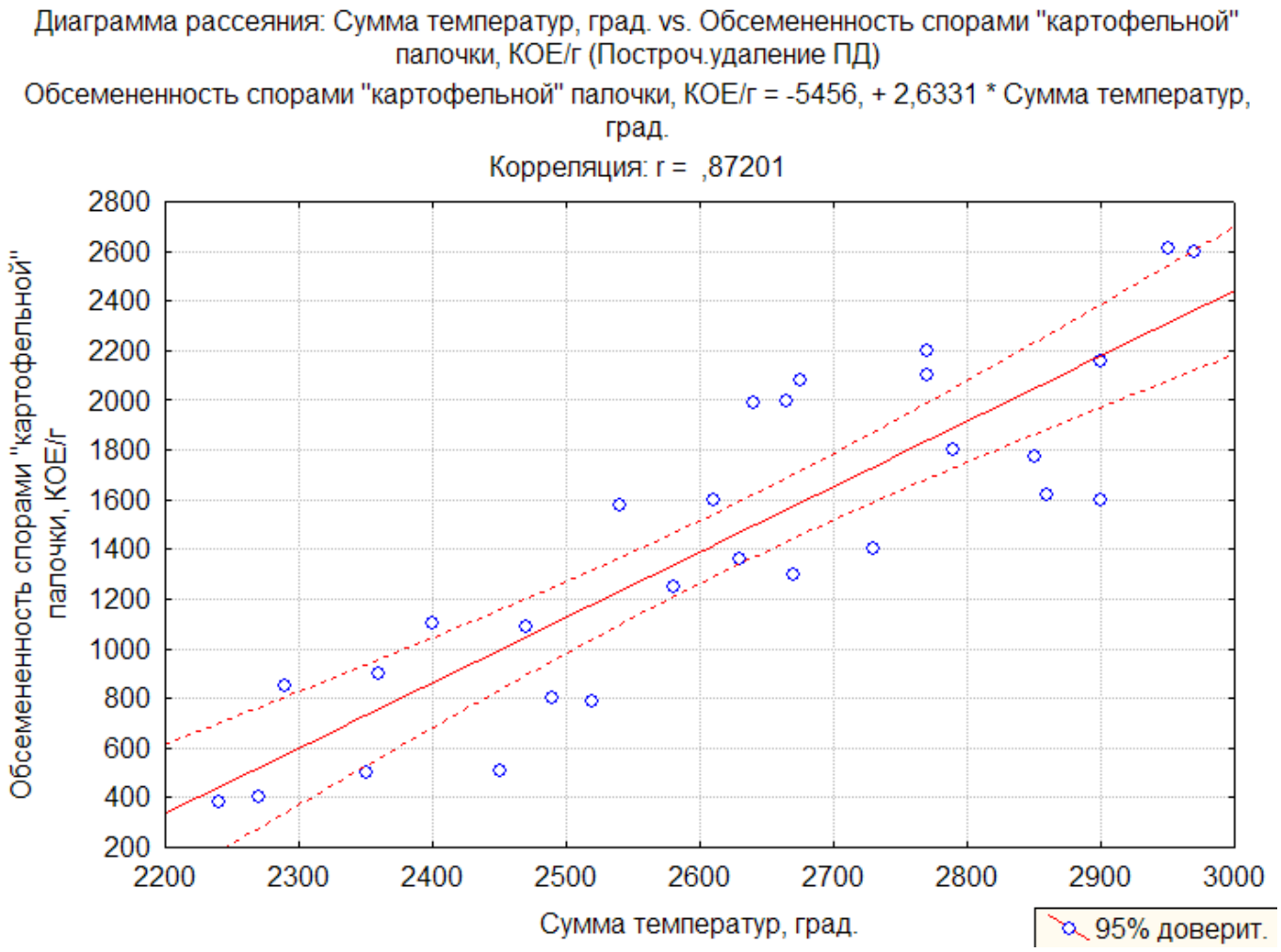


Рисунок 27 - Диаграммы рассеяния - зависимость показателя № 3 от показателя № 4

Обычно на основе значений корреляции, исключают объясняющие переменные, имеющие высокий парный коэффициент корреляции ($r > 0,8$), либо пошаговую регрессию с включением или исключением переменных.

Шаг 3. Теперь построим регрессионную модель при помощи выпадающей вкладки меню (Анализ/Множественная регрессия). В качестве зависимой переменной укажем «Обсемененность зерна», в качестве независимых – все остальные (рисунок 29).

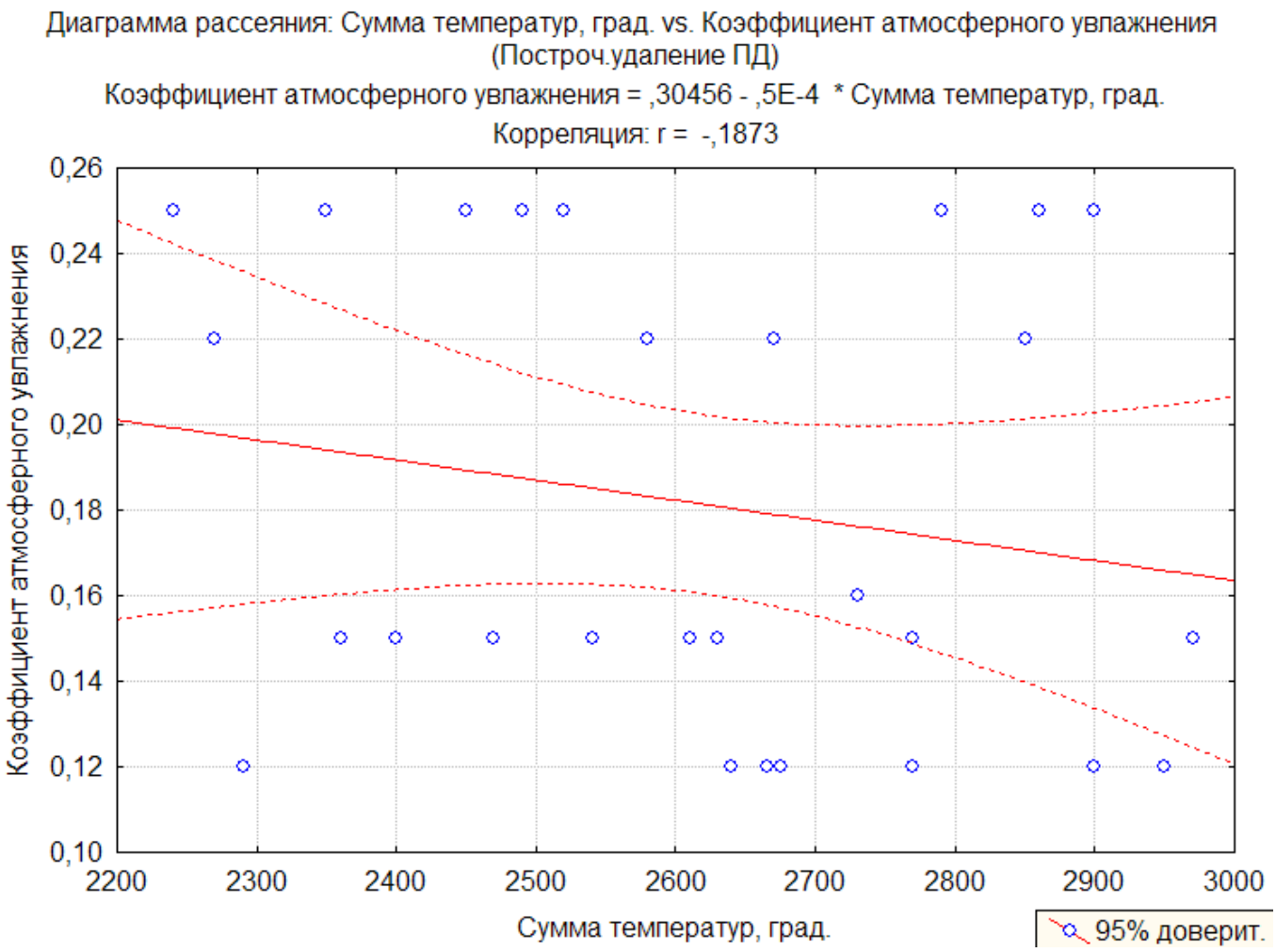


Рисунок 28 - Диаграммы рассеяния - связь показателей № 1 и № 2

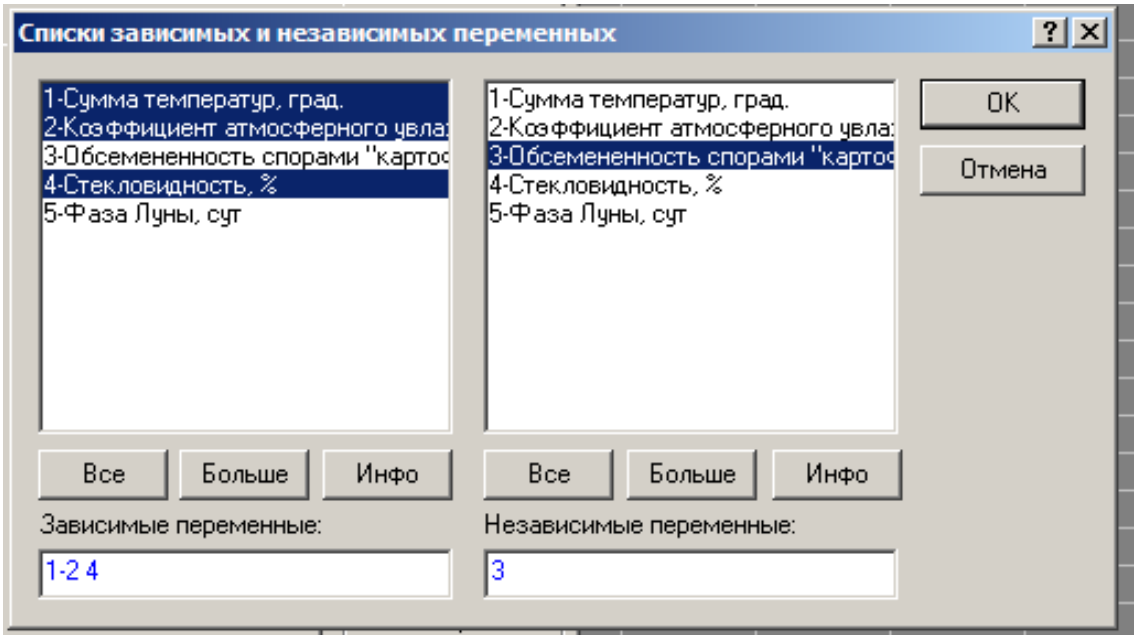


Рисунок 29 - Построение множественной регрессии

Множественную регрессию можно проводить пошагово. В этом случае в модель будут пошагово включаться (или исключаться) переменные, которые вносят наибольший (наименьший) вклад в регрессию на данном шаге (рисунок 30).

