МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
И ПРОДОВОЛЬСТВИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Учреждение образования
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ
АГРАРНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Н. Г. Серебрякова, А. П. Мириленко

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА И ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТА

Рекомендовано Учебно-методическим объединением по аграрному техническому образованию в качестве пособия для студентов учреждений высшего образования по группе специальностей 74 80 «Научные исследования и разработки, преподавание» и специальности 1-59 80 01 «Охрана труда и эргономика»

Минск
БГАТУ
2022

УДК 311(075)
ББК 60.6я7
С32

Рецензенты:
кафедра «Техническая эксплуатация автомобилей»
Белорусского национального технического университета
(кандидат технических наук, доцент,
заведующий кафедрой *А. С. Гурский*);
кандидат физико-математических наук, доцент,
заведующий кафедрой высшей математики
УО «Белорусский государственный университет
информатики и радиоэлектроники» *Е. А. Баркова*

Серебрякова, Н. Г.

С32 Статистические методы анализа и планирования эксперимента : пособие / Н. Г. Серебрякова, А. П. Мириленко. – Минск : БГАТУ, 2022. – 104 с.

ISBN 978-985-25-0149-1.

В пособии рассмотрены основные классы задач статистического анализа данных и современные технологии их решения, проведен обзор способов ввода исходных данных для статистического анализа, предложены методика подготовки данных к анализу и их визуализации, способы вывода результатов статистического анализа в виде таблиц и графиков, основные статистические модули и процедуры.

Для студентов учреждений высшего образования по группе специальностей 74 80 «Научные исследования и разработки, преподавание» и специальности 1-59 80 01 «Охрана труда и эргономика», может быть использовано магистрантами других специальностей, студентами, аспирантами и соискателями ученой степени.

УДК 311(075)
ББК 60.6я7

ISBN 978-985-25-0149-1

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 5
1. ДАННЫЕ, ОСНОВНЫЕ ПОНЯТИЯ
1.1. Типы данных 6
1.2. Кодификатор 8
1.3. Подготовка данных к статистическому анализу 10
2. ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА
2.1. Тенденции и меры 13
2.2. Оценка параметров генеральной совокупности 20
2.3. Доверительный интервал 22
3. ИССЛЕДОВАНИЕ ДАННЫХ НА ЭВМ
3.1. Описательная статистика 24
3.2. Визуальный анализ данных 29
4. СРАВНЕНИЕ ДАННЫХ
4.1. Понятие о статистической значимости
и практической значимости 32
4.2. Нулевая и альтернативная гипотезы 33
5. СРАВНЕНИЕ ДАННЫХ
С ИСПОЛЬЗОВАНИЕМ ЭВМ
5.1. Основные статистики 36
5.2. Непараметрические статистики 38
5.3. Сравнение двух независимых групп 40
6. ПОГРЕШНОСТИ, ОКРУГЛЕНИЕ
6.1. Вычисление и корректная запись приближенных чисел 42
6.2. Погрешность числа 44
6.3. Погрешность косвенного измерения 45
6.4. Порядок округления величин 47
7. СТАТИСТИЧЕСКАЯ СВЯЗЬ
7.1. Факторы и отклик 49
7.2. Типы связей и способы их описания 50
7.3. Корреляция 54
8. АНАЛИЗ СВЯЗЕЙ
ПРИ ДИСКРЕТНОМ ОТКЛИКЕ
8.1. Дискриминантный анализ 58
8.2. Кластеризация 59
8.3. Логистическая регрессия 60
8.4. ROC-анализ 66

9. РЕГРЕССИЯ
9.1. Метод наименьших квадратов 71
9.2. Множественная регрессия 74
9.3. Доверительный интервал отклика 75
10. ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА
10.1. Активный эксперимент 77
10.2. Факторы и уровни 77
10.3. Полный факторный план 80
11. ДРОБНЫЙ ПЛАН ЭКСПЕРИМЕНТА.
ПЛАНЫ ВТОРОГО ПОРЯДКА.
ЦЕНТРАЛЬНЫЙ КОМПОЗИЦИОННЫЙ ПЛАН
11.1. Дробный факторный план 83
11.2. Планы второго порядка 84
11.3. Центральный композитный план 85
11.4. Расчет факторных планов на ЭВМ 86
12. ПОНЯТИЕ О СПЕЦИАЛЬНЫХ ПЛАНАХ
ЭКСПЕРИМЕНТА
12.1. Планы для смесей 96
12.2. Планы с ограничениями 97
12.3. Многофакторный отсеивающий план 99
12.4. Методы Тагучи 100
СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ 102

ВВЕДЕНИЕ

Поскольку сегодня информация становится частью действительности, без адекватных технологий анализа данных человек оказывается беспомощным в информационной среде. Статистика позволяет компактно описать данные, понять их структуру, провести классификацию, увидеть закономерности в потоке случайных результатов эксперимента.

Особенность пособия заключается в том, что в нем всесторонне описано применение разнообразных методов статистического анализа данных.

Классификация всевозможных методов анализа данных позволяет пользователю, используя систему STATISTICA, свободно применять на практике эти методы и работать как с собственными данными, так и с предлагаемыми.

Пособие содержит:

– общие представления о методологии анализа данных;

– конкретные техники осуществления расчетов;

– правила оформления и интерпретации результатов.

Предполагается, что большинство расчетов будет производиться в пакетах прикладных программ, главным образом, STATISTICA.

Значительное внимание уделено понятийным и методологическим аспектам математической статистики, поскольку трудности в их освоении совершенно обычны и объективно обоснованы. В научном сообществе распространено мнение, что данный предмет по своему существу является *контринтуитивным*, т. е. оценка результата из общих соображений и за счет предыдущего опыта преимущественно оказывается ошибочной.

В пособии не приводятся формулы и предназначенные для ручных расчетов алгоритмы, поскольку в настоящее время они не имеют практической значимости и при необходимости могут быть получены из сети Интернет.

1. ДАННЫЕ, ОСНОВНЫЕ ПОНЯТИЯ

1.1. Типы данных

Данные – это факты и идеи, представленные в виде, пригодном для анализа средствами математики.

Следовательно, необходимо пользоваться только такой информацией, которая изначально пригодна для средств математики или может быть преобразована к пригодности. Например, повествование «Хождение за три моря», являющееся информацией, для этих целей не может быть признано данными, поскольку к математическому виду не приводится.

Выборка и случай

Очевидно, что статистический анализ данных всегда имеет дело с некоторыми выборками. Анализ отдельного случая также может быть методом научного исследования и может иметь определенную научную ценность, однако средства математического анализа данных для него неприменимы.

Таким образом, всегда будет использоваться выборка, которая будет состоять из некоторого количества случаев. Также необходимо помнить, что полученные результаты относятся исключительно к исследованной выборке, а распространение их на всю реальность является отдельной задачей.

Однородность данных и табличная форма

Данные будут использоваться в виде выборки, состоящей из случаев, но необходимо добавить еще 3 требования:

1. Все случаи должны быть извлечены из одной и той же генеральной совокупности.

2. Все случаи должны описываться одинаковыми наборами данных.

3. Данные выборки должны представляться в табличном виде, где каждая строка – это отдельный случай с набором характеризующих его элементов информации, каждый столбец – переменная, определяющая содержание элемента.

Предложенные требования не встретят проблем с пониманием, поскольку дословно соответствуют представлению данных в хорошо известном табличном процессоре Excel.

Терминология. В различных источниках и программах применяются разные слова для обозначения случаев и переменных, поэтому следует запомнить правила:

объект = случай = наблюдение = *case*,

поле = переменная = *variable*.

В различных контекстах переменная также может называться «фактор» или «предиктор».

Типы данных

Точное и правильное понимание типа используемых данных необходимо для работы с методами статистического анализа, поскольку для разных типов данных применяются разные методы и критерии. Ошибка в определении типа поставит под сомнение полученные результаты. К сожалению, в терминологии и классификации тоже есть существенные разночтения, поэтому схема типологии данных представлена в наиболее удобном виде.

Целесообразно разделить типы данных на *количественные* и *качественные*. Количественные данные изначально характеризуются численно, качественные говорят о принадлежности к некоторым категориям. Основные типы данных делятся на подтипы:

– *количественные*:

интервальные,

относительные;

– *качественные*:

номинальные,

дихотомические,

порядковые.

Интервальные данные описывают свойство численно, однако шкала измерения не позволяет оценить отношение величин, например: шкала температуры Цельсия, по которой нельзя сказать, что 30 градусов в 3 раза больше, чем 10.

Относительные – данные, для которых шкала начинается с нуля и позволяет записать отношение величин, например: шкала температуры Кельвина или длина, измеренная рулеткой.

Номинальные – данные, которые нельзя упорядочить: пол, национальность, цвет, город и т. д.

Дихотомические – подтип номинальных данных, у которых есть только две возможные категории: мужчины и женщины, а также любой признак наличия-отсутствия.

Порядковые данные также говорят о принадлежности к категориям, но к таким, что их можно выстроить по увеличению некоторого свойства. Например: типы самолетов (легкие – средние – тяжелые) выстроены по увеличению массы.

Варианты терминологии. Качественные данные также могут называться категориальными и атрибутивными; дихотомические – бинарными.

Количественные данные иногда делят на непрерывные (те, что могут иметь дробную часть) и дискретные (которые могут принимать только целочисленные значения), например, количество колес.

Данные можно:

- описать,
- показать,
- исследовать,
- сравнить,
- анализировать их связи,
- моделировать,
- прогнозировать,
- диагностировать,
- представить результат.

К тому же сделать это:

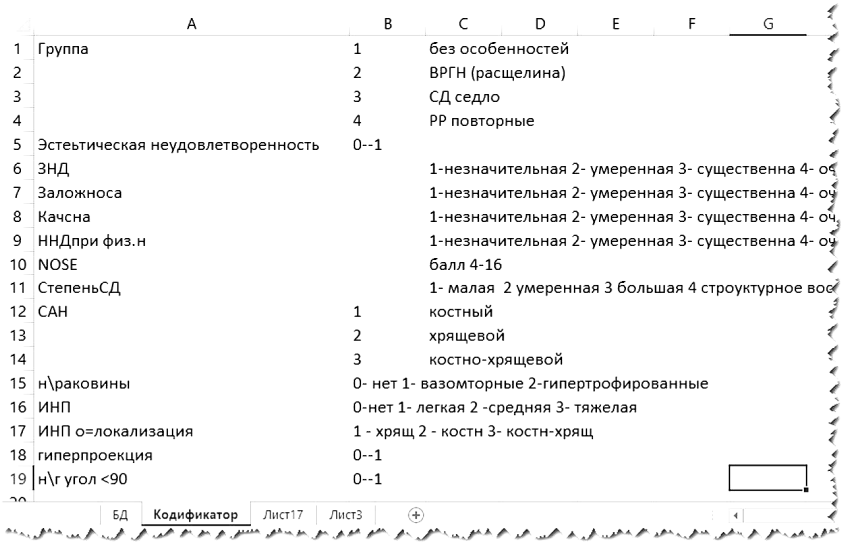
- правильно,
- методологически корректно,
- красиво,
- убедительно,
- современно.

1.2. Кодификатор

Кодификатор – свод правил по преобразованию данных в вид, пригодный для анализа средствами математики, т. е. правил, по которым делается перевод данных в численный вид. В кодификаторе перечислены переменные, и для каждой из них указан способ преобразования. Например, для переменной «Пол» правило выглядит так: «М-1; Ж-2»; значит, если объект мужского пола, то записывают 1, если женского – 2.

Кодификация может иметь более сложные правила, но все они должны представлять данные в численном виде. Для изначально количественных переменных в кодификаторе обычно указывают единицы измерения.

Практически кодификатор представляет собой отдельный лист в Excel (рис. 1.1).



	A	B	C	D	E	F	G
1	Группа	1	без особенностей				
2		2	ВРГН (расщелина)				
3		3	СД седло				
4		4	РР повторные				
5	Эстеьтическая неудовлетворенность	0--1					
6	ЗНД		1-незначительная 2- умеренная 3- существенна 4- оч				
7	Заложноса		1-незначительная 2- умеренная 3- существенна 4- оч				
8	Качсна		1-незначительная 2- умеренная 3- существенна 4- оч				
9	ННДпри физ.н		1-незначительная 2- умеренная 3- существенна 4- оч				
10	NOSE		балл 4-16				
11	СтепеньСД		1- малая 2 умеренная 3 большая 4 строуктурное вос				
12	САН	1	костный				
13		2	хрящевой				
14		3	костно-хрящевой				
15	н\раковины	0- нет 1- вазомторные 2-гипертрофированные					
16	ИНП	0-нет 1- легкая 2 -средняя 3- тяжелая					
17	ИНП о=локализация	1 - хрящ 2 - костн 3- костн-хрящ					
18	гиперпроекция	0--1					
19	н\г угол <90	0--1					

БД | Кодификатор | Лист17 | Лист3

Рис. 1.1. Фрагмент кодификатора

Преобразование данных

Нередко исследование данных требует их преобразования, которое может привести к изменению типа данных.

Основные варианты возможных преобразований:

1. Количественное поле может быть преобразовано:
– в количественное (log(x));
– дихотомическое (норма–не норма);
– порядковое (легкий–средний–тяжелый).
2. Порядковое поле может быть преобразовано в количественное:
– легкий = 1;
– средний = 2;
– тяжелый = 5.

Количественное поле может быть также преобразовано:

– в количественное поле по некоторой формуле, например, логарифма или возведения в квадрат. Применяется для более удобного представления или поиска линейных зависимостей;

– дихотомическое поле. Это всем знакомое поле «норма–не норма»;

– порядковое поле. Иногда численные данные являются избыточными, и требуется выделение категорий в виде порядкового поля.

Порядковое поле можно преобразовать в количественное, считая номера категорий числами. Однако целесообразно сделать «нелинейный» перевод, чтобы точнее отразить степень выраженности свойства.

1.3. Подготовка данных к статистическому анализу

Первоначальный этап подготовки к статистическому анализу:

1. Определение с набором полей (переменных), характеризующих объект (случай). Традиционно первым полем задается номер объекта, вторым – его название.

2. Создание кодификатора и правил преобразования (кодирования) качественных данных.

3. Осуществление кодирования и получение данных в табличной форме.

4. Обдумывание и осуществление целесообразности преобразования данных.

В абсолютном большинстве случаев подготовка данных производится в программе Excel или аналогичных табличных процессорах Google doc, Open Office.

Далее следует произвести проверку полученной таблицы.

1. Проверка на простые ошибки. Посмотреть максимальные и минимальные величины в столбце и убедиться, что они корректны. Обычно проверку выполняют с помощью сортировки.

2. Проверить форматы ячеек. Часто ячейки, которые выглядят, как числа, оказываются в текстовом формате, и в пакетах прикладных программ они не будут восприниматься.

3. Убедиться, что при отсутствии данных ячейки не заполнены, и в них не стоят нули.

Если все перечисленные процедуры выполнены, можно считать, что данные готовы к импорту в программу STATISTICA.

Для импорта данных надо в программе STATISTICA в меню **File** выбрать **Open**. Появится диалоговое окно (рис. 1.2). Скорее всего, в файле будет несколько листов, поэтому надо выбрать вторую кнопку и указать нужный лист.

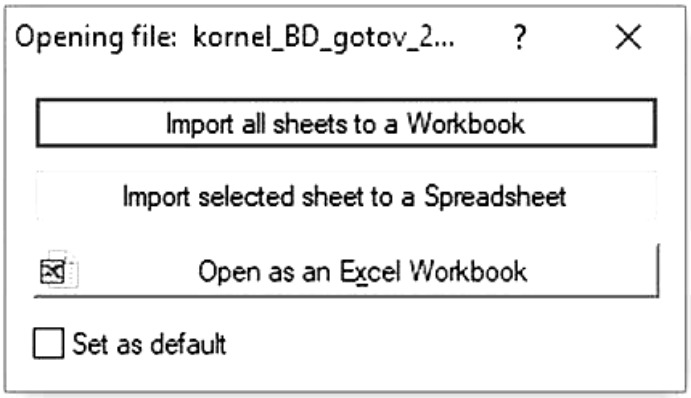


Рис. 1.2. Диалоговое окно «Импорт данных в программе STATISTICA»

В следующей форме (рис. 1.3) надо поставить две галочки, указав, что номера случаев в первом столбце, а названия переменных в первой строке.

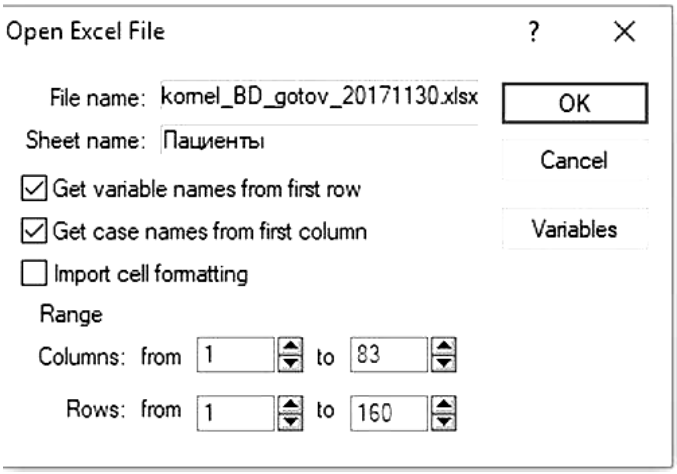
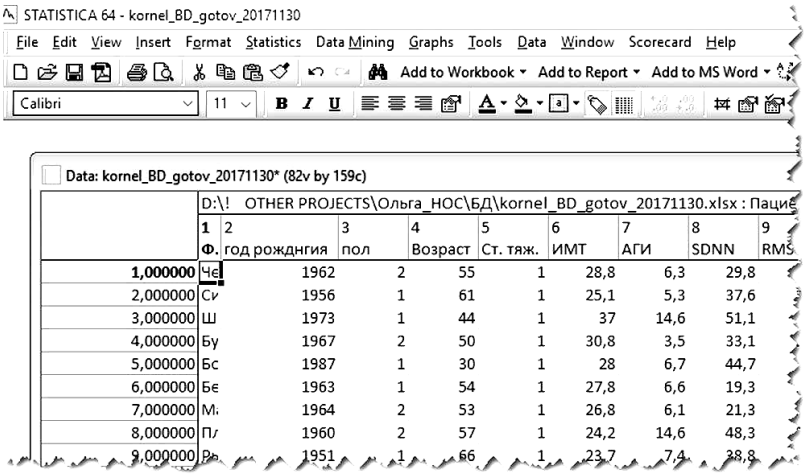


Рис. 1.3. Диалоговое окно с указанием расположения номеров случаев и названий переменных

В результате импорта данные окажутся в программе (рис. 1.4). Далее можно сохранить файл в формате **STA**.



	1 Ф.	2 год рождния	3 пол	4 Возраст	5 Ст. тяж.	6 ИМТ	7 АГИ	8 SDNN	9 RMS
1,000000	Че	1962	2	55	1	28,8	6,3	29,8	
2,000000	Си	1956	1	61	1	25,1	5,3	37,6	
3,000000	Ш	1973	1	44	1	37	14,6	51,1	
4,000000	Бу	1967	2	50	1	30,8	3,5	33,1	
5,000000	Бс	1987	1	30	1	28	6,7	44,7	
6,000000	Бе	1963	1	54	1	27,8	6,6	19,3	
7,000000	М	1964	2	53	1	26,8	6,1	21,3	
8,000000	Пл	1960	2	57	1	24,2	14,6	48,3	
9,000000	Рь	1951	1	66	1	23,7	7,4	38,8	

Рис. 1.4. Фрагмент файла с данными

Таким образом, данные готовы к началу анализа.

2. ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА

2.1. Тенденции и меры

Генеральная совокупность

Генеральная совокупность – все множество объектов, которое необходимо исследовать и относительно которого требуется делать выводы (то, что названо реальностью).

Генеральная совокупность представляет собой субъективное образование, т. е. состав ее напрямую зависит от того, что намерен включать в совокупность исследователь. Поэтому исследование должно начинаться именно с определения границ генеральной совокупности. Обычно это делается с помощью *критериев включения* и *критериев исключения*. В практике научных исследований редко говорят о генеральной совокупности, но почти обязательно описывают *материал исследования*, для которого и указывают критерии включения и исключения. Таким образом, правила включения объектов в выборку для исследования одновременно являются правилами, описывающими границу генеральной совокупности.

После определения критериев включения объектов в исследование производится сбор информации в виде данных об объектах, т. е. формируется исследуемая выборка.

Вариабельность данных

Анализ данных основан на том, что объекты генеральной совокупности имеют различия в свойствах. Несомненно, бывают совокупности объектов, не имеющие различий, например, поверхностное наблюдение говорит, что все императорские пингвины одинаковы. Но такие данные не требуют анализа, поэтому в предложенной методологии свойства исследуемых объектов всегда имеют различия.

Вариабельность данных – различия в однородных исследуемых данных или степень этих различий.

Величины, которые различны при разных наблюдениях, называют *случайными*, или *стохастическими*. По существу, различия обусловлены некоторыми внутренними или внешними причинами, рассмотрение которых не входит в задачи исследования, поэтому в статистическом анализе стохастичность величины просто считается присущим ей свойством.

Рассматривая причины различий в данных, можно выделить две основные: фактическая собственная вариабельность величины и вариабельность ее измерения.

Таким образом, в наблюдении имеются две составляющие, соотношение которых в общем случае может быть произвольным. Однако в практике научных исследований редко встречаются случаи, когда вариабельности измерительной системы и собственной вариабельности величины соизмеримы. Обычно имеют место предельные случаи, когда одна из составляющих намного больше другой. Соответственно, различны методы исследования.

В инженерных системах обычно собственная изменчивость величин невелика, а различия определяются случайной ошибкой измерения. Такие процессы и системы называют *детерминированными*, и их исследование сосредоточено на оценке точности измерения.

В сложных системах (биологических, социальных), наоборот, измерения достаточно точны, а изменчивость системы связана множеством неизвестных внутренних причин. Такие системы и описывающие их модели называют *стохастическими*.

Количественные и качественные данные

Представления количественных и качественных данных в исследуемых выборках различны.

Количественные данные всегда выражены числом, и эти числа для разных объектов различны. Следовательно, для выборки количественные данные по некоторой переменной представлены в виде набора чисел, причем их количество равно количеству объектов в выборке. Например, рост студентов (см): 168, 176, 184, 174 и т. д. Значит, в другой выборке будет другой набор чисел.

Качественные данные всегда указывают на принадлежность к категориям, поэтому для выборки качественные данные представляются как количества объектов, принадлежащих к каждой категории. Например, студенты приехали из городов: Минск – 5 чел., Гродно – 3, Брест – 4 чел. и т. д. В этом примере сумма чисел будет равна количеству объектов в выборке. Кроме абсолютных количеств, качественные данные также представляются выборочными долями.

Выборочная доля – отношение количества объектов, принадлежащих к категории, к общему количеству объектов. Например: Минск – 5/20, Гродно – 3/20 или, соответственно, 25 % и 15 %.

Тенденция и мера

Очевидно, что оперировать данными как списком чисел неудобно и неперспективно, поэтому в математической статистике введены обобщающие понятия: центральная тенденция и мера рассеяния.

Центральная тенденция

Центральная тенденция – число, описывающее всю выборку, то или иное понимание ее центра.

Для характеристики центральной тенденции используют две основные величины – среднее и медиана.

Среднее ($\bar{x}$) – величина, которая представляет собой среднее арифметическое всех наблюдений и вычисляется по формуле

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \mathbf{K} + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Иногда среднее обозначают буквой μ.

Медиана (Me) – это такое число, в котором половина из элементов выборки больше него, а другая половина меньше. Для нахождения Me выстраивают все числа по увеличению и находят расположенное посередине:

Медиана

9,5

1 3 3 8 8 9 10 11 12 14 26

К центральным тенденциям также относится *мода* – величина, на которую приходится максимальное количество наблюдений.

Мера рассеяния

Мера рассеяния – число, характеризующее степень различий в данных. Мера рассеяния имеет несколько вариантов.

