ФМБА РОССИИ
Федеральное медико-биологическое агентство

Медико-биологический университет
инноваций и непрерывного образования
ФМБЦ им. А.И. Бурназяна ФМБА России

Адрес: г. Москва, ул. Живописная, д. 46, стр. 8
Тел.: 8 (499) 190-96-92
Сайт: www.mbufmbc.ru

Посохов И.Н.

ВВЕДЕНИЕ В БИОСТАТИСТИКУ: ЛЕКЦИИ ДЛЯ КЛИНИЦИСТОВ

Монография

Москва, 2027

Федеральное медико-биологическое агентство
Федеральное государственное бюджетное учреждение
«Государственный научный центр Российской Федерации —
Федеральный медицинский биофизический центр
имени А.И. Бурназяна»
МЕДИКО-БИОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ
ИННОВАЦИЙ И НЕПРЕРЫВНОГО ОБРАЗОВАНИЯ

Посохов И.Н.

ВВЕДЕНИЕ В БИОСТАТИСТИКУ: ЛЕКЦИИ ДЛЯ КЛИНИЦИСТОВ

Монография

Москва 2027

УДК 57.087.1
ББК 60.65
П62

Посохов И.Н. Введение в биостатистику: лекции для клиницистов. — М.: ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, 2027. 146 с.

Автор:

Посохов Игорь Николаевич — доктор медицинских наук, профессор кафедры терапии МБУ ИНО ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России.

Рецензенты:

Репин Алексей Николаевич — доктор медицинских наук, профессор, профессор кафедры пропедевтики внутренних болезней с курсом терапии педиатрического факультета Сибирского государственного медицинского университета Минздрава России, Томск, Россия.

Бекболат Жолдин — PhD, MD, FESC, профессор, член проблемной комиссии по направлению «Медицина» Западно-Казахстанского медицинского университета имени Марата Оспанова, Актобе, Казахстан.

Настоящая монография представляет собой сборник материалов, которые использовались автором для чтения лекций ординаторам и аспирантам, обучающимся в Университете инноваций и непрерывного образования ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России. Сборник представляет собой структурированное практическое руководство по основам биостатистики и анализу медицинских данных. Программа курса охватывает многие аспекты цикла научного исследования — от момента разработки регистрационных карт до финальной оценки диагностической точности моделей и пригодна для первого знакомства с дисциплиной. Целевая аудитория: студенты-медики, врачи-клиницисты, ординаторы, аспиранты.

ISBN 978-5-93064-466-1 © ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, 2026

Содержание

Введение 5
Глава 1. Сбор данных в статистических исследованиях 6
1.1. Генеральная совокупность и выборка 6
1.2. Наблюдение и эксперимент 7
1.3. Опросники 7
1.4. Индивидуальная регистрационная карта 9
1.5. Типы данных 11
1.6. Ввод данных в компьютер 13
1.7. Сбор данных: краткие итоги 18
Глава 2. Графическое и табличное представление данных 20
2.1. Зачем нужна визуализация данных? 20
2.2. Столбиковые диаграммы 20
2.3. Гистограммы 22
2.4. Круговые диаграммы 23
2.5. Диаграммы рассеяния и пузырьковые диаграммы 24
2.6. Линейные диаграммы 25
2.7. Правила конструирования таблиц 26
2.8. Проценты в таблицах 27
2.9. Представление данных: краткие итоги 29
Глава 3. Описание данных 30
3.1. Переход от графиков к числовым показателям 30
3.2. Меры центральной тенденции / положения 31
3.3. Меры рассеяния 33
3.4. «Ящик с усами» и «выбросы» 35
3.5. Связь типа данных и описательной статистики 39
3.6. Описание данных: краткие итоги 40
Глава 4. Нормальное распределение 41
4.1. Ключевые характеристики нормального распределения 41
4.2. Правило трех сигм 42
4.3. Функция плотности и функция распределения 43
4.4. Стандартизированное нормальное распределение и *Z*-оценки 46
4.5. Квантили и процентили 49
4.6. Тестирование распределения на нормальность 53
4.7. Центральная предельная теорема 57
4.8. Доверительные интервалы и стандартная ошибка средней 59
4.9. Логнормальное распределение 64
4.10. Нормальное распределение: краткие итоги 67

Глава 5. Анализ качественных данных 69
5.1. Биномиальное распределение. 69
5.2. Стандартная ошибка доли и доверительные интервалы 72
5.3. Репрезентативность выборки 76
5.4. Таблицы частот и критерий Хи-квадрат. 79
5.5. Анализ качественных данных: краткие итоги. 83
Глава 6. Оценка взаимосвязей 84
6.1. Диаграмма рассеяния, корреляция 84
6.2. Линейная регрессия и коэффициент детерминации. 86
6.3. Множественная линейная регрессия 88
6.4. Оценка взаимосвязей: краткие итоги 91
Глава 7. Сравнение двух и более групп. 93
7.1. Критерий Стьюдента: сравнение независимых групп. 93
7.2. Парный *t*-тест: сравнение зависимых групп 96
7.3. Дисперсионный анализ, ANOVA 98
7.4. Непараметрические тесты в контексте сравнения групп 102
7.5. Сравнение групп: краткие итоги 105
Глава 8. Анализ выживаемости, рисков и шансов. 108
8.1. Цензурированные наблюдения 108
8.2. Функции, изучаемые в анализе выживаемости 109
8.3. Метод Каплана-Майера. 110
8.4. Сравнение выживаемости в группах. 111
8.5. Модель пропорциональных рисков Кокса 111
8.6. Hazard Ratio и Odds Ratio 112
8.7. Метаанализ и форест-плот 115
8.8. Анализ выживаемости, рисков и шансов: краткие итоги. 119
Глава 9. Диагностическая точность. 121
9.1. Чувствительность и специфичность 121
9.2. ROC-кривая и показатель AUC 124
9.3. Метод Бланда-Альтмана 128
9.4. Диагностическая точность: краткие итоги. 131
Заключение 133
Глоссарий. 134
Полезные ссылки 145
Рекомендуемая литература. 146

Введение

Уважаемые коллеги! В современной медицине ежедневно принимаются сотни критических решений: от выбора схемы лечения конкретного пациента до утверждения национальных протоколов здравоохранения. Однако, на чём базируются эти решения? В эпоху доказательной медицины интуиции, личного опыта или мнения авторитетных экспертов уже недостаточно. Фундаментом для принятия обоснованных медицинских решений становятся строго верифицированные данные.

Многие клиницисты воспринимают статистику как сухую, оторванную от реальности дисциплину, наполненную громоздкими математическими формулами. Постараемся разрушить этот стереотип. Мы будем говорить о статистике не как о математике, а как о методологии работы с данными, ведь биостатистика — это незаменимый инструмент, который позволяет клиницисту, во первых, отделять истинные клинические закономерности от случайных совпадений и информационного «шума»; во-вторых, корректно планировать собственные исследования, избегая «капканов» еще на этапе сбора данных, в-третьих, критически оценивать публикации в зарубежных и отечественных научных журналах, понимая, где за красивыми графиками скрываются методологические ошибки.

Этот курс разработан специально для будущих или состоявшихся клиницистов: теоретический математический аппарат минимизирован (насколько это возможно), а фокус смещён на прикладной анализ, интерпретацию результатов и разбор реальных клинических кейсов — от оценки результатов анкетирования до анализа выживаемости и расчета точности диагностических тестов.

Глава 1. Сбор данных в статистических исследованиях

1.1. Генеральная совокупность и выборка

Когда мы начинаем собирать данные, перед нами встает выбор. Мы можем исследовать абсолютно всех или только часть объектов. Давайте введем два главных термина.

1. ***Генеральная совокупность (Population).*** Это абсолютно все объекты или люди, которые нас интересуют. Если мы изучаем подростковую гиперактивность, то генеральная совокупность — это все подростки с таким диагнозом в мире или в стране. Объектами исследования могут быть не только люди. Это могут быть бактериальные колонии, медицинские карты, таблетки или больницы.

2. ***Выборка (Sample).*** Это относительно небольшое число объектов, которое мы реально берем для изучения.

Почему мы почти никогда не исследуем всю генеральную совокупность? Исследовать только выборку намного быстрее и дешевле. Более того, аккуратно спланированное выборочное исследование часто дает более точные результаты, чем плохо организованное сплошное исследование.

Главное требование к выборке — она должна быть ***репрезентативной***. Это значит, что она должна правильно отражать свойства всей генеральной совокупности. На практике репрезентативность достигается за счет случайного отбора участников. Чем больше размер выборки, тем точнее будут наши оценки.

В качестве примера можно привести меры по разработке в комплексном врачебном здравпункте программ обучения пациентов домашнему мониторированию артериального давления (АД). Генеральная совокупность здесь — все пациенты, посещающие здравпункт, которым требуется следить за АД. Для начала нужно узнать первоначальный уровень осведомлённости этих пациентов о правилах и распорядке измерений АД. Но один врач не может опросить каждого, кто приходит в течение года в здравпункт, поэтому он формирует случайную выборку из 31 человека. На основе их ответов исследователь сделает выводы обо всех остальных пациентах.

1.2. Наблюдение и эксперимент

Многие исследования в медицине основаны на изучении архивных записей, которые ведут врачи. Но иногда нужных нам данных в готовом виде нет. Тогда нам приходится выбирать стратегию исследования. Основных путей два: выборочное исследование и контролируемый эксперимент.

1. ***Выборочное исследование (Sample Survey)***. Здесь мы выступаем в роли пассивных наблюдателей. Мы просто фиксируем характеристики, которые уже существуют у объектов. Мы не меняем образ жизни людей и не вмешиваемся в процессы.

2. ***Спланированный эксперимент (Experiment)***. Здесь исследователь активно вмешивается в процесс. Мы целенаправленно меняем один или несколько факторов. Делается это для того, чтобы измерить их влияние на результат.

Иногда вмешательство может быть деликатным и незаметным, но всё равно давать результаты. Вернёмся в наш здравпункт. Врачи поняли, что пациенты плохо осведомлены о правилах домашнего измерения АД, и из-за этого предоставляют некачественную информацию о своих измерениях, записанную на бумаге, врачам. В распоряжении исследователей есть пара методов обучения: один метод – обучение во время традиционных «школ здоровья», второй – при помощи онлайн материалов, содержащихся в QR-коде на бланке, который заготовлен врачом и который заполняет пациент при измерении АД. Исследователь планирует эксперимент: он использует оба метода, а затем сравнивает качество и полноту информации об АД по обычным записям на бумаге и по записям в бланках. Выясняется, что цифровое ассистирование, содержащееся в QR-кодах бланков, резко поднимает качество их заполнения. На основе эксперимента исследователь принимает решение по внедрению более лучшего способа.

1.3. Опросники

Где физически брать информацию? Самый простой способ — это анкетирование. В опросе всегда участвуют три стороны: исследователь (составитель анкеты), интервьюер (тот, кто задает вопросы) и респондент (тот, кто отвечает). Все вопросы в анкетах делятся на два типа:

1. ***Фоновые вопросы***. Это пол, возраст, доход, образование. Они нужны для того, чтобы потом группировать результаты в таблицах.

2. ***Программные вопросы***. Это суть исследования. Например: «Сколько дней вы провели в стационаре за прошлый год?» или «Оцените уровень вашей боли».

По форме ответа вопросы бывают трех видов:

1. ***Закрытые вопросы***. Респондент выбирает из готовых вариантов (например, «Да/Нет»). Их очень легко обрабатывать на компьютере. Но человек может не найти нужного варианта.

2. ***Открытые вопросы***. Человек пишет ответ в свободной форме. Это дает свободу мысли. Но анализировать такие данные крайне тяжело. Специалисту придется вручную кодировать каждый текстовый ответ.

3. ***Полузакрытые вопросы***. Они содержат список готовых вариантов и строчку «Другое: _____».

Методическое правило: если вопрос содержит шкалу оценок (например, «Очень плохо — Плохо — Нейтрально — Хорошо — Отлично»), используйте от 3 до 7 вариантов. Чаще всего выбирают 5 вариантов. При этом нечетное число вариантов дает человеку возможность выбрать нейтральную середину. Если вы хотите заставить респондента принять решение, используйте четное число вариантов (4 или 6).

Данные редко бывают идеальными. В ходе любого опроса возникают искажения.

1. ***Ошибки формулировок***. Нечеткие или слишком сложные вопросы путают людей. Также опасны наводящие вопросы, которые подталкивают к определенному ответу.

2. ***Неискренность***. Люди часто хотят казаться лучше. Например, при опросе об алкоголе респонденты занижают объемы выпитого. Для выявления лжи используют контрольные вопросы: один и тот же вопрос задают разными словами в разных частях анкеты.

3. ***Проблема неответивших***. Это главный враг репрезентативности. Люди могут быть заняты или просто отказываются отвечать. Из-за этого выборка становится смещенной. Напри-

мер, если вы звоните людям только в рабочее время, вы никогда не опросите деловых людей. Результаты исследования будут искажены.

Чтобы снизить долю отказов, нужно проводить ***пилотное исследование*** (Pilot study). Протестируйте анкету на небольшой группе коллег или друзей. Это поможет найти скрытые ошибки до запуска проекта. Также полезно мотивировать участников небольшими подарками или участием в лотерее.

Выбор метода всегда зависит от баланса стоимости и качества данных:

1. ***Интернет-опросы*** (Email, SMS, формы). Самый дешевый и быстрый способ. Позволяет задавать вопросы широкой аудитории. Главный минус — огромный процент отказов. Длинные анкеты в интернете люди просто закрывают.

2. ***Почтовые бумажные анкеты***. Недорогой метод. Однако вы не можете контролировать процесс заполнения. Человек может пропустить важные пункты, и вы не сможете это исправить.

