

МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«ГОМЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ»

Кафедра медицинской и биологической физики

А. А. КОВАЛЕВ, В. А. ИГНАТЕНКО

ОСНОВЫ СТАТИСТИКИ.

КОНСПЕКТ ЛЕКЦИЙ

Учебно-методическое пособие
для студентов 3–6 курсов всех факультетов
медицинских вузов

Гомель
ГоГМУ
2018

УДК 311 (072)

ББК 22.172я73

К 56

Рецензенты:

кандидат физико-математических наук,
доцент кафедры математических проблем управления и информатики
Гомельского государственного университета им. Ф. Скорины

В. Л. Мережа;

кандидат технических наук, доцент,
заведующий кафедрой автоматизированных систем обработки информации
Гомельского государственного университета им. Ф. Скорины

В. Д. Левчук

Ковалев, А. А.

К 56 Основы статистики. Конспект лекций: учеб.-метод. пособие для студентов 3–6 курсов всех факультетов медицинских вузов / А. А. Ковалев, В. А. Игнатенко. — Гомель: ГомГМУ, 2018. — 176 с.

ISBN 978-985-588-021-0

В учебно-методическом пособии в доступной форме излагаются теоретические основы статистического анализа данных. Изложение начинается с рассмотрения самых первых этапов по первичному анализу полученных в результате эксперимента данных и последовательно раскрывает более сложные методы анализа данных, такие как дискриминантный и кластерный анализы, логистическая регрессия. В последних главах рассмотрены некоторые вопросы по планированию исследования, а также даны рекомендации по выбору методики исследования и проведения эксперимента. Может быть использовано для формирования начального базового теоретического уровня при выполнении статистического анализа данных в процессе написания научных статей, диссертационных работ, экспериментальных исследований.

Предназначено для студентов 3–6 курсов всех факультетов медицинских вузов.

Утверждено и рекомендовано к изданию научно-методическим советом учреждения образования «Гомельский государственный медицинский университет» 27 октября 2017 г., протокол № 8.

УДК 311 (072)

ББК 22.172я73

ISBN 978-985-588-021-0

© Учреждение образования
«Гомельский государственный
медицинский университет», 2018

ОГЛАВЛЕНИЕ

ГЛАВА I. ВВЕДЕНИЕ В ПРЕДМЕТ	7
§ 1. Предмет математической статистики	7
1.1. Чай к молоку или молоко к чаю или немного истории	7
1.2. Некоторые этапы формирования и развития статистики как науки.....	9
1.3. «Считай все, что можешь считать!»	12
1.4. Современное видение статистики	15
§ 2. Статистика в системе образования других стран	16
§ 3. Математическая статистика на службе у медицины	18
3.1. Определение.....	18
3.2. Первые применения статистики в медицине.....	19
3.3. Введение доказательной медицины	20
3.4. Зачем изучать основы статистического анализа?	21
3.5. Типичные примеры медицинских исследований.....	22
3.6. Что должен знать врач о статистическом исследовании	26
§ 4. Программы для статистического анализа.....	27
ГЛАВА II. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ	30
§ 1. Несколько важных определений.....	30
§ 2. Какие бывают данные?	32
§ 3. Виды биомедицинской информации.....	34
§ 4. Генеральная совокупность и выборка. Свойства выборки	35
§ 5. Вариационный ряд	38
§ 6. Распределение значений признака. Полигон частот	40
§ 7. Относительная частота встречаемости	45
§ 8. Виды распределения	47
8.1. Нормальное (колоколообразное, гауссово) распределение	47
8.2. Площадь под кривой распределения.....	48
8.3. Равномерное распределение.....	49
8.4. Асимметричное распределение	50
8.5. Полимодальное распределение.....	51
ГЛАВА III. ОПИСАТЕЛЬНАЯ СТАТИСТИКА	52
§ 1. Основные параметры распределения	55
1.1. Среднее значение.....	55
1.2. Дисперсия и стандартное отклонение	55
1.3. Правило трех стандартных отклонений.....	59
1.4. Медиана, мода, перцентили	62

§ 2. Критерии проверки распределения случайных величин на соответствие нормальному распределению	66
2.1. Критерий Шапиро — Уилка.....	67
2.2. Критерий Колмогорова-Смирнова (λ (лямбда)).....	67
2.3. Критерий по асимметрии и эксцессу.....	67
§ 3.Arteфакт.....	69
§ 4. Основные выводы по описательным статистикам	71
ГЛАВА IV. СРАВНЕНИЕ ГРУПП.....	71
§ 1. Постановка задач	71
§ 2. Нулевая гипотеза	73
§ 3. Статистическая значимость результатов	75
3.1. Уровень значимости.....	75
3.2. Ошибки первого и второго рода	77
3.3. Критерий значимости.....	78
§ 4. Дисперсионный анализ	79
4.1. Критерий Фишера.....	79
4.2. Критическое значение F	80
§ 5. Критерий Стьюдента.....	82
5.1. Ошибки в использовании критерия Стьюдента.....	86
5.2. Правила использования критерия Стьюдента.....	87
§ 6. Непараметрические критерии	88
§ 7. Основные непараметрические критерии для сравнения групп	90
7.1. Сравнение двух выборок: критерий Манна — Уитни.....	90
7.2. Сравнение наблюдений «до и после»: критерий Уилкоксона	92
7.3. Сравнение нескольких групп: критерий Крускала — Уоллиса.....	94
7.4. Повторные измерения нескольких групп: критерий Фридмана....	95
7.5. Выводы по анализу сравнения групп	96
ГЛАВА V. АНАЛИЗ КАЧЕСТВЕННЫХ ПРИЗНАКОВ.....	97
§ 1. Общие понятия	97
§ 2. Статистические процедуры над качественными признаками	97
2.1. Первичный анализ	97
2.2. Сравнение двух групп с качественными признаками	99
2.3. Таблицы сопряженности: критерий χ^2 для таблицы 2×2	100
ГЛАВА VI. ПОНЯТИЕ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА	104
§ 1. Доверительный интервал.....	104
1.1. Определение.....	104
1.2. Пример использования доверительного интервала	105
1.3. Расчет ДИ	106

1.4. Степень надежности ДИ	106
1.5. Ширина ДИ	107
1.6. Область применения ДИ.....	107
ГЛАВА VII. АНАЛИЗ СВЯЗЕЙ	108
§ 1. Корреляционный и регрессионный анализ.....	108
1.1. Постановка проблемы.....	108
1.2. Виды связей между переменными. Функциональная связь.....	109
1.3. Корреляционная связь.....	110
1.4. Этапы анализа зависимости между переменным.....	113
1.5. Первичный анализ	114
1.6. Выбросы	114
1.7. Регрессионный анализ	115
1.8. Уравнение регрессии по выборочным данным	119
1.9. Доверительная область для линии регрессии.....	121
1.10. Сравнение двух линий регрессии	122
1.11. Корреляционный анализ. Общие положения	123
1.12. Коэффициент корреляции Пирсона.....	124
1.13. Криволинейная корреляция и регрессия.....	129
1.14. Итоговая модель	131
§ 2. Множественная регрессия. Логистическая регрессия.....	133
2.1. Множественная линейная регрессия	133
2.2. Логистическая регрессия	133
ГЛАВА VIII. КЛАССИФИКАЦИЯ: ДИСКРИМИНАНТНЫЙ И КЛАСТЕРНЫЙ АНАЛИЗЫ	136
§ 1. Кластерный анализ.....	138
1.1. Определение.....	138
1.2. Цели кластеризации	140
1.3. Этапы кластерного анализа	142
1.4. Методы кластеризации	142
1.5. Пример кластерного анализа.....	144
§ 2. Дискриминантный анализ	145
2.1. Определение.....	145
2.2. Задача дискриминантного анализа	147
2.3. Предположения и ограничения.....	148
ГЛАВА IX. ФАКТОРНЫЙ АНАЛИЗ.....	149
§ 1. Понятие факторного анализа	149
1.1. Определение.....	149
1.2. Задачи и возможности факторного анализа	149

1.3. Условия применения и алгоритм факторного анализа.....	151
1.4. Ошибки использования факторного анализа	151
ГЛАВА X. АНАЛИЗ ВЫЖИВАЕМОСТИ	152
1.1. Постановка задачи.....	152
1.2. Определение.....	152
1.3. Требования к анализу выживаемости	155
1.4. Кривая выживаемости.....	155
1.5. Сравнение двух кривых выживаемости	157
ГЛАВА XI. МЕТОДИКА ПОСТРОЕНИЯ ИССЛЕДОВАНИЯ	
В МЕДИЦИНЕ	160
§ 1. Основные цели и задачи доказательной медицины.....	160
§ 2. Этапы медицинского исследования	162
2.1. Формулировка цели исследования	162
2.2. Планирование исследования	163
2.3. Типы исследования	163
2.4. Выборка	165
2.5. Необходимые свойства выборки	166
2.6. Опасность множественных сравнений.....	167
2.7. Объем выборки	168
Заключение	172
Литература	175

ГЛАВА I. ВВЕДЕНИЕ В ПРЕДМЕТ

§ 1 . Предмет математической статистики

*Высшее назначение математики... состоит в том, чтобы находить скрытый порядок в хаосе, который нас окружает.
Норберт Винер*

1.1. Чай к молоку или молоко к чаю или немного истории

Знаете ли вы, что англичане пьют чай примерно шесть раз в день и даже шутят, что Англию проще представить без королевы, чем без чаепития. Так называемый «Five o'clock tea» (чай в 17:00) стал даже своеобразной английской традицией. И именно в Англии появился обычай пить чай с молоком. Конечно, это сочетание можно объяснить тем, что смешивая чай с молоком, англичане просто охлаждали чай, из-за которого могли лопаться хрупкие стенки фарфоровой посуды. Но традиция прижилась и по сей день, и крепкий чай с молоком часто называют «чай по-английски».

Как и многие традиции эта также стала обрастать целыми ритуалами и правилами, одним из которых является то, что добавлять надо чай к молоку и никак не наоборот! Да, да, именно чай! Доливание молока в чай, по мнению английских аристократических кругов, портит аромат и вкус напитка и даже рассматривается как невежество.

С этим необычным правилом этикета при чаепитии связана одна забавная история, случившаяся с молодым ученым, довольно серьезно интересовавшимся статистикой сэром Рональдом Фишером (рисунок 1), и послужившая огромным толчком к развитию методов современной науки.

Фишер отлично учился в школе и даже получил медаль за победу в математической олимпиаде. Закончив институт, он решил связать свою жизнь с наукой и поступил на работу в биологическую лабораторию. На тот момент в лаборатории подобрался вполне дружный коллектив, душой которого была молодая исследовательница, занимающаяся изучением водорослей Мюриэль Бристол, объединившая сотрудников под эгидой послеполуденных чаепитий, без которых не проходил ни один день. На чаепитиях ученые обменивались новыми результатами исследований и делились своими идеями с коллегами в непринужденной обстановке. Однажды чаепитие было несколько омрачено. Мюриэль попросила налить ей чай с молоком, что Рональд с удовольствием сделал. Однако девушка надменно отвернула носик и отказалась пить из вежливо протянутой ей чашки. Нет смысла говорить, что Рональд Фишер был удивлен, а когда Мюриэль пояснила причину своего поведения, он был охвачен даже возмущением. А дело оказалось в том, что Бристол происходила из аристократических кругов, в которых, как вам уже известно, считалось, что добавлять надо ис-

ключительно чай к молоку. Фишер же сделал наоборот, и, по мнению Мюриэль, это несоблюдение традиции церемонии выглядело как-то поплебейски. А сам Рональд испортил вкус чая, долив в него молока.

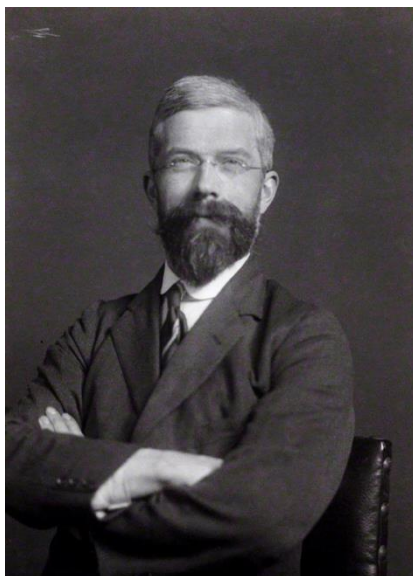


Рисунок 1 — Рональд Эйлмер Фишер

Молодой ученый не особо признавал устаревших предрассудков, а тем более связанных с сословным происхождением, да и ко всему он не понимал как Мюриэль, будучи ученым, не понимает, что вкус не меняется в зависимости от того что наливать первым? Он направил все усилия на то, чтобы доказать, что вкус чая в обоих случаях будет абсолютно одинаков, а поведение Мюриэль — пустой каприз. Однако Бристол не соглашалась и утверждала, что отлично может почувствовать разницу во вкусе.

В результате спора было заключено пари и решение провести слепой эксперимент, в ходе которого будет доказана правота одного из спорщиков. В соседней комнате при участии штатного химика приготовили разными способами несколько чашек чаю. Каково же было удивление Рональда Фишера, когда Мюриэль, несмотря на то, что во всех чашках пропорции чая и молока были одинаковы, действительно без труда определяла на вкус, в какой чашке чай разбавлен молоком, а в какой молоко чаем. По воспоминаниям одного из участников чаепития, Мюриэль правильно определила все чашки.

Фишер проиграл пари. Но проведенный эксперимент натолкнул его на новые мысли о статистических исследованиях, такие как: сколько раз надо повторить опыт, чтобы результат можно было считать достоверным? Если чашек было бы всего две, угадать метод приготовления вполне можно было и чисто случайно. Случайность могла играть немалую роль даже если чашек было и три и даже четыре. Эти размышления привели к публикации к одной из самых известных работ по статистике «Статистические методы

для научных работников», вышедшей в 1925 г. В ней были описаны основные принципы современной статистики, которыми биологи, медики и не только они пользуются по сей день.

1.2. Некоторые этапы формирования и развития статистики как науки

В вышеприведенном забавном происшествии прослеживаются многие важные вопросы, которые решает статистика, и которые дали толчок к ее дальнейшему развитию. Такие как:

- вопрос планирования эксперимента;
- определение размера выборки;
- понятие случайности событий;
- неизменное наличие вероятностной оценки результатов эксперимента;
- важность повторных измерений и многие другие.

Однако стоит заметить, что Рональд Фишер не является родоначальником статистики как науки, хотя датский историк статистических исследований Андрес Хальд (Anders Hald) и охарактеризовал его как «гения, едва не в одиночку заложившего основы современной статистики».

Не вдаваясь в серьезное исследование можно сказать, что примитивный анализ массовых явлений возник примерно ко времени возникновения государства. Изначально основной целью статистики являлось описание экономического и политического состояния государства или его части.



Первые статистические наблюдения в основном сводились к описанию экономического и политического состояния государства.

Считается, что первой опубликованной статистической информацией являются глиняные таблички Шумерского царства (а это в районе III–II тысячелетия до н. э.). Примером статистического учета также являются переписи населения в Древнем Китае, сравнение военного потенциала государств, проводимый учёт имущества граждан в Древнем Риме. Отсюда видно, что статистическое описание явлений есть необходимым условием функционирования сложных социальных систем, таких как государство.

В античном мире учитывали родившихся; в специальные списки вносились юноши, достигшие возраста военнообязанных (18 лет), а также 20 лет (возраст полноправных граждан). Составлялись земельные списки (кадастры), в которые включались сведения о строениях, рабах, скоте, инвентаре, доходах. Греческий философ Аристотель ещё за триста лет до нашей эры составил описание 157 городов государств.

Однако, если *сбор* статистических данных начался в самой глубокой древности, то их *обработка и анализ*, т. е. зарождение статистики как науки, относятся к более позднему периоду — второй половине XVII в. В это время и было введено в научный обиход само слово «**статистика**». Оно происходит от латинского *status* — состояние дел. И в науку термин «статистика» ввёл немецкий ученый **Готфрид Ахенваль** (рисунок 2) в 1746 г., предложив заменить название курса «*Государствоведение*», преподававшегося в университетах Германии, на «*Статистику*», положив тем самым начало развитию статистики как науки и учебной дисциплины.



Рисунок 2 — Готфрид Ахенваль

И уже к 1792 г. предмет статистики определяли как:



Описание состояния государства в настоящее время или в некоторый известный момент в прошлом.

Как видно из определения основным объектом изучения немецкой школы статистического анализа являлось государство с его производными и институтами, что, конечно же, является лишь частным случаем применения статистики.

Однако стоит заметить, что на 100 лет раньше немецкой описательной школы возникла школа **политических арифметиков**. Основоположниками которой были **Д. Граунт** (1620–1674), **Э. Галлей** (1656–1742) и **В. Петти** (1623–1687). В их трудах наметились два направления: **демографическое** с преобладанием вопросов страхования жизни и **статистико-экономическое**. Граунт на основе обработки бюллетеней о естественном движении населения Лондона впервые открыл некоторые *закономерности массовых общественных явлений и показал, как следует обрабатывать и анализировать большое количество первичного материала*. Он впервые попытался построить таблицу смертности для стационарного населения.



В это период был написан ряд научных работ по статистике, главные идеи которых состоят в стремлении конкретно оценить то или иное явление в условиях явной нехватки числовых данных.

Политические арифметики путем обобщения и анализа фактов стремились *цифрами охарактеризовать состояние и развитие общества, показать закономерности развития общественных явлений, проявляющиеся в большом количестве собранного материала*. Цели и задачи, которые ставили перед собой ученые, близки к современному пониманию сущности статистики.

Постепенно термин «статистика» стал использоваться довольно широко. К правлению **Наполеона Бонапарта** статистика определялась как «**бюджет вещей**».

И тем самым статистические методы были признаны полезными не только для административного управления, но и для применения на уровне отдельного предприятия.

Согласно формулировке 1833 г.:



Цель статистики заключается в представлении фактов в наиболее сжатой форме.

Во 2-й половине XIX – начале XX вв. сформировалась научная дисциплина — **математическая статистика**, являющаяся частью математики.

В этот период большой вклад в развитие дисциплины внесли такие ученые, как **Фрэнсис Гальтон** (1822–1911) и **Карл Пирсон** (1857–1936) (рисунок 3).

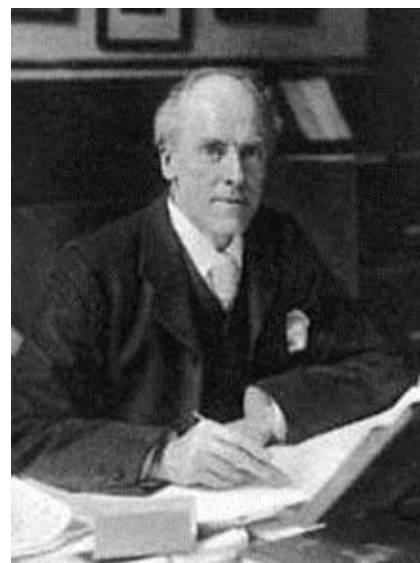
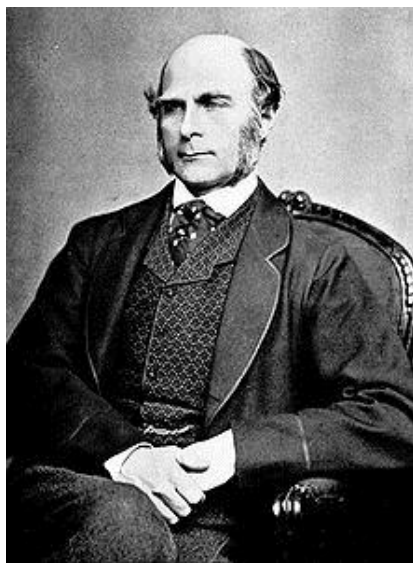


Рисунок 3 — Фрэнсис Гальтон и Карл Пирсон

1.3. «Считай всё, что можешь считать!»

В 1911 г. произошло знаменательное событие: из Лувра была украдена знаменитая картина «Джоконда» Леонардо да Винчи. Каких-либо значимых зацепок, указывавших на вора, обнаружено не было. И по сути поиски могли с легкостью зайти в тупик. Однако на стекле рамы от картины сохранились отпечатки пальцев, на тот момент метод расследования преступлений с использованием отпечатков пальцев был диковинной новинкой, но именно благодаря им и был идентифицирован и пойман злоумышленник. После этого, прогремевшего по всему миру успешного раскрытия преступления, метод *дактилоскопии*, предложенный выдающимся ученым Фрэнсисом Гальтоном, был окончательно признан, и дактилоскопия за несколько лет завоевала Европу, Америку, Японию, Россию. И это лишь одно из множества значимых нововведений, возникших благодаря уму и трудолюбию сэра Фрэнсиса Гальтона.

Родственник Чарлза Дарвина Фрэнсис Гальтон серьезно интересовался проблемой **наследственности**, к анализу которой он вскоре применил статистические методы. Он первым начал изучение однойцевых близнецов. И обнаружил, что некоторые человеческие признаки явственно передаются по наследству. Развивал учение о наследственной обусловленности индивидуально-психологических различий между людьми. Для определения наследуемости способностей он изучил генеалогию выдающихся людей и установил, что среди их родственников процент талантливых людей (со способностями выше среднего) превышал средний процент среди остального населения. Результаты своей работы Гальтон изложил в книге «Наследственный гений».

Обратившись к анализу отпечатков пальцев, он пришёл к выводу, что рисунок на кончиках пальцев у каждого человека индивидуален. Тем самым Ф. Гальтон заложил основу для новой науки.



Строго говоря, сам метод опознания преступников по их отпечаткам пальцев был разработан в 1860-х гг. Уильямом Гершелем в Индии, а его потенциальное использование в судебной практике было впервые предложено доктором Генри Фаулдсом в 1880 г., однако Гальтон впервые **доказал** именно уникальность каждого отдельного отпечатка пальца.

В 1892 г. Ф. Гальтон опубликовал в Лондоне книгу «Finger Prints» («Отпечатки пальцев»), в которой изложил результаты своих исследований. Поначалу судебные эксперты не оценили их значения, и, как уже было сказано выше, метод Ф. Гальтона начал широко применяться в Скотланд-Ярдом только в 1900 г. Но совсем скоро метод дактилоскопии стал переворотом в криминалистике. В 1905 г. отпечатки пальцев впервые были признаны в суде в качестве доказательства.

Также среди достижений Ф. Гальтона особо следует выделить разработку **метода тестов**. Тест (испытание) стал одной из основных методик в лаборатории Ф. Гальтона. Этот метод прочно вошёл в науку независимо от тех теоретических положений, которые соединял с ним сам автор. Статистический подход — применение серии тестов к большому числу исследуемых индивидов — **выдвигался как средство внедрения в психологию точных количественных методов**. Это получило большое практическое применение, а дальнейшее совершенствование техники разработки и применения тестов, распространение их на исследование важнейших психологических параметров (памяти, мышления, личности), по существу, изменили общий облик психологии, дав возможность сформировать систему объективных методов психологического исследования людей.



Сэр Фрэнсис Гальтон был очень разносторонним в своих интересах человеком, он интересовался как психологией, так и географией, занимался вопросами генетики и акустики, при этом он был очень изобретательным и наблюдательным человеком!

В 1906 г. проводилась ежегодная сельская ярмарка в Плимуте, Гальтон, интересовавшийся вопросами наследственности не мог не посетить ее.

Прогуливаясь по выставке, Гальтон наткнулся на стенд, около которого проводились соревнования по угадыванию веса. На всеобщее обозрение был выставлен откормленный бык, и собравшая толпа должна была на глазок определить вес животного. За шесть пенсов можно было приобрести билет, в который надо было внести имя, адрес и прогноз. За самые точные ответы были обещаны призы.

Счастье попытали около восьми сотен человек. Это была разношерстная публика — как мясники и фермеры, явно искушенные в оценке веса скота, так и люди, наверняка далекие от животноводства. «Участие приняли множество непрофессионалов, — писал впоследствии Гальтон в научном журнале *Nature*, — клерки и прочие из тех, кто, не имея специальных знаний о лошадях, делают ставки на бегах, опираясь на мнение газет, друзей или собственное разумение». Гальтону тут же пришла на ум аналогия с демократией, когда люди с радикально различающимися способностями и интересами получают каждый по одному голосу. «Средний участник конкурса был экипирован знаниями для точной оценки веса забитого и освежеванного быка не лучше, чем средний избиратель — для оценки качества того или иного претендента или особенностей большинства политических вопросов, по которым он голосует», — сетовал он.

Гальтону пришла в голову мысль установить, на что способен «средний избиратель», при этом он намеревался доказать, что его возможности

очень малы. Поэтому он превратил конкурс в импровизированный эксперимент. Когда соревнование закончилось и призы были розданы, Гальтон позаимствовал у его организаторов билеты и подверг их ряду статистических тестов. Гальтон рассортировал билеты с прогнозами (всего их оказалось 787 — ему пришлось исключить тринадцать билетов, ибо они были заполнены неразборчиво) в порядке убывания точности и выстроил график, чтобы удостовериться о правомочности использования средних значений. Затем он сложил все оценки участников и вывел усредненный прогноз группы. Эта цифра и представляла собой коллективную *«мудрость плимутской толпы»*. Если бы толпа была одним человеком, именно так этот человек оценил бы вес быка.

Ученый, несомненно, в силу своего мировоззрения и убежденности, полагал, что средний прогноз группы будет очень далек от истины. Казалось очевидным, что коллективное решение толпы, состоящей как из мудрецов, так и из людей посредственных и недалеких, скорее всего окажется неудачным. Но Гальтон ошибся. Толпа предположила, что вес быка, после того как его забьют и освежают, составит 547 кг. После того как его действительно забили и освеживали вес оказался — 543 кг. Иными словами, оценка толпы оказалась очень точной, коллективная ошибка составила менее 1 %. Подобная точность в оценке параметров или количества каких-либо предметов ещё не раз наблюдалась в экспериментах в разных странах, а феномену дали название **«мудрость толпы»**. (Из книги *«Мудрость толпы. Почему вместе мы умнее, чем поодиночке, и как коллективный разум влияет на бизнес, экономику, общество и государство»* Джеймс Шуровьески).

Гальтон оказал большое влияние на развитие математической статистики, разработав вместе **корреляционный** и **регрессионный анализ**. Лозунгом жизни ученого были слова: «Считай всё, что можешь считать!».

Его ученик **Карл Пирсон** также считается основателем математической статистики, и один из основоположников биометрики. Он опубликовал основополагающие труды по математической статистике (более 400 работ) и приложил много усилий для популяризации и внедрения своих результатов в математической статистике для применения их в других прикладных науках, прежде всего в биологии и медицине.

Приемником же Пирсона и наиболее влиятельным ученым XX в. в области статистики Запада является известный уже нам Р. Фишер. Который работал очень продуктивно с 1912 по 1962 гг., и многие его исследования оказали существенное воздействие на современную статистику и ее внедрение в научные исследования. С именами Пирсона и Фишера мы неоднократно будем сталкиваться на протяжении изучения курса.

1.4. Современное видение статистики

В XX в. статистику часто рассматривают, прежде всего, как самостоятельную научную дисциплину.

Подразумевая под ней совокупность методов и принципов, согласно которым проводится сбор, анализ, сравнение, представление и интерпретация числовых данных.

Она включает в себя:

- сбор сведений, характеризующих отдельные единицы каких-либо массовых совокупностей или явлений;
- выявление закономерностей, которые могут быть установлены на основе данных массового наблюдения или изучения группы объектов;
- разработка приёмов наблюдения и анализа данных.

Или другими словами и обобщенно:



Статистика — отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых явлений.

В современном мире, пронизанном разного рода информацией, результаты статистических анализов мелькают перед глазами человека ежедневно. Мы черпаем статистическую информацию из новостных каналов и колонок в Интернете; узнаем прогноз погоды, получение которого невозможно без статистических расчетов; для развлечения читаем результаты совсем несерьезных статистическим исследований, из которых мы узнаем о том, что чтобы покрыть обоями Великую китайскую стену потребуется порядка 15 840 000 рулонов обоев или что на голове блондинов (и блондинок) в среднем по 150 тыс. волос, а на голове брюнетов (и брюнеток) — по 100 тыс.; но в тоже время статистический анализ незаменим и в строгих серьезных и жизненно важных исследованиях, например, о том, как влияет метод трансплантации костного мозга на выживаемость пациента при остром лимфобластном лейкозе или возникает ли побочный эффект от приема того или иного медицинского препарата. Другими словами, статистический аппарат используется повсеместно и практически во всех областях современного человеческого знания. Без него невозможен качественный анализ полученной информации, а также невозможна грамотная организация и метод получения этой информации, чтобы можно было считать ее достоверной.

Также стоит сказать, что в определенном смысле своеобразный статистический учет присущ природе человека, он проявляется в свойстве человеческого разума упорядочивать поступающую информацию. Например, многообразие схожих явлений обязательно будет *классифицироваться* по

группам (классам) в зависимости от близости сходства между наблюдаемыми явлениями. Таким же способом человек может определять степень опасности в определенной ситуации: настораживаться при виде другого человека, внешне напоминающего образ неблагоприятного субъекта, сложившийся согласно общественному мнению или в результате личного опыта. К схожим результатам (но при изучении куда более сложных явлений) приходят *кластерный* и *дискриминантный* анализы в статистических исследованиях.

Другим примером может являться знание о том, как поведет себя тот или иной объект при определенном воздействии на него: силе нажатия, времени воздействия. Схожие исследования (но в виде строгих градаций) предполагают *регрессионный* и *корреляционный* анализы.



Как статистический анализ данных способствует упорядочиванию хаоса полученной информации о сложном объекте наблюдения, но использует при этом строгие математические методы, так и человек стремится упорядочить хаос поступающей к нему информации от внешнего мира, используя при этом данные природой возможности. Однако стоит заметить, что научные исследования предполагают стремление к объективности и определенной точности своих выводов и основаны не только на предположении, но и на доказательстве этих предположений.

Поэтому важно, так или иначе, понимать, как берутся статистические данные, каким результатам исследований можно доверять, а какие стоит воспринять как занимательную шутку, в чем состоит принцип статистического анализа полученных данных, и где могут быть допущены ошибки еще до начала анализа или даже эксперимента, когда полученные данные правдивы и значимы, а когда эксперимент стоило бы повторить для более детального и точного анализа, когда нужен анализ, а когда можно обойтись и без него и т. д.

§ 2 . Статистика в системе образования других стран

Говоря о развитии статистики как науки, нельзя не сказать, что во многих высокоразвитых странах изучение и применение статистического анализа поставлено на достаточно высокий уровень и зачастую знание основ является обязательным, так же основы статистики предлагаются к изучению еще в школе (старших классах), при этом материал преподносится достаточно простой, разнообразный и в основном так или иначе связанный с повседневностью.

Примером может служить образовательная манга «Занимательная статистика», автор Син Такахаси (рисунок 4).

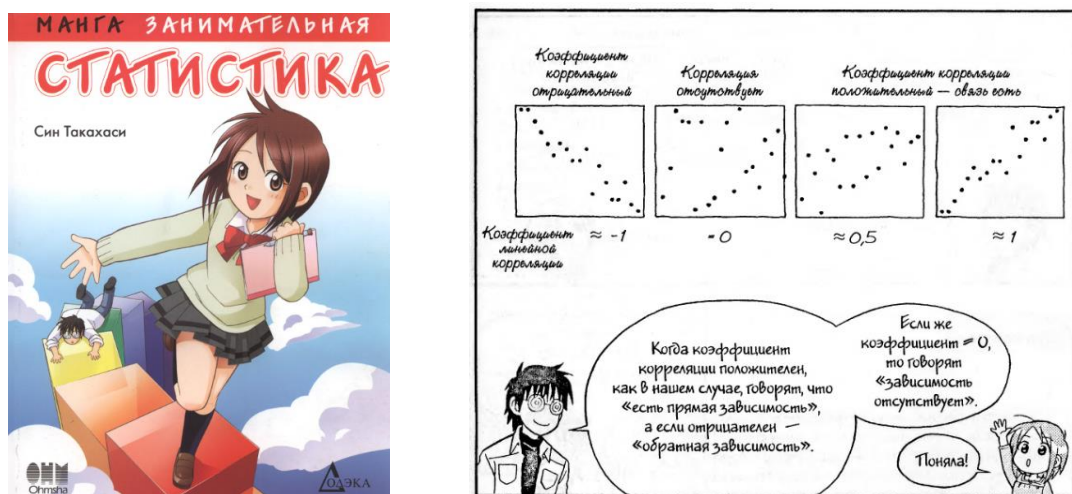


Рисунок 4 — Пример страниц книги «Занимательная статистика»

Или книга «Статистика для чайников», автор Дебора Рамси (рисунок 5).

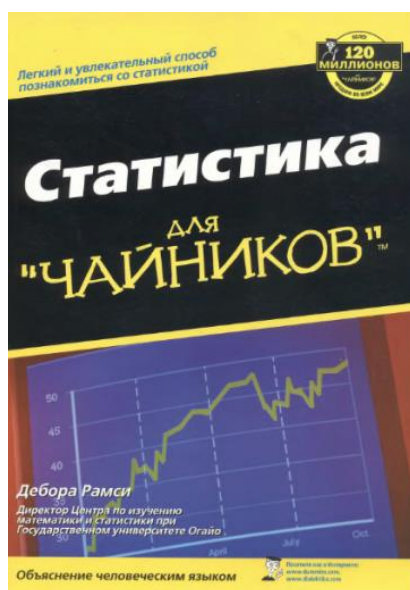


Рисунок 5 — Обложка книги «Статистика для чайников»

Данные публикации переведены на русский язык. Книги позволяют достаточно легко войти в дисциплину и предлагают определенный базис для понимания и дальнейшего изучения статистики.

Большим плюсом издания подобных учебников является демонстрация основных статистических методов наглядным и простым языком, не отягощенным большим количеством «заумных» определений и не перегруженным обильным и сложным математическим аппаратом, зачастую отпугивающим начинающих при изучении предмета.

Хорошим, но конечно же более сложным по содержанию, является учебник «Медико-биологическая статистика», автор Стентон Гланц. В котором на простых как вымышленных, так и реальных практических примерах рассмотрены основные наиболее часто используемые статистические методы, при этом автор ставит цель максимально доступно предоставить порой сложный для восприятия материал.

§ 3 . Математическая статистика на службе у медицины

3.1. Определение

Предметом изучения **математической статистики** является изучение свойств массовых явлений в экономике, биологии, медицине, технике и других областях. Чаще всего эти явления представляются **сложными**, вследствие разнообразия (варьирования) отдельных индивидуумов или единиц. И чтобы получить правильное представление об изучаемых свойствах массовых явлений и дать им определенные количественные оценки, их подвергают совместному рассмотрению и анализу. Математическая статистика используется во многих областях сферы человеческой деятельности и научно-практических интересов. Фокусом данного пособия является общий обзор методов статистического анализа, однако, учитывая профессиональные особенности, имеющиеся в сфере медицинских исследований, нельзя не сказать о **биомедицинской статистике** и сферах ее применения. Также в большинстве случаев в качестве примеров исследования будут приводиться именно медико-биологические исследования.



Биомедицинская информация — это сведения о свойствах биологических объектов и явлениях, являющихся предметами медицинских исследований, а также представления и суждения об этих свойствах и явлениях.



Биомедицинская статистика — это инструмент для анализа экспериментальных данных (биологического или медицинского исследования) и клинических наблюдений, а также язык, с помощью которого исследователь сообщает читателю полученные им результаты.

Центральной задачей математической статистики как метода исследования являются **заключения, выходящие за рамки изученного материала**, т. е. заключения о свойствах статистических совокупностей, принимая во внимание и неизученную их часть.

Метод изучения явлений, при котором приходят к обобщениям, изучив отдельные случаи этого явления, называется методом **индукции** (от частного — к общему).

Следовательно, математическая статистика использует **метод индуктивных заключений**



Исходя из этих обобщенных определений, можно определить предмет **медицинской статистики**. Предметом изучения медицинской статистики является информация, поступающая в ходе наблюдений, исследований, экспериментов так или иначе связанная с повседневной врачебной практикой. Обработка которой, позволяет делать определенные выводы о методах лечения, поведении объекта под действием того или иного препарата и т. п.

Статистическая обработка цифрового материала, полученного путем опытов, учетов и наблюдений при медико-биологических исследованиях, необходима для **проверки степени достоверности результатов и правильного их обобщения**.

3.2. Первые применения статистики в медицине

Известно, что в 17 в. врач и философ **Жан Баптист ван Хельмонт** предложил первое *клиническое испытание* с участием большого количества участников с целью оценить полезность практики кровопускания, при этом он уже тогда применил некоторые особенности исследования, которые только много позже оформились в важные критерии любого статистического значимого эксперимента. Предполагалось вовлечение в исследование 200–500 людей с разделением их на 2 группы путем жребия, где в одной группе кровопускание не проводилось бы, а в другой — его делали столько, сколько врачи считали нужным. При этом оценить эффективность кровопускания предполагалось по числу похорон в каждой группе, однако, к сожалению, в истории нет данных о реализации этого эксперимента.

И только в середине 19 в. в Париже в своих работах **Пьер Шарль Александр Луи** описал принципы статистического анализа для оценки медицинского лечения и доказал, что кровопускание является бесполезным видом лечения. Правда, стоит заметить, что это не изменило привычки врачей в то время и на протяжении последующих этапов истории человечества.

Принято считать, что первые применения статистики в медицине (хотя скорее этот пример ближе к социологии) относятся к 18 в., когда в Англии было замечено, что относительная частота смертности мужчин и женщин одного возраста, живущих примерно в одинаковых условиях, из года в год колеблется, но колеблется в весьма узких пределах.

Самым интересным здесь является замечание: **«колеблется в узких пределах»**. Всем известно, что колебания происходят, неожиданным являются узкие границы колебания, что позволяет с большой точностью

предсказать долю умерших в той или иной категории населения и послужило основой современного страхования.

Таким образом, в случайном явлении таком как смертность или, наоборот, выживаемость людей — была открыта **устойчивая закономерность**: относительная частота или доля для людей одного пола и близкого возраста примерно постоянна. Это открытие повлекло за собой множество событий, в частности возникновение современной теории страхования.

3.3. Введение доказательной медицины

Так сложилось, что вплоть до середины XX в. лечение мало влияло на сроки да и на сам факт выздоровления. Однако открытие и введение в клиническую практику инсулина, пенициллина, витамина В₁₂ изменило ситуацию очень радикально. Победа над ранее неизлечимыми болезнями породила веру в науку и стимулировала множество дальнейших исследований. В западных странах и США вызванный этой верой безграничный оптимизм породил почти столь же безграничное финансирование. На протяжении всего этого периода *врачи и исследователи получали в свое распоряжение практически неограниченные и зачастую не обусловленные конкретными целями ресурсы*. Характерной особенностью этого периода оказалось, что даже для по-настоящему действенных методов лечения отсутствуют достоверные оценки того, как часто и насколько эффективно они помогают. И тем более отсутствовали какие-либо значимые исследования о действенности новых методов лечения или препаратов.

Поэтому сложившийся подход часто не оправдывал вкладываемых средств. И чего хуже больные регулярно принимали сильно действующие препараты или подвергались врачебному вмешательству не только без серьезных оснований, но с риском возможных осложнений.

С течением времени сложившаяся ситуация несколько изменилась и столь огромное финансирование снизилось, поэтому медицинские работники и исследователи вынуждены взглянуть на используемые ими средства с точки зрения их **реальной отдачи**. Потребовались **строгие доказательства эффективности методов диагностики и лечения**. На данном этапе развития медицины приходится не только выяснить и проверить эффективность лечения, но и выяснить какому проценту больных оно помогает, и в какой степени. **А эти данные без помощи статистики и ее методов получить невозможно.**



На основании этого сформировалась так называемая **доказательная медицина**. Такой подход к медицинской практике, при котором решения о применении профилактических, диагностических и лечебных мероприятий принимаются исходя из имеющихся доказательств их эффективности и безопасности.

Естественная биологическая изменчивость, психотерапевтический эффект, субъективность оценок — все эти факторы делают **прямое суждение** об эффективности лечения ненадежным. Поэтому интуитивных и накопленных с течением времени опытных данных, не прошедших строгой анализ, на данном этапе развития доказательной медицины недостаточно.



Перевести клинический опыт на язык количественных оценок — задача **медицинской статистики**.

Стоит отметить, что статистическому анализу может быть подвергнута не только эффективность нового метода лечения, но и эффективность работы самого врача.

Так в одном исследовании было показано, что больные выписываются из стационара в среднем на 2 дня раньше, если их лечение проводилось в строгом соответствии с рекомендациями «Настольного справочника врача». Расходы на пребывание в стационаре составляют значительную часть всех медицинских расходов, поэтому сокращение сроков госпитализации (разумеется, не в ущерб больному) позволило бы сэкономить значительные средства.

3.4. Зачем изучать основы статистического анализа?

Как и перед исследователем, так и перед врачом, стремящимся расширить свои знания в профессиональной сфере, могут возникнуть следующие вопросы:

- Как найти устойчивые закономерности в опытных данных, т. е. в тех данных, которые лежат перед вами?
- Имеются ли существенные отличия между двумя методами лечения?
- Как воздействует новый препарат на определенную возрастную группу?
- Насколько эффективен новый метод лечения?
- Как доказать эффективность выбранного метода лечения, когда на результат воздействует такие факторы, как биологическая изменчивость пациентов, психотерапевтический фактор, субъективные оценки врача и т. д.?
- А также насколько достоверными можно считать исследования, опубликованные в журналах и статьях, можно ли доверять предоставленной информации?

О новых методах диагностики и лечения врачи узнают главным образом из публикации в медицинских журналах. Познания читателей в статистике обычно скромны, поэтому выводы авторов им приходится принимать на веру. Это было бы не так страшно, если бы публикации предшествовала серьезная проверка результатов. К сожалению, проводится она далеко не всегда.



Каждый из методов математической статистики имеет определенные возможности и свою ограниченную область применения. Использование метода, не соответствующего данному экспериментальному материалу, может привести к неверным обобщениям и необоснованным выводам.

Исторически сложилось так, что медики редко участвовали в обсуждении статистических вопросов, на первый взгляд далеких от врачебной практики и носящих скорее технический и теоретический характер. Однако поиск новых методов диагностики и лечения выбор наилучшего из уже принятых — везде статистические соображения играют далеко не последнюю роль. Чтобы принять полноправное участие в обсуждении этих вопросов, врач должен быть знаком с принципами и основными методами статистики.

Одной из целей данной работы является создание информационной основы для формирования у будущих врачей общепринятой базы статистических исследований, как элемента **доказательной медицины**.

3.5. Типичные примеры медицинских исследований

Рассмотрим некоторые типичные примеры применения статистических методов в медицине.

Галотан и морфин при операциях на открытом сердце¹

Галотан препарат, широко используемый при общей анестезии. Он обладает сильным действием, удобен в применении и очень надежен. Галотан — газ его можно вводить через респиратор. Поступая в организм через легкие, галотан действует быстро и кратковременно поэтому, регулируя подачу препарата можно оперативно управлять анестезией. Однако галотан имеет существенный недостаток, он угнетает сократимость миокарда и расширяет вены, что ведет к падению артериального давления. В связи с этим было предложено вместо галотана для общей анестезии применять морфин, который не снижает артериальное давление. Т. Конахан и соавт. сравнили галотановую и морфиновую анестезию у больных, подвергшихся операции на открытом сердце. В исследование включали больных, у которых не было противопоказаний ни к галотану, ни к морфину. Способ анестезии (галотан или морфин) выбирали случайным образом. Регистрировали следующие показатели параметры гемодинамики на разных этапах операции, длительность пребывания в реанимационном отделении и общую длительность пребывания в больнице после операции, а также послеоперационную летальность. Сосредоточим внимание на артериальном давлении между началом анестезии и началом операции. Именно в

¹ По книге С. Гланц «Медико-биологическая статистика».

этот период артериальное давление наиболее адекватно отражает гипотензивное действие анестетика, поскольку в дальнейшем начинает сказываться гипотензивный эффект самой операции. Артериальное давление между началом анестезии и началом операции измеряли многократно, каждый раз вычисляя среднее артериальное давление (данные заносились в таблицу) В исследование вошло 122 больных. У половины больных использовали галотан (1-я группа), у половины — морфин (2-я группа). Результаты представлены в таблице 1. Данные округлены до ближайшего четного числа.

Таблица 1 — Результаты исследования галотановой и морфиновой анестезии

Галотановая	Мофиновая	Галотановая	Мофиновая	Галотановая	Мофиновая	Галотановая	Мофиновая
42	42	60	62	68	74	74	86
44	46	60	64	68	74	76	86
46	46	60	64	68	76	78	86
46	52	60	64	68	76	78	88
48	56	62	66	70	76	78	88
48	58	62	66	70	78	80	90
50	58	62	66	70	80	80	90
50	58	62	66	70	80	82	92
52	58	62	68	70	80	82	96
54	60	64	68	72	82	84	96
54	60	64	68	72	82	86	98
56	60	66	68	72	82	90	98
58	62	66	70	72	84	94	98
58	62	66	70	74	84	98	100
58	62	66	70	74	84		
60	62	66	72	74	84		

На первый взгляд различия между двумя столбцами данных не значительны, Но результаты анализа показали, что разница статистически значима, поэтому можно заключить, что морфин в меньшей степени снижает артериальное давление, чем галотан.

Обнаружение карциномы мочевого пузыря

Результаты диагностического теста на предмет обнаружения карциномы (злокачественной опухоли) мочевого пузыря могут выглядеть следующим образом: подсчет количества (типизация) T-клеток и тест LAI. Результатами первого теста являются значения, выраженные числом, а тест LAI дает дихотомический результат: «положительно» или «отрицательно». С помощью статистического анализа возможно построить модель, которая на основании полученных анализов у пациента позволяет отнести его либо к группе больных, либо к группе здоровых.

Для этого необходимо собрать данные теста полученные в ходе работы со здоровыми людьми и заведомо больными пациентами. Результаты представлены в таблице 2.

Таблица 2 — Результаты эксперимента на обнаружение карциномы мочевого пузыря

Болен	Типизация t-клеток	LAI	Болен	Типизация t-клеток	LAI
да	48,50	положительно	да	73,50	положительно
да	55,50	положительно	нет	61,10	положительно
да	57,50	положительно	нет	62,50	отрицательно
да	58,50	положительно	нет	63,50	отрицательно
да	61,00	положительно	нет	64,50	положительно
да	61,50	положительно	нет	69,50	положительно
да	61,50	положительно	нет	70,00	отрицательно
да	62,00	положительно	нет	70,00	отрицательно
да	62,00	положительно	нет	71,00	положительно
да	62,00	положительно	нет	71,50	положительно
да	62,50	положительно	нет	71,50	отрицательно
да	63,00	положительно	нет	72,00	отрицательно
да	63,50	положительно	нет	73,00	отрицательно
да	65,00	положительно	нет	76,00	отрицательно
да	65,00	положительно	нет	72,50	отрицательно
да	66,50	отрицательно	нет	73,00	отрицательно
да	66,50	отрицательно	нет	73,50	отрицательно
да	66,50	положительно	нет	74,00	отрицательно
да	68,50	положительно	нет	75,00	отрицательно
да	69,00	отрицательно	нет	77,00	отрицательно
да	71,00	положительно	нет	77,00	отрицательно
да	71,00	положительно	нет	78,50	отрицательно
да	71,00	положительно			

На основании полученных данных при помощи методики логистической регрессии строится модель, определяющая вероятность отнесения нового пациента к категории болен или здоров. Другими словами, если посмотреть на результаты типизации T-клеток, то можно заметить, что здесь для здоровых людей значения в среднем выше, чем для больных, так же можно наблюдать и взаимосвязь с тестом LAI. Следовательно, исходя из значений, получившихся при типизации T-клеток и теста LAI, можно попытаться вывести вероятность наличия карциномы мочевого пузыря.

Эффективность диуретика²

Создатели нового препарата (диуретика) предполагают, что он увеличивает суточный диурез пропорционально принятой дозе. Для проверки этого предположения они назначают пяти добровольцам разные дозы препарата. По результатам наблюдений был построен **график зависимости**

² Вымышленный пример по книге С. Гланц «Медико-биологическая статистика».

диуреза от дозы. Зависимость видна невооруженным глазом (рисунок 6): при увеличении дозы увеличивается и суточный диурез. Исследователи поздравляют друг друга с открытием, а мир — с новым диуретиком.

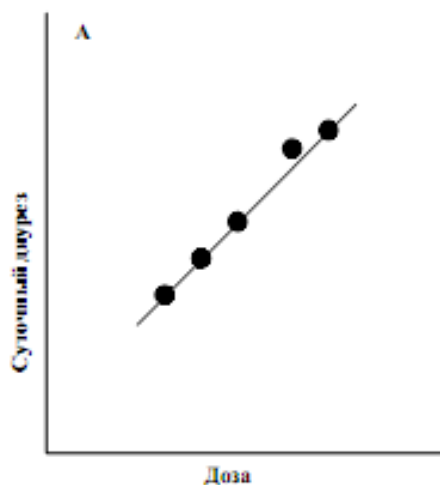


Рисунок 6 — Точечная диаграмма результатов измерения суточного диуреза по выборочным данным

На самом же деле данные позволяют достоверно утверждать лишь то, что зависимость диуреза от дозы наблюдалась у этих пяти добровольцев. То, что эта зависимость проявится у всех людей, которые будут принимать препарат, — не более чем предположение. Нельзя сказать, что оно беспочвенно — иначе, зачем ставить эксперименты?