Рассеяние также характеризуется различными способами. Это – величины дисперсии и перцентиля.

Дисперсия – величина, характеризующая квадратичное отклонение наблюдений от среднего и вычисляемая по формуле

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}.$$

Величина *S* называется *среднеквадратическим отклонением*.

Перцентиль – величина, в сторону меньшего значения которой попадает заданная доля наблюдений. Для практических целей используют *нижний квартиль* и *верхний квартиль*.

Нижний (LQ) *и верхний* (UQ) *квартили* – это 25%-ный перцентиль и 75%-ный перцентиль. Соответственно, 25 % наблюдений попадет ниже нижнего квартиля, и 25 % – выше верхнего. Медиана является 50%-ным перцентилем.

Собственно характеристикой рассеяния является расстояние от нижнего квартиля до верхнего – *интерквартильный размах*. Иногда используют термины «интерквартильный диапазон», «межквартильный размах».

К мерам рассеяния также относится термин «размах» – расстояние от минимального значения до максимального.

Mi			LQ				Me				UQ				Ma
n															*x*
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

Термины и обозначения

Применяют следующие термины и обозначения (таблица):

Таблица

Термины и обозначения

Термин		Обозначение
Русский язык	Английский язык	
Среднее	Mean	$\bar{x}$ – выборка, M – генеральная совокупность

Окончание таблицы

Термин		Обозначение
Русский язык	Английский язык	
Дисперсия	Variance	s^2 – выборка, σ^2 – генеральная совокупность
Стандартное отклонение	Standard deviation	SD
Стандартная ошибка (среднего)	Standard error	SE
Медиана (50%-ный перцентиль)	Median	Me
Нижний квартиль (25%-ный перцентиль)	Lower quartile	LQ
Верхний квартиль (75%-ный перцентиль)	Upper quartile	UQ
Размах	Range	R или Min–Max

Функция распределения. Гистограмма

Наиболее удобное и визуально понятное представление о характере стохастической количественной величины можно получить с помощью функции распределения, чаще всего имеющей вид гистограммы.

Функция распределения (плотность распределения) – функция, определяющая вероятность того, что случайная величина примет соответствующее значение.

Гистограмма – это диаграмма, в которой отдельные значения представлены столбцами различной высоты.

Поскольку применяют пакеты прикладных программ, то алгоритм построения гистограммы вручную приведен в самом кратком виде.

Алгоритм построения гистограммы

1. Расчет размаха R из n результатов измерений (размах – это разница между наибольшим X_{max} и наименьшим X_{min} значениями).

2. Определение количества интервалов k по формуле

$$k = 1 + 3{,}3 \cdot \lg n.$$

Обычно $6 < k < 20$.

3. Вычисление ширины h интервалов гистограммы.

4. Расчет границ интервалов: границы интервалов выбирают таким образом, чтобы они не совпадали с результатами измерений и крайние интервалы были заполнены.

5. Подсчет числа попаданий результатов в интервалы: полученные результаты сводят в таблицу.

6. Построение гистограммы.

Результатом будет столбчатая диаграмма, в которой по горизонтальной оси отложена исследуемая величина, по вертикальной – количество наблюдений, попавших в соответствующий интервал.

Огибающая столбчатую диаграмму кривая и будет представлять собой функцию плотности распределения вероятностей. При построении гистограммы в пакете STATISTICA огибающая кривая строится автоматически, над диаграммой выводятся параметры соответствующей функции Гаусса (рис. 2.1).

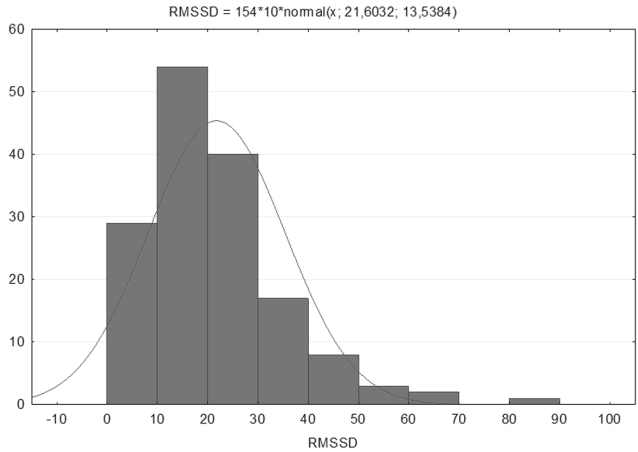


Рис. 2.1. Графическая функция плотности распределения вероятностей

Нормально распределенные количественные данные

Количественные данные могут иметь *нормальное распределение* или не иметь его. И это обстоятельство имеет в математической статистике важнейшее значение, поскольку для нормально распределенных данных существуют отдельные точные методы расчетов и дополнительные возможности по их описанию.

Стохастическая величина имеет *нормальное распределение* в том случае, если ее функция распределения соответствует функции Гаусса:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}},$$

где $f(x)$ – плотность вероятности;

$\bar{x}$ – среднее;

σ^2 – дисперсия.

Функция Гаусса на гистограмме показана линией, и о ней можно сказать, что она симметрична и куполообразна (рис. 2.1).

Фактическая гистограмма не обязательно должна точно совпадать с функцией Гаусса, достаточно, чтобы она соответствовала ей приблизительно. Для определения степени соответствия существуют специальные критерии, общее их количество – не менее 20. Но для практических целей обычно используют 3 критерия, так как они непосредственно рассчитываются в программе STATISTICA:

– Шапиро–Уилка;

– Колмогорова–Смирнова;

– Лиллиефорса.

Вывод о нормальности распределения делается по рассчитанному *уровню значимости*. Если уровень значимости $p > 0{,}05$, то считается, что значимые отличия от нормальности не обнаружены. Этот факт позволяет считать случайную величину нормально распределенной и применять к ней соответствующие методы, которые называют *параметрическими*.

Параметрические свойства нормального распределения

Нормальное распределение описывается функцией, которая основана на параметрах среднего значения и дисперсии. Следовательно,

функцию распределения можно вычислить, что дает возможность рассчитать аналитически вероятность попадания наблюдения в некоторый интервал значений.

Прежде всего, надо знать, какое количество наблюдений попадает в интервалы, кратные среднеквадратическому отклонению.

В диапазон $\pm\sigma$ попадает 68 % наблюдений, в $\pm 2\sigma$ – 95 %, в диапазон $\pm 3\sigma$ – 99 %.

В то же время можно рассчитать вероятность попадания наблюдения в произвольный диапазон. Для этого рассчитывают величину Z (т. е. отношение отклонения значения от среднего к среднеквадратической дисперсии) и определяют вероятность по таблицам критерия Z.

Уровень значимости

Важную роль в математической статистике играет диапазон значений, выпадающий за $\pm 2\sigma$. Очевидно, что вероятность получения значения большего, чем $\bar{x} + 2\sigma$, или меньшего, чем $\bar{x} - 2\sigma$, составляет только 5 %. Этот факт используют при оценке вероятности ошибки статистического вывода. То есть вывод о некотором различии делают только тогда, когда вероятность ошибки меньше 5 %. Записывают так: $p < 5$ % или $p < 0{,}05$ и называют эту величину *уровнем значимости*.

2.2. Оценка параметров генеральной совокупности

В большинстве исследований анализ выборки осуществляется для того, чтобы получить данные о генеральной совокупности. В генеральной совокупности есть некоторое неизвестное *среднее значение величины* и есть неизвестная *дисперсия* генеральной совокупности. Исследуя выборки, можно получить значения среднего и дисперсии для выборок, однако они не равны в точности величинам генеральной совокупности.

Поэтому перенос данных, полученных для выборки, на генеральную совокупность называют *оценкой параметров генеральной совокупности*. Существуют два метода: точечная оценка и интервальная оценка.

Точечная оценка

Этим термином называют простой способ: считать, что среднее генеральной совокупности равно среднему выборки, дисперсия – дисперсии выборки. Точечные оценки в научном исследовании могут применяться только для предварительных сведений или в случаях, когда не нужен научный вывод.

Интервальная оценка

Интервальная оценка предполагает, что для параметров генеральной совокупности можно найти некоторый интервал, в котором они находятся. Поскольку параметр генеральной совокупности находится в заданном интервале с некоторой вероятностью, следовательно, интервальная оценка одновременно содержит размер интервала и вероятность нахождения в нем искомого значения.

Распределение средних значений выборок. Существенную помощь здесь оказывает тот факт, что средние значения выборок, извлеченных из этой генеральной совокупности, всегда имеют нормальное распределение.

Этот факт имеет место независимо от того, является ли распределение стохастической величины в генеральной совокупности нормальным. Точнее, это справедливо для больших выборок с $n > 30$, для малых выборок применяется распределение Стьюдента, зависящее от n.

То есть, если извлечь из генеральной совокупности n выборок и для каждой из них рассчитать среднее значение, затем собрать все эти значения и посчитать их выборкой, то окажется, что эта выборка всегда распределена нормально, а ее дисперсия в $\sqrt{n}$ раз меньше дисперсии исходных выборок.

Эту дисперсию называют *ошибкой среднего* и обозначают как ОС (SE):

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}.$$

Известная дисперсия средней величины выборок позволяет сделать обратный вывод в отношении генеральной совокупности – ошибка оценки среднего генеральной совокупности также равна SE, следовательно, истинное среднее генеральной совокупности в большинстве случаев будет отличаться от точечной ошибки не более чем на SE.

Некоторое время назад принято было записывать результат исследования в виде $\mu = \overline{x} \pm$ SE, например: рост студентов равен (178 ± 3) см. Однако в последнее время такая запись считается некорректной, и предпочтительно просто указать параметры: рост студентов (SE) составил 176 см (± 3 см).

2.3. Доверительный интервал

В научных исследованиях чаще всего делают интервальную оценку в виде доверительного интервала, обычно – 95%-ный доверительный интервал. Это обусловлено тем, что в диапазон ± SE попадает только 68 % значений, и для научного вывода это – недостаточная надежность. Поэтому в большинстве случаев диапазон расширяют до ± 2SE, и вероятность правильного вывода составляет 95 %, что считается приемлемым.

95 %-ный доверительный интервал среднего, или 95 % ДИ, – это интервал, в котором среднее генеральной совокупности находится с вероятностью 95 % (рис. 2.2).

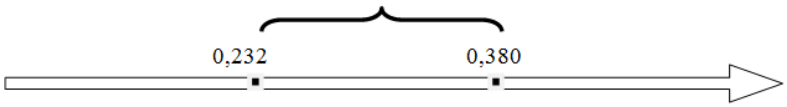


Рис. 2.2. Доверительный интервал среднего

Для различных типов данных существуют соответствующие методы расчета доверительных интервалов.

Результат исследования, в котором оценивался доверительный интервал, записывают так: рост студентов (95 % ДИ) составил 176 см (от 170 до 182 см).

Расчет ДИ вручную. Метод Вальда. 1. Рассчитать среднее значение и стандартное отклонение:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

$$S_0 = \sqrt{\frac{n}{n-1} S^2} = \sqrt{\frac{n}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}.$$

2. Рассчитать ошибку среднего или стандартное отклонение среднего:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}.$$

3. Выбрать доверительную вероятность – 0,9; 0,95; 0,99 и для $n \le 30$ найти *t*-критерий Стьюдента с $(n-1)$ степенями свободы.

Для $n > 30$ – критерий *Z*, который, соответственно, равен 1,65; 1,96; 2,58.

4. Записать доверительный интервал:

нижняя граница: $\overline{x} - Z \cdot \mathrm{SE}_{\overline{x}}$,

верхняя граница: $\overline{x} - Z \cdot \mathrm{SE}_{\overline{x}}$,

5. Написать ответ (например):

«Масса коровы равна 630 кг (95 % ДИ от 605 до 655 кг)».

Доверительный интервал медианы. В известных программах ДИ медианы не рассчитывается, поэтому при необходимости нижний и верхний пределы можно вычислить по формулам:

$$L = \frac{n}{2} - \left(z_{1-\alpha} \cdot \frac{\sqrt{n}}{2} \right),$$

$$U = 1 + \frac{n}{2} + \left(z_{1-\alpha} \cdot \frac{\sqrt{n}}{2} \right).$$

Тем самым вычислив предельные значения медианы, которые с заданной доверительной вероятностью будут находиться в этом интервале при выборке большого объема.

3. ИССЛЕДОВАНИЕ ДАННЫХ НА ЭВМ

3.1. Описательная статистика

Анализ данных проводят, используя файл с данными в формате **.*.sta**, в котором в первом столбце – идентификатор (номера) случаев, в первой строке – названия переменных.

Основные статистики. Описание данных

Исследование данных начинается с их описания, соответствующий раздел называется «*описательная статистика*». В описательной статистике исследуются тенденции, меры и другие свойства выборки, и при необходимости производится оценка параметров генеральной совокупности.

Описательная статистика количественных данных

Описательная статистика количественных данных расположена в меню по адресу **Statistics>Basic statistics>Descriptive statistics.**

При выборе этого пункта меню открывается диалоговое окно (рис. 3.1).

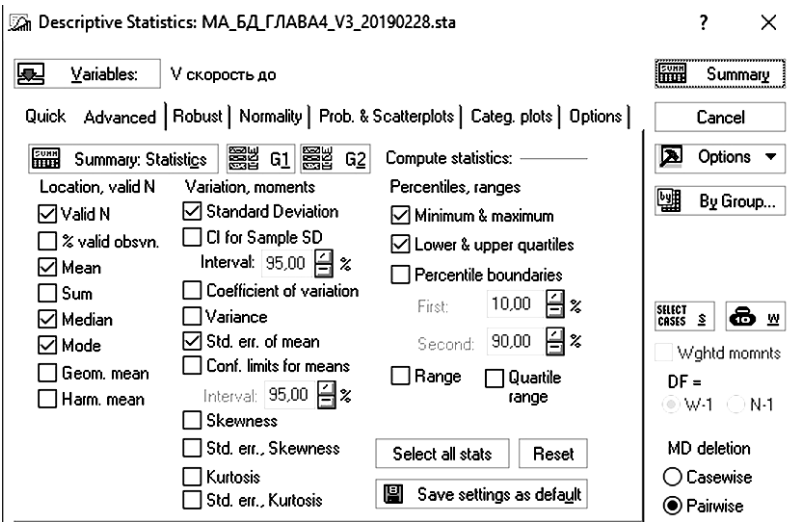


Рис. 3.1. Диалоговое окно «Описательная статистика количественных данных»

На вкладке **Advanced** выбирают переменные для анализа и указывают, какие данные должны быть вычислены. Например, в окне (рис. 3.1) показан выбор количества наблюдений *N*, среднего,

медианы, моды, среднеквадратического отклонения, стандартной ошибки среднего, минимума и максимума, нижнего и верхнего квартилей. Это – обычный максимальный набор описательных данных.

Обратите внимание: на вкладке **Advanced** также можно выбрать произвольные перцентили.

Результат расчета выдается в виде таблицы (рис. 3.2).

	Descriptive Statistics (МА_БД_ГЛАВА4_V3_20190228.sta)									
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile	Std.Dev.
V скорость до	101	278,0792	281,0000	Multiple	4	222,0000	332,0000	258,0000	298,0000	26,30805

Рис. 3.2. Результаты описательной статистики количественных данных

Доверительный интервал среднего. В этой вкладке можно найти доверительный интервал среднего, для чего надо поставить галочку в пункте **Conf. limits for mean** и оставить уровень 95 % (рис. 3.3):

Descriptive Statistics (IK_Глваа4_BD_SKV_V2)			
	Mean	**Confidence SD - -95,000%**	**Confidence SD - +95,000%**
скф	97,86857	27,06883	43,84579

Рис. 3.3. Результаты вычисления доверительного интервала среднего

Проверка на нормальность. В этом же диалоговом окне можно проверить данные на нормальность распределения, выбрав вкладку **Normality** (рис. 3.4).

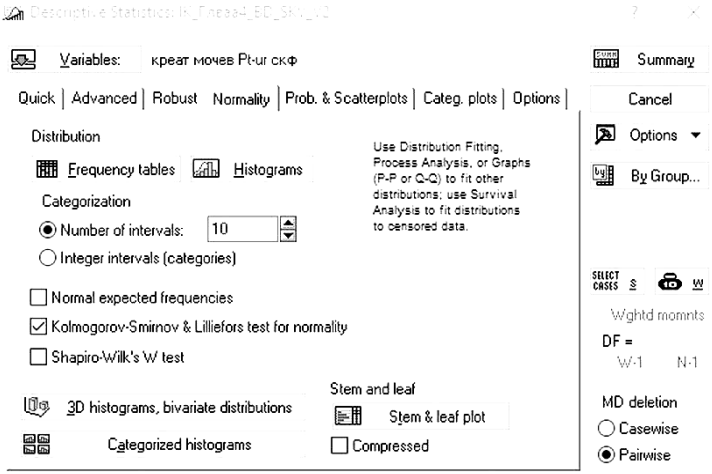


Рис. 3.4. Диалоговое окно «Проверка на нормальность»

При желании можно указать количество интервалов, потребовать, чтобы интервалы были целочисленными, и проставить галочки напротив интересующих критериев. Затем выбрать режим гистограммы, нажав на соответствующую кнопку, и получить примерный результат (рис. 3.5).

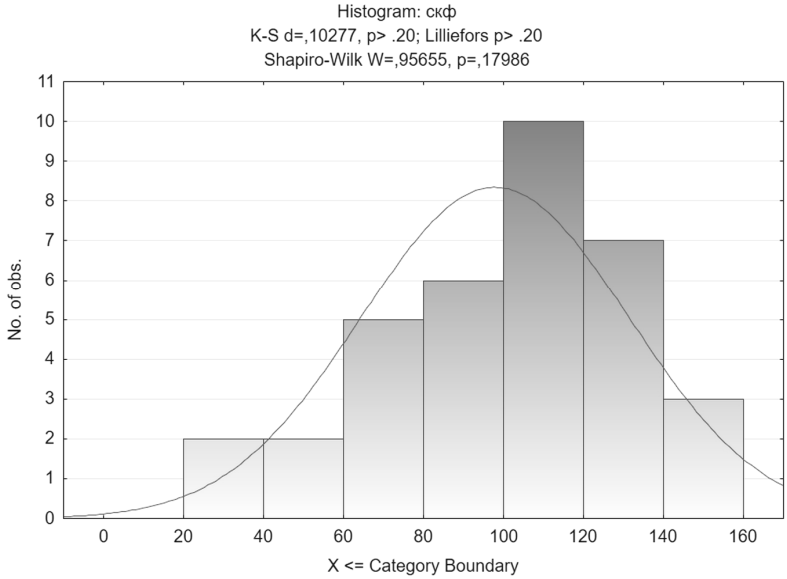


Рис. 3.5. Гистограмма распределения значений переменной

Над гистограммой видны ее название и значения критериев. Критерии Колмогорова–Смирнова и Лиллиефорса показывают уровень значимости $p > 0{,}2$, критерий Шапиро–Уилка – $p = 0{,}180$. В данном случае уровень значимости выше принятого критического значения 0,05, поэтому можно сделать вывод, что отличия распределения от нормального несущественны, и далее обращаться с переменной как с нормально распределенной.

Отчет о параметрах количественных данных. В зависимости от того, является ли величина нормально распределенной, в описании данных применяют разные наборы параметров. При нормальном распределении указывают среднее и среднеквадратическое отклонения, для данных, не имеющих нормального распределения, – медиану и квартили, иногда также Min–Max.

И, независимо от распределения величины, можно приводить доверительные интервалы среднего.

Описательная статистика качественных данных. Качественные данные отражают количество наблюдений, отнесенных к соответствующим категориям, которые могут выражаться в абсолютных и относительных количествах. Например, можно сказать, что на потоке из 100 студентов 25 – минчане (это абсолютная величина). Относительная величина – 25 % составляют минчане. Относительную величину принадлежности к категории также называют *долей*, которая аналогична вероятности. На нее распространяются понятийная система и методы вычисления для вероятностей.

Доля качественных данных аналогична среднему значению для количественных.

Таблицы частот. Распределения величины по категориям можно получить с помощью инструмента **Таблицы** частот по адресу: **Statistics>Basic statistics>Frequency Tables**. Как правило, настройки тут не требуются, достаточно выбрать переменные и получить результат в виде таблицы (рис. 3.6).

	Frequency table: креат>100 (IK_Глваа4_BD_S			
Category	Count	Cumulative Count	Percent	**Cumulative Percent**
0	26	26	72,22222	72,2222
1	9	35	25,00000	97,2222
Missing	1	36	2,77778	100,0000

Рис. 3.6. Результат распределения величины по категориям

Количество вариантов подсчитывается и в абсолютных величинах, и в относительных нарастающим итогом.

Ошибка доли. Очевидно, что доля является случайной величиной, ведь в разных выборках получают разные количества случаев с данной категорией. При этом рассеяние характеризуется величиной «ошибка доли». Однако в данном случае нет необходимости находить параметр рассеяния экспериментально, поскольку ошибка доли вычисляется аналитически:

$$\mathrm{SE}_{\text{доли}} = \sqrt{\frac{p(1-p)}{n}},$$

где p – величина доли.

Доверительный интервал доли. Доверительный интервал доли напрямую из данных в программе не вычисляется. Необходимо вручную или с помощью **Таблицы частот** найти относительное значение доли. Затем надо выбрать пункт меню **Statistics>Power Analysis>Interval Estimations>One Proportion** и получить калькулятор доверительного интервала (рис. 3.7).

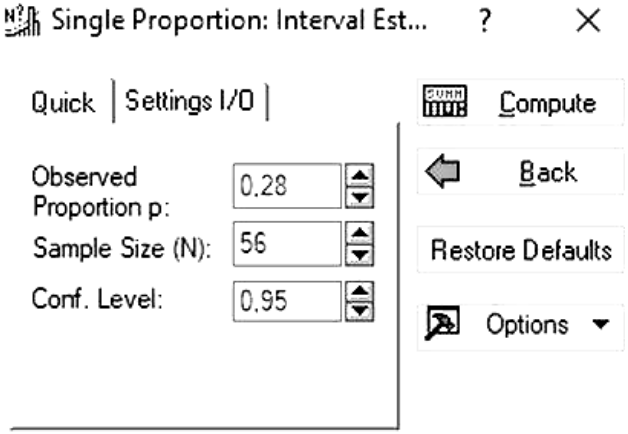


Рис. 3.7. Диалоговое окно «Доверительный интервал доли»

В калькуляторе надо указать долю и количество случаев в выборке, и после вычисления появится таблица (рис. 3.8).

	Interval Estimation (IK_Глваа4_BD_SKV_V2) One Proportion. Z (or Chi-Square) Test
	Value
Sample Proportion p	0.2857
Group Sample Size (N)	56.0000
Confidence Level	0.9500
Confidence Limits:	
Pi (Exact):	
Lower Limit	0.1730
Upper Limit	0.4221
Pi (Approximate):	
Lower Limit	0.1769
Upper Limit	0.4241
Pi (Crude):	
Lower Limit	0.1674
Upper Limit	0.4040

Рис. 3.8. Диалоговое окно «Расчет доверительного интервала»

В таблице (см. рис. 3.8), где приводятся результаты для всех методов расчета доверительного интервала, следует выбрать точный (**Exact**).

Обратите внимание: данный расчет является абсолютно правильным и, в отличие от других методов, корректно рассчитывает ДИ также для случаев, когда доля равна нулю или единице.

Отчет о параметрах качественных данных. В отчетах о качественных данных приводят таблицы частот в абсолютных и (или) относительных величинах, а также указывают ошибку доли.

При оценках параметров генеральной совокупности приводят доверительные интервалы долей.

3.2. Визуальный анализ данных

Программа STATISTICA предоставляет возможности визуального исследования данных, что целесообразно делать на первых этапах работы для выработки предварительных представлений.

Инструменты визуального исследования собраны в меню **Graphs**.

Построение гистограмм **Graphs>Histograms** аналогично функции построения гистограмм в описательной статистике, однако предоставляет больше возможностей для настроек. Существует возможность выбора различных распределений для огибающей кривой, также можно произвести проверку на нормальность с помощью тестов Шапиро–Уилка и Колмогорова–Смирнова (рис. 3.9).