Также данная опция позволяет остановиться на шаге, когда коэффициент детерминации еще не наибольший, однако уже все переменные модели являются значимыми (рисунок 31).

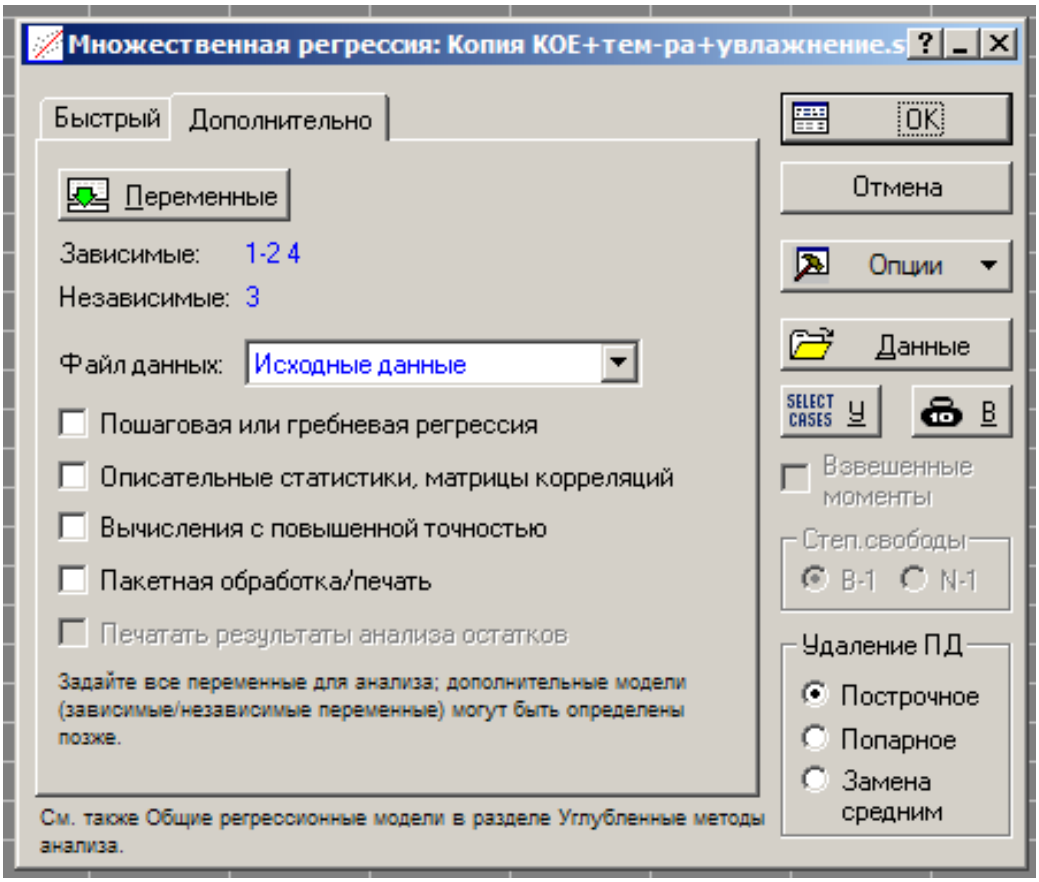


Рисунок 30 - Построение множественной регрессии

Особо стоит отметить, что пошаговая регрессия с включением, в случае, когда количество переменных больше количества наблюдений, является единственным способом построения регрессионной модели.

Установка нулевого значения свободного члена регрессионной модели используется в случае, если сама идея модели подразумевает нулевое значение отклика, когда все предикторы окажутся равными 0. Чаще всего подобные ситуации

встречаются в экономических задачах. В нашем случае свободный член мы включим в модель – рисунок 32.

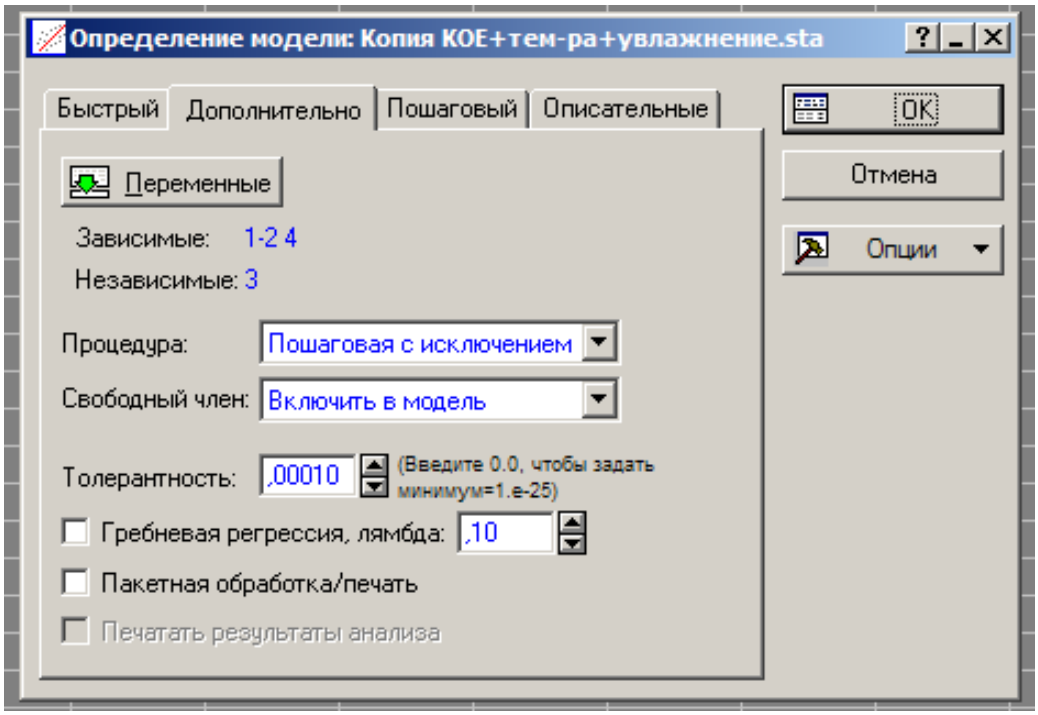


Рисунок 31 – Пошаговое построение множественной регрессии

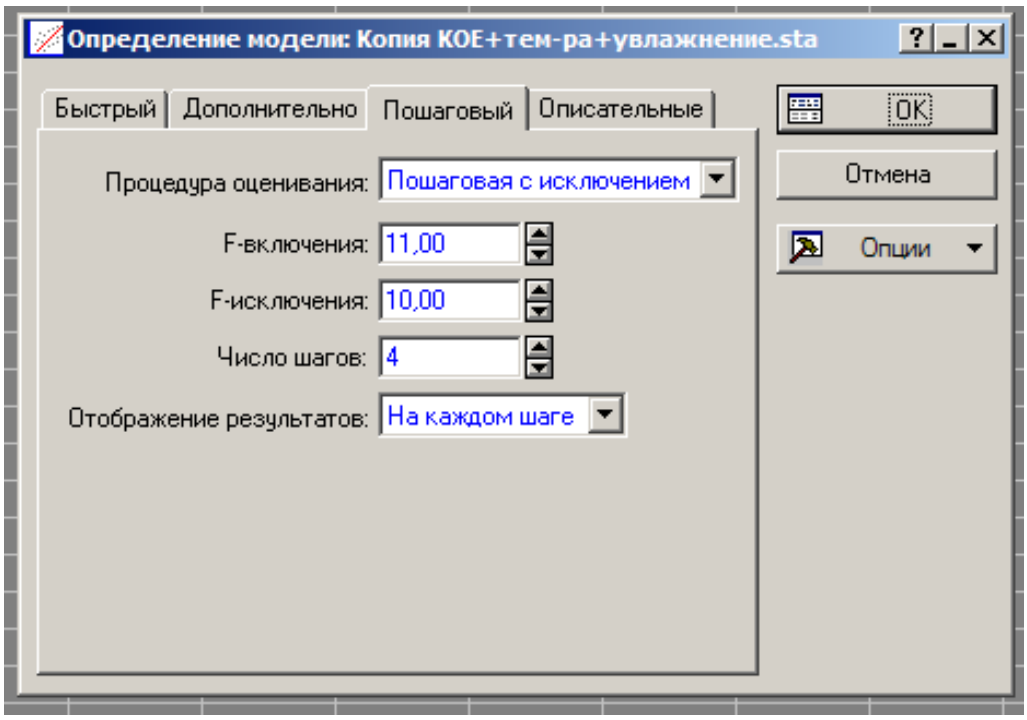


Рисунок 32 - Пошаговое построение множественной регрессии

В качестве параметров модели выберем Пошаговую с исключением (Fвкл = 11, Fвыкл = 10), с гребневой регрессией (лямбда = 0,1). И для каждой группы построим регрессионную модель.

Результаты в виде Итоговой таблицы регрессии представлены на рисунке 12. Они получены на последнем шаге регрессии.

Шаг 4. Проверка адекватности модели. Обратим внимание, что, несмотря на значимость всех переменных в регрессионной модели (p-уровень < 0.05 – подсвечены красным цветом), коэффициент детерминации R2 существенно меньше у первой группы наблюдений.