И вот препарат поступил в продажу. Все больше людей принимают его в надежде увеличить свой диурез. И что же мы видим? Мы видим данные нанесенные на второй график, который свидетельствует об отсутствии какой-либо связи между дозой препарата и диурезом (рисунок 7).

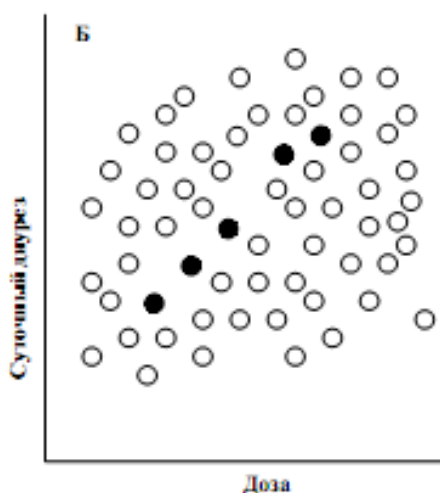


Рисунок 7 — Точечная диаграмма результатов измерения суточного диуреза, полученная в ходе исследования большого количества испытуемых

Черными кружками отмечены данные первоначального исследования. Статистика располагает методами, позволяющими оценить вероятность получения столь «непредставительной», более того, сбивающей с толку выборки.



Оказывается в отсутствие связи между диурезом и дозой препарата полученная «зависимость» наблюдалась бы примерно в 5 из 1000 экспериментов.

Поэтому, в данном случае исследователям просто **не повезло**. Если бы они применили даже самые совершенные статистические методы, это все равно не спасло бы их от ошибки.



Этот пример наглядно демонстрирует **вероятностный характер** выводов, получаемых в ходе статистического анализа. В результате применения статистического метода мы получаем **не истину в последней инстанции**, а всего лишь **оценку вероятности того или иного предположения**. Также каждый статистический метод основан на собственной математической модели и правильность результатов зависит от того насколько эта модель соответствует действительности. Т. е. от того, насколько выбранный метод математически соответствует полученным в ходе исследования данным.

3.6. Что должен знать врач о статистическом исследовании

- Как правильно сформулировать вопрос, на который вы хотите ответить с помощью статистического анализа?

Т. е. что я хочу узнать? Зачем?

- Необходимые статистические процедуры для адекватного ответа на вопрос.

Чем и как обработать полученные данные?

- Интерпретация результатов.

Что я получил?

И еще раз стоит отметить, что в результате применения статистического метода мы получаем **оценку вероятности того или иного предположения**. Имея дело с массовыми явлениями нельзя гарантировать правильность выводов и достоверность полученных результатов на 100 %, но можно значительно сократить вероятность ошибочного результата.

§ 4 . Программы для статистического анализа

Все программы статистической обработки данных можно разделить на профессиональные, полупрофессиональные (популярные) и специализированные. Статистические программы относятся к научному программному обеспечению, цена их часто недоступна индивидуальному пользователю. **Профессиональные** пакеты имеют большое количество методов анализа, в **популярных** пакетах количество функций достаточно для выполнения обычных задач. **Специализированные** же пакеты ориентированы на какую-либо узкую область анализа данных. Естественно, что создатели программных статистических пакетов заявляют, что их продукт превосходит аналоги. Определенную сложность может вызвать освоение более чем одной программы в связи с отсутствием у большинства исследователей времени на детальное изучение, что делает непростым выбор подходящего программного обеспечения.

MS Excel

Довольно популярным программным обеспечением является приложение MS Excel из пакета офисных программ компании Microsoft MS Office. Причины этого кроются в широком распространении этого программного обеспечения, наличии русскоязычной версии, тесной интеграцией с MS Word и PowerPoint.

Однако, MS Excel — это электронная таблица с достаточно мощными математическими возможностями, где некоторые статистические функции являются просто дополнительными встроенными формулами. Расчеты сделанные при ее помощи *не признаются* авторитетными биомедицинскими журналами. Также в MS Excel *невозможно построить качественные научные графики*. Безусловно, MS Excel хорошо подходит для накопления данных (для этих целей также хорошо подходит MS Access), промежуточного преобразования, предварительных статистических результатов, для построения некоторых видов диаграмм. Однако окончательный статистический анализ необходимо делать в программах, которые специально созданы для этих целей. Существует макрос-дополнение **XLSTAT-Pro** <http://www.xlstat.com> для MS Excel который, включает в себя более 50 статистических функций, включая анализ выживаемости, которых в основных случаях достаточно для обычного применения. А так же модуль *Анализ данных*.

R

Наиболее мощный бесплатный программный инструмент с огромным набором библиотек. В 2013 г. стал наиболее широко используемым программным продуктом для статистического анализа данных. С R может быть непривычно работать на начальном этапе, поскольку пользователь

должен разобраться в том, как подключать библиотеки, как выполнять и сохранять процедуры, а также из-за отсутствия привычного графического интерфейса. Но это вполне компенсируется большим количеством руководств по работе с программой, а графический интерфейс может быть установлен за дополнительную плату.

STADIA

Программа отечественной разработки с 16-и летней историей. Включает в себя все необходимые статистические функции. Она прекрасно справляется со своей задачей — статистическим анализом. Но! Программа внешне фактически не изменяется с 1996 г. Графики и диаграммы, построенные при помощи STADIA, выглядят в современных презентациях архаично. Цветовая гамма программы (красный шрифт на зеленом) очень утомляет в работе. К положительным качествам программы можно отнести русскоязычный интерфейс и наличие книг, описывающих работу. Например: Кулаичев А. П. Методы и средства анализа данных в среде Windows.

SPSS (Statistical Package for Social Science)

Самый часто используемый пакет статистической обработки данных с более чем 30 летней историей <http://www.spss.com> отличается гибкостью, мощностью, применим для всех видов статистических расчетов. Недавно вышла 13-я англоязычная версия. Существует русскоязычное представительство компании <http://www.spss.ru>, которое предлагает полностью русифицированную версию программы для Windows. Есть немало учебников на русском языке, позволяющих шаг за шагом освоить возможности SPSS, репетитор по статистике на русском языке, помогающий в выборе нужной статистической или графической процедуры для конкретных данных и задач, а также справка по SPSS Base и SPSS Tables.

STATISTICA

Производителем программы является фирма StatSoft Inc. (США) <http://www.statsoft.com>, которая выпускает статистические приложения, начиная с 1985 г. STATISTICA включает большое количество методов статистического анализа (более 250 встроенных функций), объединенных следующими специализированными статистическими модулями: *Основные статистики и таблицы, Непараметрическая статистика, Дисперсионный анализ, Множественная регрессия, Нелинейное оценивание, Анализ временных рядов и прогнозирование, Кластерный анализ, Факторный анализ, Дискриминантный функциональный анализ, Анализ длительностей жизни, Каноническая корреляция, Многомерное шкалирование, Моделирование структурными уравнениями и др.* Несложный в освоении этот статистический пакет может быть рекомендован для биомедицинских исследований любой сложности.

Российское представительство компании (<http://www.statsoft.ru/>) предлагает полностью русифицированную версию программы. Сайт компании содержит много информации по статистической обработке медицинских данных, учебник по статистике на русском языке. Сам пакет STATISTICA описан в нескольких книгах, одна из которых для медицинских работников: О. Ю. Реброва «*Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA*».

STATGRAPHICS PLUS

Довольно мощная статистическая программа. Содержит более 250 статистических функций, генерирует понятные, настраиваемые отчеты. Последнюю доступную версию можно получить на сайте <http://www.statgraphics.com>. Есть возможность скачать демо-версию. Следует отметить, что ранние версии этой программы были весьма популярны у отечественных исследователей.

PRISM

Эта программа создавалась специально для биомедицинских целей. Интуитивно понятный интерфейс позволяет в считанные минуты проанализировать данные и построить качественные графики. Программа содержит основные часто применяемые статистические функции, которых в большинстве исследований будет достаточно. Однако, как отмечают сами разработчики, программа не может полностью заменить серьезных статистических пакетов. На сайте <http://www.graphpad.com> помимо возможности ознакомления с демо-версией Prism можно получить справочник в формате PDF по биомедицинской статистике.

AtteStat

Программное обеспечение AtteStat предназначено для математико-статистического анализа данных и реализует большое число общеупотребительных и уникальных прикладных алгоритмов. Программное обеспечение AtteStat использует интерфейс электронных таблиц, например Excel. Основным достоинством данной программы является ее свободная лицензия (программа не является коммерческой), дружелюбный графический интерфейс, большое количество различных типов анализа данных, а так же интеграция с электронными таблицами и обширный справочник, описывающий методы и принцип работы с программой.

Интернет ресурсы

В настоящее время в Интернете доступны многие ресурсы, посвященные статистической обработке данных. Один из них — это статистический

портал, созданный при содействии В. П. Боровикова, автора книг по программному пакету STATISTICA <http://www.statsoft.ru/home/portal>. Российское представительство StatSoft Inc. предлагает на своем сайте бесплатный электронный учебник по статистике, который призван помочь разобраться с основными понятиями статистики и более полно представить диапазон применения статистических методов <http://www.statsoft.ru/home/download/textbook/default.htm>. На этом же сайте присутствует *Статистический медицинский советник*, который поможет правильно выбрать нужный статистический метод <http://www.statsoft.ru/home/portal/applications/medicine/medadvisor.htm>.

Из ресурсов Интернета заслуживает внимания сайт *Биометрика* <http://www.biometrica.tomsk.ru>.

Дороговизна программ не позволяет их часто менять. Поэтому имеет смысл посмотреть *демо-версии*, разобраться с работой и потом делать окончательный вывод.

Русскоязычные версии (с документацией) имеют только *SPSS* и *STATISTICA*.

ГЛАВА II. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Из предыдущей главы мы выяснили, что статистика занимается массовыми явлениями, при этом методы ее рассчитаны на то, чтобы по возможности, анализируя небольшие группы объектов, сделать вывод обо всех схожих объектах. Этот метод называется **индукцией**, т. е. от «частного к общему».

§1. Несколько важных определений

Объект исследования. Под **объектом исследования** подразумевается явление или процесс, который подвергается исследованию и анализу.



В качестве объекта исследования может выступать пациент, подопытное животное, группа людей, какие-то процессы организма, препарат, метод лечения и т. д.

Объект наблюдения — тот объект, который непосредственно участвует в ходе исследования. На основании данных, полученных в ходе наблюдения за этим объектом, делаются выводы об объекте исследования, исходя из цели исследования.

Признак³ — один из показателей, который измеряется у объекта наблюдения (исследования), интересующий исследователя как значимый в контексте конкретного исследования, выраженный числом или каким-то другим образом. Например, больной (объект наблюдения) может иметь следующие признаки: пол, возраст, рост, вес, давление, температура и т. д. Исследователь в ходе эксперимента выбирает из общего числа признаков один или несколько интересующих его и после сбора данных осуществляет анализ полученных результатов.

Переменная — в нашем изложении — синоним признака, но, стоит заметить, данный термин подчеркивает, что значение признака может варьироваться в зависимости от выбранного объекта наблюдения (единицы наблюдения) или в разных экспериментах. Например, измеряя показатели давления одного пациента, мы получили значение 120/80 мм рт. ст., а другого (из той же исследуемой группы) — 100/70 мм рт. ст. и т. д. Часто можно встретить понятие — **случайная переменная**, что в данном контексте то же самое, но лишний раз указывает на случайный характер значений изучаемого признака.

Значение переменной — конкретное значение признака. Т. е. значение интересующего исследователя признака, измеряемое у конкретного индивида (единичного представителя объекта наблюдения) из группы наблюдаемых участников эксперимента. Синонимом в данном контексте является термин **наблюдение**. Например, исследуется группа из тридцати участников, в которой измеряемым параметром является давление, в этом случае *случайная величина* (исследуемый параметр организма) — давление, а собранные данные по каждому испытуемому — *наблюдения*. Каждое отдельное значение давления — *наблюдение*. Количество наблюдений при полноценном сборе данных в этом случае будет равно количеству наблюдаемых — 30.

При исследовании реальных объектов исследователь всегда имеет дело со *случайными величинами*, другими словами, он не может точно предсказать значение измеряемого параметра⁴, можно говорить только о некой **вероятностной оценке** значений измеряемых параметров или о **вероятности наступления** некоего события.

³ В ходе изложения материала мы также иногда будем использовать слово параметр относительно какого-то исследуемого процесса (объекта наблюдения), что в данном контексте соответствует понятию «признак».

⁴ Например, выйдя на улицу и выбрав наугад первого встречного, мы лишь приблизительно можем оценить его возраст, да и то если имеем возможность его видеть, когда же данный тест проводится *вслепую*, то угадать точное и даже приблизительное значение возраста еще более затруднительно.



Важно отметить, что в силу этого все статистические методы носят вероятностный характер и выводы, которые делаются на основании методов статистики, в той или иной степени подразумевают возможность ошибки, вызванной случайным характером явлений живой природы. Однако статистические методы разработаны так, чтобы свести к минимуму вероятность ошибочного результата.

Можно сказать, что статистика анализирует массу случайных явлений и стремится делать определенные выводы об этих явлениях, обобщая их⁵.



Само собой и обобщение также носит некоторый вероятностный характер и при определенных условиях выводы, сделанные на основе данных небольшой группы, могут не распространяться за ее пределы (вспомните вымышленный пример с диуретиком).

Статистика призвана обрабатывать большое количество информации, поступающей в виде данных. И именно от типа данных зависит то, какими методами нужно воспользоваться для их обработки.

§2. Какие бывают данные?

Данные, полученные в ходе эксперимента могут быть **качественными, количественными и порядковыми**.



Для корректного использования статистических методов важно представлять, какого типа данные будут обрабатываться.

Количественные данные — признаки, которые можно выразить в числовой форме, грубо говоря, это те данные, которые можно измерить каким-то образом и выразить числом: возраст, вес, количество детей в семье и т. п. В свою очередь, они делятся на **непрерывные** и **дискретные**.

Непрерывные данные (continuous data) — такие количественные данные, которые могут принимать любое значение на непрерывной шкале. Другое название — признаки, измеряемые в интервальной шкале (температура, артериальное давление, рост). Примером может служить рост взрослого человека, он принимает любое значение в интервале от 150 см до 220 см и может выражаться в виде 178 см, 178,25 см, 182,33456 см, т. е. произвольное числовое значение на шкале в заданном интервале, при этом

⁵ Напомним, что данный метод называется **методом индукции** (от частного к общему).

точность так же может изменяться в зависимости от требований эксперимента и разрешающей способности измеряющего прибора.

Дискретные данные (discrete data) — количественные данные, принимающие, как правило, конечное число значений, хотя иногда и очень большое: количество смертей в течение года в исследуемой группе, количество пропущенных по болезни рабочих дней. Это такая переменная, которая может принимать значения только из некоторого списка определенных чисел. При этом дискретность может задаваться определенными правилами, например, персональный компьютер работает с дискретными значениями 0 и 1 и не может работать со значениями 0,34. Или количество отработанных дней может принимать значения 5, 10 или 28 или какое-то другое, но редко мы увидим запись, что человек отработал 7,28 дня. В этом примере видно, что дискретность данных задается условно, ведь если считать минуты, проведенные на работе, то значение все же может получиться выраженным и в сотых и даже тысячных. Поэтому деление данных на *дискретные* и *непрерывные* часто условное и зависит от специфики эксперимента.

Другое дело данные **качественные**, они радикально отличаются от данных количественных.

Качественные данные (классификационные, неупорядоченные, номинальные) — это признаки, которые нельзя выразить количественно, они представляют некое качество исследуемого объекта, например, диагноз, место проживания, пол, жив человек или мертв, есть температура или нет и т. п. Качественные данные можно выразить в количественной шкале, например, подсчитать количество людей с определенным диагнозом, но никак не наоборот.

Разновидностью качественных данных, выделенных в особую группу, являются данные **порядковые**.

Порядковые данные — показатели, измеряемые в шкале порядка (например, стадии болезни, оценка состояния больного — «плохо», «удовлетворительно», «хорошо»). При этом порядок изменить нельзя, только в обратном направлении, но смысл естественно от этого не меняется. Такие признаки могут быть осмысленно оцифрованы, при этом важно понимать, что порядок состояний имеет смысл. Часто к таким показателям следует относить балльные оценки, полученные при проведении тестов, анкетирования или экспертиз.



Особенность порядковых шкал — отсутствие количественного измерения расстояний между величинами на шкале (можно сказать, что течение болезни «хуже», чем среднетяжелое, при этом очень тяжелое «еще хуже», однако сложно сказать во сколько раз «хуже»).

Еще раз заметим, что для различных типов переменных применяются разные методы статистического анализа.

§ 3. Виды биомедицинской информации

В качестве небольшого отступления перечислим основные виды биомедицинской информации:

- **Первичная информация, используемая для получения изображения в медицинской диагностике.** Информация получается с использованием сложных диагностических способов, например рентгеновской автоматизированной томографии (АТ), ультразвуковой автоматизированной томографии и других методов. В процессе проведения обследования приёмник излучения (рентгеновского или ультразвукового) диагностических установок накапливает необходимые данные об исследуемом объекте, но для получения изображения с требуемым ракурсом, необходимо производить реорганизацию этих данных. Это требует большего числа вычислений, объём которых зависит от необходимой пространственной и яркостной разрешающей способности. В настоящее время для получения типовой рентгеновской томограммы требуется выполнить несколько сотен миллионов отдельных вычислительных операций. При этом обработка первичных данных должна происходить в реальном масштабе времени, т. е. с минимальной длительности процедуры от облучения пациента до получения результатов анализа.

- **Результаты индивидуального обследования отдельных пациентов в лечебных учреждениях** (поликлиника, клиника и т. д.). Это лабораторные исследования крови, мочи и др., общие рентгеновские обследования, ЭКГ и т. д. Данная информация необходима в комплексе для правильной и своевременной постановки диагноза и выбора метода лечения. Оперативное получение такой информации требует создания специализированных баз данных.

- **Информация о биологических объектах, полученная в результате исследования.** Например, исследования количества лейкоцитов в крови детей, уровень каких-либо вредных веществ в крови для различных физиологических условий и т. д.

- **Информация, получаемая в области биохимических исследований веществ, например, при синтезе новых лекарственных препаратов.** С помощью специализированных аппаратных и программных средств становится возможным детально изучать структуры сложных макромолекул и их химически активные связывающие участки и исследовать как пространственное взаимодействие рецепторов с химически активными участками потенциально полезных лекарств, так и динамику этих молекулярных взаимодействий.

Подумайте к какому типу данных можно отнести каждую группу из приведенных выше.

§4. Генеральная совокупность и выборка. Свойства выборки

Обычно исследователь в ходе эксперимента стремится сделать выводы обо всей **совокупности** объектов по какому-то интересующему его вопросу (например, какой эффект производит некий препарат на каждого человека с конкретной болезнью, на какие группы людей он оказывает побочное явление, а кому его вообще не стоит использовать).



В сущности, в этом и заключается смысл анализа: иметь представление о свойствах всего многообразия схожих по типу объектов исследования по тому или иному признаку исходя из изучения небольшого количества типовых объектов.

В ходе проведения эксперимента в распоряжение исследователя попадают значения каких-то измеряемых показателей. Обычно сами показатели называют **признаками**, а значения этих показателей — **вариантами** (или **наблюдениями**).



Например, при исследовании анестетика и его воздействии на артериальное давление во время операции, а также при анализе смертности в ходе использования данного препарата мы имеем следующее: артериальное давление — **признак**, выраженный в количественной шкале, смертность — **признак**, выраженный в качественной шкале, при этом люди, которым давался анестетик — **объект наблюдения**, с помощью которого исследователь сделает выводы о возможном влиянии данного анестетика на всех потенциальных пациентов, а сам анестетик можно назвать **предметом изучения** (или **объектом/предметом исследования**). Сами же числовые или номинальные значения признаков и есть **варианты** (наблюдения).

В статистике весь массив объектов, выводы на который распространяются по результатам исследования, обозначают как **генеральную совокупность**. И она обычно представляет собой очень большое число элементов, исследователь в силу различных факторов не может осуществить эксперимент над всеми элементами генеральной совокупности, поэтому он останавливается на достаточном количестве объектов, по возможности **характеризующим всю генеральную совокупность**. Это количество исследуемых объектов называются **выборкой** (обычно в формулах число элементов выборки обозначается как n , а число элементов генеральной совокупности — N).



Другими словами, **выборка** — это небольшое количество элементов из генеральной совокупности, но которых достаточно для формирования выводов обо всей совокупности.

Если это условие выполняется, то такую выборку называют **репрезентативной (представительной)**.



Выборка является **репрезентативной**, если с ее помощью можно охарактеризовать всю генеральную совокупность.

Репрезентативность — очень важное свойство выборки, если выборка не является репрезентативной, то исследователь может сделать ошибочные выводы обо всех объектах исследования (всей генеральной совокупности). Стоит заметить, что в медицинских исследованиях часто бывает так, что выборки имеют очень небольшой объем, порядка 10–20 элементов.

Обеспечение репрезентативности выборки — один из главных аспектов при планировании статистического исследования. При недостаточно качественном выполнении данного условия имеется большой шанс получить превратные представления об исследуемом объекте.



Классический пример нерепрезентативной выборки, произошедший в 1936 г. в США во время президентских выборов.

Известный журнал «Литэри дайджест», до этого весьма успешно прогнозировавший результаты предыдущих выборов, на этот раз ошибся в своих прогнозах, хотя и провел достаточно серьезную работу для обеспечения точности результата. Сотрудники журнала разослали несколько миллионов письменных вопросов подписчикам о том, кому отдаст предпочтение читатель «Литэри дайджест», а также ее респондентам, которых они выбрали из телефонных книг и из списков регистрации автомобилей (с целью расширить выборку и тем самым улучшить ее репрезентативность). Всего вернулось обратно 1/4 от общего числа разосланных бюллетеней, согласно которым голоса распределились следующим образом: 57 % отдали первенство кандидату от республиканцев по имени Альф Лэндон, а 43 % отдали предпочтение действующему президенту — демократу Франклину Рузвельту. И журнал с уверенностью сделал прогноз о том, что победит Лэндон.

На выборах же победил Ф. Рузвельт, который набрал почти 60 % голосов. В чем же заключалась ошибка прогноза? А она была в следующем: «Литэрари дайджест» захотели увеличить репрезентативность выборки. При этом они знали, что большинство их подписчиков относят себя к республиканцам, и поэтому решили расширить выборку за счёт респондентов, выбранных ими из телефонных книг и автомобильных регистрационных списков, т. е. предположительно абсолютно случайным образом. Но не учли существующих реалий и фактически отобрали ещё больше сторонников республиканцев, потому что во времена Великой депрессии иметь автомобили и телефоны мог позволить себе средний и высший класс. А это и были по большей части республиканцы, а не демократы.

Из этого примера видно насколько порой сложно реализовать правильное исследование. И что какая-либо неучтенная деталь может исказить все результаты до противоположности. Вспомните случай из жизни Роберта Фишера, описанный выше, и его размышления над тем, что имея две чашки с одинаковыми пропорциями молока и чая, но по-разному приготовленные, результат можно попросту угадать и сравните с приведенным примером и тем насколько скрупулёзно может быть выстроено исследование для того чтобы предсказать событие всего-то при двух возможных вариантах его исхода⁶. Классическим примером двух вариантов исхода событий является подбрасывание монетки (орел или решка), вероятность угадать что же выпадет составляет 50 %. Конечно условия даже для такого простого эксперимента можно довести до абсурда! Как свидетельствуют предположения некоторых, возможно не совсем серьезных или даже шуточных, теорий, если брать во внимание более сложные комбинации различных факторов, оказывающих влияние на ход эксперимента, то вероятность выпадения «орла» может оказаться выше, нежели «решки»⁷. Но это уже излишнее усложнение, в подробности которого особого смысла вда-

⁶ На самом деле ситуация с предсказанием результата выборов, строго говоря, не является просто случаем равновероятного возникновения события, как в примере с черными и белыми шарами в мешке (при условии, что мы имеем одинаковое количество черных и белых шаров, вероятность достать белый шар равна 50 %) и говорить о нем как о тривиальном не приходится. И сравнение с опытами Фишера не совсем корректно.

⁷ Эксперименты показали, что бросание монеты — процесс не вполне случайный; это измеряемое событие, подчиняющееся законам физики. Если каждая из монет ставится в одинаковые исходные условия и подвергается воздействию одинакового изначального усилия, ее вращение (спин) даст абсолютно равные шансы на выпадение орлом или решкой. Однако даже маломальская разница в условиях — скорость и угол вращения, расстояние от монеты до земли, какой стороной монета лежала вверх до броска — повлияет на результат. Стэнфордский эксперимент показал, что с усреднением по многим броскам такие изменения являются довольно существенными и могут помешать вероятности «50:50» (Из книги Стивена Фрая «Книга всеобщих заблуждений»).

ваться нет. Важно лишний раз уяснить себе, что для проведения грамотного статистического анализа или эксперимента необходимо стараться учитывать некоторые подводные камни, которые могут повлиять на выводы и результаты. И еще раз не забывать, что статистика является довольно точной наукой, но выводы, сделанные с ее помощью, носят все же *вероятностный* характер.

Однако вернемся к выборке и генеральной совокупности. Мы уже выяснили, что выборка должна уметь характеризовать всю совокупность, но помимо этого она должна обладать, по меньшей мере, еще одним важным свойством, которое скорее является необходимым для обеспечения ее репрезентативности.

Этим свойством является ее **случайный характер (рандомизация)**. Это означает, что каждый член генеральной совокупности равновероятно может попасть в выборку для проведения эксперимента.



Т. е. вероятность оказаться в выборке одинакова для всех членов генеральной совокупности.

Осуществить **рандомизацию** выборки необходимо с целью снижения возможной подтасовки результатов. Например, если исследуется воздействие препарата на артериальное давление и в генеральную совокупность входят люди разной возрастной группы, но с одинаковыми показаниями к препарату, стоит учитывать, что исследователь может выбрать людей помладше, тем самым улучшить показатели воздействия препарата, и подобный отбор уже не является случайным. Следовательно, выводы могут оказаться завышенными или заниженными, слишком оптимистичными или наоборот.

§5. Вариационный ряд

Двигаясь дальше необходимо остановиться на возможных вариантах отображения результатов эксперимента. Выше было показано, что для проведения эксперимента из всей генеральной совокупности выбирается небольшая группа объектов, которая может описать всю совокупность. Именно с этой группой и проводится эксперимент⁸. Исходя из цели исследования, снимаются необходимые данные. Здесь важно отметить, что данные нельзя перемешивать, т. е. нельзя представить таблицу, где в одной и

⁸ Строго говоря, эксперимент порой провести невозможно, и исследователь довольствуется только сбором данных имеющихся, например, в архивах, но это уже вопрос планирования исследования.

той же колонке содержатся показатели и давления, смертности, веса и роста исследуемых объектов⁹.

Данные по каждому исследуемому признаку (давление, рост, вес и т. п.) представляются в виде чисел (или если они качественные, то в виде номинальных значений признака), впоследствии сводимые в общую таблицу. Для большего удобства далее мы в основном будем рассматривать количественный тип данных.

Каждое отдельное число, полученное в ходе эксперимента, называется **вариантой**.

Варианта — это конкретное значение случайной переменной X_i (конкретное значение признака), т. е. величины, изменяющиеся под влиянием многих случайных причин. Или значение одного наблюдения.

Получив большое количество вариантов, их заносят в таблицы. Для более удобного отображения и возможности первичной оценки, полученных результатов данные стоит как-то упорядочить. Упорядоченное представление вариант (наблюдений) может быть следующим: **вариационный ряд, гистограмма распределения, кривая распределения, вариационная кривая**.

Вариационный ряд — это упорядоченная запись значений признака по отдельным особям изученной группы.



Другими словами, при измерении значений какого-то признака (например, температуры тела, роста или артериального давления) у разных членов исследуемой группы значение этого признака будут различными. Упорядоченная запись этих значений и называется **вариационным рядом**.

Пример:

В результате исследования группы людей на предмет влияния правильности метода лечения на сроки госпитализации (где переменной является число дней госпитализации) были получены следующие результаты (таблица 3):

Таблица 3 — Исходные данные эксперимента

Количество дней госпитализации	2,5	2	4	3	3	3,5	5	4,5	3,5	4	4	4,5	4	2,5	4,5	4,5	3,5	4,5	5,5	5	5	5,5	3
--------------------------------	-----	---	---	---	---	-----	---	-----	-----	---	---	-----	---	-----	-----	-----	-----	-----	-----	---	---	-----	---

Вариационный ряд в этом случае имеет вид (таблица 4):

Таблица 4 — Вариационный ряд по данным эксперимента

Количество дней госпитализации	2	2,5	2,5	3	3	3	3,5	3,5	3,5	4	4	4	4	4,5	4,5	4,5	4,5	4,5	5	5	5	5,5	5,5
--------------------------------	---	-----	-----	---	---	---	-----	-----	-----	---	---	---	---	-----	-----	-----	-----	-----	---	---	---	-----	-----

⁹ Это, казалось бы, банальное замечание, но на него все же стоит обратить внимание.

Как видно числа в вариационном ряду располагаются в определенном порядке: по возрастанию или убыванию, что позволяет даже без вычислений с приближением определить величину некоторых параметров (речь о которых пойдет ниже). Вариационный ряд уже нагляднее демонстрирует то, сколько раз та или иная переменная встречается в полученных данных эксперимента, какое значение минимальное, а какое максимальное, даже при беглом просмотре вариационного ряда специалист может сделать определенные выводы об исследуемой совокупности. В некоторых случаях внимательное рассмотрение вариационного ряда избавляет от необходимости расчета точных показателей.

§ 6. Распределение значений признака. Полигон частот

Вариационный ряд удобная форма представления, но она далеко не такая простая для анализа и внешне не очень информативная, так как записей в вариационном ряду может быть достаточно много, и, даже упорядочив, работать с ними довольно сложно. Поэтому имеются более удобные формы отображения.

Таблица частот — такая запись значения признака, в которой напротив каждого значения варианты записывается **частота ее встречаемости** в ходе эксперимента.

Под **частотой** подразумевается сколько раз то или иное значение появилось в выборке в ходе проведения эксперимента или сбора данных.

Пример:

В результате исследования был получен следующий вариационный ряд (таблица 5):

Таблица 5 — Данные эксперимента в виде вариационного ряда

Количество дней госпитализации	2	2,5	2,5	3	3	3	3,5	3,5	3,5	4	4	4	4	4,5	4,5	4,5	4,5	4,5	5	5	5	5,5	5,5
--------------------------------	---	-----	-----	---	---	---	-----	-----	-----	---	---	---	---	-----	-----	-----	-----	-----	---	---	---	-----	-----

Теперь запишем его в виде таблицы частот (таблица 6):

Таблица 6 — Таблица частот

Количество дней	Частота
2	1
2,5	2
3	3
3,5	3
4	4
4,5	5
5	3
5,5	2

Еще более наглядным представлением результатов эксперимента является **график распределения** (иногда его называют **графиком распределения частот**, **графическим представлением частоты встречаемости** или просто **распределением**).



График распределения — графическое представление (в виде столбчатой гистограммы или кривой линии) того, как часто (сколько раз) появляется в результатах эксперимента то или иное значение переменной.



Каждая генеральная совокупность характеризуется своим распределением значений исследуемой переменной (признака). Выборка также характеризуется распределением признака (**выборочное распределение**), и если она репрезентативна, то ее распределение такое же, как и у совокупности.

Для построения графика распределения на оси X (горизонтальной) отмечаются значения исследуемого признака (в нашем примере «Количество дней госпитализации»), по оси Y (вертикальной) — отмечается сколько раз то или иное значение появилось в ходе исследования.

Приведем пример построения графика распределения в виде столбчатой гистограммы.

Имея тот же вариационный ряд, составим таблицу частот и построим график распределения значений признака (рисунок 8):

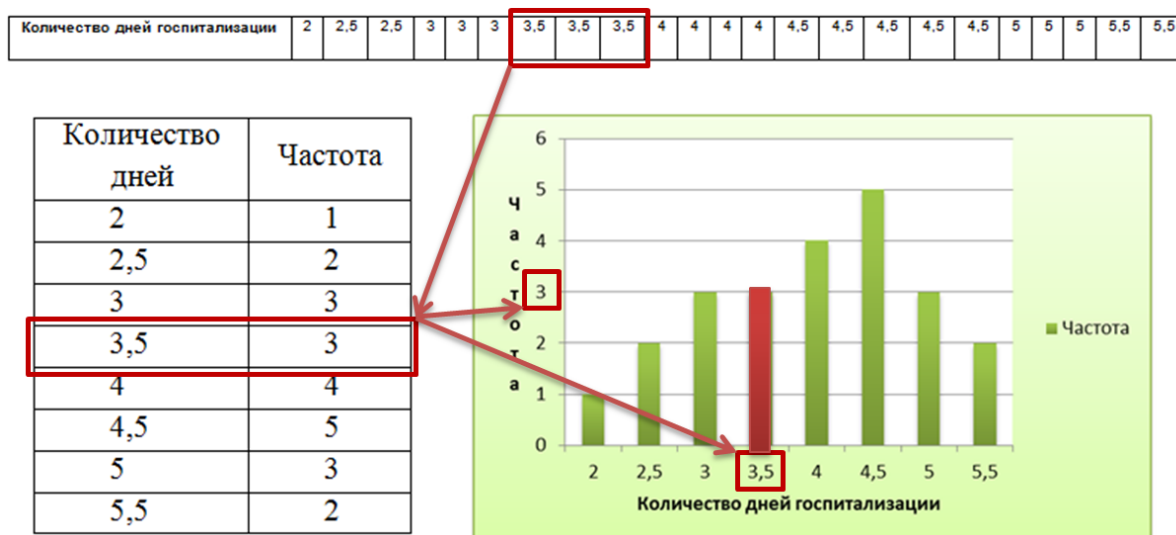


Рисунок 8 — Пример построения графика распределения признака

Другой пример графика распределения по тем же данным, построенный в программе Statistica, приведен ниже (рисунок 9).

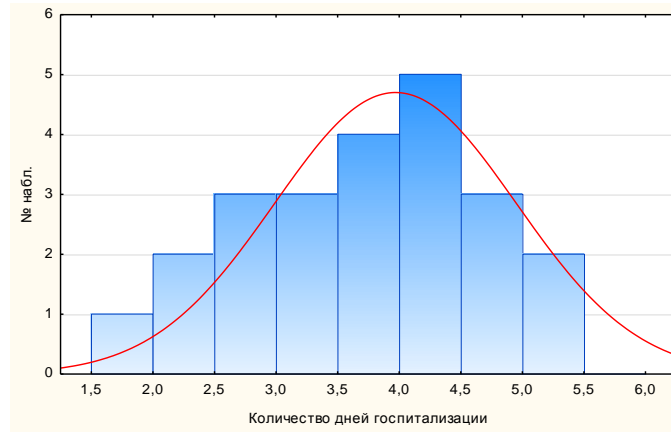


Рисунок 9 — График распределения значений признака с огибающей линией (подогнанная кривая распределения)

На втором графике видна огибающая линия. Довольно часто график распределения представляется именно в виде огибающей, но использование столбчатой диаграммы в условиях небольшого количества членов выборки все же более предпочтительно.

Столбчатую диаграмму чаще всего называют **гистограммой**, огибающую линию — **кривой распределения частот**. И ту и другую форму графического отображения обычно именуют просто *распределением*, под этим словом чаще всего подразумевают именно графическое представление.

Далее, если принять, что рассматриваемая случайная величина (признак исследуемого объекта) непрерывна, то увеличивая количество измерений и при этом уменьшая размер интервалов (карманов), мы можем получить следующие графики (рисунок 10):

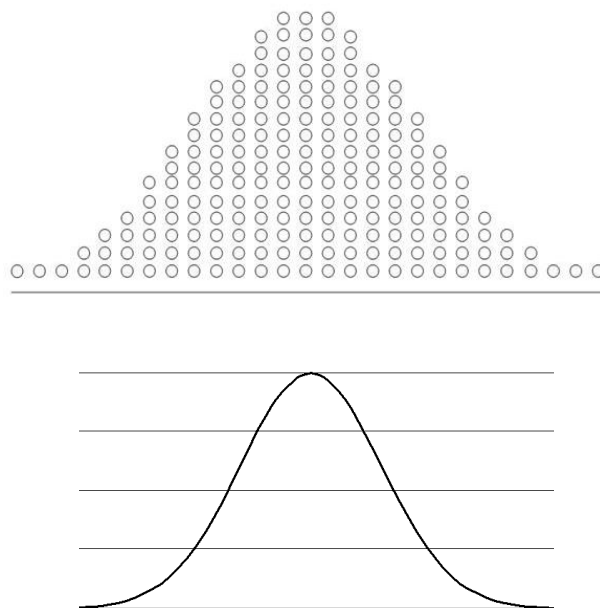


Рисунок 10 — Детально представленный пример графика распределения большой выборки («кружки» — каждое отдельное наблюдение) и получившаяся кривая распределения

Графическое представление распределения значений признака очень важно. С его помощью можно понять, с каким типом распределения мы имеем дело, какие значения признака встречаются наиболее часто, какое значение минимальное, а какое является максимальным. Это наиболее приемлемая форма представления результатов эксперимента на первичной стадии анализа.

В вышеприведенном примере (с количеством дней госпитализации) графики демонстрируют распределение дискретной случайной величины. Принцип построения такого графика для непрерывной случайной величины тот же, но в этом случае для получения столбца гистограммы необходимо учитывать, что непрерывная величина может принимать любое значение на заданном интервале. Для лучшего понимания представьте, что количество дней госпитализации из предыдущего примера измеряется не в дискретной, а в непрерывной шкале, тогда в ходе эксперимента мы могли бы получить значения такие, как: 2,6 дня, 2,8 дня, 3,65 дня и т. д. То есть вариационный ряд мог бы иметь следующий вид (таблица 7):

Таблица 7 — Данные эксперимента, представленные в виде непрерывной случайной величины

Количество дней госпитализации														
2,5	2,6	2,8	3	3,1	3,2	3,25	3,3	3,4	3,45	3,5	3,5	3,6	3,65	3,7
3,8	3,85	3,9	3,9	3,9	4	4,1	4,2	4,3	4,3	4,4	4,4	4,45	4,46	4,48
3,7	4,5	4,5	4,7	4,8	4,9	4,9	5	5,1	5,2	5,3	5,4	5,5		

Видно, что большинство значений в выборке встречаются только один раз. В этом случае, если подсчитать частоту встречаемости каждого значения, то таблица частот встречаемости выглядит несколько странно.

Изобразим фрагмент таблицы частот и часть гистограммы на ее основе (таблица 8, рисунок 11):

Таблица 8 — Фрагмент таблицы частот для таблицы 6

Количество дней госпитализации	Частота встречаемости	Количество дней госпитализации	Частота встречаемости	Количество дней госпитализации	Частота встречаемости
2,5	1	3,1	1	3,4	1
2,6	1	3,2	1	3,45	1
2,8	1	3,25	1	3,5	1
3	1	3,3	1		

Гистограмма имеет вид:

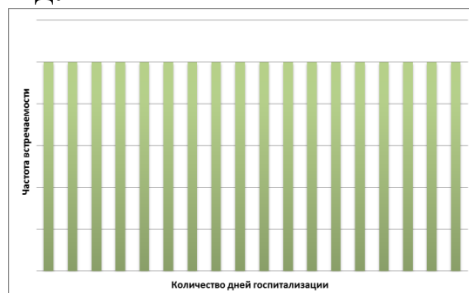


Рисунок 11 — График распределения частот без использования карманов

Другими словами, график стал очень малоинформативным. И судить о типе распределения не представляется возможным. Конечно, если мы соберем еще данных и увеличим выборку до 300 человек, то это может улучшить наше положение.

Такой подробный и в определенном смысле неестественный и тривиальный пример представлен здесь с целью знакомства с понятием **кармана**. Карманы используются для построения гистограммы распределения непрерывных значений признака.



Карман — область на горизонтальной оси X, в которую «помещаются» значения переменной, соответствующие заданному интервалу. В этом случае частота встречаемости того или иного значения переменной определяется как сумма частот переменных, соответствующих интервалу кармана. При необходимости интервалы карманов можно уменьшать или увеличивать. Чаще всего эти интервалы делаются одинаковыми, но это не является строгим требованием.

Применим карманы к нашему вариационному ряду. Зададим их следующим образом: от минимального значения равного 2,5, до максимального равного 5,5 с интервалом кармана 0,5. Мы получим следующие карманы: [2,5–3); [3–3,5); [3,5–4); [4–4,5); [4,5–5); [5–5,5].



Строго говоря, количество интервалов и их размер необходимо рассчитывать, исходя из размера выборки, но зачастую такая процедура не обязательна и карманы можно задать самостоятельно, тем более, что программные пакеты статистической обработки могут сделать это за вас.

Фрагмент вариационного ряда с обозначенными интервалами карманов [2,5–3); [3–3,5); [3,5–4) представлен ниже (таблица 9):

Таблица 9 — Вариационный ряд с обозначенными интервалами карманов

Количество дней госпитализации		Количество дней госпитализации		Количество дней госпитализации
2,5	}	3,1	}	3,5
2,6		3,2		3,6
2,8		3,25		3,65
3		3,3		3,7
		3,4		3,7
	3,45	3,8	}	3,85
	3,5	3,9		3,9
		3,9		3,9
		3,9		3,9
		3,9		4

Тогда таблица частот примет вид (таблица 10):
Таблица 10 — Таблица частот по данным таблицы 8

Количество дней госпитализации (интервалы карманов)	Частота
2,5	1
3	3
3,5	8
4	10
4,5	11
5	5
5,5	5

И соответственно гистограмма распределения (рисунок 12):

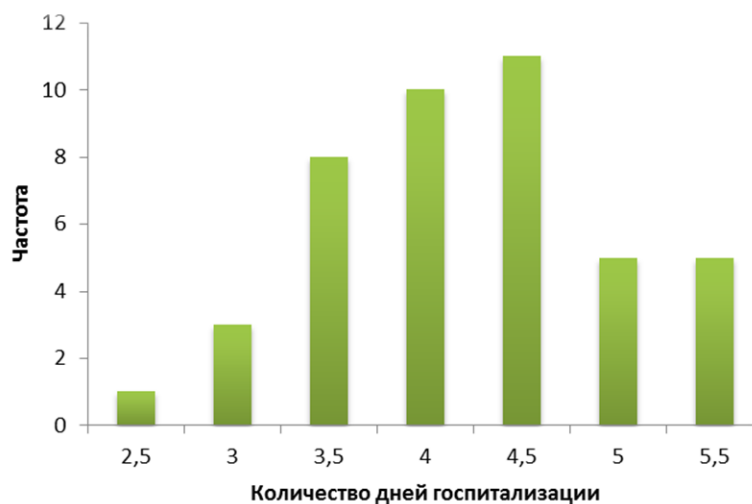


Рисунок 12 — Гистограмма распределения частот по данным таблицы частот

Как видите, график похож на тот, что был построен при использовании дискретных данных.

§7. Относительная частота встречаемости

Выше мы рассматривали так называемую **абсолютную частоту встречаемости** значения в выборке. Однако довольно часто вместо абсолютной частоты встречаемости на графике изображают **относительную частоту встречаемости**, выражаемую в долях или процентном соотношении.



Абсолютная частота — это величина, определяющая сколько раз какое-либо конкретное значение появляется в определенном наборе данных. **Относительная частота** — это доля того или другого значения признака в выборке, этот показатель довольно часто выражается в процентах.

Относительная частота встречаемости конкретного члена выборки (или генеральной совокупности) объемом N определяется следующим образом:

$$f = M/N,$$

где M — количество элементов выборки с заданным конкретным значением.

Из выше приведенного примера рассчитаем относительную частоту встречаемости дней госпитализации со значением 4,5. В этом случае абсолютная частота равна 5.

$$f = 5 / (1 + 2 + 3 + 3 + 4 + 5 + 3 + 2) = 5 / 23 = 0,2174.$$

Относительная частота встречаемости по количеству дней госпитализации со значением 4,5 дня равна 0,22, если это значение выразить в процентах, то получается 22 %.

То есть 22 % от всех участников эксперимента были выписаны спустя 4,5 суток после начала лечения

Подсчитав все относительные частоты можно получить (таблица 11):

Таблица 11 — Пример таблицы с рассчитанными относительными частотами

Количество дней	Частота встречаемости	Относительная частота встречаемости	Относительная частота встречаемости (%)
2	1	0,0435	4,35 %
2,5	2	0,0870	8,70 %
3	3	0,1304	13,04 %
3,5	3	0,1304	13,04 %
4	4	0,1739	17,39 %
4,5	5	0,2174	21,74 %
5	3	0,1304	13,04 %
5,5	2	0,0870	8,70 %
Сумма:	23	1	100 %

Построим гистограмму (рисунок 13):



Рисунок 13 — Гистограмма на основании данных об относительных частотах

Смысл использования относительных частот встречаемости заключается в том, что довольно часто необходимо выразить количество членов выборки с разными значениями исследуемого признака в процентном соотношении, указать возможное соотношение значений признака уже в генеральной совокупности, при условии, что выборка репрезентативна.

Также обратите внимание, что сумма относительных частот равна 1, а их процентного соотношения соответственно 100 %.

Забегая вперед, следует сказать, что площадь под кривой распределения всегда равна 1 (естественно, если при этом используется выражение частоты встречаемости признака в виде относительной частоты встречаемости).

Также: площадь ограниченной области под кривой распределения равна доле и вероятности возникновения признака с заданными значениями. Эта особенность будет рассмотрена ниже.

§8. Виды распределения

Рассмотрим внешний вид основных типов распределения значений признака. Здесь мы не будем рассматривать все типы распределения – их довольно много — остановимся на наиболее часто встречающихся.

В большинстве случаев в медико-биологических исследованиях встречаются следующие виды распределения:

- Нормальное.
- Асимметричное.
- Равномерное.
- Полимодальное.

8.1. Нормальное (колоколообразное, гауссово) распределение (рисунок 14, 15)

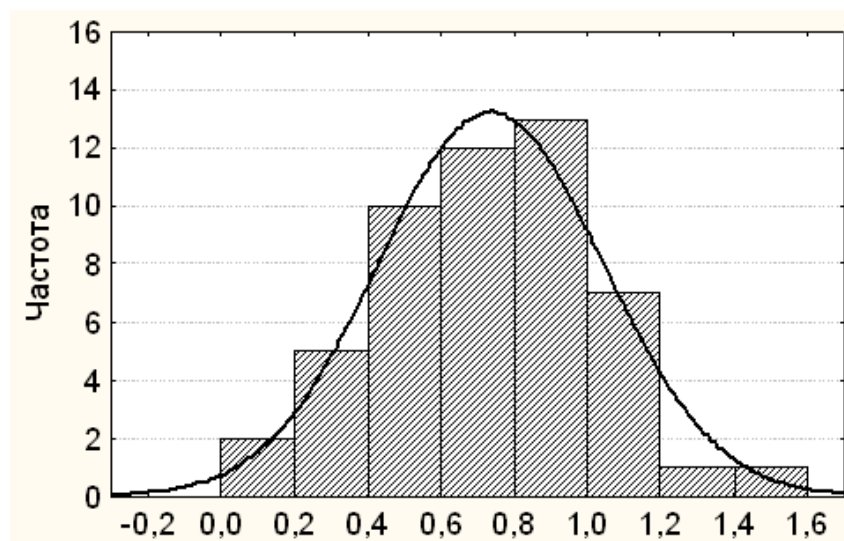


Рисунок 14 — Внешний вид нормального распределения (столбчатая гистограмма)

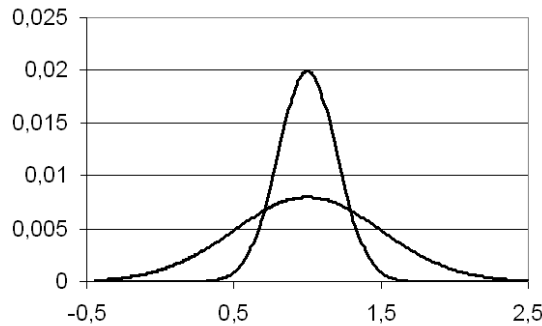


Рисунок 15 — Графики нормального распределения с разными значениями отклонения относительно среднего значения

При нормальном распределении наиболее часто в выборке встречаются значения близкие по величине к среднему по выборке и располагающиеся симметрично ему (значений больше среднего и значений меньше среднего приблизительно одинаковое количество). Или если выразить в процентном соотношении (используя относительные частоты встречаемости), то можно говорить, что наибольший процент значений признака находится в районе среднего значения, тогда как всего несколько процентов — по краям кривой.



Нормальное распределение подразумевает, что большая часть значений признака находится в районе так называемого среднего значения.

Нормальное распределение встречается очень часто, его основные свойства и особенности будут раскрываться ниже.

8.2. Площадь под кривой распределения

При изучении распределений как теоретической базы статистических заключений наибольший интерес представляет площадь под кривой нормального распределения. Эту площадь можно представить как интеграл от функции $f(x)$.