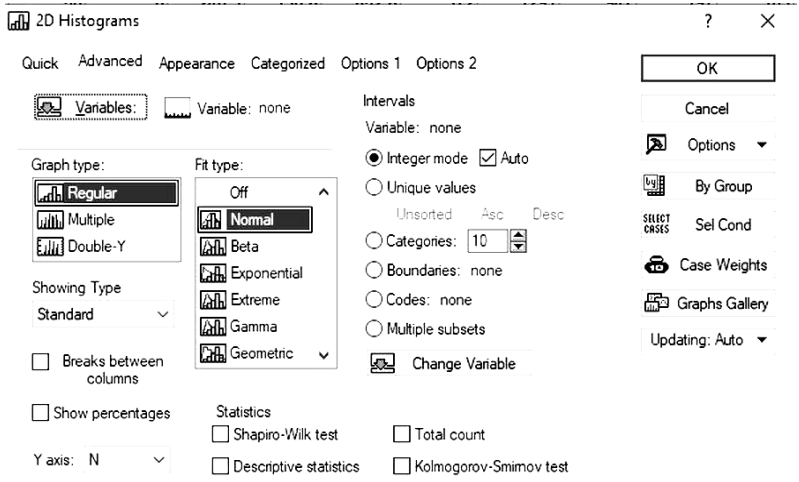


Рис. 3.9. Диалоговое окно с инструментами для визуального исследования

Этот модуль также позволяет выводить на одну гистограмму несколько величин, что дает возможности для создания качественных иллюстраций, а также для визуального сравнения распределений (рис. 3.10).

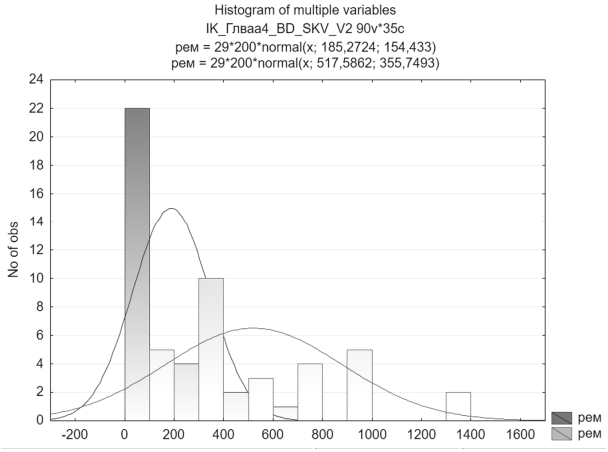


Рис. 3.10. Графическое сравнение распределений

Исследование связей между переменными целесообразно начинать с визуального модуля рассеяния **Graphs>Scatterplots**, в котором достаточно задать переменные для осей *X* и *Y* (рис. 3.11).

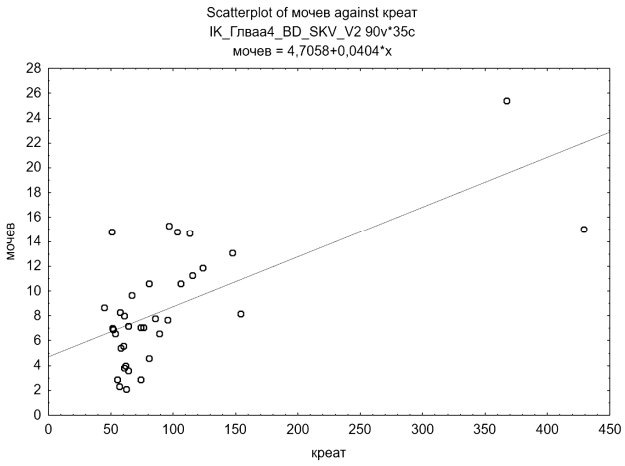


Рис. 3.11. Диаграмма рассеяния для переменных

Модуль **Graphs>Surface Plots** позволяет строить трехмерные графики зависимостей (рис. 3.12).

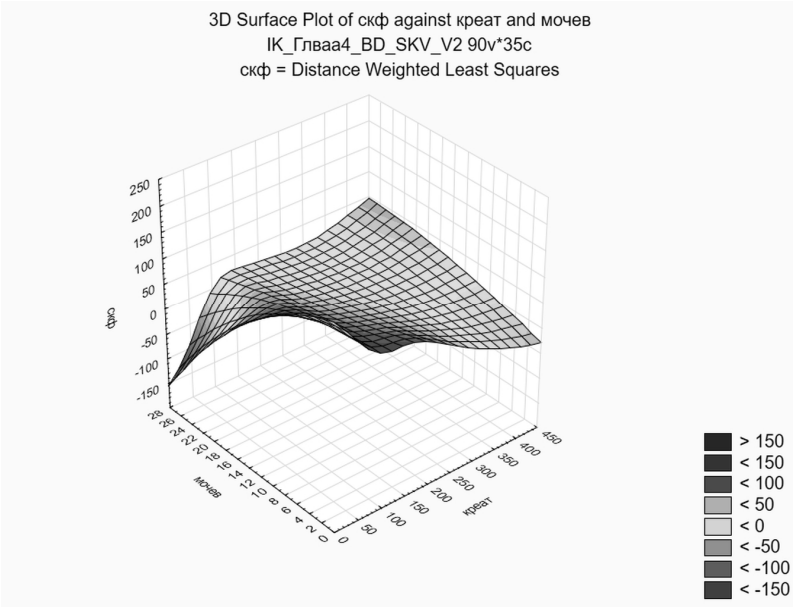


Рис. 3.12. График поверхности

В модулях **Graphs>2D Graphs** и **Graphs>3D XYZ Graphs** существует некоторое множество прочих инструментов и представлений.

4. СРАВНЕНИЕ ДАННЫХ

4.1. Понятие о статистической значимости и практической значимости

Сравнение данных – наиболее распространенная форма анализа данных, и его можно рассмотреть на примере следующего списка типовых задач.

1. Доказать, что между группами нет различий, с целью включения их в исследование как «однородные группы» (группы, происходящие из одной генеральной совокупности) и утверждать, что различные проявления связаны с некоторым «эффектом», а не с изначальными различиями.

2. Доказать, что есть различия в группах, относящихся к разным категориям. Например, что студенты из Минской области успевают лучше, чем студенты из Гродненской области.

3. Доказать, что один препарат уничтожает долгоносика лучше, чем другой.

4. Доказать, что после применения препарата популяция долгоносика уменьшилась.

Понятие о статистической значимости и практической значимости

Сравнение данных – исследование, направленное на выявление различий, доказательство их неслучайности, оценку статистической и практической значимости различий.

Определение включает несколько понятий, каждое из которых требует пояснения.

Доказательство неслучайности. Анализ данных оперирует со стохастическими величинами, т. е. они содержат различия. Если взять несколько выборок из одной генеральной совокупности, то они будут различны, у них будут разные средние значения, но их различие считается случайным. Следовательно, неслучайные различия – это различия более чем случайные, т. е. средние выборок различаются больше, чем обычно различаются средние однородных выборок.

Статистическая значимость различий означает, что вероятность ошибки в утверждении о наличии различий не больше, чем заданный уровень значимости. Уровень значимости обычно принимают равным 0,05.

Практическая значимость различий – профессиональная оценка важности достигнутых различий. Например, можно получить статистически значимое улучшение успеваемости на 0,1 балла, но нет возможности доказать, что это практически важно.

Возможность доказать наличие статистической значимости различий зависит от размера выборок. На малых выборках различия могут быть большими, но не иметь статистической значимости, что видно из формулы ошибки среднего:

$$SE = \frac{\sigma}{\sqrt{n}}.$$

Формула уменьшается с увеличением числа наблюдений пропорционально квадратному корню. Следовательно, надо разумно выбирать размер групп, чтобы еще до начала исследования оценить, будут ли доказаны различия. Для этих целей в пакете STATISTICA существует раздел анализа данных **Анализ мощности исследования**.

Таким образом, исследование различий связано с двумя проблемами: с одной стороны, надо доказать, что различия действительно существуют, а не являются следствием случайного отклонения, с другой – доказать, что достигнутые различия имеют практический смысл.

4.2. Нулевая и альтернативная гипотезы

Аксиоматика теории вероятностей изначально была связана со значительными сложностями и приобрела приемлемый для математики вид только в 80-е гг. XX в. Одновременно она окончательно перестала быть понятной кому-либо, кроме профессиональных математиков. Но и на более ранних этапах необходимость обеспечения математической корректности обусловила построение сложных логических конструкций, в частности в вопросе проверки значимости различий.

Термины и определения

Нулевая гипотеза – принимаемое по умолчанию предположение о том, что различия отсутствуют, т. е. не существует связи между двумя наблюдаемыми событиями.

Уровень значимости – вероятность отклонить нулевую гипотезу, если на самом деле она истинна (ошибка первого рода). Как правило, $p = 0{,}05$.

Альтернативная гипотеза – предположение о том, что различия не случайны.

Обратите внимание: требуется доказательство истинности альтернативной гипотезы.

Логика доказательства

1. Выдвигают нулевую гипотезу (различия случайны).

2. Вычисляют уровень p – вероятность случайного различия.

3. Сравнивают полученную вероятность с допустимым уровнем значимости, например с 0,05.

4. Делают выводы:

– если $p < 0{,}05$, гипотеза опровергнута,
вывод: «Различия статистически значимы»;

– если $p > 0{,}05$, гипотеза не опровергнута,
вывод: «Различия не обнаружены».

При этом понимают: может, их нет, может, не хватило размера выборки – мощности исследования.

Таким образом, различия считаются доказанными, если принимается альтернативная гипотеза, а она принимается, если не удалось принять нулевую гипотезу.

Также трудным для понимания оказывается факт, что можно доказать только наличие различий, а доказать отсутствие различий принципиально невозможно. Следовательно, в первом случае принято говорить, что *различия статистически значимы*, во втором (вместо фразы «различия отсутствуют») – *статистически значимые различия не обнаружены*.

Вычисление вероятности случайного различия

Идея вычисления вероятности случайности различий состоит в следующем: разность между средними выборок сравнивается со среднеквадратическим отклонением этой разности и вычисляется вероятность того, что эта разность случайна.

Если разность средних больше, чем 1,96σ, считается, что она слишком велика, и вероятность случайности отклонения составляет не более 0,05.

Следовательно, достаточно сравнить фактическую разность средних в группах со среднеквадратическим отклонением их разности:

$$t = \frac{\overline{x_1} - \overline{x_2}}{S_{\overline{x_1} - \overline{x_2}}},$$

$$S_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Здесь *t*-значение критерия Стьюдента, распределение которого при больших выборках ($n > 30$) совпадает с нормальным (*справочно*: значение *t*-критерия можно посмотреть в таблицах для *Z*. Для малых выборок, для уровня значимости 0,05 – в таблицах критерия Стьюдента, при условии, что число степеней свободы $\nu = n_1 + n_2 - 2$).

Точное сравнение случайных величин

Случайные величины характеризуются не только средними, но и параметрами рассеяния. Средние могут быть равны, дисперсии – существенно различаются.

Поэтому для доказательства однородности выборок следует также проверять равенство дисперсий. Проверка осуществляется по критерию Фишера, распределению которого соответствует отношение дисперсий:

$$t = \frac{\overline{x_1} - \overline{x_2}}{S_{\overline{x}_1-\overline{x}_2}},$$

$$F = \frac{S_1^2}{S_2^2}.$$

Логика проверки совпадает с методикой проверки различий средних. Выдвигают нулевую гипотезу, вычисляют значение *F*-критерия и по таблицам находят соответствующую ему вероятность различий. Если она оказывается меньше критического значения 5 %, нулевую гипотезу отвергают и говорят, что доказаны различия.

Сравнение количественных величин, не имеющих нормального распределения

Если количественные величины не имеют нормального распределения, их сравнение с использованием критерия Стьюдента некорректно, и для них применяют так называемые *непараметрические критерии*.

В частности, при использовании компьютерных программ обычно применяют критерии Манна–Уитни и Колмогорова–Смирнова.

5. СРАВНЕНИЕ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ЭВМ

5.1. Основные статистики

Виды сравнений

Инвентаризацию вариантов сравнения проводят, исходя из того, какие различные методы для этого понадобятся. Применяемые методы зависят от трех параметров:

1. Количество групп.
2. Различия групп.
3. Тип данных.

Количество групп. Различные методы (критерии) применяются для двух групп и для нескольких групп. Если есть несколько групп, сначала следует провести совместное сравнение, которое покажет, есть ли среди них различия, а затем провести попарное сравнение – для выяснения, между какими именно. Также существует задача сравнения среднего группы с числом.

Различия групп. В первом варианте происходит сравнение разных групп, т. е. «сравнение независимых групп». Во втором случае – сравнение различных состояний одной и той же группы, например, до и после воздействия, т. е. «сравнение зависимых групп».

Тип данных. Применяемые критерии зависят от того, какой тип данных в группах. Укрупненно применяемые методы делятся:

– на параметрические методы, которые применяются для нормально распределенных данных и для долей;

– непараметрические методы – для всех остальных данных.

Основные статистики (параметрические методы)

Независимые группы, нормально распределенные величины.

С целью проведения сравнения нормально распределенных величин в независимых группах в программе STATISTICA следует выбрать пункт меню **Statistics>Basic statistics>t-test independent by groups** (рис. 5.1).

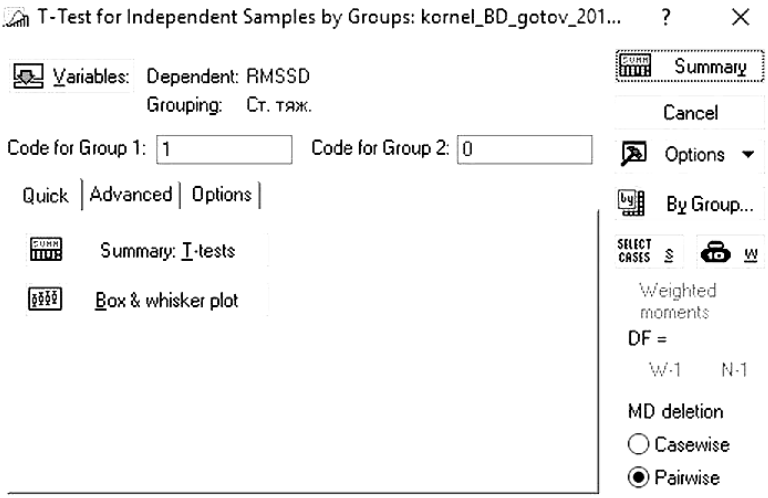


Рис. 5.1. Диалоговое окно «Сравнение нормально распределенных величин в независимых группах»

В диалоговом окне надо задать переменную, по которой идет сравнение, и переменную, которая определяет группы. Далее выбрать результаты и получить их в виде таблицы (рис. 5.2).

T-tests; Grouping: Ст. тяж. (kornel_BD_gotov_20171130)
Group 1: 1
Group 2: 0

Variable	**Mean 1**	Mean 0	t-value	df	p	Valid N 1	Valid N 0	Std.Dev. 1	Std.Dev. 0	F-ratio Variances	p Variances
RMSSD	24.21628	33.64848	-3.00915	74	0.003580	43	33	11.07646	16.22313	2.145197	0.020817

Рис. 5.2. Результаты сравнения нормально распределенных величин в независимых группах

Очевидно, что в результатах рассчитаны средние в группах и *t*-критерий по их разности. В последнем столбце выведена вероятность того, что различия случайны, и ее значение меньше, чем заданный уровень значимости 0,05.

В программе STATISTICA также можно наблюдать визуальное отображение рассеяния в виде так называемых «ящиков с усиками» (рис. 5.3). Маленький квадратик в центре обозначает среднее, собственно ящик – среднее плюс-минус среднеквадратическое отклонение, усики – диапазон, в который попадает 95 % наблюдений.

В частности, на этом графике хорошо видно, что параметры групп существенно различаются, и уже по визуальному отображению можно сказать, что они будут статистически значимы.

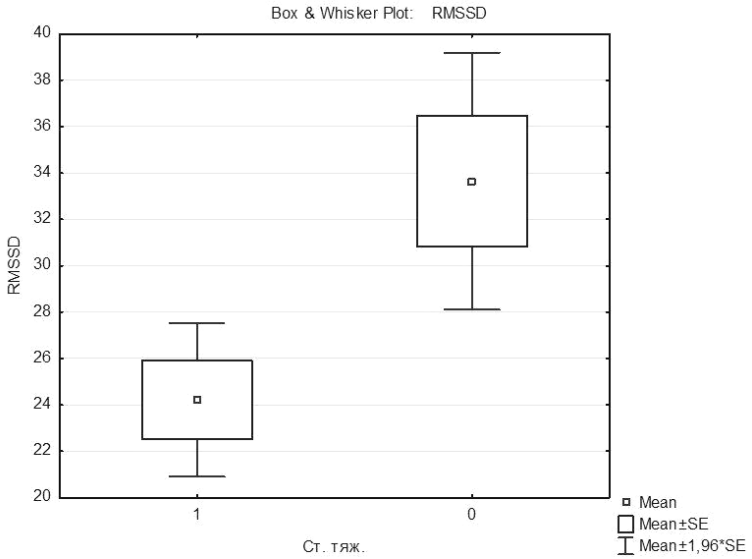


Рис. 5.3. «Ящики с усиками» для визуального отображения рассеяния

Зависимые группы, нормально распределенные величины.

В случае, когда исследуются зависимые группы, следует выбрать пункт **Statistics>Basic statistics>t-test dependent samples**. Здесь надо указать две переменные, которые и отображают разные состояния. Обычно это сравнение до и после воздействия для выяснения действенности.

Выходные формы в этом модуле аналогичны предыдущим (см. рис. 5.2).

5.2. Непараметрические статистики

Непараметрические статистики расположены по адресу **Statistics>Nonparametrics**. Здесь находятся пять методов.

Сравнение долей. Таблица 2×2. Следует помнить, что доля представляет собой отношение количества наблюдений некоторой категории к количеству всех наблюдений и по смыслу совершенно соответствует вероятности.

Сравнение долей не работает автоматически, поэтому сначала необходимо посчитать варианты вручную или взять частоты из таблицы сопряженности (**Statistics>Basic statistics>multiple response tables**) (рис. 5.4).

N=31 Caspase <250	креат>50 1.	креат>50 0.	Row Totals
0.	9	5	14
1.	11	6	17
All Grps	20	11	31

Рис. 5.4. Значения частот в таблице сопряженности

Из данных рис. 5.4 видно, что в группе из 14 наблюдений с Caspase=0 встретилось 9 случаев с креат=1 и 5 случаев с креат=5. Аналогично, в группе с Caspase=1 встретилось 11 случаев с креат=1 и 6 случаев с креат=6.

Далее выбирают пункт **2×2 Tables** в группе непараметрических методов и заносят полученные частоты в соответствующий пункт меню (рис. 5.5).

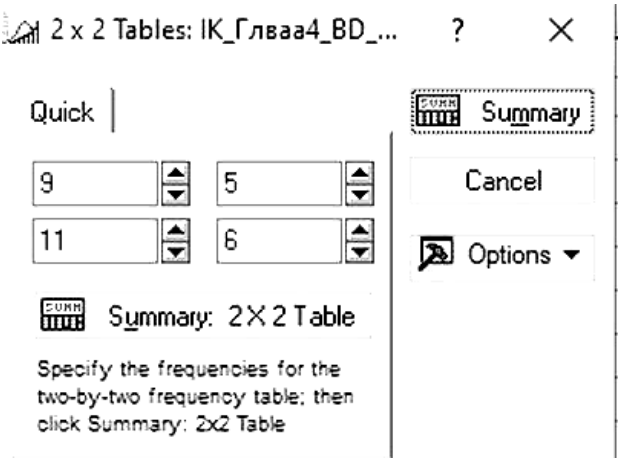


Рис. 5.5. Диалоговое окно «2×2 Tables»

В результате получен расчет для четырех критериев и соответствующие уровни вероятности (рис. 5.6). Это – критерии χ-квадрат, χ-квадрат с поправкой Йетса, точный критерий Фишера и критерий МакНамара.

	Column 1	Column 2	Row Totals
Frequencies, row 1	9	5	14
Percent of total	29,032%	16,129%	45,161%
Frequencies, row 2	11	6	17
Percent of total	35,484%	19,355%	54,839%
Column totals	20	11	31
Percent of total	64,516%	35,484%	
Chi-square (df=1)	,00	p= ,9806	
V-square (df=1)	,00	p= ,9809	
Yates corrected Chi-square	,12	p= ,7242	
Phi-square	,00002		
Fisher exact p, one-tailed		p= ,6361	
two-tailed		p=1,0000	
McNemar Chi-square (A/D)	,27	p= ,6056	
Chi-square (B/C)	1,56	p= ,2113	

Рис. 5.6. Результаты расчета для четырех критериев

При больших группах и частотах рассчитанные величины вероятности примерно одинаковы, при малых группах вступают в силу ограничения по применимости критериев. Здесь же достаточно знать, что точный критерий Фишера работает всегда, в том числе и для малых значений частот.

Обратите внимание: данные – для одностороннего и двустороннего случаев. Односторонний тест применим, если точно известно, какая величина больше, но поскольку это трудно обосновать, то надо брать результаты для двустороннего теста.

5.3. Сравнение двух независимых групп

Для сравнения двух независимых групп следует выбрать пункт **Statistics>Nonparametrics>Comparing two independent samples**.

Этот метод подходит для сравнения количественных данных, не имеющих нормального распределения, а также с некоторыми ограничениями для порядковых данных (рис. 5.7).

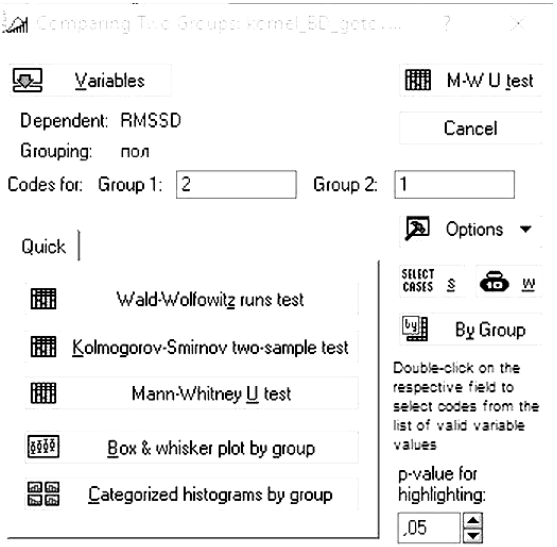


Рис. 5.7. Диалоговое окно «Сравнение двух независимых групп»

В этом диалоговом окне задаются переменные – исследуемая и группирующая, и далее выбирается форма результата. Предлагаются три критерия, из которых наиболее популярным является критерий Манна–Уитни (рис. 5.8).

Mann-Whitney U Test (kornel_BD_gotov_20171130)
By variable пол
Marked tests are significant at p <.05000

variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-value	Z adjusted	p-value	Valid N Group 1	Valid N Group 2	2*1sided exact p
RMSSD	2891.500	9043.500	2071.500	-0.857049	0.391418	-0.857069	0.391408	40	114	0.391755

Рис. 5.8. Результаты расчета по критерию Манна–Уитни

В выходной форме приводятся значения критерия **U** и вероятности.

Программа позволяет получить визуальное представление рассеяния в виде «ящиков с усиками», а также гистограммы для каждой группы.

Обратите внимание: в данном случае будут выводиться ящики с медианой, квартилями и размахом.

Другие непараметрические сравнения. Также в группе непараметрических методов предлагается сравнение нескольких независимых групп, двух зависимых и нескольких независимых. Диалоговые окна и предлагаемые возможности аналогичны случаю с двумя независимыми группами.

6. ПОГРЕШНОСТИ, ОКРУГЛЕНИЕ

6.1. Вычисление и корректная запись приближенных чисел

Вариабельность наблюдений включает в себя собственную вариабельность измеряемой величины и вариабельность процесса измерения, которые суммируются в результате наблюдения. Соответственно, задачи исследования разделяются на две группы.