3. ***Телефонные интервью***. Более дорогой метод. Появляется живой контакт. Живой интервьюер может объяснить непонятный вопрос, что повышает качество данных. Телефонные интервью могут использоваться при анализе выживаемости или повторных госпитализаций.

4. ***Личные интервью***. Самый дорогой, но самый качественный метод. Интервьюер может использовать карточки и визуальные материалы. Люди реже отказываются от диалога при личной встрече. Главный риск — интервьюер может случайно внушить ответ своим тоном или мимикой.

1.4. Индивидуальная регистрационная карта

Представьте, что вы крупный координатор многоцентрового исследования-регистра. Исследование настолько интересно, что сотрудничать с вами захотели десятки врачей в разных исследовательских центрах. Как собрать данные всех пациентов без хаоса? Как сделать так, чтобы все исследователи записывали симптомы одинаково? Для этого и нужна индивидуальная регистрационная карта (ИРК). В английском языке её называют Case Report Form (CRF). Это официальный документ исследования. В

него исследователь записывает все данные пациента. Эти данные предусмотрены протоколом испытания.

ИРК — это мост между клинической практикой и статистикой. Если этот мост построен плохо, все исследование пойдет на дно, а репутация координатора окажется под угрозой.

У ИРК есть три важнейшие задачи.

1. ***Сбор данных по протоколу***. Каждое исследование имеет свой строгий план (протокол). В CRF включены только те вопросы, которые нужны для науки. Лишняя информация туда не вносится.

2. ***Стандартизация информации***. Все врачи проекта должны собирать данные в одном формате. ИРК не позволяет писать текст в свободной форме. Документ предлагает четкие варианты ответов.

3. ***Обеспечение верификации данных***. Данные из ИРК проверяют медицинские мониторы (в случае клинических исследований) или комиссия по проверке первичной документации (в случае диссертации).

Сегодня в мире медицины существуют два формата карт. Давайте сравним их.

1. ***Бумажная ИРК***. Это классический вариант. Документ печатают на бумаге. Врач заполняет его от руки обычной ручкой. Но этот вариант имеет минусы: бумагу можно потерять или залить кофе; почерк врача бывает трудно разобрать; потом эти данные нужно вручную переносить в компьютер, а это удваивает риск ошибок.

2. ***Электронная ИРК***. Это современный стандарт. Врач вносит данные сразу в специальную компьютерную программу через браузер. Система может проверять ошибки «на лету». Если врач введет нереальное давление (например, 300), программа сразу выдаст предупреждение.

Хорошая ИРК всегда логична. Она повторяет путь пациента в исследовании. Обычно карта для клинических исследований состоит из нескольких стандартных модулей.

1. ***Демография и скрининг***. Сюда входит: уникальный номер пациента в проекте; дата рождения или возраст; пол и расовая принадлежность.

2. ***Соответствие критериям***: Список критериев включения пациента в исследование. Список критериев исключения. (Врач должен подтвердить каждый пункт в виде ответов «Да» или «Нет»).

3. ***Исходный статус***: диагноз и анамнез болезни; результаты первых анализов и осмотра; сопутствующая терапия (какие лекарства пациент уже принимает).

4. ***Визиты и лечение***: дата каждого визита; жизненные показатели (пульс, давление, температура); дозировка исследуемого препарата; результаты повторных лабораторных тестов.

5. ***Нежелательные явления***: фиксация любых побочных эффектов; степень тяжести этих эффектов; связь побочного эффекта с исследуемым лекарством.

Плохой дизайн карты может погубить хорошую научную идею. Чтобы этого избежать, для составления ИРК использую несколько правил: во-первых, необходим минимум свободного текста: статистические программы почти не умеют анализировать длинные сочинения врачей. Вместо текста используйте закрытые вопросы с вариантами ответов. Во-вторых, четкие единицы измерения. Возле заголовка каждого числового поля должны стоять единицы. Например: «Вес (кг)» или «Гемоглобин (г/л)». В-третьих, разделение связанных параметров. Так, не следует писать артериальное давление в одну ячейку через слэш. Необходимы два отдельных поля: систолическое давление и диастолическое давление. Наконец, простые форматы дат: один стандарт для всего документа. Например: «ДД.ММ.ГГГГ».

ИРК для диссертации может отличаться по структуре, например, если исследование в ней в основном диагностического характера. Но при её составлении и дальнейшем заполнении всё же следует также придерживаться вышеизложенных принципов.

1.5. Типы данных

Прежде чем нажать первую клавишу на клавиатуре, мы должны четко понимать, ***что именно*** мы вводим. В биостатистике любой показатель (переменная) относится к одному из четырех уровней измерения. Эта классификация определяет не только формат ввода, но и математические операции, которые выполняются при статистическом анализе в будущем.

1. ***Номинальный уровень*** (Nominal variables). Это самые простые качественные данные. Они представляют собой взаимно исключающие категории, порядок перечисления которых абсолютно произволен. Примеры: Пол (мужской, женский), раса (европеоидная, негроидная, монголоидная), тип бактерий (кокки, бациллы, спириллы), исход лечения (выздоровление, смерть). Категории — это просто ярлыки. Если в таблице поменять местами строки «Мужской» и «Женский», суть информации не изменится. Математически мы можем их только подсчитывать (определять частоту и проценты).

2. ***Порядковый уровень*** (Ordinal data). Качественные данные, у которых категории имеют естественную, логическую последовательность. Однако расстояние между уровнями математически измерить нельзя. Примеры: Степень тяжести побочного эффекта (легкая, умеренная, тяжелая); шкала согласия Ликерта (полностью не согласен, не согласен, нейтрален, согласен, полностью согласен); стадии рака (I, II, III, IV). Мы точно знаем, что умеренная степень тяжести хуже, чем легкая. Но мы не можем утверждать, что расстояние от легкой до умеренной степени точно такое же, как от умеренной до тяжелой.

3. ***Интервальный уровень*** (Interval data). Это уже числовые, количественные данные. Здесь разница (интервал) между двумя числами имеет четкий физический смысл, но точка «ноль» выбрана условно. Пример: Температура по шкале Цельсия. Разница между 10 °C и 20 °C точно такая же, как между 20 °C и 30 °C. Но ноль градусов по Цельсию не означает полное отсутствие тепловой энергии (это просто точка замерзания воды). Поэтому мы не имеем права говорить, что 20 °C «в два раза теплее», чем 10 °C.

4. ***Относительный уровень*** (Ratio data). Высший уровень числовых данных. Здесь существует абсолютный, истинный ноль, означающий полное отсутствие измеряемого свойства. Примеры: Рост, вес, возраст, концентрация глюкозы в плазме крови, артериальное давление, стаж курения. Так как ноль здесь настоящий, все математические операции легитимны. Мы можем с уверенностью заявить, что пациент с весом 100 кг ровно в два раза тяжелее пациента с весом 50 кг.

При проектировании базы данных под каждую переменную выделяется строго одна колонка, и тип этой колонки должен строго соответствовать её уровню измерения.

1.6. Ввод данных в компьютер

Перейдем к практической реализации. Как структурировать элементы, которые не являются непосредственно клиническими параметрами, но необходимы для организации исследования?

Запомните правило: свободный текстовый ввод в базе данных должен быть сведен к абсолютному минимуму. Текстовые поля (строковые переменные) допустимы исключительно для: комментариев исследователя в конце карты или для текстового раскрытия графы «Другое: [вписать вручную]» в опросниках.

Ни в коем случае не используйте свободный текст как замену для категориальных данных. Если вы допустите вольность вводить диагноз текстом, то в одном случае будет написано «ГБ II ст.», во втором — «Гипертоническая болезнь 2 стадии», а в третьем — «Артериальная гипертензия». Для компьютера это будут три совершенно разные болезни, и вы проведете недели, вручную исправляя эти записи.

Каждому объекту или пациенту должен быть присвоен ***уникальный идентификатор*** (ID). Никогда не используйте фамилии в качестве ключевого поля.

Если Вы проводите одноцентровое исследование, используйте простые сквозные номера (длиной 2–4 знака). Лайфхак: начинайте нумерацию не с единицы, а, например, со 100. Тогда все ваши номера будут строго трехзначными («100», «101», «102»). Если начать с «001», «002», то при импорте или сохранении некоторые программы часто сбрасывает ведущие нули, превращая коды в «1» и «2», что ломает структуру и сортировку (возможно, вы уже видели такую сортировку: 118, 119, 12, 120, 121 и тд).

При многоцентровых исследованиях ID обязательно должен быть составным и включать код клинического центра и порядковый номер пациента через дефис. Например, код «03-104» однозначно указывает: это 104-й пациент из 3-го филиала.

Если из-за специфики исследования вам разрешено хранить персональные данные (ФИО и адреса), никогда не пишите их одной длинной строкой. ФИО нужно разбивать либо на формат «Фамилия, Имя Отчество» в одной ячейке, либо выделять под фамилию отдельную колонку. Полезно заводить отдельные микрополя для префиксов (профессор, доцент, г-н) и (это чаще для зарубежных исследований) суффиксов (мл., III). Адрес следует дробить на независимые переменные: «Страна», «Область», «Город», «Улица/дом», «Почтовый индекс». Только в таком случае

это поможет мгновенно отсортировать пациентов по городам (или улицам) чтобы сделать автоматическую почтовую рассылку приглашений на осмотр.

Ввод качественных (категориальных) данных — это зона повышенного риска. Давайте разберем, как избежать классических капканов. Старые статистические программы требовали кодировать категории строго последовательными цифрами (например: «1 = Мужской», «2 = Женский»). Современный софт работает и с текстом. Лучшая мировая практика сегодня — использовать короткие, понятные, текстовые аббревиатуры (например, «М» / «Ж»). Такой подход делает базу данных «самодокументируемой». Если через три года вы откроете файл, вы сразу поймете структуру, в то время как цифровые коды «1» и «2» заставят вас искать расшифровки.

Большинство медицинских исследователей используют Microsoft Excel как хранилище данных. Excel — прекрасный инструмент, но у него есть колоссальный минус для статистики: ему абсолютно все равно, что вы вводите в ячейку. Вы можете ввести пол для первого пациента как «М», для второго — «Male», для третьего — «male» (с маленькой буквы), для четвертого поставить цифру «2», а для пятого — «м» (на кириллице). Excel даже не поморщится и воспримет этот набор как пять абсолютно разных, независимых категорий. Более того, существует невидимый враг — скрытые пробелы. Если вы случайно нажмете пробел после буквы («М »), компьютер зафиксирует, что категория «М » принципиально отличается от категории «М». Вы не заметите этого глазами, но анализ будет испорчен.

Решение: если вы вынуждены работать в Excel, обязательно зайдите в настройки файла и включите функцию «Автозаполнение значений ячеек». Тогда при вводе первых букв программа сама предложит уже существующий в этой колонке вариант, что защитит вас от опечаток.

Когда у вашей переменной больше двух уровней, критически важно ответить на методологический вопрос: это вопрос с одиночным выбором («Выберите только один вариант») или с множественным («Выберите все подходящие варианты»)?

1. ***Одиночный выбор***. Мы создаем одну колонку. Например, переменная «всегда измеряю АД два раза в день» со значениями: «SD» (полностью не согласен), «D» (не согласен), «NO» (нет мнения), «A» (согласен), «SA» (полностью согласен).

2. ***Множественный выбор***: Представьте, что мы фиксируем типы бактерий, обнаруженные в мазке пациента. У одного человека может быть одновременно найдено несколько видов. Категорически запрещено вводить их через запятую в одну ячейку («кокки, бациллы»)! Такая колонка станет абсолютно непригодной для математической обработки. Как сделать правильно? Мы обязаны создать столько независимых колонок (переменных), сколько существует вариантов ответов. Каждая колонка становится бинарной (дихотомической) переменной с ответами «Да / Нет» (или «1 / 0»).

При переносе в компьютер чистых чисел (рост, вес, лабораторные маркеры) исследователи регулярно совершают две концептуальные ошибки.

1. ***Избыточное округление или искусственная группировка***. Всегда вносите числовые данные с максимальной точностью, которую обеспечивал измерительный прибор. Если весы показали «74,3 кг», вносите именно «74,3». Не нужно округлять до «74». Позже на компьютере вы сможете округлить результаты в любой момент, но «разукрупнить» обратно небрежно внесенные данные невозможно.

Еще хуже — самостоятельная группировка непрерывных данных в интервалы при вводе. Не создавайте в базе данных колонку «Возраст» со значениями «20–29 лет», «30–39 лет» и т.д. Запишите точный возраст в годах: «23», «35», «38». Перевести точные года в возрастные группы на этапе анализа софт сможет за полсекунды, но, если вы сразу ввели интервалы, вы никогда не сможете рассчитать точный средний возраст вашей выборки.

2. ***Смешивание чисел со сленгом и единицами измерения***. Никогда не пишите единицы измерения в ячейку с числом. Запись «12,5 г/дл» превращает всю колонку в текстовую. Пишите просто число «12,5», а единицы измерения вынесите в заголовок столбца. Не объединяйте два числа в одной колонке: попытка записать артериальное давление как «120/80» уничтожает числовую природу данных из-за символа слэша. Правильное решение — создать два столбца: систолическое и диастолическое АД.