Как было сказано выше, площадь под кривой распределения всегда равна 1 (при выражении частоты встречаемости в виде относительных значений), а площадь ограниченная какими-то значениями признака соответствует вероятности или доле (рисунок 16).

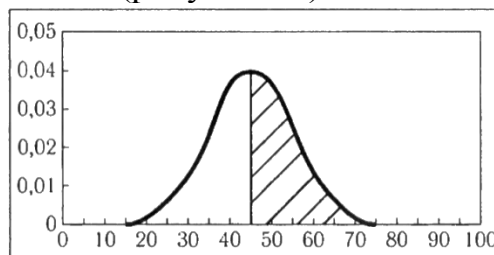


Рисунок 16 — График нормального распределения с разделенной пополам областью под кривой

На графике изображено распределение случайной величины. Оно соответствует нормальному распределению, если разделить область под кривой пополам, то обе половины будут равной площади (0,5), отсюда можно говорить, что вероятность возникновения значений признака больших 45 (согласно графику) равна 0,5, следовательно, доля членов выборки со значением больше 45 также равна 0,5 (т. е. половине от всех членов выборки).

Если же мы захотим узнать какая вероятность возникновения признака со значениями больше 63, то изобразив это на графике (рисунок 17):

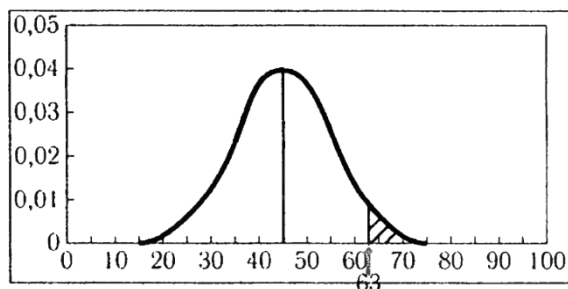


Рисунок 17 — График нормального распределения с заштрихованной областью под кривой выше значения 63

Видно, что доля таких членов выборки существенно меньше и вычислив площадь под кривой получим около 3,5 %, соответственно меньше 63 равна $100 \% - 3,5 \% = 96,5 \%$

8.3. Равномерное распределение

Внешний вид равномерного распределения в виде непрерывной кривой и столбчатой диаграммы представлен ниже (рисунки 18, 19).



Рисунок 18 — Внешний вид равномерного распределения

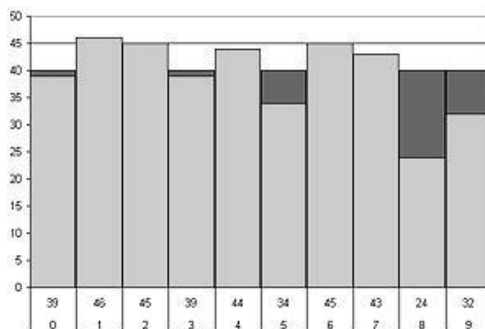


Рисунок 19 — Равномерное распределение значений признака в виде столбчатой гистограммы

При равномерном распределении все значения признака в выборке встречаются приблизительно одинаково часто или имеют примерно равное значение относительной частоты встречаемости.

8.4. Асимметричное распределение

Асимметричное распределение (если асимметрия левосторонняя — **логнормальное распределение**) имеет вид (рисунок 20):

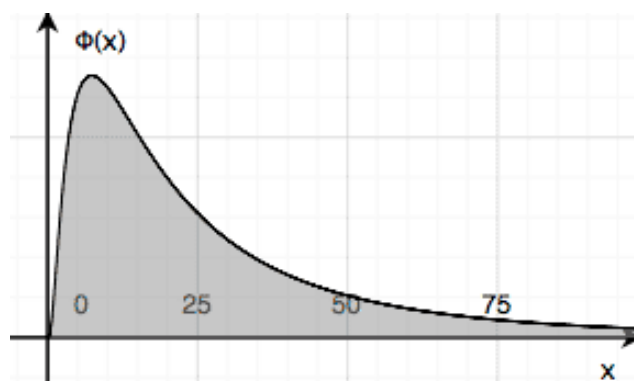


Рисунок 20 — Внешний вид асимметричного распределения

Оно характеризуется сдвигом наибольшего количества встречаемых значений в сторону относительно среднего значения и как следует из названия не является симметричным.



Если функцию $f(x)$ логнормального распределения преобразовать на ее десятичный логарифм $Lg f(x)$, то в этом случае полученная функция будет иметь нормальное распределение и характеризоваться теми же параметрами.

Используя графическое представление такой случай можно продемонстрировать следующим образом (рисунок 21):

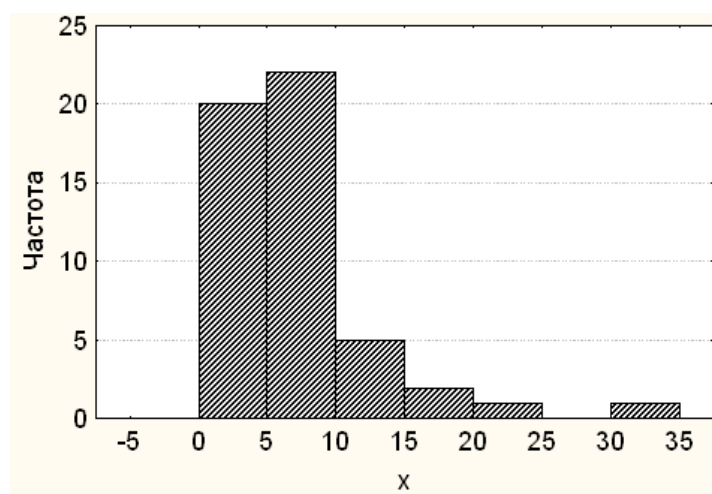


Рисунок 22 — Асимметричное распределение в виде столбчатой гистограммы

Если рассчитать логарифм десятичный от x и построить распределение получившихся значений, то мы получим следующий график (рисунок 22):

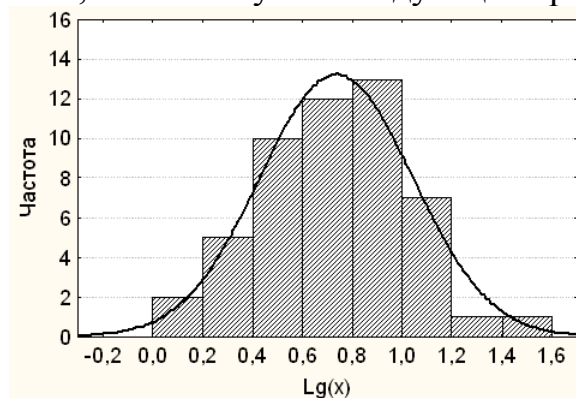


Рисунок 22 — Распределение $\lg(x)$:
 x — значения признака, имеющие логнормальное распределение

График соответствует нормальному распределению.

8.5. Полимодальное распределение

Полимодальное распределение может иметь вид соответствующий двум и более совмещенным распределениям (например, нормальным), каждое из которых можно охарактеризовать отдельно.

Полимодальное распределение имеет вид (рисунок 23):

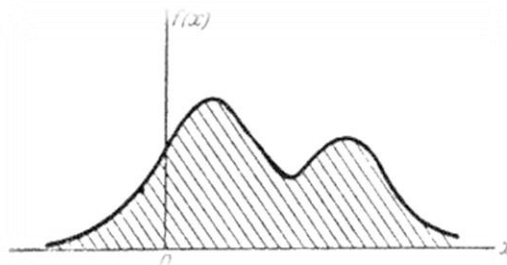


Рисунок 23 — Внешний вид двумодального распределения

С использованием столбчатой гистограммы и огибающей кривой (рисунок 24):

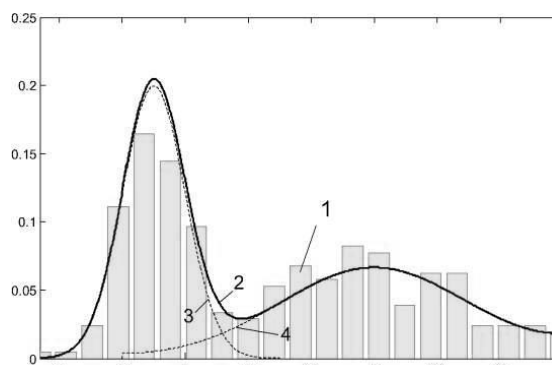


Рисунок 24 — Пример двумодального распределения: 1 — гистограмма; 2 — кривая распределения; 3 — график 1-го нормального распределения; 4 — график второго нормального распределения

На графике можно увидеть примеры двумодального распределения, у которого часть данных распределена по нормальному закону распределения с одними параметрами и часть данных имеет тот же закон распределения, но с другими параметрами.

Полимодальное распределение может быть обусловлено действием нескольких скрытых факторов. Или оно возможно из-за неправильного построения исследования, например, выборка не является достаточно репрезентативной.



В зависимости от типа распределения выбираются методы статистического анализа. Если распределение является нормальным или логнормальным, то применяют методы так называемой, **параметрической статистики**.

ГЛАВА III. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Мы выяснили, что для формирования выводов при помощи методов статистического анализа необходимо иметь достаточно большое количество исследуемого материала. Насколько большей должна быть выборка исследователь должен решить еще до начала эксперимента. Также, как было показано в предыдущей главе, полученные данные, так или иначе требуется упорядочить для их более наглядного отображения и нанести полученный результат на график, получив гистограмму распределения частот.

Зачастую эти данные представлены достаточно большим массивом чисел, показателей или другими возможными значениями проявления признака.

Например, при исследовании влияния анестетика на падение артериального давления при операции на открытом сердце исследователь получает таблицу результатов, в которой перечислены значения давления у каждого пациента выборки (например, до и во время проведения операции), выживаемость после операции (умер пациент или нет) и т. п.

Так как работать с большим массивом данных сложно и неудобно, и ко всему прочему многие важные детали могут ускользнуть от исследователя или, не обратив внимание на них, можно сделать неверные выводы¹⁰, то были разработаны процедуры, позволяющие проанализировать имеющийся массив данных и по возможности сократить количество информации, но не навредив ее качеству. То есть иметь всю необходимую информацию о полученных данных эксперимента, но при этом не отягощать себя громоздкими таблицами и списками результатов эксперимента. Это значительно улучшает восприятие материала и соответственно облегчает работу с ним. Данная процедура называется **описательной статистикой**. С некоторыми из ее этапов вы уже знакомы.

¹⁰ Вспомните пример с анестетиком из первой главы, внешне цифры в обеих колонках мало различимы, однако при дальнейшем анализе выяснилось, что различия существенны.

Описательная статистика является одним из первых этапов статистического анализа. И позволяет кратко и емко описать полученные данные.

Она включает в себя:

- Формирование таблиц результатов анализа (строго говоря, это предварительный этап).
- Проверка данных на возможное наличие артефактов (выбросов).
- Построение графика распределения значений признака (гистограмма частот).
- Расчет основных параметров распределения.
- Формирование выводов относительно полученных данных эксперимента о принадлежности их к тому или иному типу распределения и как следствие — выбор метода дальнейшего анализа.



Мы не будем останавливаться на первом пункте этапа описания данных, его достаточно емко мы рассмотрели в предыдущей главе. Напомним, что все варианты упорядочиваются в виде **вариационного ряда**, находятся **минимальные** и **максимальные** значения, далее строится **таблица частот встречаемости** (*абсолютные частоты*), при необходимости рассчитываются *относительные частоты*, строится **график распределения**.

Вспомним основные типы распределения, наиболее часто встречающиеся в медико-биологических исследованиях (рисунок 25, 26, 27, 28):

Нормальное распределение:

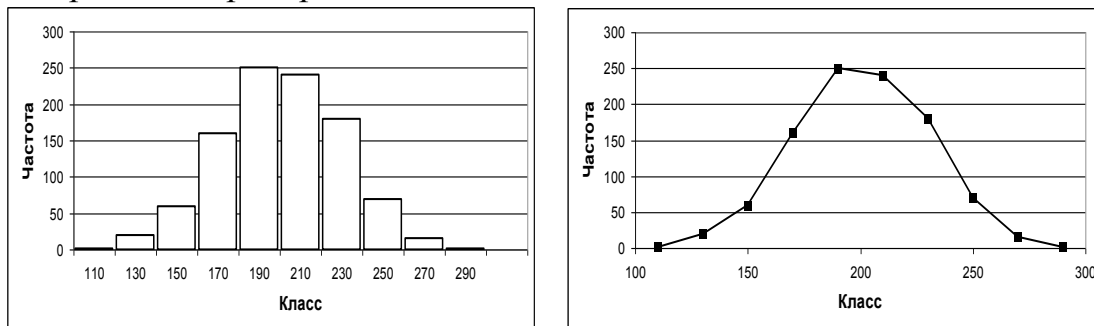


Рисунок 25 — Внешний вид нормального распределения значений признака

Равномерное распределение:

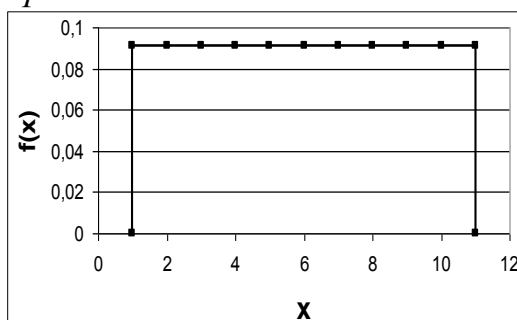


Рисунок 26 — Равномерное распределение

Асимметричное распределение (если асимметрия левосторонняя – логнормальное распределение):

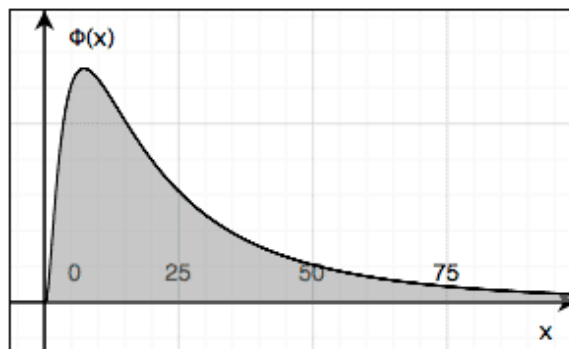


Рисунок 27 — Асимметричное распределение

Полимодалное распределение:

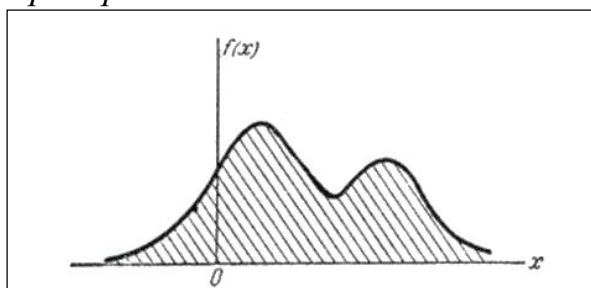


Рисунок 28 — Двумодальное распределение

На этом этапе стоит остановиться. Тип распределения для дальнейшего анализа имеет очень большое значение. Грубо говоря, все возможные типы распределения значений признака можно разделить на два: нормальное распределение и распределение отличное от нормального. Это достаточно условное разделение, правильным тоном является изучение каждого конкретного случая, но для дальнейшего понимания излагаемого материала достаточно запомнить именно этот момент. Если распределение нормальное, то для анализа данных используются одни методы, если отличное от нормального — другие. В литературе также встречаются полярные точки зрения. С одной стороны, некоторые авторы рассматривают только нормальное распределение, при этом утверждая, что при увеличении количества данных эксперимента распределение будет стремиться к нормальному. Другие же авторы, напротив, рекомендуют не пользоваться методами, разработанными для нормального распределения, а сразу приступать к оценке данных, используя альтернативные методы анализа, согласно которым тип распределения не имеет существенного значения.

Тем не менее дальше мы будем в основном говорить о распределении нормальном, так как оно более просто для понимания. Другими словами рассмотрим методы **параметрической статистики** и в отдельной главе — **непараметрические** методы.

Рассмотрим основные параметры распределения.

§1. Основные параметры распределения

1.1. Среднее значение

Среднее значение можно определить по следующей формуле:

$$\mu = \frac{\sum X}{N},$$

т. е. отношение суммы значений всех переменных к их количеству (N — количество для совокупности, n — для выборки, X — значение переменной).

Среднее значение — величина абстрактная. В выборке может не быть значений равных среднему значению по выборке, так же среднее значение может иметь величину, которую признак вообще принять не может, например «в среднем 5,8 человека на место» и т. п. Помимо того, что среднее значение — абстракция, оно очень мало характеризует выборку (особенно это проявляется при распределении отличном от нормального). Для достаточной информации о выборке необходимо знать еще некоторые параметры. И в первую очередь — **разброс** значений признака относительно среднего значения.

1.2. Дисперсия и стандартное отклонение

Величина, характеризующая разброс значений от среднего называется **дисперсией**. Она определяется по следующей формуле:

$$D = \frac{\sum (X - \mu)^2}{N},$$

где X — конкретное значение признака, μ — среднее значение выборки (генеральной совокупности); N — количество членов выборки¹¹ (генеральной совокупности).



Дисперсия — средний квадрат отклонения значений выборки от среднего по выборке. Ее величина показывает насколько в среднем отличаются полученные значения признака от среднего значения по выборке.

Еще раз стоит отметить, что это усредненная величина.

Как видно из формулы, дисперсия имеет размерность равную квадрату размерности измеряемой величины. Т. к. оперировать квадратом размерности величины не удобно (например если варьируемый признак имеет размерность $см$, то дисперсия измеряется в $см^2$), на практике чаще используют корень квадратный от дисперсии называемый **стандартным отклонением**.

¹¹ Строго говоря для выборки следует писать $n - 1$, где n — и есть количество элементов в выборке.

Стандартное отклонение также характеризует разброс значений, но измеряется в той же размерности, что и сами значения (например, сантиметры, метры, килограммы и т. п.). Можно сказать, что при представлении данных обе эти величины — синонимы и в результатах исследования достаточно указать значение одной из них.

Для лучшего понимания смысла стандартного отклонения (дисперсии) рассмотрим рисунок 29:

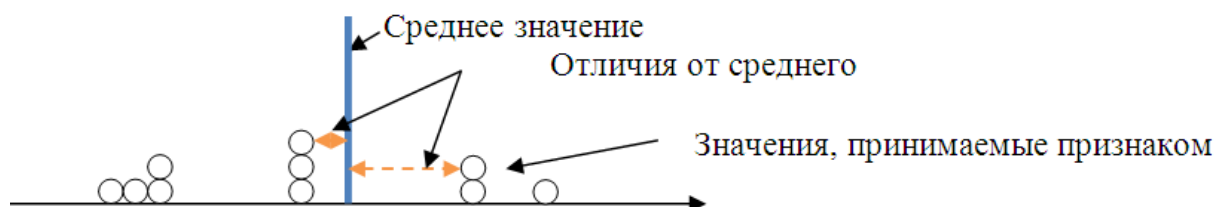


Рисунок 29 — Изображение нормального распределения значений признака с нанесенными на него средним значением и расстоянием данных эксперимента относительно среднего значения

Из него видно, что каждое значение признака отличается от среднего значения в большую или меньшую сторону, причем либо на большую величину (большая разность по модулю), либо на меньшую (меньшая разность по модулю).



Стандартное (среднеквадратичное) отклонение позволяет усредненно охарактеризовать эти различия по всей выборке и выражает их одним числом, что дает достаточное представление о среднем разбросе значений относительно среднего значения μ .

Для генеральной совокупности стандартное отклонение вычисляется по формуле:

$$\sigma = \sqrt{D} = \sqrt{\frac{\sum (X - \mu)^2}{N}}.$$

Для выборки (с чем обычно и имеет дело исследователь) формула имеет вид:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{n - 1}}.$$

В литературе дисперсия часто обозначается как D или σ^2 , стандартное отклонение как σ или s^{12} .

¹² Для выборочных данных.



Стандартное отклонение — важный статистический показатель, но когда сообщаются статистические результаты, о нем часто забывают. Без этого показателя вы видите только часть информации относительно данных.

Почему же было необходимо ввести еще одну величину для характеристики распределения? В статистике часто приводят следующий пример: представьте себе человека, одной ногой стоящего в ведре с ледяной водой, а второй — в ведре с кипятком. В среднем несчастный должен чувствовать себя отлично! Но учтите разницу двух температур для каждой его ноги! А именно эту разницу и показывает значение стандартного отклонения и «старается скрыть» среднее значение.

Другой пример. Данные о средней зарплате могут не в полной мере отражать реальное положение дел в компании, если разброс окладов очень большой. Кто-то ест мясо, а кто-то капусту, в среднем вместе едят голубцы¹³. В обоих случаях среднее значение является далеко не полной характеристикой реального положения вещей.



Важность замечаний в следующем: нельзя полагаться только на знание *среднего значения* без учета величины стандартного отклонения и как выяснится далее без знания типа распределения. Представления об объекте могут быть представлены в искаженном виде.

Изобразим этот тезис с помощью простейшего вымышленного примера. В некоторой организации имеется два отдела по два человека в каждом. В первом отделе у одного сотрудника оклад 200 рублей, у второго 1800 рублей. Среднее значение по отделу равно 1000 рублей. В другом отделе несколько иная ситуация: у одного сотрудника — 900 рублей, у второго — 1100. При этом среднее значение также равно 1000 рублей. Как видно в обоих отделах средние значения одинаковы, но при этом значительно различаются сами данные (совершенно разное расстояние между минимумом и максимумом). Так же совершенно не учитывается разброс значений в группе относительно среднего значения (очевидно, что в первом отделе он значительно больше, чем во втором). И если в отчетных данных будет представлено только среднее значение, без учета разброса относительно него, то информация явно будет неполной.

¹³ Строго говоря, данный пример не совсем корректный, так как для описания качественных признаков требуются иные параметры и средние значения со стандартным отклонением в этом случае не подойдут, так как их можно использовать только для количественного типа данных.

Выведем следующее правило:



Чем больше разброс значений, тем больше по величине стандартное отклонение (дисперсия).

Продemonстрируем данное правило с помощью графика нормального распределения¹⁴. В этом случае явно видна зависимость ширины «колокола» от величины стандартного отклонения: чем больше дисперсия (или стандартное отклонение), тем шире «колокол», т. е. разброс значений признака больше (рисунок 30).

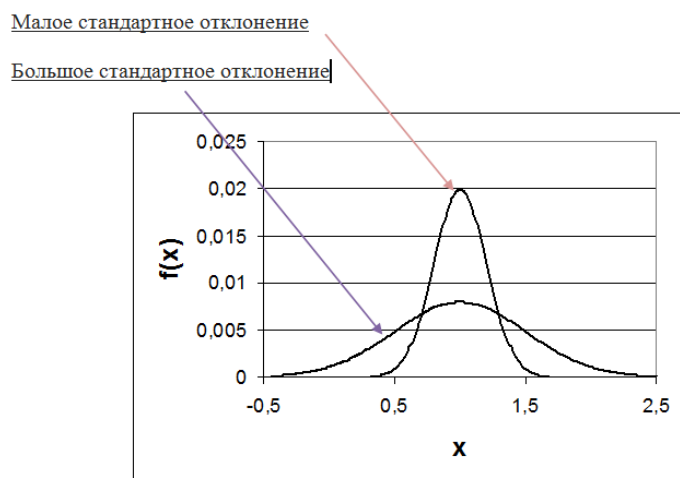


Рисунок 30 — Графики нормального распределения с разными значениями дисперсии

Теперь можно поговорить о некоторых свойствах нормального распределения.

Рассмотрим следующий график (рисунок 31):

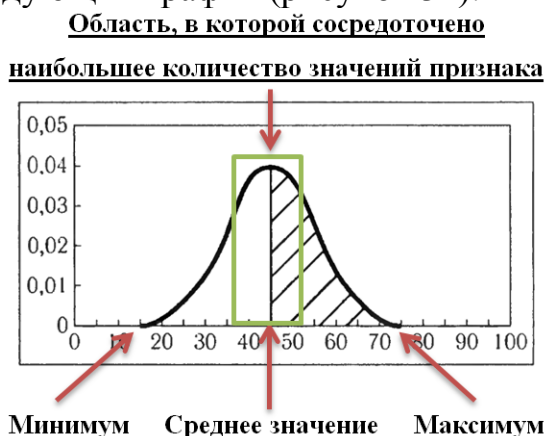


Рисунок 31 — График нормального распределения с обозначением области наибольшего сосредоточения данных эксперимента

¹⁴ Забегая вперед, заметим, что для любого другого типа распределения (отличного от нормального) среднее значение и стандартное отклонение мало информативны, и для описания таких типов распределения используют другие параметры.

Здесь представлено идеальный случай нормального распределения. Исходя из этого рисунка видно, что:

- Данные разбросаны симметрично относительно среднего значения.
- Площадь заштрихованной фигуры равна площади не заштрихованной.
- Нормальное распределение подразумевает, что большая часть значений признака находится в районе среднего значения (μ).

Более детально это можно изобразить следующим образом (рисунок 32):

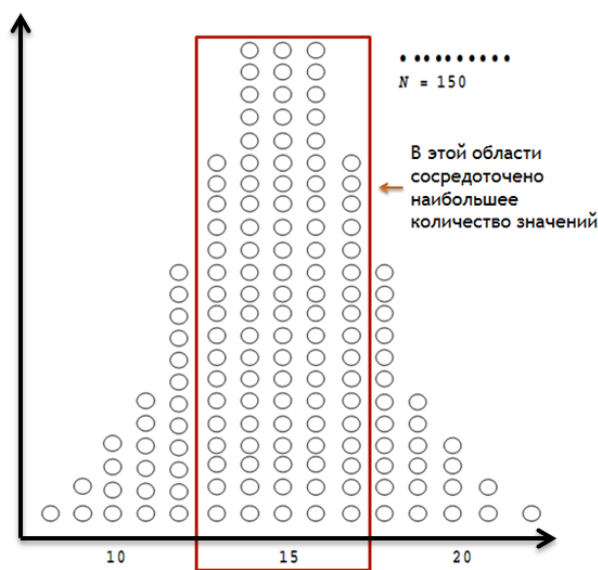


Рисунок 32 — Еще один график нормального распределения с детальным обозначением области наибольшего сосредоточения данных эксперимента

Выведем отсюда важное правило:



Если имеет место нормальное распределение признака, то наиболее часто в выборке встречаются значения близкие по величине к среднему значению по выборке μ и расположены они **симметрично** относительно среднего значения.

Далее, познакомившись со стандартным отклонением, можно вывести еще некоторые свойства нормального распределения.

1.3. Правило трех стандартных отклонений

Было обнаружено, что при нормальном распределении значений признака «большая часть» результатов, располагающаяся в пределах одного стандартного отклонения (см. область на графике ниже). То есть по обе стороны от среднего значения, количество значений признака (в процентном отношении) выборки в пределах одного стандартного отклонения всегда одно и то же и не зависит от величины стандартного отклонения. И со-

ответствует 68 % исследуемой группы (т. е. 34 % ее элементов располагается слева и 34 % — справа от средней) (рисунок 33):

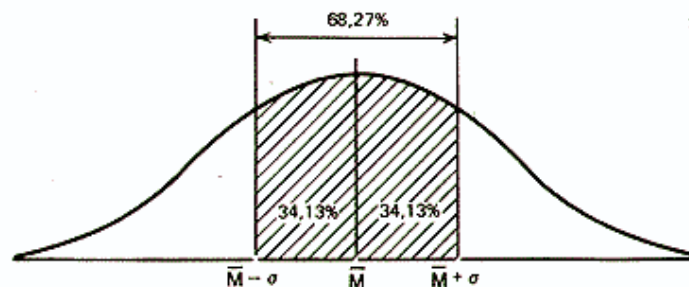


Рисунок 33 — График нормального распределения с обозначением области, равной интервалу от $-\sigma$ до $+\sigma$

Точно так же рассчитали, что 94,45 % элементов популяции при нормальном распределении не выходит за пределы двух стандартных отклонений от средней (рисунок 34):

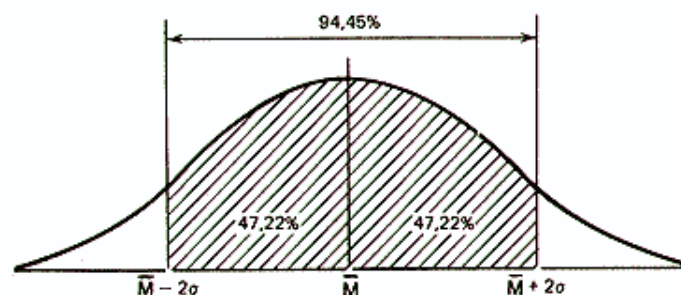


Рисунок 34 — График нормального распределения с обозначением области, равной интервалу от -2σ до $+2\sigma$

В пределах трех стандартных отклонений уместается почти вся популяция (все полученные в ходе эксперимента данные) — 99,73 % (рисунок 35).

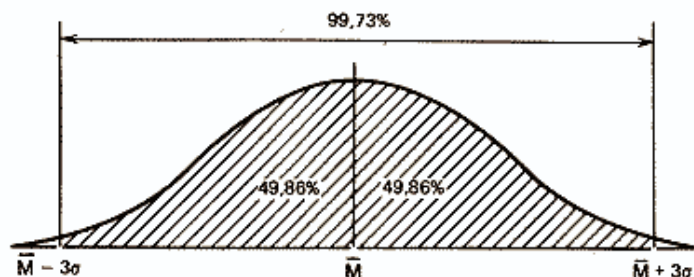


Рисунок 35 — График нормального распределения с обозначением области, равной интервалу от -3σ до $+3\sigma$

Отсюда сформулируем правило трех стандартных отклонений (правило трех сигм) (рисунок 36).



Практически все значения нормально распределённой случайной величины лежат в интервале от -3σ до $+3\sigma$. Более строго — приблизительно с 99,73 % вероятностью значение нормально распределённой случайной величины лежит в интервале от -3σ до $+3\sigma$.

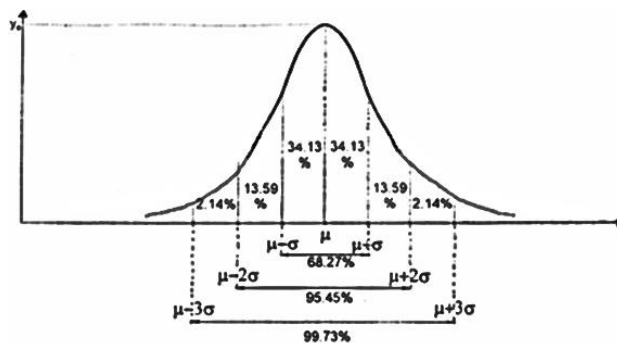


Рисунок 36 — Иллюстрация правила трех стандартных отклонений

Более подробно данное правило можно описать следующим образом: в случае нормального распределения 68 % наблюдаемых значений отклоняются от среднего значения μ не более чем на величину стандартного отклонения σ , 95 % значений не выйдут из пределов $\mu \pm 2\sigma$ и практически все значения уместятся в пределы $\mu \pm 3\sigma$. Вероятность отклонения за пределы 3σ равна $0,0026 \approx 0,003$, т. е. такое событие наступит только в среднем в 3 случаях из 1000 испытаний.

Следовательно, зная среднее значение и величину стандартного отклонения при условии нормальности распределения случайной величины (признака) можно говорить о том, в каких пределах находится большая часть результатов измерения. А также зная только лишь среднее значение и стандартное отклонение можно судить о максимальных и минимальных значениях в выборке, каких значений будет больше, а какие вероятно совсем не будут там встречаться. Например: средний рост людей в выборке 174 см, а стандартное отклонение — 10 см, отсюда следует, что в основном в выборке присутствуют люди ростом в интервале от 164 до 184 и маловероятно встретить человека ростом 207 см или 145 см.

Исходя из всего вышесказанного можно вывести еще одно очень важное правило:



Если имеет место нормальное распределение признака, то выборку можно исчерпывающе охарактеризовать параметрами: средним значением (математическим ожиданием), дисперсией или стандартным отклонением (среднеквадратичным отклонением). Или короче: нормальное распределение полностью характеризуется средним значением μ и стандартным отклонением σ .

Какой можно сделать из этого вывод? Он достаточно просто, но и очень полезен. Зная **среднее значение** и **стандартное отклонение** группы данных, распределенных по нормальному закону, можно отложить в сторону громоздкие таблицы исходных данных, а сообщить только значения этих двух параметров.

Но еще раз стоит обратить внимание, что данное правило работает, только если распределение нормальное или близко к нормальному.

4.1. Медиана, мода, перцентили

Выше мы указали, что в случае нормальности распределения значений признака среднее значение и стандартное отклонение могут достаточно полно характеризовать полученные в ходе исследования данные. Но давайте рассмотрим следующий пример.

Пусть имеется некоторая организация, которой необходимо предоставить отчет о зарплате ее сотрудников. Были собраны данные за 10 месяцев по заработной плате каждого из сотрудников, для каждого сотрудника было рассчитано среднее значение его заработка за этот период и уже эти средние значения были сведены в таблицу 12.

Таблица 12 — Исходные данные эксперимента

Заработная плата, руб				
210	285	290	285	580
200	290	295	290	590
200	295	300	295	610
200	300	305	300	600
200	305	310	305	615
250	310	300	310	1800
200	300	320	320	1900
250	320	310	500	2500
300	310	295	510	2000
320	295	280	500	2400
310	320	270	550	
295	310	275	570	
280	295	280	550	
270	280	285	600	
275	270	280	570	
280	275	285	280	

Построим график распределения (рисунок 37):

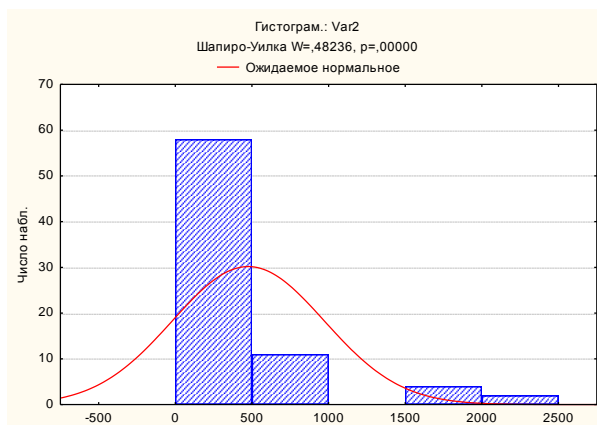


Рисунок 37 — График распределения значений данных из таблицы 12

Налицо явная асимметрия распределения. Теперь рассчитаем среднее значение и стандартное отклонение (рисунок 13):

Таблица 13 — Описательные статистики

Среднее	456,9
Стандартное отклонение	470,1
Дисперсия выборки	221038,8

Посмотрите на данные в исходной таблице и сравните их со средним значением. Можно сказать, что около 70 % опрошенных имеют оклад ниже среднего и только небольшое количество значительно выше или приближается к среднему значению. Другими словами, среднее значение дает искаженное представление о реальном состоянии дел в организации.

Исходя из этого примера, можно заключить, что если распределение отлично от нормального (например, значения распределены несимметрично относительно среднего) параметры «*среднее значение*» и «*дисперсия*» не являются информативными и могут ввести в заблуждение как самого исследователя, так и читателя результатов исследования.

Снова обратите внимание на график асимметричного распределения (рисунок 38):



Область, в которой сосредоточено наибольшее количество значений признака

Рисунок 38 — График асимметричного распределения с обозначением области наибольшего сосредоточения значений признака

На графике видно, что среднее значение мало характеризует распределение, большая часть значений признака находится далеко не в районе среднего значения, и, сообщив информацию о среднем без указания типа распределения, можно ввести в заблуждение читателя результатов вашего исследования.

Поэтому в случае распределения, отличного от нормального, совокупность обычно описывается с помощью **моды** и **медианы**, а также **процентилей**.



Медиана — значение, которое делит распределение пополам, в результате справа и слева от него находится равное число значений. То есть, по сути, она показывает реальную середину распределения значений признака, не зависимо от типа этого распределения. **Мода** — наиболее часто встречающееся значение. Используется как дополнение к медиане, она еще более детально описывает имеющееся распределение (рисунок 39).

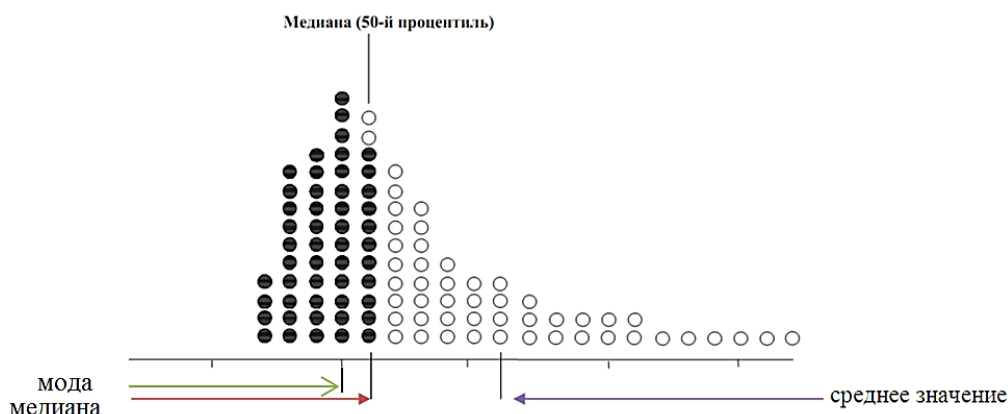


Рисунок 39 — Изображение данных, имеющих асимметричное распределение с обозначением параметров

Иногда весь диапазон значений разбивают на четыре интервала — **перцентили** (процентили) (рисунок 40).

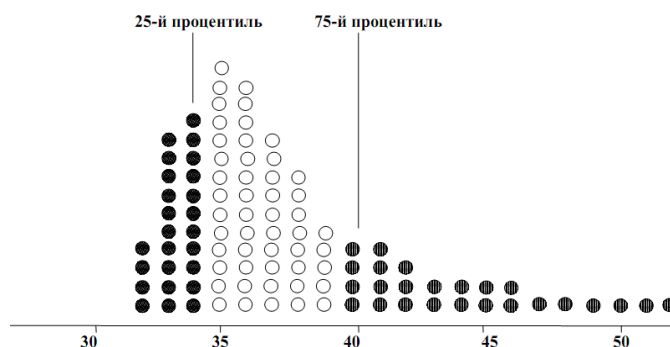


Рисунок 40 — Изображение данных, имеющих асимметричное распределение с обозначением перцентилей

Выделяют перцентили 25-й и 75-й, так же называемые *квартелями*. Однако перцентиль можно рассчитать и 20-й (80-й) и любой другой в зависимости от целей исследователя.

Данные параметры более точно характеризуют отличное от нормального распределение, и они предпочтительнее, нежели среднее значение и дисперсия.

Посчитаем значение моды и медианы для вышеприведенного примера (таблица 14):

Таблица 14 — Описательные статистики со значениями моды и медианы

Среднее	456,9
Медиана	300
Мода	310
Стандартное отклонение	470,1

Из таблицы видно, что середина распределения приходится на значение 300 рублей, а наиболее часто встречающееся значение — 310 рублей. Понятно, что данные характеристики более правдивы, нежели среднее значение.



Еще раз отметим, что *среднее значение* и *стандартное отклонение* пригодны для описания только *нормального распределения*. Тогда как для других типов распределений эти параметры дают не совсем корректное описание. Однако все же имеет смысл рассчитывать и их, но при этом не забывать о типе имеющегося распределения.

В случае с нормальным распределением значение медианы, моды и среднего значения обычно близки по своей величине. На графике это можно представить следующим образом (идеальный случай) (рисунок 41):

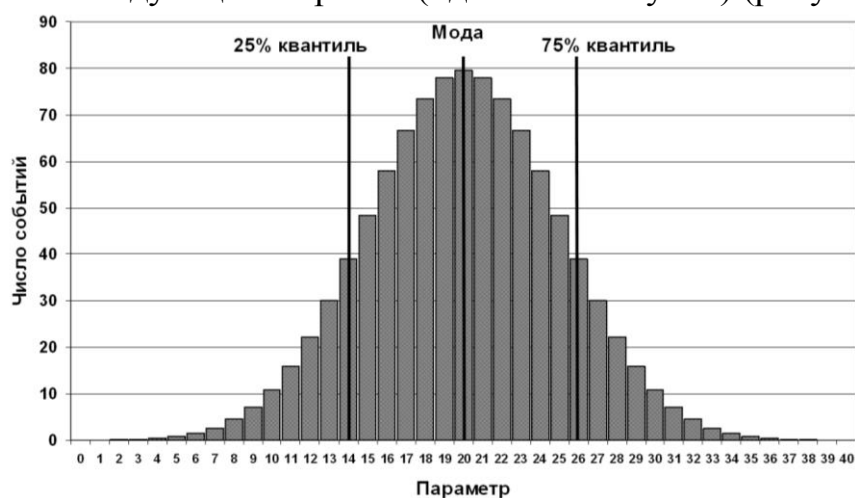


Рисунок 41 — Гистограмма нормального распределения с обозначением моды и 25-го и 75-го квантилей

Как видно из графика, медиана и среднее значение равны 20, таким же по величине является и наиболее часто встречаемое в совокупности значение признака — мода.



Отсюда можно заключить, что если при расчетах основных параметров распределения значения *моды, медианы и среднего значения* будут близки по своей величине, то вероятнее всего мы имеем дело с нормальным распределением (это свойство можно охарактеризовать как косвенный признак нормальности распределения).

В природе наиболее часто встречается распределение признака по *нормальному закону*. Однако в медицинских исследованиях так бывает не всегда. Очень часто речь идет о том, что исследователь не может однозначно сказать, что распределение является нормальным (или вообще как-то однозначно судить о типе распределения), этому могут быть несколько причин, наиболее частой из которых является недостаточное количество данных собранных в ходе исследования.



Для распределения, не являющегося нормальным, **параметрические** методы неприменимы, их использование может привести к серьезным ошибкам в выводах об исследуемой совокупности. В таких случаях разумнее воспользоваться **непараметрическими** или **ранговыми** методами, которые можно применять для любых распределений.

§2. Критерии проверки распределения случайных величин на соответствие нормальному распределению

На практике обычно пользуются внешним видом графика полученного распределения. И, исходя из этого, определяют к какому его отнести типу. Но бывает так, что внешнего вида не достаточно и для определения типа распределения полученных данных и требуются дополнительные расчёты критериев нормальности. Числовое значение этих критериев (критерии нормальности) указывает на принадлежность распределения к нормальному распределению¹⁵. Без внедрения математического аппарата рассмотрим некоторые из этих критериев.

¹⁵ Как пользоваться критериями — тема для отдельной главы. И, чтобы облегчить восприятие излагаемого материала, в данном пособии не рассматривается. Интересующимся можем порекомендовать книги Ребровой О. Ю. «Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA» или «Основы статистики: лабораторный практикум» Жученко Ю. М., Ковалева А. А., Игнатенко В. А. Важно только отметить, что критерий демонстрирует: насколько рассматриваемое распределение соответствует нормальному распределению и выражается в виде числа, а также и числового значения достоверности того, что распределение является нормальным.

2.1. Критерий Шапиро — Уилка

Критерий Шапиро — Уилка (W) используется для проверки предположения: что случайная величина X распределена нормально и является одним наиболее эффективных критериев проверки нормальности распределения. При этом данный критерий может использоваться для **малых выборок** (от 5 и выше), что является его неоспоримым преимуществом в среде медико-биологических исследований.

2.2. Критерий Колмогорова-Смирнова (λ (лямбда))

Критерий λ предложен советскими учеными А. Н. Колмогоровым и Н. В. Смирновым. Он может применяться для определения достоверности расхождения между фактическими и теоретическими распределениями, а также различий между любыми двумя распределениями частот одного и того же признака даже в том случае, когда число классов и число данных у этих распределений неодинаково. Единственным условием применения критерия лямбда является достаточная численность элементов в выборке — не менее нескольких десятков данных.

2.3. Критерий по асимметрии и эксцессу

В тех случаях, когда какие-нибудь причины влияют на появление значений признака, отличающихся от средней величины в сторону уменьшения или увеличения, образуются **асимметричные распределения**.



Асимметрия представляет собой числовое отображение степени отклонения графика распределения признака от симметричного графика распределения.

При асимметрии распределение имеет увеличенные (против симметричного расположения) частоты слева или справа относительно среднего значения (рисунок 42).

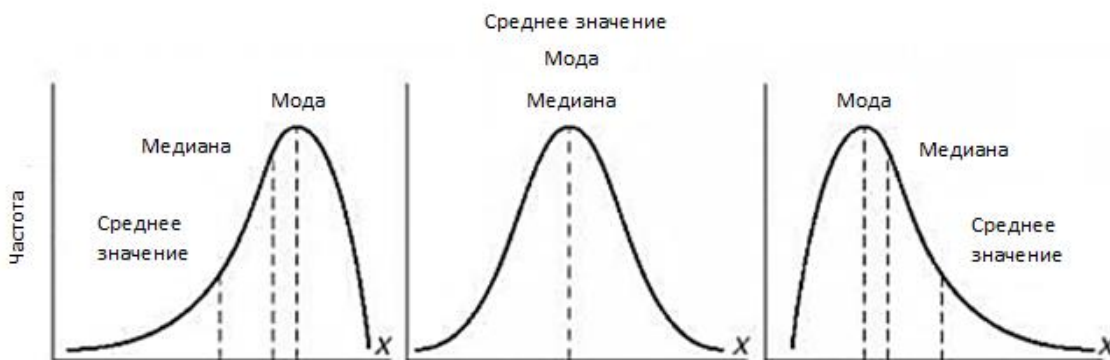


Рисунок 42 — Демонстрация асимметричности распределения значений относительно симметричного распределения

Перечислим основные правила анализа значения асимметрии (рисунок 43):

- Асимметрия симметричного распределения близка или равна 0.
- Если асимметрия больше 0, то чаще в распределении встречаются значения меньше среднего. Такая асимметрия называется **положительной** или **левосторонней**.
- Если асимметрия меньше 0, то в распределении чаще встречаются значения больше среднего. Такая асимметрия называется **отрицательной** или **правосторонней**.

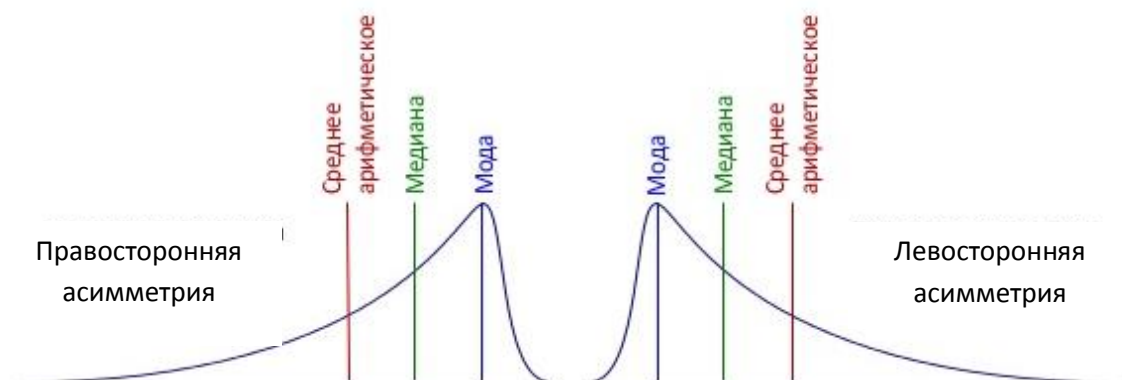


Рисунок 43 — Правосторонняя и левосторонняя асимметрия



Эксцесс — показатель остроты пика графика распределения.

При расчете эксцесса обращают внимание на следующее (рисунок 44):

- Эксцесс симметричного распределения равно 0.
- Если эксцесс больше 0, то график называется **островершинным**.
- Если эксцесс меньше 0, то график называется **плосковершинным**.



Рисунок 44 — Эксцессивные распределения ($E > 0$ и $E < 0$) и нормальное распределение

В тех случаях, когда какие-нибудь причины благоприятствуют преимущественному появлению и средних, и крайних значений признака, образуются положительные эксцессивные распределения, имеющие вид острой пирамиды с расширенным основанием. При отрицательном эксцессе в центре распределения имеется не вершина, а впадина, причем распределение становится **двумодальным** (см. выше), а кривая распределения (вариационная кривая) — **двувершинной**.

В силу вышесказанного, значения **асимметрии** и **эксцесса** так же могут быть использованы для проверки нормальности распределения.

Для выяснения достоверности того, что изучаемое распределение отличается от нормального именно в сторону асимметрии или эксцесса, применяют обычный в статистике метод сравнения показателей с их ошибками репрезентативности.