1. Превалирует вариабельность системы. К этой группе относятся медико-биологические, социальные процессы, а также сложные технические системы.

2. Превалирует вариабельность измерения. Обычно в инженерных конструкциях собственная вариабельность невелика, и результат зависит в основном от процесса измерения. Во второй группе задач вариабельность принято называть погрешностью.

Погрешность измерения – отклонение измеренного значения величины от ее истинного (действительного) значения. Погрешность измерения является характеристикой точности измерения.

Виды погрешностей

С точки зрения понятий и величин существуют три варианта описания погрешностей.

Абсолютная погрешность ΔA – разность между измеренным и истинным значениями.

Относительная погрешность $\sigma = \frac{\Delta A}{A}$ – отношение абсолютной погрешности к истинному значению.

Приведенная погрешность $\gamma = \frac{\Delta A}{A_n}$ – отношение абсолютной погрешности к так называемому номинирующему значению – диапазону от минимального до максимального значения шкалы.

Определение погрешности прибора

Погрешность прибора является нормированной величиной, которая определяется классом точности прибора.

Класс точности – обобщенная характеристика средств измерений, определяемая пределами допускаемых основных и дополнительных погрешностей, а также рядом других свойств, влияющих на точность осуществляемых с их помощью измерений. Класс точности указывается на шкале прибора (рисунок).

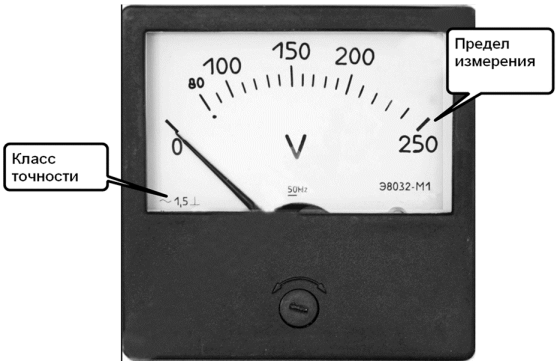


Рис. Шкала вольтметра с указанием мест размещения основных характеристик

Если прибор не может обеспечить свою паспортную точность, он запрещается к использованию. Процедура регулярной проверки точности называется *поверкой прибора* и осуществляется специализированными службами.

После установки пригодности прибора к использованию, определения его класса точности, можно вычислить погрешность измерений:

$$\Delta A = A_n \cdot \text{класс точности}.$$

Обратите внимание: эта погрешность для прибора постоянна и не зависит от измеряемой величины. Из чего следует, что для уменьшения относительной погрешности надо выбирать приборы так, чтобы измерение проводилось преимущественно во второй половине шкалы.

Погрешность снятия измерения

Кроме собственной погрешности прибора, существует также погрешность снятия измерения, которая связана с тем, что шкала дискретна. Поэтому нельзя измерить нечто точнее, чем деление шкалы. Принято считать, что *погрешность снятия измерения равна половине деления шкалы*.

Суммарная погрешность измерения

Имеются две погрешности, которые для установления суммарной погрешности необходимо сложить.

Поскольку теория погрешностей и математическая статистика развивались независимо, они не имеют общей понятийной системы. Если в математической статистике существует оценка, какая доля

наблюдений попадет в интервал $\bar{x} \pm k\sigma$ (например, при $k = 1$ – это 67 %, при $k = 2$ – 90 %) , то в теории погрешностей о вероятности попадания в интервал погрешности умалчивают. Однако при этом погрешность как ошибка измерения все равно остается стохастической величиной со своими параметрами рассеяния, причем следует ожидать, что распределение этой величины нормальное.

Значит, погрешности следует складывать не как числа, а как дисперсии или ортогональные векторы, т. е. геометрически, по теореме Пифагора. Тогда суммарная ошибка равна:

$$A_{\sum} = \left| A^2_{\text{прибора}} + A^2_{\text{снятия}} \right|.$$

На практике чаще используется прямое сложение погрешностей:

$$A_{\sum} = A_{\text{прибора}} + A_{\text{снятия}}.$$

Следовательно, можно использовать 2 варианта.

6.2. Погрешность числа

Погрешностью обладают не только измерения, но и числа, что понятно на примере числа π.

Погрешность констант

Если для вычисления взята величина 3,14, то следует понимать, что она отличается от истинного значения на 0,001 59…

Эту разницу и следует учитывать: $\pi = 3{,}14$, ошибка $\Delta\pi = 0{,}0016$. Отсюда относительная погрешность числа:

$$\sigma = \frac{\Delta\pi}{\pi} = 0{,}05\ \%.$$

Погрешность данных из справочника

Типичное представление данных в справочнике имеет следующий вид: «Плотность чистого золота равна 19,32 г/см³».

Принято считать, что погрешность округления приближенного числа равна половине единицы того разряда, до которого это число было округлено. Если воспользоваться данной рекомендацией, то можно увидеть, что плотность была округлена до второго десятичного разряда, следовательно, погрешность равна 5 единицам третьего десятичного разряда, и величину плотности следует записать как

$$\rho_A = (19{,}32 \pm 0{,}005) \text{ г/см}^3.$$

Следовательно, относительная погрешность составит 0,03 %.

6.3. Погрешность косвенного измерения

Часто искомая величина не измеряется непосредственно, например, не существует приборов для прямого измерения объема. Поэтому для получения объема параллелепипеда измеряют три величины – длину, ширину и высоту – и перемножают их.

Косвенные измерения – это измерения, при которых значение величины можно найти, используя зависимость между этой величиной и величинами, подвергаемыми прямым измерениям.

Очевидно, что погрешность косвенных измерений требует специального вычисления. Для определения правил вычисления косвенной погрешности применяется дифференциальное исчисление.

Вычисление погрешности косвенного измерения

Пусть исследуемую величину s определяют по результатам прямых измерений других независимых физических величин: x, y, z.

$$s = f\left(\bar{x}, \bar{y}, \bar{z}\right).$$

Каждая величина имеет погрешности: Δx, Δy, Δz.

Предполагается, что величины являются случайными с нормальным распределением, тогда к ним применимы следующие свойства.

1. Среднее значение

$$\bar{s} = f\left(\bar{x}, \bar{y}, \bar{z}\right).$$

3. Погрешности

$$\Delta s = \left| f_x'^2 \Delta x^2 + f_y'^2 \Delta y^2 + f_z'^2 \Delta z^2 \right|,$$

где f_x', f_y', f_z' – частные производные по x, y и z в точке $\left(\bar{x}, \bar{y}, \bar{z}\right)$.

4. Косвенную погрешность можно вычислить двумя способами.

Прямое вычисление косвенной погрешности. При прямом вычислении следует вычислить частные производные в данной точке и затем вычислить по формуле ошибку функции.

Пример 1

Требуется найти объем цилиндра. В результате измерений получены следующие значения:

d = (4,01 ± 0,03) мм,

h = (8,65 ± 0,02) мм,

π = 3,14 ± 0,002.

Обратите внимание: погрешность прямых измерений зависит от применяемых инструментов.

Объем цилиндра вычисляется по формуле

$$V = \frac{\pi d^2 h}{4},$$

где d – диаметр цилиндра;

h – высота цилиндра.

$$V = 3{,}14 \cdot (4.01)^2 \cdot 8{,}65 / 4 = 109{,}19 \text{ мм}^3.$$

Для вычисления погрешности сначала определим частные производные:

$$v'_\pi = \frac{d^2 h}{4} = 34{,}7,$$

$$v'_d = \frac{2\pi dh}{4} = 54{,}5,$$

$$v'_h = \frac{\pi d}{4} = 12{,}6.$$

Далее по формуле вычислим:

$$\sqrt{(34{,}7 \cdot 0{,}002)^2 + (54{,}5 \cdot 0{,}03)^2 + (12{,}6 \cdot 0{,}02)^2} = 1{,}66 \text{ мм}^3.$$

Следовательно, ΔV = 1,66 мм^3, δV = 1,66 / 109,19 = 1,5 %.

Вычисление косвенной погрешности по правилам. В большинстве случаев исследуемая функция записывается через типовые операции, и для вычисления погрешности можно воспользоваться следующими формулами:

Рабочая зависимость	Формула погрешности
1. $s = A \cdot x + B \cdot y + C \cdot z$	1. $\Delta s = \sqrt{(A \cdot \Delta x)^2 + (B \cdot \Delta y)^2 + (C \cdot \Delta z)^2}$
2. $s = Ax^{\pm\alpha} y^{\pm\beta} z^{\pm\gamma}$	2. $\delta s = \sqrt{(\alpha \cdot \delta x)^2 + (\beta \cdot \delta y)^2 + (\gamma \cdot \delta z)^2}$

Рабочая зависимость	Формула погрешности
3. $s = \ln x$	3. $\Delta s = \dfrac{\Delta x}{x}$
4. $s = e^x$	4. $\delta s = \Delta x$
5. $s = A \cdot \sin\varphi$	5. $\delta s = A \cdot \cos\varphi \Delta\varphi$

Здесь Δx – абсолютная погрешность, δs – относительная погрешность.

Пример 2

В задаче с объемом цилиндра формула представляет собой произведение степеней, поэтому для нее применимы формулы п. 2.

$$\delta V = \sqrt{\left(1 \cdot \frac{0,002}{3,14}\right)^2 + \left(2 \cdot \frac{0,03}{4,01}\right)^2 + \left(1 \cdot \frac{0,02}{8,65}\right)^2} = 0,15 = 1,5\ \%.$$

Соответственно, абсолютная погрешность $\Delta V = V \cdot \delta V = 1,66$ мм^3.

6.4. Порядок округления величин

Округление величин, имеющих погрешность, подчиняется определенным правилам, которые можно выразить алгоритмом (таблица).

Таблица

Алгоритм вычисления погрешности

Операция	Пример
Дано число	$x = 0,008\,452 \pm 0,000\,035$
1. Вынести за общую скобку множитель вида 10^k так, чтобы среднее было записано числом от 1 до 10	$x = (8,452 \pm 0,035) \cdot 10^{-3}$
2. Округлить погрешность (в скобках): 1) до одной значащей цифры, если эта цифра больше 2; 2) до двух значащих цифр, если она меньше двух; 3) встречаются и другие правила	$x = (8,452 \pm 0,04) \cdot 10^{-3}$

Окончание таблицы

Операция	Пример
3. Округлить число (в скобках), соответствующее среднему значению: до последнего разряда, соответствующего последнему разряду погрешности	$x = (8{,}45 \pm 0{,}04) \cdot 10^{-3}$
4. При желании, перевести в десятичную систему или посчитать относительную погрешность	$x = 0{,}008\ 45 \pm 0{,}000\ 04$
5. При необходимости вычислить относительную погрешность	$\sigma_x = \dfrac{0{,}000\,04}{0{,}008\,45} = 0{,}0047 = 0{,}5\,\%$

Приведенная таблица иллюстрирует основные алгоритмы округления величин, имеющих погрешность.

7. СТАТИСТИЧЕСКАЯ СВЯЗЬ

7.1. Факторы и отклик

Исследование связей между величинами – одна из основных задач анализа данных.

Связь – взаимообусловленность существования явлений, разделенных в пространстве и (или) во времени [5].

Статистическая связь – связь между переменными, на которую накладывается воздействие случайных факторов. В результате действия такой связи изменения одной переменной приводят к изменениям другой не детерминировано [6].

Статистическая связь – соотношение между двумя переменными, при котором изменение значения одной переменной влечет изменение распределения другой переменной [7].

Связь корреляционная – статистическая связь между двумя (или более) количественно выраженными случайными величинами, мера жесткости (степени приближения к строгой функциональной зависимости), которой измеряется коэффициент корреляции или корреляционным отношением [8].

Связь статистическая – вероятностная зависимость между двумя (или многими) случайными величинами, не имеющая в общем случае строго функционального характера и возникающая тогда, когда одна из величин зависит не только от другой (или других), но и от нескольких меняющихся.

Главным в этих определениях является следующее:

1. Анализ данных не позволяет говорить о причинно-следственных связях, только о статистических.

2. Статистические связи не точные, а приблизительные, со случайной вариацией.

3. «После» не значит «вследствие».

4. «Одинаково» не значит «вследствие».

Можно принять следующее определение: *статистическая связь* – обнаруженный факт того, что различным характеристикам одной стохастической величины неслучайно соответствуют различные стохастические характеристики другой.

Факторы и отклик

Методы анализа данных не говорят напрямую о причинно-следственных связях, но в конкретном исследовании предполагается, что какие-то переменные будут считаться влияющими (x_1, x_2, x_n), и ка-

кие-то – подверженными влиянию (y_1, y_2, y_n) (рис. 7.1). (Так же как и в функциональном анализе: какую-то переменную считают аргументом, какую-то – функцией.)

Переменные, которые считают влияющими, называются *факторами*.

Переменные, которые подвержены влиянию, называются *откликом*. Обычно в качестве отклика рассматривается одна переменная.

x_1	Объект исследования	y_1
x_2		y_2
x_n		y_n

Рис. 7.1. Факторы и отклик

Варианты терминов:

1. Фактор=аргумент=независимая переменная= =предиктор=independent variable.

2. Отклик=функция=зависимая переменная=dependent variable.

Разумеется, знак равенства здесь не надо понимать, как абсолютную идентичность, обычно разные термины применяют в разных контекстах.

7.2. Типы связей и способы их описания

Способы описания статистических связей представлены в табл. 7.1.

Таблица 7.1

Статистическая связь и способы ее описания	Словесный	При увеличении переменной…
	Визуальный	Графически
	Численный	Коэффициент корреляции…
	Классификационный	Такие-то объекты принадлежат классу…
	Диагностический	Если…, то…
	Аналитический	Если значение переменно равно…, то риск…
	Формульный	$y = f(x)$

Каждому варианту, начиная с численного способа, соответствуют определенные методы анализа данных, например, зависящие от типа величин, от вида отклика (количественный, качественный, номинальный, порядковый или дихотомический) (табл. 7.2).

Таблица 7.2

Зависимость выбора метода анализа данных от вида отклика

Вид отклика	Значение отклика	Метод
Дихотомический	0 или 1	Логистическая регрессия
Порядковый	1, 2, 3 ..	Дискриминантный анализ
Номинальный	1, 2, 3 ….	Классификация
Количественный	Число	Регрессия

Словесный способ

Этот метод не считается предпочтительным, но при сложностях с получением зависимостей в ином виде необходимо описывать связь между переменными текстом, например: с увеличением переменной *X* до середины шкалы происходит монотонное увеличение переменной *Y*, далее значения *Y* выходят на плато.

Визуальный способ

Визуализация данных в программе STATISTICA описана в разделе 3. Различные формы визуального отображения данных позволяют иллюстрировать статистическую связь между переменными.

Дискретный фактор, количественный отклик. Наиболее удобным визуальным представлением в данном случае является формат «ящик с усиками» (рис. 7.2).

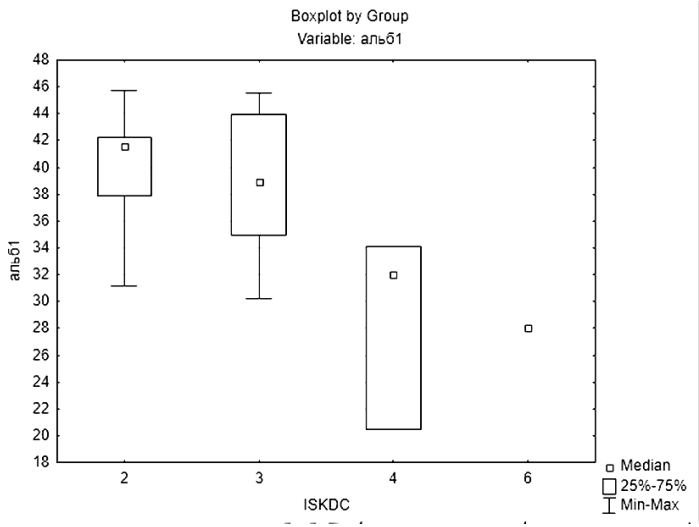


Рис. 7.2. Графическое изменение рассеяния отклика с увеличением фактора

Из данных графика видно, как с увеличением фактора меняется рассеяние отклика.

Две количественные переменные. При проведении анализа связи двух количественных переменных может помочь отображение взаимного рассеяния (рис. 7.3).

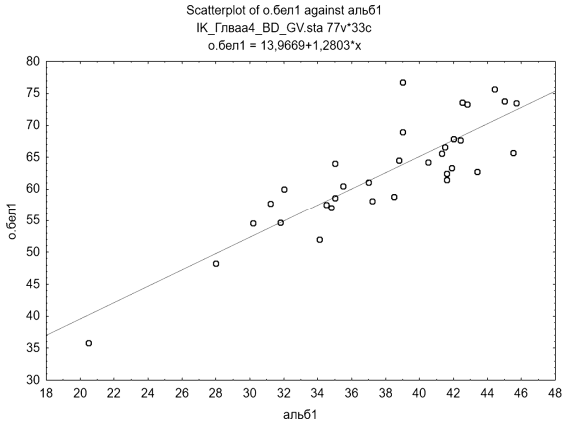


Рис. 7.3. Графический характер статистической связи переменных

Из данных графика виден характер статистической связи переменных.

Три количественные переменные. Для отображения взаимосвязи трех количественных переменных в модуле **Graphs** существуют несколько форматов. Прежде всего, это трехмерная поверхность функции (рис. 7.4).

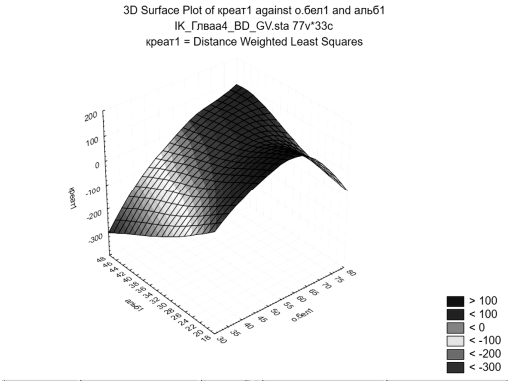


Рис. 7.4. График поверхности функции

Другим вариантом отображения трехмерной зависимости является контурный график (рис. 7.5).

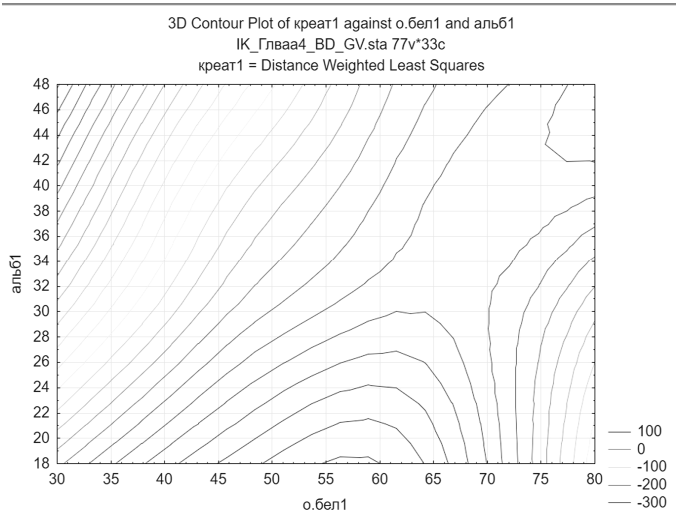


Рис. 7.5. Контурный график

Как вариант – топографический график (рис. 7.6).

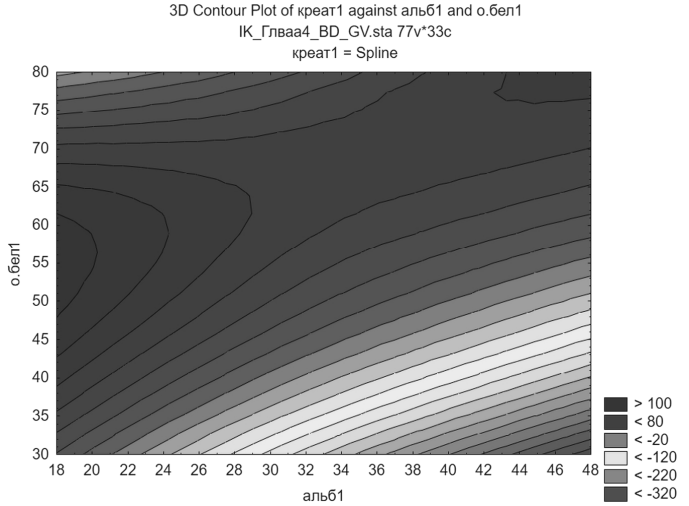


Рис. 7.6. Топографический график

7.3. Корреляция

Корреляция – мера связи между двумя переменными. Коэффициент корреляции может изменяться от –1,00 до +1,00. Значение коэффициента корреляции –1,00 соответствует абсолютной обратно пропорциональной связи. Значение +1,00 соответствует абсолютной пропорциональной связи. Значение 0,00 соответствует полному отсутствию связи.

Сила связи и значимость связи

Корреляционная связь характеризуется двумя параметрами:

– коэффициент корреляции отражает силу связи;

– значимость коэффициента корреляции отражает его неслучайное отличие от нуля.

То есть при расчете коэффициента корреляции одновременно сравнивают его с нулем и находят вероятность того, что отличие от нуля не случайно. Версию о статистической значимости коэффициента корреляции признают верной, если эта вероятность меньше критического уровня (обычно 0,05).

Таким образом, к рассмотрению принимаются только те коэффициенты корреляции, которые статистически значимы, и затем рассматривают силу связи.

Выбор варианта корреляции

Методы оценки корреляционных зависимостей используют в зависимости от типа переменных (табл. 7.3).

Таблица 7.3

Методы оценки корреляционных зависимостей в зависимости от типа переменных

Тип шкалы		Мера связи
Переменная *X*	Переменная *Y*	
интервальная, или отношений	интервальная, или отношений	Коэффициент Пирсона
ранговая, интервальная, или отношений	ранговая, интервальная, или отношений	Коэффициент Спирмена
ранговая	ранговая	Коэффициент Кендалла
дихотомическая	дихотомическая	Коэффициент φ, двухполевая корреляция

Окончание таблицы 7.3

Тип шкалы		Мера связи
Переменная X	Переменная Y	
дихотомическая	ранговая	Рангово-бисериальный коэффициент
дихотомическая	интервальная, или отношений	Бисериальный коэффициент
интервальная	ранговая	Не разработан

В абсолютном большинстве случаев используют так называемую корреляцию Спирмена, подходящую для количественных и порядковых переменных.

Ограниченные возможности корреляции

Исследование корреляционных связей позволяет быстро оценить стохастический процесс, однако корреляционные методы хорошо выявляют только линейные зависимости. Различные варианты рассеяния и соответствующие им коэффициенты приведены на рис. 7.7.

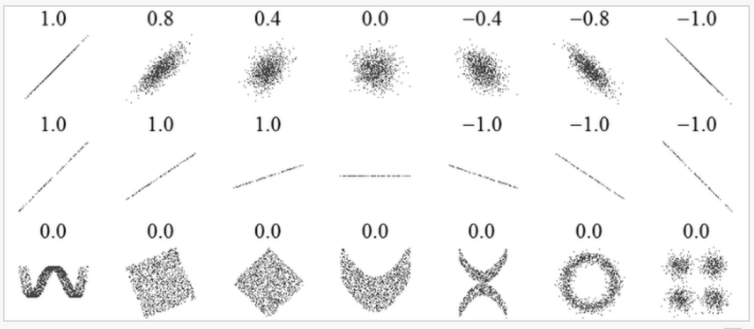


Рис. 7.7. Диаграммы вариантов рассеяния и значения соответствующих им коэффициентов

Следует знать, что сложные немонотонные зависимости корреляционный метод различает плохо.

Коэффициент корреляции и сила связи

Существуют общепринятые способы интерпретации силы корреляционной связи в зависимости от значения коэффициента (табл. 7.4).

Таблица 7.4

Сила корреляционной связи в зависимости от значения коэффициента

Значение коэффициента	Сила корреляционной связи
До 0,2	Очень слабая
До 0,5	Слабая
До 0,7	Средняя
До 0,9	Высокая
Свыше 0,9	Очень высокая

Расчет коэффициентов корреляции

Расчет в программе STATISTICA размещается по адресу: **Statistics>Nonparametric>Correlations** (рис. 7.8).