Коэффициент детерминации показывает, по сути, какая доля дисперсии отклика объясняется влиянием предикторов в построенной модели. Чем ближе R2 к 1, тем лучше модель.

F-статистика Фишера (таблица А.3) используется для проверки гипотезы о нулевых значениях коэффициентов регрессии. Гипотеза отклоняется при малом уровне значимости. В нашем случае (рисунок 33) значение F-статистики = 109,69 при уровне значимости p < 0,00000, т.е. гипотеза об отсутствии линейной связи отклоняется.

N=28	Итоги регрессии для зависимой переменной: Обсемененност R= ,96541326 R2= ,93202275 Скорректир. R2= ,92352560 F(3,24)=109,69 p<,00000 Станд. ошибка оценки: 179,22					
	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(24)	p-уров.
Св.член			-3779,12	598,6290	-6,31296	0,000002
Сумма температур, град.	0,790776	0,054573	2,39	0,1648	14,49012	0,000000
Коэффициент атмосферного увлажнения	-0,424631	0,054990	-5108,97	661,6135	-7,72199	0,000000
Стекловидность, %	-0,019263	0,054228	-1,26	3,5444	-0,35522	0,725524

Рисунок 33 – Итоговая таблица построения множественной регрессии

Шаг 5. Теперь проведем анализ остатков полученной модели. Результаты, полученные при анализе остатков, являются важным дополнением к значению коэффициента детерминации при проверке адекватности построенной модели.

В окне, представленном на рисунке 34, на вкладке Остатки/предсказанные/наблюдаемые значения нажмем на кнопку Анализ остатков, и далее нажмем на кнопку Остатки и предсказанные (рисунок 35).

Кнопка Анализ остатков будет активна, только если регрессия получена на последнем шаге. Чаще оказывается важным получить регрессионную модель, в которой значимы все предикторы, чем продолжить построение модели (увеличивая коэффициент детерминации) и получить незначимые предикторы.

В этом случае, когда регрессия не останавливается на последнем шаге, можно искусственно задать количество шагов в регрессии.

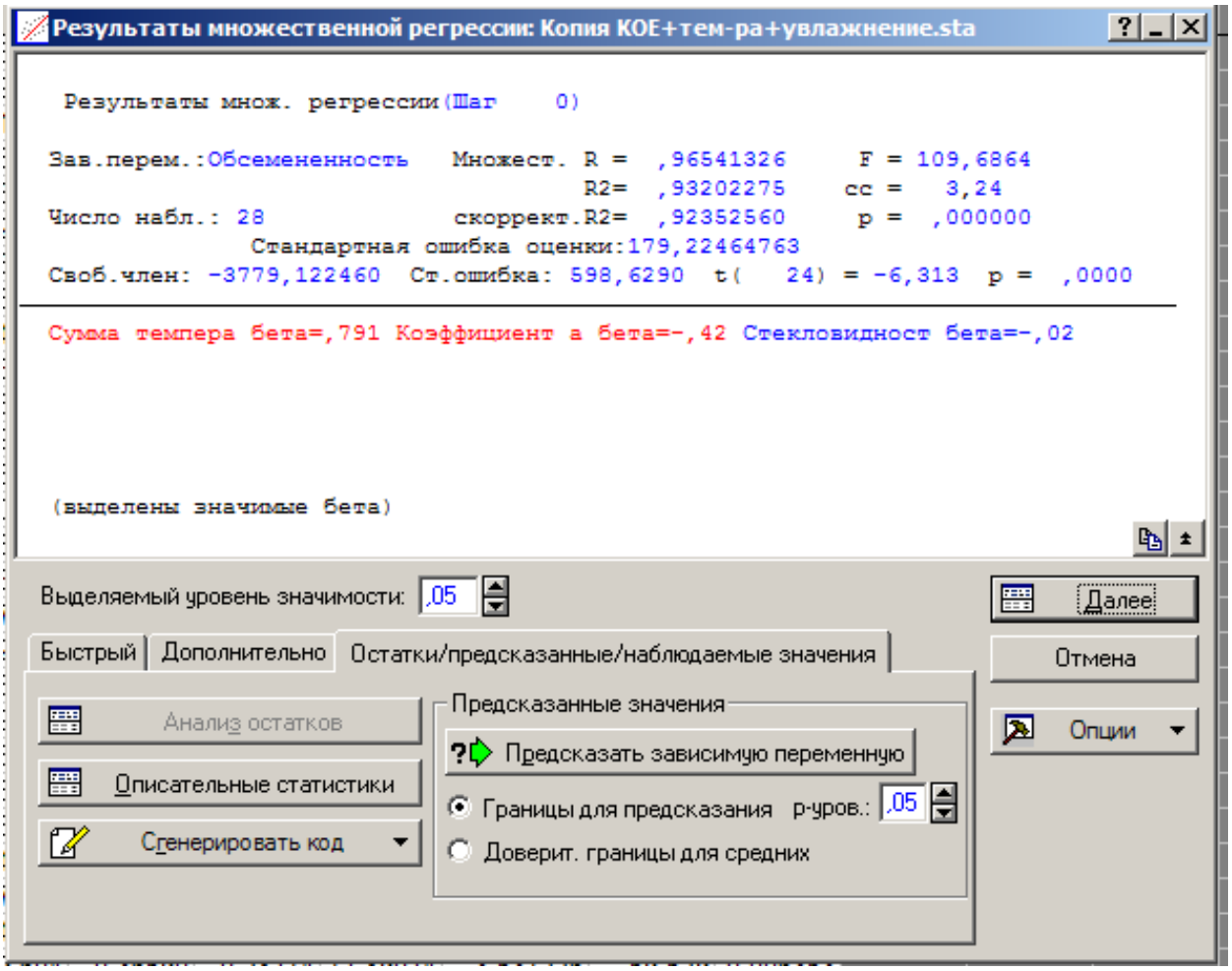


Рисунок 34 – Результаты множественной регрессии

Прокомментируем результаты, представленные на рисунке 35. Важным является столбец с Остатками (разница первых 2-х столбцов). Большие остатки по

многим наблюдениям и наличие наблюдения с маленьким остатком может указывать на последнее как на выброс.

Другими словами анализ остатков нужен для того, чтобы отклонения от предположений, угрожающие обоснованности результатов анализа, могли быть легко обнаружены.

Предсказанные значения и остатки (Копия КОЕ+тем-ра+увлажнение.sta)
Зависимая перемен.: Обсемененность спорами "картофельной" палочки, КОЕ/г

Набл. No.	**Наблюд. Значение**	Предск. Значение	Остатки	Станд. предск.	Станд. Остатки	Стд.Ош. предск.	Махалан. расст.	Удален. остатки	Кука расст.
C:1	400,000	408,736	-8,736	-1,64969	-0,04962	65,94547	2,823536	-10,161	0,000156
C:2	900,000	979,081	-79,081	-0,73795	-0,44916	59,68850	2,138851	-89,350	0,009866
C:3	500,000	448,257	51,743	-1,58651	0,29389	64,77859	2,690674	59,844	0,005213
C:4	1100,000	1074,873	25,127	-0,58482	0,14271	54,95294	1,665992	27,839	0,000812
C:5	800,000	783,529	16,471	-1,05055	0,09355	55,90326	1,757751	18,318	0,000364
C:6	1600,000	1577,781	22,219	0,21911	0,12620	39,05922	0,364537	23,369	0,000289
C:7	1300,000	1366,655	-66,655	-0,11839	-0,37858	43,11639	0,654929	-70,908	0,003242
C:8	2000,000	1861,558	138,442	0,67275	0,78632	50,99203	1,300485	151,118	0,020598
C:9	1400,000	1814,469	-414,469	0,59747	-2,35407	38,99384	0,360091	-435,848	0,100196
C:10	2100,000	2113,011	-13,011	1,07471	-0,07390	53,70493	1,547878	-14,346	0,000206
C:11	1800,000	1501,969	298,031	0,09792	1,69274	65,10980	2,728145	345,246	0,175284
C:12	2200,000	1960,949	239,051	0,83163	1,35775	43,77449	0,704735	254,802	0,043156
C:13	1600,000	1765,396	-165,396	0,51903	-0,93941	76,24067	4,098541	-203,568	0,083557
14	2600,000	2439,908	160,092	1,59728	0,90928	65,29974	2,749719	185,625	0,050967
C:15	380,000	184,829	195,171	-2,00761	1,10852	75,81628	4,042335	239,600	0,114470
C:16	850,000	963,508	-113,508	-0,76285	-0,64469	78,96100	4,466278	-142,086	0,043663
C:17	510,000	687,737	-177,737	-1,20368	-1,00950	57,68785	1,934315	-199,113	0,045767
C:18	1090,000	1242,509	-152,509	-0,31684	-0,86621	47,63027	1,011709	-164,552	0,021309
C:19	790,000	855,373	-65,373	-0,93571	-0,37130	55,02161	1,672570	-72,448	0,005512
C:20	1580,000	1410,145	169,855	-0,04887	0,96473	42,07821	0,577891	180,144	0,019932
C:21	1250,000	1151,124	98,876	-0,46293	0,56159	41,21432	0,515217	104,609	0,006448
C:22	1990,000	1801,688	188,312	0,57704	1,06956	51,15175	1,314695	205,673	0,038394
C:23	1360,000	1625,677	-265,677	0,29567	-1,50897	38,75729	0,344072	-279,207	0,040621
C:24	2080,000	1885,506	194,494	0,71103	1,10468	51,01667	1,302675	212,321	0,040701
C:25	1770,000	1797,719	-27,719	0,57070	-0,15744	58,60949	2,027673	-31,174	0,001158
C:26	2160,000	2424,335	-264,335	1,57238	-1,50136	63,44231	2,541438	-303,779	0,128844
C:27	1620,000	1669,604	-49,604	0,36590	-0,28174	71,89014	3,537226	-59,529	0,006353
C:28	2610,000	2544,075	65,925	1,76380	0,37444	68,52819	3,126043	77,695	0,009834
Минимум	380,000	184,829	-414,469	-2,00761	-2,35407	38,75729	0,344072	-435,848	0,000156