Работа с датами и временем в медицине всегда окружена «подводными камнями». В исследовании, в котором ведётся наблюдение в динамике, постоянно нужно вычислять какие-то временные интервалы: время от начала лечения до ремиссии,

возраст на момент операции, длительность пребывания на койке. Современный статистический софт хранит дату и время не как текст, а как единую переменную, обозначающую конкретную точку на непрерывной линии времени. Внутри компьютера любая дата — это число дней (и долей дня), прошедших с определенной условной нулевой точки. В Microsoft Excel нулевая точка — это полночь на 31 декабря 1899 года. Соответственно, дата 21 ноября 2012 года, 18:00, физически хранится в памяти компьютера как число «41234,75» (дробная часть «,75» означает, что прошло ровно 3/4 суток, то есть 18 часов). В SPSS отсчет идет от 14 октября 1582 года (день введения Григорианского календаря). В SAS нулевой точкой выбрано 1 января 1960 года.

Методическое правило: никогда не считайте интервалы времени вручную перед вводом в компьютер! Вносите чистые даты (например, дату рождения пациента `DOB` и дату визита `VisDt`) и делегируйте расчет софту. Программа сделает это мгновенно и без ошибок, автоматически учитывая високосные года и разное количество дней в месяцах.

В клинической практике мы постоянно сталкиваемся с тем, что пациент не помнит точную дату события: «Я перенес операцию где-то в марте 2011 года». Однако компьютерные типы данных требуют ввода полной структуры (день, месяц, год) и отказываются принимать неполные строки.

В биостатистике для этого разработано «***правило середины***»: если неизвестен конкретный день месяца — мы условно устанавливаем его на 15-е число («15.03.2011»). Если неизвестны ни день, ни месяц (пациент помнит только год) — мы устанавливаем дату на середину года: «30 июня» или «1 июля» («01.07.2011»). Математически доказано, что на больших выборках погрешность такого допущения носит случайный характер и взаимно уничтожается, стремясь к нулю. При этом в базе данных рекомендуется создать бинарную переменную-индикатор (например, «Оригинальные_данные» со значениями «Да/Нет»), чтобы при детальном анализе вы могли отделить истинные даты от реконструированных.

Идеальных исследований не бывает — в реальных проектах всегда есть пропуски. Как их обозначать?

1. ***Категориальные пропуски***. Для качественных данных лучше всего вводить явный текстовый код уровня (например,

«Missing», «Unknown», «N/A»), чтобы при анализе причина отсутствия информации учитывалась.

2. ***Числовые пропуски***. В XX веке ученые массово использовали цифровые маркеры пропусков: «99», «999» или «9999». Предполагалось, что эти числа выбиваются из общего ряда. Но это опасная практика! Для обозначения пропущенных числовых данных оставляйте ячейку абсолютно пустой. Весь современный софт (включая Excel, R и SPSS) по умолчанию интерпретирует пустые ячейки как системные пропуски и автоматически корректно исключает их из текущих математических расчетов, не ломая структуру базы.

Когда ввод данных завершен, наступает этап верификации. Ошибки и опечатки при переносе информации из бумажных карт в компьютер происходят всегда, их средняя частота составляет от 1% до 5%. Перед запуском тестов базу необходимо «вычистить».

Есть пять простых и эффективных стратегий «чистки»:

1. ***Анализ экстремальных значений*** (Min/Max). Заставьте программу показать минимальное и максимальное значение для каждого числового столбца. Это занимает секунды, но мгновенно вскрывает опечатки. Если максимальное значение гемоглобина в выборке оказалось «145 г/дл» вместо «14,5», или вес пациента зафиксирован как «587 кг» вместо «58,7», вы сразу поймете, что произошел сдвиг десятичной запятой.

2. ***Сортировка массивов***. Отсортируйте колонку по возрастанию или убыванию. Числовые выбросы и опечатки в словах сразу окажутся в самом верху или в самом низу списка, став видимыми для глаз.

3. ***Построение частотных таблиц***. Запустите расчет частот для каждой категориальной переменной. Если на 200 пациентов вы увидите: «Male — 105», «Female — 93», «male — 1», «Femle — 1», вы мгновенно обнаружите две скрытые опечатки в кодировании пола.

4. ***Поиск скрытых символов***. Используйте стандартную функцию поиска в Excel для выявления случайных запятых в числовых столбцах (если система настроена на точки) или для поиска лишних пробелов.

5. ***Метод визуального картирования (карта листа)***. При использовании специализированных надстроек (например, PopTools для Excel) можно трансформировать таблицу в миниа-

тюрную карту, где каждая ячейка сжимается до размеров одного пикселя и окрашивается в цвет, соответствующий типу данных (например: формулы — зеленые, числа — синие, текст — красный, пустые ячейки — белые). Если посреди длинной синей полосы чисел вдруг вспыхнет красная точка, это означает, что при вводе числа вы случайно нажали букву, превратив всю ячейку в текст. Вы найдете ошибку визуально за секунду.

Завершающий аккорд правильной организации данных — создание словаря данных, или файла метаданных («данных о данных»).

Словарь данных — это независимый технический файл, который содержит паспорт каждой переменной, присутствующей в вашей основной базе. Он жизненно необходим для обеспечения преемственности науки. Если руководитель проекта или аналитик сменится, новый человек сможет войти в работу за один день, не гадая, что означают те или иные шифры.

Каждая строка в словаре данных описывает одну колонку основной базы и включает: системное имя переменной (например, «PAT_ID», «BPS_v1», «GLU_post»), короткое имя (до 8–10 символов) без пробелов, используемое для написания кода и формул, полное описание (например, «Систолическое артериальное давление на первом визите, мм рт. ст.»), тип данных, правила кодирования (например: «M = Male», «F = Female»; или «0 = Нет», «1 = Да»), количество знаков после десятичной точки или маска даты (ДД.ММ.ГГГГ).

1.7. Сбор данных: краткие итоги

- Любое исследование начинается с четкого планирования. Ошибки сбора данных невозможно исправить математическими формулами в конце.
- Мы изучаем небольшую случайную выборку, чтобы оценить параметры всей генеральной совокупности. Выборка обязана быть репрезентативной.
- В выборочных исследованиях мы пассивно наблюдаем за объектами. В экспериментах мы сами меняем условия, чтобы увидеть их влияние.
- Хорошая анкета состоит из закрытых или полузакрытых вопросов, не перегружает респондента и обязательно проходит проверку в пилотном исследовании.

- Высокий процент отказов портит качество данных. Бороться с ним нужно обучением интервьюеров, повторными звонками в разное время суток и поощрением участников.
- ИРК — это главный инструмент сбора данных в клиническом исследовании.
- Карта собирает только ту информацию, которая указана в протоколе.
- Структура ИРК строго соответствует визитам пациента.
- Правильный дизайн карты исключает свободный текст и стандартизирует все числа и даты.
- Ввод данных — это не механический перенос символов, а первый и важнейший этап статистического анализа.
- Каждая переменная должна строго соответствовать своему уровню измерения (номинальному, порядковому, интервальному или относительному).
- Категориальные данные кодируются короткими, самодокументируемыми текстовыми кодами. Множественный выбор категорически запрещено писать в одну ячейку — он дробится на систему бинарных колонок.
- Числа вносятся с максимальной точностью прибора, без ручного укрупнения в интервалы. Единицы измерения выносятся в заголовки.
- Пропущенные числовые значения обозначаются исключительно пустыми ячейками.
- База данных в обязательном порядке проходит процедуру очистки (Data Cleaning) через анализ минимумов/максимумов и частотные таблицы, а также снабжается подробным словарем данных.

Глава 2. Графическое и табличное представление данных

2.1. Зачем нужна визуализация данных?

Любое статистическое исследование начинается со сбора массива информации, но сухие столбцы цифр в базах данных не способны рассказать историю сами по себе. Чтобы обнаружить скрытые паттерны, структуры, тренды и взаимосвязи, нам необходимы инструменты представления данных. Графический метод — это не просто красивое оформление отчета. В первую очередь это незаменимый инструмент первичного разведочного анализа. Графики позволяют исследователю «почувствовать» распределение данных «изнутри» и мгновенно обнаружить невероятные, экстремальные значения или логические ошибки ввода (например, выявить запись о человеке очень высокого роста, имеющем при этом подозрительно малый вес).

2.2. Столбиковые диаграммы

Столбиковые диаграммы (иногда их называют столбчатыми диаграммами) — один из наиболее популярных и интуитивно понятных способов визуализации. Они превосходно резюмируют табличную информацию, переводя числовые величины в длину прямоугольных столбиков.

Допустим, мы изучаем проблемы недостатка питания у подростков и студентов. Нас интересуют средний рост и вес молодых пациентов в зависимости от пола. В тексте это выглядит так: Средний рост девочек составляет 148,8 см, а вес 48,2; средний рост мальчиков 163,9, вес 56,4.

Построив классическую столбиковую диаграмму, где вертикальная ось *Y* начинается строго от 0, мы наглядно увидим, что мальчики в среднем лишь немного выше и тяжелее девочек. Пропорции столбиков честно отражают реальность (рисунок 1).

Столбиковые диаграммы — излюбленный инструмент для графического «манипулирования» («честной лжи») в отчетах. Представьте, что какой-то исследователь хочет искусственно подчеркнуть и драматизировать разницу в росте, чтобы выгодно показать свои идеи о влияния питания на рост девочек. Для этого достаточно «отрезать» нижнюю часть оси *Y*, начав шкалу не с 0, а, например, со 140 см.



Рис. 1. Средний рост и вес.

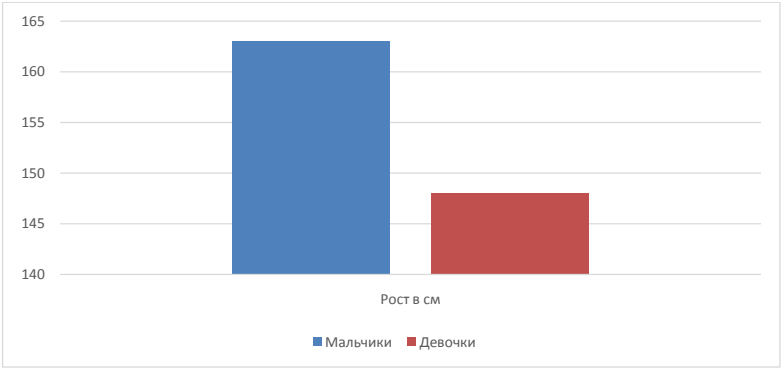


Рис. 2. Средний рост и вес.

При таком масштабировании, как на рисунке 2, столбик девочек едва приподнимется над осью, а столбик мальчиков окажется гигантским. У читателя возникнет ложное ощущение, что мальчики выше девочек чуть ли не в три-четыре раза.

Таким образом, важно соблюдать правило: при анализе и конструировании столбиковых диаграмм всегда проверять начало координат на осях. Срезание оси *Y* искажает визуальное восприятие пропорций и нарушает правила академической презентации.

2.3. Гистограммы

Очень часто начинающие исследователи путают столбиковую диаграмму и гистограмму. Запомните: гистограмма — это особый тип столбиковой диаграммы, разработанный исключительно для отображения частот непрерывных числовых данных. Если столбиковая диаграмма сравнивает дискретные категории (например, мальчиков и девочек), то гистограмма группирует непрерывную переменную (например, рост) в последовательные интервалы («карманы») и показывает, сколько наблюдений попало в каждый интервал. С её помощью мы оцениваем форму распределения: где находится центр, каков разброс значений, симметрично ли оно.

Чтобы построить гистограмму, непрерывный ряд данных разбивается на непересекающиеся смежные интервалы. Крайне важно определить, к какому интервалу принадлежит пограничная точка (обычно правая граница включается, а левая — нет). Допустим, к интервалу от 100 до 119 относится 3 подростка, от 120 до 129 тоже 3, от 140 до 159 – 10; от 160 до 179 – 10; и от 180 до 200 – 4. Гистограмма будет выглядеть как показано на рис. 3.

Как определить оптимальное количество интервалов (столбиков)? График должен быть читаемым: стандартное число столбиков обычно варьируется от 3 до 13. Чем больше объем выборки (n), тем больше интервалов мы можем себе позволить. Имеется грубое примерное правило выбора количества столбиков:

при n = 10 около 3 столбиков;
при n = 100 около 7 столбиков;
при n = 1000 около 10 столбиков;
при n = 10000 около 13 столбиков.

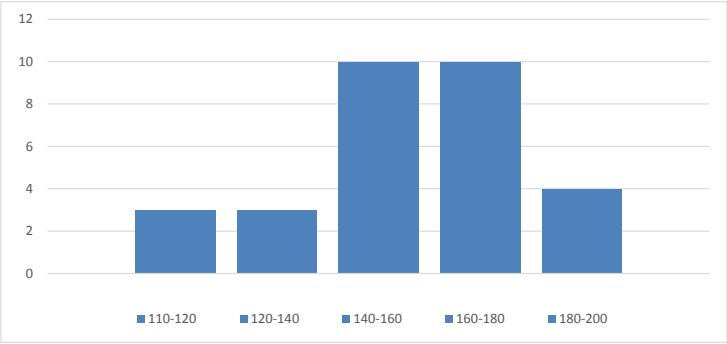


Рис. 3. Гистограмма.

Иногда в ПО можно встретить гистограммы с разной шириной столбиков. Категорически избегайте этого! Нарушение одинаковой ширины ломает восприятие площади столбиков, читатель дезориентируется, а адекватное чтение шкалы частот по оси *Y* становится невозможным. Только равные интервалы!

2.4. Круговые диаграммы

Круговые диаграммы идеальны для демонстрации долей («кусков пирога»), которые группы составляют от единого целого. Чаще всего они применяются для визуализации частот качественных признаков. Рассмотрим распределение по полу в нашей выборке: 13 девочек и 17 мальчиков. Эту информацию можно подать и в виде столбиков, и в виде круга. На рис. 4 представлены обе диаграммы, несущие одинаковую информацию.