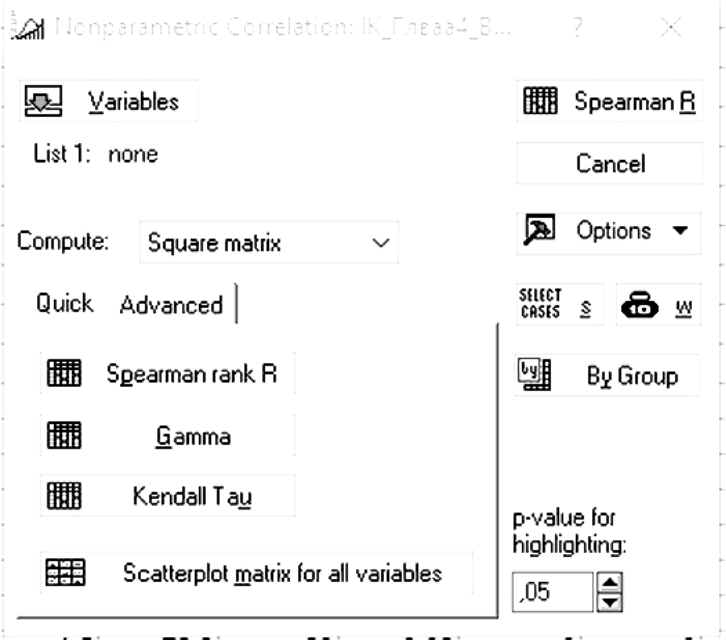


Рис. 7.8. Диалоговое окно «Расчет коэффициентов корреляции»

Кнопкой **Variables** надо выбрать переменные.

Программа позволяет кнопкой **Compute** выбрать три варианта отображения.

– квадратная матрица покажет результат «все против всех»;

– два списка дадут попарные корреляции между переменными из двух списков;

– детальный отчет также позволяет получить точные значения вероятностей.

Выходная форма результатов выглядит следующим образом (рис. 7.9).

Variable	креат1	мочев1	о.бел1	альб1
креат1	1,000000	0,163567	0,608340	0,337011
мочев1	0,163567	1,000000	-0,111773	-0,171807
о.бел1	0,608340	-0,111773	1,000000	0,840451
альб1	0,337011	-0,171807	0,840451	1,000000

Рис. 7.9. Результаты расчета коэффициентов корреляции

Выделены коэффициенты, которые статистически значимы, т. е. доказано, что они отличны от нуля.

8. АНАЛИЗ СВЯЗЕЙ ПРИ ДИСКРЕТНОМ ОТКЛИКЕ

8.1. Дискриминантный анализ

В целом ряде задач отклик является дискретным. В некоторых задачах он может быть дихотомическим – случилось событие или не случилось, подтвердился диагноз или нет и т. п. В других задачах отклик может иметь несколько значений, обычно – это отнесение объектов к некоторым классам. Соответственно, для данных случаев применяются различные методы анализа связей.

Дискриминантный анализ

Дискриминантный анализ – раздел вычислительной математики, представляющий собой набор методов статистического анализа для решения задач распознавания образов, который используется для принятия решения о том, какие переменные разделяют возникающие наборы данных.

То есть этот метод применяют, когда отклик уже известен и следует выяснить, какие переменные могут описать возникшее разделение. Наиболее дискриминантный анализ подходит для случаев порядкового отклика.

Функции классификации. Описывают правдоподобие того, что объект с заданными свойствами относится к данной группе при данных значениях признаков согласно построенной классификации:

$$C_j = c_{j0} + c_{j1}X_1 + \ldots + c_{jp}X_p,$$

где C – функция классификации;

$j = 1, \ldots;$

c – коэффициенты функций классификации;

X – переменные-признаки;

p – количество признаков.

Для каждого объекта (в том числе нового) можно вычислить значение всех функций классификации. В зависимости от того, какое значение больше, выясняется, к такой группе относится объект.

Пример

Факторы: средний балл студента, количество задолженностей, живет ли в общежитии, были ли выговоры за нарушение дисциплины, рост, вес.

Отклик: оператор телефона – А1, МТС, Life.

Модуль дискриминантного анализа находится по адресу: **Statistics>Multivariate Exploratory Techniques> Discriminant Function Analysis**.

8.2. Кластеризация

Кластеризация – методы разбиения множества объектов на группы (кластеры). Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны как можно более отличаться. Главная особенность кластеризации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма. Похожими считаются объекты, между которыми малые расстояния в пространстве переменных. Причем в разных вариантах методов в качестве расстояния может браться Евклидово расстояние, его квадрат, расстояние «городских кварталов», расстояние Чебышева и т. д.

Из всех методов кластеризации рекомендуется к использованию *метод К-средних*, который расположен по адресу: **Statistics>Multivariate Exploratory Techniques> Cluster Analysis>K–means clustering.**

Задачи кластеризации данных могут быть самыми различными. Например, с помощью кластеризации решается задача автоматической классификации научной литературы.

Также кластеризация поможет разбить общее множество на классы, которые далее исследуются на некоторое воздействие. Предполагается, что объекты внутри класса будут давать подобный отклик на воздействие, а реакция классов между собой будет различаться. Например, это может быть исследование воздействия препарата или лекарства.

Диалоговое окно кластеризации выглядит следующим образом (рис. 8.1):

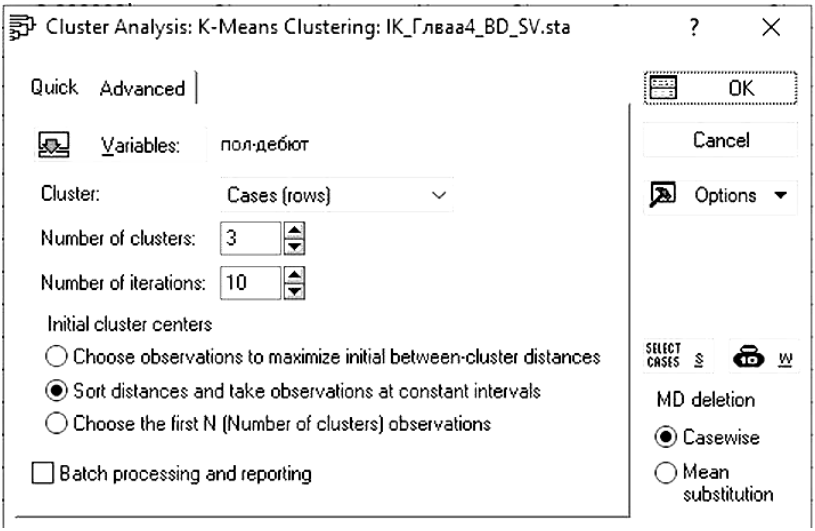


Рис. 8.1. Диалоговое окно «Кластеризация»

Изначально надо выбрать переменные, в пространстве которых будет происходить кластеризация. Следует иметь в виду, что количество переменных должно быть в 8–10 раз меньше количества наблюдений.

Затем необходимо выбрать количество кластеров и количество итераций.

8.3. Логистическая регрессия

Логистическая регрессия, или логит-модель (англ. *logit model*), – это статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события, причем функция вероятности описывается логистической кривой.

Таким образом, логистическая регрессия используется, когда отклик дихотомический, и может быть полезна в задачах предсказания вероятности, диагностики, исследования зависимости вероятности от факторов, оценки отношения шансов и отношения рисков. Очевидно, данный метод имеет широкие возможности и, соответственно, – широкое применение.

Логистическая кривая описывается функцией

$$p(z) = \frac{1}{1+e^{-z}},$$

где $z = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$, т. е. z представляет собой линейную комбинацию переменных или факторов.

Особенности логистической функции состоят в том, что при изменении z от $-\infty$ до $+\infty$ функция меняется от 0 до 1, т. е. в пределах, необходимых для описания вероятности.

При этом основное изменение происходит в диапазоне ± 6z, на участке ± 2z функция практически линейна (рис. 8.2).

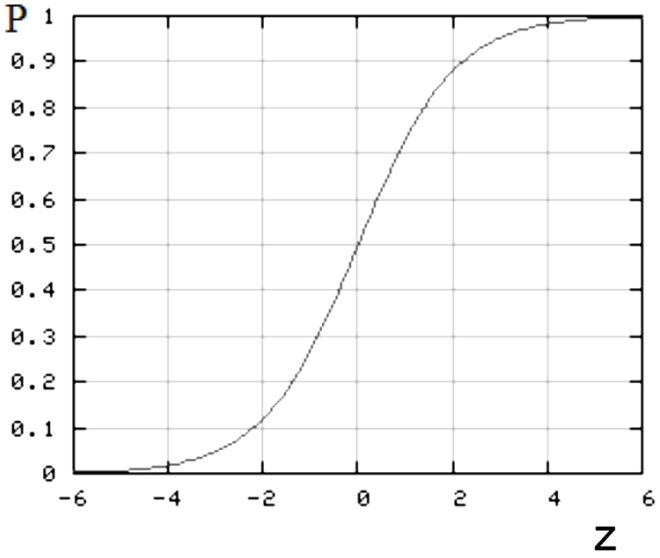


Рис. 8.2. Вид логистической кривой

Перечисленные особенности обусловили удобство и широкое применение метода.

Метод логистической регрессии наиболее удобно реализован программе SPSS и находится там по адресу: **Анализ>Регрессия>Логистическая**. Диалоговое окно выглядит следующим образом (рис. 8.3):

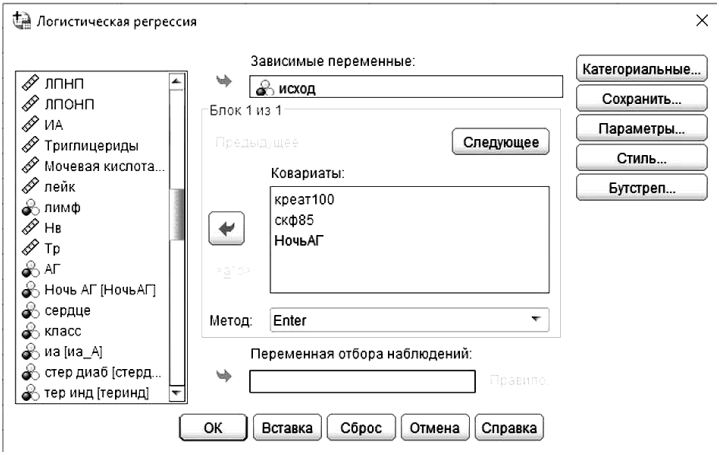


Рис. 8.3. Диалоговое окно «Реализация метода логистической регрессии»

В качестве зависимой переменной требуется выбрать ту, которая считается откликом, в качестве ковариат – предполагаемые факторы, влияющие на отклик.

Если среди факторов содержатся номинальные категориальные переменные, их надо указать кнопкой **Категориальные**. Программа сама преобразует их в набор дихотомических переменных.

Далее кнопкой **Параметры** целесообразно указать вывод доверительных интервалов для коэффициентов (рис. 8.4).

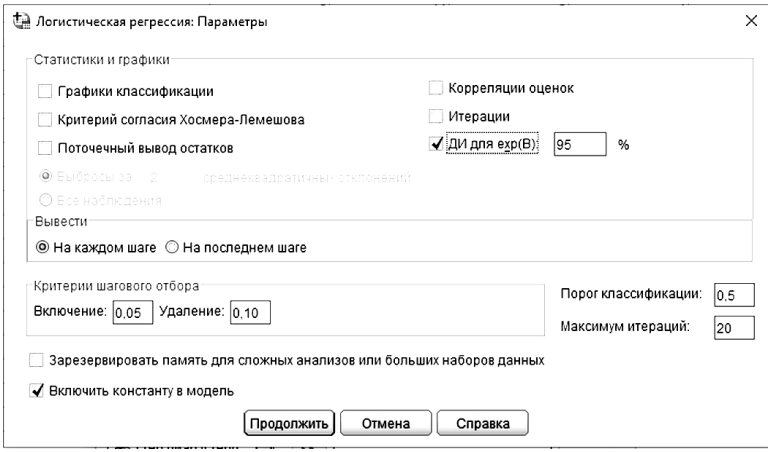


Рис. 8.4. Диалоговое окно «Вывод доверительных интервалов для коэффициентов»

Нажав **ОК** и запустив программу на исполнение, можно получить характерный для программы SPSS длинный вывод всевозможных таблиц и сведений, из которых интерес представляет таблица «Переменные в уравнении» (рис. 8.5).

Переменные в уравнении

								95% доверительный интервал для EXP(B)	
		B	Среднеквадратичная ошибка	Вальд	ст.св.	Знач.	Exp (B)	Нижняя	Верхняя
Шаг 1[a]	Группа	-2,355	,804	8,579	1	,003	,095	,020	,459
	альб	1.478	.737	4.017	1	.045	4.383	1,033	18,594
	Константа	1,520	1,023	2,208	1	,137	4,571		

Рис. 8.5. Переменные в уравнении

В приведенном примере (см. рис. 8.5) **Константа** – это коэффициент b_0, коэффициенты b_1 и b_2 – при переменных **альб** и **Группа**.

Прежде всего следует смотреть столбец **Знач**, в котором отображается уровень значимости коэффициентов b. В данном случае (см. рис. 8.5) видно, что уровни значимости коэффициентов при переменных меньше 0,05, следовательно, можно считать их отличными от нуля. Такое уравнение можно считать применимым для дальнейшего исследования.

Если уровень значимости при каком-либо коэффициенте будет больше принятого предельного уровня 0,05, то такое уравнение непригодно, и следует продолжить поиск статистически значимых факторов.

Если предположить, что данное уравнение принято, то требуется рассмотреть, какие еще данные есть в таблице (см. рис. 8.5). В столбце В выводятся собственно коэффициенты, входящие в модель.

$$p(x_n) = \frac{1}{1+e^{(-(1{,}52 + 1{,}478 \cdot \text{альб} - 2{,}355 \cdot \text{Группа}))}}.$$

По этому уравнению можно рассчитать вероятность наступления события (отклик=1) при различных значениях факторов.

Далее приводятся значения среднеквадратичной ошибки и значения критерия Вальда значимости коэффициентов.

В следующих столбцах выведены значения экспонент коэффициентов b и доверительные интервалы для них.

Риски и шансы

Задачи определения риска наступления событий широко распространены. При этом параллельно с рисками для оценки событий употребляются шансы, пришедшие в анализ данных из карточных игр.

Риск (*risk*) – вероятность наступления события. Очевидно, что риск может меняться от 0 до 1.

Шанс (*odds*) – отношение вероятности наступления события к вероятности не наступления. Шанс может изменяться от 0 до ∞.

Обычно риски и шансы исследуются не сами по себе, а сравниваются для разных групп или при разных значениях факторов (табл. 8.1).

Таблица 8.1

Фактор риска

Событие	Да	Нет	
Есть	*A*	*B*	*A* + *B*
Нет	*C*	*D*	*C* + *D*
	A + *C*	*B* + *D*	*N*

Очевидно, что при наличии фактора риска (фактор=1) наступило С событий из А + С случаев, при отсутствии фактора риска (фактор=0) – В событий из случаев.

В соответствии с определениями, при наличии фактора риска:

Риск = A/(A + C).

Шанс = A/C.

При отсутствии фактора риска:

Риск = B/(B+D).

Шанс = B/D.

Чтобы сравнить величину влияния фактора на результат, применимы еще два понятия:

1. *Относительный риск* (*relative risk*) – отношение риска (при наличии фактора) к риску (при его отсутствии):

$$\mathrm{RR} = \frac{\mathrm{A}/\left(\mathrm{A}+\mathrm{C}\right)}{\mathrm{B}/\left(\mathrm{B}+\mathrm{D}\right)}.$$

2. *Отношение шансов* (*odds ratio*) – отношение шанса (при наличии фактора) к шансу (при его отсутствии):

$$\mathrm{OR} = \frac{\mathrm{A}\cdot\mathrm{D}}{\mathrm{B}\cdot\mathrm{C}}.$$

Пример (результаты исследований, основанных на понятиях рисков и шансов)

Курение влияет на предрасположенность к заболеваниям легких.

В проведенном исследовании из 300 курящих у 225 были проблемы с легкими, у 75 – не было. При этом из 700 некурящих проблемы были у 75, у 625 проблем с легкими не было (табл. 8.2).

Таблица 8.2

		Да	Нет	
Проблемы с легкими	Есть	225	75	300
	Нет	75	625	700
		300	700	1000

Относительный риск

Риск = 225/300 = 75 % (у курящих).

Риск = 75/700 = 11 % (у некурящих).

Следовательно, относительный риск заболеваний легких, связанный с курением, составляет: RR = 75/11 = 6,8.

Отношение шансов

Шанс = 225/75 = 3 (у курящих).

Шанс = 75/625 = 0,12 (у некурящих).

Следовательно, отношение шансов OR = 3/0,12 = 25.

Очевидно, что отношение шансов сильнее выделяет зависимость.

Отношение шансов и логистическая регрессия

Логистическая регрессия может использоваться в качестве инструмента для вычисления отношения шансов и его доверительного интервала.

Для этого фактор следует закодировать как дихотомический. В этом случае при получении уравнения регрессии со значимым коэффициентом b при данном факторе отношение шансов будет равно Exp(b). Что не сложно доказать, подставив значения в уравнение регрессии.

Обратите внимание: в выходной форме содержится не только Exp(b), но и его доверительный интервал, т. е. непосредственно из выходной формы логистической регрессии получено отношение шансов с его доверительным интервалом. Если доверительный интервал OR не включает 1, то отношение шансов является статистически значимым.

8.4. ROC-анализ

ROC-анализ представляет собой графический метод оценки качества диагностического метода и выбора дискриминационного порога для разделения диагностируемых классов.

По существу, задачи диагностики относятся к задачам исследования связей при дихотомическом отклике. В данном случае дихотомическим откликом-событием является факт подтверждения диагноза или прогноза.

Диагностические характеристики

В диагностических и прогностических задачах используется понятийная система, имеющая свою специфику.

Для диагностики (прогнозирования) применяют некоторый фактор, считая, что если фактор = 1, то диагноз подтвердится (событие случится), если фактор = 0, то диагноз не подтвердится (или событие не случится).

Есть некоторый эффективный метод, которым можно проверить, подтвердился ли диагноз (случилось ли событие), так называемый «золотой стандарт», и его результат принято считать заведомо истинным (табл. 8.3):

Таблица 8.3

Значение диагностического фактора	Отклик – фактический диагноз	
	1	0
1	ИП	ЛП
0	ЛО	ИО

В результате заполняется табличка, в ячейках которой находятся следующие параметры:

ИП (*истинно положительные*) – те случаи, когда фактор = 1 и диагноз подтвердился.

ЛП (*ложно положительные*) – здесь фактор = 1, но диагноз не подтвердился.

ЛО (*ложно отрицательные*) – диагностический фактор = 0, но диагноз подтвердился.

ИО (*истинно отрицательные*) – фактор = 0, и диагноз не подтвердился.

Также добавляются следующие диагностические характеристики:

$$\text{Чувствительность} = \frac{\text{Истинно-положительный}}{\text{Истинно-положительный + Ложно-отрицательный}}$$

$$\text{Специфичность} = \frac{\text{Истинно-отрицательный}}{\text{Ложно-положительный + Истинно-отрицательный}}$$

$$\text{Доля ложно-позитивных} = \frac{\text{Ложно-положительный}}{\text{Ложно-положительный + Истинно-отрицательный}}$$

$$\text{Доля ложно-негативных} = \frac{\text{Ложно-отрицательный}}{\text{Истинно-положительный + Ложно-отрицательный}}$$

$$\text{Прогностическая ценность положительного результата} = \frac{\text{Истинно-отрицательный}}{\text{Истинно-положительный + Ложно-положительный}}$$

$$\text{Прогностическая ценность отрицательного результата} = \frac{\text{Истинно-отрицательный}}{\text{Ложно-отрицательный + Истинно-отрицательный}}$$

$$\text{Точность} = \frac{\text{Истинно-положительный + Истинно-отрицательный}}{\text{Все положительные + Все отрицательные}}$$

Главные характеристики здесь – чувствительность и специфичность, соответственно, способность диагностировать заболевания и способность не ставить неверные диагнозы.

ROC-анализ в SPSS

Чтобы воспользоваться ROC-анализом, надо иметь переменную отклика дихотомического типа и переменную с рассчитанной вероятностью того, что отклик = 1. Ее можно создать вручную, однако логистическая регрессия дает возможность сделать это автоматически, для чего требуется запустить логистическую регрессию кнопкой **Сохранить** и поставить галочку напротив параметра **Вероятности** (рис. 8.6).

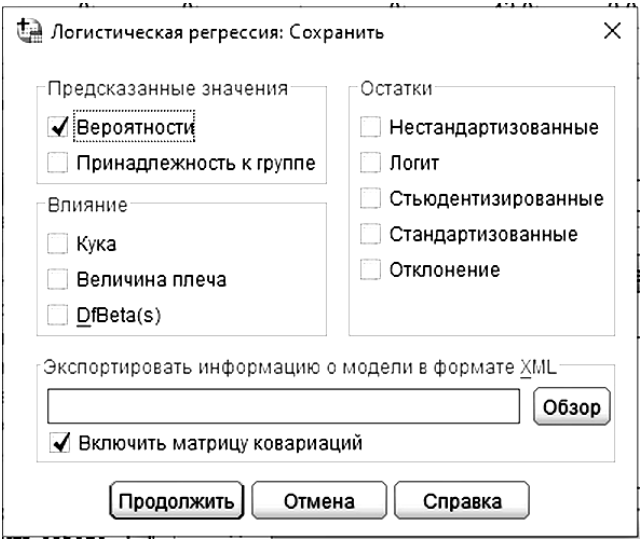


Рис. 8.6. Диалоговое окно «ROC-анализ в SPSS»

В этом случае программа автоматически добавит новую переменную **PRE_n**, поместит ее в конце и запишет в нее рассчитанные значения вероятностей.

Окно ROC-анализа выглядит следующим образом (рис. 8.7):

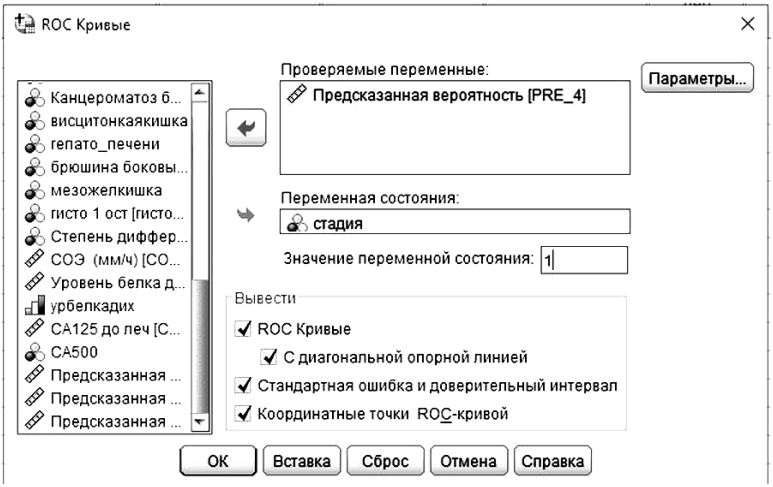


Рис. 8.7. Диалоговое окно «ROC-кривые в SPSS»

В поле **Проверяемая переменная** поместить рассчитанные вероятности, в поле **Переменная состояния** – переменную отклика. Также надо поставить галочки во все чек-боксы. В результате получен график с ROC-кривой (рис. 8.8).