Рисунок 35 - Остатки и предсказанные значения регрессионной модели

Набл. No.	Предсказанные значения и остатки (Копия КОЕ+тем-ра+увлажнение.sta) Зависимая перемен.: Обсемененность спорами "картофельной" палочки, КОЕ/г										
	Наблюд. Значение	**Предск. Значение**	Остатки	Станд. предск.	Станд. Остатки	Стд.Ош. предск.	Махалан. расст.	Удален. остатки	Кука расст.	UBC =V2+1.96*V6	LBC =V2-1.96*V6
C:1	400,000	408,736	-8,736	-1,64969	-0,04962	65,94547	2,823536	-10,161	0,000156	537,988921	279,482679
C:2	900,000	979,081	-79,081	-0,73795	-0,44916	59,68850	2,138851	-89,350	0,009866	1096,07056	862,09164
C:3	500,000	448,257	51,743	-1,58651	0,29389	64,77859	2,690674	59,844	0,005213	575,222936	321,290864
C:4	1100,000	1074,873	25,127	-0,58482	0,14271	54,95294	1,665992	27,839	0,000812	1182,58076	967,165238
C:5	800,000	783,529	16,471	-1,05055	0,09355	55,90326	1,757751	18,318	0,000364	893,09909	673,95831
C:6	1600,000	1577,781	22,219	0,21911	0,12620	39,05922	0,364537	23,369	0,000289	1654,33707	1501,22493
C:7	1300,000	1366,655	-66,655	-0,11839	-0,37858	43,11639	0,654929	-70,908	0,003242	1451,16312	1282,14688
C:8	2000,000	1861,558	138,442	0,67275	0,78632	50,99203	1,300485	151,118	0,020598	1961,50238	1761,61362
C:9	1400,000	1814,469	-414,469	0,59747	-2,35407	38,99384	0,360091	-435,848	0,100196	1890,89693	1738,04107
C:10	2100,000	2113,011	-13,011	1,07471	-0,07390	53,70493	1,547878	-14,346	0,000206	2218,27266	2007,74934
C:11	1800,000	1501,969	298,031	0,09792	1,69274	65,10980	2,728145	345,246	0,175284	1629,58421	1374,35379
C:12	2200,000	1960,949	239,051	0,83163	1,35775	43,77449	0,704735	254,802	0,043156	2046,747	1875,151
C:13	1600,000	1765,396	-165,396	0,51903	-0,93941	76,24067	4,098541	-203,568	0,083557	1914,82771	1615,96429
14	2600,000	2439,908	160,092	1,59728	0,90928	65,29974	2,749719	185,625	0,050967	2567,89549	2311,92051
C:15	380,000	184,829	195,171	-2,00761	1,10852	75,81628	4,042335	239,600	0,114470	333,428909	36,2290912
C:16	850,000	963,508	-113,508	-0,76285	-0,64469	78,96100	4,466278	-142,086	0,043663	1118,27146	808,74434
C:17	510,000	687,737	-177,737	-1,20368	-1,00950	57,68785	1,934315	-199,113	0,045767	800,804986	574,668614
C:18	1090,000	1242,509	-152,509	-0,31684	-0,86621	47,63027	1,011709	-164,552	0,021309	1335,86433	1149,15367
C:19	790,000	855,373	-65,373	-0,93571	-0,37130	55,02161	1,672570	-72,448	0,005512	963,215056	747,530344
C:20	1580,000	1410,145	169,855	-0,04887	0,96473	42,07821	0,577891	180,144	0,019932	1492,61829	1327,67171
C:21	1250,000	1151,124	98,876	-0,46293	0,56159	41,21432	0,515217	104,609	0,006448	1231,90407	1070,34393
C:22	1990,000	1801,688	188,312	0,57704	1,06956	51,15175	1,314695	205,673	0,038394	1901,94543	1701,43057
C:23	1360,000	1625,677	-265,677	0,29567	-1,50897	38,75729	0,344072	-279,207	0,040621	1701,64129	1549,71271
C:24	2080,000	1885,506	194,494	0,71103	1,10468	51,01667	1,302675	212,321	0,040701	1985,49867	1785,51333
C:25	1770,000	1797,719	-27,719	0,57070	-0,15744	58,60949	2,027673	-31,174	0,001158	1912,5936	1682,8444
C:26	2160,000	2424,335	-264,335	1,57238	-1,50136	63,44231	2,541438	-303,779	0,128844	2548,68193	2299,98807
C:27	1620,000	1669,604	-49,604	0,36590	-0,28174	71,89014	3,537226	-59,529	0,006353	1810,50867	1528,69933
C:28	2610,000	2544,075	65,925	1,76380	0,37444	68,52819	3,126043	77,695	0,009834	2678,39025	2409,75975

Рисунок 36 - Остатки и предсказанные значения регрессионной модели + 2 границы 0,95 доверительного интервала

В конце приведем график, иллюстрирующий данные, полученные из таблицы на рисунке 36. Здесь добавлены 2 переменные: UCB и LCB, представляющих собой 0,95 верхний и нижний доверительный интервал

$$UBC = V2 + 1{,}96 \cdot V6, \tag{13}$$

$$LBC = V2 - 1{,}96 \cdot V6. \tag{14}$$

И удалены четыре последних наблюдения. Построим линейный график с переменными (Графики/2М Графики/Линейные графики для переменных):

1) наблюдаемое значение (V1);

2) предсказанное значение (V2);

3) UCB (V9);

4) LCB (V10).

Результат представлен на рисунке 37. Теперь видно, что построенная регрессионная модель довольно неплохо отражает реальную обсемененность зерна спорами, особенно на результатах недавнего прошлого. Это означает, что в ближайшем будущем реальные значения могут быть приближены модельными.

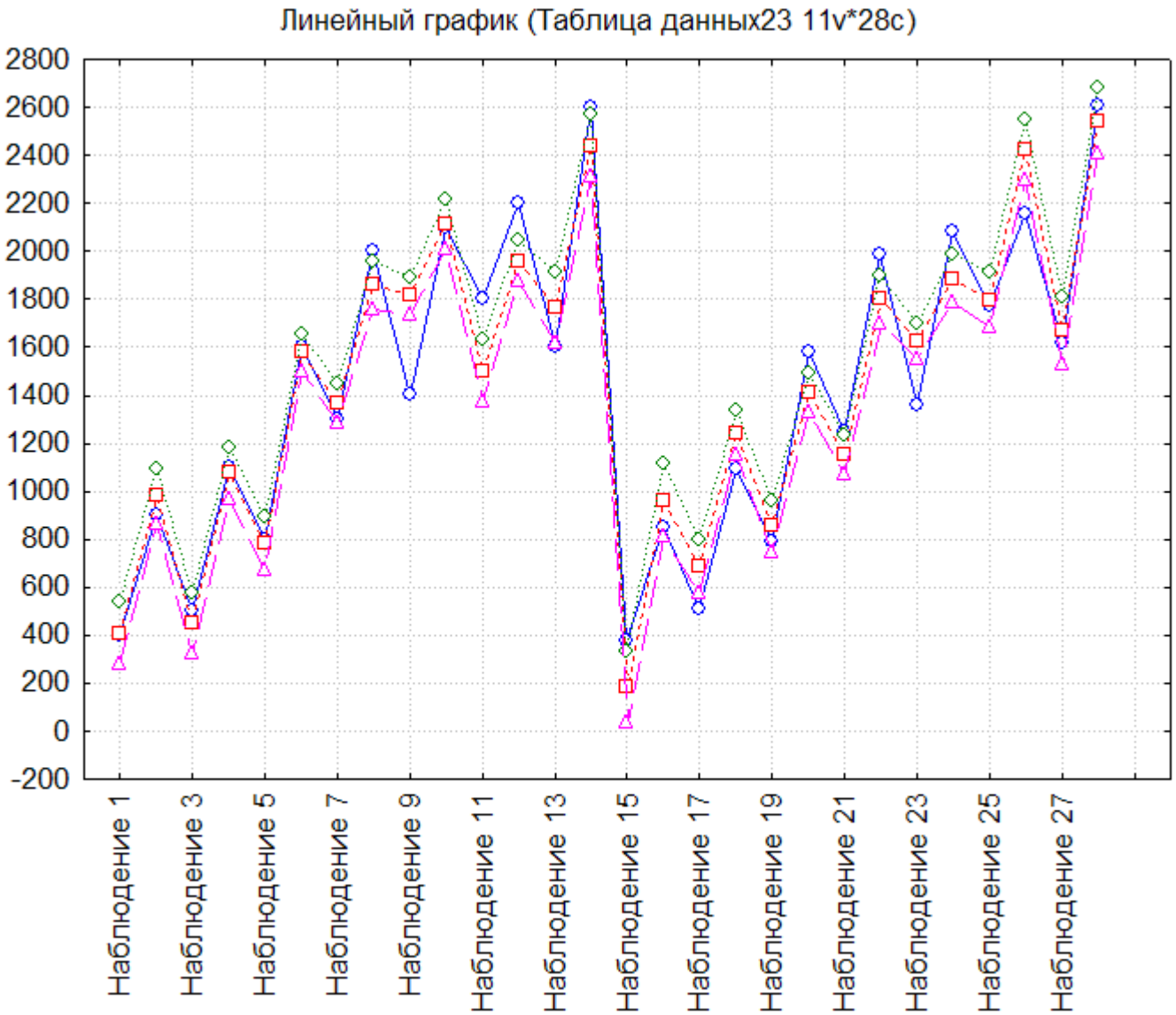


Рисунок 37 – Наблюдаемые и предсказанные значения

Следует помнить, что все методы регрессионного анализа позволяют обнаружить только числовые зависимости, а не лежащие в их основе причинные связи. Поэтому ответ на вопрос о значимости переменных в полученной модели остается за экспертом в данной области, который, в частности, способен учесть влияние факторов, возможно, не вошедших в данную таблицу.