Показатели асимметрии и эксцесса с их ошибками репрезентативности определяются по следующим формулам:

$$A = \frac{\sum x^3}{n \cdot \sigma^3}; \quad s_A = \sqrt{\frac{6}{n}}; \quad t_A = \frac{A}{s_A} \geq 3;$$

$$E = \frac{\sum x^4}{n \cdot \sigma^4} - 3; \quad s_E = 2 \cdot \sqrt{\frac{6}{n}}; \quad t_E = \frac{E}{s_E} \geq 3,$$

где A — показатель асимметрии; $\sum x^3 = \sum (X_i - \mu)^3$ — сумма кубов отклонений от средней арифметической (центральных отклонений); σ^3 — стандартное отклонение, возведенное в третью степень; E — показатель эксцесса; $\sum x^4 = \sum (X_i - \mu)^4$ — сумма четвертых степеней центральных отклонений; σ^4 — четвертая степень стандартного отклонения; n — общее число данных в эмпирическом распределении; s_A, s_E — ошибки репрезентативности показателей асимметрии и эксцесса; t_A, t_E — критерии достоверности выборочных показателей асимметрии и эксцесса.



Показатели асимметрии и эксцесса свидетельствуют о достоверном отличии распределения от нормального в том случае, если они превышают свою ошибку репрезентативности в три и более раз.

§3. Артефакты

Бывает так, что в ходе сбора данных для исследования в процессе статистического анализа обнаруживаются данные, имеющие значения явно невозможные (не маловероятные, а именно *невозможные*) в рамках исследования, но которые могут значительно повлиять на ход эксперимента. Такие значения признака необходимо проверить на их принадлежность к **артефактам**.

Артефакты (или выбросы) — такие записанные значения признака, которые резко отличаются от всех других значений признака в группе.



Проверка артефактов должна проводиться всегда перед началом обработки полученных первичных данных.

Если подтвердится, что резко выделяющееся значение (например, записанное значение роста 263 см обычного человека) действительно не может относиться к объектам данной группы и попало в записи вследствие ошибок внимания, то такой артефакт необходимо исключить из обработки. Проверка артефактов может производиться по критерию, равному нормированному отклонению выброса.

$$T = \frac{\hat{X}_i - \mu}{\sigma} \geq T_{st},$$

где T — критерий выброса; X_i — выделяющееся значение признака (или очень большое или очень малое); μ , σ — средняя и стандартное отклонение, рассчитанные для группы, включающей артефакт; T_{st} — стандартные значения критерия выбросов, определяемых по таблице 15:

Таблица 15 — Стандартные значения критерия выбросов (T_{st})

n	T_{st}	n	T_{st}	n	T_{st}	n	T_{st}
2	2,0	16–20	2,4	47–66	2,8	125–174	3,2
3–4	2,1	21–28	2,5	67–84	2,9	175–349	3,3
5–9	2,2	29–34	2,6	85–104	3,0	350–599	3,4
10–15	2,3	35–46	2,7	105–124	3,1	600–1500	3,5

Примечание: n — выборка. Если $T \geq T_{st}$, то анализируемое значение признака является выбросом. Альтернатива $T < T_{st}$ не позволяет исключить из анализа значение признака.



Артефакты могут являться следствием ошибки в ходе эксперимента, неправильной записи результата, сбоя измерительных приборов и т. д. Проверка на наличие выбросов желательна при предварительном анализе полученных данных, так как их наличие может существенно повлиять на конечные выводы об исследуемой совокупности. В принципе, если исследователь знает границы возможных результатов и какие-то полученные значения сильно выбиваются из этих границ, он может исключить их из анализа, не проводя дополнительной проверки вышеописанным способом.

Стоит заметить, что иногда исследователя могут интересовать и сами выбросы как периодически возникающие аномальные явления, конечно, если они не являются следствием ошибки.

§4. Основные выводы по описательным статистикам

Подведем итоги описания данных. Описательная статистика позволяет обобщать первичные результаты, полученные при наблюдении или в ходе эксперимента. Процедуры здесь сводятся к:

- упорядочиванию значений признака;
- построению распределения их частот;
- выявлению основных параметров распределения (например, среднего значения, моды), центральных тенденций распределения;
- оценке разброса данных по отношению к найденной центральной тенденции.

Когда совокупность подчиняется нормальному распределению, она исчерпывающе описывается такими параметрами распределения, как: средним и стандартным отклонением. Когда же распределение сильно отличается от нормального, более информативны медиана и перцентили.

Проверкой распределения на принадлежность нормальному закону может служить построение графика, значения коэффициентов асимметрии и эксцесса, а также критерии нормальности, например критерии Шапиро — Уилка или Колмогорова — Смирнова.

От типа распределения признака зависит дальнейший ход исследования и работа с данными эксперимента.

ГЛАВА IV. СРАВНЕНИЕ ГРУПП

§1. Постановка задачи

В предыдущей главе вы познакомились с первичным анализом данных, полученных в результате исследования. Результаты описательных статистик вы часто можете встретить на страницах научно-популярных изданий, в новостных колонках и т. д. Теперь вы знаете, что информация о среднем значении не является достаточной и для того, чтобы судить о выводах, сделанных в ходе эксперимента необходимо знать тип распределения, значения стандартного отклонения, желательно иметь представление о числе элементов выборки и насколько добросовестно эта выборка была сформирована. Однако описательная статистика — это лишь начальный этап в обработке данных в ходе научного исследования. Обычно выводы, к которым стремится исследователь выходят за рамки первичного анализа данных. И описательная статистика является только инструментом, чтобы понять, какими методами анализа можно воспользоваться дальше.

Вспомните задачу про два анестетика. Ее цель заключалась в определении, какой анестетик благоприятнее (меньше снижает артериальное давление) влияет на организм человека в ходе операции на открытом сердце. Были сформированы две группы по 60 человек в каждой. У первой группы анестезирующим средством был *галотан*, у второй — *морфин* (его как раз и проверяли как альтернативное менее негативно влияющее средство). В результате эксперимента были сформированы две таблицы результатов (таблица 16), в каждой из которых было по 60 цифр — значений сниженного давления. Необходимо выяснить, какой же из анестетиков наименее вреден для пациента.

Глядя на исходные таблицы результатов вряд ли можно что-то судить об различии препаратов, поэтому проведем первичный анализ данных. Результаты запишем в таблицу:

Таблица 16 — Описательные статистики по данным эксперимента

Галотановая		Морфиновая	
Среднее	66,52	Среднее	73,26
Медиана	66,00	Медиана	71,00
Мода	60,00	Мода	62,00
Стандартное отклонение	12,21	Стандартное отклонение	14,15
Дисперсия выборки	148,98	Дисперсия выборки	200,16
Эксцесс	-0,02	Эксцесс	-0,68
Асимметричность	0,20	Асимметричность	0,04

Из таблицы видно: средние значения значительно отличаются друг от друга, поэтому можно предположить, что анестетики на самом деле отличаются по степени воздействия и морфин менее «склонен» понижать артериальное давление. Значения моды, медианы и среднего приблизительно равны, асимметрия имеет значение близкое к нулю, что указывает на нормальность распределения признака в обеих группах и как следствие правомочности учитывать различия средних значений. Для большей надежности построим графики распределения частот и рассчитаем *критерий Шапиро — Уилка* (рисунок 45).

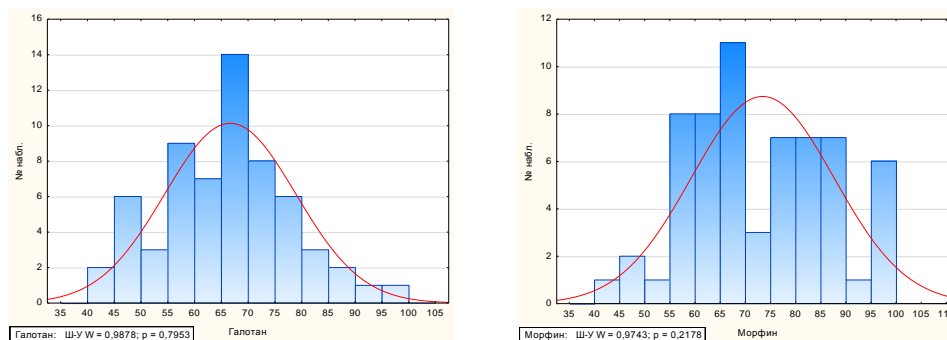


Рисунок 45 — Графики распределения по данным исследования

Судя по графикам, мы имеем дело с нормальным распределением (хотя внешний вид второго графика мог бы вызвать вопросы). Об этом же свидетельствуют значения критерия Шапиро — Уилка, а также малые значения асимметрии и эксцесса. Таким образом, как было отмечено выше, мы вправе судить о выборках используя среднее значение и значение стандартного отклонения. Видно, что средние значения значительно отличаются и, уже исходя из этого, исследователь может попытаться заключить, что группы действительно различаются и, стало быть, морфиновая анестезия более безопасна, чем галотановая. Но возникает вопрос: насколько статистически значимы видимые различия средних, насколько правомочно заключать вывод, опираясь только на данные описательной статистики? Какова вероятность ошибочного заключения? Выяснением данных вопросов занимаются методы **сравнения групп**.

В ходе исследований группы часто сравниваются между собой. Такая задача может возникнуть, если необходимо выявить, какой метод лечения более эффективный или, при изучении воздействия нового препарата, поставить эксперимент, сравнив препарат с аналогом, или исследовать эффективность препарата, поставив эксперимент с плацебо, метод сравнения групп используется на одном из этапов проверки модели при регрессионном анализе, элементы этих методов используются в ходе дискриминантного анализа и т. д. В данной главе мы рассмотрим основные методы сравнения групп, возможные ошибки и интерпретацию результатов. Стоит заметить, что сравнивать можно не только две, но и большее количество групп между собой. Для большей наглядности и упрощения дальнейшего изложения материала будем говорить о сравнении двух групп.

Следует начать с того, что в подобных вышеприведенному примеру исследованиях одна из групп испытуемых принимается за **контрольную** — например, группа, которую лечили эталонным методом лечения или давали препарат плацебо вместо испытуемого лекарства (в примере такой группой можно назвать группу с галотановой анестезией), другая группа — **экспериментальная**, новый метод лечения, новый препарат и т. д. (в примере такой группой можно назвать группу с морфиновой анестезией).

Заметим, что эти названия скорее условны и используются для детального описания исследования, а также для наглядного представления результатов эксперимента. Для понимания механизма анализа это не принципиально.

§2. Нулевая гипотеза

Прежде чем двигаться дальше необходимо ввести несколько новых понятий. Оставим пока в стороне наш пример и остановимся на разъяснении понятия **нулевой гипотезы**, так как оно довольно часто используется и дальше по тексту, и вообще в ходе любого описания анализа данных.

Результатом исследования обычно является некий вывод и предположение, впоследствии подкрепленное строгим анализом, например: каковы самые распространенные побочные эффекты определенного вида хирургического вмешательства, каков коэффициент выживаемости после определенного лечения и действительно ли новое экспериментальное лекарство улучшает состояние пациента на длительный срок без побочных эффектов.



Однако прежде чем получить какой-либо вывод (утверждение), необходимо правильно спланировать сам ход исследования, при этом одним из важных этапов является формулировка исходного предположения, правильность которого и проверяется в ходе исследования. Такое предположение называется **гипотезой**.



Гипотеза (др.-греч. ὑπόθεσις — предположение) — утверждение, предполагающее доказательство. А статистическая процедура, предназначенная для проверки утверждения — **проверка гипотезы**.

В нашем примере исследователь утверждает, что, *возможно*, морфиновая анестезия меньше снижает давление, чем галотановая. Именно это утверждение (гипотезу) и необходимо проверить в ходе анализа.

После сбора данных мы получили таблицу, для дальнейшего описания наших действий в статистике принято использовать понятие **нулевой гипотезы**. Понятие нулевой гипотезы вводится для акцентирования внимания на том, что:



При проведении статистического исследования, необходимо учитывать возможное отсутствие закономерности, а статистической значимости в полученных данных эксперимента может и не быть вовсе. **Нулевая гипотеза** — это, по сути, предположение, что результата (конечной цели любого исследования) не существует.

Попытаемся разъяснить данное утверждение на нашем примере. Перед процедурой сравнения групп, мы видим, что средние значения различаются между собой, видимые различия имеют и данные в исходных таблицах сбора результатов, и мы можем предположить, что действие анестетиков на самом деле различается. Однако **нулевая гипотеза** будет утверждать обратное: видимые различия случайны и группы различаются между собой статистически *незначимо*. И в ходе дальнейшего анализа необходимо или доказать обратное, или принять верность нулевой гипотезы.

Поэтому при выполнении анализа по сравнению групп нулевую гипотезу можно определить как:



Предположение, что **исследуемые факторы** не оказывают никакого влияния на значения признака и полученные различия между исследуемыми группами случайны¹⁶.

Помимо этого в качестве нулевой гипотезы может выступать также гипотеза обратная утверждению о том, что значения признака распределены по нормальному закону распределения. Часто в качестве нулевой гипотезы выступают гипотезы об отсутствии взаимосвязи или корреляции между исследуемыми переменными, об отсутствии различий (однородности) в распределениях (параметрах распределений) двух и(или) более выборках и т. д.

Опять же стоит сказать, что само введение в описание анализа понятия нулевой гипотезы скорее больше формальность, чем необходимость, но оно позволяет более грамотно описать ход вашего исследования и анализа для дальнейшего понимания его другими людьми, а также правильное восприятие вами чьих-то экспериментов. К этому можно добавить и возможность более лаконично описать сам результат исследования.

В литературе иногда можно встретить и понятие **альтернативной гипотезы**. Альтернативная гипотеза обратна нулевой гипотезе. Однако обычно формулируется именно нулевая гипотеза, т. е. тот тезис относительно исследования, который мы хотим опровергнуть.

§3. Статистическая значимость результатов

3.1. Уровень значимости

Исследователь в подавляющем большинстве случаев работает с *выборочными данными*, и в силу этого абсолютно избежать ошибок в статистических исследованиях и выводах на их основе **невозможно**. Получив результат по выборочным оценкам, исследователь должен установить: можно ли доверять полученному результату. А что если, проведя новый эксперимент, результат его окажется совершенно противоположным? Необходимо выяснить насколько возможным является **ошибочное заключение** вследствие использования статистических методов для анализа. Ведь возможности оценить всю *генеральную совокупность* нет. И нет уверенности в том, что наш вывод является достоверным. Так же возникает вопрос, как снизить **вероятность ошибочных выводов**? Статистические методы позволяют оценить вероятность возможного возникновения подобных ошибок.

¹⁶ В нашем примере **исследуемым фактором** является тип анестетика, **признаком**, на который влияет данный фактор — давление пациента во время операции.

Начнем с того, что перед началом анализа необходимо задать **максимально приемлемую вероятность ошибки** и, уже опираясь на ее значение, можно делать выводы согласно цели исследования. Не вдаваясь в сложные теоретические обоснования, можно сказать: если мы представим, что у нас есть возможность сформировать очень большое количество выборок из одной генеральной совокупности и проделать эксперимент с каждой из них, то мы можем выяснить *в скольких случаях из всех будут сделаны ошибочные выводы*¹⁷. И как следствие получить точную вероятность ошибочного результата в ходе работы с нашей конкретной выборкой. Но так как подобный эксперимент провести не имеет возможности в силу многих причин, то было принято задавать именно максимально допустимую вероятность ошибочного результата.

Другими словами, мы допускаем наличие ошибки и максимальную вероятность ее возникновения устанавливаем сами. И если эта вероятность мала, то мы заключаем, что результаты эксперимента статистически значимы. Эта вероятность и есть **уровень значимости**.

Если перейти к более строгой формулировке, то:



Уровень значимости — максимально приемлемая вероятность отвергнуть *верную* нулевую гипотезу¹⁸. Обозначается греческой буквой α .

Обычно α принимают равным 0,05 (5 %) или 0,01 (1 %). Что говорит нам следующее:



Мы допускаем возможную ошибку в нашем исследовании, но вероятность ее возникновения не превышает 5 % (1 %). Или, другими словами, в результате работы с выборочными данными мы рискуем ошибиться только в 5 % (1 %) случаев.

Это, разумеется, еще не означает, что если в ходе исследования мы пришли к заключению о значимом действии изучаемых факторов, то они на самом деле действуют (это вопрос прежде всего планирования эксперимента), но, во всяком случае, маловероятно, что результат обусловлен случайностью.

Если в ходе исследования мы получили результат, который *отвергает нулевую гипотезу*, при уровне значимости 5 %, то можно сказать следующее:



Если бы нулевая гипотеза была справедлива, то вероятность получить наблюдаемые результаты была бы меньше 5 %.

¹⁷ Естественно при условии, что результаты экспериментов нам будут известны.

¹⁸ Например, найти различия там, где их нет.

В ходе расчетов мы также получаем значение **вероятности ошибочного результата** для нашего конкретного случая и если она окажется меньше уровня значимости, то мы отвергаем нулевую гипотезу. В принятой системе обозначений это записывается как $p < 0,05$. При этом подразумевая, что вывод может быть ошибочным, но вероятность этой ошибки равна p и она меньше максимально допустимой, другими словами ошибка слишком маловероятна.

Таким образом, p есть вероятность ошибочно отвергнуть верную нулевую гипотезу.

Опираясь на наш пример с анестетиками (сравнение групп), можно было бы заключить, что гипотеза об отсутствии различия двух препаратов на давление вряд ли справедлива, т. е. различия статистически значимы (при 5 % уровне значимости). Разумеется, этот вывод по сути своей носит вероятностный характер. Не исключено, что мы ошибочно признаем один препарат отличающимся от другого, т. е. находим различия там, где их нет. Однако мы можем утверждать, что вероятность подобной ошибки не превышает 5 %.

3.2. Ошибки первого и второго рода

Различают ошибки **первого и второго** рода.



Если мы ошибочно отклоняем верную нулевую гипотезу, например, находим различия там, где их нет, то это называется **ошибкой I рода**. Максимальная приемлемая вероятность ошибки I рода и есть **уровень значимости**.

Обычно α принимают равным 0,05 (т. е. 5 %), однако можно взять и какой-нибудь другой уровень значимости, например 0,1 или 0,01.



Если же мы не отклоняем нулевую гипотезу, когда она не верна, т. е. не находим различий там, где они есть на самом деле, то это — **ошибка II рода**.

Какая из ошибок более предпочтительна? Это зависит от самого исследования и областей, где в дальнейшем используются его результаты. Можно продемонстрировать на примере диагностики какого-либо серьезного заболевания. Мы имеем модель, согласно которой определяем диагноз и отнесение пациента к той или иной группе, исходя из каких-то входных данных. Модель работает на основании сравнения групп, сравнивая входные данные предполагаемого больного с данными здорового человека. Если последствия диагностики заболевания данной моделью носят

жизненно важный характер, и она совершила *ошибку первого рода*, т. е. нашла различия там, где их нет, и ошибочно определила здорового человека как больного (гипердиагностика), то в этом случае нет ничего плохого, и дальнейшие наблюдения и тесты позволят исправить этот недочет. Но если эта же модель совершит *ошибку второго рода*, т. е. не найдет различий там, где они есть, это может оказаться причиной отсутствия своевременного медицинского вмешательства. Однако, если некая модель диагностики определяет не столь жизненно важные последствия и тем более сам процесс дальнейших анализов и процедур является дорогостоящим или имеет неприятные побочные эффекты, то в этом случае совершение *ошибки второго рода* более предпочтительно, чем совершение *ошибки первого рода*.

Так же стоит отметить, что при уменьшении вероятности ошибки первого рода растет вероятность ошибки второго рода и наоборот. Это иллюстрируется формулой определения ошибки второго рода:

$$\beta = 1 - \alpha.$$

Но не будем очень сильно заострять внимание на данной проблеме. На данный момент достаточно знать, что обычно $\alpha = 0,05$ или $0,01$. И значение уровня значимости задается перед анализом полученных данных.

Возвращаясь к сравнению групп можно сказать, что если в ходе расчетов мы получили значение вероятности ошибки p меньше чем заданный уровень значимости, то группы скорее всего различаются между собой. Однако для полной уверенности в данном выводе необходимо рассчитать еще одну величину.

3.3. Критерий значимости

Упрощенно говоря, **критерий значимости** — это число, которое мы рассчитываем в ходе анализа сравнения групп, позволяющее определить различаются группы между собой или нет.

Полученное числовое значение критерия значимости указывает на то, принимается или отвергается нулевая гипотеза.

Делаемый нами вывод на основе **критерия значимости** тесно связан со значением **уровня значимости**. При одном и том же значении критерия значимости, но при разных уровнях значимости выводы о различии групп могут быть неодинаковыми, при том, что группы и результаты измерения не изменились. В одном случае мы можем сделать вывод о различии сравниваемых групп, в другом — отклонить эту гипотезу. А связано это с тем, что в зависимости от уровня значимости выбирается так называемое **критическое значение критерия значимости**, с которым сравнивается рассчитанное значение критерия значимости.



Если расчетный критерий оказался больше критического, то группы различаются между собой и при условии, что вероятность ошибки меньше заданного уровня значимости, эти различия признаются статистически значимыми. Если же рассчитанный критерий оказался меньше критического, то группы не отличаются друг от друга.

Мы дали общий алгоритм работы с критерием значимости при анализе сравнения групп. Этих критериев существует достаточно много. Некоторые из них подходят только для данных с нормальным распределением, для некоторых это не имеет значения. Какие-то критерии можно использовать только для сравнения двух групп, какие-то можно использовать, сравнивая несколько групп между собой. Ниже будут описаны некоторые из критериев значимости, которые используются наиболее часто, а также алгоритм и области их применения.

§4. Дисперсионный анализ

4.1. Критерий Фишера

Критерий Фишера (F) или дисперсионный анализ, используется очень часто. Это **параметрический критерий**, т. е. его можно использовать только в случае нормального распределения значений признака в обеих сравниваемых группах. Так же он может работать с несколькими группами.

Здесь мы не будем подробно расписывать вывод формулы для получения числового значения критерия Фишера. Интересующимся дадим отсылку к третьей главе книги Стентона Гланца «Медико-биологическая статистика».

Общей формулой определения критерия Фишера является следующая:

$$F = \frac{\text{Дисперсия совокупности, оцененная по выборочным средним}}{\text{Дисперсия совокупности, оцененная по выборочным дисперсиям}}$$

где, **дисперсия совокупности, оцененная по выборочным средним** = $m \times \sigma_{x-}^2$; в этом случае m — количество групп; σ_{x-}^2 — дисперсия, оцененная по выборке из средних значений каждой группы.

А дисперсия совокупности, оцененная по выборочным дисперсиям, есть сумма дисперсий каждой из сравниваемых групп.

Короче говоря, и числитель, и знаменатель в формуле для F — это оценки одной и той же величины — **дисперсии совокупности** σ^2 , по-

этому если различий между группами нет, то значение F должно быть близко к 1¹⁹.

Если F значительно превышает 1, нулевую гипотезу²⁰ следует отвергнуть. И признать, что группы различаются между собой. Если же значение F близко к 1, нулевую гипотезу следует принять и отклонить гипотезу о различии групп. Осталось понять, начиная с какой именно величины F следует отвергать нулевую гипотезу.

4.2. Критическое значение F

Если извлекать случайные выборки из нормально распределенной совокупности, значение F будет меняться от опыта к опыту²¹. Выше мы указали, что рассчитанное значение критерия значимости сравнивается с неким **критическим значением**.



Критическое значение критерия это такое значение критерия, начиная с которого мы отвергаем нулевую гипотезу.

Критическое значение критерия определяется уровнем значимости α , а также еще двумя параметрами, которые называются **внутригрупповым** и **межгрупповым числом степеней свободы** и обозначаются греческой буквой ν («ню»).



Степени свободы — числа, показывающие количество элементов варьирования, которые могут принимать произвольные значения, не изменяющие заданных характеристик.

Степени свободы пришлось ввести из практического опыта вычислений. Дело в том, что в выборке у нас n данных, скажем 10, но когда мы с ними начинаем манипулировать, то между этими десятью значениями (в ряду) оказывается 9 промежутков, т. е. $(n-1)$.

¹⁹ Другими словами дисперсионный анализ основан на сравнении дисперсий внутри каждой из сравниваемых групп с разбросом средних значений между группами. Если средние в генеральных совокупностях на самом деле не различаются, то разброс средних между выборками (по сути, являющихся в данном случае выборками из одной и той же генеральной совокупности) должен быть сопоставим с разбросом данных внутри групп. Если же группы на самом деле различаются, то разброс средних будет больше разброса внутри групп.

²⁰ Нулевой гипотезой в дисперсионном анализе является предположение о равенстве средних во всех исследуемых группах, а полученные в результате эксперимента различия статистически незначимы.

²¹ Это обусловлено тем, что данные в выборках будут разными, а, следовательно, и разными значения дисперсии в каждой из выборок.



Межгрупповое число степеней свободы — это число групп минус единица $\nu_{\text{меж}} = m - 1$. **Внутригрупповое число степеней свободы** — это произведение числа групп на численность каждой из групп минус единица $\nu_{\text{вну}} = m(n - 1)$.

Вычислить критическое значение F довольно сложно, поэтому пользуются таблицами критических значений F для разных α , $\nu_{\text{меж}}$ и $\nu_{\text{вну}}$.



Определив на основании значения **уровня значимости** и **степеней свободы** величину **критического значения критерия Фишера $F_{\text{кр}}$** делается вывод о различии между группами. Если

$$F > F_{\text{кр}}$$

и

$$p \text{ (вероятность ошибки)} < \alpha,$$

то нулевая гипотеза отвергается и различия между группами статистически значимы, в противном случае различия — случайны или незначимы.

Следует учитывать, что даже при обнаруженной статистической значимости различий исследователь может ошибаться, но максимальная вероятность ошибки равна уровню значимости (т. е. незначительна). Если оказалось, что p столь велико, что превышает уровень значимости, то различия признаются *статистически незначимыми*, правда в этом случае возможно возникновение *ошибки второго рода*, т. е. не выявление различия там, где оно есть.

Проведем дисперсионный анализ, сравнив группы из нашего примера про два анестетика. Мы ограничимся результатами, полученными при помощи MS Excel.

Как вы помните, данные в обеих группах распределены по *нормальному закону распределения*, *средние* имели значительное различие. Выберем *уровень значимости* равный 0,05. **Нулевая гипотеза:** *различие групп случайно и статистически незначимо, влияние двух анестетиков на снижение давления одинаково.*

Результаты дисперсионного анализа (таблица 17):

Таблица 17 — Значения критерия Фишера, критического значения критерия Фишера и достоверности

F	p	F критическое
8,07	0,0053	3,92

Из таблицы видно, что F больше чем $F_{\text{критическое}}$, а p значительно **меньше** уровня значимости. Следовательно нулевая гипотеза отклоняется и различие между группами статистически значимо.

В нашем примере мы рассматривали воздействие **одного фактора** (анестетик), бывает и так, что на группы влияют большее количество факторов. Дисперсионный анализ, в котором проверяется влияние одного фактора, называется **однофакторным**. При изучении влияния более чем одного фактора используют **многофакторный** дисперсионный анализ.

Дадим алгоритм дисперсионного анализа при анализе сравнения групп. Общая последовательность действий такова:

✓ Формирование исследуемых групп (например, первая группа принимает препарат, вторая - плацебо).

✓ Формирование таблиц результатов измерения (снимаются исследуемые показатели первой группы и показатели второй группы).

✓ Выполняется первичный анализ полученных данных (описательная статистика).

✓ Если данные в обеих группах распределены по *нормальному закону распределения* (или близкому к таковому), то выбирается метод анализа, в нашем случае — *дисперсионный анализ*.

✓ Выбирается значение *уровня значимости* (в зависимости от строгости исследования: 0,1 или 0,05 или 0,01). Следует помнить об *ошибках I и II рода*.

✓ Формулируется *нулевая гипотеза* (различие между группами незначимо или является следствием случайности).

✓ Рассчитывается *критерий Фишера F* и *вероятность ошибочного результата P* (вероятность ошибочно отвергнуть верную нулевую гипотезу, то есть найти различия там, где их нет).

✓ На основании значения *уровня значимости* и *степеней свободы* определяется величина *критического значения критерия Фишера $F_{кр}$* (значение критерия, начиная с которого мы отвергаем нулевую гипотезу). Сравниваются между собой F и $F_{кр}$, а также P и α , если $F > F_{кр}$ и $P < \alpha$, то нулевая гипотеза отвергается и различия между группами статистически значимы, в противном случае различия – случайны или незначимы. Следует учитывать, что даже при обнаруженной статистической значимости различий исследователь может ошибаться, но допустимая вероятность равна *уровню значимости* (т.е. незначительна).

Заметим, что дисперсионный анализ дает точный результат только при условии *нормального распределения* признака в обеих группах. Так же он может быть использован для сравнения нескольких групп.

§5. Критерий Стьюдента

Нередко нужно сравнить только две группы. В этом случае можно применить **критерий Стьюдента**. **Критерий Стьюдента** — это частный случаи дисперсионного анализа. Он чрезвычайно популярен, использование его можно обнаружить в большом количестве научных публикаций.



Чем меньше (по абсолютной величине) t , тем больше вероятность нулевой гипотезы и тем самым отсутствия различий между группами. Чем больше t , тем больше оснований отвергнуть нулевую гипотезу и считать, что различия статистически значимы.

Величина *критерия Стьюдента*, начиная с которой отвергается нулевая гипотеза, по аналогии с *критерием Фишера* называется **критическим значением критерия Стьюдента** ($t_{кр}$).



Если значение *t-критерия Стьюдента* по модулю больше чем его *критическое значение* (найденное по таблице критических значений или рассчитанное при помощи программного обеспечения)

$$t > t_{кр}$$

для заданного уровня значимости, то нулевая гипотеза отвергается, и различия считаются статистически значимыми.

При этом следует учитывать, что если бы группы представляли собой две случайные выборки из одной и той же совокупности, то вероятность получить наблюдаемые различия (или более сильные) равна 0,05 (или другому значению выбранного уровня значимости). Следовательно, ошибочный вывод о существовании различий мы будем делать в 5 % случаев. Застраховаться от подобных ошибок можно приняв уровень значимости не 0,05, а к примеру 0,01. Однако даже в этом случае ошибочные выводы о существовании различий все же не исключены — их вероятность снижается до 1 % и в тоже время вероятность не найти различия там где они есть теперь повысилась. При использовании специализированного программного обеспечения так же обычно рассчитывается вероятность ошибочного результата p , если она оказалась меньше, чем заданный *уровень значимости*, то различия признаются статистически значимыми.

Критические значения t (подобно критическим значениям F они сведены в таблицу) зависят не только от уровня значимости, но так же и от числа *степеней свободы* $\nu = 2(n - 1)$, где n — объем выборок.



Чем больше объем выборок, тем меньше критическое значение t . Это следствие правила — чем больше выборка, тем менее выборочные оценки зависят от случайных отклонений и тем точнее представляют исходную совокупность.

Рассмотрим применение критерия Стьюдента на примере решения задачи об анестетике.

Выполнив процедуру описательной статистики мы получили таблицу 18.

Таблица 18 — Описательная статистика по данным в исследуемых группах

Галотановая		Морфиновая	
Среднее	66,52	Среднее	73,26
Стандартная ошибка	1,55	Стандартная ошибка	1,80
Медиана	66,00	Медиана	71,00
Мода	60,00	Мода	62,00
Стандартное отклонение	12,21	Стандартное отклонение	14,15
Дисперсия выборки	148,98	Дисперсия выборки	200,16
Экссесс	-0,02	Экссесс	-0,68
Асимметричность	0,20	Асимметричность	0,04

Различие дисперсий незначительное (в идеале конечно это стоит проверить используя **критерий Левена**), как выяснилось ранее распределение значений нормальное, следовательно мы можем воспользоваться t-критерием Стьюдента. Рассчитаем его, результаты приведем в таблице 19:

Таблица 19 — Значения критерия Стьюдента и значимости

t-критерий Стьюдента	-2,8411
p(T<=t) одностороннее	0,0026
t критическое одностороннее	1,6574
p(T<=t) двухстороннее	0,0053
t критическое двухстороннее	1,9796

Полученное значение t-критерия берем по модулю и сравниваем с $t_{кр}$. Видно, что расчетное значение *критерия Стьюдента* больше, нежели чем его критическое значение, так же $p(T \leq t)$ значительно меньше заданного *уровня значимости* $\alpha = 0,05$. Отсюда можно заключить, что различие групп статистически значимо.

Выше был приведен пример сравнения двух выборок с одинаковым количеством членов. Такое бывает далеко не всегда. Формула для расчета *критерия Стьюдента* для **разновеликих выборок** приведена ниже:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

где s^2 — **объединенная оценка дисперсий**, рассчитывается по формуле:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

где s_1 и s_2 — стандартные отклонения первой и второй выборок соответственно.

Число степеней свободы в этом случае $\nu = n_1 + n_2 - 2$.

5.1. Ошибки в использовании критерия Стьюдента

Критерий Стьюдента в первую очередь предназначен для сравнения двух групп. Однако зачастую его используют для оценки различий большего числа групп посредством попарного их сравнения. При этом редко учитывают возникающий при этом эффект **множественных сравнений**.

Рассмотрим пример (по С. Гланц Медико-биологическая статистика). Исследуют влияние препаратов А и Б на уровень глюкозы плазмы. Исследование проводят на трех группах — получавших препарат А, получавших препарат Б и получавших плацебо В. С помощью критерия Стьюдента проводят три парных сравнения: группу А сравнивают с группой В, группу Б — с группой В и наконец А с Б. Получив достаточно высокое значение t в каком-либо из трех сравнений сообщают что « $P < 0,05$ ». Это означает, что вероятность ошибочного заключения о существовании различий не превышает 5 %. Но это неверно: вероятность ошибки значительно превышает 5 %. Разберемся подробнее. В исследовании был принят 5 % уровень значимости. Значит, вероятность ошибиться при сравнении групп А и В — 5 %. Казалось бы все правильно. Но точно также мы ошибемся в 5 % случаев при сравнении групп Б и В. И наконец при сравнении групп А и Б ошибка возможна также в 5 % случаев. Следовательно, вероятность ошибиться хотя бы в одном из трех сравнений, составит не 5 %, а значительно больше. Итак, в нашем исследовании вероятность ошибиться хотя бы в одном из сравнений составляет примерно 15 %. При сравнении четырех групп число пар и соответственно возможных попарных сравнений еще больше и равно 6. Поэтому при уровне значимости в каждом из сравнений 0,05 вероятность ошибочно обнаружить различие хотя бы в одном равна уже не 0,05, а примерно $6 \times 0,05 = 0,30$. И когда исследователь, выявив таким способом «эффективный» препарат, будет говорить про 5 % вероятность ошибки, на самом деле эта вероятность равна 30 %. В такой ситуации необходимо воспользоваться поправкой Бонферрони. Эта поправка учитывает попарное сравнение и уменьшает уровень значимости, тем самым делая критерий более строгим. Метод поправки Бонферрони утверждает, что для уменьшения ложноположительных результатов, необходимо отклонить только те нулевые гипотезы, для которых $p < \alpha/m$, где m — количество попарных сравнений²². Таким образом, если с помощью крите-

²² Эта процедура называется коррекцией на множественное сравнение. Примером возможных ошибочных выводов при отсутствии учета коррекции может служить опыт Джозеф Райна, который в 1950 г. проводил эксперименты по выявлению людей с экстрасенсорными способностями. Каждый испытуемый должен был угадать цвет каждой карты в последовательности из 10 карт. В таком опыте вероятность случайно угадать 9 или 10 карт из 10 составляет $11 \cdot 2^{-10} \approx 0.01$ — достаточно редкое событие, и его можно признать маловероятным, поэтому Райн считал его свидетельством того, что испытуемый — экстрасенс. Однако, поскольку в опыте участвовало 1000 человек, вероятность того, что хотя бы один из них случайно угадает 9 или 10 карт составляла $1 - (1 - 11 \cdot 2^{-10})^{1000} \approx 0.9998$. Действительно, из 1000 испытуемых 12 человек угадали 9 из 10 карт, двое — все 10 карт. В последующих экспериментах ни один из них не подтвердил свои способности.

рия Стьюдента сравниваются между собой три группы, то парных сравнений будет: 1 со 2, 2 с 3, 3 с 1, итого — 3 сравнения. Следовательно, при уровне значимости равном 0,05 нулевая гипотеза будет отклоняться только в случае, когда $p < 0,05/3 = 0,016$.

Это важное замечание, которое необходимо учитывать при анализе статей в научных журналах. Большое количество авторов публикаций, так или иначе, обращаются к *критерию Стьюдента* в своих исследованиях и ошибочно могут выдать желаемое за действительное, особенно когда доказывают свою правоту статистическими методами. И если в ходе статистического анализа был неверно использован какой-либо из критериев значимости, то ошибка может накапливаться и в дальнейших расчетах. Поэтому необходимо достаточно четко осознавать области применения критериев и если критерий, был неверно применен, т. е. смысл оценивать выводы, описанные в статье более скептически.

5.2. Правила использования критерия Стьюдента:

Перечислим основные правила использования критерия Стьюдента:

- Критерий Стьюдента может быть использован только в случае выборок с нормально распределенными значениями признака.
- Критерий Стьюдента может быть использован для проверки гипотезы о различии средних только для двух групп.
- Желательно, что бы в обеих группах величина стандартного отклонения была приблизительно одинаковой (это оценивается с помощью критерия Левена).
- Если схема эксперимента предполагает большее число групп, рекомендуется воспользоваться дисперсионным анализом.
- Если критерий Стьюдента был использован для проверки различий между несколькими группами, то истинный уровень значимости можно получить, разделив уровень значимости на число возможных сравнений (поправка Бонферрони).

Общая последовательность действий при использовании *критерия Стьюдента*:

1. Формирование таблиц результатов измерения (записываются и организуются в таблицу исследуемые показатели первой группы и показатели второй группы).

Еще примеры. Пусть сравниваются способности нового и уже существующего лекарства влиять на ряд симптомов некоторого заболевания, причём, поскольку интерес представляют выводы отдельно по каждому симптому, проверяются покомпонентные гипотезы превосходства нового лекарства над старым. Чем больше список рассматриваемых симптомов, тем более вероятно, что новый препарат окажется лучше старого хотя бы в отношении одного из симптомов. Пусть проверяется безопасность нового лекарства путём проверки наличия большого количества побочных эффектов. Чем больше побочных эффектов рассматривается, тем более вероятно, что хотя бы один из них будет признан связанным с употреблением лекарства.

2. Выполняется анализ полученных данных (описательная статистика).
3. Проверка соответствия распределения данных нормальному закону распределения.
4. Выбирается значение *уровня значимости* (в зависимости от строгости исследования: 0,1 или 0,05 или 0,01).
5. Формулируется *нулевая гипотеза* (различие между группами незначимо или является следствием случайности).
6. Рассчитывается значение *критерия Стьюдента t* (значение критерия, начиная с которого мы отвергаем нулевую гипотезу) и *вероятность ошибочного результата P* (вероятность ошибочно отвергнуть верную нулевую гипотезу, т. е. найти различия там, где их нет).
7. На основании значения уровня значимости и количества элементов выборки определяется величина *критического значения критерия $t_{кр}$* (значение критерия, начиная с которого мы отвергаем нулевую гипотезу).
8. Сравниваются между собой t и $t_{кр}$, а также P и α , если $t > t_{кр}$ и $P < \alpha$, то нулевая гипотеза отвергается и различия между группами статистически значимы, в противном случае различия — случайны или незначимы. Следует учитывать, что даже при обнаруженной статистической значимости различий исследователь может ошибаться, но допустимая вероятность равна *уровню значимости*.

§6. Непараметрические критерии

Для сравнения двух или более групп, например для определения эффективности одного или нескольких методов лечения, используется **дисперсионный анализ**, в частности **критерий Стьюдента**. Эти критерии основаны на допущении, что наблюдаемый признак подчиняется **нормальному закону распределения**. Более того, для применимости этих методов требуется, чтобы сравниваемые совокупности имели приблизительно одинаковые дисперсии. Отличаться могут только средние значения. По их различию и судят о различии совокупностей.



Применяя тот или иной метод, нужно быть уверенным, что допущения, на которых он основан, выполняются хотя бы приближенно. Иначе велик риск, что, выполнив, казалось бы, правильную последовательность действий, в результате мы придем к ошибочным выводам.

Следует указать, что условия применимости *дисперсионного анализа и критерия Стьюдента* выполняются часто, но далеко не всегда. В одних случаях слишком велика разница дисперсий, в других распределение далеко от нормального. В такой ситуации следует воспользоваться так называемыми **непараметрическими методами**.



Непараметрические критерии — это критерии, основанные на **рангах**. Ранговые критерии не нуждаются в предположениях о типе распределения. Единственное требование состоит в том, чтобы тип распределения в сравниваемых совокупностях был приблизительно одинаковым. При этом не обязательно знать, с каким распределением мы имеем дело и каковы его параметры.

Эта особенность ранговых методов сделала данные критерии очень популярными в медицинских исследованиях. Когда исследователь ограничен небольшим объемом выборки о типе распределения судить довольно сложно. Некоторые авторы даже рекомендуют не прибегать к параметрическим критериям, а сразу же обращать свое внимание на непараметрические их аналоги.

Рассмотрим общие принципы непараметрических методов сравнения групп.

Непараметрические методы заменяют реальные значения признака **рангами**.

Каждому значению признака в группе присваивается свой ранг в зависимости от величины значения признака. Проиллюстрируем это на примере таблицы 20:

Таблица 20 — Результаты эксперимента с обозначенными рангами значений

Значения признака	5	10	17,2	17	18	21	21,8	25	24,6
Ранги	1	2	4	3	5	6	7	9	8

В таблице представлено задание рангов для выборки из девяти членов. Каждый непараметрический критерий работает с рангами по своему, поэтому мы не будем здесь подробно останавливаться. Перечислим этапы сравнения групп непараметрическими методами, минуя построение таблиц и первичный анализ данных.

1. Формируем *нулевую гипотезу* об отсутствии различий между сравниваемыми группами.
2. Задаем приемлемый *уровень значимости* для конкретного исследования (0,05 или какой-то другой).
3. Рассчитываем значение (эмпирическое значение) выбранного непараметрического критерия.
4. По таблице или при помощи программного обеспечения находим *критическое значение* выбранного критерия для заданного уровня значимости.
5. Если значение эмпирического критерия больше его критического значения, то различия между группами признаются значимыми.
6. Так же рассчитывается *вероятность ошибочного результата P*.
7. Если *P* оказывается меньше уровня значимости α , то различия между группами признаются статистически значимыми.

При использовании непараметрических критериев большая часть информации о распределении сохраняется, но нет необходимости знать, что это за распределение. Исследователя куда меньше интересуют параметры распределения, отпадает и необходимость равенства дисперсий.

Следует заметить, если выполняется условие нормальности распределения, параметрические критерии обеспечивают наибольшую чувствительность. И им стоит отдать предпочтение. Однако если же это условие не выполняется хотя бы приблизительно, их чувствительность существенно снижается, и непараметрические критерии дают больше шансов выявить реально существующие различия или признать их отсутствие.

Еще раз остановимся на нормальности распределения признака. Простейший способ выявить тип распределения состоит в том, чтобы построить график. Нарисовав график, посмотрите, насколько похож он на нормальное распределение:

- ✓ Похожа ли его форма на график нормального распределения.
- ✓ Симметричен ли он относительно среднего значения.
- ✓ Покрывают ли интервал, равный плюс-минус двум (трем) стандартным отклонениям от среднего практически все наблюдения?
- ✓ Сравните графики для разных групп.
- ✓ Близок ли разброс значений. Выясните, насколько близки по величине значения дисперсий в обеих группах.

Если ответы на все вопросы утвердительны (допускаются небольшие отклонения), то наилучшим вариантом будет воспользоваться **параметрическим критерием**. В противном случае следует использовать **непараметрический критерий**.

§7. Основные непараметрические критерии для сравнения групп

7.1. Сравнение двух выборок: критерий Манна — Уитни

U-критерий Манна — Уитни используется для оценки различий между двумя независимыми выборками (тип данных может быть либо *количественным*, либо *порядковым*), он является аналогом параметрического критерия Стьюдента. Однако этот критерий работает не с самими данными эксперимента, а с их рангами. Поэтому изначально данные эксперимента упорядочиваются, и каждому из них присваивается свой ранг. *U-критерий Манна — Уитни* рассчитывается следующим образом:

1. Данные обеих групп объединить и упорядочить по возрастанию. Ранг 1 присвоить наименьшему из всех значений, ранг 2 — следующему и так далее. Наибольший ранг присвоить самому большому среди значений в обеих группах. Если значения совпадают, им присваивают один и тот же средний ранг (например, если два значения поделили 3-е и 4-е места, обоим присваивают ранг 3,5). Общее количество рангов в этом случае равно $N = n_1 + n_2$, где n_1 и n_2 — количество членов в 1 и 2 выборках соответственно.

2. Поделить полученный ряд на два, состоящих соответственно из единиц первой и второй выборок. Подсчитать отдельно сумму рангов, пришедшихся на долю элементов первой выборки, и отдельно — на долю элементов второй выборки.

3. Определить большую из двух ранговых сумм (T_x), соответствующую выборке с n_x элементами.

4. Рассчитать *U-критерий Манна — Уитни* по формуле:

$$U = n_1 \cdot n_2 + \frac{n_x \cdot (n_x + 1)}{2} - T_x$$

5. Полученное значение U сравнивают с критическими значениями $U_{кр}$ для данных n_1 и n_2 . *Критические значения критерия Манна — Уитни* приведены в таблице его **критических значений**. Столбец критических значений содержит пары чисел. Различия статистически значимы, если U не больше первого из них или не меньше второго. Например, когда в одной группе 3 человека, а в другой 6, различия статистически значимы, если $U \leq 7$ или $U \geq 23$.

6. Если U меньше или равно первому из них либо больше или равно второму, то нулевая гипотеза отвергается (различия статистически значимы).

Критические значения критерия *U Манна — Уитни* для $\alpha = 0,05$ приведены в таблице 21:

Таблица 21 — Таблица критических значений критерия Манна — Уитни

n2	n1													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Ограничения применимости критерия:

- Используется для сравнения только количественных или порядковых типов данных. Использование критерия для сравнения групп с качественными признаками недопустимо.

- В каждой из выборок должно быть не менее 3 значений признака. Допускается, чтобы в одной выборке было два значения, но во второй тогда не менее пяти.

- В выборочных данных не должно быть совпадающих значений (все числа — разные) или таких совпадений должно быть очень мало.

- Если необходимо сравнить несколько групп попарно, то рекомендуется воспользоваться поправкой Бонферрони.

7.2. Сравнение наблюдений «до и после»: критерий Уилкоксона

Если необходимо сравнить результаты измерений, сделанные у одной и той же группы, но в разные моменты времени, например, в случае до эксперимента и после него, то можно применить **критерий Уилкоксона** для сравнения двух групп.

Критерий Уилкоксона (также используются названия *T-критерий Уилкоксона*, *критерий Вилкоксона* и т. п.) — непараметрический статистический критерий, используемый для сравнения двух связанных (парных) выборок с количественными признаками, измеренными в непрерывной или в порядковой шкале.



То есть **T-критерий Уилкоксона** используется для оценки различий между двумя рядами измерений, выполненных для одной и той же совокупности исследуемых, но в разных условиях или в разное время.

Классическим примером ситуации, в которой может применяться *T-критерий Уилкоксона* для связанных совокупностей, является исследование «до-после», когда сравниваются показатели до и после лечения. Например, при изучении эффективности антигипертензивного средства сравнивается артериальное давление до приема препарата и после приема. При этом изучаемый признак может быть измерен как в количественной непрерывной, так и в порядковой шкале.

Принцип критерия следующий. Для каждого члена выборки вычисляют величину изменения признака. Все изменения упорядочивают по абсолютной величине (без учета знака). Затем рангам приписывают знак изменения и суммируют эти «знаковые ранги» — в результате получается значение *T-критерия Уилкоксона*. Как видно, используется информация об абсолютной величине изменения и его знаке (т. е. уменьшении или увеличении наблюдаемого признака). Как в случае с *критерием Манна — Уитни*, здесь также можно перечислить все возможные величины *T* и найти критическое значение. Исходно ранги присваиваются в соответствии с абсолютной величиной изменения.

Последовательность шагов:

1. Вычислить величины изменений наблюдаемого признака. Отбросить пары наблюдений, которым соответствует нулевое изменение.

2. Упорядочить изменения по возрастанию их абсолютной величины и присвоить соответствующие ранги. Рангами одинаковых величин назначить средние тех мест, которые они делят в упорядоченном ряду.

3. Присвоить каждому рангу знак в соответствии с направлением изменения: если значение увеличилось — «+», если уменьшилось — «-».

4. Вычислить сумму рангов с присвоенными им знаками — *T-критерий Уилкоксона*²³.

5. Сравнить полученную величину *T* с критическим значением. Если она больше критического значения, изменение показателя статистически значимо.

Фрагмент таблицы критических значений *T-критерия Уилкоксона* приведен ниже (таблица 22):

Таблица 22 — Таблица критических значений критерия Уилкоксона

Число исследуемых, n	p=0.05	p=0.01
5	0	—
6	2	—
7	3	0
8	5	1
9	8	3
10	10	5
11	13	7
12	17	9
13	21	12
14	25	15
15	30	19
16	35	23
17	41	27
18	47	32
19	53	37
20	60	43
21	67	49
22	75	55

Условия и ограничения применения *T-критерия Уилкоксона*:

- Число исследуемых при использовании *T-критерия Уилкоксона* должно быть не менее 5 и не более 50.

- Данный критерий используется только в случае сравнения двух вариационных рядов (двух связанных групп). Аналогом *T-критерия Уилкоксона* для сравнения трех и более связанных совокупностей является **Критерий Фридмана**.