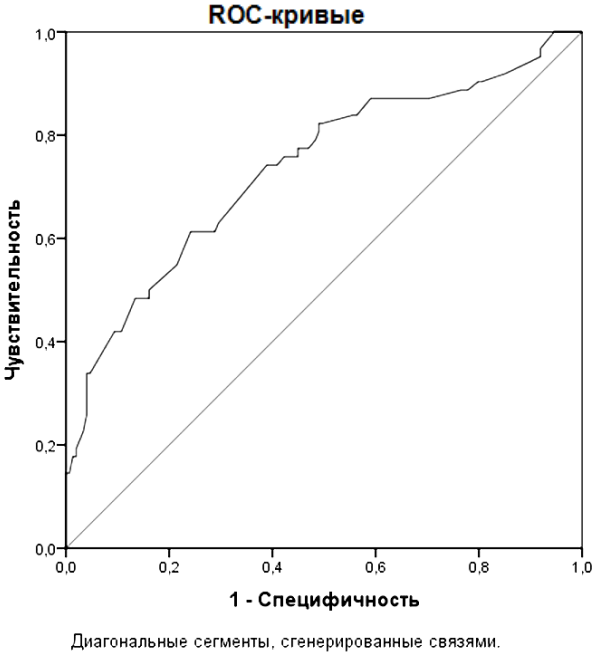


Рис. 8.8. Вид ROC-кривых в SPSS

Полученная кривая (рис. 8.8) показывает зависимость чувствительности от специфичности, точнее – от величины (1 – специфичность). Выбор пары «чувствительность–специфичность» зависит от того, какое значение вероятности, рассчитанное по логистической функции, принимать пороговым, т. е. если расчет показал вероятность больше пороговой, можно считать, что диагноз подтвердился.

Естественно, с увеличением чувствительности уменьшается специфичность, и выбранная точка является неким компромиссом. Пару «чувствительность–специфичность» выбирают по таблице с точными координатами кривой (рис. 8.9). Например, в этом фрагменте можно выбрать разделяющую вероятность 0,12 и соответствующие ей: чувствительность, равную 0,935, и специфичность, равную 0,886.

,0000000	1,000	1,000
,0825736	1,000	,993
,0912275	1,000	,973
,1006886	1,000	,946
,1086834	,968	,919
,1140665	,952	,919
,1222472	,935	,886

Рис. 8.9. Фрагмент таблицы координат ROC-кривой

Кроме того, в выдаче присутствуют данные по точности метода (рис. 8.10).

Область	Стандартная Ошибка[a]	Асимптотическая знч.[b]	Асимптотический 95% доверительный интервал	
			Нижняя граница	Верхняя граница
,735	,040	,000	,656	,813

Рис. 8.10. Фрагмент таблицы точности метода

Площадь под кривой равна предсказательной точности математической модели. В данном примере (см. рис. 8.10) она составляет 73,5 % и имеет 95%-ный доверительный интервал: 65,6 %–81,3 %.

9. РЕГРЕССИЯ

9.1. Метод наименьших квадратов

Методы регрессии применяются в том случае, когда и факторы, и отклик являются количественными.

Регрессия в математической статистике – это зависимость среднего значения одной величины y от другой величины (или нескольких величин) x. В отличие от строгой функциональной зависимости $y = f(x)$ в регрессионной модели одному и тому же значению величины x могут соответствовать несколько значений величины y, т. е. при фиксированном значении x величина y имеет некоторое случайное распределение.

Очевидно, в общем случае существует один отклик – y, и может быть несколько факторов $x_1 \ldots x_n$. Функция $f(x)$ может быть произвольной.

Линейная регрессия

Имеются две непрерывные переменные $x = (x_1, x_2, ..., x_n)$, $y = (y_1, y_2, ..., y_n)$. Точки размещены на двумерном графике рассеяния, и считается, что соотношение является линейным, если данные аппроксимируются прямой линией.

Если предположить, что y зависит от x, причем изменения в y вызываются именно изменениями в x, можно определить линию регрессии (регрессия y на x), которая лучше всего описывает прямолинейное соотношение между этими двумя переменными (рис. 9.1).

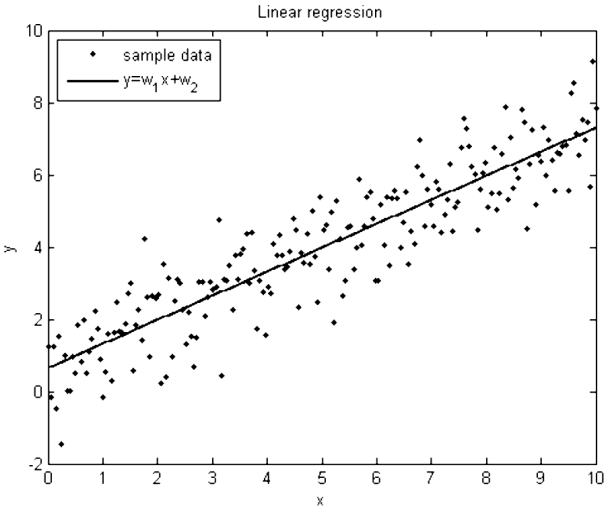


Рис. 9.1. График рассеяния

Математическое уравнение, которое оценивает линию простой (парной) линейной регрессии, следующее:

$$Y = a + bx,$$

где Y – зависимая переменная, или переменная отклика – значение, которое ожидают для y (в среднем), если знают величину x, т. е. это – «предсказанное значение y»;

a – свободный член (пересечение) линии оценки (это значение Y, когда $x = 0$);

b – коэффициент регрессии, представляющий собой величину, на которую Y увеличивается в среднем, если увеличивают x на 1;

x – независимая переменная, или предиктор.

Метод наименьших квадратов.
Нахождение коэффициентов регрессии

Метод наименьших квадратов – алгоритм, по которому находят коэффициенты регрессии, т. е. находится такая линия, что сумма квадратов отклонений наблюдаемых откликов минимальна.

Проверка данных и оценка регрессии

Квадраты расстояний легко записываются аналитически, и задача имеет аналитическое решение, что позволяет быстро рассчитать коэффициенты вручную. С приходом ЭВМ несложно было бы написать алгоритм минимизации расстояний, однако традиция сохранилась.

Большие отклонения более значимы, поэтому квадратичная зависимость автоматически лучше отслеживает большие расстояния.

Выбросы

Данные, участвующие в регрессионном анализе, являются стохастическими и имеют некоторое отклонение от средней величины, однако могут встречаться наблюдения, которые отстоят от среднего далеко. Их называют выбросами. Например, это можно сказать о наблюдениях за пределами $\pm 3\sigma$. Квадрат таких отклонений вносит большое возмущение в расчет, и полученные коэффициенты оказываются смещенными.

Как правило, выбросы не являются следствием собственной вариабельности процесса, но связаны с ошибками измерений или записи.

Поэтому перед расчетом регрессии следует провести проверку на выбросы и исключить сомнительные наблюдения.

Проверка значимости коэффициентов регрессии

Проверка статистической значимости отличия коэффициентов от нуля позволяет считать их значимыми.

При ручном расчете сравнивают отношение коэффициента b к его стандартной ошибке с критическими значениями критерия Стьюдента. При расчетах на ЭВМ проверка значимости коэффициентов производится автоматически.

Оценка качества линейной регрессии: коэффициент детерминации R^2. Из-за линейного соотношения Y и X ожидают, что Y изменяется по мере того, как изменяется X, и называют это вариацией, которая обусловлена или объясняется регрессией.

Однако на практике, кроме этой обусловленной вариации, существует еще дополнительная (или остаточная) вариация, связанная со стохастическим существом процесса.

Соотношение обусловленной вариации и случайной определяет качество регрессии и говорит о том, насколько точно можно предсказать отклик по фактору.

Долю общей дисперсии, которая объясняется регрессией, называют *коэффициентом детерминации*, обозначают R^2 и считают, что он позволяет субъективно оценить качество уравнения регрессии.

В парной линейной регрессии эта величина является квадратом коэффициента корреляции.

Проверка качества модели по критерию F. Для определения статистической значимости коэффициента детерминации проверяется гипотеза: «Коэффициент детерминации равен нулю». Для проверки нулевой гипотезы используется статистика:

$$F = \frac{R^2(n-2)}{1-R^2},$$

которая при справедливости H_0 имеет F-распределение Фишера с $\nu_1 = 1$, $\nu_2 = n - 2$ степенями свободы.

Вычисленный критерий F сравнивается с критическим значением $F_{кр}$:

– если $F_{набл} < F_{кр}$, то нет оснований для отклонения H_0, т. е. коэффициент детерминации незначим (или: уравнение регрессии незначимо);

– если $F_{набл} > F_{кр}$, то нулевая гипотеза отклоняется, т. е. R^2 статистически значим (или: уравнение регрессии значимо в целом).

9.2. Множественная регрессия

Множественная регрессия отличается только тем, что в ее уравнении не один, а несколько предикторов. Принцип минимизации квадратов расстояний остается, только расстояние вычисляется в многомерном пространстве.

Нелинейная регрессия

В уравнении нелинейной регрессии в качестве функции может браться произвольная зависимость. Обычно это – показательная функция, логарифмическая или полином. Не существуют методы автоматического определения функции регрессии, и ее форму должен задавать исследователь (рис. 9.2).

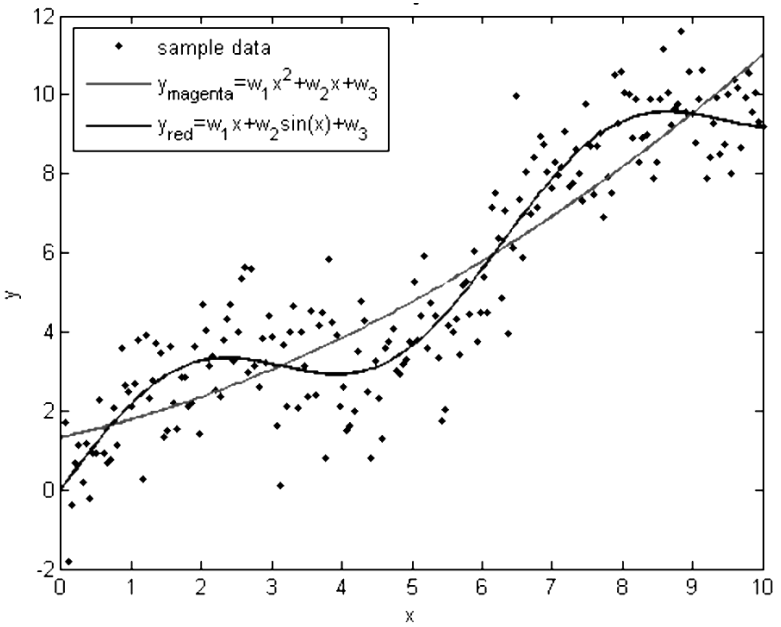


Рис. 9.2. Описание фактического рассеяния наблюдений с использованием квадратичной и тригонометрической функций

Для вычисления коэффициентов нелинейной функции обычно проводят пребразование данных так, чтобы для новых переменных получалась линейная зависимость.

9.3. Доверительный интервал отклика

Регрессионная функция позволяет предсказывать по факторам значение отклика, поэтому применимо понятие доверительного интервала, который должен рассчитываться.

Отклик имеет некоторые границы, в которые он должен попадать:

$$Y = \beta \cdot X + a \pm \varepsilon.$$

Регрессия в программе STATISTICA

В программе STATISTICA парная и множественная регрессии расположены в одном модуле по адресу: **Statistics>Multiple Linear Regression** (рис. 9.3).

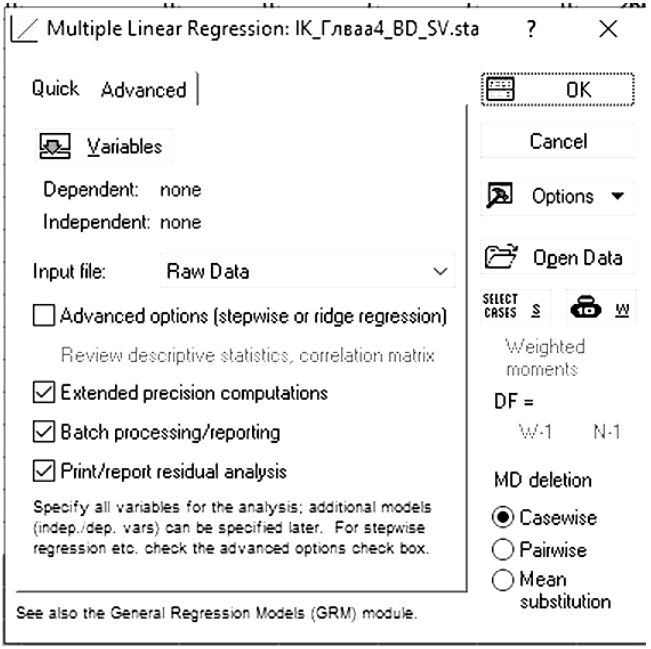


Рис. 9.3. Диалоговое окно «Statistics>Multiple Linear Regression»

Здесь надо задать зависимую переменную и один или несколько факторов. Также можно поэкспериментировать с простановкой галочек и изучить, какие выходные формы появляются дополнительно (рис. 9.4).

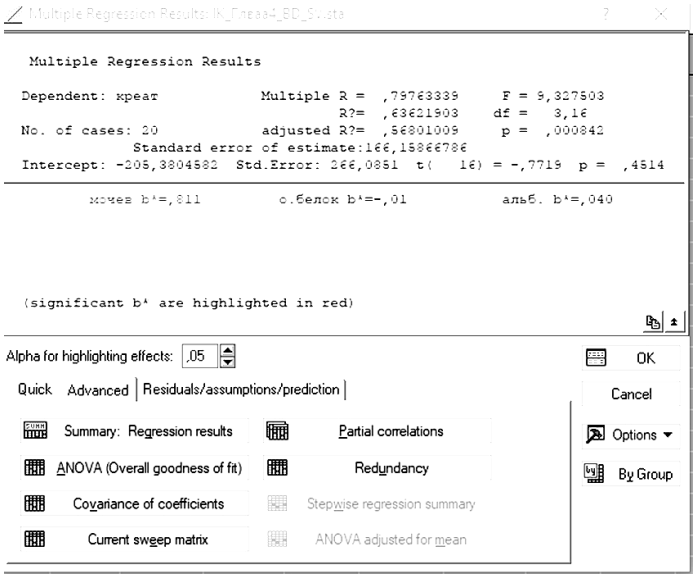


Рис. 9.4. Диалоговое окно «Парная и множественная регрессии»

Вверху выходной формы (рис. 9.4) приводятся значение коэффициента детерминации R^2 и результаты проверки его значимости. В данном случае критерий F = 9,33, уровень значимости (p = 0,0008) существенно меньше критического 0,05.

Следовательно, модель статистически значима в целом, т. е. модель адекватна.

Основные результаты следующие (рис. 9.5):

R= ,79763339 R?= ,63621903 Adjusted R?= ,56801009
F(3,16)=9.3275 p<,00084 Std.Error of estimate: 166,16

N=20	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			-205,380	266,0851	-0,771860	0,451448
мочев	0,811252	0,165032	31,082	6,3230	4,915724	0,000155
о.белок	-0,006276	0,227158	-0,139	5,0418	-0,027630	0,978299
альб.	0,039606	0,238056	1,470	8,8335	0,166374	0,869947

Рис. 9.5. Таблица результатов

Очевидно, что статистически значим только один из коэффициентов при факторах. Из этого следует, что хотя модель в целом признана адекватной, оставлять в ней незначимые факторы нельзя, следует их из модели исключить и, возможно, попробовать включить другие.

10. ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

10.1. Активный эксперимент

Активный эксперимент

Планирование эксперимента (англ. *experimental design techniques*) – это комплекс мероприятий, направленных на эффективную постановку опытов. Основная цель планирования эксперимента – достижение максимальной точности измерений при минимальном количестве проведенных опытов и сохранении статистической достоверности результатов.

По приведенному определению можно сделать несколько выводов:

1. Планирование эксперимента предполагает *активный эксперимент*, в котором исследователь сам задает значения факторов, и только отклик остается стохастическим.

2. Полученные в процессе спланированного эксперимента результаты будут основываться на регрессии, причем в общем случае это будет многофакторная нелинейная регрессия.

3. Планирование эксперимента актуально там, где проведение опытов трудоемко или требует много времени. В промышленности это – оптимизация конструкции, требующая для каждого опыта изготовления новых образцов. В сельском хозяйстве опыт может проводиться раз в год.

10.2. Факторы и уровни

Планирование эксперимента предполагает, что существует некоторый объект исследования с неизвестными свойствами, существует отклик (одна или несколько переменных) и существует некоторое количество переменных, влияющих на отклик и называемых факторами.

В результате исследования необходимо получить функцию, описывающую взаимосвязь отклика и факторов. В общем виде она выглядит следующим образом:

$$Y = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} b_{iu} x_i x_u + \sum_{i=1}^{k} b_{ii} x_i^2 + \mathbf{K}$$

Это – общий вид многофакторной квадратичной функции. В ней существует свободный член, первая степень факторов, квадраты факторов,

а также всевозможные сочетания их произведений. Абсолютное большинство задач ограничиваются функцией второй степени.

Традиционно количество факторов обозначают буквой k.

В задачах планирования эксперимента факторы могут принимать несколько фиксированных значений, которые называют *уровнями*.

Количество таких уровней обозначают буквой n.

Значит, для исследования всех вариантов сочетаний значений факторов требуется n^k опытов.

Уровни. Необходимое количество уровней

Уравнение регрессии в общем случае описывает некоторую многомерную поверхность в пространстве факторов. Ее форма зависит от количества точек, через которые она должна пройти.

Если заданы два уровня, зависимость будет представлять собой линию или плоскость. Такой вариант может быть хорош там, где ожидается монотонная зависимость отклика от факторов. Если решается задача оптимизации, поверхность будет иметь некоторые максимумы или минимумы (она должна быть выпуклой или вогнутой), следовательно, должна определяться минимум тремя точками. Случай с двумя перегибами потребует четырех точек, однако задачи такого рода редки, поэтому на практике исследователи чаще всего сталкиваются с задачами, в которых требуются планы с двумя или тремя уровнями.

Кодирование факторов

Областью определения факторов называется диапазон изменения их значений, принятый при реализации плана эксперимента:

$$X_i \in [X_{i\min}; X_{i\max}].$$

Для 2-факторного эксперимента область определения представляет собой прямоугольник (рис. 10.1).

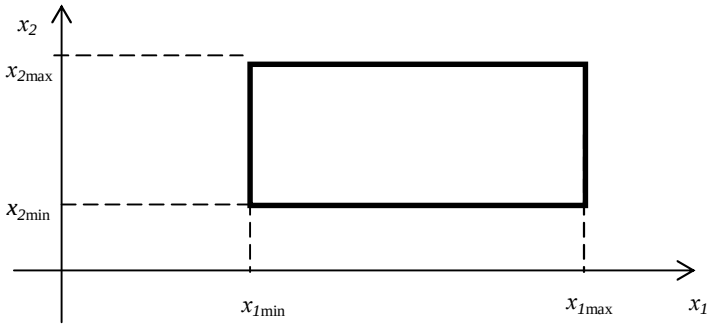


Рис. 10.1. Область определения факторов для двухфакторного эксперимента

Уровнем фактора называется его значение, фиксируемое в эксперименте. Экспериментатор может устанавливать любой уровень фактора в пределах его области определения.

Различают верхний, нижний и нулевой уровни. Верхний и нижний уровни соответствуют границам области определения $X_{i\max}$ и $X_{i\min}$. Нулевой уровень соответствует середине интервала:

$$X_{i0} = \frac{X_{i\min} - X_{i\max}}{2}.$$

Интервалом варьирования называют величину, равную максимальному отклонению уровня фактора от нулевого:

$$\Delta X_i = X_{i0} - X_{i\min} = X_{i\max} - X\ .$$

Кодирование. Для дальнейшего планирования эксперимента целесообразно перейти от натуральных значений факторов к кодированным:

$$x_i = \frac{X_i - X_{i0}}{\Delta X_i}.$$

Кодированные значения любого фактора на нижнем, верхнем и нулевом уровнях составляют:

$$x_{i\min} = -1, \quad x_{i\max} = 1, \quad x_{i0} = 0.$$

Область определения кодированных факторов для 2-факторного эксперимента представляет собой квадрат, для 3-факторного – куб, для *k*-факторного – *k*-мерный куб.

Таким образом, в данном методе все переменные в обязательном порядке перекодируются, превращаются в нормированные и безразмерные. Например, в других науках обычно проводится нормирование «на единицу», т. е. интервал равен единице, нормированный фактор изменяется от 0 до 1. В данном методе интервал изменения оказывается равным 2 (рис. 10.2).

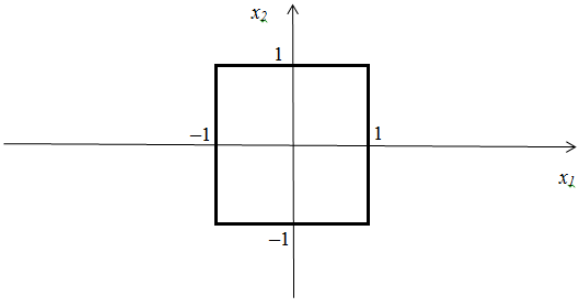


Рис. 10.2. Область определения кодированных факторов для двухфакторного эксперимента

10.3. Полный факторный план

Полный факторный план (эксперимент) (ПФЭ) – это эксперимент, в котором реализуются все возможные неповторяющиеся комбинации уровней факторов.

Для полного факторного плана требуется n^k опытов. Необходимое количество опытов при разных количествах факторов и уровней показано в таблице.

Таблица

Необходимое количество опытов при разных количествах факторов и уровней

n	k		
2	4	8	16
3	9	27	81
4	16	64	256

Матрица планирования эксперимента

Для проведения опытов составляют матрицу планирования, которая в случае двух уровней выглядит следующим образом:

Номер опыта	Фактор				Параметр
	x_1	x_2	x_3	x_4	
1	+1	–1	–1	–1	Y_1
2	–1	–1	–1	–1	Y_2
3	+1	+1	–1	–1	Y_3

Номер опыта		Фактор				Параметр
		x_1	x_2	x_3	x_4	
ПФЭ 2^2	4	–1	+1	–1	–1	Y_4
	5	+1	–1	+1	–1	Y_5
	6	–1	–1	+1	–1	Y_6
	7	+1	+1	+1	–1	Y_7
ПФЭ 2^3	8	–1	+1	+1	–1	Y_8
	9	+1	–1	–1	+1	Y_9
	10	–1	–1	–1	+1	Y_{10}
	11	+1	+1	–1	+1	Y_{11}
	12	–1	+1	–1	+1	Y_{12}
	13	+1	–1	+1	+1	Y_{13}
	14	–1	–1	+1	+1	Y_{14}
	15	+1	+1	+1	+1	Y_{15}
ПФЭ 2^4	16	–1	+1	+1	+1	Y_{16}

Уровни факторов записываются перебором, т. е. так, что первый фактор меняет знак при каждом переходе, второй – после каждого второго перехода, третий – после каждого четвертого и т. д.

Матрицы ПФЭ обладают рядом свойств, позволяющих проверить правильность их составления:

1. Свойство симметричности. Каждый фактор в матрице на верхнем уровне встречается столько же раз, сколько и на нижнем:

$$\sum_{u=1}^{n} x_{iu} = 0,$$

где i – номер опыта;

n – количество опытов, $n = 2^k$.

2. Свойство нормировки. Каждый фактор в матрице встречается только на уровнях –1 и +1:

$$\sum_{u=1}^{n} x_{iu}^2 = n.$$

3. Свойство ортогональности. Суммы почленных произведений двух любых столбцов равны нулю:

$$\sum_{u=1}^{n} x_{iu} \cdot x_{ju} = 0,$$

где j – номер столбца.

4. Свойство ротабельности. Точки в матрице выбираются так, что точность предсказания параметра одинакова во всех направлениях.

Расчет полного факторного плана на ЭВМ

В программе STATISTICA планирование эксперимента находится по адресу **Statistics>Industrial Statistics & Six Sigma>Experimental Design** (рис. 10.3).