2.14 Выводы по результатам регрессионного анализа

Итак, для модели с гребневой регрессией (лямбда равна 0,1) в результате проведения множественной регрессии для всех наблюдений, зависимая переменная № 3 «Обсемененность зерна» представима как

$$x_3 = 2,39 \cdot x_1 - 5068,76 \cdot x_2 - 3912,33 \ , \qquad (15)$$

где x_3 - зараженностью зерна пшеницы спорами «картофельной» палочки, КОЕ/г;

x_1 - показатели суммы температур, градусы;

x_2 - коэффициента атмосферного увлажнения.

Связь с показателем стекловидности оказался незначительным (рисунок 38).

Параметры модели: пошаговая с включением, гребневая регрессия с параметром лямбда 0,1. Результат получен на последнем шаге регрессии [16].

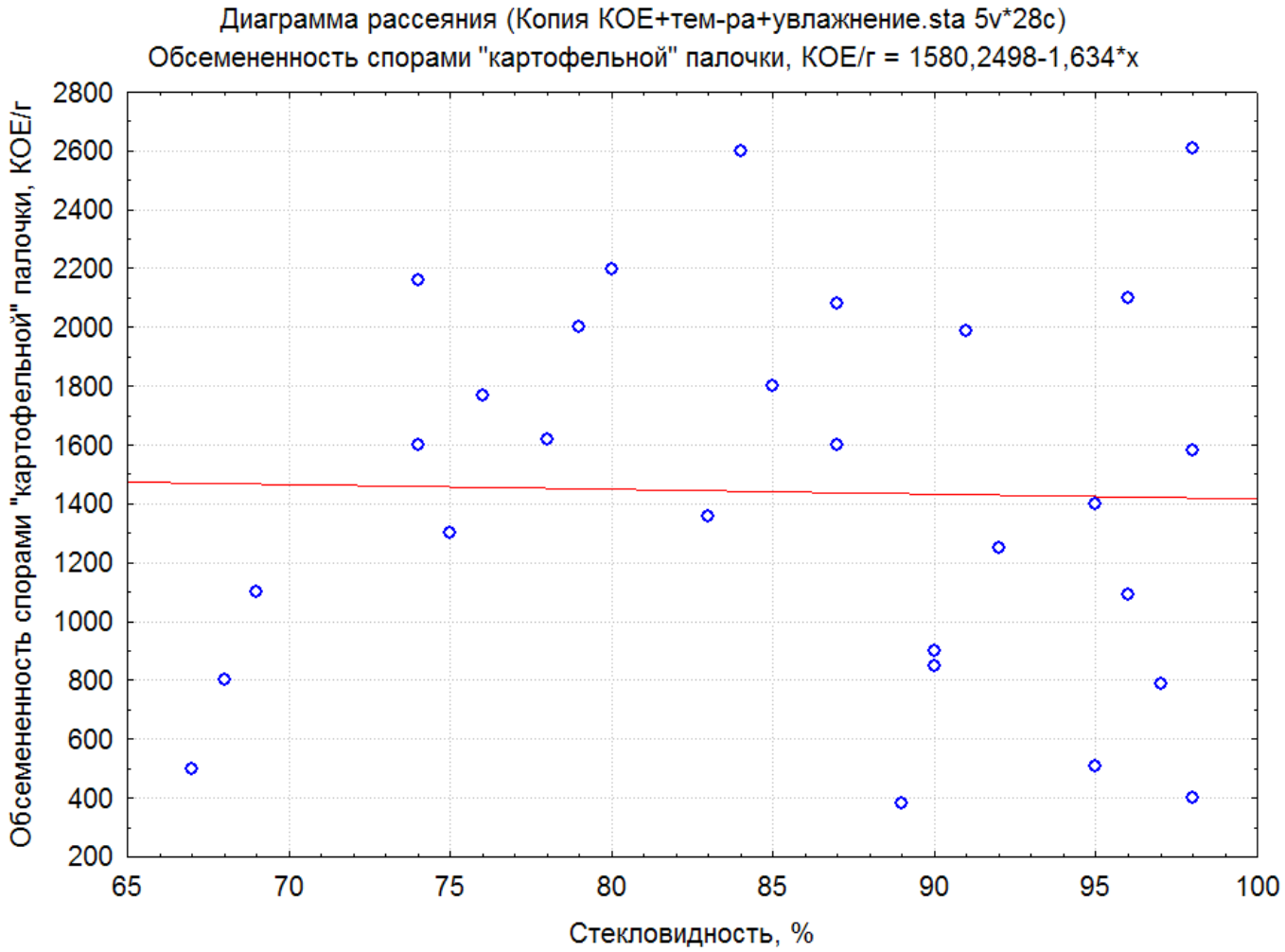


Рисунок 38 – Диаграмма рассеяния показателя стекловидности и обсемененности зерна

2.15 Задания для самостоятельной работы по теме «Регрессионный анализ»

Требуется построить регрессионные модели для показателей качества зерна и выпускаемой продукции на основе данных таблиц 11, 12, 13, 14.

Таблица 11 – Показатели качества зерна и хлеба

Номер эксперимента	Микротвердость зерна, кг/мм²	Объемный выход хлеба, см³/100 г муки	Формоустойчивость хлеба
1	11,2	0,33	0,33
2	11,1	0,33	0,37
3	11,3	0,38	0,34
4	11,6	0,36	0,3
5	11,9	0,4	0,32
6	11,9	0,41	0,32
7	12,4	0,4	0,45
8	12,6	0,43	0,35
9	13	0,44	0,41
10	13,5	0,47	0,47
11	13,5	0,45	0,49
12	14,2	0,48	0,52
13	14	0,48	0,42
14	14,1	0,51	0,46
15	14	0,53	0,55
16	14,5	0,51	0,53
17	14,6	0,54	0,55
18	14,6	0,53	0,53
19	14,5	0,53	0,5
20	14,7	0,55	0,57
21	15	0,52	0,56
22	14,8	0,53	0,58

Продолжение таблицы 11

Номер эксперимента	Микротвердость зерна, кг/мм²	Объемный выход хлеба, см³/100 г муки	Формоустойчивость хлеба
23	15,3	0,55	0,51
24	15,3	0,56	0,52
25	15,1	0,54	0,52
26	15,5	0,52	0,55
27	16	0,49	0,5
28	16	0,51	0,47
29	16,2	0,51	0,49
30	16,5	0,47	0,48
31	17,1	0,48	0,46
32	17	0,48	0,45
33	17,2	0,47	0,44
34	17,4	0,43	0,39
35	17,7	0,43	0,4
36	17,7	0,4	0,34
37	18,4	0,4	0,39
38	18,4	0,37	0,39
39	19,2	0,35	0,34
40	19,5	0,34	0,32
41	19,2	0,32	0,30
42	19,1	0,31	0,29

Таблица 12 – Показатели качества зерна и реологические свойства теста из муки, смолотой из этого зерна

Номер эксперимента	Твердозерность, %	Содержание клейковины, %	Газообразующая способность муки, мл CO2 на 100 г	Водопоглотительная способность муки, %
1	9,9	21,3	1324	53
2	10,5	21,1	1322	59
3	11	23,1	1317	57
4	12	19,6	1350	58
5	13,2	29,9	1374	68
6	14	22,7	1368	57
7	14,5	22,4	1387	62
8	15	19,3	1396	63
9	15,8	21,7	1462	68
10	16,6	29,1	1465	75
11	17	22,2	1420	69
12	18	22,1	1501	77
13	18,6	19,5	1521	76
14	19	20,9	1512	82
15	9,9	22,2	1273	55
16	11	20,4	1303	52
17	10,4	25,6	1351	60
18	9,3	29,7	1275	52
19	10,1	28,6	1327	51
20	10,6	27,8	1374	59

Продолжение таблицы 12

Номер эксперимента	Твердозерность, %	Содержание клейковины, %	Газообразующая способность муки, мл CO2 на 100 г	Водопоглотительная способность муки, %
21	10,1	30,8	1357	59
22	11,2	32,5	1344	58
23	11	32,3	1352	61
24	10,5	31,1	1404	66
25	11,5	32,7	1348	61
26	10,2	34,7	1424	61
27	11,1	39,5	1426	70
28	11,6	41,8	1430	69
29	11,7	40,4	1380	62
30	11,2	23,4	1334	54
31	12,4	19,9	1350	52
32	13,6	29,4	1374	63
33	14,1	22,2	1362	54
34	14,8	22,7	1380	67

Таблица 13 – Характеристики твердой яровой пшеницы Оренбургской области по показателям качества

Показатели	2010	2011	2012	2013
Натура, г/л	811	820	792	805
Стекловидность зерна, %	95	95	91	84
Содержание сырой клейковины, %	32	24	30,5	30,5
Число падения, с	320	300	360	300

Таблица 14 – Характеристики мягкой яровой пшеницы Оренбургской области по показателям качества

Показатели	2010	2011	2012	2013
Натура, г/л	779	811	775	781
Стекловидность зерна, %	86,3	79,5	88,1	68,2
Содержание сырой клейковины, %	32,1	35,3	30,2	31,2
ИДК, у.е.	85	90	85	95
Число падения, с	260	280	210	230

Список использованных источников

1 Гмурман, В.Е. Теория вероятностей и математическая статистика: учебное пособие для бакалавров / В.Е. Гмурман. – Москва : Юрайт, 2013. - 479 с.

2 Горлач, Б.А. Теория вероятностей и математическая статистика: учебное пособие / Б.А. Горлач. – Санкт-Петербург : Лань, 2013. - 320 с.

3 Калинина, В.Н. Теория вероятностей и математическая статистика: учебник для бакалавров / В.Н. Калинина. – Москва : Юрайт, 2013. - 472 с.

4 Кобзарь, А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. – Москва : ФИЗМАТЛИТ, 2012. - 816 с.

5 Колемаев, В.А. Теория вероятностей и математическая статистика: учебник / В.А. Колемаев, В.Н. Калинина. – Москва : КноРус, 2013. - 376 с.