Круговые диаграммы имеют преимущества и недостатки. Плюс в том, что, если групп мало (2–3 категории), круговая диаграмма выглядит максимально иллюстративно. Её ключевое преимущество — на ней невозможно «сманипулировать» осями и обрезать нижнюю часть, как на столбиковых диаграммах. Целое всегда остается 100-процентным кругом.

Минус в том, что круговая диаграмма становится абсолютно нечитаемой, если число групп превышает 6–7 категорий. Человеческий глаз плохо дифференцирует близкие по площади углы секторов. Если категорий много, исследователю безоговорочно следует вернуться к столбиковой диаграмме.

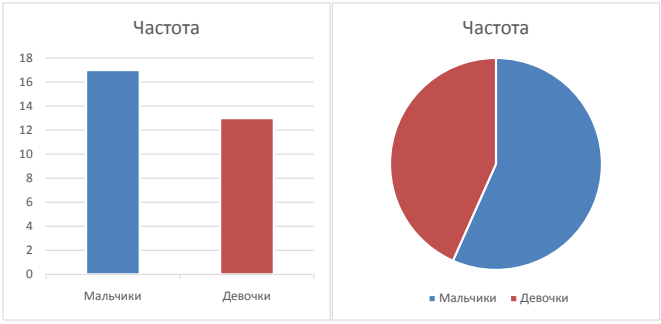


Рис. 4. Столбиковая и круговая диаграммы

2.5. Диаграммы рассеяния и пузырьковые диаграммы

Когда перед исследователем стоит задача изучить взаимосвязь между двумя непрерывными количественными переменными, на сцену выходит диаграмма рассеяния.

В примере с изучением питания логично предположить наличие связи между ростом и весом подростка. Мы строим график, где: ось *X* (горизонтальная) — независимая переменная (или, условно, фактор, причина), то есть рост; ось *Y* (вертикальная) — зависимая переменная (условно, эффект или следствие), вес.

Каждый подросток отображается на плоскости в виде отдельной точки с координатами (*x*, *y*). Разброс точек покажет нам тренд: если облако точек устремлено из левого нижнего угла в правый верхний, связь положительная — чем выше подросток, тем он тяжелее (рис. 5а).

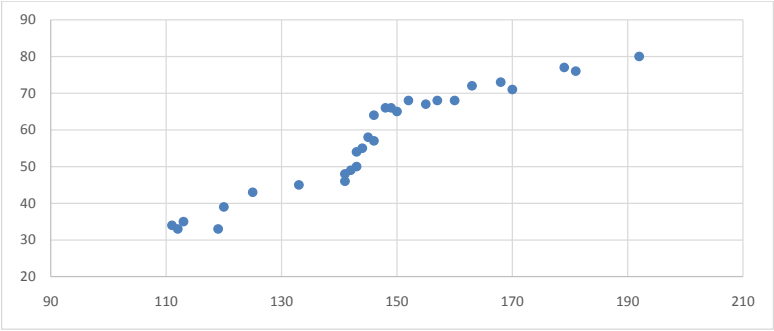


Рис. 5а. Диаграмма рассеяния

Однако, следует помнить (этот факт мы рассмотрим подробно позже), что с помощью графиков или статистических тестов можно доказать наличие математической связи (корреляции), **но невозможно доказать наличие причинно-следственной связи**. Пресса часто совершает эту ошибку, заявляя, что раз факторы взаимосвязаны на графике, то один строго порождает другой. Статистика же на деле лишь фиксирует сонаправленность изменений (мы к этому ещё вернёмся в дальнейшем).

Есть возможность, не прибегая к 3D-моделированию, отобразить на плоскости не две, а сразу три переменные: с помощью

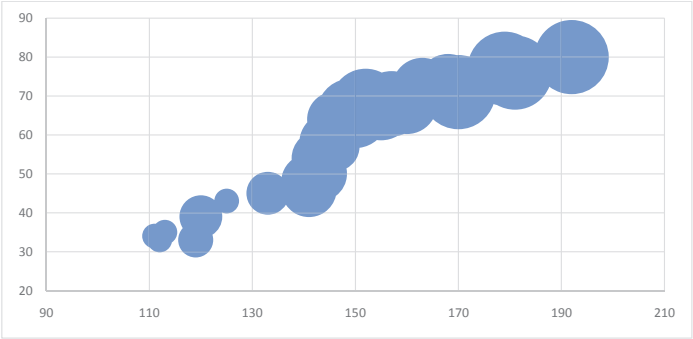


Рис. 5б. Пузырьковая диаграмма

пузырьковой диаграммы (рис. 5б). Это модернизированная диаграмма рассеяния, где вместо точек рисуются круги (пузырьки), а их размер кодирует значение третьего признака — например, возраст подростка. Большие пузырьки — старшие дети, маленькие — младшие.

Есть методическая особенность кодирования размера пузырьков: если значения третьей переменной различаются значительно, то с точки зрения визуальной справедливости её значение нужно делать пропорциональным площади круга (так как человеческий глаз оценивает объект объёмно). Однако, если разница между минимумом и максимумом третьего признака невелика (например, отличается максимум в 2 раза, как возраст от 10 до 17 лет), рекомендуется делать пропорциональным диаметр пузырька. Иначе визуальная разница между площадями кругов сотрется, и анализировать график будет трудно.

2.6. Линейные диаграммы

Линейные диаграммы незаменимы, когда нам нужно показать динамику, непрерывное изменение или тренд. Ось X на таких графиках почти всегда представляет собой непрерывную шкалу времени, возраста или стажа. Снова обратимся к примеру изучения недостатка питания, связав средний вес подростков в каждой возрастной группе со значением возраста. Поставив полученные точки на графики и соединив эти точки непрерывной линией, мы получим наглядную траекторию изменения веса с возрастом.

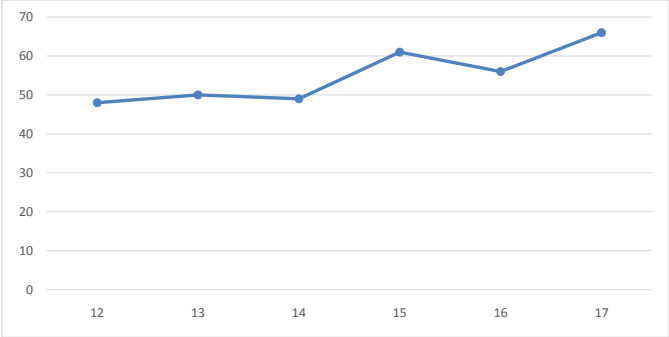


Рис. 6. Линейная диаграмма

2.7. Правила конструирования таблиц

Таблицы — это второй фундаментальный способ структурирования и презентации данных. Хорошая таблица должна быть самодостаточной и понятной читателю без обращения к основному тексту статьи. Обязательные компоненты таблицы:

1. ***Название таблицы***: Четкое, лаконичное, отражающее суть (например, «Количество детей по полу и возрасту»).
2. ***Заголовки столбцов***: Верхний уровень группировки.
3. ***Заголовки строк***: Боковой уровень группировки.
4. ***Ячейки*** (тело таблицы): Самое сердце таблицы. В них могут содержаться абсолютные частоты, проценты или средние значения других переменных.
5. ***Сноски***: располагаются в самом низу. Содержат информацию об источнике данных или важные комментарии (например, используемые сокращения).

Таблица всегда имеет как минимум два измерения: строки и столбцы. Они могут компоноваться по-разному. В одном случае пересечение строк и столбцов дает частоту распределения (табл. 1). В другом — одно из измерений может содержать несколько различных расчетных значений для одного признака (минимум, среднее, максимум) или средние значения для разных переменных (табл. 2).

Таблица 1

Пример двумерной таблицы абсолютных частот, n=30

Пол	Возраст, лет			
	12–13	14–15	16–17	Всего
Девочки	5	6	2	13
Мальчики	6	8	3	17
Всего	11	14	5	30

Таблица 2

Пример двумерной таблицы со средними значениями

Пол	Среднее значение		
	Роста, см	Веса, кг	Возраста, лет
Девочки	148,8	48,2	13,8
Мальчики	163,9	56,4	14,2

2.8. Проценты в таблицах

Зачастую представление данных в абсолютных частотах малоинформативно, особенно если мы хотим сравнить нашу выборку с генеральной совокупностью (например, данными регистра) для оценки её репрезентативности. В таких случаях частоты переводят в проценты.

Есть три основных способа расчета процентов в таблицах.

1. ***Общий процент***. Рассчитывается от общего объема всей выборки (n=30). Сумма всех внутренних ячеек дает 100%. Например, девочки 12–13 лет составляют: 16.7% (табл. 3).

2. ***Строчный процент***. За 100% принимается итог по каждой строке. Проценты суммируются до 100% горизонтально. Это необходимо, когда мы хотим изучить распределение признака внутри определенной группы (табл. 4).

Таблица 3

Общий процент подростков по возрасту и полу

Пол	Возраст, лет			
	12–13	14–15	16–17	Всего
Девочки	16,7%	20,0%	6,7%	43,3%
Мальчики	20,0%	26,7%	10%	56,7%
Всего	36,7%	46,7%	16,7%	100,0%

Таблица 4

Строчный процент подростков по возрасту и полу

Пол	Возраст, лет			
	12–13	**14–15**	**16–17**	**Всего**
Девочки	28,5%	46,2%	15,4%	100%
Мальчики	35,3%	47,1%	17,6%	100%

Таблица 5

Столбцовый процент подростков по возрасту и полу

Пол	Возраст, лет		
	12–13	**14–15**	**16–17**
Девочки	45,5%	42,9%	40,0%
Мальчики	54,5%	57,1%	60,0%
Всего	100,0%	100,0%	100,0%

3. ***Столбцовый процент***. За 100% принимается итог по каждому столбцу. Суммирование идет вертикально.

Как исследователю понять, какой тип процентов (строчный или столбцовый) выбрать для публикации? Используйте тот тип процентного «расщепления», где за 100% принимается переменная, выступающая в роли предполагаемой «причины» (фактора).

Рассмотрим пример. Подростков опросили, занимаются ли они кардиотренировками (Причина). Также они субъективно оценили свою физическую форму (Следствие: Плохая, Средняя, Хорошая). Мы ожидаем, что именно тренировки влияют на форму, а не наоборот. Значит, мы обязаны использовать строчный процент, сделав в строках «Занятие кардиотренировками» (Да/Нет) расчёты со строчным процентом (табл. 6).

Таблица 6

Строчный процент подростков по возрасту и полу

Занятие кардио-тренировками	Физическая форма				
	Плохая	**Средняя**	**Хорошая**	**Всего**	**n**
Нет	40%	40%	20%	100%	15
Да	20%	40%	40%	100%	15

По этой таблице сразу виден тренд: среди тех, кто тренируется, доля людей с хорошей формой в два раза выше (40% против 20%).

Важное предупреждение: Будьте предельно осторожны с процентами на малых выборках! Если ваша группа составляет всего несколько человек, расчет процентов теряет научную ценность. Запись «10%» звучит солидно, но, если за ней скрывается всего 1 подросток из 10, такое представление данных вводит в заблуждение. Сопровождайте проценты указанием абсолютного объема группы (n).

2.9. Представление данных: краткие итоги

- Графики — мощный инструмент выявления закономерностей и поиска ошибок, но они требуют строгого контроля масштаба осей.
- Гистограмма строится исключительно для непрерывных количественных данных и требует строго равных интервалов.
- Диаграмма рассеяния визуализирует связь двух переменных, но сама по себе не доказывает причинность. Пузырьковая диаграмма позволяет изящно добавить на плоскость третье измерение.
- Проценты в таблицах должны выбираться осознанно, опираясь на логику «фактор–результат», и применяться с осторожностью при малых объемах выборки.

Глава 3. Описание данных

3.1. Переход от графиков к числовым показателям

На прошлой лекции мы подробно разбирали, как представить собранную информацию с помощью графиков и таблиц. Мы видели, что визуализация, например гистограмма, дает мгновенное интуитивное представление о структуре данных. Однако в реальной научной и практической работе нам недостаточно просто посмотреть на график. Представьте, что вам нужно сравнить климат или уровень доходов в пяти разных странах. Нам необходимо выразить ключевые свойства этих графиков в виде конкретных, точных чисел. В статистике такие ключевые параметры называют ***выборочной статистикой*** (Sample Statistics), или показателями дескриптивной (описательной) статистики. «Сворачивать» распределения данных можно описывая две их главные характеристики:

1. ***Меры положения*** (Measures of Location), или иначе, анализ центральной тенденции.
2. ***Меры рассеяния*** (Measures of Dispersion), или разброс данных вокруг этого центра.

Прежде чем мы перейдем к формулам, давайте поймем философию статистической изменчивости (вариации). Статистика — это наука об описании изменчивости в данных. Её можно разделять на два принципиальных источника: ***систематическую и случайную изменчивость***.

В качестве иллюстрации рассмотрим четыре гистограммы, отображающие ежедневную среднюю температуру в четырех разных географических локациях.

Две верхние гистограммы имеют очень широкое основание (большой разброс). Это говорит о высокой изменчивости климата — там бывают как очень холодные, так и очень жаркие дни. Две нижние гистограммы, напротив, сжаты к центру (малый разброс). Климат в этих регионах исключительно стабилен. При этом две левые гистограммы центрированы вокруг отметки 10º C (холодный климат), а две правые гистограммы — вокруг 20º C (теплый климат). Положение центра распределения на оси X отражает систематическую изменчивость (климатическую зону). А степень ширины или разброса графика отражает случайную изменчивость (непостоянство погоды в этой зоне).