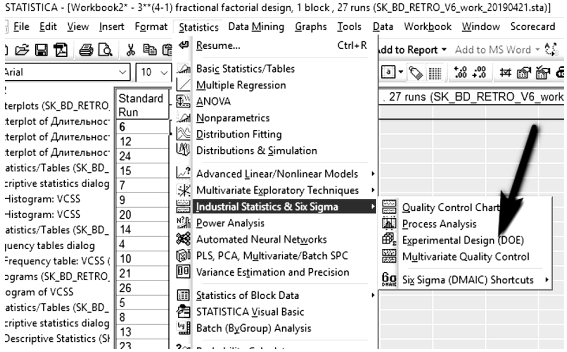


Рис. 10.3. Диалоговое окно «Statistics>Industrial Statistics & Six Sigma>Experimental Design»

Далее в диалоговом окне надо выбрать **2**(*k*-*p*) standard design** или **3**(*k*-*p*) and Box-Behnken designs**, соответственно, для 2-уровневого и 3-уровневого планов (рис. 10.4).

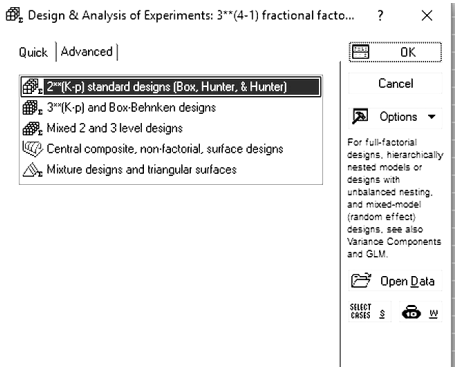


Рис. 10.4. Диалоговое окно «2**(*k*–*p*) standard design» или «3**(*k*–*p*) and Box-Behnken designs»

11. ДРОБНЫЙ ПЛАН ЭКСПЕРИМЕНТА. ПЛАНЫ ВТОРОГО ПОРЯДКА. ЦЕНТРАЛЬНЫЙ КОМПОЗИЦИОННЫЙ ПЛАН

11.1. Дробный факторный план

Количество опытов в полном факторном плане представляет собой степенную зависимость от числа факторов, поэтому с увеличением числа факторов резко возрастает количество опытов ПФЭ: при пяти факторах оно равно 32, при шести – 64 и т. д. Очевидно, выполнить такое количество опытов технически сложно.

Для преодоления данной проблемы были разработаны *дробные факторные планы*, в которых используется уменьшенное количество экспериментов.

Разрешение дробного факторного плана – это количество факторов полного плана, которому соответствует дробный план. Дробный план обозначают как **2**(*k*–*p*)**, и это означает, что у плана **2** уровня ***k*** факторов, и сокращается он на ***p*** порядков, в результате чего его матрица становится эквивалентной полному факторному плану, в котором **(*k*–*p*)** факторов. Это количество называют *разрешением дробного плана*. В данном случае количество опытов сокращается в 2^p раз. Например, имея 11 факторов, можно сократить план на 7 порядков и получить разрешение, равное 4. Таким образом, вместо $2^{11} = 2048$ опытов произведено $2^4 = 16$ (рис. 11.1).

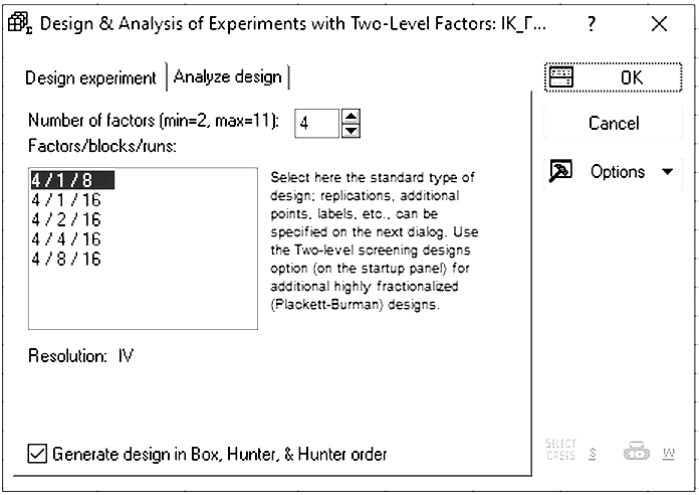


Рис. 11.1. Диалоговое окно «Блок реализации полного факторного плана и дробных планов эксперимента в программе STATISTICA»

Сокращение опытов основано на предположении о том, что в функции отклика тройственное (и более) взаимодействие факторов (т. е. их произведения) не вносит существенного вклада. В любом случае, алгоритм построения плана реализуется программно, и получают его уже в готовом виде.

В программе STATISTICA полный факторный план и дробные планы эксперимента реализованы в одном блоке

Так при выборе двухуровневых планов **2**(*k*–*p*)** и четырех факторов полному плану будет соответствовать вариант $2^4 = 16$, т. е. **4/1/16**. Далее расчеты для полного факторного плана и дробных планов одинаковы.

11.2. Планы второго порядка

Планами второго порядка называют планы с тремя уровнями, которые, соответственно, обеспечивают получение функции отклика в виде полинома второго порядка.

В результате после нормирования данных кроме точек +1 и –1 получают также точки 0 (рис. 11.2).

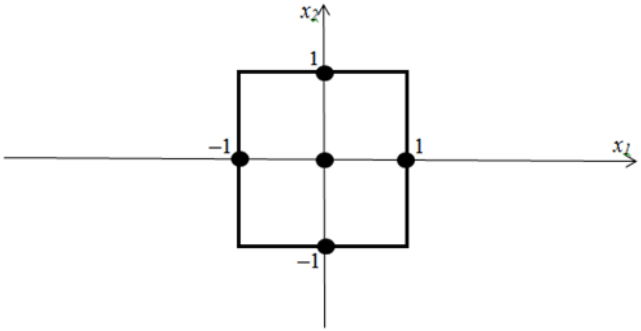


Рис. 11.2. Планы второго порядка

При этом количество экспериментов для полного плана составит 3^k, для дробного плана – $3^{(k-p)}$, и функция отклика будет иметь следующий вид:

$$Y = b_0 + \sum_{i=1}^{k} b_i X_i + \sum_{i,j=1}^{C} b_{ij} X_i X_j + \sum_{i=1}^{k} b_{ii} X_i^2 .$$

Следует отметить, что даже дробные планы второго порядка требуют большого количества опытов, поэтому после появления центральных планов на практике применяются редко.

11.3. Центральный композитный план

Ортогональный центрально-композиционный план (ОЦКП) второго порядка – специальный план эксперимента, который, кроме трех уровней, включает дополнительно «звездные точки».

Свойство ортогональности обеспечивается тем, что матрица планирования остается диагональной. План называется центральным, поскольку все точки расположены симметрично относительно центра плана.

В ОЦКП входят: ядро (полный факторный план первого порядка с 2^k точками), по две «звездных» точки для каждого фактора и одна центральная точка. Таким образом, общее количество опытов вычисляется по формуле

$$n = 2^k + 2k + 1.$$

«Звездные» точки отстоят от центра на расстояние, которое называют плечом. Плечо рассчитывается так, чтобы план был ортогональным (рис. 11.3).

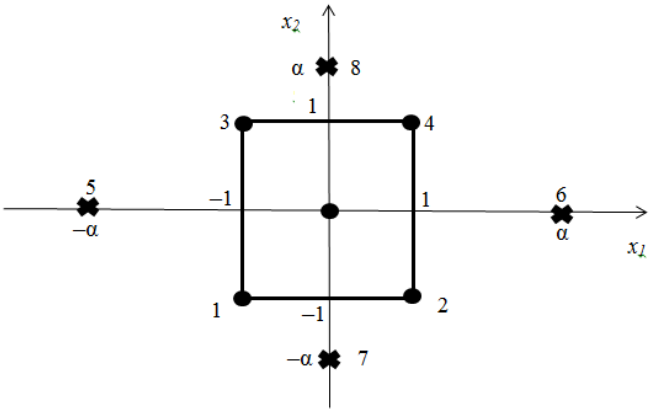


Рис. 11.3. Ортогональный центрально-композиционный план (ОЦКП) второго порядка

На практике оказывается, что ОЦКП обеспечивает получение функции отклика практически с той же степенью точности, что и соответствующий план второго порядка, при этом количество опытов оказывается намного меньше, особенно с увеличением количества факторов (таблица).

Таблица

Количество опытов	Количество факторов				
	2	3	4	5	6
ОЦКП	9	15	25	43	77
ПФЭ 3^2	9	27	81	243	729

11.4. Расчет факторных планов на ЭВМ

Специфика расчетов факторных планов состоит в том, что они осуществляются в два этапа. На первом этапе получают и редактируют матрицу плана, затем производят эксперименты по плану, на втором этапе вносят результаты эксперимента и проводят их анализ.

Получение матрицы дробного факторного плана

Следует помнить, что для получения матрицы плана эксперимента не нужны никакие исходные данные. Достаточно знать количество факторов и ограничение на число опытов. При выборе дробного плана необходимо указать вариант с соответствующим числом факторов и количеством опытов.

При выборе ОЦКП надо перейти по адресу: **Statistics>Industrial Statistics & Six Sigma>Experimental Design>Central composite, non-factorial, surface designs** (рис. 11.4).

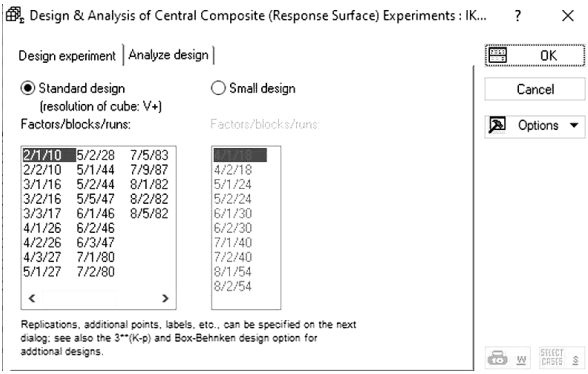


Рис. 11.4. Диалоговое окно «Выбор ОЦКП»

Далее в диалоговом окне надо выбрать план, соответствующий количеству факторов. Очевидно, здесь для каждого количества есть варианты, различающиеся количеством блоков. В базовом случае требуется один блок, поэтому выбор плана оказывается однозначным. Например, при 4 факторах – это 4/1/26, и потребуется 26 опытов.

Существует также вариант сокращенного плана (*small design*), в котором количество опытов почти в 2 раза меньше. Здесь исследователь должен из профессиональных оснований выбирать, что лучше: уменьшить количество факторов на единицу или уменьшить точность при использовании малого плана.

После нажатия **OK** переходят в диалоговое окно (рис. 11.5).

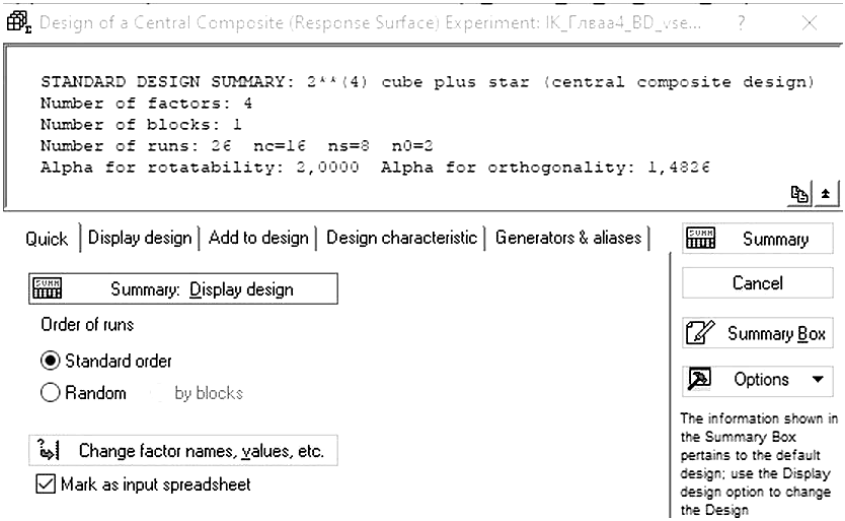


Рис. 11.5. Диалоговое окно «Сокращенный ОЦКП»

В верхней части окна представлен отчет о выборе плана и данные по значению плеча для «звездных» точек. Кнопкой **Summary Display Design** переходят к матрице результатов, причем можно выбрать варианты стандартного порядка опытов по номерам или случайного чередования номеров опытов для исключения влияния человеческого фактора (рис. 11.6).

Standard Run	2**(4) central composite, nc=16 ns=8 n0= F 1	F 2	F 3	F 4
1	-1,00000	-1,00000	-1,00000	-1,00000
2	-1,00000	-1,00000	-1,00000	1,00000
3	-1,00000	-1,00000	1,00000	-1,00000
4	-1,00000	-1,00000	1,00000	1,00000
5	-1,00000	1,00000	-1,00000	-1,00000
6	-1,00000	1,00000	-1,00000	1,00000
7	-1,00000	1,00000	1,00000	-1,00000
8	-1,00000	1,00000	1,00000	1,00000
9	1,00000	-1,00000	-1,00000	-1,00000
10	1,00000	-1,00000	-1,00000	1,00000
11	1,00000	-1,00000	1,00000	-1,00000
12	1,00000	-1,00000	1,00000	1,00000
13	1,00000	1,00000	-1,00000	-1,00000
14	1,00000	1,00000	-1,00000	1,00000
15	1,00000	1,00000	1,00000	-1,00000
16	1,00000	1,00000	1,00000	1,00000
17	-2,00000	0,00000	0,00000	0,00000
18	2,00000	0,00000	0,00000	0,00000
19	0,00000	-2,00000	0,00000	0,00000
20	0,00000	2,00000	0,00000	0,00000
21	0,00000	0,00000	-2,00000	0,00000
22	0,00000	0,00000	2,00000	0,00000
23	0,00000	0,00000	0,00000	-2,00000
24	0,00000	0,00000	0,00000	2,00000
25 (C)	0,00000	0,00000	0,00000	0,00000
26 (C)	0,00000	0,00000	0,00000	0,00000

Рис. 11.6. Матрица результатов

Данная матрица имеет абстрактный вид, в котором нет названий факторов, и величины представлены в нормированном виде. Этой матрицей уже можно пользоваться на практике, пересчитав при этом в рабочем журнале ее в реальные имена и числа. Однако можно отредактировать матрицу кнопкой **Change factor names, values etc** (рис. 11.7).

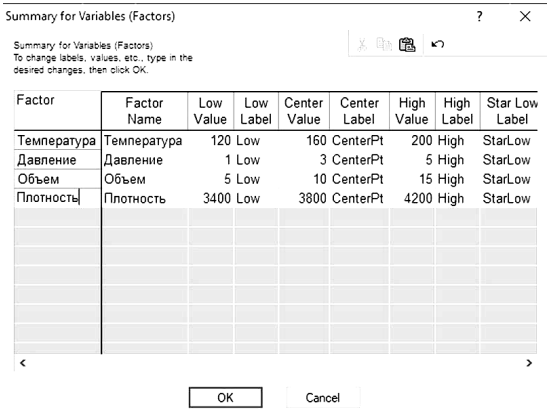


Рис. 11.7. Матрица результатов

Необходимо задать имена факторов и величины минимальных и максимальных значений в опытах, тогда можно получить план эксперимента, пригодный для непосредственного использования (рис. 11.8).

Standard	2**(4) central composite, nc=16 ns=8 n0=2 Runs=26 (			
Run	**Температура**	Давление	Объем	Плотность
1	120,0000	1,00000	5,00000	3400,000
2	120,0000	1,00000	5,00000	4200,000
3	120,0000	1,00000	15,00000	3400,000
4	120,0000	1,00000	15,00000	4200,000
5	120,0000	5,00000	5,00000	3400,000
6	120,0000	5,00000	5,00000	4200,000
7	120,0000	5,00000	15,00000	3400,000
8	120,0000	5,00000	15,00000	4200,000
9	200,0000	1,00000	5,00000	3400,000
10	200,0000	1,00000	5,00000	4200,000
11	200,0000	1,00000	15,00000	3400,000
12	200,0000	1,00000	15,00000	4200,000
13	200,0000	5,00000	5,00000	3400,000
14	200,0000	5,00000	5,00000	4200,000
15	200,0000	5,00000	15,00000	3400,000
16	200,0000	5,00000	15,00000	4200,000
17	80,0000	3,00000	10,00000	3800,000
18	240,0000	3,00000	10,00000	3800,000
19	160,0000	-1,00000	10,00000	3800,000
20	160,0000	7,00000	10,00000	3800,000
21	160,0000	3,00000	0,00000	3800,000
22	160,0000	3,00000	20,00000	3800,000
23	160,0000	3,00000	10,00000	3000,000
24	160,0000	3,00000	10,00000	4600,000
25 (C)	160,0000	3,00000	10,00000	3800,000
26 (C)	160,0000	3,00000	10,00000	3800,000

Рис. 11.8. План эксперимента

Анализ результатов эксперимента

Подготовка данных. Данные должны состоять из матрицы плана эксперимента с добавленным столбцом отклика, в который заносят результаты. Обычно их формируют в Excel, предварительно скопировав в него полученный план эксперимента.

Затем данные экспортируют в программу STATISTICA в обычном порядке (рис. 11.9).

	D:\!! KAFEDRA\!		МАГИСТРЫ\Тема 7	
	1 P	2 N	3 Семян кг/г	4 Урожай
1	30	30	200	37,0464
2	60	9,546215	140	31,38867
3	60	60	240,90757	36,10537
4	60	60	140	70,96444
5	60	60	39,09243	41,11284
6	30	90	80	41,21307
7	110,4538	60	140	15,70973
8	90	90	200	16,71169
9	60	110,4538	140	28,88867
10	90	30	80	23,37836
11	90	90	80	20,87836
12	30	30	80	40,71307
13	90	30	200	17,21169
14	60	60	140	71,46444
15	30	90	200	37,5464
16	9,546215	60	140	49,06761

Рис. 11.9. Фрагмент таблицы «Матрица исходных данных»

Далее обращаются по адресу: **Statistics>Industrial Statistics & Six Sigma>Experimental Design>Central composite, non-factorial, surface designs**, только сейчас требуется выбрать вкладку **Analyze Design** (рис. 11.10).

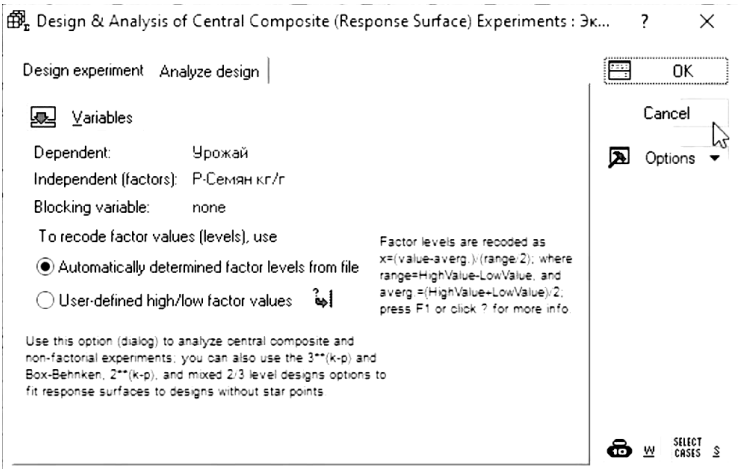


Рис. 11.10. Диалоговое окно «Выбор вкладки Analyze Design»

В окне (рис. 11.10) выбирают в качестве независимых переменных факторы, в качестве независимой переменной – отклик, в данном случае *Урожай*.

Оценка регрессии. Нажатие **ОК** открывает окно «Результаты» (рис. 11.11).

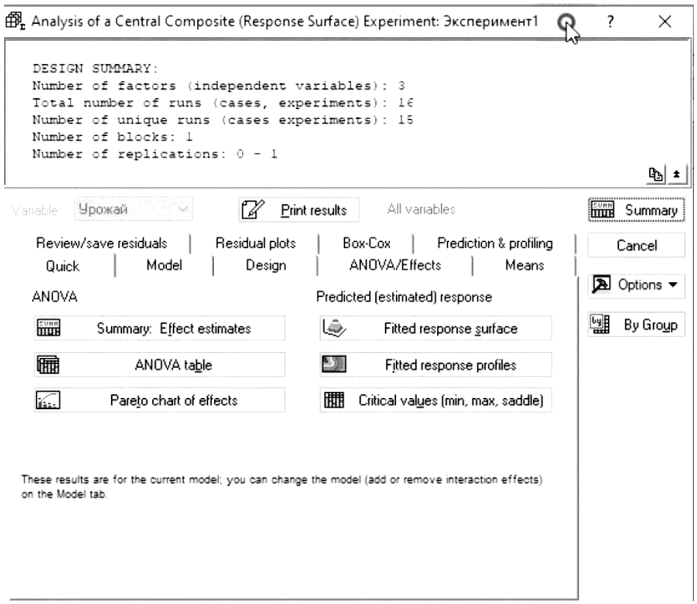


Рис. 11.11. Диалоговое окно «Результаты»

Кнопкой **Summary** открывают окно «Описание функции отклика» (рис. 11.12):

Effect Estimates; Var.:Урожай; R-sqr=,99602; Adj.,99004 (Эксперимент1)
3 factors, 1 Blocks, 16 Runs; MS Residual=2,871109
DV: Урожай

Factor	Effect	Std.Err.	t(6)	p	-95,% Cnf.Limt	+95,% Cnf.Limt	Coeff.	Std.Err. Coeff.	-95,% Cnf.Limt	+95,% Cnf.Limt
Mean/Interc.	71,4171	1,194635	59,7815	0,000000	68,4939	74,3402	71,4171	1,194635	68,4939	74,3402
(1)P (L)	-19,6883	0,917021	-21,4698	0,000001	-21,9321	-17,4444	-9,8441	0,458511	-10,9661	-8,7222
P (Q)	-28,4323	1,113403	-25,5364	0,000000	-31,1567	-25,7079	-14,2161	0,556701	-15,5783	-12,8539
(2)N (L)	-0,9086	0,917021	-0,9908	0,360023	-3,1525	1,3352	-0,4543	0,458511	-1,5762	0,6676
N (Q)	-30,0233	1,113403	-26,9653	0,000000	-32,7477	-27,2989	-15,0116	0,556701	-16,3738	-13,6494
(3)Семян кг/г(L)	-3,8205	0,917021	-4,1662	0,005903	-6,0644	-1,5767	-1,9103	0,458511	-3,0322	-0,7883
Семян кг/г(Q)	-24,0338	1,113403	-21,5859	0,000001	-26,7582	-21,3094	-12,0169	0,556701	-13,3791	-10,6547
1L by 2L	-1,0000	1,198146	-0,8346	0,435894	-3,9318	1,9318	-0,5000	0,599073	-1,9659	0,9659
1L by 3L	-0,7500	1,198146	-0,6260	0,554386	-3,6818	2,1818	-0,3750	0,599073	-1,8409	1,0909
2L by 3L	0,5000	1,198146	0,4173	0,690971	-2,4318	3,4318	0,2500	0,599073	-1,2159	1,7159

Рис. 11.12. Описание функции отклика

Коэффициенты, которые статистически значимо отличаются от нуля и, соответственно, имеют уровень значимости $p < 0{,}05$, выделены.

Проверка на адекватность. Проверка на адекватность по существу является проверкой соответствия уравнения тем данным, из которых оно было получено. Для этого кнопкой **ANOVA table** переходят к выходной форме с оценками по критерию Фишера (рис. 11.13).

Factor	SS	df	MS	F	p
(1)P (L)	1323.444	1	1323.444	460.9522	0.000001
P (Q)	1872.268	1	1872.268	652.1064	0.000000
(2)N (L)	2.819	1	2.819	0.9818	0.360023
N (Q)	2087.665	1	2087.665	727.1283	0.000000
(3)Семян кг/г(L)	49.835	1	49.835	17.3575	0.005903
Семян кг/г(Q)	1337.791	1	1337.791	465.9494	0.000001
1L by 2L	2.000	1	2.000	0.6966	0.435894
1L by 3L	1.125	1	1.125	0.3918	0.554386
2L by 3L	0.500	1	0.500	0.1741	0.690971
Error	17.227	6	2.871		
Total SS	4325.846	15			

Рис. 11.13. Оценки по критерию Фишера

В окне (рис. 11.13) выделены светлым цветом коэффициенты, статистически значимо отличные от нуля, соответствуя оценкам по критерию Стьюдента из предыдущей таблицы (см. рис. 11.12).