6 Кочетков, Е.С. Теория вероятностей и математическая статистика: учебник / Е.С. Кочетков, С.О. Смерчинская, В.В. Соколов. – Москва : Форум, НИЦ ИНФРА-М, 2013. - 240 с.

7 Краснов, М.Л. Вся высшая математика. Теория вероятностей. Математическая статистика. Теория игр: учебник / М.Л. Краснов, А.И. Киселев, Г.И. Макаренко. – Москва : ЛКИ, 2013. - 296 с.

8 Кремер, Н.Ш. Теория вероятностей и математическая статистика: учебник для студентов вузов / Н.Ш. Кремер. – Москва : ЮНИТИ-ДАНА, 2012. - 551 с.

9 Кричевец, А.Н. Математическая статистика для психологов: учебник для студ. учреждений высш. проф. образования / А.Н. Кричевец, А.А. Корнеев, Е.И. Рассказова. – Москва : ИЦ Академия, 2012. - 400 с.

10 Семенов, В.А. Теория вероятностей и математическая статистика: учебное пособие / В.А. Семенов. - Санкт-Петербург : Питер, 2013. - 192 с.

11 Сидняев, Н.И. Теория вероятностей и математическая статистика: учебник для бакалавров / Н.И. Сидняев. – Москва : Юрайт, ИД Юрайт, 2011. - 219 с.

12 Спирина, М.С. Теория вероятностей и математическая статистика: учебник для студ. учреждений сред. проф. образования / М.С. Спирина, П.А. Спирин. – Москва : ИЦ Академия, 2012. - 352 с.

13 Бычкова, С.Г. Социально-экономическая статистика: учебник для бакалавров / С.Г. Бычкова. – Москва : Юрайт, 2013. - 591 с.

14 Яковлева, А.В. Экономическая статистика: учебное пособие / А.В. Яковлева. – Москва : ИЦ РИОР, 2013. - 95 с.

15 Балдин, К.В. Общая теория статистики: учебное пособие / К.В. Балдин, А.В. Рукосуев. – Москва : Дашков и К, 2012. - 312 с.

16 Батракова, Л.Г. Теория статистики: учебное пособие / Л.Г. Батракова. – Москва : КноРус, 2013. - 528 с.

17 Громыко, Г.Л. Теория статистики: Практикум / Г.Л. Громыко. – Москва : НИЦ ИНФРА-М, 2013. - 238 с.

18 Ефимова, М.Р. Общая теория статистики: учебник / М.Р. Ефимова, Е.В. Петрова, В.Н. Румянцев. – Москва : ИНФРА-М, 2013. - 416 с.

19 Лысенко, С.Н. Общая теория статистики: учебное пособие / С.Н. Лысенко, И.А. Дмитриева. – Москва : ИД ФОРУМ, НИЦ ИНФРА-М, 2013. - 208 с.

20 Костин, В. Н. Теория эксперимента : учебное пособие / В. Н. Костин, В. В. Паничев; М-во образования и науки Рос. Федерации, Федер. гос. бюджет. образоват. учреждение высш. проф. образования "Оренбург. гос. ун-т", Каф. прогр. обеспечения вычисл. техники и автоматизир. систем. - Оренбург : Университет, 2014. - 212 с. : табл. - Библиогр.: с. 207-208. - Прил.: с. 209-212. - ISBN 978-5-4417-0415-1.

21 Килов, А. С. Планирование экспериментов и обработка экспериментальных данных : методические указания к практическому занятию для студентов, обучающихся по программам высшего профессионального образования по направлению подготовки 150700.62 Машиностроение / А. С. Килов; М-во образования и науки Рос. Федерации, Федер. гос. бюджет. образоват. учреждение высш. проф. образования "Оренбург. гос. ун-т", Каф. материаловедения и технологии материалов. - Оренбург : ОГУ, 2014. - 35 с.

22 Годин, А. М. Статистика: учебник / А. М. Годин. – Москва: Дашков и К°, 2012. – 451 с.

23 Елисеева, И. И. Статистика : учебник для бакалавров / И. И. Елисеева. – Москва : Юрайт: ИД Юрайт, 2011. – 565 с.

24 Ниворожкина, Л. И. Статистика: учебник для бакалавров: учебник / Л. И. Ниворожкина. – Москва : Дашков и К°: Наука–Спектр, 2011. – 415 с.

25 Тумасян, А. А. Статистика промышленности: учебное пособие / А. А. Тумасян, Л. И. Василевская. – Минск: Новое знание. – Москва : Инфра–М, 2012. – 429 с.

Приложение А
(обязательное)

Проверка гипотез

Под гипотезой подразумевается некоторое предположение о случайной величине (функции распределения, математической модели и прочие). Примером может служить гипотеза об типе закона распределения.

Проверка статистических гипотез – один из разделов математической статистики. Необходимость выдвижения гипотез возникает при обработке или интерпретации результатов наблюдений. При проверке гипотезы необходимо установить, насколько экспериментальные результаты согласуются с выдвинутой гипотезой, после чего принять или отвергнуть гипотезу.

Правило, в соответствии с которым принимается или отвергается данная гипотеза, называется статистическим критерием. Построение критерия сводится к выбору подходящей функции от результата наблюдений, служащей мерой расхождения между экспериментальными и гипотетическими законами. При решении вопроса о принятии или отклонении какой-либо гипотезы с помощью какого-либо статистического критерия, основанного на результатах эксперимента, могут быть допущены ошибки двух типов.

Ошибка «первого рода» совершается тогда, когда гипотеза отвергается, а на самом деле она верна; «второго рода» – когда гипотеза принимается, а на самом деле она не верна. Результаты проверки гипотезы никогда не могут служить доказательством абсолютной справедливости и правильности гипотезы. Они означают лишь то, что гипотеза с заданной вероятностью не противоречит результатам эксперимента. Поэтому при проверке гипотезы нужно заранее допустить возможность ошибочного решения.

Соответственно, ошибку второго рода иногда называют пропуском события или ложноотрицательным срабатыванием - человек болен, но анализ крови

этого не показал, или у пассажира имеется холодное оружие, но рамка металлодетектора его не обнаружила (например, из-за того, что чувствительность рамки отрегулирована на обнаружение только очень массивных металлических предметов).

Слово «отрицательный» в данном случае не имеет отношения к желательности или нежелательности самого события. Термин широко используется в медицине. Например, тесты, предназначенные для диагностики заболеваний, иногда дают отрицательный результат (т.е. показывают отсутствие заболевания у пациента), когда на самом деле пациент страдает этим заболеванием. Такой результат называется ложноотрицательным.

В других областях обычно используют словосочетания со схожим смыслом, например, «пропуск события», и т.п. В информационных технологиях часто используют английский термин false negative без перевода.

Степень чувствительности системы защиты должна представлять собой компромисс между вероятностью ошибок первого и второго рода. Где именно находится точка баланса, зависит от оценки рисков обоих видов ошибок.

Вероятность того, что гипотеза будет отвергнута, хотя на самом деле она верна, называют уровнем значимости и обозначают q. Тогда величина $P = 1 – q$, называемая статистической надежностью, характеризует вероятность выполнения статистического критерия при условии, что гипотеза верна. В технических задачах, как правило, выбирают $q = 0,05$ или $0,01$, что соответствует уровням значимости 5 % и 1 %.

Ошибки первого и второго рода являются большой проблемой в системах биометрического сканирования, использующих распознавание радужной оболочки или сетчатки глаза, черт лица и т.д. Такие сканирующие системы могут ошибочно отождествить кого-то с другим, «известным» системе человеком, информация о котором хранится в базе данных (к примеру, это может быть лицо, имеющее право входа в систему, или подозреваемый преступник и т.п.). Противоположной ошибкой будет неспособность системы распознать легитимного зарегистрированного пользователя, или опознать подозреваемого в преступлении.

Таблица А.1 - Значения критерия Стьюдента (t-критерия). Критические значения коэффициента Стьюдента (t-критерия) для различной доверительной вероятности p и числа степеней

f	p							
	0,999	0,998	0,995	0,99	0,98	0,95	0,90	0,80
1	636,6190	318,3060	127,6560	63,6560	31,8200	12,7060	6,3130	3,0770
2	31,5990	22,3270	14,0890	9,9240	6,9640	4,3020	2,9200	1,8850
3	12,9240	10,2140	7,4580	5,8400	4,5400	3,1820	2,35340	1,6377
4	8,6100	7,1730	5,5970	4,6040	3,7460	2,7760	2,13180	1,5332
5	6,8630	5,8930	4,7730	4,0321	3,6490	2,5700	2,01500	1,4759
6	5,9580	5,2070	4,3160	3,7070	3,1420	2,4460	1,9430	1,4390
7	5,4079	4,7850	4,2293	3,4995	2,9980	2,3646	1,8946	1,4149
8	5,0413	4,5008	3,8320	3,3554	2,8965	2,3060	1,8596	1,3968
9	4,7800	4,2968	3,6897	3,2498	2,8214	2,2622	1,8331	1,3830
10	4,5869	4,1437	3,5814	3,1693	2,7638	2,2281	1,8125	1,3720
11	4,4370	4,0240	3,4960	3,1050	2,7180	2,2010	1,7950	1,3630
12	4,1780	3,9290	3,4284	3,0845	2,6810	2,1788	1,7823	1,3562
13	4,2200	3,8520	3,3725	3,1123	2,6503	2,1604	1,7709	1,3502
14	4,1400	3,7870	3,3257	2,9760	2,6245	2,1448	1,7613	1,3450
15	4,0720	3,7320	3,2860	2,9467	2,6025	2,1314	1,7530	1,3406
16	4,0150	3,6860	3,2520	2,9200	2,5830	2,1190	1,7450	1,3360
17	3,9650	3,6458	3,2224	2,8982	2,5668	2,1098	1,7396	1,3334
18	3,9216	3,6105	3,1966	2,8784	2,5514	2,1009	1,7341	1,3304
19	3,8834	3,5794	3,1737	2,8609	2,5395	2,0930	1,7291	1,3277
20	3,8495	3,5518	3,1534	2,8453	2,5280	2,0860	1,7247	1,3253