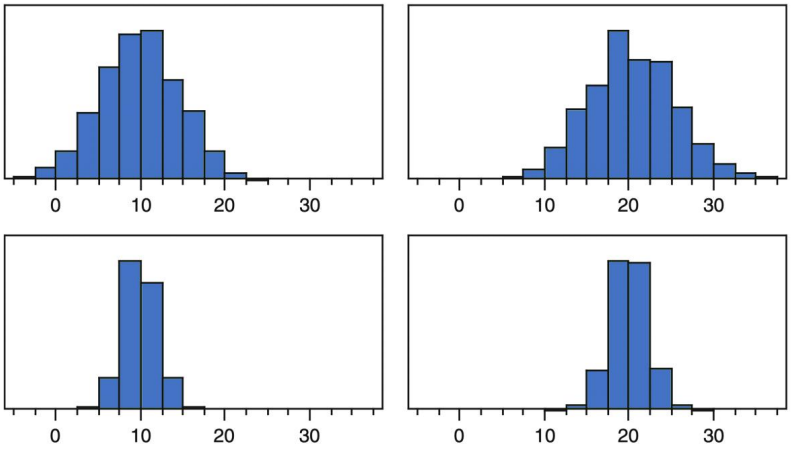


Рис. 7. Центральная тенденция и рассеивание

В статистических терминах используют как слово «центральная тенденция», так и синоним, «положение»» («Location»). Вместо слова «разброс» — чаще используют слово рассеяние или «дисперсия».

3.2. Меры центральной тенденции / положения

Как математически описать, где находится центр наших данных? Для этого существуют три классических показателя: среднее арифметическое, медиану и моду.

1. ***Среднее значение*** (Average / Mean) – это самый знакомый всем показатель. В обиходе его называют средним арифметическим, а в статистике обозначают символом (читается как «x с чертой», x-bar).

Важное терминологическое различие: строго говоря, в академической статистике слово «average» относится к характеристике выборки (конкретной группы людей, которую мы измерили), в то время как слово «mean» обозначает математическое ожидание для всей популяции (генеральной совокупности). Однако на практике эти термины часто (особенно в русском языке) используют как синонимы.

Среднее значение рассчитывается по формуле:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

где n — количество наблюдений в выборке.

Например, имеем данные о возрасте четырех детей: 3, 5, 6, 4 года. Сумма = 18. Среднее значение = 18 / 4 = 4,5 года.

Главный недостаток среднего в том, что оно чрезвычайно чувствительно к экстремальным значениям (выбросам). Если мы измеряем среднюю зарплату в отделе из 10 человек, где 9 получают скромные деньги, а один является миллиардером, этот единственный человек «разрушит» всю картину. Среднее значение станет огромным, но оно абсолютно не будет отражать реальный уровень жизни обычных сотрудников.

2. ***Медиана***. Этот показатель обозначается как M и является значением, которое находится строго посередине упорядоченного ряда данных. Она делит выборку на две равные части: 50% наблюдений лежат ниже медианы, 50% — выше.

Алгоритм поиска медианы таков: В начале следует отсортировать данные по возрастанию. Если количество наблюдений (n) нечетное, медиана — это центральное число выборки (Пример: есть ряд 12, 12, 13, 13, 13, 14, 14, 14, 14, 14, 14, 15, 15, 15, 16, 17, 17. Центральным является 9-е по счету число. M = 14).

Если количество наблюдений (n) четное, единого центрального числа нет. В этом случае медиана определяется как среднее арифметическое двух центральных чисел (пример: для ряда 3, 4, 5, 6 центральными являются 4 и 5. M = (4 + 5) / 2 = 4,5).

Главное преимущество медианы: она устойчива к экстремальным выбросам. Тот самый миллиардер из примера выше никак не изменит медиану, ведь он просто останется крайней точкой на правом фланге ряда.

3. ***Мода*** — это просто самое часто встречающееся значение в выборке. Для её поиска не нужны вычисления, достаточно подсчитать частоту каждого значения. (Пример: В упомянутом выше ряду чаще всего встречается число 14 [5 раз]. Значит, мода = 14). Главный недостаток моды: Она полезна только тогда, когда у нас много повторяющихся дискретных значений. Если же мы измеряем точные непрерывные величины (например, вес с точностью до грамма: 56.42 кг, 56.81 кг и т.д.), все значения в выборке могут оказаться уникальными и встретиться по одному разу. В таком случае мода теряет всякий практический смысл.

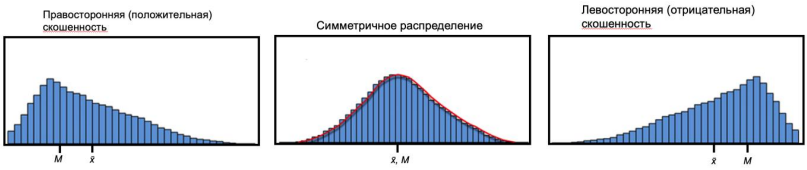


Рис. 8. Гистограммы с различными формами распределения данных

Взаимоотношение между средним значением и медианой позволяет нам без всяких графиков понять форму распределения данных (рис. 8).

Симметричное распределение: Левая и правая стороны гистограммы зеркальны (например, идеальное нормальное распределение). В этом случае среднее значение часто примерно равно медиане и часто также примерно равно моде.

Правосторонняя (положительная) скошенность: Распределение имеет длинный «хвост» из-за аномально больших значений в правой части. Классический пример — доходы населения. Большинство (это показывают высокие столбцы на гистограмме, их высота отражает высокую частоту) людей зарабатывает средние или небольшие суммы (ближе к началу оси X), и лишь очень немногие — сверхвысокие. Из-за длинного хвоста высоких доходов среднее значение искусственно утягивается вверх. Среднее значение больше медианы.

Левосторонняя (отрицательная) скошенность: Распределение имеет «хвост» из экстремально малых значений в правой части и больших – в левой. Пример: большинство (высокие столбцы) смертей приходится на пожилой возраст (дальше от начала оси X), а случаи смерти в молодом возрасте образуют вытянутый левый хвост. Эти малые значения «тянут» среднее вниз. Здесь среднее значение меньше медианы.

3.3. Меры рассеяния

Знать, где находится центр — это лишь половина дела. Нам жизненно необходимо понимать, насколько кучно или, наоборот, хаотично данные разбросаны вокруг этого центра. Для этого статистика использует меры рассеяния.

1. ***Размах***. Самая простая и очевидная мера. Размах — это разность между максимальным и минимальным значениями в

выборке. Минус: Этот показатель опирается всего на два крайних числа. Если в выборке есть хотя бы один случайный экстремальный выброс, размах мгновенно исказится и создаст ложное впечатление о гигантском рассеянии всех данных.

2. ***Дисперсия (Variance) и стандартное отклонение*** – это фундаментальные показатели в статистике. Они измеряют отклонение каждого конкретного значения выборки (x_i) от вычисленного среднего значения ().

Если мы просто сложим все отклонения, их сумма всегда будет равна нулю, так как положительные и отрицательные отклонения взаимно уничтожатся. Чтобы избавиться от знака «минус», отклонения возводят в квадрат. Выборочная дисперсия, s^2 вычисляется по формуле:

$$s^2 = \frac{(x_i - \bar{x})^2}{n - 1}$$

Обратите внимание: в знаменателе используется (n – 1), а не n. Деление на (n – 1) (число степеней свободы) необходимо для получения несмещенной оценки дисперсии именно для выборки.

Дисперсия измеряется в «квадратных единицах» (например, если мы измеряли рост в см, дисперсия будет в см^2), что крайне неудобно для интерпретации. Чтобы вернуться к исходным единицам измерения, из выборочной дисперсии извлекают квадратный корень и получают стандартное отклонение.

$$s = \sqrt{s^2} = \sqrt{\frac{(x_i - \bar{x})^2}{n - 1}}$$

Стандартное отклонение показывает, на какую величину в среднем отклоняются индивидуальные значения нашей выборки от среднего арифметического.

3. ***Интерквартильный размах*** (Interquartile Range — IQR). По аналогии с медианой, которая делит выборку пополам, мы можем разделить упорядоченные данные на четыре равные части (по 25% наблюдений в каждой). Точки раздела называют квартилями: первый квартиль (Q_1) обозначает, что 25% данных лежат ниже этой точки; второй квартиль (Q_2) – это и есть наша медиана (50% данных); третий квартиль (Q_3) –75% данных лежат ниже этой точки (или 25% — выше нее).

Соответственно, интерквартильный размах — это расстояние между третьим и первым квартилями:

$$IQR = Q_3 - Q_1$$

IQR охватывает центральные 50% нашей выборки, отсекая по 25% самых крайних значений с каждого бока.

Подобно медиане, IQR абсолютно устойчив к выбросам и экстремальным значениям.

4. ***Коэффициент вариации*** / Относительный разброс. Если стандартное отклонение измеряется в абсолютных единицах (см, кг), то как нам сравнить разброс двух совершенно разных признаков? Например, у кого изменчивость веса сильнее: у мышей или у слонов? Очевидно, что абсолютное отклонение в килограммах у слонов будет огромным, но в относительных масштабах их изменчивость может быть одинаковой. Для этого вычисляют коэффициент вариации (CV) — отношение стандартного отклонения к среднему значению, выраженное в процентах:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Это безразмерная величина, позволяющая сравнивать вариабельность любых процессов и выборок между собой.

3.4. «Ящик с усами» и «выбросы»

Давайте попробуем освоить один из самых мощных инструментов разведочного анализа данных — диаграмму «ящик с усами». Нам надо научиться читать структуру распределения по графику, рассчитывать его ключевые компоненты вручную и профессионально идентифицировать аномальные значения (выбросы). В 1977 году выдающийся американский математик и статистик Джон Тьюки (John Tukey) предложил гениальный графический инструмент, который буквально совершил революцию в визуализации данных. Он назвал его Box-and-Whisker Plot (диаграмма «ящик с усами», или просто боксплот).

«Ящик с усами» — это геометрическое отображение главных числовых характеристик выборки, которое позволяет с одного взгляда оценить:

- Где находится центр данных;
- Насколько данные однородны или разбросаны;
- Симметрично ли распределение или оно скошено;
- И самое главное — есть ли в выборке «выбросы» (аномалии).

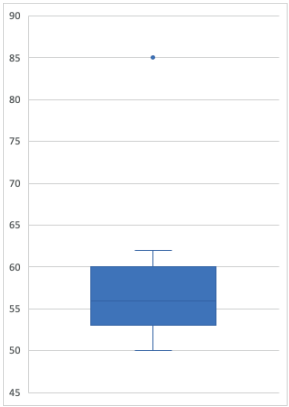


Рис. 9. Диаграмма «ящик с усами»

Давайте построим этот график снизу вверх (рис. 9). В основе боксплота лежит деление упорядоченного массива данных на квартили — четыре равные части по 25% наблюдений в каждой.

1. ***Ящик***. Сам прямоугольный ящик очерчивает границы центральных 50% ваших данных. Нижняя (или левая) граница ящика — это первый квартиль (Q_1). Ниже этой отметки лежат самые маленькие 25% всех значений выборки. Верхняя (или правая) граница ящика – это третий квартиль (Q_3). Выше этой отметки лежат самые большие 25% значений.

Линия внутри ящика – это второй квартиль (Q_2), более известный как медиана. Она делит центральные 50% данных строго пополам.

Важная деталь: высота (или ширина) самого ящика в точности равна IQR (интерквартильному размаху): IQR = $Q_3 – Q_1$. Это зона, где сосредоточена самая плотная и репрезентативная часть выборки.

2. «***Усы***». От границ ящика в обе стороны расходятся прямые линии — те самые «усы». Вопреки распространенному заблуждению новичков, усы далеко не всегда тянутся к абсолютному минимуму и максимуму выборки. Длина усов строго ограничена правилом Тьюки, чтобы отсечь аномалии. По стандарту длина одного уса не может превышать полтора интерквартильных размаха (1,5 IQR).

Нижний ус заканчивается на минимальном реальном значении выборки, которое попадает в диапазон до границы: $[Q_1 - 1{,}5 \times \text{IQR}]$; Верхний ус заканчивается на максимальном реальном значении выборки, которое попадает в диапазон до границы: $[Q_3 + 1{,}5 \times \text{IQR}]$. Если за пределами усов данных больше нет, то ус просто упирается в реальный минимум или максимум. Но если там есть экстремальные точки, усы «замораживаются» на рассчитанной границе, а всё, что дальше, объявляется выбросами.

Выброс — это наблюдение в массиве данных, которое стоит настолько далеко от остальных значений, что вызывает обоснованное подозрение, не является ли оно ошибкой измерения или уникальным единичным феноменом.

Представьте, что мы измеряем пульс у группы студентов в покое. У тридцати человек он равен 65–85 ударам в минуту. И вдруг у тридцать первого студента прибор фиксирует пульс 240. Или мы исследуем доходы врачей в клинике: у всех зарплата в районе 70–150 тысяч рублей, а у главного врача, который также владеет акциями сети частных клиник доход 12 миллионов. Это классические выбросы.

Математический критерий Тьюки пригоден для поиска выбросов. Точка признается выбросом, если она лежит за пределами указанных выше интервалов. На диаграмме «ящик с усами» эти значения не прячутся внутри усов. Они обозначаются отдельными точками, звездочками или крестиками далеко за пределами линий «усов».

Прежде чем удалять выброс из исследования, ученый обязан разобраться, откуда он взялся. Выбросы делятся на три категории.

1. ***Ошибки ввода*** или технические сбои: например, медсестра случайно поставила лишний ноль в результатах анализа крови, или датчик температуры на мгновение потерял контакт со средой. Такие выбросы подлежат безусловному удалению или исправлению.