Визуальный анализ поверхности отклика. Модуль анализа результатов эксперимента по факторным планам позволяет непосредственно в нем же проводить визуальный анализ поверхности отклика и предлагает для этого два варианта: трехмерную поверхность и топографическую карту поверхности (рис. 11.14).

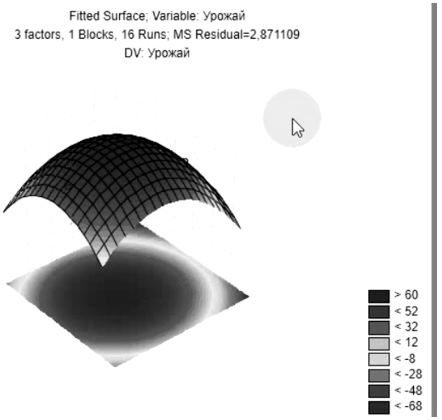


Рис. 11.14. Трехмерная поверхность и топографическая карта поверхности

В данном случае имеются три фактора, но в трехмерной проекции можно построить зависимость только для двух. Поэтому модуль позволяет выбрать два фактора, по которым будет строиться поверхность, а третий фактор зафиксировать на некоторой величине. Рассматривая поверхности при разных значениях третьего фактора, можно провести полный визуальный анализ.

Топографическая форма сходна по функционалу, но имеет другое отображение (рис. 11.15).

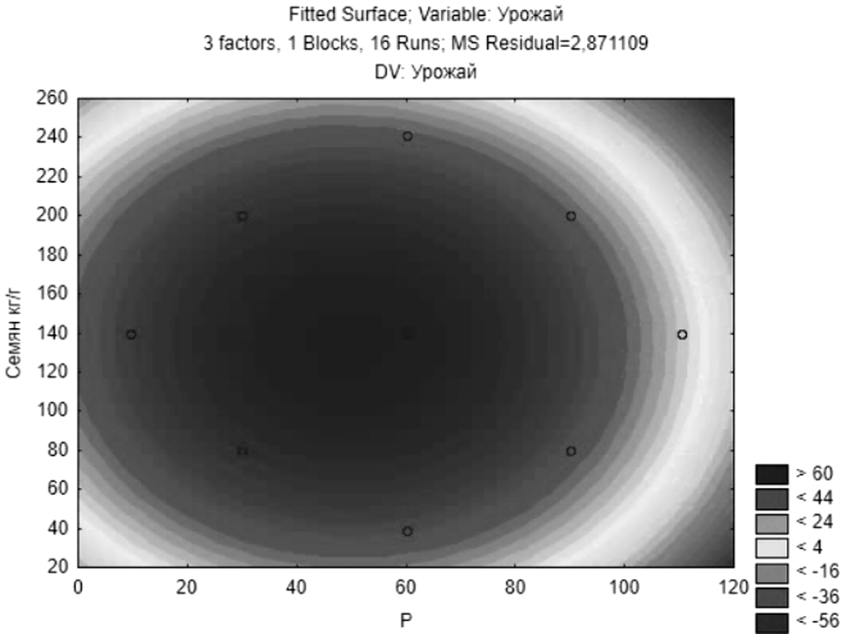


Рис. 11.15. Топографическая форма

В том и другом случаях (см. рис. 11.14, 11.15) на графики наносятся точки факторного плана.

Анализ остатков. Данный инструмент позволяет производить анализ разности между фактическими значениями и значениями, полученными по определенной формуле. Это – необходимая часть анализа, в котором требуется доказать, что остатки симметричны, нормально распределены и т. д.

Анализ экстремумов. В отдельном разделе **Critical values** можно найти точные значения функции, соответствующие критическим значениям – минимумам, максимумам и перегибам.

Модификация модели. Алгоритм метода наименьших квадратов рассматривал самое общее уравнение регрессии, однако имеется возможность вручную ограничить входящие в регрессию члены. Это имеет смысл сделать в отношении тех членов, чьи коэффициенты оказались незначимыми.

Данная операция производится во вкладке **Model** (рис. 11.16).

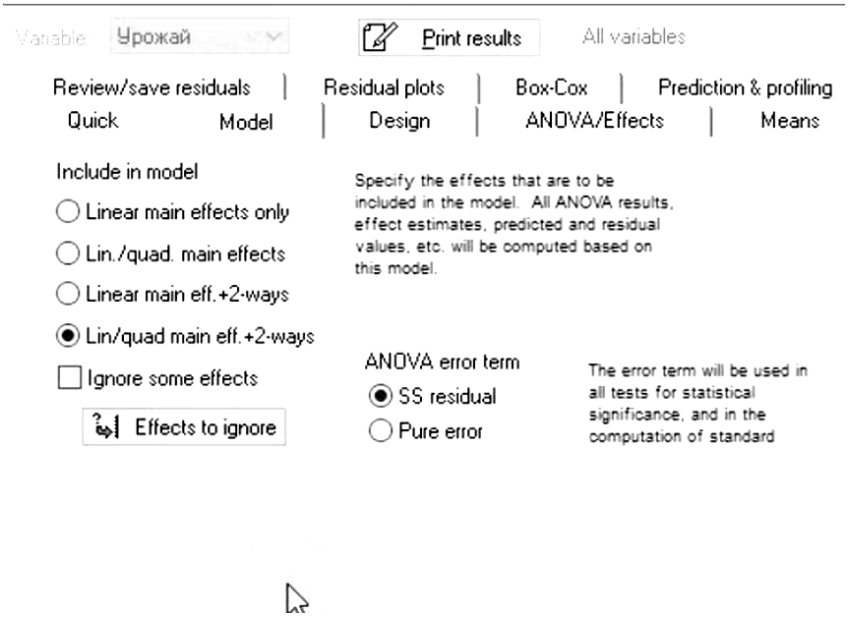


Рис. 11.16. Диалоговое окно «Ручное ограничение входящих в регрессию членов»

В разделе **Include in model** имеется возможность выбрать уровень модели, а также указать, какие эффекты желательно исключить.

Уравнение регрессии. Целью настоящего исследования является получение аналитической формы функции отклика в виде регрессии соответствующего уровня. Чтобы найти коэффициенты аналитической формы регрессии, надо во вкладке **ANOVA/effects** использовать кнопку **Regression coefficients** (рис. 11.17).

Factor	Regressn Coeff.	Std.Err.	t(6)	p	-95,% Cnf.Limt	+95,% Cnf.Limt
Mean/Interc.	-88,4484	6,868048	-12,8782	0,000013	-105,254	-71,6429
(1)P (L)	1,6298	0,097516	16,7137	0,000003	1,391	1,8685
P (Q)	-0,0158	0,000619	-25,5364	0,000000	-0,017	-0,0143
(2)N (L)	2,0003	0,097516	20,5125	0,000001	1,762	2,2389
N (Q)	-0,0167	0,000619	-26,9653	0,000000	-0,018	-0,0152
(3)Семян кг/г(L)	0,9070	0,052256	17,3562	0,000002	0,779	1,0348
Семян кг/г(Q)	-0,0033	0,000155	-21,5859	0,000001	-0,004	-0,0030
1L by 2L	-0,0006	0,000666	-0,8346	0,435894	-0,002	0,0011
1L by 3L	-0,0002	0,000333	-0,6260	0,554386	-0,001	0,0006
2L by 3L	0,0001	0,000333	0,4173	0,690971	-0,001	0,0010

Рис. 11.17. Аналитическая форма функции отклика

Здесь показаны коэффициенты регрессии при линейных (L) и квадратичных (Q) членах уравнения, уровень статистической значимости их отличия от нуля, их ошибки и доверительные интервалы. Последние обстоятельства позволяют вычислить стандартные ошибки и доверительные интервалы отклика.

Например: коэффициент при P(L) равен 1,63; при P(Q) – 0,016; при N(L) – 2,00 и т. д.

Например, получена функция отклика, которая выглядит следующим образом:

$$\text{Урожай} = -88{,}45 + 1{,}63\text{P} - 0{,}016\text{P}^2 + 2{,}00\text{N} - 0{,}017\text{N}^2 + \\ + 0{,}91\text{Семян} - 0{,}0033\text{Семян}^2.$$

Таким образом, получена удовлетворительная функция отклика.

12. ПОНЯТИЕ О СПЕЦИАЛЬНЫХ ПЛАНАХ ЭКСПЕРИМЕНТА

12.1. Планы для смесей

Специальные вопросы возникают, когда анализируются смеси компонентов, которые в сумме должны давать константу. Например, если требуется оптимизировать вкус фруктового пунша, состоящего из соков пяти фруктов, то сумма долей всех соков в каждой смеси должна быть равна 100 %. Такая задача оптимизации смесей часто встречается в производстве пищи, при очистке или производстве химикатов или лекарств. Разработан ряд планов специально для анализа смесей.

Например: имеются 6 вариантов смесей из трех компонентов.

Сумма долей компонентов для каждой смеси равна 1,0, поэтому значения компонентов в каждой смеси могут интерпретироваться как пропорции. Если нанести эти данные на график в виде обычной трехмерной диаграммы рассеяния, станет очевидно, что точки образуют треугольник в трехмерном пространстве. Только точки внутри треугольника, где сумма значений компонентов равна 1, представляют настоящие смеси. Следовательно, можно просто наносить данные только в треугольник (в данном случае – двумерный), чтобы изображать значения компонентов (пропорции) для каждой смеси (рис. 12.1).

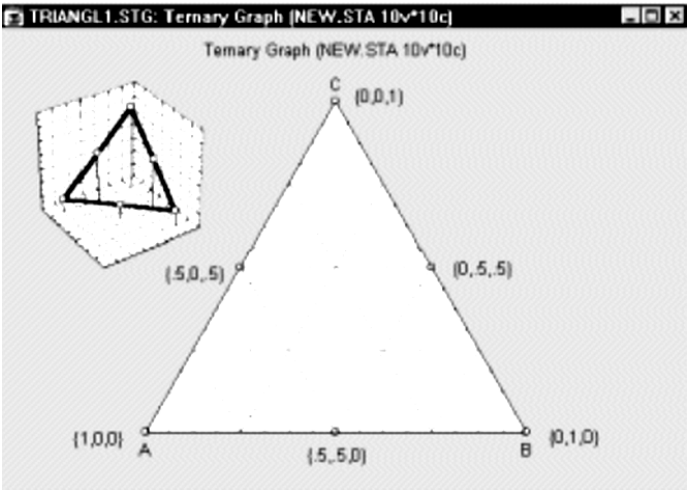


Рис. 12.1. Трехмерная диаграмма рассеяния

Вершина, соответствующая конкретному фактору, представляет собой чистую смесь, т. е. состоящую только из данного компонента. Поэтому координата компонента соответствующей вершине равна 1 (или 100 %, или любой другой величине в зависимости от шкалирования) и равна 0 для всех других компонентов. На стороне, противоположной соответствующей вершине, значение данного компонента равно 0, для других компонентов – 0,5 (или 50 % и т. д.).

Можно теперь добавить четвертое измерение и нанести на график значения зависимой переменной или функцию (поверхность) для каждой точки внутри треугольника. Поверхность отклика может быть представлена либо в трехмерном пространстве, где предсказываемый отклик (оценка Taste-Вкуса) наносится, как расстояние поверхности от плоскости треугольника, либо представлена в виде контурной диаграммы, где контуры равной высоты наносятся в двумерном треугольнике (рис. 12.2).

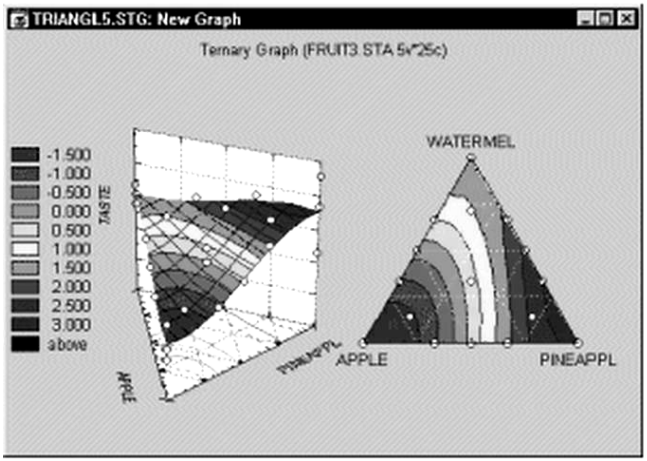


Рис. 12.2. Контурная диаграмма

Модуль для работы с планами для смесей находится по адресу: **Statistics>Industrial Statistics & Six Sigma>Experimental Design>Mixture design and triangular surfaces**.

12.2. Планы с ограничениями

В некоторых задачах существуют ограничения на содержание в смеси некоторых компонентов.

Для планов требуются точки-вершины, т. е. чистые смеси, состоящие из одного компонента. На практике такие точки обычно неприемлемы, так как не могут производиться из соображений стоимости или по причине других ограничений.

Например: требуется изучить эффект воздействия пищевых добавок на вкус фруктового пунша. Дополнительный ингредиент может варьироваться в узких пределах: не может превышать некоторого процента от общей массы. Очевидно, что фруктовый пунш, составленный только из чистой добавки, не будет на самом деле пуншем. Такого рода ограничения весьма типичны.

Или: в смеси из трех компонентов на компонент А наложено ограничение *x*A.3. Общая сумма для трехкомпонентной смеси должна быть равной 1. Это ограничение может быть показано на треугольной диаграмме в виде прямой с треугольными координатами для ***x*A =.3**, т. е. прямой, параллельной стороне треугольника, противоположной вершине А (рис. 12.3).

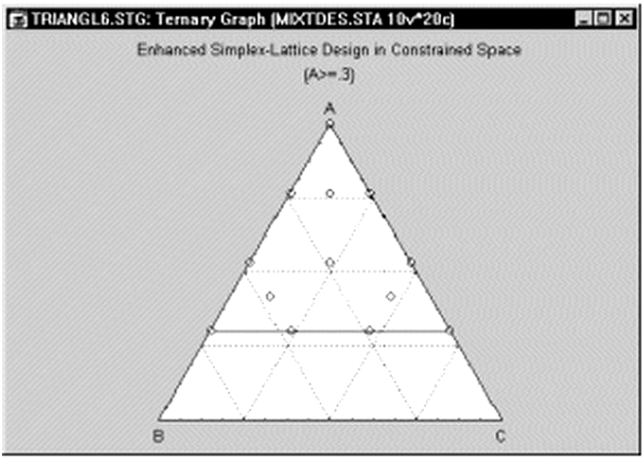


Рис. 12.3. Треугольная диаграмма (на компонент А наложено ограничение *x*A.3)

Если имеются ограничения снизу и сверху (что часто бывает в экспериментах на смесях), то стандартные симплекс-вершинные и симплекс-центроидные планы не могут быть построены, поскольку область, определяемая ограничениями, не является больше треугольником, и центры (вершины) могут не принадлежать области определения. Существует общий алгоритм нахождения точек-вершин и центроидов для таких планов с ограничениями (рис. 12.4).

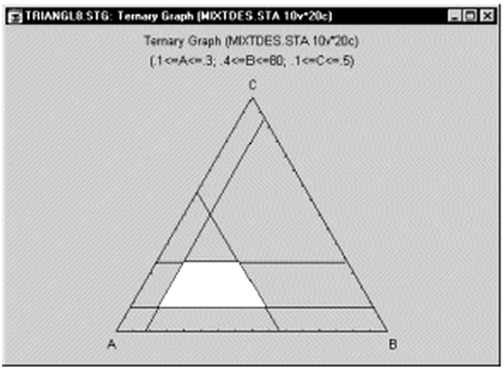


Рис. 12.4. Треугольная диаграмма алгоритма нахождения точек-вершин и центроидов для планов с ограничениями

12.3. Многофакторный отсеивающий план

В некоторых задачах имеется множество неизвестных факторов, поэтому на первом этапе надо отсеять заведомо незначимые факторы. Для решения этой задачи применяются планы Р. Л. Плакетта и Д. П. Бермана (с матрицей Адамара) и насыщенные дробные факторные планы с числом факторов до 127.

Если необходимо просеять большое число факторов, которые могут быть потенциально важными (т. е. связаны с изучаемой зависимой переменной), следует использовать план, который позволил бы тестировать наибольшее число главных эффектов при наименьшем числе наблюдений, т. е. построить план разрешения III с наименьшим числом наблюдений. Один из способов планирования такого эксперимента состоит в смешивании всех взаимодействий с новыми главными эффектами. Такие планы часто называют насыщенными, поскольку вся информация в них используется для оценки параметров, не оставляя степеней свободы для оценки эффекта (члена) ошибок ДА. Поскольку дополнительные факторы создаются приравниванием («присвоением псевдонимов») новых факторов к взаимодействиям в полной факторной модели, то эти планы всегда будут состоять из 2**k опытов (4, 8, 16, 32 и т. д.). Плакетт Р. Л. и Берман Д. П. показали, как полная факторная модель может быть разбита, чтобы получить насыщенные планы, в которых число опытов кратно 4, а не степени 2. Это позволяет оценить изменчивость случайных эффектов и тестировать оценки параметров на статистическую значимость.

Такие планы иногда называют планами с матрицей Адамара.

12.4. Методы Тагучи

Методы Тагучи значительно отличаются от традиционных процедур контроля качества и промышленного эксперимента. Особенно важными в них являются следующие понятия:

– функция потери качества;

– отношение сигнал/шум (С/Ш);

– ортогональные массивы.

Функции потери качества

Тагучи начинает с вопроса: что такое качество? Если новый автомобиль теряет скорость в центре перекрестка, подвергая участников движения риску, то принято говорить, что автомобиль не обладает высоким качеством. Понятие, противоположное качеству, более простое: это общие потери для индивида и для общества, обусловленные функциональной изменчивостью и неблагоприятными побочными эффектами, связанными с соответствующим продуктом. Следовательно, чем больше потери качества, тем ниже оно само.

Разрывная функция потерь. Чтобы сформулировать гипотезу об общем классе и форме функции потерь, достаточно предположить, что имеется особая идеальная точка высшего качества, например, автомобиль без каких-либо проблем, т. е. высокого качества. Обычно в статистическом контроле процессов принято определять уровень допуска вокруг номинальной идеальной точки производственного процесса. Согласно традиционной точке зрения, используемой в методах СКП, если находиться внутри допуска, то не возникнет проблем с качеством. Другими словами, внутри зоны допуска потери качества равны нулю. Если выйти за его пределы, то потери качества объявляются неприемлемыми. Так, согласно традиционной точке зрения, функция потерь качества является разрывной порогообразной функцией: если находиться внутри зоны допуска, то потери качества пренебрежимы, если выйти за пределы допуска, то потери станут неприемлемыми.

Квадратичная функция потерь. Существенной является разница между автомобилем, который безотказно работал в течение года после покупки, и автомобилем, у которого отмечались поломки. Если постепенные отклонения от номинала дают непропорциональное увеличение потерь, то, скорее всего, это квадратичные увеличения.

Если фактические потери качества являются квадратичной функцией отклонения от номинального значения, то цель прилагаемых усилий состоит в том, чтобы минимизировать квадрат отклонения

или дисперсию продукта относительно его номинальной (идеальной) спецификации, а не число единиц внутри границы допуска (как это принято в традиционных процедурах анализа процессов).

Отношения сигнал/шум (С/Ш)

Измерение потери качества. Даже если заключить, что функция потерь квадратична, то точно неизвестно, как измерять сами потери. Но любая выбранная мера должна отражать квадратичную природу функции.

Сигнал, шум и управляющие факторы. Продукт идеального качества всегда должен откликаться одинаковым образом на управляющие сигналы. Когда поворачивают ключ зажигания автомобиля, то ожидают, что стартер провернет двигатель, и он заведется. В автомобиле идеального качества процесс зажигания всегда происходит одним и тем же образом, например, после трех поворотов ключа зажигания двигатель заводится. Если в ответ на один и тот же сигнал – поворот ключа зажигания – наблюдается случайная изменчивость процесса, то можно говорить о качестве худшем, чем идеальное. Например, из-за таких неконтролируемых факторов, как низкая температура, влажность, изношенность двигателя автомобиль может иногда завестись только после 20 попыток и даже не завестись совсем. Этот пример иллюстрирует ключевой принцип измерения качества по Тагучи: минимизация изменчивости реакции продукта в ответ на факторы шума с максимизацией изменчивости в ответ на управляющие факторы.

Факторы шума – это факторы, которые находятся вне контроля оператора. В примере с автомобилем эти факторы включают колебания температуры, различия в качестве бензина, изношенность двигателя и т. д. Управляющие факторы – это факторы, которые устанавливаются или управляются оператором машины для ее использования по назначению (поворот ключа зажигания запускает двигатель, и автомобиль может начать движение). Итак, целью усилий по улучшению качества является установка наилучших значений управляющих факторов, которые включены в производственный процесс для того, чтобы максимизировать отношение С/Ш; поэтому факторы в эксперименте выступают как управляющие.

С/Ш-отношения. Таким образом, качество может быть рассмотрено с точки зрения отклика продукта на шумы и управляющие факторы. Идеальный продукт будет реагировать только на сигналы оператора, но не будет реагировать на случайный шум (погоду, температуру, влажность и т. д.). Следовательно, цель усилий по совершенствованию качества может рассматриваться как попытка максимизировать отношение С/Ш соответствующего продукта.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

Основной

1. Боровиков, В. П. Популярное введение в современный анализ данных и машинное обучение на STATISTICA / В. П. Боровиков. – М. : Горячая линия–Телеком, 2020. – 353 с.

2. Онлайн калькуляторы по статистике [Электронный ресурс]. – Режим доступа: https://www.math.semestr.ru/group/group_manual.php. – Дата доступа: 16.01.2022.

3. Основы статистики [Электронный ресурс]. – Режим доступа: https://www.stepik.org/course/76/promo. – Дата доступа: 16.01.2022.

4. StatSoft : электронный учебник по статистике [Электронный ресурс]. – Режим доступа: http://www.statsoft.ru/resources/STATISTICA_text_book.php. – Дата доступа: 07.01.2022.

Дополнительный

1. Лопатников, Л. И. Экономико-математический словарь / Л. И. Лопатников. – М. : Дело, 2003.– 427 с.

2. Российская геологическая энциклопедия: в 3 т. / Федеральное агентство по недропользованию (Роснедра), Рос. гос. геологоразведочный ун-т, Рос. акад. естественных наук, Ин-т геолого-экономических проблем; редкол.: Е. А. Козловский [и др.]. – М. : [б. и.]; СПб : ВСЕГЕИ, 2010–2012. – 230 с.

3. Социологический словарь / Акад. учеб.-науч. центр РАН–МГУ им. М. В. Ломоносова; отв. ред. Г. В. Осипов, Л. Н. Москвичев; учен. секр. О. Е. Чернощек. – М. : Норма, 2008. – 606 с.

4. Философский энциклопедический словарь / ред.-сост.: Е. Ф. Губский, Г. В. Кораблева, В. А. Лутченко. – М. : ИНФРА-М, 2007. – 575 с.

ДЛЯ ЗАМЕТОК

Учебное издание

Серебрякова Наталья Григорьевна,
Мириленко Андрей Петрович

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА И ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТА

Пособие

Ответственный за выпуск *Н. Г. Серебрякова*
Редактор *Т. В. Каркоцкая*
Компьютерная верстка *Д. А. Пекарского*
Дизайн обложки *Д. О. Бабаковой*

Подписано в печать 4.04.2022. Формат $60\times84^{1}/_{16}$.
Бумага офсетная. Ризография.
Усл. печ. л. 6,04. Уч.-изд. л. 4,73. Тираж 99 экз. Заказ 7.

Издатель и полиграфическое исполнение:
учреждение образования
«Белорусский государственный аграрный технический университет».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий
№ 1/359 от 09.06.2014.
№ 2/151 от 11.06.2014.
Пр-т Независимости, 99–1, 220023, Минск.