Продолжение таблицы А.1

f	p							
	0,999	0,998	0,995	0,99	0,98	0,95	0,90	0,80
21	3,8190	3,5270	3,1350	2,8310	2,5170	2,0790	1,7200	1,3230
22	3,7921	3,5050	3,1188	2,8188	2,5083	2,0739	1,7117	1,3212
23	3,7676	3,4850	3,1040	2,8073	2,4999	2,0687	1,7139	1,3195
24	3,7454	3,4668	3,0905	2,7969	2,4922	2,0639	1,7109	1,3178
25	3,7251	3,4502	3,0782	2,7874	2,4851	2,0595	1,7081	1,3163
30	3,6460	3,3852	3,0298	2,7500	2,4573	2,0423	1,6973	1,3104
40	3,5510	3,3069	3,9712	2,7045	2,4233	2,0211	1,6839	1,3030
42	3,5370	3,2960	2,6930	2,6980	2,4180	2,0180	1,6820	1,3200
44	3,5258	3,2861	3,9555	2,6923	2,4141	2,0154	1,6802	1,3010
46	3,5150	3,2771	3,9488	2,6870	2,4102	2,0129	1,6767	1,3000
48	3,5051	3,2689	3,9426	2,6822	2,4056	2,0106	1,6772	1,2990
50	3,4060	3,2614	3,9370	2,6778	2,4033	2,0086	1,6759	1,2980
55	3,4760	3,2560	2,9240	2,6680	2,3960	2,0040	1,6730	1,2997
60	3,4602	3,2317	3,9146	2,6603	2,3901	2,0003	1,6706	1,2958
80	3,4160	3,1950	2,8870	2,6380	2,3730	1,9900	1,6640	1,2820
90	3,4019	3,1833	2,8779	2,6316	2,3885	1,9867	1,6620	1,2910
100	3,3905	3,1737	2,8707	2,6259	2,3642	1,9840	1,6602	1,2901
150	3,3566	3,1455	2,8482	2,6090	2,3515	1,9759	1,6551	1,2872
200	3,3398	3,1315	2,8385	2,6006	2,3451	1,9719	1,6525	1,2858
250	3,3299	3,1232	2,8222	2,5966	2,3414	1,9695	1,6510	1,2849
300	3,3233	3,1176	2,8279	2,5923	2,3388	1,9679	1,6499	1,2844
400	3,3150	3,1107	2,8227	2,5882	2,3357	1,9659	1,6487	1,2837
500	3,3100	3,1060	2,8190	2,7850	2,3330	1,9640	1,6470	1,2830

Таблица А.2 - Критические значения коэффициента корреляции r

n	p			
	0,001	0,01	0,05	0,1
5	0,991	0,959	0,878	0,805
6	0,974	0,917	0,811	0,729
7	0,951	0,875	0,754	0,669
8	0,925	0,834	0,707	0,621
9	0,898	0,798	0,666	0,582
10	0,872	0,765	0,632	0,549
11	0,847	0,735	0,602	0,521
12	0,823	0,708	0,576	0,497
13	0,801	0,684	0,553	0,476
14	0,780	0,661	0,532	0,458
15	0,760	0,641	0,514	0,441
16	0,742	0,623	0,497	0,426
17	0,725	0,606	0,482	0,412
18	0,708	0,590	0,468	0,400
19	0,693	0,575	0,456	0,389
20	0,679	0,561	0,444	0,378
21	0,665	0,549	0,433	0,369
22	0,652	0,537	0,423	0,360
23	0,640	0,526	0,413	0,352
24	0,629	0,515	0,404	0,344
25	0,618	0,505	0,396	0,337
26	0,607	0,496	0,388	0,330
27	0,597	0,487	0,381	0,323

Продолжение таблицы А.2

n	p			
	0,001	0,01	0,05	0,1
28	0,588	0,479	0,374	0,317
29	0,579	0,471	0,367	0,311
30	0,570	0,463	0,361	0,306
31	0,562	0,456	0355	0,301
32	0,554	0,449	0,349	0,296
33	0,547	0,442	0,344	0,291
34	0,539	0,436	0,339	0,287
35	0,532	0,430	0,334	0,283
36	0,525	0,424	0,329	0,279
37	0,519	0,418	0,325	0,275
38	0,513	0,413	0,320	0,271
39	0,507	0,408	0,316	0,267
40	0,501	0,403	0,312	0,264
41	0,495	0,398	0,308	0,260
42	0,490	0,393	0,304	0,257
43	0,484	0,389	0,301	0,254
44	0,479	0,384	0,297	0,251
45	0474	0,380	0,294	0,248
46	0,469	0,376	0,291	0,246
47	0,465	0,372	0,288	0,243
48	0,460	0,368	0,285	0,240
49	0,456	0,365	0,282	0,238
50	0,451	0,361	0,279	0,235

Продолжение таблицы А.2

n	p			
	0,001	0,01	0,05	0,1
51	0,447	0,358	0,276	0,233
52	0,443	0,354	0,273	0,231
53	0,439	0,351	0,271	0,228
54	0,435	0,348	0,268	0,226
55	0,432	0,345	0,266	0,224
56	0,428	0,341	0,263	0,222
57	0,424	0,339	0,261	0,220
58	0,421	0,336	0,259	0,218
59	0,418	0,333	0,256	0,216
60	0,414	0,330	0,254	0,214
61	0,411	0,327	0,252	0,213
62	0,408	0,325	0,250	0,211
63	0,405	0,322	0,248	0,209
64	0,402	0,320	0,246	0,207
65	0,399	0,317	0,244	0,206
66	0,396	0,315	0,242	0,204
67	0,393	0,313	0,240	0,203
68	0,390	0,310	0,239	0,201
69	0,388	0,308	0,237	0,200
70	0,385	0,306	0,235	0,198
80	0,361	0,286	0,220	0,185
90	0,341	0,270	0,207	0,174

Продолжение таблицы А.2

n	p			
	0,001	0,01	0,05	0,1
100	0,324	0,256	0,197	0,165
110	0,310	0,245	0,187	0,158
120	0,297	0,234	0,179	0,151
130	0,285	0,225	0,172	0,145
140	0,275	0,217	0,166	0,140
150	0,266	0,210	0,160	0,135
200	0,231	0,182	0,139	0,117
250	0,207	0,163	0,124	0,104
300	0,189	0,149	0,113	0,095
350	0,175	0,138	0,105	0,088
400	0,164	0,129	0,098	0,082
450	0,155	0,121	0,092	0,078
500	0,147	0,115	0,088	0,074
600	0,134	0,105	0,080	0,067
700	0,130	0,101	0,078	0,062

Таблица А.3 - Критические значения критерия Фишера F для уровней статистической значимости p <= 0,05 (df1 - число степеней свободы в числителе, df2 - число степеней свободы в знаменателе)

df2	df1									
	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,36	2,30

Продолжение таблицы А.3

df2	df1								
	11	12	14	16	18	20	30	40	50
1	243	244	245	246	248	250	251	252	253
2	19,40	19,41	19,42	19,43	19,44	19,48	19,47	19,47	19,49
3	8,76	8,74	8,71	8,69	8,66	8,62	8,60	8,58	8,56
4	5,93	5,91	5,87	5,84	5,80	5,74	5,71	5,70	5,66
5	4,70	4,68	4,64	4,60	4,56	4,50	4,46	4,44	4,40
6	4,03	4,00	3,96	3,92	3,87	3,81	3,77	3,75	3,71
7	3,60	3,57	3,52	3,49	3,44	3,38	3,34	3,32	3,28
8	3,31	3,28	3,23	3,20	3,15	3,08	3,05	3,03	2,98
9	3,10	3,07	3,02	2,98	2,93	2,86	2,82	2,80	2,76
10	2,94	2,91	2,86	2,82	2,77	2,70	2,67	2,64	2,59
11	2,82	2,79	2,74	2,70	2,65	2,57	2,53	2,50	2,45
12	2,72	2,69	2,64	2,60	2,54	2,46	2,42	2,40	2,35
13	2,63	2,60	2,55	2,51	2,46	2,38	2,34	2,32	2,26
14	2,56	2,53	2,48	2,44	2,39	2,31	2,27	2,24	2,19
15	2,51	2,48	2,43	2,39	2,33	2,25	2,21	2,18	2,12
16	2,45	2,42	2,37	2,33	2,28	2,20	2,16	2,13	2,07
17	2,41	2,38	2,33	2,29	2,23	2,15	2,11	2,08	2,02
18	2,58	2,51	2,46	2,41	2,19	2,11	2,07	2,04	1,98
19	2,55	2,48	2,43	2,38	2,15	2,07	2,02	2,00	1,94
20	2,52	2,45	2,40	2,35	2,12	2,04	1,99	1,96	1,90
21	2,49	2,42	2,37	2,32	2,09	2,00	1,96	1,93	1,87
22	2,47	2,40	2,36	2,30	2,07	1,98	1,93	1,91	1,84

Продолжение таблицы А.3

df2	df1									
	1	2	3	4	5	6	7	8	9	10
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,38	2,32	2,28
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26
25	4,24	3,38	2,99	2,76	2,60	2,49	2,41	2,34	2,28	2,24
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,30	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16
32	4,15	3,30	2,90	2,67	2,51	2,40	2,32	2,25	2,19	2,14
34	4,13	3,28	2,88	2,65	2,49	2,38	2,30	2,23	2,17	2,12
36	4,11	3,26	2,86	2,63	2,48	2,36	2,28	2,21	2,15	2,10
38	4,10	3,25	2,85	2,62	2,46	2,35	2,26	2,19	2,14	2,09
40	4,08	3,23	2,84	2,61	2,46	2,34	2,25	2,18	2,12	2,07
42	4,07	3,22	2,83	2,59	2,44	2,32	2,24	2,17	2,11	2,06
44	4,06	3,21	2,82	2,58	2,43	2,31	2,23	2,16	2,10	2,05
46	4,05	3,20	2,81	2,57	2,42	2,30	2,22	2,14	2,09	2,04
48	4,04	3,19	2,80	2,56	2,41	2,30	2,21	2,14	2,08	2,03
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02
55	4,02	3,17	2,78	2,54	2,38	2,27	2,18	2,11	2,05	2,00
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99
65	3,99	3,14	2,75	2,51	2,36	2,24	2,15	2,08	2,02	1,98
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97
80	3,96	3,11	2,72	2,48	2,33	2,21	2,12	2,05	1,99	1,95