2. ***Ошибки отбора***: в выборку здоровых людей случайно затесался пациент с тяжелой хронической патологией сердца. Его показатели истинны, но он не принадлежит к изучаемой популяции. Его нужно исключить на этапе верификации критериев включения.

3. ***Уникальная естественная вариация***: это реальное, безошибочное значение. Например, гениальный спортсмен с уникальным объемом легких. Такие данные удалять нельзя — именно они могут содержать в себе ключ к новому научному открытию.

Чтобы закрепить материал, давайте рассчитаем и нарисуем боксплот для небольшой выборки. Допустим, мы измерили вес 11 подростков (в кг):

50, 52, 53, 54, 55, 56, 57, 58, 60, 62, 85.

Шаг 1: сортировка (уже выполнена по возрастанию). Количество элементов $n = 11$ (нечетное число);

Шаг 2: находим медиану (Q_2). Центральным элементом является 6-я точка, то есть $Q2 = 56$. Она делит массив на две половины по 5 элементов.

Шаг 3: находим первый (Q_1) и третий (Q_3) квартили. Q_1 — это середина нижней половины выборки (состоящей из чисел 50, 52, **53**, 54, 55), то есть 53. Q_3 — это медиана верхней половины выборки (состоящей из чисел 57, 58, **60**, 62, 85), то есть 60.

Шаг 4: рассчитываем интерквартильный размах, $IQR = Q_3 - Q_1 = 60 - 53 = 7$.

Шаг 5: определяем лимиты для «усов» и ищем выбросы. Шаг усов: $1{,}5 \times IQR = 1{,}5 \times 7 = 10{,}5$.

Нижний предел уса: $Q_1 - 10{,}5 = 53 - 10{,}5 = 42.5$. Наше минимальное значение (50) значительно больше. Значит, нижний ус упрется ровно в точку 50.

Верхний предел уса: $Q_3 + 10{,}5 = 60 + 10{,}5 = 70{,}5$. Посмотрим на наш ряд. Точки 57, 58, 60, 62 укладываются в лимит. А вот точка 85 находится далеко за пределами границы 70,5!

Итог визуализации: мы рисуем ящик от 53 до 60. Проводим линию медианы на уровне 56. Нижний ус тянется до точки 50. Верхний ус тянется до максимального легитимного значения — то есть до точки 62. А аномальное значение 85 мы рисуем в виде одинокой точки сверху. Мы математически доказали, что подросток весом 85 кг является «выбросом» для данной группы.

Как читать форму распределения по «ящику»? Опытный аналитик смотрит на боксплот так же, как кардиолог на кардиограмму. Форма ящика рассказывает о симметрии данных:

1. Если медиана лежит строго по центру ящика, а усы имеют одинаковую длину — перед нами симметричное распределение (например, нормальное).

2. Если медиана смещена к левой (нижней) границе, а правый ус и правая половина ящика сильно вытянуты вперед (и за ними идут точки-выбросы) — это правосторонняя скошенность. Это значит, что большинство людей имеют небольшие показатели, но есть группа лиц с экстремально высокими значениями (ситуация с доходами населения).

3. Если медиана прижата к правому (верхнему) краю, а левый ус длиннее — это левосторонняя скошенность (смещение в сторону малых величин).

3.5. Связь типа данных и описательной статистики

Выбор статистического показателя в описательной статистике жестко продиктован типом данных, с которыми вы работаете. Ошибки на этом этапе делают весь последующий анализ бессмысленным.

Необходимо учитывать базовое разделение данных на два типа (см. ранее рассмотренные уровни в типах данных):

1. ***Количественные данные***: результаты подсчетов или измерений. Числа здесь имеют физический смысл, с ними можно производить математические операции.

2. ***Качественные / Категориальные данные***: обозначают принадлежность к группе или категории. Бывают номинальными (пол, цвет глаз) или порядковыми (уровень удовлетворенности: низкий, средний, высокий).

Соответствие типов данных и допустимых дескриптивных статистик представлено в табл. 7.

Запомните критическое правило: попытка рассчитать «средний пол» или «средний цвет глаз» — это абсурд. Для номинальных данных единственной мерой центра является мода. Для сильно скошенных экономических показателей (зарплаты) среднее значение подменяется медианой, чтобы избежать искажения картины из-за единичных богатых представителей.

Таблица 7

Типы данных и допустимые показатели статистики

Тип данных	Пример	Допустимые меры положения (центральной тенденции)	Допустимые меры рассеяния
Количественные (Симметричные)	Рост, температура	Среднее значение, Медиана	Стандартное отклонение, Дисперсия, Размах, IQR
Количественные (Скошенные)	Доходы населения	Медиана (Среднее искажается!)	IQR (Стандартное отклонение искажается!)
Качественные (Порядковые)	Успеваемость (неуд, уд, хорошо, отлично)	Медиана, Мода	Иногда квартили
Качественные (Номинальные)	Пол (мальчики/девочки)	Только Мода	Не применимы (считаются только частоты, %)

3.6. Описание данных: краткие итоги

- Описательная статистика сворачивает массивы данных с помощью двух параметров: меры положения (центр) и меры рассеяния (разброс).
- Среднее арифметическое и стандартное отклонение — идеальная пара для описания красивых, симметричных распределений без выбросов.
- Медиана и интерквартильный размах (IQR) — это устойчивая пара, которая спасает исследователя в случае скошенных распределений или наличия грубых ошибок и «выбросов».
- «Ящик» визуализирует центральные 50% данных (IQR), защищая нас от искажений, которые создают экстремальные точки, а линия внутри ящика — это медиана, устойчивое сердце вашей выборки.
- Все точки, оказавшиеся за пределами усов — это официальные кандидаты в выбросы, требующие пристального внимания исследователя.
- Прежде чем запустить расчет описательной статистики в Excel или другой программе, вы обязаны определить тип ваших данных (количественные или качественные), чтобы выбрать математически обоснованный показатель.

Глава 4. Нормальное распределение

4.1. Ключевые характеристики нормального распределения

Представьте себе ситуацию: вы проводите бессонные ночи за написанием рукописи своей важной работы, и чтобы не спать, покупаете в магазине кофе любимого сорта (другой вы не покупаете) в пачках, на которых написано «500 г». Если вы придете домой и взвесите её на очень точных электронных весах, вы вряд ли обнаружите там ровно 500,0000 г. Скорее всего, вес составит 501,2 г или 498,7 г. Если мы взвесим сотни таких пачек, то заметим, что в среднем вес действительно «крутится» около 500 г, а сильные отклонения (например, меньше 450 г или больше 550 г) практически не встречаются. Вес колеблется в разумных пределах случайной вариации (например, от 490 до 510 г).

Эта изменчивость описывается непрерывным статистическим распределением. Главным, самым мощным и часто встречающимся в природе распределением является ***нормальное распределение***. Именно оно идеально описывает не только физические параметры (рост, вес, температура), но и массу процессов в медицине, да и в экономике, менеджменте и администрировании.

График нормального распределения представляет собой симметричную, колоколообразную кривую, внешне напоминающую сглаженную гистограмму. Нормальное распределение уникально тем, что оно полностью задается всего двумя параметрами, которые мы уже умеем рассчитывать:

1. ***Среднее значение*** (μ, мю) — определяет центр распределения (положение пика на оси X).

2. **Стандартное отклонение** (σ, сигма) — определяет разброс данных (ширину «колокола»).

Поскольку речь идет о параметрах генеральной совокупности (всех пачек кофе, производимых фабрикой), в статистике их принято обозначать греческими буквами, в отличие от латинских выборочных характеристик ($\overline{x}$, s).

Если мы меняем только μ, график сдвигается влево или вправо по оси X, не меняя формы (рис. 10). Если мы меняем σ, меняется «сжатость» графика. При малой σ колокол узкий и высокий (данные кучные), при большой σ — широкий и плоский (данные сильно рассеяны).

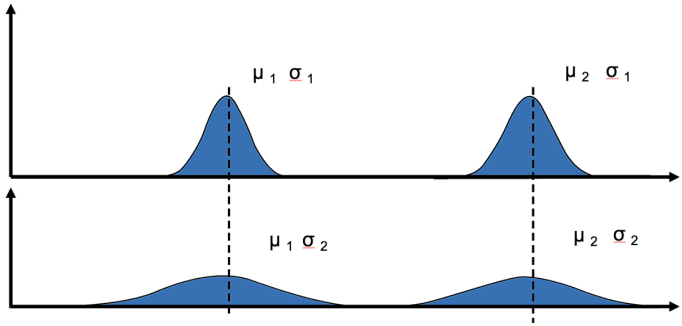


Рис. 10. Средние и стандартные отклонения

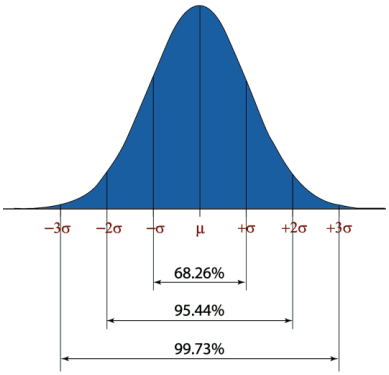


Рис. 11. Вероятности в нормальном распределении

4.2. Правило трех сигм

Для любого нормального распределения строго выполняются следующие математические пропорции:

1. 68% всех данных всегда лежат в пределах одного стандартного отклонения от среднего: $\mu \pm 1\sigma$;

2. 95% (точнее, 95.44%) всех данных лежат в пределах двух стандартных отклонений: $\mu \pm 2\sigma$;

3. 99.7% (практически всё распределение) укладывается в рамки трех стандартных отклонений: $\mu \pm 3\sigma$.

4.3. Функция плотности и функция распределения

Давайте ещё раз вспомним про типы данных. Нам относительно хорошо понятно про вероятности дискретных случайных величин. Количество детей в семье, оценка на экзамене — все эти величины принимают изолированные, четкие значения, которые можно пересчитать. Для них легко составить таблицу вероятностей: вероятность получить «пятерку» — 0,3, вероятность получить «четверку» — 0,5 и так далее. Сумма всех вероятностей в такой таблице всегда равна единице. Но что делать, если мы переходим к непрерывным случайным величинам, как в том примере с точным весом пачки кофе?

Попробуйте ответить на вопрос: какова вероятность того, что мы купим пачку кофе весом ровно 500,0000000… грамм? Не 500,0000001 и не 499,999999, а абсолютно точно: 500,0000000…? Математический ответ вас удивит: эта вероятность равна нулю. Поскольку число возможных значений непрерывной величины бесконечно велико и их невозможно пересчитать, вероятность попасть в одну конкретную изолированную точку всегда стремится к нулю ($P(X = x) = 0$).

Как же описывать случайную величину, если вероятность каждого её отдельного значения равна нулю? Статистики нашли изящный выход: нужно оценивать вероятность попадания не в точку, а в определённый интервал (например, от 495 до 505 гр). Для этого и были созданы два фундаментальных инструмента: функция плотности вероятности и функция распределения.

1. ***Функция плотности*** (Probability Density Function, обозначается малой f, $f(x)$ и иногда pdf(x)) — тот самый классический колокол. Она показывает форму распределения, но сама по себе для расчетов вероятностей неудобна. По большому счёту, функция плотности нужна в основном для того, чтобы рисовать красивые иллюстрации в учебниках.

2. ***Функции распределения*** (в зарубежной литературе её называют Cumulative Distribution Function или CDF) — имеет более широкое применение. В реальных прикладных задачах мы практически всегда используем именно функцию распределения. Обычно её обозначают заглавной буквой $F(x)$, иногда cdf(x). Функция распределения $F(x)$ — это функция, которая для любого числа x задает вероятность того, что случайная величина X примет значение, меньшее или равное этому x: $F(x) = P(X \leq x)$

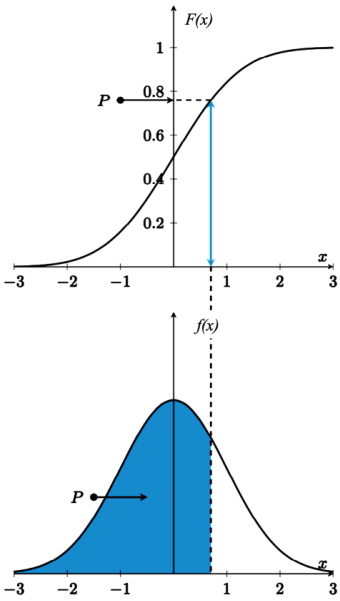


Рис. 12. Сравнение функций распределения и плотности

Слово «кумулятивная» означает «накопительная». Представьте, что мы движемся по оси X слева направо (от $-\infty$ до $+\infty$) со специальным «сачком» и собираем, накапливаем вероятности всех точек, которые остались позади нас. Значение $F(x)$ на оси Y показывает ***накопленную*** площадь (вероятность) от крайнего левого хвоста до точки x. Для непрерывных величин график $F(x)$ обычно представляет собой плавную, красивую S-образную кривую.

Давайте сопоставим графики обеих функций для самого популярного в мире распределения — нормального (гауссовского) колокола. Посмотрите на их взаимосвязь на рисунке 12: Центр распределения (Медиана / Матожидание): на графике плотности $f(x)$ это самая высокая точка — вершина колокола. На графике распределения $F(x)$ в этой же точке высота равна половине — 0,5 (накоплено 50% вероятности). Левый хвост: на графике плотности значения близки к нулю (события маловероятны). На графике распределения линия идет горизонтально низко, почти касаясь нуля (вероятность практически не прирастает). Правый

хвост: на графике плотности кривая снова падает к нулю. На графике распределения S-образная кривая выходит на плато на уровне 1,0 (мы собрали почти всю возможную вероятность, расти дальше некуда).