²³ Существует вариант *критерия Уилкоксона*, в котором суммируют только положительные или только отрицательные знаковые ранги. На выводе это никак не сказывается, однако значение *T*, естественно, получается другим. Поэтому важно знать, на какой вариант критерия рассчитана имеющаяся в вашем распоряжении таблица критических значений.

7.3. Сравнение нескольких групп: критерий Крускала — Уоллиса

Непараметрическим аналогом *критерия Фишера* в дисперсионном анализе, но не требующим предположения о нормальности распределения, является **критерий Крускала — Уоллиса** (Краскела — Уоллиса).

Критерий Крускала — Уоллиса представляет собой обобщение критерия Манна — Уитни.

Последовательность анализа:

1. Все значения, независимо от того, какой выборке они принадлежат, упорядочивают по возрастанию. Каждому значению присваивается ранг — номер его места в упорядоченном ряду. (Совпадающим значениям присваивают общий ранг, равный среднему тех мест, которые эти величины делят между собой в общем упорядоченном ряду).

2. Далее найти суммы рангов R для каждой группы i отдельно. При отсутствии межгрупповых различий средние ранги групп должны оказаться близки. Напротив, если существует значительное расхождение средних рангов, то гипотезу об отсутствии межгрупповых различий следует отвергнуть. Значение **критерия Крускала — Уоллиса H** и является мерой такого расхождения средних рангов.

3. Вычислить критерий **Крускала — Уоллиса H** по формуле

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

где n — общее количество испытуемых (сумма по всем выборкам); R_i — сумма рангов для i -й выборки; n_i — количество членов i -й выборки.

4. Сравнить вычисленное значение H с найденным по таблице **критическим значением $H_{кр}$** для числа степеней свободы, на единицу меньшего числа групп. Если вычисленное значение H окажется больше **критического**, различия групп статистически значимы.

Ограничения применимости критерия:

- Используется для сравнения только количественных или порядковых типов данных. Использование критерия для сравнения групп с качественными признаками недопустимо.

- Для того, чтобы диагностировать различия на уровне значимости больше, чем $0,05$, необходимо, чтобы в каждой выборке было не менее 3 наблюдений, или чтобы хотя бы в одной из них было 4 наблюдения, а в двух других — по 2; при этом неважно, в какой именно выборке сколько испытуемых, а важно именно само соотношение 4:2:2.

- При сравнении групп числом больше чем две, критерий указывает только на наличие различий между группами, но не указывает между какими группами было оно выявлено. Что бы понять какие группы различаются между собой необходимо провести сравнения попарно, используя критерий Манна — Уитни с поправкой Бонферрони.

7.4. Повторные измерения нескольких групп: критерий Фридмана

Если необходимо сравнить несколько групп в случае «до и после», например одна и та же группа испытуемых последовательно подвергается нескольким методам лечения или просто наблюдается в разные моменты времени, применяют **дисперсионный анализ повторных измерений**²⁴. Но как вы помните, для возможности использовать дисперсионный анализ группы должны иметь нормальное распределение признака. Если данное условие не выполняется, то лучше воспользоваться непараметрическим вариантом критерия Фишера для повторных измерений — **критерием Фридмана**.

Логика критерия Фридмана довольно проста. Каждый больной ровно один раз подвергается каждому методу лечения (или наблюдается в фиксированные моменты времени). Результаты наблюдений у каждого больного упорядочиваются. Получается столько упорядоченных рядов, сколько имеется членов в выборке. Далее для каждого момента наблюдения вычисляется сумма рангов. Если разброс сумм велик — различия статистически значимы.

Порядок расчета **критерия Фридмана** рассмотрим на примере следующей таблицы²⁵ (таблица 23):

Пять испытуемых исследуются по четырём тестам.

Таблица 23 — Результат тестирования

Номер испытуемого	Оценки в баллах по проведённым тестам			
	Тест А	Тест В	Тест С	Тест D
1	3.6	4.1	2.9	3.5
2	3.8	4.2	3.7	4.6
3	3.3	3.8	3	3.7
4	3.8	3.3	3.4	2.7
5	4	3.6	1.9	3.1

Расположим значения для каждого члена выборки (испытуемого) по возрастанию, каждому значению присваиваем ранг.

Для каждого из моментов наблюдения (методов лечения или в нашем примере для каждого отдельного теста) подсчитаем сумму присвоенных ему рангов.

В результате получаем (таблица 24):

Таблица 24 — Значения рангов для результатов исследования

Номер испытуемого	Ранги тестов (по строкам)			
	Тест А	Тест В	Тест С	Тест D
1	2	1	4	3
2	3	2	4	1
3	3	1	4	2
4	1	3	2	4
5	1	2	4	3
Сумма рангов:	10	9	18	13

²⁴ Не рассматривается в рамках данного курса.

²⁵ Пример взят с сайта <http://matstats.ru/>.

Значение $\chi_{эмн}^2$ вычисляем по формуле:

$$\chi_{эмн}^2 = \frac{12}{n(c+1)c} \cdot \sum_j T_j^2 - 3n(c+1),$$

где c — количество моментов наблюдений; n — размер выборки, T_j — сумма рангов по j -му моменту наблюдения.

Определяем **критическое значение** $\chi_{кр}^2$, зависящее от *уровня значимости* α и *степени свободы* $k=c-1$ по таблице критических значений.

Если рассчитанное значение $\chi_{эмн}^2$ превышает **критическое** $\chi_{кр}^2$ — различия статистически значимы.

7.5. Выводы по анализу сравнения групп

В предыдущих разделах мы рассмотрели основные критерии, используемые для сравнения групп. Непараметрические критерии построены на том же принципе что и их параметрические аналоги, но могут быть использованы для сравнения групп с распределением отличным от нормального. Важно помнить, что параметрические критерии (**критерий Фишера** и **критерий Стьюдента**) используются только при выполнении некоторых условий: *нормальность распределения признака, приблизительно равное значение дисперсий* в группах. Если эти условия не выполняются, следует выбрать непараметрический аналог данных критериев²⁶.

Перечислим основные этапы статистического сравнения групп:

1. Формулируется *нулевая гипотеза*, т. е. предполагаем, что наблюдаемые различия случайны.
2. Выбирается критерий, наиболее подходящий для решения задачи.
3. Рассчитывается значение выбранного критерия.
4. По таблице находится критическое значение критерия для заданного уровня значимости (α).
5. Найденное *значение критерия* для наших данных сравнивается с *критическим значением*: если вычисленное значение больше, признаем различия статистически значимыми.

²⁶ Выбор между параметрическими и непараметрическими методами определяется прежде всего характером данных. Имея дело с порядковыми признаками, не остается ничего, кроме как воспользоваться непараметрическими методами. Тут могут помочь как общие соображения, так и графическое представление данных. Даже если нет веских оснований сомневаться в нормальности распределения, но данных мало, или вы не хотите делать никаких предположений о типе распределения — воспользуйтесь непараметрическими методами.

ГЛАВА V. АНАЛИЗ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

§1. Общие понятия

Статистические процедуры, с которыми вы познакомились в предыдущих главах, предназначены для анализа количественных признаков (некоторые из них также подходили и для признаков порядковых). Над значениями количественных признаков можно производить арифметические операции. Говоря об этих признаках можно делать замечание, что их значения изменились на какую-то величину или среднее значение признака в одной группе больше среднего значения в другой. Эти признаки можно упорядочить, расположив их в порядке возрастания или убывания. Но, как вам известно, есть признаки, которые нельзя выразить в числовой форме. Например, можно быть мужчиной либо женщиной, больным либо здоровым. Можно быть инженером, врачом, юристом, рабочим и так далее. Здесь мы имеем дело с неким *качеством* исследуемого объекта и признаки, характеризующие объект таким образом, называются **качественными признаками**. Эти признаки не связаны между собой никакими арифметическими соотношениями (их нельзя сложить, поделить, умножить, сказать на сколько один признак больше другого и т. д.), их также нельзя упорядочить.



Единственный способ описания качественных признаков состоит в том, чтобы подсчитать число объектов, имеющих одно и то же значение. Кроме того, можно подсчитать, какая **доля** от общего числа объектов приходится на то или иное значение.

Напомним, что существует еще один вид признаков. Это порядковые признаки. Их можно упорядочить, но производить над ними арифметические действия также нельзя. Пример порядкового признака — состояние больного тяжелое, средней тяжести, удовлетворительное.

§2. Статистические процедуры над качественными признаками

2.1. Первичный анализ

Рассмотрим более подробно анализ данных с качественными признаками.

Если часть объектов исследуемой группы характеризуется одним значением признака, а вторая часть другим, то можно подсчитать какую **долю** (p) или **процент от общего количества** объектов в группе составляют объекты той или иной группы.

Другими словами, если известно общее число членов группы N с признаком M , то долю p этих членов можно выразить формулой:

$$p = M/N$$

или в процентном соотношении

$$p = (M/N) \times 100 \%$$

Например, если в группе из 100 человек 30 человек — женщины, а 70 — мужчины, то доля p (процент) женщин в группе равен $30/100 = 0,3$ или 30 %, соответственно мужчин — 70 %. Разумеется, группы могут состоять не только из двух классов (значений признака).



В некотором смысле доля p аналогична среднему μ по совокупности. При условии, что исследуемая совокупность состоит только из двух классов, достаточно указать численность одного из них, тогда если доля одного класса во всей совокупности равна p , то доля другого равна $1 - p$.

Помимо вычисления долей часто строятся различного рода графики и диаграммы, иллюстрирующие соотношение долей, например круговая диаграмма (рисунок 47):

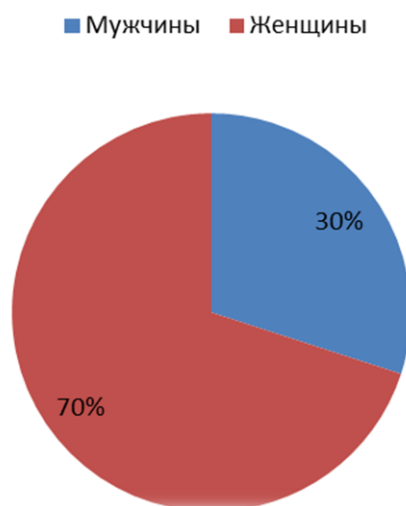


Рисунок 47 — Диаграмма процентного соотношения мужчин и женщин в группе

В отчете так же предоставляется значение моды в исследуемой группе, т. е. то значение исследуемого признака, которое встречалось наибольшее количество раз в группе.

Собственно к этим простым процедурам и сводится *первичный анализ* качественных признаков.

2.2. Сравнение двух групп с качественными признаками

Как и в случае с количественными признаками на первичном анализе собранных качественных данных статистический анализ в исследованиях не заканчивается. Довольно часто необходимо сравнить две группы, характеризующиеся качественным признаком между собой. Для проведения этой процедуры используется **z-критерий**. Принцип сравнения групп с качественными признаками в целом схож со сравнением групп с количественными признаками. Первоначально формулируется *нулевая гипотеза* об отсутствии различий между группами и задается приемлемый *уровень значимости* α . Далее *находится эмпирическое значение подходящего критерия значимости* и сравнивается с его *критическим значением*, взятым из таблицы критических значений для заданного *уровня значимости*. Если эмпирический критерий *больше* критического, то различия между группами признаются *статистически значимыми*, если наоборот — *незначимыми*. Также рассчитывается *вероятность ошибки P* и сравнивается с заданным *уровнем значимости*, если P меньше α , то различия значимы.

Рассмотрим *z-критерий* для сравнения двух групп. Этот критерий аналогичен *критерию Стьюдента t* для количественных признаков и рассчитывается по формуле:

$$z = \frac{\text{Разность выборочных долей}}{\text{Стандартная ошибка разности выборочных долей}}$$

или

$$z = \frac{p_1 - p_2}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{p_1 - p_2}{\sqrt{s_{\hat{p}_1}^2 + s_{\hat{p}_2}^2}},$$

где p_1 и p_2 — доля исследуемого признака в первой группе и во второй соответственно.

Если n_1 и n_2 — объемы двух выборок, то

$$s_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} \quad \text{и} \quad s_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} ;$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

О статистически значимом различии долей можно говорить, если значение z окажется «большим»²⁷.

²⁷ Не вдаваясь в подробное описание причин, отметим, что чаще используют не выше-приведенную формулу для *z-критерия*, а учитывают еще **поправку Йейтса на непрерывность**. Это связано с излишним «оптимизмом» *критерия z*. С учетом этой поправки формула для z имеет следующий вид: $z = \frac{|p_1 - p_2| - 1/2(1/n_1 + 1/n_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$.

Критическое значение z -критерия находится по таблице критических значений в зависимости от количества членов выборки (вычисляется *число степеней свободы*) и выбранного *уровня значимости*. Однако если сравнивать с критическим значением критерия Стьюдента, то $t_{кр}$ подчиняется *распределению Стьюдента*, а $z_{кр}$ — *стандартному нормальному распределению*. При увеличении числа степеней свободы распределение Стьюдента стремится к *нормальному*, критические значения z можно найти в последней строке таблицы критических значений для *критерия Стьюдента*.

Сравнив полученное значение z -критерия с его критическим значением, делается вывод о статистической значимости различий между группами. Если критическое значение больше эмпирического, то группы признаются различающимися.



Необходимо уточнить, что если выборки небольшие, то *критерий z неприменим*, и нужно воспользоваться **точным критерием Фишера**.

2.3. Таблицы сопряженности: критерий χ^2 для таблицы 2×2

Довольно часто исследуемый объект может иметь *несколько качественных признаков*, например: человек болен или здоров и при этом принял или не принял соответствующее лекарство. В этом случае данные эксперимента сводятся в так называемые **таблицы сопряженности**.

Общий вид *таблицы сопряженности* представлен ниже (таблица 25):

Таблица 25 — Общий вид таблицы сопряженности

	Исход есть (1)	Исхода нет (0)	Всего
Фактор риска есть (1)	A	B	A + B
Фактор риска отсутствует (0)	C	D	C + D
Всего	A + C	B + D	A + B + C + D

Пример заполненной *таблицы сопряженности* (таблица 26):

Таблица 26 — Таблица сопряженности по данным исследования вакцинации

Прививка (фактор риска)	Заболел (исход)	
	да	нет
да	5	18
нет	15	8



Таблица сопряженности — представляет собой таблицу m на n , где m — значения, принимаемые первой переменной, n — значения, принимаемые второй переменной.

Иногда содержимое таблицы сопряженности представляется в процентном соотношении относительно общего числа испытуемых (таблица 27):

Таблица 27 — Таблица сопряженности с рассчитанными долями

Прививка (фактор риска)	Заболел (исход)	
	да	нет
да	10,9 %	39,1 %
нет	32,6 %	17,4 %

В таблице выше $m = 2$ (заболел: да или нет), $n = 2$ (прививка: да или нет). Такие таблицы называются **таблицами сопряженности 2×2** .

Цифры в ячейках таблицы указывают на количество членов выборки с соответствующими значениями признака. То есть из примера таблицы следует, в результате эксперимента были получены следующие данные: количество заболевших наблюдаемых составило 20 человек, из них пятерым была сделана прививка, а 15 — нет; не заболевших — 26, из них 18-ти была сделана прививка, а 8-ми — нет.

Помимо более наглядной записи результатов эксперимента таблицы сопряженности можно использовать и для решения некоторых задач. Например, выше приведенная таблица позволяет оценить влияет ли наличие прививки на факт возникновения заболевания. Для этого случайным образом были отобраны пациенты, одним из которых была сделана прививка, а другим — нет. Спустя некоторое время собираются данные о случаях заболевания у всех членов эксперимента и заносятся в таблицу результатов исследования. Далее строится *таблица сопряженности*, которая позволяет оценить влияние исследуемого фактора (*фактор риска*) — наличие прививки в выше приведенных таблицах — на возникновение *исхода* (заболевание в приведенных таблицах).

Для понимания смысла анализа таблиц сопряженности необходимо учитывать, что если исследуемый фактор не оказывает воздействия (например, прививка не влияет на наличие заболевания), то значения в ячейках таблицы сопряженности будут примерно одинаковыми (исходя из выше приведенного примера, количество заболевших не будет зависеть от наличия или отсутствия прививки)²⁸. Это будет наблюдаться при верности нулевой гипотезы. Если же цифры в ячейках таблицы будут существенно отличаться, то вероятно исследуемый фактор оказывает воздействие. В этом случае необходимо проверить насколько значимо влияние этого фактора.

Поэтому проверка начинается с того, что строится еще одна таблица, в которую записываются приблизительные значения, как если бы была верна нулевая гипотеза. То есть как если бы исследуемый фактор не оказы-

²⁸ При условии, что прививку сделали половине испытуемых. Более подробно это можно описать, используя следующий пример: пусть в комнате находится 100 человек, из которых 30 являются мужчинами, а 70 — женщинами. Если известно, что из этих 100 человек 10 увлекаются искусством, то в случае, если увлечение не зависит от пола, мы будем ожидать, что из 10 увлекающихся искусством 3 являются мужчинами, а 7 — женщинами. Сопоставляя эти ожидаемые частоты с наблюдаемыми частотами, мы можем судить о том, действительно ли два номинальных признака не связаны.

вал бы никакого влияния на исход события, другими словами, между переменными не наблюдается какой-либо взаимосвязи. Эти рассчитанные значения называются **ожидаемыми значениями**. А значения в таблице, построенной по данным эксперимента — **наблюдаемые значения**.

Расчет *ожидаемых значений* для каждой из ячеек таблицы сопряженности (при условии справедливости *нулевой гипотезы об отсутствии взаимосвязи*) производится путем перемножения сумм строк и столбцов с последующим делением полученного произведения на общее число наблюдений. Общий вид таблицы **ожидаемых значений** представлен ниже (таблица 28):

Таблица 28 — Расчет ожидаемых значений

	Исход есть (1)	Исхода нет (0)	Всего
Фактор риска есть (1)	$(A+B)*(A+C) / (A+B+C+D)$	$(A+B)*(B+D) / (A+B+C+D)$	A + B
Фактор риска отсутствует (0)	$(C+D)*(A+C) / (A+B+C+D)$	$(C+D)*(B+D) / (A+B+C+D)$	C + D
Всего	A + C	B + D	A+B+C+D

В случае нашего примера (таблица 29):

Таблица 29 — Ожидаемые значения для примера с исследованием вакцинации

Прививка	Заболел	
	да	нет
да	10	13
нет	10	13

На сравнении *ожидаемых* и *наблюдаемых* значений и основан **критерий χ^2 (критерий согласия Пирсона)**, который и является *критерием значимости* для анализа таблиц сопряженности.

Критерий χ^2 (читается «хи-квадрат») не требует никаких предположений относительно параметров совокупности, из которой извлечены выборки, — это *непараметрический критерий*.

Определяется критерий следующим образом²⁹:

$$\chi^2 = \sum \frac{(H-O)^2}{O},$$

где *H* — наблюдаемое число в клетке таблицы сопряженности; *O* — ожидаемое число в той же клетке. Суммирование проводится по всем клеткам таблицы.

²⁹ Приведенная формула для χ^2 в случае таблицы 2×2 (то есть при 1 степени свободы) дает несколько завышенные значения, т.е. повышается вероятность совершить *ошибку I рода*. Это вызвано тем, что теоретическое распределение χ^2 непрерывно, тогда как набор вычисленных значений χ^2 дискретен. На практике это приведет к тому, что нулевая гипотеза будет отвергаться слишком часто. Чтобы компенсировать этот эффект, в формулу вводят **поправку Йейтса**:

$$\chi^2 = \sum \frac{(H-O-\frac{1}{2})^2}{O}. \text{ Поправка Йейтса применяется только при } \nu = 1, \text{ т. е. для таблиц } 2 \times 2.$$

Нулевая гипотеза в подобных задачах звучит следующим образом: переменные не связаны между собой, т.е. являются независимыми и исследуемый фактор (фактор риска) не оказывает значимого влияния на исход события (нет значимой взаимосвязи между фактором риска и исходом), видимые различия в клетках таблицы сопряженности случайны и статистически незначимы.

Полученное значение критерия χ^2 характеризует значимость этих различий. Расчетное значение критерия сравнивается с его **критическим значением критерия $\chi^2_{кр}$** (значение находится по таблице для заданного уровня значимости) и если полученное значение критерия χ^2 больше критического, то нулевая гипотеза об отсутствии связи между переменными отклоняется.

Применение критерия χ^2 правомерно, если ожидаемое число в любой из клеток больше или равно 5 (в противном случае мы вынуждены **использовать точный критерий Фишера**). Это условие аналогично условию применимости критерия z . Критическое значение χ^2 зависит от размеров *таблицы сопряженности*, то есть от числа строк таблицы и числа возможных столбцов таблицы. Размер таблицы выражается числом степеней свободы v :

$$v = (r - 1)(c - 1),$$

где r — число строк, а c — число столбцов.



При работе с таблицами большего чем 2×2 размера критерий χ^2 применим, если все ожидаемые числа не меньше 1 и доля клеток с ожидаемыми числами меньше 5 не превышает 20 %. При невыполнении этих условий критерий χ^2 может дать ложные результаты. В этой ситуации есть выход: можно собрать дополнительные данные, однако это не всегда осуществимо или объединить несколько строк или столбцов, что тоже может противоречить условиям эксперимента.

Для расчета критерия χ^2 можно воспользоваться как специализированным программным обеспечением, так и *онлайн-калькуляторами*, примером такого калькулятора может являться калькулятор на сайте medstatistic.ru, находящийся на страницах <http://medstatistic.ru/calculators/calchi.html> — для расчета таблиц сопряженности 2 на 2 и <http://medstatistic.ru/calculators/calchit.html> — для расчета критерия χ^2 произвольных таблиц сопряженности.

Приведем общую последовательность действий анализа данных с несколькими качественными признаками:

1. По имеющимся данным построить *таблицу сопряженности*.
2. Подсчитать число объектов в каждой строке и в каждом столбце, найти, какую *долю* от общего числа объектов составляют эти величины.

3. Зная эти доли, подсчитать с точностью до двух знаков после запятой *ожидаемые числа* — количество объектов, которое попало бы в каждую клетку таблицы, если бы связь между строками и столбцами отсутствовала (как если бы не было значимой взаимосвязи между фактором риска и исходом).

4. Найти величину, характеризующую различия наблюдаемых и ожидаемых значений — рассчитать критерий χ^2 . Если таблица сопряженности имеет размер 2×2 , желательно применить *поправку Йеутса*.

5. Вычислить число *степеней свободы*, и выбрать *уровень значимости*.

6. Определить критическое значение χ^2 . Сравнить полученное значение с *критическим* значением.

7. Если полученное значение критерия χ^2 больше *критического*, то нулевая гипотеза об отсутствии связи между переменными отклоняется, в противном случае мы остаемся в рамках нулевой гипотезы.

Стоит заметить, что помимо критерия χ^2 часто рассчитывают **коэффициент ϕ** (фи) и **коэффициент Крамера** — V , которые показывают величину силы взаимосвязи между рассматриваемыми переменными. При этом V — Крамера принимает значения от 0 до 1 и служит характеристикой силы связи между переменными.

ГЛАВА VI. ПОНЯТИЕ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА

§1. Доверительный интервал

1.1. Определение

В главе, посвященной сравнению групп, мы познакомились с *критерием Стьюдента*, с его помощью мы сравнивали две группы между собой, однако данный критерий используется и во многих других случаях. В данной главе с помощью данного критерия мы можем оценить еще один важный статистический термин, используемый в статистике в качестве интервальной оценки статистических параметров — **доверительный интервал**.



Большинство статистических параметров совокупности оценивается посредством выборочных данных и, как следствие, используется только для **приблизительной** оценки какой-либо характеристики изучаемой совокупности, например, средний доход семьи, средний возраст заболевших и т. д. Выборочные данные позволяют сделать только **предположение** о реальном состоянии дел. Возникает вопрос на сколько качественное это предположение?



Конечно же самым качественным предположением было бы *полное отсутствие предположения* — т. е., если бы мы могли точно определить необходимые нам параметры. Но определить точное значение параметров, без исследования всей совокупности, невозможно, так как в большинстве случаев это была бы очень сложная и дорогостоящая задача.

Часто говорят: «*Быть статистиком — значит никогда не говорить, что ты уверен. Главное — просто подобраться поближе к истине*». Конечно хочется быть уверенным, что полученные результаты посредством анализа выборочных данных как можно точнее отражают действительность, ведь на исследования были затрачены и средства и время. Однако, если мы будем утверждать, что полученные нами результаты точно характеризуют всю совокупность, то этим самым мы вводим в заблуждение тех, кому сообщаем результаты наших исследований, поэтому правильнее было бы сообщать не «точное» значение параметра, а его значение *плюс/минус какая-то величина погрешности (предел погрешности)*, воспользовавшись данными выборки.



Статистический показатель плюс/минус предел погрешности — это и есть **доверительный интервал**. **Доверительный интервал** — интервал, который покрывает неизвестный параметр с заданной надёжностью (заданной точностью). Смысл доверительного интервала достаточно ясен: мы не знаем точно, чему равна некоторая величина, но можем указать интервал, в котором она находится (с заданной вероятностью).

Приведем пример применения доверительных интервалов (по книге С. Гланц Медико-биологическая статистика).

1.2. Пример использования доверительного интервала.

Рассмотрим типичную задачу. В ней исследуется препарат диуретик, вся *генеральная совокупность* состоит из 200 человек. Одной группе дается препарат, второй — плацебо.

Средний диурез при приеме плацебо составил $\mu_n = 1200$ мл, при приеме диуретика — $\mu_d = 1400$ мл. Таким образом, препарат увеличивает суточный диурез на $\mu_d - \mu_n = 1400 - 1200 = 200$ мл. Это есть *точное значение*, так как исследовалась вся генеральная совокупность.

Однако если мы воспользуемся *выборками* для оценки величины эффекта, в каждую из которых входило по 10 человек, то в этом случае данные эксперимента будут следующими: плацебо (контрольная группа) средний диурез составил 1180 мл, а в группе, получавшей диуретик, —

1400 мл. Среднее увеличение диуреза в данном опыте: 220 мл. Как и всякая выборочная оценка, эта величина отличается от истинного увеличения суточного диуреза, как мы выяснили равного 200 мл. И если бы мы, основываясь на выборочных данных, сказали, что препарат увеличивает суточный диурез в среднем на 220 мл, то упустили бы из виду неопределенность, присущую выборочной оценке и ко всему прочему несколько завысили результат.

В этом случае правильнее будет рассчитать **доверительный интервал** — он покажет не одно число, скорее всего не совпадающее с истинным, а диапазон чисел, куда истинное попадает почти наверняка (например, с вероятностью 95 %).

1.3. Расчет ДИ

Приведем формулу вычисления доверительного интервала для среднего значения по совокупности:

$$X - t_{0,05} \cdot \frac{s}{\sqrt{n}} < \mu < X + t_{0,05} \cdot \frac{s}{\sqrt{n}},$$

где X — среднее значение по выборке; $t_{0,05}$ — значение критерия Стьюдента для заданного *уровня значимости* α и *степени свободы* $\nu = n - 1$ (n — количество членов выборки); а s — стандартное отклонение, подсчитанное по выборке.

1.4. Степень надежности ДИ

Доверительные интервалы используются довольно часто. Но не стоит забывать, что их значение также находится по выборочным данным, а это значит, что истинный оцениваемый параметр может попасть в данный интервал, а может и не попасть в некотором проценте случаев.



Степень надежности доверительного интервала характеризует то, с какой вероятностью найденный доверительный интервал покрывает истинное значение оцениваемого параметра.

Обычно доверительные интервалы рассчитывают с 99, 95 или 90 % надежностью.



Таким образом, получив 95 % доверительный интервал можно сказать, что выборочные данные позволяют с 95 % надежностью утверждать, что истинное значение параметра находится в пределах заданного интервала. Иначе говоря, быть на 95 % уверенным — это значит, что если вы сделаете много выборок и каждый раз, исходя из результатов, определите доверительный интервал, то в 95 % случаев полученные доверительные интервалы попадут точно в цель, т. е. будут действительно отражать истинный параметр.

1.5. Ширина ДИ

Расстояние от нижней границы до верхней границы доверительного интервала называется **шириной доверительного интервала**.

Исходя из формулы приведенной выше, ширину доверительного интервала для среднего значения можно записать как:

$$\left(X - t_{0,05} \cdot \frac{s}{\sqrt{n}} ; X + t_{0,05} \cdot \frac{s}{\sqrt{n}} \right).$$

Понятно, что для более точной оценки параметра желательно, чтобы доверительный интервал был как можно уже. Однако это получается далеко не всегда. Как добиться того, чтобы доверительный интервал был достаточно узким? Обдумать этот вопрос необходимо до того, как собирать данные, так как после окончания сбора данных ширина доверительного интервала уже установлена. Поэтому укажем факторы, влияющие на ширину интервала:

- Размер выборки (чем больше размер выборки, тем точнее и уже доверительный интервал).
- Степень изменчивости в генеральной совокупности (чем ниже степень изменчивости, тем выше возможность более точно оценить параметр).
- Уровень доверия (99 % или 95 % этот параметр можно задать уже при расчете интервала, если уровень доверия велик, то доверительный интервал почти наверняка содержит истинное значение рассматриваемого параметра).

1.6. Область применения ДИ

В заключении перечислим наиболее часто рассчитываемые доверительные интервалы:

- среднего значения;
- для разности средних значений;
- доли;
- разности долей;
- значений;
- линии регрессии.

Принципы расчета каждого из них схожи между собой, однако формулы несколько отличаются. Пример доверительного интервала для линии регрессии мы рассмотрим в следующей главе.

ГЛАВА VII. АНАЛИЗ СВЯЗЕЙ

§1. Корреляционный и регрессионный анализ

1.1. Постановка проблемы

Часто исследователя интересует возможность предсказать поведение (значение) одной переменной *в зависимости* от значения другой. В процессе исследования можно заметить, что изменчивость одного признака находится в некотором соответствии с изменчивостью другого. Определив характер зависимости, исследователь в процессе дальнейшего практического применения данных эксперимента опирается уже не на свой личный опыт или опыт своих коллег, или некое интуитивное знание, но *на знание характера зависимости*, подтвержденное методами статистического анализа, результаты которого сведены в таблицы или представлены с помощью наглядных графиков.



Зная характер зависимости между переменными, можно с заданной степенью точности (или в пределах рассчитанного доверительного интервала) предсказывать поведение одной переменной (зависимой) при изменении значения другой (независимой).

В популярных научных статьях и заметках заключения, сделанные на основе обнаруженной зависимости, могут быть представлены следующим образом:

- Сидячие виды деятельности связаны с увеличением веса и возрастанием риска диабета у женщин.
- Выражение гнева может иметь обратную взаимосвязь с риском сердечного приступа или удара (те, кто не сдерживают свой гнев, рискуют меньше).
- У мужчин потребление спиртного в умеренных дозах снижает риск сердечных заболеваний.
- Незамедлительное лечение позволяет приостановить развитие глаукомы.
- Или предположение, что аспирин может предотвратить образование полипов у больных раком толстой кишки.

За многими из этих выводов могут стоять большое количество проанализированных данных эксперимента, длительные наблюдения или сложные клинические испытания. Однако, выявление и измерение связи между признаками, характеризующими изучаемые явления или процессы, является важнейшей частью исследования, а зачастую и его целью.

Различают **функциональную** и **корреляционную** связи. Для лучшего понимания дальнейшего материала коротко рассмотрим первый тип связи.

1.2. Виды связей между переменными. Функциональная связь

При наличии **функциональной связи** изменение величины одного признака неизбежно вызывает совершенно *определенные* изменения величины другого признака.

Примером такой связи может служить зависимость площади круга от его радиуса, при различных значениях радиуса площадь так же принимает различные значения, при этом значение площади будет *неизменным*, если радиус круга имеет *одно и то же* значение. Так же мы со стопроцентной точностью можем говорить, что если радиус примет определенное значение, то площадь круга примет строго определенное значение и никакое иное. Еще примеры: площадь треугольника точно определяется его высотой и основанием, скорость падения есть функция времени падения и ускорения силы тяжести, скорость протекания определенной химической реакции находится в зависимости от температуры.



Можно сказать, что функциональная связь может быть выражена посредством точной математической формулы.

На приведенном ниже графике представлена зависимость переменной y от переменной x , выраженная формулой $y = \frac{x^2}{2}$ (рисунок 48).

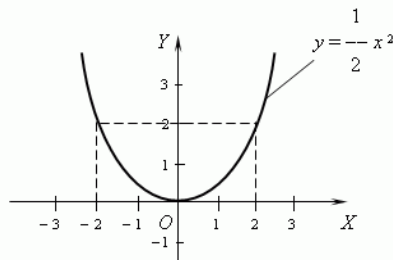


Рисунок 48 — Пример графика функциональной связи между переменными

Здесь каждому определенному значению x соответствует строго определенное значение y . Например, если $x = 2$, то $y = \frac{2^2}{2} = 2$. Что так же можно увидеть на графике данной функции. Иными словами мы всегда можем определить *точное* значение y , задав необходимое нам значение x , более того, даже если по каким-то причинам мы не имеем возможности сделать точный расчет, но располагаем графическим отображением зависимости, то мы всегда можем точно определить значение y (как это показано на рисунке). Мы довольно подробно остановились на данных особенностях функциональной зависимости несмотря на то, что они в целом тривиальны, но именно на этих свойствах функциональных зависимостей опирается анализ и процесс построения модели при исследовании зависимостей корреляционных.

1.3. Корреляционная связь

Функциональная связь между явлениями присуща неживой природе. Или если условия эксперимента близки к идеальным (когда предполагается, что никаких посторонних влияний нет), что достаточно сложно осуществимо.

В биологических науках чаще приходится иметь дело со связью между явлениями, когда *одной и той же* величине одного признака соответствует *ряд* варьирующих значений другого признака (сравните с функциональной связью, когда каждому конкретному значению одной переменной соответствует строго определенное значение другой), что обусловлено чрезвычайным многообразием взаимодействия различных явлений живой природы.



То есть при изменении значения одного признака, зависимый признак может принимать в разных ситуациях (экспериментах) неодинаковые значения (*ряд* варьирующих значений признака), они могут быть близки по величине, но очень редко одинаковыми. Каждому определенному значению первого признака соответствует не одно значение второго признака, а целое распределение этих значений. Такого рода связь носит название **корреляционной** (correlation—соответствие, соотносительность) или **корреляция**.

В качестве простого примера часто приводят наблюдение за сверчками: частота трелей сверчка зависит от температуры, когда на улице холодно, сверчки поют не так часто. Другой пример: количество преступлений (на душу населения) связано с количеством полицейских на данной территории.



В то время как функциональная связь имеет место в каждом отдельном наблюдении, корреляционную связь можно обнаружить только при **многочисленном сопоставлении** признаков.

Также следует помнить, что обнаружение корреляции между сопоставляемыми явлениями не говорит еще о существовании причинной связи между ними.

Приведем пример простейшей корреляционной зависимости (таблица 30). В группе из 15 человек были измерены рост и вес у каждого члена группы, данные внесли в таблицу попарно, что означает: для каждого отдельного участника эксперимента записывается именно его рост и соответствующий ему вес, перепутывать данные ни в коем случае нельзя!

Таблица 30 — Данные эксперимента

№	Рост, см	Вес, кг
1	175	67
2	177	68
3	179	68
4	181	70
5	183	69
6	168	57
7	169	60
8	170	61
9	171	63
10	172	64
11	180	69
12	181	71
13	182	74
14	179	68
15	175	65

Построим *точечную диаграмму* по данным таблицы 30 (рисунок 49):

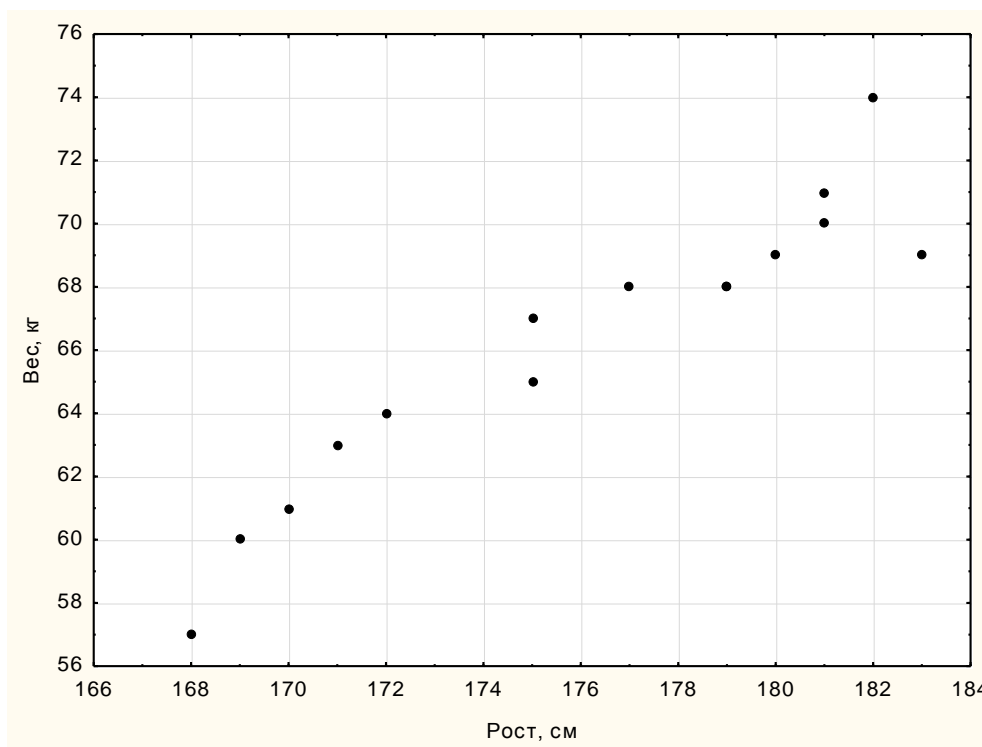


Рисунок 49 — Точечная диаграмма по выборочным данным

Из таблицы видно, что значению роста 181 см соответствуют значения веса 70 и 71 кг, что также отражено на графике. Можно заметить, что на лицо имеется некоторая зависимость: при увеличении значений роста значение веса так же увеличивается. Если мы будем увеличивать

выборку, например до 125 человек, то можем получить следующий график (рисунок 50):

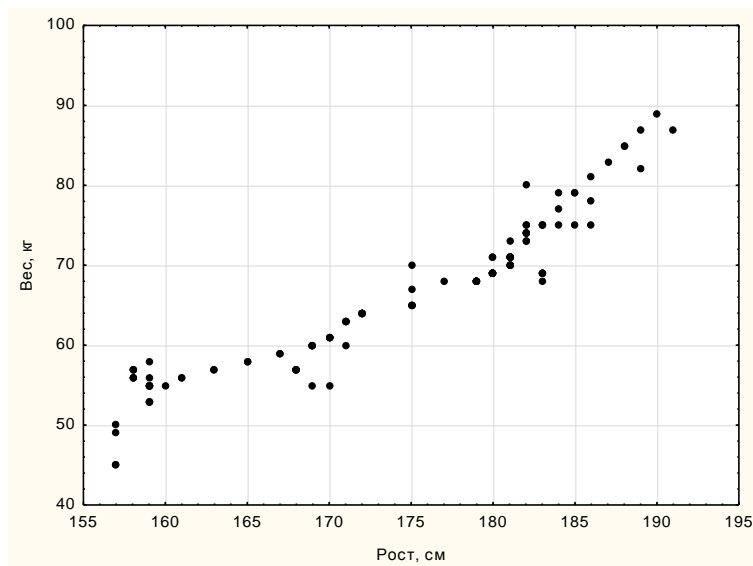


Рисунок 50 — Точечная диаграмма при изучении большой выборки

Что еще больше уверяет в наличествующей взаимосвязи. Однако в приведенном примере нельзя выявить точную математическую формулу, при подстановке в которую значения роста мы бы получили совершенное определенное и единственное значение веса; **каждому определенному значению роста будет соответствовать несколько близких по величине значений веса**³⁰, но мы не можем сказать точно какое. При этом обратите внимание, что если мы возьмем другую выборку (рисунок 51), то данные изменятся, но если зависимость все же имеется, то она будет иметь схожий характер (схожие тенденции).

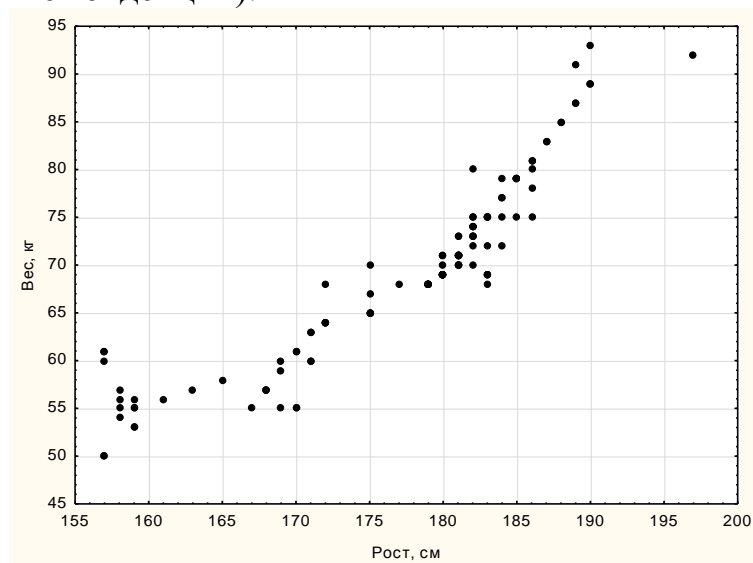


Рисунок 51 — Вариант точечной диаграммы при изучении большой выборки

³⁰ Конечно при условии, что выборка однородна.

Видно, что характер зависимости сохранился, но значения несколько отличаются от данных на предыдущем рисунке.



Данный пример достаточно полно характеризует свойства корреляционной зависимости между переменными, а именно следующее: имея дело с явлениями живой природы, исследователь не может однозначно определить вид (математическую формулу) взаимосвязи между ними, как это выглядит при функциональной зависимости. Причиной может являться как сложные механизмы самого взаимодействия, так и влияющие факторы, о которых исследователь не знает или не может исключить из эксперимента или повседневного практического опыта. Однако он может обнаружить некую имеющуюся взаимосвязь и определить ее характер, что в дальнейшем послужит построению модели, которая с заданной точностью будет описывать выявленную зависимость.

В связи с этим исследователь вынужден использовать *приблизительную характеристику* зависимости, отражающую имеющуюся связь с некоторой точностью, каковая не становится менее значимой в силу своей «*неточности*» и «*несовершенности*» с точки зрения математической формулы.

Остановимся более подробно на методиках анализа статистических зависимостей между переменными.

1.4. Этапы анализа зависимости между переменными

В общем случае любой анализ вероятной зависимости сводится к следующим этапам:

1. Сбор данных.
2. Внесение данных в таблицы (попарно).
3. Первичный анализ предполагаемой зависимости.
4. Исключение из выборки артефактов (выбросов).
5. Регрессионный анализ.
6. Корреляционный анализ.
7. Проверка соответствия модели экспериментальным данным.

Рассмотрим эти этапы по порядку. Минуем этап сбора данных, так как это входит в раздел о планировании эксперимента. Относительно занесения данных в таблицу стоит еще раз обмолвиться, что полученные значения необходимо вносить **попарно**, то есть каждому члену выборки (каждому наблюдению) соответствует своя пара данных. Так же заполнение таблиц должно быть аккуратным, что позволит избежать некоторых осложнений в процессе анализа, о которых речь пойдет ниже. После занесения данных в таблицы их желательно отсортировать по возрастанию и проверить тип распределения каждого признака. Далее необходимо осуществить беглый **первичный анализ**.

1.5. Первичный анализ

Первичный анализ необходим для определения наиболее подходящих методов дальнейшего анализа зависимостей, что подразумевает знание основных статистических параметров, вида распределения, типа исследуемых данных. После данные эксперимента наносятся на график в виде точечной диаграммы, где по оси X откладываются значения независимой переменной, а по оси Y — зависимой. Уже по первичному анализу можно судить о наличии какой-либо взаимосвязи (рисунок 52).

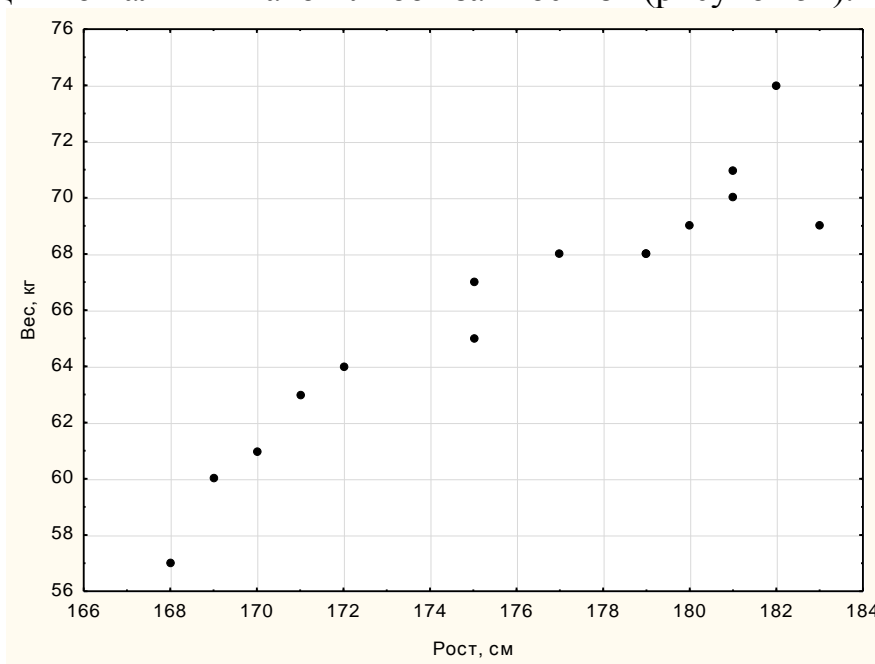


Рисунок 52 — Точечная диаграмма на первичном этапе анализа

Однако этого не достаточно, необходимо еще определить тип зависимости и построить модель, описывающую эту взаимосвязь, а также определить ее *значимость*, то есть соответствие эмпирическим данным, так же довольно часто определяют *силу зависимости* между переменными.

Но прежде чем идти дальше выясним, что же такое **выбросы (артефакты)**.

1.6. Выбросы

В силу возможных ошибок эксперимента или записи результатов исследования в таблицу могут попасть значения явно контрастирующие с остальными значениями данных эксперимента. Такие значения именуются *выбросами*.

По определению, выбросы являются нетипичными, резко выделяющимися наблюдениями. Выбросы могут существенно повлиять на все дальнейшие вычисления и значительно исказить результаты. Поэтому даже наличие единичного выброса при обработке результатов исследования нежелательно. Обычно считается, что выбросы представляют собой *слу-*

чайную ошибку, которую следует контролировать. К сожалению, не существует общепринятого метода автоматического удаления выбросов. Чтобы не быть введенными в заблуждение полученными значениями, необходимо на построенном графике зависимости (рисунок 53) проверить каждую точку по свойствам напоминающую выброс. Если будет сделан вывод, что данная пара значений на самом деле является выбросом, то ее следует удалить из дальнейшего анализа. После этого построить график снова, но уже без выбросов и продолжить анализ.

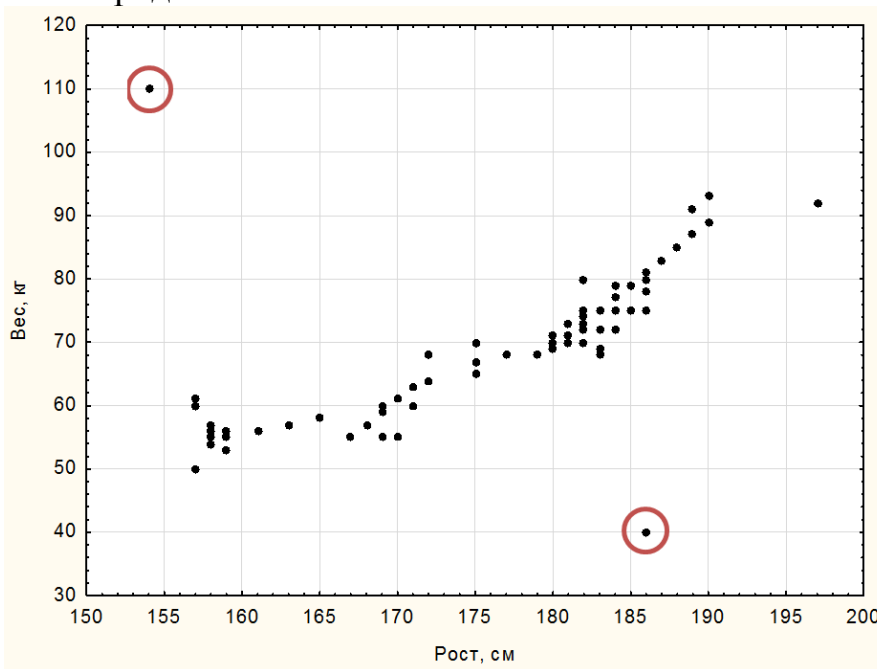


Рисунок 53— Пример графика с присутствующими выбросами

После первичных процедур перейдем (если мы выяснили, что это имеет смысл, т. е. увидели на графике возможную зависимость) к построению модели, описывающей характер зависимости.