Продолжение таблицы А.3

df2	df1								
	11	12	14	16	18	20	30	40	50
23	2,45	2,38	2,32	2,28	2,04	1,96	1,91	1,88	1,82
24	2,43	2,36	2,30	2,26	2,02	1,94	1,89	1,86	1,80
25	2,41	2,34	2,28	2,24	2,00	1,92	1,87	1,84	1,77
26	2,39	2,32	2,27	2,22	1,99	1,90	1,85	1,82	1,76
27	2,37	2,30	2,25	2,20	1,97	1,88	1,84	1,80	1,74
28	2,36	2,29	2,24	2,19	1,96	1,87	1,81	1,78	1,72
29	2,35	2,28	2,22	2,18	1,94	1,85	1,80	1,77	1,71
30	2,34	2,27	2,21	2,16	1,93	1,84	1,79	1,76	1,69
32	2,32	2,25	2,19	2,14	1,91	1,82	1,76	1,74	1,67
34	2,30	2,23	2,17	2,12	1,89	1,80	1,77	1,71	1,64
36	2,28	2,21	2,15	2,10	1,87	1,78	1,72	1,69	1,62
38	2,26	2,19	2,14	2,09	1,85	1,76	1,71	1,67	1,60
40	2,04	2,00	1,95	1,90	1,84	1,74	1,69	1,68	1,59
42	2,02	1,99	1,94	1,89	1,82	1,73	1,68	1,64	1,57
44	2,01	1,98	1,92	1,88	1,81	1,72	1,66	1,63	1,56
46	2,00	1,97	1,91	1,87	1,80	1,71	1,65	1,62	1,54
48	1,99	1,96	1,90	1,86	1,79	1,70	1,64	1,61	1,53
50	1,98	1,95	1,90	1,85	1,78	1,69	1,63	1,60	1,52
55	1,97	1,93	1,88	1,83	1,76	1,67	1,61	1,58	1,50
60	1,95	1,92	1,86	1,81	1,75	1,65	1,59	1,56	1,48
65	1,94	1,90	1,85	1,80	1,73	1,63	1,57	1,54	1,46
70	1,93	1,89	1,84	1,79	1,72	1,62	1,56	1,53	1,45
80	1,91	1,88	1,82	1,77	1,70	1,60	1,54	1,51	1,42

Продолжение таблицы А.3

df2	df1									
	1	2	3	4	5	6	7	8	9	10
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92
125	3,92	3,07	2,68	2,44	2,29	2,17	2,08	2,01	1,95	1,90
150	3,91	3,06	2,67	2,43	2,27	2,16	2,07	2,00	1,94	1,89
200	3,89	3,04	2,65	2,41	2,26	2,14	2,05	1,98	1,92	1,87
400	3,86	3,02	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85
600	3,85	3,01	2,62	2,39	2,22	2,11	2,02	1,95	1,89	1,85
1000	3,85	3,00	2,61	2,38	2,22	2,10	2,02	1,95	1,89	1,84
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83

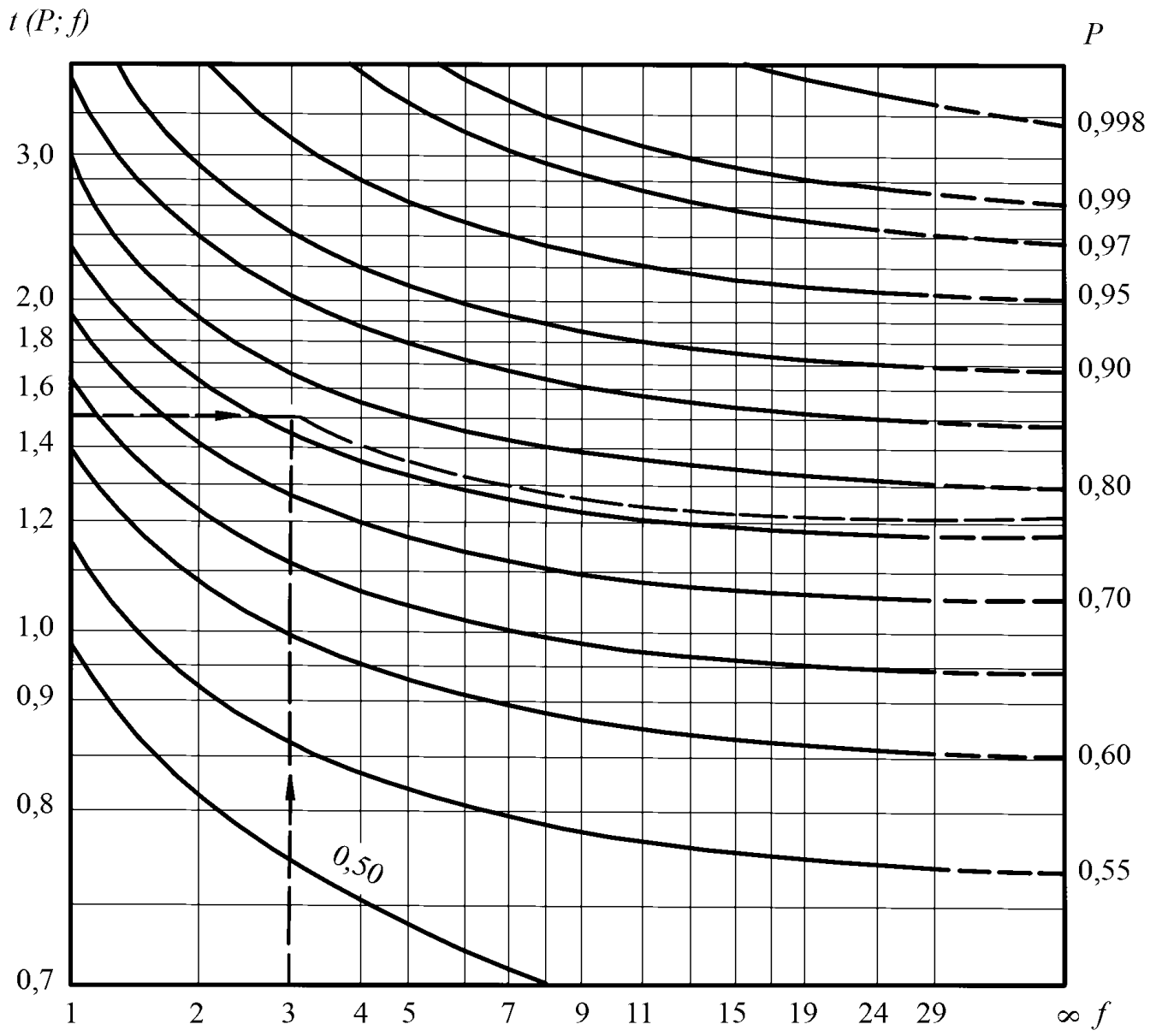


Рисунок А.1 – Зависимость критерия Стьюдента от доверительной вероятности Р и числа степеней свободы f

Основные формулы, используемые при обработке результатов исследований

Точность измерений оценивают с помощью следующих критериев, разработанных для малого числа определений.

1. Выборочное среднее (среднее арифметическое) – математическое ожидание (используется для простейшего прогнозирования)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \qquad \text{(A.1)}$$

где n – число измерений.

2. Единичные отклонения – отклонения отдельных измерений от среднего арифметического

$$E_i = x_i - \bar{x} \,. \qquad \text{(A.2)}$$

Алгебраическая сумма единичных отклонений $\sum E_i$ равна нулю.

3. Выборочная дисперсия (рассеяние)

$$S^2 = \frac{\sum_{i=1}^{n} (\bar{x} - x_i)^2}{k}, \qquad \text{(A.3)}$$

где $k = n - 1$, если $0 < n < 50$, при $n \geq 50$ параметр $k = n$.

4. Выборочное среднеквадратичное (стандартное) отклонение

$$S = \sqrt{S^2} \,. \qquad \text{(A.4)}$$

5. Коэффициент вариации (относительное стандартное отклонение)

$$W = S \cdot 100 / \bar{x}. \qquad \text{(A.5)}$$

6. Выборочная дисперсия среднего значения

$$S_{\bar{x}}^2 = \sum_{i=1}^{n} E_i^2 / n(n-1). \qquad \text{(A.6)}$$

7. Средняя квадратичная ошибка среднего арифметического или стандартное отклонение среднего результата

$$S_{\bar{x}} = \sqrt{S_{\bar{x}}^2}\ ;\ S_{\bar{x}} = S / \sqrt{n}\ . \qquad \text{(A.7)}$$

8. Точность измерения среднего результата

$$E_\alpha = t_\alpha S_{\bar{x}}, \qquad \text{(A.8)}$$

где α - коэффициент надежности, принимают равным 0,95;

t_α - коэффициент Стьюдента или коэффициент нормированных отклонений.

9. Интервальные значения среднего результата

$$\bar{x} \pm E_\alpha. \qquad \text{(A.9)}$$

10. Относительная погрешность среднего результата, %,

$$\pm E_\alpha \cdot 100 / \bar{x}. \qquad \text{(A.10)}$$

11. Числовая характеристика, выражающая линейную взаимосвязь двух случайных величин Y и X по совместным наблюдениям $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, называется коэффициентом корреляции

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{(A.11)}$$

Учебное пособие

Павел Викторович Медведев
Виталий Анатольевич Федотов

МАТЕМАТИЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

ISBN 978-5-7410-1772-2