Важнейшие свойства функции распределения F(x):

1. Её значения всегда лежат в диапазоне от 0 до 1: $0 \leq F(x) \leq 1$. Вероятность не может быть отрицательной или превышать 100%.

2. Она никогда не убывает: если мы двигаемся вправо, наш «сачок» может только пополняться новыми вероятностями (или оставаться прежним, если на каком-то участке значений нет), но накопленная вероятность не может уменьшиться.

3. В самом начале пути, у левого края, накопленная вероятность равна нулю, то есть $F(-\infty) = 0$. В самом конце пути, когда мы прошли все возможные значения, мы собрали 100% вероятностей: $F(+\infty) = 1$.

Как использовать $F(x)$ на практике? Главная ценность функции распределения заключается в том, что она позволяет мгновенно находить вероятность попадания случайной величины в любой интервал $[a; b]$. Формула выглядит так:

$$P\ (a \leq X \leq b) = F(b) - F(a)$$

Логика проста: Чтобы узнать вероятность того, что рост человека будет от 170 до 180 см, нужно взять накопленную вероятность для 180 см (все, кто ниже 180 см) и вычесть из неё накопленную вероятность для 170 см (все, кто ниже 170 см). Останутся как раз люди интересующего нас диапазона.

Вернёмся к функции плотности вероятности. Математически функция плотности — это просто производная от функции распределения:

$$f(x) = F'(x)$$

Если говорить образно, функция распределения $F(x)$ показывает, сколько вероятности накопилось к данному моменту, а функция плотности $f(x)$ показывает, с какой скоростью эта вероятность накапливалась в конкретной точке x.

Для сравнения, необходимо также уяснить ***важнейшие свойства функции плотности f(x)***:

1. Она всегда неотрицательна: $f(x) \geq 0$. График плотности никогда не опускается ниже оси X.

2. Условие нормировки (главное свойство): Вся площадь под графиком плотности вероятности (от – ∞ до + ∞) строго равна единице:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Этот интеграл — математическое выражение того факта, что случайная величина со 100%-й вероятностью примет хоть какое-нибудь значение из всех возможных.

Как рассчитывать вероятности с помощью плотности? Поскольку $f(x)$ — это плотность, вероятность попадания в интервал $[a; b]$ равна площади криволинейной трапеции под графиком плотности на этом участке. В терминах математики вероятность — это определенный интеграл:

$$P\ (a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

Возвращаясь к нашему примеру с пачкой кофе: вероятность того, что она будет весить ровно 500,0000… гр — это площадь «линии» шириной в ноль. Площадь линии равна нулю. Но площадь под колоколом между 495 и 505 гр имеет конкретный физический размер — это и есть искомая вероятность.

4.4. Стандартизированное нормальное распределение и Z-оценки

Представьте, что вы — проректор и работаете в приемной комиссии престижного инновационного университета. К вам поступают заявления от двух абитуриентов из разных стран с разными системами национальных экзаменов. Первый абитуриент сдавал национальный экзамен типа «А» и набрал 75 баллов. Второй абитуриент сдавал экзамен типа «Б» и набрал 1150 баллов. Кто из них справился лучше и более достоин зачисления?

В исходном виде эти числа абсолютно несопоставимы. Это классическая статистическая проблема сравнения «яблок с апельсинами». Чтобы ответить на вопрос, нам нужно знать контекст: какое среднее значение признака в каждой стране и насколько силен разброс результатов.

Допустим, мы собрали эти данные: для экзамена «А»: средний балл (μ) равен 60, а стандартное отклонение (σ) равно 5; для экзамена «Б»: средний балл (μ) равен 1000, а стандартное отклонение (σ) равно 100.

Теперь картина становится яснее: первый студент оторвался от среднего значения своей группы на 15 баллов (75 – 60), а второй — на 150 баллов (1150 – 1000). Но из-за разных масштабов (сигм) мы все еще не можем сопоставить их напрямую.

Для решения подобных задач в статистике используется метод стандартизации, а универсальной мерой сравнения выступает *Z*-оценка.

Z-оценка (стандартизированное отклонение) — это безразмерный статистический показатель, который показывает, на какое количество стандартных отклонений конкретное наблюдаемое значение (X) отклоняется от μ в большую или меньшую сторону.

Математически расчет *Z*-оценки выглядит очень просто:

$$Z = \frac{X - \mu}{\sigma}$$

где X — исходное значение признака (сырые данные), μ — среднее значение в популяции (генеральной совокупности) и σ — стандартное отклонение в популяции.

Примечание: если мы работаем не с целой популяцией, а с выборкой, формула сохраняет свой смысл, но записывается латинскими буквами:

$$Z = \frac{x - \bar{x}}{s}$$

где $\bar{x}$ — выборочное среднее, а s — выборочное стандартное отклонение.

Логика формулы:

1. Шаг первый: Центрирование ($X - \mu$). Мы вычитаем среднее значение, чтобы узнать чистый сдвиг от центра. Если результат положительный — наше число выше среднего, если отрицательный — ниже.

2. Шаг второй: Нормирование (деление на σ). Мы делим полученный сдвиг на «эталонный метр» распределения — стандартное отклонение. Этим действием мы уничтожаем физические единицы измерения (сантиметры, рубли, граммы, баллы) и переходим к универсальному масштабу «шагов-сигм».

Если мы возьмем абсолютно любой массив данных, имеющий форму нормального распределения (рост людей, вес деталей, результаты тестов), и для каждой точки рассчитаем *Z*-оценку, мы совершим математическое преобразование распределения. Мы получим ***стандартизированное нормальное распределение***.

Оно обладает уникальными, фиксированными характеристиками: Его среднее значение всегда строго равно нулю ($\mu = 0$). Это логично, ведь если мы подставим в формулу значение X, в точности равное среднему μ, мы получим $Z = 0$. Центр колокола перемещается в начало координат. Его стандартное отклонение всегда строго равно единице ($\sigma = 1$). Дисперсия также равна единице ($\sigma^2 = 1$).

Благодаря стандартизации, эмпирическое правило (правило трех сигм) для Z-распределения звучит максимально просто:

1. Примерно 68% всех возможных Z-оценок лежат в интервале от –1 до +1.

2. Примерно 95% всех Z-оценок лежат в интервале от –2 до +2 (если точнее, 95% площади находятся между –1,96 и +1,96).

3. Примерно 99,7% данных укладываются в диапазон от –3 до +3. Любая Z-оценка, которая выходит за рамки ±3, считается экстремальной редкостью (выбросом).

Возвращаемся к абитуриентам, теперь у нас достаточно знаний, чтобы разрешить спор двух абитуриентов. Давайте переведем их сырые результаты в стандартизированные Z-оценки.

Абитуриент 1 (Экзамен «А»): $X = 75$, $\mu = 60$, $\sigma = 5$.

$$Z_1 = \frac{75 - 60}{5} = +3$$

Результат этого студента находится на расстоянии трех положительных стандартных отклонений от среднего. Он сдал экзамен феноменально блестяще, обогнав почти 99,7% всех остальных тестируемых в своей стране.

Абитуриент 2 (Экзамен «Б»): $X = 1150$, $\mu = 1000$, $\sigma = 100$.

$$Z_2 = \frac{1150 - 1000}{100} = +1{,}5$$

Этот студент тоже молодец, его результат выше среднего. Но он оторвался от центра лишь на 1,5 стандартных отклонения. В нормальном распределении Z-оценка +1,5 означает, что он обошел примерно 93% сверстников.

Вердикт приемной комиссии: Первый абитуриент ($Z = +3{,}0$) показал гораздо более высокий уровень относительно своей группы, чем второй абитуриент (Z = +1,5). Выбор очевиден.

Помимо сравнения разнородных шкал, Z-оценки активно применяются в самых разных сферах человеческой деятельности.

1. ***Поиск аномалий и выбросов***. В Data Science и анализе данных *Z*-оценка — один из первых инструментов очистки данных. Если алгоритм видит в базе транзакций или медицинских показателей строчку, у которой $Z > 3{,}5$ или $Z < -3{,}5$, эта сессия или пациент автоматически помечаются как аномалия, требующая проверки на ошибку оборудования или, допустим, какой-то обман.

2. ***Объединение метрик в бизнесе***. Представьте, что вам нужно составить рейтинг эффективности филиалов банка. У вас есть три показателя: объем выручки (в миллионах), количество новых клиентов (в персонах) и уровень удовлетворенности (в процентах). Вы не можете их просто сложить. Но если вы переведете показатели каждого филиала в *Z*-оценки, вы сможете рассчитать для каждого отделения среднюю *Z*-оценку. Физический смысл этого индекса: «На сколько в среднем данный филиал работает лучше или хуже, чем стандартный филиал сети».

3. ***Медицинская диагностика***. Когда пациентам проводят исследование плотности костей (денситометрию) для диагностики остеопороза, прибор выдает результат не в физической массе минералов, а в виде *Z*-предикторов (сравнение со средней нормой для людей того же возраста и пола). Снижение *Z*-оценки ниже –2,0 свидетельствует о патологическом разрежении костной ткани.

4.5. Квантили и процентили

Разбирая нормальное распределение, мы научились решать «прямую задачу»: у нас есть конкретное значение признака (например, пачка кофе весом 495 граммов), и мы ищем вероятность встретить пачку с таким весом или меньше. Математически мы брали точку на оси *X* и искали накопленную площадь (вероятность) под колоколом слева от неё. Однако в реальной практике исследователь гораздо чаще сталкивается с «обратной задачей», например:

- Мы принимаем на работу дворников и проводим тест на IQ среди кандидатов. Нам нужно отсеять 90% людей с рядовыми результатами и пригласить на собеседование только топ-10% лучших. Какой пороговый балл теста нам нужно установить?

- На приеме ребенок в возрасте 5 лет имеет рост 100 см. Какая доля детей в популяции ниже его, и укладывается ли этот показатель в здоровую норму?

Во всех этих случаях нам известна доля данных (вероятность) на оси *Y,* и мы хотим найти пороговое значение на оси *X*, которое эту долю отсекает. Инструментом для решения таких задач служат квантили и их самое популярное подмножество — процентили.

Что такое квантиль? Давайте дадим строгое, но интуитивно понятное определение. Представьте, что мы взяли все наблюдения нашей выборки или популяции и отсортировали их по возрастанию: от самых маленьких до самых больших.

Квантиль уровня p (где $0 \geq p \geq 1$) — это такое число (значение признака), ниже которого лежит ровно доля p всех наблюдений распределения. Соответственно, выше этого числа лежит доля $1 — p$ наблюдений.

Обозначается квантиль обычно как x_p или Q_p.

Если мы посмотрим на кумулятивную функцию распределения $F(x)$ (ту самую S-образную кривую, которую мы изучали ранее), то поиск квантиля — это просто движение в обратную сторону:

1. Мы находим долю p (например, 0,75) на вертикальной оси *Y.*
2. Ведем горизонтальную линию до пересечения с S-образной кривой.
3. Опускаемся вниз на ось *X* и находим значение квантиля $x_{0,75}$.

В зависимости от того, на сколько равных частей мы делим наше упорядоченное распределение, квантили называют по-разному:

1. ***Медиана***: это квантиль уровня 0,5 ($x_{0,5}$). Она делит распределение ровно на две равные половины. В силу идеальной симметрии нормального распределения, медиана здесь совпадает со средним арифметическим (μ).

2. ***Квартили***: они делят данные на 4 равные части (по 25% или 0,25 в каждой).

Q_1 (первый или нижний квартиль) = квантиль уровня 0,25.

Q_2 (второй квартиль) = Медиана = квантиль уровня 0,50.

Q_3 (третий или верхний квартиль) = квантиль уровня 0,75.

3. ***Децили***: делят распределение на 10 равных частей (уровни 0,1, 0,2, … 0,9).

Когда статистики общаются с заказчиками, менеджерами или врачами, язык долей (числа вроде 0,95 или 0,01) часто заменяют на более привычный язык процентов. Так мы переходим к процентилям (перцентилям).

Процентиль — это тот же самый квантиль, но выраженный не в долях единицы, а в процентах (от 1-го до 99-го процентиля): k-й процентиль (обозначается P_k) — это значение, ниже которого падает k% всех наблюдений. Так,

- 90-й процентиль (P_{90}) — это барьер, который делит данные в отношении 90% (снизу) к 10% (сверху). Если ваш результат теста находится на 90-м процентиле, это значит, что вы написали тест лучше, чем 90% участников, и лишь 10% справились лучше вас.
- 50-й процентиль (P_{50}) — это не что иное, как медиана.
- 25-й и 75-й процентили (P_{25} и P_{75}) — это первый и третий квартили соответственно.

В педиатрии процентили — главный инструмент оценки развития. Врач не просто измеряет рост ребенка, он смотрит в процентильную таблицу. Если пятилетний мальчик находится на уровне 5-го процентиля по росту, это серьезный сигнал для эндокринолога: лишь 5% здоровых детей в популяции имеют такой рост или ниже, остальные 95% — выше. Если же он на уровне 50-го процентиля — он идеальный «среднестатистический» ребенок.

Как рассчитать точное значение квантиля для любого нормального распределения, если мы знаем только его среднее (μ) и стандартное отклонение (σ)? На помощь приходит метод стандартизации.

Вспомним формулу Z-оценки, которая превращает любое нормальное распределение в стандартизованное (где центр = 0, а шаг = 1):

$$Z = \frac{X - \mu}{\sigma}$$

Если мы перепишем эту формулу, чтобы выразить реальное значение X, мы получим:

$$X = \mu + Z \times \sigma$$

Это фундаментальное уравнение! Оно означает, что для поиска любого квантиля нам достаточно знать, сколько шагов-сигм (Z) вправо или влево нужно отступить от центра (μ).