1.7. Регрессионный анализ

Для того, чтобы математически описать предполагаемую зависимость (построить *модель*) данные эксперимента наносятся на график, после чего определяется *вид зависимости* (линейная, криволинейная и т. п.) и *уравнение*, с помощью которых можно описать имеющуюся зависимость, а также *степень достоверности* полученной модели.

Не вдаваясь в возможные вариации видов зависимости, остановимся на простой *линейной зависимости*. Это означает, что имеющуюся корреляционную взаимосвязь можно описать *прямой линией* и *ее уравнением с заданной точностью*.

Рассмотрим следующий пример. Как и прежде работаем с предполагаемой зависимостью между ростом и весом. Нанесем результаты измерения на график (рисунок 54).

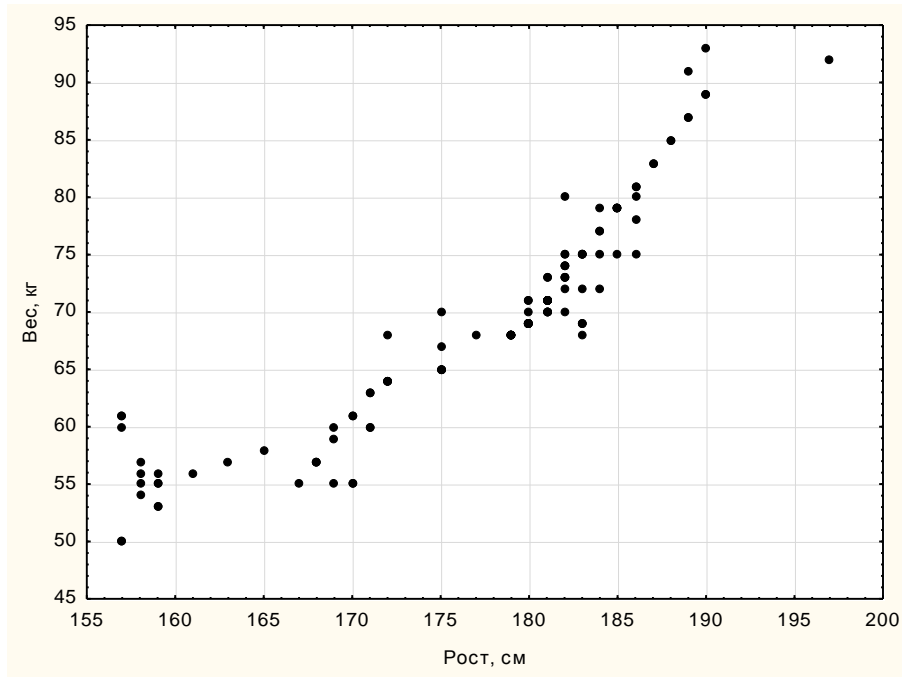


Рисунок 54 — Точечная диаграмма результатов измерения *роста* и *веса* в группе

Из графика видно, что зависимость между переменными имеет скорее всего линейный характер, т. е. ее можно описать используя линейную математическую модель: прямую линию и уравнение прямой линии. Как результат можно говорить, что зависимость между переменными имеет линейный характер.

Изобразим это на графике (рисунок 55):

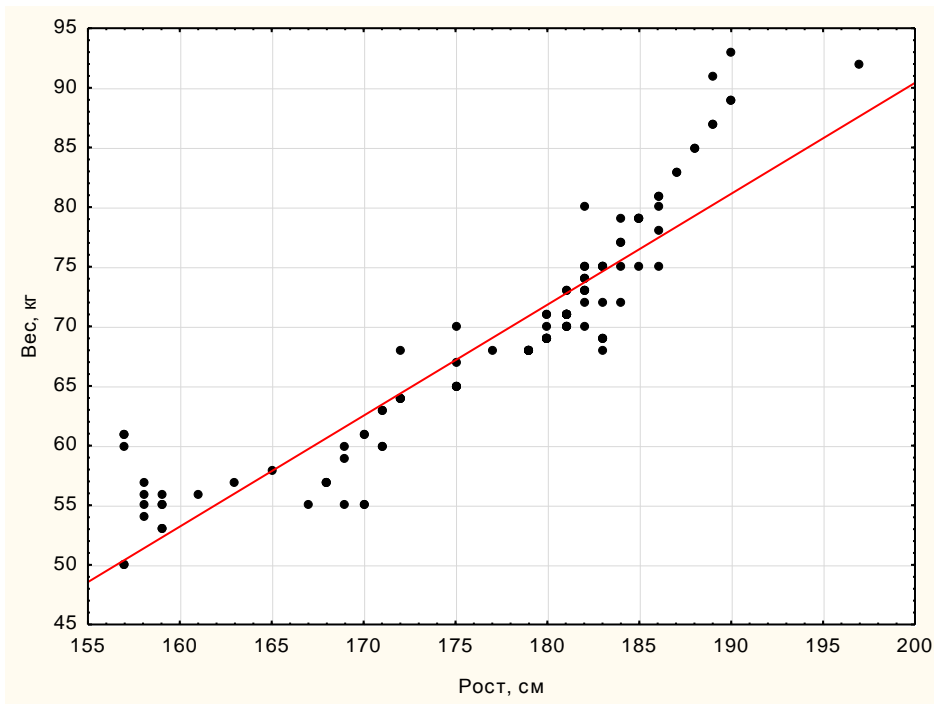


Рисунок 55 — Точечная диаграмма с обозначенной линией регрессии

Прямая линия на графике изображает характер зависимости между переменными и может быть выражена уравнением прямой. Уравнение прямой линии выглядит следующим образом:

$$y=ax+b.$$

Линия на графике называется **прямой регрессии**, уравнение — **уравнением регрессии**.

А само построение линии регрессии и ее уравнения называется **регрессионным анализом**.

В уравнении регрессии одна из переменных, x , называется *независимой* переменной, а другая, y , — *зависимой*. Набор значений y , соответствующих определенному значению x , обозначим $y(x)$ ³¹. Коэффициенты a и b — **параметры уравнения регрессии**. Если при увеличении одной переменной увеличивается значение другой, то зависимость называют **прямой**, если же переменные ведут себя наоборот — **обратной**.



Смысл регрессионного анализа в следующем: весь массив данных (точек на графике) мы заменяем уравнением (прямой или кривой линией), характеризующим и описывающим наличествующую связь с заданной точностью.

Регрессионный анализ предполагает построение прямой (или кривой, если зависимость криволинейная) регрессии, построенной **методом наименьших квадратов** (рисунок 56).

Кратко опишем данный метод. Линия, проходящая через точки на графике строится таким образом, чтобы сумма квадратов расстояний (вычисленных по оси Y) от наблюдаемых точек до прямой являлась минимальной.

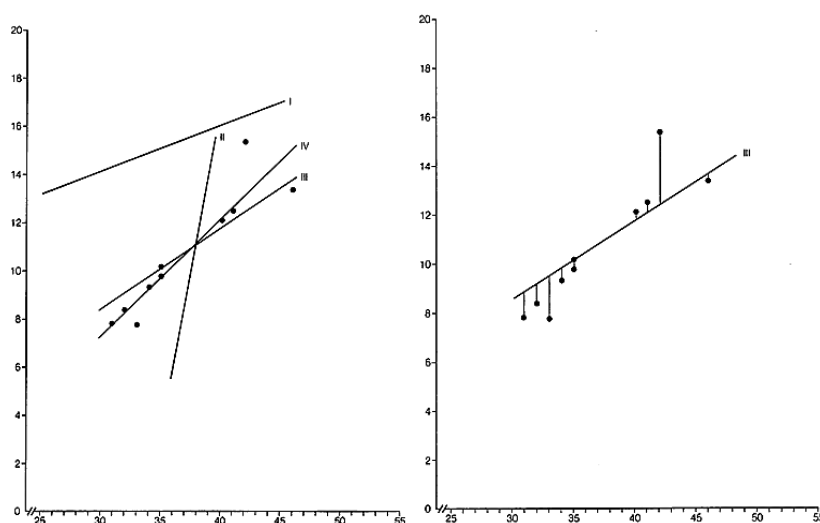


Рисунок 56 — Иллюстрация метода наименьших квадратов

³¹ Читается y от x .

В результате регрессионного анализа мы получаем модель, с помощью которой по значению одного признака (независимая переменная) мы предсказываем значение (поведение) второго признака (зависимая переменная).

Однако еще раз стоит отметить, что *только в условиях эксперимента* мы произвольно меняем независимую переменную и смотрим, как меняется зависимая. При этом речь действительно идет о зависимости, т. е. о причинной связи. В прочих же случаях выявление статистической связи двух переменных указывает на возможность причинной связи, но не доказывает ее. Разобраться в причинах и следствиях вообще невозможно чисто статистическими методами. Необходимо, в частности, найти биологический механизм, порождающий выявленную связь.

Например, эпидемиологические данные о связи пассивного курения с заболеваемостью ишемической болезнью сердца (ИБС) еще не доказывают, что пассивное курение способствует развитию ИБС. Может быть и то и другое — следствие какой-либо неизвестной причины, например нервной обстановки в рабочем коллективе. Однако экспериментальные данные о том, что пассивное курение и отдельные компоненты табачного дыма вызывают поражение сердца у лабораторных животных, говорят в пользу именно причинной связи.

Это замечание опять же выводит нас за рамки рассматриваемого вопроса и в очередной раз указывает на *вероятностную оценку* статистических выводов. Тем не менее мы можем пользоваться этими выводами пусть и приближенно, зачастую этого уже достаточно. Например, выявив характер зависимости между необходимой дозой препарата и весом пациента, мы, зная вес данного конкретного пациента, можем отмерять ему необходимое количество препарата, так же можем предупредить о возможных побочных эффектах, если доза была превышена или о возможном отсутствии эффекта, если количество принятого препарата недостаточно.

Так же если модель достаточно точная, то ее можно использовать в приблизительной оценке значения зависимой переменной при определенном значении переменной независимой. Например, исходя из выше приведенного графика (данных о росте и весе), мы получили следующее уравнение:

$$y = -95,653 + 0,9304 \cdot x,$$

где y – значение веса, x – значение роста.

Следовательно, если мы хотим узнать приблизительный вес человека ростом 185 см, то подставив это значение в данную формулу получим

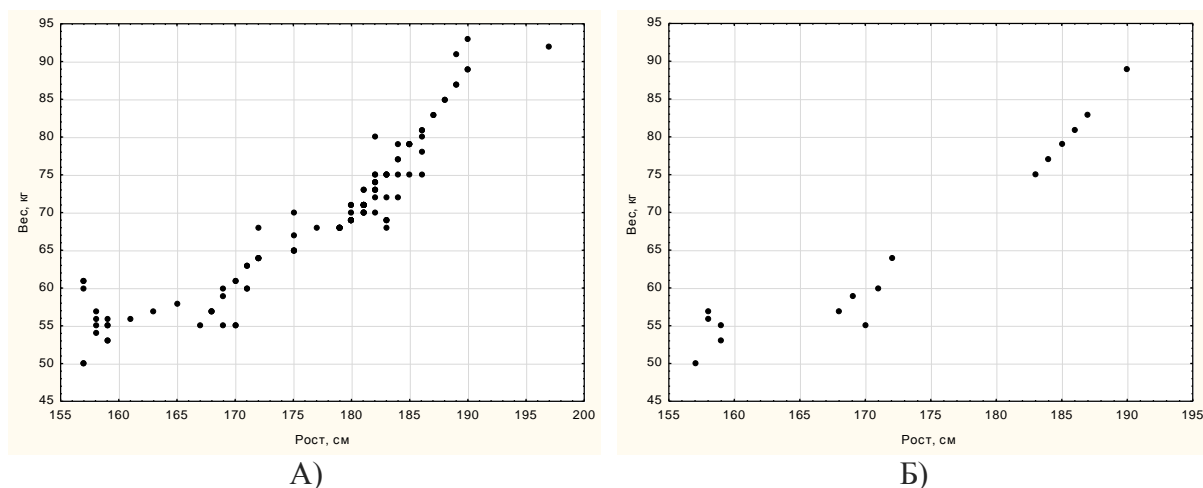
$$-95,653 + 0,9304 \cdot 185 \text{ см} = 76 \text{ кг} .$$

Конечно это приближенная оценка, но тем не менее она довольно точно отражает реальное состояние дел.

Очень важно, чтобы выборка, по которой будут делаться выводы о наличии взаимосвязи, была репрезентативной, в противном случае найденная зависимость может давать неточные результаты. Поэтому вкратце коснемся выборочных оценок при регрессионном анализе.

1.8. Уравнение регрессии по выборочным данным

В реальной жизни редко удается получить данные обо всей совокупности, и исследователю приходится довольствоваться выборочными данными (рисунок 57).



**Рисунок 57 — Пример графиков зависимости между переменными:
А) генеральная совокупность; Б) выборка**

При этом выборки могут значительно отличаться друг от друга. Другая выборка из той же генеральной совокупности (рисунок 58):

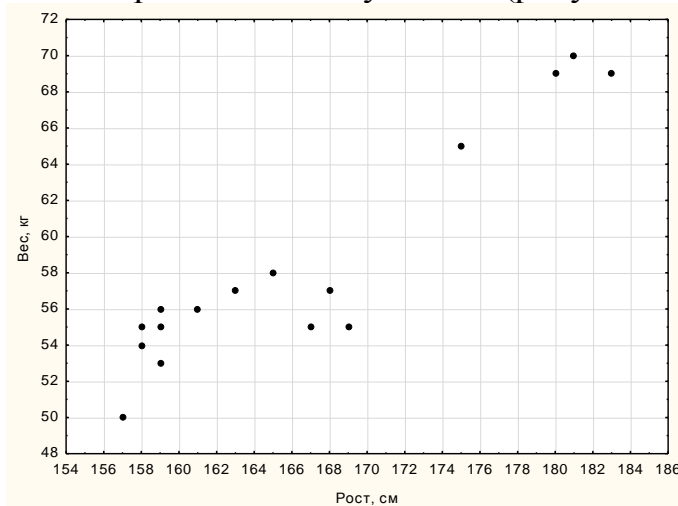


Рисунок 58 — Другая выборка из той же генеральной совокупности

Коэффициенты уравнения для этих выборок рассчитываются также методом наименьших квадратов. И если мы построим линии регрессии и уравнения регрессии для совокупности и выборок, то получим небольшие различия (рисунок 59).

Видно, что уравнения регрессии отличаются как у данных по совокупности от выборочных данных, так и между самими выборками.

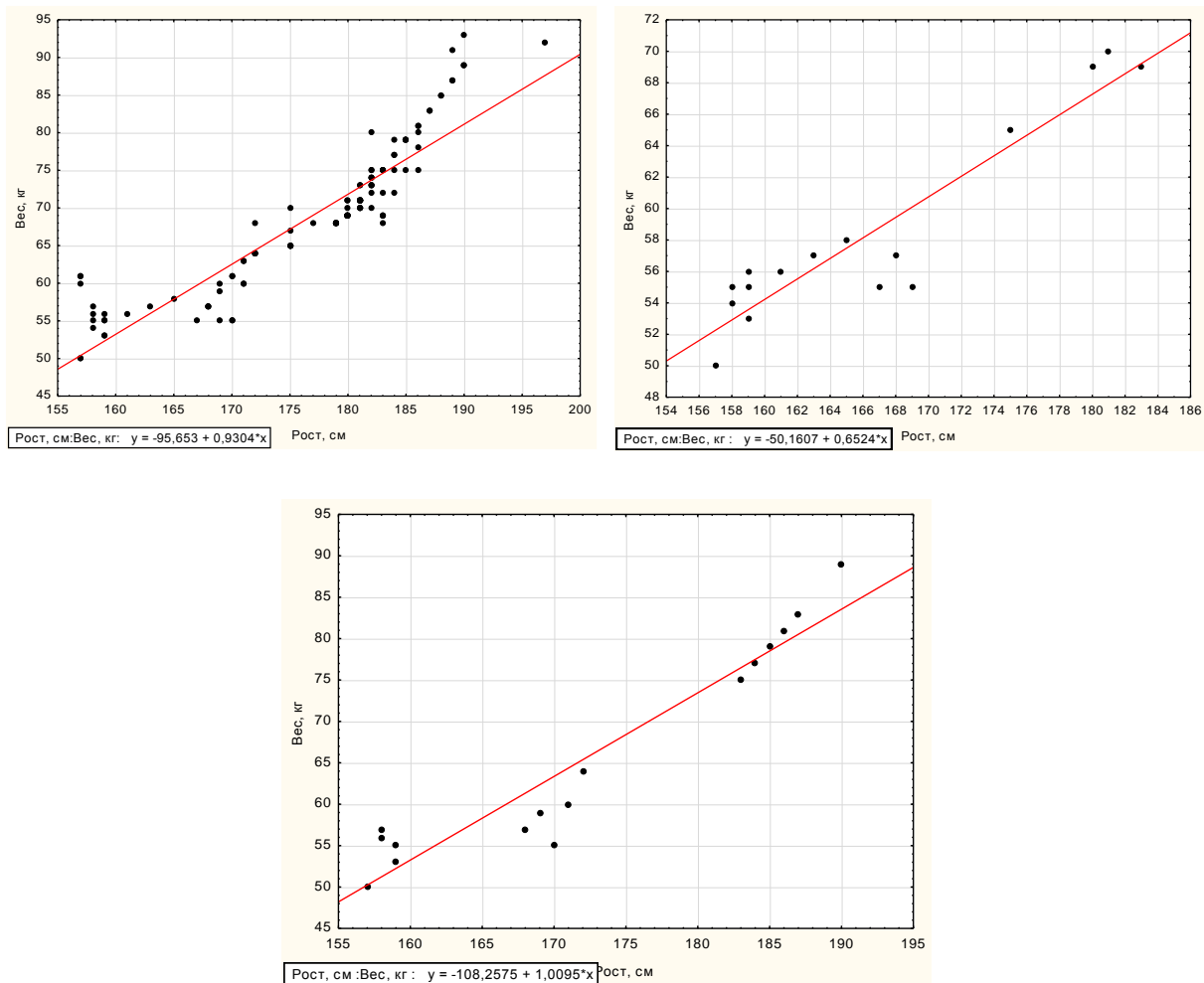


Рисунок 59 — Линии регрессии для генеральной совокупности и двух выборок из данной совокупности

Как видно из графиков линии регрессии имеют несколько разный наклон, еще больше различие заметно при рассмотрении уравнений регрессии:

Уравнение регрессии для совокупности: $y = -95,653 + 0,9304 \cdot x$.

Уравнение регрессии для первой выборки: $y = -50,1607 + 0,6524 \cdot x$.

Уравнение для второй выборки: $y = -108,2575 + 1,0095 \cdot x$.

Различия небольшие, но они имеются. Следовательно, разные выборки будут давать разные оценки коэффициентов регрессии. Выборочные параметры уравнения регрессии a и b — не более чем оценки *истинных* коэффициентов регрессии α и β .

Если построить все возможные выборки, получится совокупность всех значений a и b . Их *средние* равны α и β , а *стандартные отклонения* — σ_a и σ_b . Эти стандартные отклонения называются стандартными ошибками коэффициентов регрессии. Стандартные ошибки коэффициентов регрессии используются при проверке гипотез о **статистической значимости** выяв-

ленной зависимости³² и вычисления *доверительных интервалов* для кривой регрессии.

1.9. Доверительная область для линии регрессии

Так как мы не знаем истинных величин коэффициентов регрессии α и β , нам известны только их *оценки* a и b . Иначе говоря, истинная прямая регрессии может пройти выше или ниже, быть более крутой или пологой, чем построенная по выборочным данным. Поэтому более полным представлением результатов регрессионного анализа будет представление их с указанием **доверительного интервала для коэффициентов регрессии**, а также **доверительной области для самой линии регрессии** (рисунок 60).

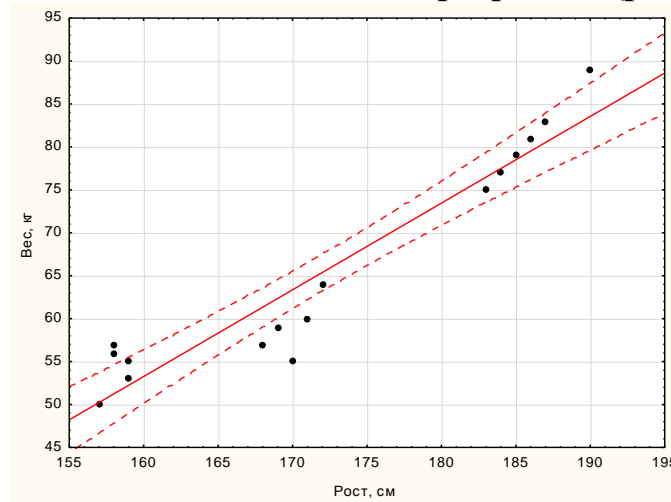


Рисунок 60 — Точечная диаграмма с нанесенной линией регрессии и доверительным интервалом для линии регрессии

На рисунке пунктирной линией изображена **95 % -ная доверительная область** для линии регрессии, построенной по выборочным данным.



95 %-ная доверительная область говорит о том, что мы точно не знаем, под каким наклоном пройдет реальная линия регрессии, если бы она была построена по данным всей генеральной совокупности, но мы можем с 95-процентной вероятностью говорить, что она будет находиться в пределах данной доверительной области.

³² В качестве критерия используется t-критерий Стьюдента (здесь мы не будем описывать методику расчета данного критерия для уравнений регрессии, приведем лишь некоторые замечания). Найденное по данным наблюдений значение t-критерия сравнивается с критическим значением, определяемым по таблицам распределения Стьюдента. Так же специализированные статистические пакеты рассчитывают достоверность p , значение которой сравнивается с уровнем значимости. Если выяснилось, что коэффициенты статистически не значимы, то велика вероятность того, что влияние независимой переменной на зависимую в генеральной совокупности вообще отсутствует или это влияние значительно отличается от того, что было замечено при анализе выборочных данных. В этом случае необходимо увеличить объем выборки, тогда возможно коэффициент станет статистически значимым и заодно уточнится его значение, и как следствие — точность полученной модели и способности ее прогнозирования.

1.10. Сравнение двух линий регрессии

Исследование редко ограничивается построением кривой регрессии, довольно часто требуется сравнить несколько линий регрессии между собой. Рассмотрим пример сравнения двух линий регрессии, рассчитанных по двум выборкам (пример взят из С. Гланц. Медико-биологическая статистика).

Мышечная сила при ревматоидном артрите

Причины ограниченной подвижности при ревматоидном артрите разнообразны: болезненность суставов, их тугоподвижность, атрофия мышц. Каков вклад каждого из этих факторов? Пытаясь ответить на этот вопрос П. С. Хелливелл и С. Джексон исследовали в частности, связь между мышечной массой и силой. В исследовании приняли участие 25 больных ревматоидным артритом (1-я группа) и 25 здоровых (2-я группа). Рассчитывали площадь поперечного сечения предплечья, а силу сжатия кисти определяли ручным динамометром. Результат измерений показан на рисунке 61. Кружки — результаты здоровых, квадратики — больных ревматоидным артритом.

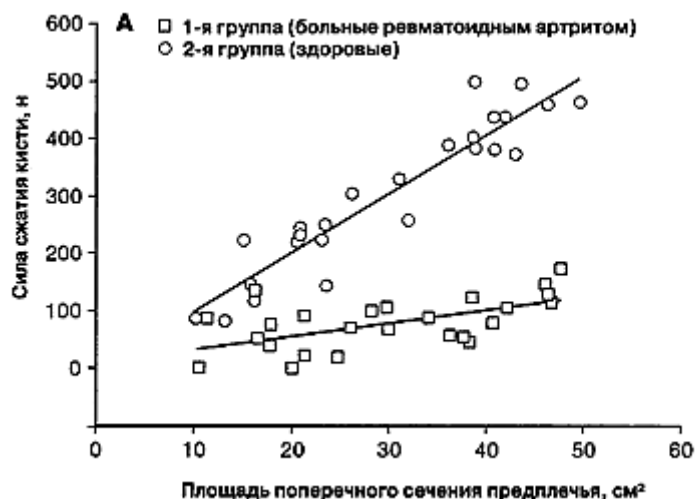


Рисунок 61 — Точечные диаграммы данных эксперимента в обеих группах и линии регрессии для каждой из групп

Как видно из графиков между площадью поперечного сечения предплечья (мышечная масса) и силой сжатия имеется однозначная *линейная зависимость*. Однако у больных артритом линия регрессии наклонена значительно ниже, чем у здоровых, что говорит о меньшей силе сжатия при тех же значениях мышечной массы. Это может указывать на атрофию мышц у больных ревматоидным артритом. Однако для того, чтобы однозначно судить об этом наблюдении, необходимо узнать значимо ли статистически различие между линиями регрессии. Для этого можно воспользоваться следующим методом: объединим обе группы испытуемых в одну и построим общую линию регрессии (рисунок 62).

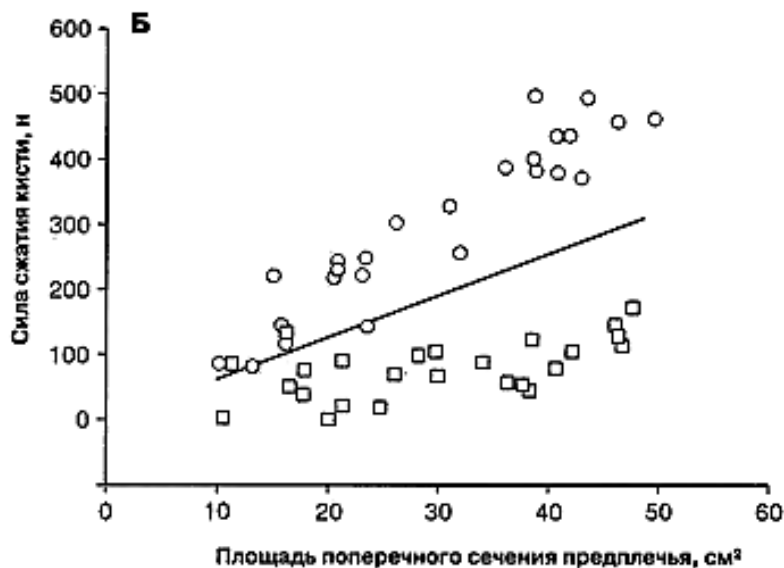


Рисунок 62 — Общая линия регрессии для обеих групп

Если разброс точек относительно общей линии, значительно превышает разброс относительно двух отдельных линий, тем самым это укажет на значимость этих различий, то отличия в наклоне у линий регрессии следует считать значимыми.

Минуя промежуточные вычисления (см. подробнее главу 8 книги С. Гланц Медико-биологическая статистика), находим *критерий Фишера F* и сравниваем его с табличным критическим значением для заданного уровня значимости и рассчитанных степеней свободы. Эмпирическое значение критерия больше критического, следовательно различия между разбросом точек для каждой отдельной кривой и разбросом точек для общей следует признать статистически значимыми. Линии регрессии на самом деле различаются углом наклона, который больше в группе здоровых.

1.11. Корреляционный анализ. Общие положения

Регрессионный анализ позволяет оценить, как одна переменная зависит от другой и каков разброс значений зависимой переменной вокруг прямой, определяющей зависимость. Эти оценки и соответствующие доверительные интервалы позволяют предсказать значение зависимой переменной и определить точность этого предсказания. Результаты регрессионного анализа можно представить только в наглядной цифровой или графической форме.

Однако иногда такая подробная информация излишняя и довольно сложна для восприятия, а исследователя интересует не предсказание значения одной переменной по значению другой, а просто *характеристика тесноты (силы) связи* между ними, при этом выраженная одним числом. Эта характеристика называется **коэффициентом корреляции**, обычно ее

обозначают буквой r . Здесь конечно стоит заметить, что для полного анализа зависимости между переменными регрессионный анализ все же необходим, как, впрочем, и расчет коэффициента корреляции. Конечно, представление результатов анализа остается на усмотрении исследователя, но сообщения только о тесноте связи между переменными на наш взгляд не достаточно.

Корреляция означает, что между двумя числовыми переменными наблюдается определенная линейная взаимосвязь.



Корреляционная связь не является точной зависимостью одного признака от другого, поэтому она может иметь различную степень — от полной независимости до очень сильной связи. Кроме того, характер связи между разными признаками может быть различен. Поэтому возникла необходимость определять форму, направление и степень корреляционных связей.

По форме корреляционная зависимость может быть **прямолинейной** или **криволинейной**, по направлению — **прямой** или **обратной**. Степень корреляции измеряется различными показателями, введенными для установления силы связи между количественными и качественными признаками. **Коэффициент корреляции** может принимать значения от -1 до $+1$.

Знак коэффициента корреляции показывает **направление связи** (прямая или обратная), а абсолютная величина — **тесноту связи**.

Чем ближе значение коэффициента корреляции к 1 (-1), тем теснее связь между переменными. Коэффициент, равный -1 , определяет столь же жесткую связь, что и равный 1 . В отсутствие связи коэффициент корреляции равен нулю. Промежуточные значения ($0,8$; $-0,5$; $0,25$) указывают на наличие связи в большей или меньшей степени, при этом, чем ближе значение по модулю к нулю, тем менее тесная связь между переменными. Знак коэффициента указывает на направление связи: « $-$ » — **обратная** связь, « $+$ » — **прямая**. Если связь между переменными является прямой, то при увеличении независимой переменной увеличивается значение зависимой переменной, при обратной связи наблюдается обратная закономерность — при увеличении значения независимой переменной значение зависимой уменьшается (рисунок 63).

1.12. Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона предназначен для описания линейной связи количественных признаков; как и регрессионный анализ, он требует нормальности распределения. Проиллюстрируем с помощью диа-

грамм рассеяния примеры зависимости между переменными и силу взаимосвязи, выраженную при помощи коэффициента корреляции.

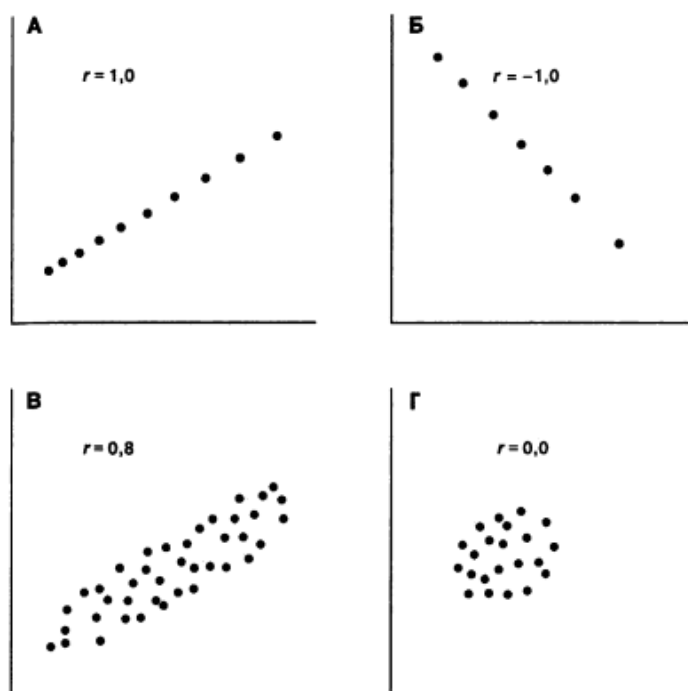


Рисунок 63 — Точечные диаграммы групп с разным значением коэффициента корреляции

Как видно из рисунков, чем выше значение коэффициента корреляции, тем больше точки стремятся к виду прямой линии, тогда как при $r = 0$, зависимость полностью отсутствует.



Коэффициент корреляции Пирсона (r) представляет собой меру линейной зависимости двух переменных. Когда говорят просто о «коэффициенте корреляции», почти всегда имеют в виду коэффициент корреляции Пирсона. Если возвести его в квадрат, то полученное значение **коэффициента детерминации** (r^2) представляет долю вариации, общую для двух переменных (иными словами, степень зависимости или связанности двух переменных). Чтобы оценить зависимость между переменными, нужно знать как *величину корреляции*, так и ее *значимость*.

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

где X и Y — средние значения переменных X и Y . Выражение для r «симметрично» — поменяв местами X и Y , мы получим ту же величину. Коэффициент корреляции принимает значения от -1 до $+1$.

Для оценки тесноты, или силы, корреляционной связи, по значению коэффициента корреляции довольно просто, обычно говорят, что если абсолютное (по модулю) значение $r < 0,3$ — это свидетельствуют о *слабой* связи, значения r от 0,3 до 0,7 — о связи *средней* тесноты, значения $r > 0,7$ — о *сильной* связи.

Более точную оценку силы корреляционной связи можно получить, если воспользоваться *таблицей Чеддока* (таблица 31):

Таблица 31 — Оценка силы корреляции по шкале Чеддока

Абсолютное значение r	Теснота (сила) корреляционной связи
Менее 0.3	Слабая
От 0.3 до 0.5	Умеренная
От 0.5 до 0.7	Заметная
От 0.7 до 0.9	Высокая
Более 0.9	Весьма высокая

Найдем коэффициент корреляции для приведенного выше примера с анализом зависимости между *ростом* и *весом* (рисунок 64).

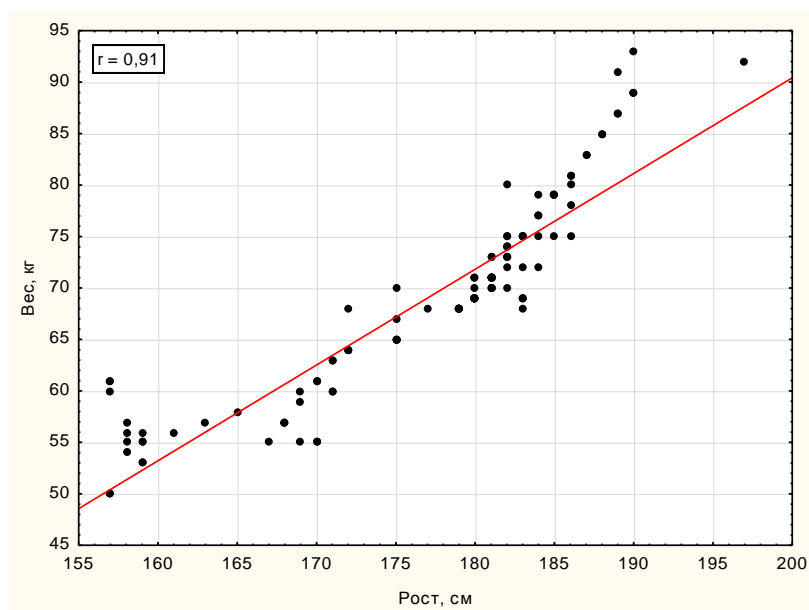


Рисунок 64 — График по данным эксперимента со значением коэффициента корреляции

Значение r получилось равным 0,91, что говорит о сильной линейной зависимости между переменными. Связь является прямой (так как значение коэффициента корреляции имеет знак «+»).

Может возникнуть вопрос, а почему для описания тесноты связи нельзя воспользоваться только регрессионным анализом, ведь он достаточно информативен и имеет наглядное визуальное представление, а в качестве меры тесноты связи можно было бы использовать остаточное стан-

дартное отклонение. Однако если поменять местами зависимую и независимую переменные, то остаточное стандартное отклонение, как и другие показатели регрессионного анализа, будет иным. Получается, что связь роста с весом одна, а веса с ростом — другая. Асимметричность регрессионного анализа — вот что мешает непосредственно использовать его для характеристики силы связи. Коэффициент корреляции, хотя его идея вытекает из регрессионного анализа, свободен от этого недостатка.

После вычисления значения коэффициента корреляции необходимо проверить его значимость, насколько найденная зависимость или теснота связи статистически значима. Оценка *статистической значимости* коэффициента корреляции осуществляется при помощи t-критерия, рассчитываемого по следующей формуле:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}},$$

где n — объем выборки. Полученное значение t сравнивается с критическим значением при определенном уровне значимости и числе степеней свободы $n-2$. Если t превышает $t_{крит}$, то делается вывод о статистической значимости выявленной корреляционной связи.

Условия и ограничения применения *критерия Пирсона*:

- Сопоставляемые показатели должны быть измерены в *количественной* шкале.

- Посредством коэффициента корреляции Пирсона можно определить лишь наличие и силу *линейной* взаимосвязи между величинами. Прочие характеристики связи, например, более точное определение направления связи (прямая или обратная), характер изменений (прямолинейный или криволинейный), а также наличие зависимости одной переменной от другой — определяются при помощи *регрессионного анализа*.

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

- Критерий корреляции Пирсона является *параметрическим*, поэтому условием его применения служит нормальное распределение сопоставляемых переменных. В случае необходимости корреляционного анализа показателей, распределение которых отличается от нормального, в том числе измеренных в порядковой шкале, следует использовать *коэффициент ранговой корреляции Спирмена*³³.

Отметим, что все же следует различать понятия *зависимости* и *корреляции*. Зависимость величин обуславливает наличие корреляционной связи

³³ Будет рассмотрен ниже.

между ними, однако наличие корреляционной связи все же не говорит еще о наличии зависимости³⁴.

Рассмотрим использование коэффициента корреляции на следующем примере.

Животный жир и рак молочной железы (из книги С. Гланц Медико-биологическая статистика)

По результатам опытов на лабораторных животных установлено, что высокое содержание животного жира в рационе повышает риск рака молочной железы. Можно ли говорить о подобной взаимосвязи у людей?

К. Кэррол собрал данные о потреблении животных жиров и смертности от рака молочной железы по 39 странам. Результат представлен на рисунке 65. Коэффициент корреляции между потреблением животных жиров и смертностью от рака молочной железы оказался равен 0,90. При этом корреляция значима на уровне значимости 0,001. Таким образом, при уровне значимости 0,001 можно утверждать, что существует корреляция между потреблением животных жиров и смертностью от рака молочной железы.

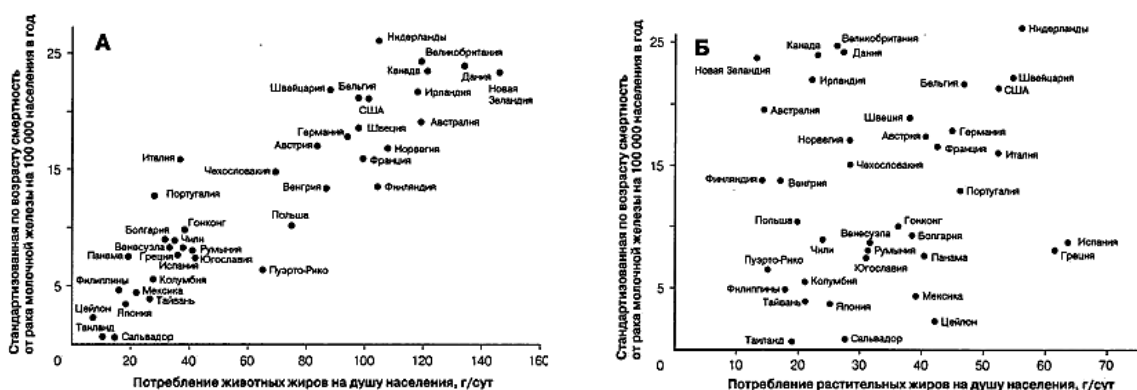


Рисунок 65 — Графики результатов исследования

34 Например, рост ребенка зависит от его возраста, т. е. чем старше ребенок, тем он выше. Если мы возьмем двух детей разного возраста, то с высокой долей вероятности рост старшего ребенка будет больше, чем у младшего. Данное явление и называется зависимостью, подразумевающей причинно-следственную связь между показателями. Разумеется, между ними имеется и корреляционная связь, означающая, что изменения одного показателя сопровождаются изменениями другого показателя.

В другой ситуации рассмотрим связь роста ребенка и частоты сердечных сокращений (ЧСС). Как известно, обе эти величины напрямую зависят от возраста, поэтому в большинстве случаев дети большего роста (а значит и более старшего возраста) будут иметь меньшие значения ЧСС. То есть, корреляционная связь будет наблюдаться и может иметь достаточно высокую тесноту. Однако, если мы возьмем детей одного возраста, но разного роста, то, скорее всего, ЧСС у них будет различаться несущественно, в связи с чем можно сделать вывод о независимости ЧСС от роста.

Приведенный пример показывает, как важно различать фундаментальные в статистике понятия связи и зависимости показателей для построения верных выводов (пример взят с сайта <http://medstatistic.ru>).

Далее была предпринята проверка связи смертности с потреблением растительных жиров. Соответствующие данные приведены на том же рисунке ниже. В этом случае коэффициент корреляции равен 0,15. И как оказалось даже при уровне значимости 0,10 вычисленное значение критерия значимости меньше критического. Следовательно, корреляция статистически не значима.

Таким образом, риск рака молочной железы статистически значимо связан с потреблением животных, но не растительных жиров. Значит ли это, что животный жир способствует развитию рака молочной железы? Пока нет. Ведь обе рассматриваемые переменные могут зависеть от какой-то третьей. В **обсервационном исследовании** (такое исследование, в котором исследователь собирает данные путём простого наблюдения событий в их естественном течении, не вмешиваясь в них активно), каковым является работа Кэррола, такую возможность отвергнуть нельзя. Однако экспериментальные данные, о которых мы упомянули выше, — сильный аргумент в пользу именно причинно-следственной связи. Вообще истолкование результатов регрессионного и корреляционного анализа зависит от того, в каком исследовании были получены данные — обсервационном или **экспериментальном** (метод исследования некоторого явления в управляемых наблюдателем условиях, он отличается от наблюдения активным взаимодействием с изучаемым объектом). Если мы обнаружили связь переменных в обсервационном исследовании, то это еще не значит, что одна из них влияет на другую. Возможно, их согласованные изменения — результат действия какого-то неизвестного фактора или суммы нескольких факторов. В экспериментальном же исследовании, произвольно меняя одну из переменных, мы можем быть уверены, что связь, если она будет выявлена, является причинной.

Впрочем, осторожность не помешает и в этом случае. В самом деле, трудно менять только одну переменную. Увеличивая содержание жира в рационе, мы либо увеличиваем общую калорийность, либо снижаем содержание белков и углеводов. Можно ли поручиться, что канцерогенное действие оказывает именно жир, а не дисбаланс питательных веществ?

Мы описали общий принцип анализа связи между переменными на примере линейной зависимости. Однако переменные далеко не всегда имеют линейную связь, а зачастую связь необходимо обнаружить между качественными типами данных. Эти вопросы будут рассмотрены в следующих параграфах.

1.13. Криволинейная корреляция и регрессия

В качестве дополнения кратко рассмотрим **криволинейную регрессию**.



Криволинейная связь между признаками — это такая связь, при которой **равномерным** изменениям первого признака соответствуют **неравномерные** изменения второго, причем эта неравномерность имеет определенный закономерный характер. Другими словами зависимость имеет вид не прямой линии, а какой-то кривой (параболы, экспоненциальной или логарифмической функции).

При графическом изображении криволинейных связей, когда по оси абсцисс откладывают значения первого признака (аргумент — независимая переменная), а по оси ординат — значения второго признака (функция — зависимая переменная) и полученные точки соединяют, получают изогнутые линии (рисунок 66). Характер изогнутости зависит от природы коррелируемых признаков.

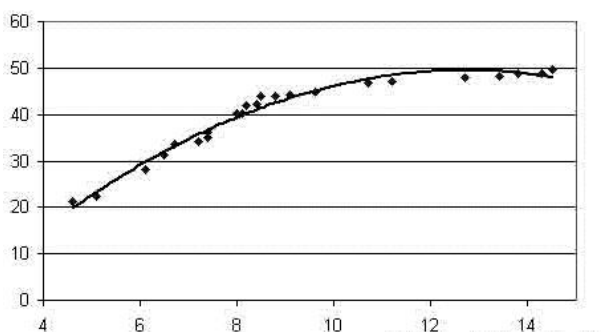


Рисунок 66 — Пример криволинейной зависимости между переменными

Если связь между изучаемыми явлениями носит характер кривой линии, что легко установить по графику, то *коэффициент корреляции* непригоден в качестве меры связи между переменными.

В этом случае *коэффициент корреляции* может указать на отсутствие связи там, где налицо сильная зависимость, но имеющая вид *кривой линии*.

Естественно коэффициент корреляции Пирсона, предназначенный для анализа линейной взаимосвязи, а также и сама прямая линия регрессии не будут характеризовать имеющуюся зависимость (рисунок 67).

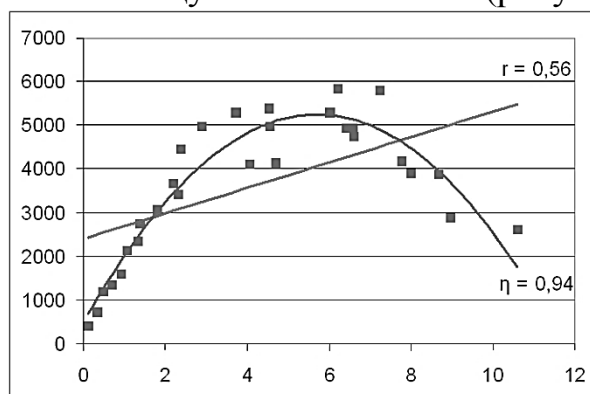


Рисунок 67 — Характеристика криволинейной зависимости с использованием кривой и прямой линий

Из рисунка видно, что прямая линия не характеризует имеющуюся криволинейную зависимость между переменными, коэффициент корреляции так же имеет низкое значение и незначим.

Поэтому необходим новый показатель, который бы правильно измерял степень криволинейной зависимости. Таким показателем является **корреляционное отношение**, обозначаемое греческой буквой η (эта).



Корреляционное отношение измеряет степень корреляции при любой ее форме и может измерять степень криволинейных и прямолинейных связей.

При отсутствии связи значение *корреляционного отношения* равно нулю, а при тесной (функциональной связи) — единице. Другими словами, чем ближе к единице значение *корреляционного отношения*, тем теснее, ближе к функциональной зависимости связь между признаками.



В отличие от *коэффициента корреляции*, который дает одинаковую меру связи признаков (первого со вторым и второго с первым), *корреляционное отношение* второго признака по первому обычно не бывает равно *корреляционному отношению* первого признака по второму. Поэтому крайне важно определить, какая выборка является независимой, а какая зависимой.

При работе с криволинейной регрессией задачей исследователя является **подбор вида функции**, которая бы наиболее четко отражала наличествующую взаимосвязь между переменными (иначе: значение квадрата корреляционного отношения было бы максимально возможным).

1.14. Итоговая модель

Определив тип взаимосвязи и силу связи между переменными, рассчитав все необходимые параметры и проверив их значимость, исследователь формирует **модель**, по которой он может с определенной вероятностью предсказывать поведение зависимого объекта (значения *зависимого признака*), при изменении внешнего воздействия (изменении значений *независимого признака*).

Подведем итог нашему анализу зависимости между ростом и весом (рисунок 68).

Мы получили, что зависимость между переменными линейная и прямая, при этом она довольно тесная (*коэффициент корреляции* $r = 0,91$ и значим на уровне $0,05$). *Уравнение регрессии* имеет вид: $y = -95,653 + 0,9304 \cdot x$, при этом коэффициенты его также значимы на *уровне значимости* $0,05$. Это и есть наша *модель*³⁵, по которой мы можем предсказывать поведение одной переменной в зависимости от поведения другой.

³⁵ Так же следует проверить ее значимость. Один из наиболее часто используемых вариантов проверки заключается в использовании F — критерия Фишера.

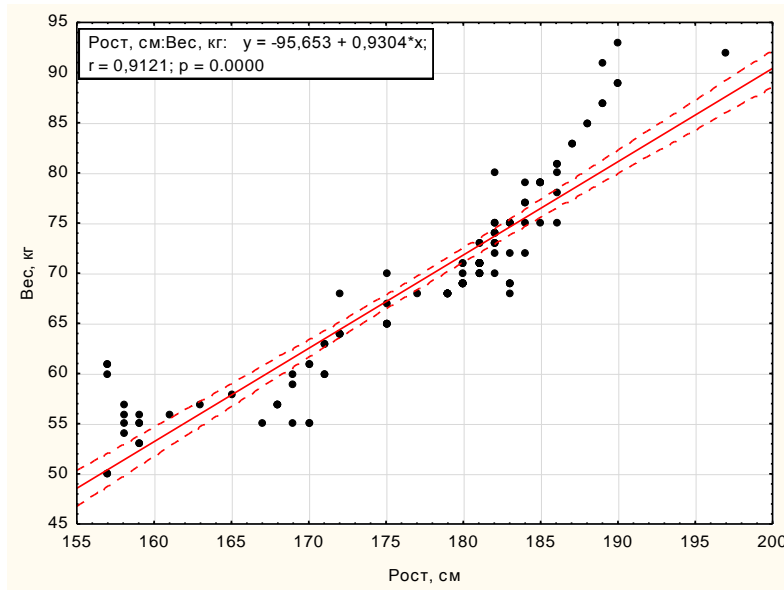


Рисунок 68 — Диаграмма рассеяния с нанесенной прямой регрессии, доверительной областью, уравнением регрессии, значением коэффициента корреляции и значимостью модели

Но только потому, что у нас есть модель, мы не можем подставлять в нее любое значение x и по ней рассчитать значение y .

Почему? Кто может поручиться, что зависимость будет работать и за пределами той области, по которой были собраны данные? Это очень важный момент! Вспомните пример про сверчков, очевидно, что вряд ли по мере того, как температура будет повышаться до бесконечности, сверчки будут стрекотать все быстрее и быстрее, без ограничений. В определенный момент сверчок умрет от перегрева. Точно так же сверчки не выживут в слишком холодную погоду, значит, нельзя подставлять в уравнение слишком маленькие значения независимой переменной и рассчитывать, что модель сработает.



Ни в коем случае нельзя строить предположения, используя значения независимой переменной, которых нет в пределах собранных вами данных.

Статистики называют это экстраполяцией. И она далеко не всегда допустима. Поэтому предсказание поведения зависимой переменной на основании модели будет достоверным только в пределах полученных данных (значений независимой переменной) в ходе эксперимента. Тогда как вне этих пределов поведение может оказаться непредсказуемым.

И еще. Модель описывает общую взаимосвязь между переменными (x и y), на самом деле вы предсказываете не y (значение зависимой переменной), а **ожидаемое значение y** для любого конкретного значения x (независимой переменной). В любом случае модель необходимо проверить на практике.

§2. Множественная регрессия. Логистическая регрессия

Вкратце рассмотрим еще два типа анализа зависимости между переменными.

2.1. Множественная линейная регрессия

Выше мы рассмотрели зависимость между двумя признаками. Таким образом мы можем проследить, влияет ли увеличение дозы препарата на какой-либо интересующий нас параметр здоровья, предсказать значение одной переменной от значений другой и т. п. Однако довольно часто может быть ситуация, когда на одну зависимую переменную могут оказывать влияние несколько независимых. В случае если все переменные в исследовании являются количественными, то для анализа зависимостей между ними используют методы **множественной регрессии**.

Здесь мы не будем вдаваться в достаточно сложное математическое описание данного метода. Смысл множественной регрессии схож с описанной выше линейной регрессией. В результате анализа данных исследователь получает *модель*, описывающую зависимость между несколькими переменными вида:

$$y = b_1 \times x_1 + b_2 \times x_2 + \dots + b_n \times x_n + a,$$

где n — количество независимых переменных, обозначенных как x_1 и x_n , a — некоторая константа, как и в случае уравнения прямой при линейной парной регрессии.

Сложностью наглядного представления множественной регрессии является невозможность представить ее результаты графически при количестве независимых переменных больше двух.

Рассчитанные коэффициенты оцениваются с использованием *критериев значимости*, например *критерия Стьюдента*. Получившуюся модель необходимо обязательно проверить на практике. Необходимо учитывать, что чем сложнее полученная модель, тем менее вероятна ее стабильная работоспособность. Другими словами, если мы обнаружили, что зависимая переменная имеет взаимосвязь с пятью или большим количеством независимых переменных, то имеет смысл более критично отнестись к результату и предпринять достаточно тщательную проверку модели.

2.2. Логистическая регрессия

Часто возникает ситуация, когда необходимо построить модель, которая предсказывает возможность наступления события в зависимости от состояния одного или нескольких параметров исследуемого объекта. Другими словами, исследователь с некоторой долей вероятности может предсказать возможный исход, имея статистически обоснованную модель поведения наблюдаемого объекта. При этом под исходом подразумевается, что событие либо **наступило** («да»), либо **не наступило** («нет»), т. е. зависи-

мая переменная может принимать только два значения. Такая переменная называется **бинарной**. В задачах подобного типа *бинарная переменная* зависит от одной или нескольких переменных любого типа, будь то количественные переменные, качественные или порядковые. Примером может служить модель определения успешности сдачи теста в зависимости от возраста и опыта тестируемого, определение наличия заболевания в зависимости от результатов анализов по интересующим врача параметрам организма пациента и т. д.

Процедура, позволяющая строить подобные модели, называется **логистической регрессией**. Опять же в данном курсе мы опишем данный метод довольно кратко и только с целью ознакомления.



Логистическая регрессия, или **логит-регрессия**, (англ. logit model) — это статистическая модель, используемая для предсказания вероятности возникновения некоторого события в зависимости от одной или нескольких независимых (объясняющих) переменных.

Можно сказать, что *логистический регрессионный анализ* позволяет строить статистическую модель для **прогнозирования вероятности наступления** события по имеющимся данным (например, факторам риска).

Для этого вводится *зависимая переменная (объясняемый признак)*, принимающая лишь одно из двух значений (бинарная переменная, коды для такой переменной задаются как числа: 0 (событие не произошло) и 1 (событие произошло)), и несколько *независимых переменных* (также называемых **объясняющие признаки, предикторы** или **регрессоры**), на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной. С помощью логистической регрессии можно оценивать вероятность того, что событие наступит для конкретного испытуемого (больной/здоровый, возврат кредита/дефолт и т. д.).

Приведем несколько примеров из медицины: прогнозирование возникновения инфаркта миокарда по данным о курении, возрасте, наличии артериальной гипертонии. В данном случае *объясняемая бинарная переменная* — возникновение инфаркта миокарда (может принимать только два значения: да или нет, инфаркт миокарда возник или нет), тогда как возраст, данные о курении, наличие артериальной гипертонии — *объясняющие переменные, предикторы*. Или модель обнаружения карциномы (злокачественной опухоли) мочевого пузыря, в этом случае *объясняющим переменными* могут быть количество (типизация) Т-клеток и тест LAI. Результатами первого теста являются значения, принадлежащие к интервальной шкале, а тест LAI дает **дихотомический результат**: «положительно» или «отрицательно». В зависимости от значения этих переменных

модель определяет принадлежность тестируемого к группе больных или здоровых.

Не вдаваясь глубоко в математический аппарат метода, приведем его алгоритм.

Математическая модель множественной регрессии (общий случай) выглядит следующим образом:

$$y = a + b_1 \times x_1 + b_2 \times x_2 + \dots + b_n \times x_n,$$

где n — количество независимых переменных, обозначенных как x_1 и x_n , a — некоторая константа, как и в случае уравнения прямой при линейной парной регрессии.

Однако данное уравнение не совсем годится для моделирования исхода события для бинарной переменной, так как неизбежно значение функции (y) будет принимать значения больше 1 и меньше 0, а это некорректно с точки зрения поставленной задачи. Для того, чтобы модель работала правильно, уравнение необходимо преобразовать к такому виду, при котором функция будет принимать не бинарные значения, а значения, выражающие **вероятность наступления события**. То есть вместо предсказания значения бинарной переменной, мы предсказываем значение непрерывной переменной со значениями на отрезке $[0,1]$ при любых значениях независимых переменных (рисунок 69). Это достигается применением следующего регрессионного уравнения (**логистическое преобразование**):

$$p = 1 / (1 + e^{-y}),$$

где p — вероятность того, что произойдет интересующее событие; e — основание натуральных логарифмов 2,71... (число Эйлера); y — стандартное уравнение регрессии ($y = a + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$).

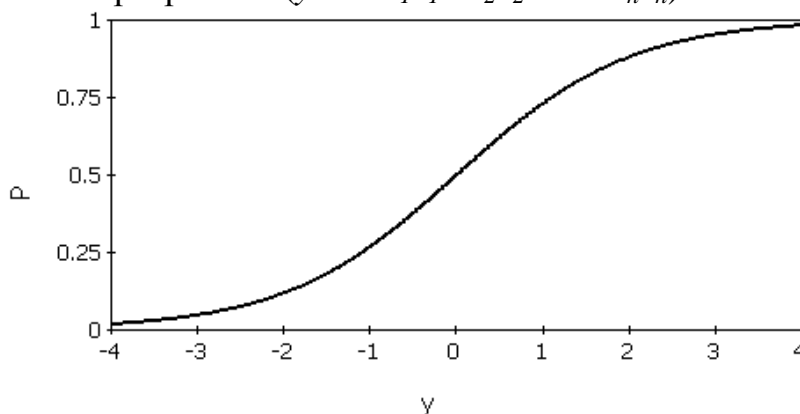


Рисунок 69 — Логистическая кривая

Найдя значения коэффициентов регрессии (a , b_1 , b_2 и т. д.) и подставив интересующие нас значения объясняющих переменных (x_1 , x_2 и т. д.) мы получим **вероятность наступления** смоделированного события (например болен пациент или нет и т. д.). В зависимости от того, с каких

значений вероятности p мы условились принимать их как положительный исход, а с каких — отрицательный, мы делаем вывод о принимаемом значении бинарной переменной. То есть, если после постановки в уравнение модели данных анализа из вышеприведенного примера таких, как: *Типизация t-клеток* = 72, а *LAI* = «да» и получим, что значение вероятности $p = 0,881047$, то в этом случае можно говорить, что бинарная переменная (болен (1)/здоров (0)) примет значение болен (1).

Условия применения метода:

- Зависимый признак — качественный и бинарный, т. е. имеет только два возможных значения: 1 — наличие заболевания и 0 — отсутствие заболевания.

- Независимые признаки — может принимать любые значения и иметь любой тип данных.

- Независимые признаки должны быть также независимыми друг от друга, между ними не должно быть корреляции или ассоциаций.

В результате *логит-регрессионного анализа* получается модель (оценка регрессионных коэффициентов), которую чаще всего используют для оценки вероятности исхода события. Для оценки качества модели логистической регрессии существует эффективный инструмент **ROC-анализа**.

ГЛАВА VIII. КЛАССИФИКАЦИЯ: ДИСКРИМИНАНТНЫЙ И КЛАСТЕРНЫЙ АНАЛИЗЫ

В этой главе будут рассмотрены более глубокие и сложные методы статистического анализа данных. Мы осветим их в общих чертах, менее подробно, чем методы в предыдущих главах.

Довольно часто приходится сталкиваться с необходимостью определить тот или иной изучаемый объект к некой уже определенной группе объектов по некоторым общим свойствам и признакам. Или же наоборот большую группу внешне разнородных объектов разделить на значительно меньшее количество подгрупп, объединив схожие по какому-то количеству признаков объекты. При этом необходимо учитывать довольно большое количество переменных, зачастую больше двух, что может вызвать немалые затруднения в силу сложности математических расчетов. Процедуры, позволяющие осуществить такие манипуляции с большим массивом данных, называются **классификацией**.



Классификация — разделение рассматриваемой совокупности объектов или явлений на однородные в определенном смысле группы.

Приведем простейший пример. Представьте, что у нас есть несколько фруктов. Они отличаются по форме (шарообразная и конусообразная) и по цвету (зеленые и красные) (рисунок 70):



Рисунок 70 — Исходные данные

Как разбить их на группы, объединив похожие плоды? Самый очевидный способ — разделить их по форме (рисунок 71):



Рисунок 71 — Классификация по форме

Но, с другой стороны, можно сгруппировать фрукты по цветам (рисунок 72):



Рисунок 72 — Классификация по цвету

Но можно сформировать больше групп, основываясь на цвете и на форме фрукта (рисунок 73):



Рисунок 73 — Классификация по форме и цвету

А если появится новый, неопознанный фрукт (рисунок 74)?

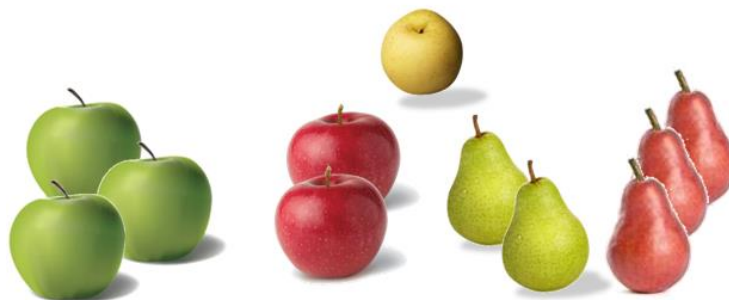


Рисунок 74 — Неопознанный фрукт

В какую группу его отнести? Или выделить под него новую группу?

В научных исследованиях задачи возникают куда более сложные и трудоемкие нежели, чем в выше приведенном примере. При наличии огромных массивов разнородных данных осуществить подобное разделение на группы (классифицировать объекты) — непростая задача.

Различают классификацию при **наличии обучающих выборок** (*с учителем*) (**дискриминантный анализ**) и классификацию **без обучения** (*без учителя*). К классификации *без обучения* относят методы *автоматической классификации* (**кластерный анализ**).

§1. Кластерный анализ

1.1. Определение

В простейшем примере, который был представлен выше, мы объединяли фрукты в группы. *Критерии объединения* в отдельную группу были различными. В первом случае мы объединили фрукты по форме, при этом у нас получилось две группы, так же две группы мы получили, разделив фрукты по цвету, ну а в третьем мы разделили фрукты как по форме, так и по цвету и в результате получили уже четыре группы.



Процесс разделения массива объектов на группы схожих между собой объектов и называется **классификацией** или **кластерным анализом**. А сами группы — **классами**.

Приведем более строгое определение:



Кластерным анализом называются разнообразные формализованные процедуры разделения объектов на группы, называемые классами, при этом сходство между объектами внутри группы (класса) значительно больше, чем между самими классами.



Это значит, что **классы** (довольно часто можно встретить термин **кластер**) выделяются таким образом, чтобы в один класс попадали максимально схожие объекты, а объекты различных классов максимально отличались друг от друга. **Кластер** — это тип объектов, схожих по определенному признаку

С классами (кластерами) вы сталкивались неоднократно, это классы животных в биологии, классы химических элементов. Если вы взглянете на географическую карту и увидите на ней горы или посмотрите на звездное небо и увидите там созвездия, то поймете, что такое кластеры.

На деле *кластерный анализ* — это обобщённое название достаточно большого набора алгоритмов, используемых при создании *классификации*.

Для разбиения объектов на классы рассчитывается некий *количественный показатель сходства*. Он рассчитывается заданным способом на основании данных, характеризующих объекты. И как было продемонстрировано выше, классы могут формироваться в зависимости от интересующих нас характеристик (классификация фруктов по форме или по цвету), а также мы сами можем задавать *глубину* классификации, т. е. увеличивать (или уменьшать) количество классифицирующих признаков (классификация фруктов и по форме, и по цвету, при этом мы получили уже четыре класса). Однако необходимо учитывать, что рассматриваемые объекты могут обладать довольно большим количеством признаков.

Конечно классификацией наука занимается довольно давно, но внедрение статистических методов произошло совсем недавно. Сам предмет *кластерного анализа* (от англ. «cluster» — гроздь, пучок, группа) был сформулирован в 1939 г. психологом Робертом Трионом. Основными идейными разработчиками кластерного анализа являются американские систематики Роберт Сокэл и Питер Снит. Одно из важнейших их достижений в этой области — книга «Начала численной таксономии», выпущенная в 1963 г. В соответствии с основной идеей авторов, *классификация должна строиться не на смешении плохо формализованных суждений о сходстве и родстве объектов, а на результатах, сделанных посредством обработки результатов математических вычислений сходства/отличия классифицируемых объектов*. Для выполнения этой задачи нужны были соответствующие процедуры, разработкой которых и занялись авторы. На данный момент *кластерный анализ* приобрел значительную популярность при анализе больших массивов данных. С его помощью можно значительно упростить исходные данные, объединив схожие объекты в классы и далее исследовать каждый класс отдельно.

Наиболее часто методы кластерного анализа используются в социологии, маркетинговых исследованиях, экономике, биологии, медицине, археологии. Лидирующей же в развитии кластерного анализа наукой является

ся биология. Например, разбиение животных на различные виды, чтобы содержательно описать различия между ними (рисунок 75). В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. Заметьте, что в этой классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. Человек имеет больше сходства с другими приматами (т. е. с обезьянами), чем с «отдаленными» членами семейства млекопитающих (например собаками) и т. д.

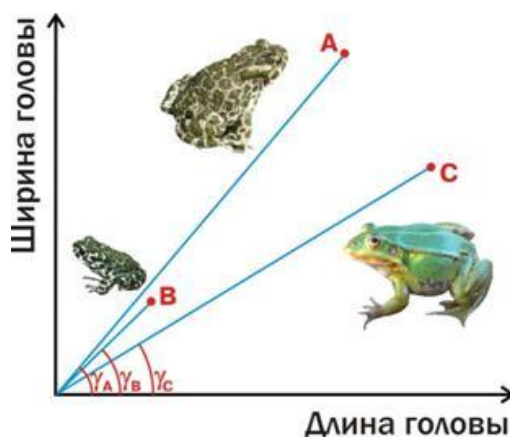


Рисунок 75 — Пример классификации внешне схожих земноводных по двум классифицирующим признакам: длина головы и ее ширина

В биологических науках кластеризация имеет множество приложений в самых разных областях. Например, в биоинформатике с помощью неё анализируются сложные сети взаимодействующих генов, состоящие порой из сотен или даже тысяч элементов. Кластерный анализ позволяет выделить подсети, узкие места, концентраторы и другие скрытые свойства изучаемой системы, что позволяет, в конечном счете, узнать вклад каждого гена в формирование изучаемого феномена.

В области медицины используется кластеризация заболеваний, лечения заболеваний или симптомов заболеваний. В области психиатрии правильная диагностика кластеров симптомов, таких как паранойя, шизофрения и т. д., является решающей для успешной терапии. В археологии с помощью кластерного анализа исследователи пытаются установить таксономию каменных орудий, похоронных объектов и т. д.

1.2. Цели кластеризации

Исходя из вышеописанного, определим основные цели кластерного анализа:

- Понимание данных путём выявления кластерной структуры (разбиение выборки на группы схожих объектов). Что позволяет упростить дальнейшую обработку данных и принятие решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»).

- Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера и рассматривать его отдельно, при этом учитывать, что он характеризует весь свой класс.

- Обнаружение новизны. Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

В первом случае число кластеров стараются сделать поменьше. Во втором случае важнее обеспечить высокую степень сходства объектов внутри каждого кластера, а кластеров может быть сколько угодно. В третьем случае наибольший интерес представляют отдельные объекты, не вписывающиеся ни в один из кластеров.

Во всех этих случаях может применяться **иерархическая кластеризация**, когда крупные кластеры дробятся на более мелкие, те в свою очередь дробятся ещё мельче и т. д. Такие задачи называются задачами таксономии. Результатом таксономии является *древовидная иерархическая структура*. При этом каждый объект характеризуется перечислением всех кластеров, которым он принадлежит, обычно от крупного к мелкому. Внешне это выглядит следующим образом (рисунок 76):

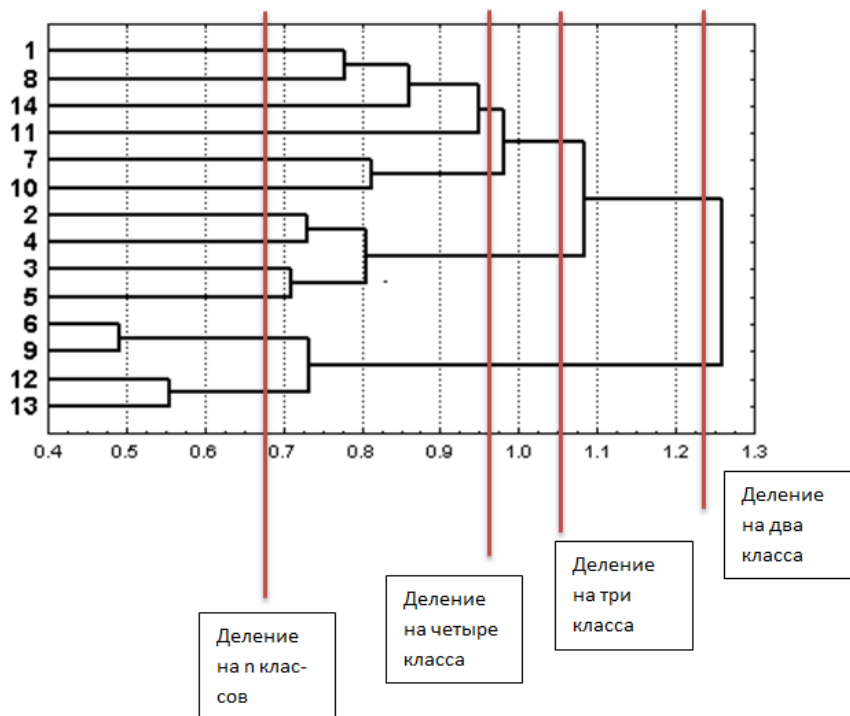


Рисунок 76 — Древоподобная структура с обозначенными классами

Из рисунка видно, что объекты 1 и 8 принадлежат к одному классу, но если подняться на уровень выше, то видно, что к более высокому уровню помимо объектов 1 и 8 принадлежат также объект 14, еще выше — 1, 8, 14 и 11 и так далее до изначального разделения на два класса (объекты 6, 9, 12, 13 относятся к первому классу, остальные — ко второму).



Важно еще раз отметить, что задача кластеризации не является тривиальной. Сложность задач кластерного анализа состоит в том, что реальные объекты являются многомерными, т. е. описываются не одним, а несколькими параметрами, и объединение объектов в группы проводится в пространстве многих измерений, что весьма непросто. Кроме того, данные могут иметь нечисловой характер.

1.3. Этапы кластерного анализа

Перечислим основные этапы кластеризации:

1. Выбор сравнимых друг с другом объектов.
2. Выбор множества признаков (характеристик), по которому будет проводиться сравнение и описание объектов по этим признакам.
3. Вычисление меры сходства между объектами (или меры различия объектов) в соответствии с избранной метрикой.
4. Группировка объектов в кластеры с помощью той или иной процедуры объединения.
5. Проверка применимости полученного кластерного решения (проверка построенной модели).

1.4. Методы кластеризации

Остановимся немного на методах кластеризации, не вдаваясь в сложные теоретические и математические выкладки можно сказать, что в целом методы кластеризации делятся на **агломеративные** (от слова агломерат — скопление) и **итеративные** (дивизивные, от слова division — деление, разделение).

В *агломеративных*, или *объединительных*, методах происходит последовательное объединение наиболее близких объектов в один кластер. Процесс такого последовательного объединения можно показать на графике в виде **дендрограммы**, или **дерева объединения**, с внешним видом которого вы уже познакомились. Это удобное представление позволяет наглядно представить кластеризацию агломеративными алгоритмами.

На каждом шаге ее составления алгоритм находит два самых близких объекта по расстоянию, подсчитанному специальным методом. Условно говоря, расстояние — это определенная мера различия между объектами. Это расстояние откладывается по оси *y*. В итоге, исходя из расстояний на *дендрограмме*, можно определить необходимое количество групп. В нашем примере с фруктами это выглядит следующим образом (рисунок 77):



Рисунок 77 — Дендрограмма

Исходными данными для анализа могут быть как сами объекты и их параметры, так и уже рассчитанная **матрица расстояний** между объектами.



Расстояние между объектами — одна из мер сходства, чем меньше расстояние между объектами, тем они более схожи.

Так же визуальной интерпретацией кластерного анализа могут быть следующие графики (рисунок 78), на которых цветом обозначены классы и объекты, к ним отнесенные.

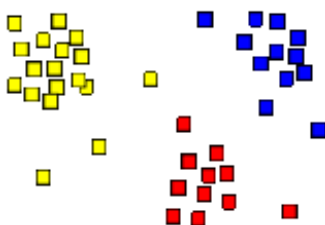


Рисунок 78 — Итог кластерного анализа

Обратите внимание, что количество *классов необходимо выбрать заранее*. А так же на то, что некоторые объекты можно отнести к разным классам, так по своим параметрам они подходят как к одному классу, так как и к другому. Пример такого перекрытия представлен на рисунке ниже (рисунок 79).

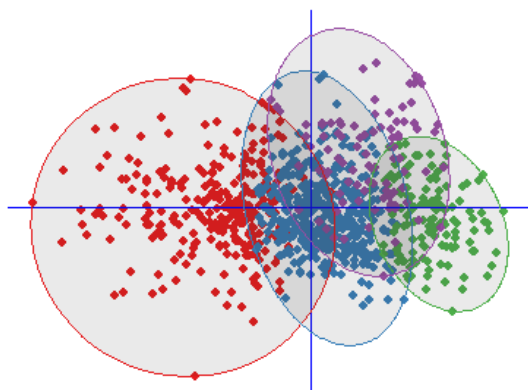


Рисунок 79 — Пример перекрытия классов

Решив применить в своем исследовании кластерный анализ, нужно помнить, что выделенные в ходе анализа кластеры могут быть **неустойчивыми**. Это значит, что от выборки к выборке классификация может несколько меняться. Поэтому нужно проверить результаты на другой группе

объектов или через определенный промежуток времени вычислить погрешность измерения. Более того, лучше всего использовать кластерный анализ на больших выборках, подобранных методом рандомизации или стратификации, ведь только так можно сделать научный вывод, применяя индукцию.

1.5. Пример кластерного анализа

Приведем небольшой пример *кластерного анализа*.

Исходный файл данных содержит следующую информацию об автомобилях и их владельцах (таблица 32):

- марка автомобиля — первая переменная;
- стоимость автомобиля — вторая переменная;
- возраст водителя — третья переменная;
- стаж водителя — четвертая переменная;
- возраст автомобиля — пятая переменная;

Таблица 32 — Исходная таблица данных для кластеризации

	Price	Age	Experience	Age Car
Acura	0,521	25	3	10
Aud	0,866	24	3	1
BMW	0,496	29	3	4
Buick	0,614	50	25	9
Corvette	1,235	62	38	15
Chrysler	0,614	43	21	9
Dodge	0,706	26	1	5
Eagle	0,614	20	1	1
Ford	0,706	54	10	11
Honda	0,429	38	8	7
Isuzu	0,798	27	5	3
Mazda	0,126	51	20	10
Mercedes	1,051	46	25	4
Mitsub.	0,614	28	2	7
Nissan	0,429	31	6	6
Olds	0,614	45	16	4
Pontiac	0,614	40	16	2
Porsche	3,454	41	8	8
Saab	0,588	29	5	2
Toyota	0,059	36	13	1
VW	0,706	38	15	6
Volvo	0,219	42	19	4

Целью данного анализа является разбиение автомобилей и соответствующих им владельцев на классы, каждый из которых соответствует определенной группе риска.

Наблюдения, попавшие в одну группу, характеризуются *одинаковой вероятностью наступления страхового случая*, которая впоследствии оценивается страховой компанией.

В ходе расчетов была получена следующая дендрограмма (рисунок 80):

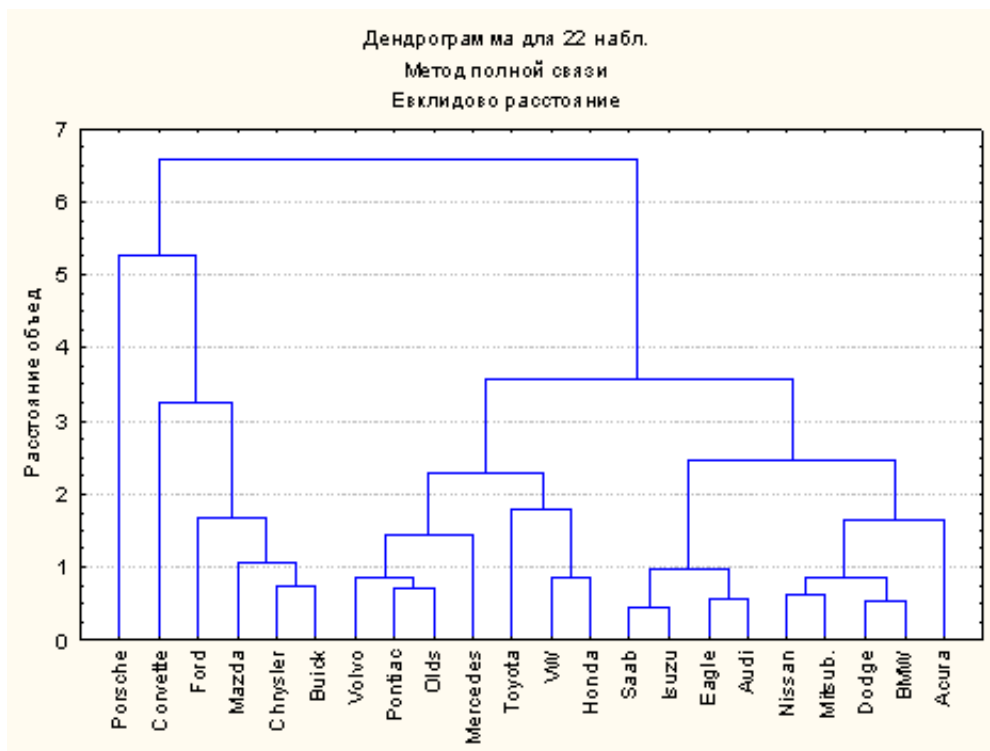


Рисунок 80 — Полученная дендрограмма

Рассмотрение результата кластеризации с помощью дендрограммы начинается сверху (для вертикальной дендрограммы) с каждого автомобиля в своем собственном кластере.

Двигаясь вниз по древовидной структуре видно, что автомобили, которые «теснее соприкасаются друг с другом» объединяются и формируют кластеры. Каждый узел дендрограммы представляет объединение двух или более кластеров, положение узлов на вертикальной оси определяет расстояние, на котором были объединены соответствующие кластеры.

В результате проведения процедуры классификации для каждого кластера экспертом может быть определена вероятность наступления страхового случая.

§2. Дискриминантный анализ

2.1. Определение

В предыдущем параграфе мы рассмотрели цели и возможности одного из методов классификации — *кластерного анализа*. Это так называемая *классификация без учителя*, или, как ее еще называют, *без обучающей выборки*. Другими словами, мы сами задаем количество классов и в зависи-

мости от интересующих нас параметров делим массив объектов на соответствующие классы. При этом, получив древовидную иерархическую структуру, мы можем остановиться на любом уровне классификации. Однако бывает ситуация, когда некий объект, который до этого не участвовал в кластерном анализе необходимо отнести к той или иной группе (кластеру), которые уже имеются в распоряжении исследователя. Такая классификация называется *классификация с учителем* или *классификация с обучающей выборкой*. В этом случае **заранее подготовленные классы и есть такая обучающая выборка**. Исходя из нашего примера с фруктами, мы получили четыре класса, и если появится новый объект, например шаровидный и зеленого цвета, то мы его отнесем к классу зеленых яблок. Подобными задачами занимается **дискриминантный анализ**.

Дискриминантный анализ является одним из методов многомерного статистического анализа.



Цель дискриминантного анализа состоит в том, чтобы на основе измерения различных характеристик (признаков, параметров, **дискриминантных переменных**) объекта *классифицировать* его, т. е. отнести к одной из нескольких заранее определенных групп (классов) некоторым оптимальным способом. Под оптимальным способом понимается либо минимум средних потерь, либо минимум вероятности ложной классификации.

Этот вид анализа является **многомерным**, так как измеряется несколько параметров объекта, по крайней мере больше одного, например температура, влажность в технологическом процессе или давление, состав крови, температура больного и т. д.

Типичные области применения дискриминантного анализа — биология, медицина, управление производством, экономика, геология, контроль качества.

В медицине объектом исследования является пациент, когда по результатам измерений различных параметров, проведения диагностических тестов врач определяет, например, необходимо ли хирургическое вмешательство при лечении. Медик может регистрировать различные переменные, относящиеся к состоянию больного, чтобы выяснить, какие переменные лучше предсказывают, что пациент, вероятно, выздоровел полностью (группа 1), частично (группа 2) или совсем не выздоровел (группа 3).

Биолог может записать различные характеристики сходных типов (групп) цветов, чтобы затем провести анализ дискриминантной функции, наилучшим образом разделяющей типы или группы.

В экономике важно решение по отнесению клиента к определенному классу при выдаче кредита. Для оценки финансового состояния своих кли-

ентов при выдаче им кредита банк классифицирует их по надежности на несколько категорий по ряду признаков. И в случае, когда следует отнести клиента к той или иной категории используют процедуры дискриминантного анализа.

Довольно интересно применение дискриминантного анализа в борьбе с терроризмом. Последствия действий против террористов могут привести к одному из двух результатов: успешное освобождение заложников и случаи, когда заложники пострадали. Здесь имеются следующие *дискриминантные переменные*: число террористов, степень поддержки, количество оружия и т. д. Очевидно, лицам, ведущим переговоры, важно классифицировать возникшую ситуацию, с тем, чтобы определить свои действия.

Очень удобно использовать дискриминантный анализ при обработке результатов тестирования. Так при выборе кандидатов на определенную должность можно всех опрошенных претендентов разделить на две группы — удовлетворяющих и неудовлетворяющих предъявляемым требованиям.

2.2. Задача дискриминантного анализа

Исходя из выше перечисленных примеров, определим задачу дискриминантного анализа. Предположим, имеется n объектов с m характеристиками. В результате измерений каждый объект характеризуется m характеристиками: $x_1 \dots x_m$, $m > 1$. Задача состоит в том, чтобы по результатам измерений отнести объект к одной из нескольких ранее определенных групп (классов) G_1, \dots, G_k , при этом классов естественно должно быть не менее двух ($k \geq 2$).

Иными словами, нужно построить такое *решающее правило*, которое позволяет по результатам измерений параметров объекта указать группу, к которой его следует отнести. Число групп заранее известно, также известно, что объект заведомо принадлежит к определенной группе.



Для отнесения объекта к той или иной группе используют одну или несколько переменных, по которым и определяется классификация. Эти переменные называются *дискриминирующими переменными*.

Рассмотрим простой пример. Предположим, что вы измеряете рост в случайной выборке из 50 мужчин и 50 женщин. Женщины в среднем не так высоки, как мужчины, и эта разница должна найти отражение для каждой группы средних значений для переменной *Рост*. Поэтому переменная *Рост* (*дискриминантная переменная*) позволяет вам провести *дискриминацию* между мужчинами и женщинами лучше, чем, например, вероятность, выраженная следующими словами: «Если человек большой, то это, скорее всего, мужчина, а если маленький, то это, вероятно, женщина».



Основная идея дискриминантного анализа заключается в том, чтобы определить: отличаются ли совокупности по среднему значению какой-либо переменной (или линейной комбинации переменных), и затем использовать эту дискриминирующую переменную, чтобы предсказать для новых членов их принадлежность к той или иной группе.

Другими словами, целью является построение «*модели*», позволяющей лучше всего **предсказать**, к какой совокупности будет принадлежать тот или иной объект.



Поставленная таким образом задача о дискриминантной функции может быть перефразирована как задача *дисперсионного анализа*. Можно спросить, в частности, являются ли две или более совокупности значимо отличающимися одна от другой по среднему значению какой-либо конкретной переменной.

В дискриминантном анализе можно задать **априорные вероятности** принадлежности объекта к определенному классу. На практике эти вероятности оцениваются из массива экспериментальных данных.

Так как массив экспериментальных данных накапливается, то оценки постепенно уточняются. При этом можно учесть различные факторы, влияющие на принадлежность объекта к определенному классу, например, если поступает мука в хлебное производство, то можно учесть сезонные факторы: вероятность того, что мука будет лучшего качества осенью выше той же вероятности весной.

2.3. Предположения и ограничения

Дискриминантный анализ «работает» при выполнении ряда предположений:

- Так как задача дискриминации сводится к различию средних значений и схожа с методами *дисперсионного анализа*, то важным является предположение о том, что наблюдаемые величины — измеряемые характеристики объекта — имеют *нормальное распределение*. Это предположение следует проверять. Впрочем, умеренные отклонения от этого предположения не являются фатальными.

- Предположение об *однородности* дисперсий наблюдаемых переменных в разных классах (отличие между классами имеется только в средних). Умеренные отклонения от этого предположения также допустимы.

Дискриминантный анализ может быть проведен и когда основные предположения не выполняются (предположение о нормальности и однородности дисперсий). Задача состоит в интерпретации результатов. В конечном счете, наиболее важным критерием правильности построенного *классификатора* является практика. И если окажется, что в результате построен *классификатор*, «работающий» на практике, то это будет достижением.

ГЛАВА IX. ФАКТОРНЫЙ АНАЛИЗ

§1. Понятие факторного анализа

Довольно интересным методом статистического анализа является анализ **факторный**. Данный тип анализа достаточно сложен как с математической точки зрения, так и теоретически. Поэтому его мы так же рассмотрим в общих чертах.

1.1. Определение

Факторный анализ — это так же многомерный метод анализа, применяемый для изучения взаимосвязей между значениями переменных (признаков) и влиянием на них **скрытых факторов**. При этом предполагается, что известные переменные зависят от меньшего количества неизвестных переменных и случайной ошибки.

Более строго определение звучит следующим образом:



Факторный анализ (от лат. factor — действующий, производящий и греч. analysis — разложение, расчленение) — совокупность методов, которые на основе объективно существующих корреляционных взаимосвязей признаков (или объектов) позволяют выявлять латентные (скрытые) обобщающие характеристики структуры изучаемых объектов и их свойств.

Другими словами, весь массив исследуемых данных можно сократить, объединив связанные между собой (коррелирующие) признаки в отдельные факторы.

1.2. Задачи и возможности факторного анализа

Факторный анализ позволяет решить две важные проблемы исследователя. Во-первых, описать объект измерения *компактно*, тем самым сократив объем анализируемых данных и в то же время *всесторонне*. Во-вторых, с помощью факторного анализа возможно выявление *скрытых переменных (факторов)*, отвечающих за наличие линейных статистически значимых взаимосвязей между наблюдаемыми переменными.

Таким образом, можно выделить 2 цели факторного анализа:

- определение взаимосвязей между переменными, в некотором смысле — это классификация переменных по степени их корреляции между собой;
- сокращение числа переменных, необходимых для описания данных. Под сокращением числа переменных (**редукцией**) понимается переход от многих исходных количественных признаков, характеризующих исследуемый объект к **пространству факторов**, число которых в свою очередь

значительно меньше числа исходных количественных признаков. Например, от исходных 10–20 количественных признаков производится переход к 3–5 факторам. Факторный анализ позволяет обнаружить некоторые взаимосвязи между переменными и избавиться от лишнего «шума», который может возникнуть и накапливаться при рассмотрении всех переменных без объединения их в факторы. Такая «очищенная» от «шума» структура взаимосвязей позволяет сконцентрировать в некоторых из новых переменных-факторах значительно больше информации, нежели в отдельно взятой исходной количественной переменной. Благодаря этому отдельный фактор можно рассматривать как совокупность наиболее сильно взаимосвязанных между собой исходных признаков. Вследствие этого факторный анализ позволяет вычленил из всего многообразия исходных признаков отдельные группы таких взаимосвязанных исходных признаков. При этом сформированный фактор можно описать как-то более обобщенно и емко и в тоже время более точно, применив к нему соответствующее прилагательное или наименование (гормональный, спирометрический, гематологический, гуморальный и т. п.).

Итак, с одной стороны выделение факторов позволяет выделить подгруппы взаимосвязанных количественных признаков, а с другой — более наглядно представить взаимное расположение имеющихся подгрупп наблюдений (состояние здоровья (здоров, болен), группы патологии, степень тяжести заболевания и т. п.) в осях этих новых, более информативных признаков.



При анализе в один фактор объединяются сильно коррелирующие между собой переменные, как следствие происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов. После объединения *коррелированность* признаков внутри каждого фактора между собой будет выше, чем их коррелированность с компонентами из других факторов. Эта процедура также позволяет выделить латентные переменные.

Например, анализируя оценки, полученные по нескольким шкалам, исследователь замечает, что они сходны между собой и имеют высокий коэффициент корреляции, он может предположить, что существует некоторая латентная переменная, с помощью которой можно объяснить наблюдаемое сходство полученных оценок. Такую **латентную переменную** называют **фактором**. Данный фактор влияет на многочисленные показатели других переменных, что приводит нас к возможности и необходимости выделить его как наиболее общий, более высокого порядка.



Основной смысл факторного анализа заключается в выделении из всей совокупности переменных только небольшого числа латентных независимых друг от друга группировок, внутри которых переменные связаны сильнее, чем переменные, относящиеся к разным факторам.

1.3. Условия применения и алгоритм факторного анализа

Практическое выполнение факторного анализа начинается с проверки его условий. В обязательные условия факторного анализа входят:

- Все признаки должны быть количественными.
- Число наблюдений должно быть не менее чем в два раза больше числа переменных.
- Выборка должна быть однородна.
- Исходные переменные должны иметь нормальное распределение.
- Факторный анализ осуществляется по коррелирующим переменным.

Алгоритм факторного анализа довольно сложен. Поэтому мы еще раз отметим, что в отдельные факторы объединяются переменные, сильно взаимодействующие между собой, т. е. имеют высокую степень корреляции. Что и позволяет объединить их в группу (означенную как фактор).

1.4. Ошибки использования факторного анализа

Ошибки использования факторного анализа можно условно разделить на две части: **ошибки применения** и **ошибки описания**. В первом случае авторы достаточно подробно описывают процедуру факторного анализа, не понимая при этом, что выполненный ими анализ ошибочен, поскольку авторская реализация противоречит основным положениям факторного анализа. Во втором случае авторы просто констатируют сам факт использования факторного анализа, не сообщая читателю достаточной информации, необходимой для критической оценки надёжности описываемых выводов. В этом случае даже корректно выполненный факторный анализ может не достичь своей цели, поскольку читатель не имеет возможности оценить по достоинству описанные результаты.

ГЛАВА X. АНАЛИЗ ВЫЖИВАЕМОСТИ

1.1. Постановка задачи

Наиболее важной и объективной характеристикой метода лечения или лекарственного препарата является средняя продолжительность жизни пациентов с момента начала лечения или средняя продолжительность ремиссии заболевания. Для описания средних времен жизни или ремиссии можно было бы использовать стандартные параметрические и непараметрические методы. Однако в анализируемых данных есть существенная особенность — могут найтись пациенты, которые в течение всего периода наблюдения выжили или заболевание все еще находится в стадии ремиссии. Так же может образоваться группа больных, контакт с которыми был потерян еще до завершения эксперимента (например, их перевели в другие клиники). При использовании стандартных методов оценки среднего, о которых мы говорили выше, эту группу пациентов пришлось бы исключить, тем самым, потеряв с трудом собранную важную информацию. К тому же, большинство этих пациентов являются «выжившими» (выздоровевшими) в течение того времени, которое их наблюдали, что скорее всего свидетельствует в пользу нового метода лечения (лекарственного препарата). Потеря с трудом собранной информации нежелательна и это плохо сказывается на результатах исследования. Поэтому желание исследователя использовать все данные, т. е. анализировать как *полные времена жизни*, так и *неполные*, является вполне естественным.

Для того, чтобы избежать подобных проблем и предназначены **методы анализа выживаемости**, которые позволяют изучать **неполные**, или так называемые **цензурированные** данные. Впервые метод *анализа выживаемости* был использован для проведения анализа времен жизни больных лептоспирозом, где были выявлены статистические закономерности течения заболевания, обоснованы периоды времени наиболее опасные для жизни. И свое название метод получил в результате первоначального использования в медицинских, биологических исследованиях и страховании, но затем стал широко применяться в социальных и экономических науках, а также в промышленности в инженерных задачах — анализ надежности и времен отказов³⁶.

1.2. Определение

Определим основную задачу *анализа выживаемости*:



Анализ выживаемости занимается моделированием процессов наступления **терминальных** (критических) событий для элементов той или иной совокупности (изначально таким критическим событием был выбран смертельный исход для элементов совокупности живых существ).

³⁶ StatSoft.Inc. (2001). Электронный учебник по статистике. М. StatSoft. WEB: <http://www.StatSoft.ru/home/textbook/default.htm>.

Так, в рамках медицинских исследований анализ выживаемости может отвечать на такие вопросы, как «какова будет доля выживших среди пациентов спустя некоторое время после применённых техник лечения?», «какие темпы смертности будут наблюдаться среди выживших?», «какие факторы воздействуют на увеличение или уменьшение шансов на выживание?» и проч.



Однако в целом анализ выживаемости представляет собой построение моделей, описывающих данные о времени наступления совершенно любого события, не обязательно связанного с отрицательным результатом. Но терминология осталась той же, что и при наблюдении до смертельного исхода.

Для ответа на соответствующие вопросы необходимо иметь возможность чётко определить «**время жизни**» объекта наблюдения³⁷. В случае с биологическим выживанием «смерть» однозначна, но в иных случаях наступление терминального события не всегда возможно локализовать в отдельном моменте времени.

Так как живой организм может умереть лишь один раз, то традиционно в рамках данного подхода рассматриваются лишь единичные и единовременные терминальные события.



Наблюдения от момента операции до летального исхода называются **полными**. Наблюдения, которые содержат неполную информацию, называются **неполными** или **цензурированными**.

Например, «пациент А был жив по крайней мере 3 года после того, как переехал в другой город и контакт с ним был потерян». Это пример *цензурированного наблюдения*: информация о том, что пациент был жив 3 года, важна и может быть использована для построения оценок³⁸.

Использование цензурированных наблюдений составляет специфику рассматриваемого метода. С его помощью исследуются вероятностные ха-

³⁷ Период пребывания объекта наблюдения в исследуемой совокупности до наступления терминального события

³⁸ Конечно, можно было использовать только полные времена жизни и не учитывать тех участников исследования, которые выбыли из исследования, но тогда мы имели бы в своем распоряжении очень мало наблюдений и соответственно неточные оценки. Использование, наряду с полными наблюдениями, неполных или цензурированных наблюдений является главной особенностью методов анализа выживаемости.

рактические интервалы времени между последовательным возникновением критических событий³⁹.

Для визуальной демонстрации «времени жизни» объекта наблюдения удобно использовать следующий график (рисунок 81):

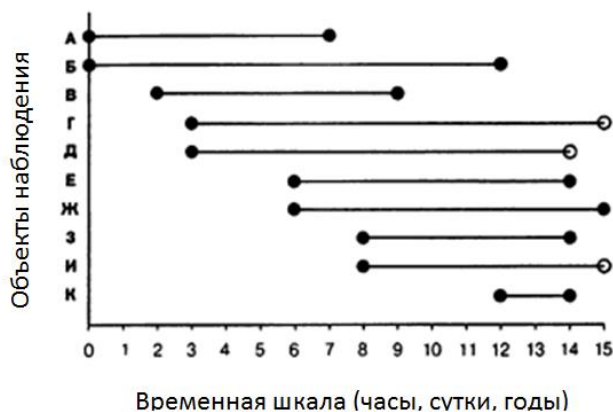


Рисунок 81 — График, демонстрирующий «время» жизни объекта наблюдения

На рисунке ход исследования показан в обычной шкале времени. Жизнь пациента представлена горизонтальным отрезком. Левый конец отрезка — это начало наблюдения. На правом конце отрезка — черный или белый кружок. Черный кружок означает, что пациент умер (наступило событие) и, таким образом, продолжительность его жизни нам известна. Белый кружок означает, что исследование закончилось до его смерти либо он куда-то уехал — словом, выбыл из-под наблюдения. Относительно выбывших нам известно только то, что они прожили не меньше определенного срока.

Зачастую ход исследования удобнее представить так, будто все пациенты начали наблюдаться одновременно. Теперь на шкале времени не просто астрономическое время, а время наблюдения каждого пациента (рисунок 82).

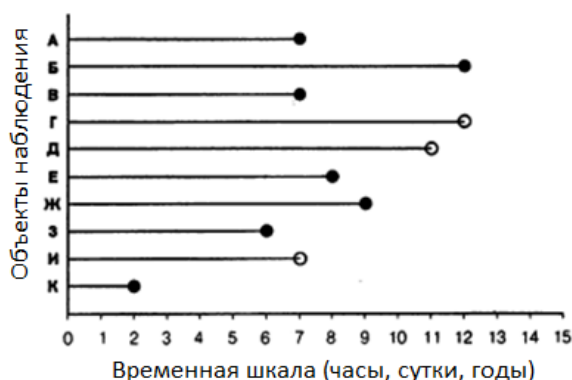


Рисунок 82 — График времени наблюдения каждого пациента

³⁹ Такого рода исследования еще называют *анализом длительностей до момента прекращения* (duration или time until failure), которые можно определить как интервалы времени между началом наблюдения за объектом и моментом прекращения (failure), при котором объект перестает отвечать заданным для наблюдения свойствам. Цель исследований — определение условных вероятностей, связанных с длительностями до момента прекращения.

1.3. Требования к анализу выживаемости

Для анализа выживаемости используются особые методы. Но прежде чем в общих чертах с ними познакомиться, необходимо сформулировать требования, которым должны удовлетворять все исследования выживаемости:

- для всех наблюдаемых известно время начала наблюдения;
- для всех наблюдаемых известно время окончания наблюдения, а также — умер он или выбыл;
- выбор наблюдаемых произведен случайно.

1.4. Кривая выживаемости

*«На достаточно большом промежутке времени шанс на выживание каждого из нас приближается к нулю»
Фильм «Бойцовский клуб»*

Важным моментом при решении задач на выживаемость является построение **кривой выживаемости**.



Кривая выживаемости задает (изображает графически) вероятность пережить любой из моментов времени после некоторого начального события. Эту вероятность обычно называют просто выживаемостью. **Выживаемость $S(t)$** — это вероятность прожить более определенного времени t с момента начала наблюдения.

Для всей генеральной совокупности исследуемых объектов эта величина рассчитывается следующим образом:

$$S(t) = \frac{\text{Число переживших момент } t}{\text{Объем совокупности}}$$

Часто *кривая выживаемости* применяется для изучения продолжительности жизни. Однако кривыми такого рода можно описать продолжительность самых разнообразных процессов. Тогда в качестве исхода будет выступать не смерть, а другое интересующее нас событие, не всегда нежелательное. Например, можно изучать срок лечения какого-либо заболевания (исход — ремиссия), длительность лечения бесплодия или эффективность контрацепции (исход в обоих случаях — наступление беременности), долговечность протеза (исход — поломка).

Изобразим кривую выживаемости (рисунок 83):

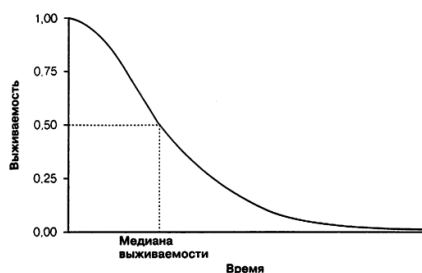


Рисунок 83 — Внешний вид кривой выживаемости

Интерпретация рисунка несложная. По вертикальной оси отчается значение выживаемости, выраженное в долях или процентах. Максимальное значение по этой шкале равно 1 (если размерность — доли) или 100 % (если размерность — проценты). По горизонтальной оси откладывается время наблюдения. В начальный момент времени выживаемость естественно равна 1, постепенно с течением времени это значение будет уменьшаться. При очень больших значениях времени кривая стремится к нулю, что говорит о низкой выживаемости на большом отрезке времени и большой вероятности наступления критического события.



Время, за которое значение функции выживаемости примет значение равное 0,5, называется **медианой выживаемости**.

Для выборочных данных кривая выживаемости будет иметь несколько иной вид (рисунок 84):

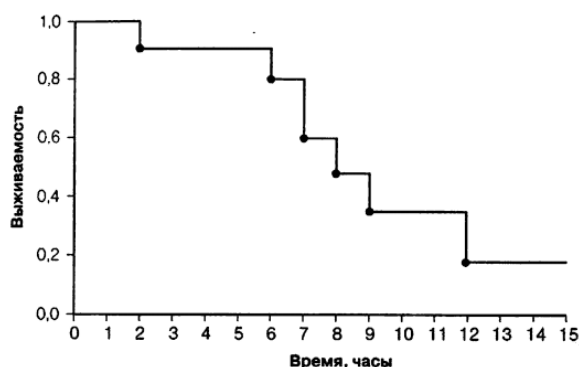


Рисунок 84 — Кривая выживаемости для выборочных данных

Это связано с тем, что при подобных исследованиях мы почти всегда имеем дело с вышеописанными цензурированными данными. Так как часть данных может быть утеряна в процессе проведения эксперимента, и мы знаем только то, что на какой-то момент времени до момента утраты возможности наблюдения объекты наблюдения были еще живы (терминальное событие для них еще не наступило). В тех случаях, когда имеет место выбывание, мы не сможем воспользоваться выше приведенной формулой. Вместо этого для каждого момента времени, когда произошла хотя бы одна смерть, оценим вероятность пережить этот момент. Такой оценкой будет отношение числа переживших этот момент к числу наблюдавшихся к этому моменту. Тогда, согласно правилу умножения вероятностей, вероятность пережить некоторый момент времени для каждого вступившего в исследование будет равна произведению этих оценок от нулевого до данного момента. Каждой ступеньке на кривой выживаемости соответствует наступление критического события для одного или нескольких членов выборки.

При исследовании выборки полученная нами кривая выживаемости на самом деле представляет собой **оценку кривой выживаемости**. Поэтому необходимо *определить стандартную ошибку* и *доверительный интервал* кривой выживаемости. Отложив на графике доверительные интервалы выживаемости для каждого момента времени, мы увидим расширяющийся «рукав» — доверительную область для выживаемости. Причину расширения доверительной области можно объяснить тем, что чем меньше остается наблюдаемых, тем больше ошибка (рисунок 85).

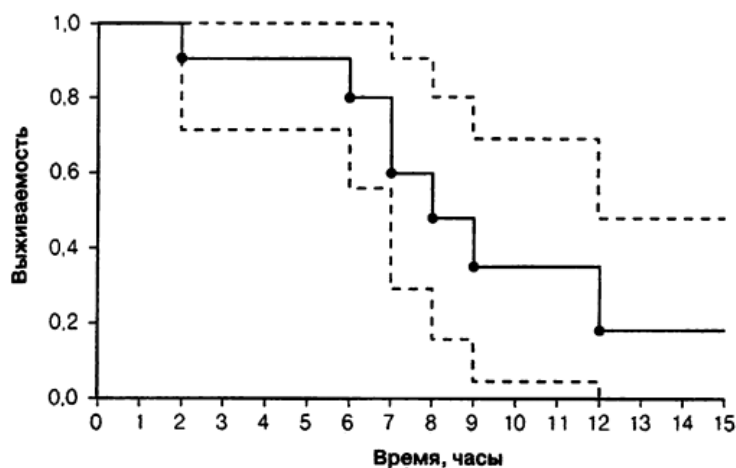


Рисунок 85 — Кривая выживаемости и доверительные области для этой кривой

1.5. Сравнение двух кривых выживаемости

Приведем практический пример. Часто необходимо сравнить несколько кривых выживаемости. Для каждой исследуемой группы строится своя кривая выживаемости, затем кривые наносятся на общий график и происходит сравнение кривых. Естественно, если кривые отличаются друг от друга, то необходимо проверить насколько значимо это различие. Следующий пример иллюстрирует решение данной проблемы.

Трансплантация костного мозга при остром лимфобластном лейкозе взрослых (по книге С. Гланц — «Медико-биологическая статистика»)

При остром лимфобластном лейкозе мутация предшественника лимфоцитов приводит к появлению клона лейкозных клеток, способных неограниченно делиться. В отличие от обычных лимфоцитов, лейкозные клетки функционально неактивны и не обладают защитными свойствами. Размножаясь в костном мозге, они подавляют нормальное кроветворение, в результате развиваются иммунодефицит, анемия и тромбоцитопения. Без лечения острый лимфобластный лейкоз неизбежно приводит к смерти. Задача лечения — полностью уничтожить лейкозные клетки. Этого можно достичь с помощью облучения и химиотерапии. Однако при этом уничтожаются и нормальные кроветворные клетки. Чтобы компенсировать это

побочное действие лечения, используют трансплантацию костного мозга. Для трансплантации лучше всего подходит костный мозг близкого родственника (аллотрансплантация). К сожалению, не всегда есть у кого его взять. Поэтому применяется и другой способ, так называемая ауто-трансплантация, когда костный мозг берут у самого больного. Из полученного костного мозга специальными методами удаляют лейкозные клетки и по завершении курса лучевой и химиотерапии его вновь вводят больному. Н. Вей с соавт. сравнили выживаемость после ауто- и аллотрансплантации. В исследование включали больных старше 15 лет с подтвержденным диагнозом острого лимфобластного лейкоза после достижения первой полной ремиссии. Больным, у которых не было подходящих родственников, проводили ауто-трансплантацию (1-я группа), остальным — аллотрансплантацию (2-я группа). Исследование продолжалось 11 лет. Полученные данные представлены в таблице 33 ниже. Выбывшие в ходе исследования помечены знаком «+». После проведения расчета выживаемости для каждой из групп были получены кривые, изображенные на рисунке 86. Видно, что выживаемость в 1-й группе хуже, чем во 2-й (кривая у пациентов с ауто-трансплантацией ближе к нулевой отметке, чем кривая у пациентов с алло-трансплантацией). Необходимо понять на сколько значимо это различие.

Таблица 33 — Результаты исследования методов трансплантации костного мозга

Ауто-трансплантация (1-я группа n = 33)		Алло-трансплантация (2-я группа n = 21)	
Месяцы после пересадки	Число смертей и выбытий	Месяцы после пересадки	Число смертей и выбытий
1	3	1	1
2	2	2	1
3	1	3	1
4	1	4	1
5	1	6	1
6	1	7	1
7	1	12	1
8	2	15+	1
10	1	20+	1
12	2	21+	1
14	1	24	1
17	1	30+	1
20+	1	60+	1
27	2	85+	2
28	1	86+	1
30	2	87+	1
36	1	90+	1
38+	1	100+	1
40+	1	119+	1
45+	1	132+	1
50	3		
63+	1		
132+	2		

Построим кривые выживаемости для обеих групп.

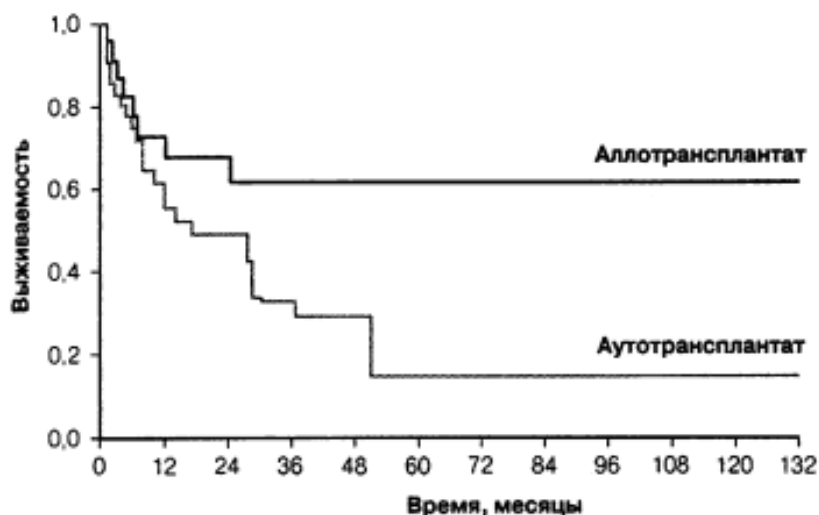


Рисунок 86 — Кривые выживаемости, полученные в ходе исследования

Нулевая гипотеза состоит в том, что в обеих группах выживаемость одинакова. Если бы не было выбывания и все больные наблюдались равное время, то вполне подошел бы анализ таблиц сопряженности из раздела «Анализ качественных признаков». Если бы все больные наблюдались до самой смерти, можно было бы сравнить выживаемость в обеих группах с помощью непараметрических методов для сравнения групп, например *критерия Манна — Уитни* или *Крускала — Уоллиса*. В реальной жизни подобные ситуации редки, и, как мы уже говорили, выбывание практически неизбежно. Для сравнения кривых выживаемости нужны специальные методы. Здесь мы не будем вдаваться в подробности математических выкладок. Алгоритм их схож с методами, приведенными в предыдущих главах. В ходе расчетов мы находим соответствующий *критерий значимости* (**логранговый критерий** или **критерий Гехана**). А также его *критическое значение* для заданного уровня значимости. После вычисления критерия значимости и сравнения его с критическим делается вывод о значимости полученных различий. Если эмпирический критерий оказался больше критического, то различия кривых следует признать значимыми.

В заключение отметим, что значение анализа выживаемости чрезвычайно велико. Если, например, препарат снижает уровень холестерина, то это еще не значит, что он позволяет продлить жизнь больного или отдалить появление стенокардии, — речь, следовательно, идет о показателе процесса. Напротив, если доказано, что препарат продлевает жизнь, то речь идет о показателе результата, имеющем несомненную клиническую значимость.

Глава XI. Методика построения исследования в медицине

На страницах данного пособия вы познакомились со многими методами статистической обработки, узнали том, в каких случаях одни методы применимы, а в каких — нет. В начале книги мы вскользь коснулись проблемы *планирования эксперимента и построения исследования*, частично указав на вероятностную оценку полученного результата анализа, важность того, каким образом формируются выборочные данные и т. д. В этой главе мы поговорим об этих проблемах более подробно и более тесно свяжем методику планирования исследования с практикой доказательной медицины.

§1. Основные цели и задачи доказательной медицины

Напомним вкратце основные проблемы, с которыми сталкивается современная *доказательная медицина*.

Современное общество характеризуется лавинообразно возрастающим количеством информации и не меньшими требованиями к ее качеству и, в конечном счете, полезности. Такие же требования предъявляются и к технологиям. Значительно улучшенное качество жизни подразумевает высокие стандарты в сфере обслуживания, к коей в определенном смысле относится и здравоохранение. Растут требования пациентов к качеству медицинского обслуживания, не меньшие требования предъявляются и к лекарственным препаратам. Однако с примерами использования ненадежной медицинской информации как пациенты, так и врачи сталкиваются ежедневно. Вспомните многочисленные рекламные ролики, говорящие о новых препаратах, но при этом эффективность данных препаратов часто может быть подвергнута сомнению. Может возникнуть вопрос, была ли вообще строго проверена их эффективность, проводилось ли серьезное исследование прежде чем препарат поступил в продажу и т. д. Если для пациента данный вопрос может быть не столь важен, в том смысле, что он зачастую доверяет преподносимой информации, то для практикующего врача знания об этих подводных камнях является обязательным. Необходимо понимать, каким образом были получены результаты, на каких больных проводились исследования, какой был тип исследования и т. д. Ко всему прочему сами исследования в сфере врачебной практики обязаны отвечать не меньшим требованиям: правильно построены, а выводы, часто жизненно важные, полученные в результате исследования, должны быть максимально приближены к действительности, при этом вероятность ошибочного вывода должна быть сведена к минимуму.

Для подведения итога вышесказанного дадим определение доказательной медицины.



Доказательная медицина (англ. Evidence-based medicine — медицина, основанная на доказательствах) — термин описывает такой подход к медицинской практике, при котором решения о применении профилактических, диагностических и лечебных мероприятий принимаются исходя из полученных доказательств, их эффективности и безопасности, и предполагающий поиск, сравнение, обобщение и широкое распространение полученных доказательств для использования в интересах больных.

Можно дать и другие определения, отражающие тот же смысл. **Доказательная медицина** — это совокупность методологических подходов к проведению клинических исследований, оценке и применению их результатов. В узком смысле **доказательная медицина** — это способ (разновидность) медицинской практики, когда врач применяет в ведении пациента только те методы, полезность которых доказана в доброкачественных исследованиях. Если совсем упростить, то можно сказать, что это медицина, основанная на методах, чья эффективность доказана (В. В. Власов, Е. Н. Семернин, П. В. Мирошенков Доказательная медицина и принципы методологии).

Из определения видно, что одним из ключевых моментов является проведение «**доброкачественного исследования**». Опишем общую схему такого исследования и так же разберем некоторые из этапов, отметив при этом встречающиеся в ходе их проведения сложности и ошибки (рисунок 87).

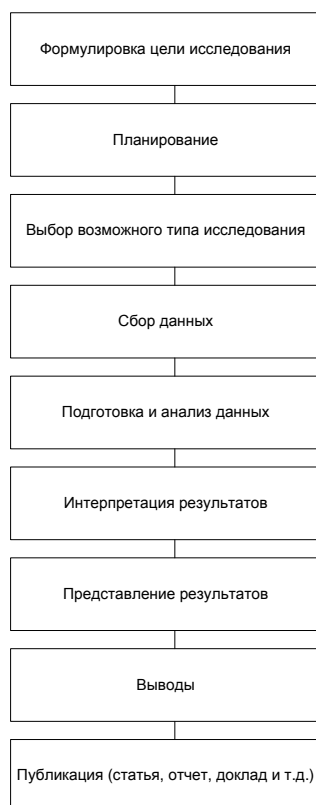


Рисунок 87 — Общая схема исследования

§2. Этапы медицинского исследования

2.1. Формулировка цели исследования

Прежде чем начать исследование необходимо четко понимать его цель. Уже на этом этапе можно сделать достаточно серьезные замечания. С формулировкой цели исследования вроде все понятно, но, увы, у немалой части исследований сбор данных предшествует формулировке вопроса, на который они должны собственно ответить. В такой ситуации исследователя неизменно подстерегают малоприятные моменты. Ситуация может выглядеть следующим образом: данные собраны и остается только вычислить значение p (вероятность ошибочного результата), но тут обнаруживается, что это значение существует не само по себе, а лишь в связи с проверкой «какой-то там гипотезы». И самое обескураживающее — чтобы проверить гипотезу, ее, оказывается, нужно заранее сформулировать.

Не многие исследователи обременяют себя необходимостью еще до начала сбора данных осознать цели исследования и подлежащие проверке гипотезы.



Очень важно понять, что вы хотите от исследования, какой вопрос необходимо решить. Когда будет сформулирована конкретная цель станет понятно, каким должен быть тип предстоящего эксперимента и какие потребуются данные. Тогда и легко определится нужный метод анализа, принцип сбора данных, достаточных для анализа и т. д.

Немногие поступают таким образом. Поэтому неудивительно, что, когда настает время проводить статистический анализ и рассчитать критерии значимости и вероятность p , исследователь обнаруживает, что собранные им данные мало связаны с проверяемой гипотезой, и вообще нарушают предпосылки известных ему статистических методов анализа. Но не начинать же все с начала! Поэтому для устранения и сглаживания статистических несообразностей на этом завершающем этапе призывается специалист, который оставляет от массива собранных данных немного, хоть как-то пригодное для анализа, заменяет неприменимые *параметрические методы* неприменимыми, но менее чувствительными *непараметрическими* или предлагает вместо одной гипотезы перейти к нескольким, пригодным для статистической проверки. В результате отчет об исследовании приобретает приемлемый вид. Однако само исследование не становится более осмысленным. И результат его выглядит несколько подозрительным и, возможно, натянутым. Способ избежать этого прост и состоит в том, чтобы задуматься о том, как анализировать данные: в начале, а не в конце исследования.

2.2. Планирование исследования

Выше этот этап уже был частично рассмотрен. Стоит сказать сразу, что он очень зависит от предыдущего этапа. От правильно спланированного исследования во многом зависит точность выводов в результате его проведения. До возникновения *доказательного подхода* в медицине исследования не планировались, а чаще всего «случались». Другими словами выбор типа исследования, сбор данных, формирование выборки происходило не по строгому плану, а «как получится» уже по ходу исследования. Нет смысла говорить, что без четко очерченного плана исследования сам процесс может получиться довольно скомканным и подверженным многим случайным факторам, не учтенным из-за отсутствия плана реализации.

Планирование исследования условно можно разбить на два этапа:

- Определение типа исследования, обеспечение достоверности и возможности обобщения результатов, сведение к минимуму возможности случайных ошибок.
- Определение объемов выборок.

2.3. Типы исследования

Дадим определения основным типам исследования, используемым в клинической практике⁴⁰:

- **Описание отдельных случаев** — состоит в описании редкого наблюдения, так называемого «классического» случая («классические» случаи, кстати, никогда не бывают частыми) или *нового феномена*. Это наиболее старый способ медицинского исследования. Несмотря на узкую направленность данного способа исследования он также важен в медицине, так как описание редких случаев или явлений никак нельзя недооценивать.

- **Описание нескольких объектов** — исследование, включающее обычно описательную статистику группы больных, отобранных по какому-либо исследуемому признаку. Описательные исследования используются, например, в эпидемиологии для изучения влияния неконтролируемых факторов на возникновение заболевания.

- **Ретроспективное исследование** (исследование *случай — контроль*) — исследование, в котором по архивным данным или опросу его участников формируют группы из этих участников (больных) с определенным заболеванием и без него, а затем оценивается сила воздействия исследуемого фактора. Такие исследования чаще выдвигают научные гипотезы, а не проверяют их. Преимуществом исследования данного типа являются его относительная простота, дешевизна и быстрота выполнения. Однако подобные исследования чреваты возникновением множества возможных систематических ошибок. Наиболее существенными из них можно считать

⁴⁰ По О. Ю. Реброва Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA.

систематические ошибки, связанные с отбором участников исследования, и систематическую ошибку, возникающую при измерении.

• **Одномоментное (поперечное) исследование** — описательное исследование, включающее однократно обследуемые группы участников и проводимое с целью оценки распространенности того или иного исхода, течения заболевания, а также эффективности диагностики. Наличие или отсутствие болезни (или наличие или отсутствие других признаков) определяется у каждого члена выборки в один момент времени (<http://medspecial.ru>). Такие исследования относительно просты и недороги. Основной проблемой является трудность формирования выборки, адекватно отражающей типичную ситуацию в изучаемой популяции больных (репрезентативной выборки).

• **Проспективное (когортное, продольное) исследование** — исследование, в котором выделенная группа участников наблюдается в течение определенного времени. Сначала выделяют группу (или несколько групп, например лиц, подвергшихся **фактору риска**⁴¹, и лиц, не подвергшихся ему, а затем проводят наблюдение за ней (ними) и сбор данных. Другими словами группы формируются еще до начала проведения наблюдения или эксперимента. В этом заключается отличие от ретроспективного исследования, в котором группы выделяют после сбора данных. Такой вид исследований используют для выявления факторов риска, прогностических факторов, причин заболеваний, для определения уровня заболеваемости. Проспективные исследования весьма трудоемки, так как должны проводиться в течение длительного времени, выборки должны быть достаточно велики в связи с тем, что выявляемые события (например, возникновение новых случаев заболевания) могут быть достаточно редки. Основные проблемы, возникающие при проведении *проспективного исследования*, следующие:

✓ вероятность изучаемых событий зависит от способа формирования выборки (например, наблюдаемые участники из группы риска имеют большую вероятность заболеть, чем участники из неорганизованной популяции);

✓ при выбывании участников в ходе исследования необходимо выяснять, не связано ли это с изучаемым исходом или фактором; с течением времени могут изменяться сила и характер воздействия изучаемого фактора (например, интенсивность курения как фактора риска развития ишемической болезни сердца);

✓ необходимо добиваться одинакового объема обследования групп воздействия и контроля, чтобы свести к минимуму возможность более раннего выявления заболеваний (следовательно, лучшего прогноза) в более тщательно обследуемой группе.

⁴¹ **Фактор риска** — это какое-либо свойство или особенность человека или какое-либо воздействие на него, повышающее вероятность развития болезни или травмы.

• **Рандомизированное контролируемое исследование** — при данном типе исследования участники случайным образом делятся на группы, в одной из которых проводится исследуемое вмешательство (наличествует исследуемый фактор), а в другой (контрольной) применяются стандартные методики или плацебо (например, исследование проводится с целью оценки эффективности и безопасности нового лекарственного препарата или расширения показаний к применению уже известного лекарственного препарата). Рандомизированные контролируемые испытания проводятся с середины XX в. Они являются основой доказательной медицины. Наиболее известный вариант рандомизированного исследования — **клиническое испытание**⁴².

• **Мета-анализ** — анализ объединенных результатов нескольких *клинических испытаний* одного и того же вмешательства при одном и том же заболевании. Такой подход обеспечивает большую статистическую чувствительность (мощность), чем в каждом отдельном исследовании, за счет увеличения размера выборки. *Мета-анализ* используется для обобщения результатов многих испытаний, зачастую противоречащих друг другу.

Как видно типов исследования не мало, и каждый из них имеет свои недостатки и положительные стороны. Поэтому выбор возможного типа анализа является крайне важной процедурой при построении исследования.

2.4. Выборка

О выборке мы много говорили выше. Правильность формирования выборки нельзя переоценить. Однако необходимо еще раз обратить внимание на важнейшее свойство выборки — ее **представительность** (*репрезентативность*). Любой статистический метод исходит из предположения, что выборка извлечена из совокупности **случайно** (рандомизация). Это основное условие репрезентативности выборки. Если же это условие не выполняется (то есть если выборка непредставительна), никакой, даже самый изощренный статистический метод не даст правильного результата. Выше мы останавливались на примере нерепрезентативной выборки, но даже если выборка представительна, то какую совокупность она представляет? Ведь больные в крупных медицинских учреждениях, где обычно проводятся клинические испытания, мало напоминают тех, с которыми встречается врач общей практики. И это лишь одна из возможных проблем на этапе планирования исследования.

⁴² **Клиническое испытание** — это проспективное сравнительное исследование эффективности двух вмешательств и более (лечебных, профилактических) или диагностического метода, в котором группы испытуемых формируются случайно. При этом обычно существует гипотеза, возникающая до проведения исследования относительно эффективности испытываемых методов, которая и проверяется в ходе испытания.

2.5. Необходимые свойства выборки

Напомним, что означает рандомизация при формировании выборки. Все статистические методы исходят из предположения, что данные извлечены из совокупности *случайно*. Это значит, что вероятность оказаться в выборке одинакова для всех членов совокупности. Казалось бы, все просто, но на практике обеспечить равную вероятность попадания в любую из исследуемых групп совсем не так просто, как кажется на первый взгляд. Примером может служить хотя бы уже то, что в самом начале эксперимента важно исключить всякое влияние человека, а это не просто.



Необходимо знать, что если у кого-то есть возможность хоть как-то повлиять на конечный результат, она будет использована, осознанно или нет. И такое вмешательство неизбежно приведет к нарушению сопоставимости групп и к искажению результатов исследования.

Следует тщательно продумать, как сделать такое влияние невозможным для всех участников исследования, и прежде всего для самого исследователя.

Например, при исследовании нового метода лечения задачей рандомизации является обеспечение такого подбора больных, чтобы *контрольная группа* ни в чем не отличалась от *экспериментальной*, кроме рассматриваемого метода лечения.

Однако этого обычно мало. Велика и роль самого больного, в особенности, если он рассчитывает, что новый метод лечения ему на самом деле поможет. Для исключения влияния этого фактора предназначен **слепой метод исследования**. В идеале это **двойной слепой метод**: ни больной, ни наблюдающий его врач не знают, какой из способов лечения был применен. Двойной слепой метод не всегда осуществим, поэтому используют также **простой слепой** (примененный способ лечения известен врачу, но не больному или наоборот) и **частично слепой** (и врач, и больной располагают лишь частью информации) методы.

В любом случае информацию об исследовании, которой могут располагать участники исследования, следует свести к минимуму.

Исследователь редко стремится обмануть других, но легко становится жертвой самообмана. При этом форма самообмана может быть весьма изощрённой.

Рассмотрим такой пример: больных, госпитализированных по нечетным дням месяца, определяют в экспериментальную группу, по четным — в контрольную. Можно ли считать такую рандомизацию достаточной? Разумеется, нет. Врач может влиять на дату госпитализации, следовательно, состав групп будет неслучайным.

Для рандомизации недостаточно, чтобы выбор не зависел от исследователя. Он должен быть независим и от самих подопытных. Приведем пример из области лабораторных исследований.

Двадцать крыс, сидящих в клетке, нужно разделить на две группы. Выпустим из клетки десять крыс и назовем их контрольной группой. Представительна ли она? Скорее всего, нет. Вероятно, первыми из клетки выбегут самые сильные и агрессивные особи.

Есть только один способ получить случайную выборку — воспользоваться для этого достоверно случайным процессом, например *бросанием игральной кости* или *генерацией случайных чисел*.

Однако великие открытия, во многом кардинально изменившие состояние медицины в середине XX в., такие как открытие пенициллина, на тот момент не подвергались строгой проверке с помощью рандомизированных исследований.

Да и исторические факты свидетельствуют, что порой сами обстоятельства способны натолкнуть на серьезную переоценку общепринятых методов лечения. Например, французский военный хирург Амбруаз Парэ в полном соответствии с предписаниями и рекомендациями своего времени лечил огнестрельные раны кипящим маслом. Но однажды, в одну из битв 1536 г., масла на всех раненых солдат не хватило. Вышло так, что некоторому количеству раненых Парэ сделал перевязку, не обработав при этом рану рекомендуемым целительным средством. Утром он с удивлением обнаружил, что солдаты, чьи раны перед перевязкой были обработаны по всем правилам, корчатся от боли, тогда как просто перевязанные «прекрасно отдохнули и не испытывали болей». Понятно, что какого-либо дополнительного исследования проводить не имело смысла.

Так же стоит отметить, что рандомизация далеко не всегда возможна. Вспомните рассмотренный выше пример о сравнении двух кривых выживаемости полученных в ходе исследования методов трансплантации костного мозга при остром лимфобластном лейкозе у взрослых. Одним больным пересаживался костный мозг близких родственников, другим — их собственный. Случайно распределить больных по двум этим группам невозможно, поскольку не у каждого найдется родственник-донор (аллотрансплантация), плюс данная процедура выглядит несколько не этично, так как изначально предполагалось, что аллотрансплантация костного мозга предпочтительнее. К счастью для экспериментаторов само по себе наличие или отсутствие близких родственников не влияет на течение заболевания.

Ситуация, когда разделить больных случайным образом невозможно, в медицинских исследованиях возникает довольно часто. В таких случаях надо стремиться сделать группы максимально схожими по всем известным прогностическим факторам.

2.6. Опасность множественных сравнений

Выше мы уже говорили про эффект множественных сравнений. Напомним о нем еще раз. В исследовании часто может присутствовать множественное сравнение, причем этот эффект нередко присутствует скрыто и исследователь не учитывает наличие этого эффекта и в результа-

те, сам того не подозревая, многократно занижает вероятность *ошибочно* выявить мнимый эффект. Приведем типичный пример такого случая. Если в ходе исследования был получен ожидаемый результат, естественно он будет опубликован. А если бы результат был отрицательным? Скорее всего, публикация бы не появилась на свет, однако велик шанс, что будет предпринята повторная проверка, потом еще и, в конце концов, в одной из многих проверок будет обнаружен желаемый «эффект».

Далее. При широкомасштабных исследованиях различных методов лечения, используемых прежде всего при хронических болезнях, таких, например, как ишемическая болезнь сердца и сахарный диабет. Результатом исследования является описание огромного числа разнообразных признаков. Данные подвергаются различным группировкам с целью выяснения наиболее информативных признаков, в наибольшей степени влияющих на конечный и наиболее важный показатель — *выживаемость*. Понятно, что при значительном числе возможных группировок не составит труда выделить группы, на которых тот или иной метод лечения будет наиболее эффективен. Однако данные процедуры могут привести к превратным заключениям в силу того, что при такой перетасовке данных наблюдений резко возрастает вероятность ошибочного результата и если это не учитывать, то легко выявить что-то чего на деле просто нет.

Приведем пример исследования коронарного шунтирования. Среди наблюдавшихся больных в целом не было выявлено статистически значимых различий в выживаемости между оперированными и не оперированными больными. Однако стоило разделить наблюдения на подгруппы, как оказалось, что хирургическое вмешательство обеспечивает более высокую выживаемость среди «больных с поражением ствола левой коронарной артерии». Интерпретация подобных результатов требует крайней осторожности.

Сходная картина наблюдается, когда в данных, полученных для анализа одних факторов, обнаруживается связь между другими. Возможно, это реально существующая связь, но, возможно, что никакой связи нет, а видимый эффект вызван эффектом множественных сравнений, когда попарно сравнивая все со всем, исследователь непременно найдет какую-нибудь статистическую зависимость, в данном случае при наличии возможности имеет смысл сделать дополнительный анализ, но уже на предмет наличия реальных взаимосвязей.

Этих примеров вполне достаточно, что бы понять насколько порой сложно осуществить только самый начальный этап исследования. Но это далеко еще не все возможные аспекты подготовки исследования. Рассмотрим такое свойство выборки, как ее размер.

2.7. Объем выборки

Проблема определения **объема выборки** не менее важна вышеперечисленных. От объема выборки зависит то, насколько по данным, полученным в

ходе ее анализа, мы сможем судить о всей генеральной совокупности. Если выборка слишком мала, то велика вероятность, что мы получим слишком отличающиеся от общей картины результаты, однако если делать выборку слишком большой, то это вызывает трудности как финансового плана, так и самого сбора информации. В некотором смысле тема подбора нужного количества испытуемых в выборку является достаточно сложной.



Основной смысл определения объема выборок заключается в том, чтобы иметь в результате исследования высокую вероятность выявления реальных статистически значимых различий между выборками, т. е. обеспечить **статистическую мощность (чувствительность)** исследования. Другими словами, необходимо определить, каков должен быть достаточный объем выборок, чтобы мы получили статистически значимый результат.

Рассмотрим эту проблему более подробно. Вопрос о размере групп тесно связан как с научными, так и с этическими сторонами исследования. Очевидно, что от количества объектов исследования зависит и мощность статистических методов, применяемых для обработки результатов эксперимента. Но помимо достоверности результатов, оптимальный подбор числа испытуемых в клинических исследованиях необходим для того, чтобы возможный риск для пациентов был оправдан положительным вкладом в существующее понимание той или иной научной проблемы.

Для того, чтобы получить представление о размере выборки и мощности исследования приходится делать довольно много допущений. Информацию, необходимую для оценки объёма выборки, получают либо из результатов собственных предыдущих исследований, либо из литературных источников, а иногда исследователь не имеет ни того, ни другого. Но в любом случае, необходимо знать **минимальную величину эффекта**, которая в данном исследовании будет считаться достаточной. На основании этой величины можно будет сделать предположение способности исследования дать ответ на интересующий нас вопрос (<http://www.sciencefiles.ru>).

Как вы помните, для определения наличия эффекта в статистических методах используются **критерии значимости**. Отсюда следует, что в определенном смысле мощность статистического исследования напрямую связана с мощностью используемого критерия⁴³. А *мощность*, или чув-

⁴³ Чувствительностью (мощностью) критерия называют его способность обнаружить различия. Чтобы оценить чувствительность критерия, нужно задать величину различий, которые он должен выявлять. Если в результате проверки гипотезы о существовании различий был сделан вывод об их отсутствии, необходимо проверить, была ли чувствительность критерия достаточной для обнаружения таких различий. Так же важным замечанием является то, что для различных статистических критериев и методов анализа чувствительность вычисляется по разным формулам.

ствительность, статистического критерия зависит от нескольких показателей, таких как:

- **Уровень значимости.** Чем меньше этот уровень, тем ниже чувствительность. Имеется в виду то, что при уменьшении значения уровня значимости критерий становится более строгим и вероятность ошибочного результата значительно снижается, однако в этом случае критерий может не обнаружить действительно существующих различий, т. е. может возникнуть ошибка второго рода.

- **Объем выборки.** Чем больше испытуемых в исследовании, тем более чувствительными становятся статистические критерии. Однако и здесь есть смысл сделать оговорку. Например, мы имеем дело с двумя группами, в которых необходимо определить количество пациентов для исследования, для того чтобы проследить изменение значения интересующего параметра до и после начала терапии. Математически эта задача похожа на задачу сравнения двух независимых групп. Но будет использован другой критерий значимости. И уже только в силу специфики такого исследования для статистически достоверного заключения понадобится меньше пациентов, поскольку вариация разности между наблюдениями «до и после» значительно меньше, чем вариация одиночных наблюдений. Это объясняется наличием корреляции между последовательными измерениями (<http://statistica.ru>).

- **Соотношение величины эффекта (величины различий между контрольной и опытной группами) и стандартного отклонения.** Чем выше это значение, тем чувствительнее критерий.

Как видно, ещё до начала исследования необходимо определить, какого размера потребуется выборка, чтобы при заданном уровне значимости и величине эффекта получить достаточную мощность или чувствительность критериев. Например, при изучении препарата, снижающего артериальное давление, исследователи могут ожидать обнаружения эффекта в 15 мм рт. ст. с вероятностью 80 % при уровне значимости 0,05. Затем, определившись со своей целью, исследователи могут приступить к определению оптимальной численности групп, которые будут вовлечены в работу. Естественно, что в это время уже должен быть решён вопрос о том, какой статистический критерий использовать для проверки гипотезы. Следует иметь в виду, что если уже после проведения эксперимента исследователи решат обнаружить изменения давления меньшие, чем выбранные ранее 15 мм рт. ст., то при сохранении всех прочих параметров неизменными мощность уменьшится. Следовательно, обратные действия приведут к увеличению мощности критериев.

Для определения необходимого объема выборки можно воспользоваться математическими формулами⁴⁴ или возможностями программных пакетов для статистического анализа.

⁴⁴ Здесь мы не будем приводить используемые для расчетов формулы, подробнее этот вопрос рассмотрен на странице <http://statistica.ru/local-portals/medicine/planirovanie-meditinskikh-issledovaniy/>.

Таким образом, для оценки требуемого объема выборки для данного исследования, необходимо:

- понять, переменные какого типа (количественные, качественные, и т. д.) будут измерять эффект в данном исследовании;
- выбрать, исходя из специфики исследования, подходящий план;
- оценить величину различий между эффектами, приемлемую для данного исследования с клинической точки зрения;
- выбрать подходящий статистический критерий для последующего анализа интересующих событий, это определит выбор конкретных формул для расчета;
- определить, односторонний или двусторонний критерий будет использоваться;
- оценить по данным литературы, пилотному исследованию или результатам сходных исследований величины показателей, входящих в выбранные для расчета формулы;
- увеличить рассчитанные по формулами значения объема с учетом возможного исключения в процессе исследования (<http://statistica.ru/local-portals/medicine/planirovanie-meditsinskikh-issledovaniy/>).

Как видно, определение объема исследуемых групп не является тривиальной задачей. И требует достаточно серьезного подхода. Способность применить статистический подход в медицине не сводится к заучиванию нескольких формул и умению отыскать табличное значение. Как и любая творческая деятельность, применение статистических методов и интерпретация полученных результатов требуют глубокого понимания в сути дела, знания как возможностей и ограничений используемых методов, так и самой проблемы решаемой в ходе исследования.

Не менее важным моментом является проблема формирования самих групп: *контрольной* и *экспериментальной*. Если с экспериментальной в принципе все достаточно ясно — например, та группа, которой дается исследуемый препарат, то контрольная группа имеет некоторую специфику и ее необходимо учитывать. Исследование обязательно должно быть проведено с использованием плацебо⁴⁵. Тогда как исследования, когда контрольной группе совсем ничего не дается уже не проводятся. Так же на данный момент, конечно же, нельзя серьезно рассматривать исследования,

⁴⁵ **Плаце́бо** — физиологически инертное вещество, используемое в качестве лекарственного средства, положительный лечебный эффект которого связан с бессознательным психологическим ожиданием пациента. Плацебо не способно действовать непосредственно на те условия, ради изменения которых исследуют препарат. Кроме того, термином «эффект плацебо» называют само явление немедикаментозного воздействия не только препарата, а, например, облучения (иногда используют разные «мигающие» аппараты, лазеротерапию и др.). В качестве вещества для плацебо часто используют лактозу. Степень проявления плацебо-эффекта зависит от внушаемости человека и внешних обстоятельств «лечения», например, от размера и яркого цвета таблетки, степени доверия врачу, авторитета клиники.

в которых исследуемый препарат сравнивается со своим предшественником или схожими конкурентами (<http://www.sportmedicine.ru>).

Другими словами подход к планированию исследования должен осуществляться особенно тщательно и не стоит упускать различного рода «мелочи», как можно заключить из предыдущего материала, ошибки, возникающие вследствие их накопления, могут существенно повлиять на результат всего исследования.

ЗАКЛЮЧЕНИЕ

На страницах данного пособия мы рассмотрели наиболее часто используемые методы статистической обработки, а также, по возможности коснулись, важных вопросов, связанных со статистической обработкой данных. Основной целью пособия являлось — донести эту информацию в максимально простой форме, при этом не упустить важные детали и возможные трудности, с которыми зачастую сталкиваются те, кто, так или иначе, обращается к методам статистического анализа. Содержимое пособия ни в коем случае не является полным описанием используемых на современном этапе развития науки методик работы с данными и не претендует на место достаточного и единственного руководства для изучения статистики, тем более, что некоторых важных моментов мы коснулись довольно поверхностно. Для более полного знакомства с теоретическими и практическими особенностями дисциплины рекомендуется самостоятельно изучить некоторые источники, которые мы привели в списке используемой литературы и которые мы часто цитировали при нашем изложении материала.

В заключение приведем рекомендации, взятые нами из книги «Медико-биологическая статистика» под авторством Стентона Гланца.

Мы говорили, что значение статистических методов возрастает по мере возрастания требований к обоснованию эффективности используемых современной медициной методов лечения, диагностики и т.д. И выяснили, что статистическое обоснование зачастую оказывается важнейшим фактором, определяющим решение в пользу предлагаемого метода.

В то же время сами медики довольно редко занимаются статистическим обоснованием своих исследований зачастую в силу того, что их познания в этой области достаточно невелики.

Обычно вся статистическая сторона дела перепоручается консультантам, нередко действительно разбирающимся в статистике, но имеющим довольно смутное представление о медицинских вопросах. Единственный выход состоит в том, чтобы медики наконец сами занялись статистическим анализом, поскольку именно они знают цели исследования и несут за него ответственность.

Примеров не совсем корректных статистических анализов, используемых в исследовании, достаточно много, однако результаты исследования были «обоснованы», публикация вышла в свет и не исключено, что рекомендации, в ней описанные, будут применяться на практике, но эффективность их возможно окажется значительно ниже заявленной. Поэтому серьезный врач, особенно исследователь, не должен принимать за чистую монету все, что пишется в журналах (В. В. Власов «Как читать медицинские статьи»).

Знакомясь с материалами очередного исследования, обратите внимание, названы ли:

- подлежащая проверке гипотеза;
- использованные данные и способ их получения (включая используемый метод рандомизации);
- совокупность, которую представляют используемые в исследовании выборки;
- тип распределения выборочных данных;
- статистические методы, использованные для оценки гипотезы.

Чем ближе публикация соответствует данным требованиям, тем вернее можно положиться на приведенные в ней выводы. Напротив, очень мало доверия заслуживает статья, в которой использованные методы не указаны вовсе или упоминаются некие «стандартные методы».

Затрагивая вопрос об этичности исследований на людях (или лабораторных животных), хочется подчеркнуть, что чем менее грамотно и добросовестно проведено исследование, тем менее оно этично как по отношению к тем пациентам, которые в нем участвовали, так и ко всем возможным пациентам, лечение которых может зависеть от результатов данного исследования.

Не этичен любой вводящий в заблуждение результат. Неэтично подвергать людей страданиям и мучить лабораторных животных ради получения данных, на основании которых невозможно сделать какой-либо вывод. Неэтично выполнять такие исследования, опровержение которых потребует чьих-то сил, здоровья и средств.

Конечно, тщательная проработка статистической стороны исследования не освобождает исследователя от обязанности тщательно продумать эксперимент с врачебной точки зрения, свести риск и страдания больных к минимуму. Более того, она даже не гарантирует, что в исследовании будут получены глубокие и новаторские результаты. Иными словами, статистическая корректность — это необходимое, но еще не достаточное условие успеха исследования.

Как же изменить исследовательскую практику к лучшему? Прежде всего, старайтесь быть более активными. Если это от вас зависит, по возможности, не подпускайте к исследованиям людей, несведущих в статистике, а также тех, кто не смыслит в медицине. Встретив явные статистические ошибки и недочеты в журнале, пишите редактору. Не стесняйтесь задавать вопросы своим коллегам. Не поддавайтесь гипнозу наукообразия — докапывайтесь до сути дела. Когда вас осыпают мудреными терминами, задайте какой-нибудь тривиальный вопрос, на который вы сами можете ответить. Это позволит понять насколько человек, забрасывающий вас сложной терминологией, сам разбирается в сути дела и в тонкостях исследуемого вопроса.

Но самое главное, чтобы ваши собственные исследования были максимально хороши с точки зрения планирования и применения статистических методов.

ЛИТЕРАТУРА

1. *Гланц, С.* Медико-биологическая статистика / С. Гланц; пер. с англ. — М.: Практика, 1998. — 459 с.
2. *Платонов, А. Е.* Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы / А. Е. Платонов. — М.: Издательство РАМН, 2000. — 52 с.
3. *Жученко, Ю. М.* Информационные технологии в биологии и химии: лабораторный практикум для студентов вузов по специальности 1-31 01 01 «Биология» / Ю. М. Жученко. — Гомель: ГГУим. Ф. Скорины, 2010. — 148 с.
4. *Реброва, О. Ю.* Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О. Ю. Реброва. — 3-е изд. — М.: МедиаСфера, 2006. — 312 с.
5. *Халафян, А. А.* STATISTICA 6. Статистический анализ данных / А. А. Халафян. — 3-е изд. — М.: ООО Бином-Пресс, 2007. — 512 с.
6. *Жученко, Ю. М.* Основы статистики: лабораторный практикум / Ю. М. Жученко, А. А. Ковалев, В. А. Игнатенко. — Гомель: ГомГМУ, 2016. — 176 с.
7. Обработка экспериментальных данных в MS Excel: методические указания к выполнению лабораторных работ для студентов дневной формы обучения / сост. Е. Г. Агапова, Е. А. Битехтина. — Хабаровск: Изд-во Тихоокеан. гос. ун-та, 2012. — 32 с.
8. STATISTICA 6.0 — фирменное руководство. Компания StatSoft. Электронная публикация, 1995.
9. StatSoft.Inc. Электронный учебник по статистике. М. StatSoft.WEB: <http://www.StatSoft.ru/home/textbook/default.htm>, 2001.
10. *Городин, В. Н.* Моделирование и анализ времен жизни больных лептоспирозом / В. Н. Городин, А. А. Халафян // Известия высших учебных заведений. Северокавказский регион. Естественные науки, 2008. — № 2.
11. *Власов, В. В.* Как читать медицинские статьи / В. В. Власов.
12. *Мерков, А. Б.* Об анализе выживаемости / А. Б. Мерков.
13. *Гайдышев, И. П.* Программное обеспечение анализа данных AtteStat. Руководство пользователя / И. П. Гайдышев.
14. *Лялин, В. С.* Статистика: теория и практика в Excel: учеб. пособие / В. С. Лялин, И. Г. Зверева, Н. Г. Никифорова. — М.: Финансы и статистика; ИНФРА-М, 2010. — 448 с.
15. *Сырцова, Л. Е.* Основы эпидемиологии и статистического анализа в общественном здоровье и управлении здравоохранением: учеб. пособие / Л. Е. Сырцова, И. И. Косаговская, М. В. Авксентьева. — М., 2003.
16. *Максимов, С. И.* Статистический анализ и обработка данных с применением MS Excel и SPSS: учеб.-метод. пособие / С. И. Максимов. — Минск: РИВШ, 2012. — 114 с.

17. *Климентьева, О. В.* Логистическая регрессия и ROC-анализ в скоринге / О. В. Климентьева, М. А. Милосердова.
18. *Леонов, В. П.* Основные понятия понятия ROC-анализа. Лекция в Якутске, 13.11.2009. Презентация.
19. *Власов, В. В.* Доказательная медицина и принципы методологии / В. В. Власов, Е. Н. Семернин, П. В. Мирошенков. — Мир медицины, 2001.
20. *Костюкевич, А. А.* Конспект лекций по курсу «Автоматизированные системы обработки биомедицинской информации» / А. А. Костюкевич. — Минск, 2006.
21. *Боровиков, В.* Statistica. Искусство анализа данных на компьютере: для профессионалов / В. Боровиков. — 2-е изд. — СПб.: Питер, 2003. — 688 с.
22. *Юнкеров, В. И.* Математико-статистическая обработка данных медицинских исследований / В. И. Юнкеров, С. Г. Григорьев. — СПб.: ВМедА, 2002. — 266 с.
23. *Зайцев, В. М.* Прикладная медицинская статистика / В. М. Зайцев, В. Г. Лифляндский, В. И. Маринкин. — СПб.: ФОЛИАНТ, 2003. — 432 с.
24. *Рамси, Дебора.* Статистика для "чайников" / Дебора Рамси; пер. с англ. — М.: И. Д. Вильяме, 2008. — 320 с.
25. *Наследов, А. Д.* SPSS: Компьютерный анализ данных в психологии и социальных науках / А. Д. Наследов. — СПб.: Питер, 2005. — 416 с.
26. *Наследов, А. Д.* Математические методы психологического исследования. Анализ и интерпретация данных: учеб. пособие / А. Д. Наследов. — СПб.: Речь, 2004. — 392 с.
27. *Такахаси, Син.* Занимательная статистика. Манга / Син Такахаси; пер. с яп. Е. А. Захаровой, Коги Муцуми. — М.: Додэка-XXI, 2010. — 224 с.
28. *Шуровьески, Д.* Мудрость толпы. Почему вместе мы умнее, чем поодиночке, и как коллективный разум влияет на бизнес, экономику, общество и государство / Д. Шуровьески.
29. *Фрай, С.* Книга всеобщих заблуждений / С. Фрай.
30. Triola, Mario F. Elementary statistics / Mario F. Triola. — 10th ed. — 770 с.
31. Интернет ресурс <http://www.sciencefiles.ru>.
32. Интернет ресурс <http://statistica.ru>.
33. Интернет ресурс <http://medstatistic.ru>.
34. Интернет ресурс <http://www.sportmedicine.ru>.
35. Интернет ресурс <http://www.biometrica.tomsk.ru>.
36. Интернет ресурс <http://matstats.ru>.
37. Интернет ресурс <http://medspecial.ru>.

Учебное издание

Ковалев Алексей Алексеевич
Игнатенко Валерий Андреевич

**ОСНОВЫ СТАТИСТИКИ.
КОНСПЕКТ ЛЕКЦИЙ**

**Учебно-методическое пособие
для студентов 3–6 курсов всех факультетов
медицинских вузов**

Редактор *Т. М. Кожемякина*
Компьютерная верстка *Ж. И. Цырыкова*

Подписано в печать 02.04.2018.
Формат 60×84¹/₈. Бумага офсетная 80 г/м². Гарнитура «Таймс».
Усл. печ. л. 20,46. Уч.-изд. л. 22,37. Тираж 150 экз. Заказ № 151.

Издатель и полиграфическое исполнение:
учреждение образования «Гомельский государственный медицинский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий № 1/46 от 03.10.2013.
Ул. Ланге, 5, 246000, Гомель