Давайте запомним несколько критически важных, классических квантилей стандартизованного нормального распределения, которые в статистике называют *Z*-критическими значениями (z_p):

1. Для 95-го процентиля (P_{95} или квантиль 0,95): нам нужно найти такую точку *Z*, чтобы слева осталось 95% площади, а в правом хвосте — одинокие 5%. Если заглянуть в статистические таблицы или использовать Excel-функцию НОРМ.СТ.ОБР(0,95), мы получим $Z \approx 1{,}645$. Формула: $X_{0,95} = \mu + 1{,}645\sigma$.

2. Для 97,5-го процентиля ($P_{97,5}$ или квантиль 0,975) эта точка оставляет в правом хвосте всего 2,5% площади. Значение *Z* для нее равно 1,96. Запомните это число, оно — «священный Грааль» построения доверительных интервалов и теории проверки гипотез, с которыми мы познакомимся позже. Почему оно так важно? Потому что, если мы возьмем симметричный интервал от –1,96 до +1,96, внутри него поместится ровно 95% всей площади нормального колокола (отсекая по 2,5% с каждого хвоста). Формула: $X_{0,975} = \mu + 1{,}96\sigma$.

3. Для 99-го процентиля (P_{99}) $Z \approx 2{,}33$. Только 1% экстремально больших значений лежит вне этой отметки.

Чтобы закрепить теорию, решим практическую задачу из области управления качеством и логистики.

Условие: производитель изготавливает автоматические мониторы артериального давления, и среди потребителей оказалось большое количество людей, скрупулёзно и часто проверяющих точность оборудования. Срок точной работы пневмодатчика в мониторах распределен нормально со средним значением $\mu = 4$ года и стандартным отклонением $\sigma = 0{,}6$ года. Производителю очень хотелось бы, чтобы в маркетинговых целях гарантийный срок был не 1 год, а побольше. Какой срок можно установить, чтобы делать бесплатное гарантийное обслуживание не более чем для 2% проданных мониторов?

Решение: производителю нужно найти такое пороговое значение времени *X* (гарантийный срок), чтобы левее него (то есть датчик сломался раньше гарантии) оказалось лишь 2% распределения. Для этого он ищет второй процентиль ($P_{0,02}$) имеющегося распределения: сначала определяет *Z*-оценку для квантиля уровня 0,02. Поскольку эта точка лежит в левом хвосте (меньше 50%), *Z*-оценка обязательно будет отрицательной. С помощью таблиц или Excel (=НОРМ.СТ.ОБР(0,02)) можно обнаружить,

что $Z \approx -2{,}05$. Это значит, что 2% самых плохих датчиков «умирают» раньше, чем средний срок минус 2,05 стандартных отклонений. Переведем стандартизованную Z-оценку в реальные годы по нашей формуле: $X = \mu + Z \times \sigma$, то есть $X = 4 + (-2.05) \times 0.6 = 4 - 1{,}23 = 2{,}77$ года.

Таким образом, чтобы минимизировать издержки на гарантийный ремонт до желаемых 2%, производитель должен установить гарантию сроком примерно 2,7 года (или округлить в меньшую сторону — например, до 2,5 лет, обеспечив себе дополнительный запас прочности).

4.6. Тестирование распределения на нормальность

Когда аналитик или исследователь планирует применить к собранным данным такие мощные инструменты, как *t*-критерий Стьюдента (для сравнения средних двух групп), дисперсионный анализ (ANOVA), или линейный регрессионный анализ, он обязан сначала проверить выполнение ключевого допущения — распределены ли исходные данные нормально? Если это допущение нарушено, результаты вышеперечисленных тестов (так называемых параметрических методов) могут оказаться ложными и привести к неверным научным выводам. В случае серьезных отклонений от нормальности нам придется использовать другие, более осторожные методы — непараметрические (например, критерий Манна-Уитни или Краскела-Уоллиса), которые не требуют конкретной формы распределения, но обладают чуть меньшей статистической мощностью.

Если распределение не является идеальным нормальным колоколом, оно может деформироваться в двух направлениях: искривляться вбок или сжиматься/растягиваться по вертикали. Эти искажения математически измеряются третьим и четвертым моментами распределения — асимметрией и эксцессом.

Асимметрия (Skewness) измеряет степень скошенности (несимметричности) распределения относительно его центра. В идеальном нормальном распределении левая половина зеркально повторяет правую, и коэффициент асимметрии равен строго нулю (Skewness = 0). В этом случае среднее значение, медиана и мода совпадают.

Если данные скошены, мы получаем один из двух вариантов (рис. 13):

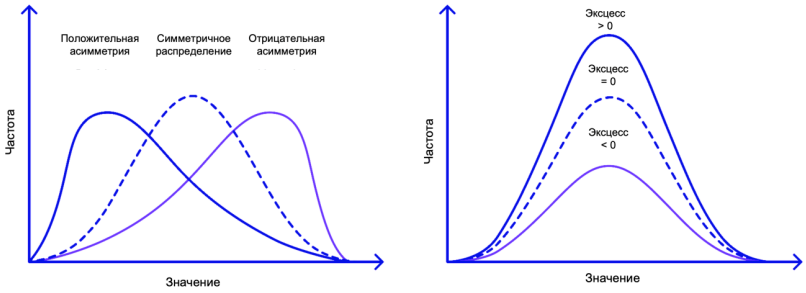


Рис. 13. Асимметрия и эксцесс

1. Положительная (правосторонняя) асимметрия (Skewness > 0): длинный «хвост» распределения вытягивается в правую (положительную) сторону. Это означает, что в выборке присутствует небольшое количество аномально больших значений. Классический, уже известный нам пример — доходы населения: большинство зарабатывает небольшую сумму, но несколько миллионеров утягивают «хвост» графика и среднее арифметическое далеко вправо.

2. Отрицательная (левосторонняя) асимметрия (Skewness < 0): длинный «хвост» графика уходит влево, в сторону меньших значений. Это говорит о наличии редких, аномально низких показателей. Пример — возраст людей при проявлении возрастных заболеваний (большинство заболевает в старости, но есть редкие ранние случаи).

Эксцесс (Kurtosis) измеряет «островершинность» распределения или, говоря более современным языком Data Science, «тяжесть хвостов» (heavy tails). Он показывает, насколько часто в распределении встречаются экстремальные выбросы по сравнению с нормальным колоколом. В статистических программах (Excel, R, Python) обычно рассчитывается так называемый избыточный эксцесс (Excess Kurtosis), из которого вычтено число 3. Это сделано для удобства: у идеального нормального распределения избыточный эксцесс равен строго нулю (Kurtosis = 0).

Типы распределений по уровню эксцесса (рисунок 13):

1. Мезокуртическое (Mesokurtic): Kurtosis = 0. Это эталонное нормальное распределение.
2. Лептокуртическое (Leptokurtic): Kurtosis > 0. График имеет острый пик в центре и очень «тяжелые», длинные хвосты.

Это значит, что данные «кучкуются» около центра, но при этом в выборке аномально много экстремальных выбросов.

3. Платикуртическое (Platykurtic): Kurtosis < 0. График приплюснутый, с плоской вершиной и легкими, короткими хвостами. Данные распределены более равномерно, а экстремальные выбросы почти отсутствуют.

Первое правило проверки: если при расчете ваших данных в программе коэффициенты асимметрии и избыточного эксцесса находятся в интервале от –1 до +1, распределение можно условно считать близким к нормальному. Если значения выходят за рамки ±2, отклонение от нормальности считается существенным.

Но никогда не доверяйте только сухим цифрам. Прежде чем делать выводы, посмотрите на свои данные глазами. Для этого существуют пара классических графиков.

1. ***Гистограмма с наложением кривой нормальности***. Вы строите стандартную гистограмму частот и просите программу наложить сверху идеальный нормальный колокол, рассчитанный по среднему и сигме вашей выборки. Вы сразу увидите, есть ли явная скошенность или провалы в данных.

2. ***График Квантиль-Квантиль*** (*Q-Q* Plot / Quantile-Quantile Plot) — это самый профессиональный и чуткий график для проверки нормальности. Принцип построения: Программа берет ваши реальные данные, сортирует их по возрастанию и рассчитывает их реальные квантили (ось *Y*). Затем она моделирует математически идеальные квантили, которые должны были бы быть, если бы выборка была строго нормальной (ось *X*). Каждое наблюдение наносится в виде точки на эту координатную сетку. На графике также проводится прямая диагональная линия под углом 45 градусов. При идеальной нормальности все точки вашей выборки лежат строго на диагональной линии. Если есть выбросы и тяжелые хвосты, то точки на краях графика (слева снизу и справа сверху) резко загибаются, отрываясь от линии вверх или вниз («S-образный» изгиб). При скошенности точки выгибаются дугой (аркой) выше или ниже опорной линии.

Когда графиков недостаточно и требуется строгое математическое подтверждение («доказано с уровнем значимости p»), применяют формальные статистические тесты. Во всех этих тестах формулируется стандартная пара гипотез:

1. Нулевая гипотеза (H_0): «данные распределены нормально».

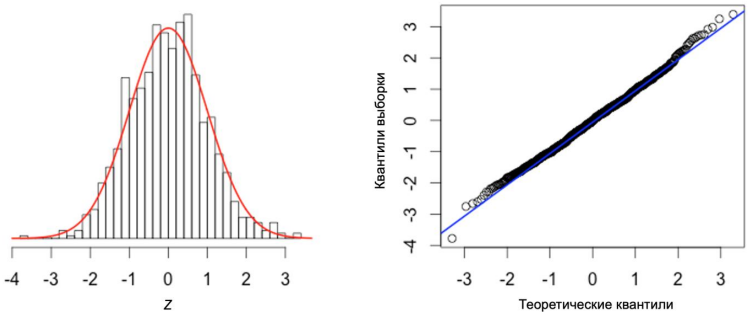


Рис. 14. Гистограмма с наложением кривой нормальности и график *Q-Q* (квантиль-квантиль)

2. Альтернативная гипотеза (H_1): «распределения данных отлично от нормального».

Обратите внимание! В отличие от большинства медицинских или биологических тестов, где ученый стремится отвергнуть H_0 и найти различия, в тестах на нормальность наша цель — НЕ отвергать H_0. Нам нужно получить значение p-value > 0,05. Только в этом случае мы можем сказать: «У нас нет оснований отвергнуть нормальность, данные чисты, мы можем применять t-критерий». Если $p < 0{,}05$, нормальность официально опровергнута.

Рассмотрим пару тестов, которые вы встретите в пакетах SPSS, R или Python:

1. ***Тест Шапиро-Уилка*** (Shapiro-Wilk Test): на сегодняшний день это самый мощный и авторитетный тест для выборок малого и среднего объема (примерно до 50–1000 наблюдений). Он детально анализирует структуру выборки на основе упорядоченных статистик. Если в данных есть даже небольшие отклонения, тест это «заметит».

2. ***Тест Колмогорова-Смирнова*** (с поправкой Лиллиефорса) — этот тест сравнивает реальную кумулятивную функцию распределения ваших данных с теоретической нормальной кривой и ищет точку их максимального расхождения. Классический тест Колмогорова-Смирнова требует, чтобы параметры μ и σ были известны заранее. Если мы рассчитываем их по самой выборке, тест становится слишком «слепым» и пропускает нарушения. Чтобы это исправить, используют поправку Лиллиефорса (Lilliefors Correction). Тест хорош для больших выборок ($n > 1000$).

4.7. Центральная предельная теорема

Представьте, что вы фанатичные исследователи физиологии аорты, и до вас дошла новость о том, что долгожданное оборудование экспертного класса, о котором вы мечтали, наконец-то доступно и ничто не мешает предаться интересному делу – новому и непревзойдённому по точности способу измерения скорости пульсовой волны в аорте. Дело настолько интересное, что вы готовы приступить к нему, не дожидаясь окончания летнего отпуска, и принимать пациентов несколько дней, чтобы узнать, каковы же показатели, которое даёт это оборудование. В эти дни удалось исследовать показатели у 20 призывников, которые пришли все вместе, у 6 абитуриентов, и у 4 очень пожилых пенсионеров, которые уже не выбираются на дачу, и их привезли на такси. Если мы построим график распределения полученных показателей, то увидим абсолютно хаотичную картину. Гистограмма дико скошена, асимметрична и бесконечно далека от красоты гладкого нормального распределения. А ведь многие методы классической статистики требуют, чтобы данные были распределены нормально! Как же теперь писать диссертацию? Опускать руки? К счастью, нет.

В игру вступает центральная предельная теорема (ЦПТ) — одно из достижений человеческой мысли. Она утверждает: даже если исходный мир полон хаоса и асимметрии, в нем существует скрытый порядок, который неизбежно рождает нормальное распределение. Чтобы понять ЦПТ, давайте проведем мысленный эксперимент, но будем представлять нашу деятельность в сильно ускоренном, как в кино, темпе, когда за минуту проносятся месяцы работы. Кадры исследований сменяются кадрами, в которых регулярно строятся гистограммы. К 100-му пациенту вы начинаете чувствовать, что ситуация как-то налаживается, но из интереса продолжаете, и вот уже диаграмма построена из 300 наблюдений, а затем и из 1000.

Что мы увидим, когда закончится период дач и отпусков, и исследование пройдёт большое количество всевозможных разнообразных пациентов? Мы обнаружим чудо: рваная, скошенная форма исходного графика полностью исчезла. Перед нами предстанет симметричный колокол нормального распределения.

Центральная предельная теорема утверждает, что распределение выборочных значений стремится к нормальному по мере того, как увеличивается объем выборки (n), независимо от того,

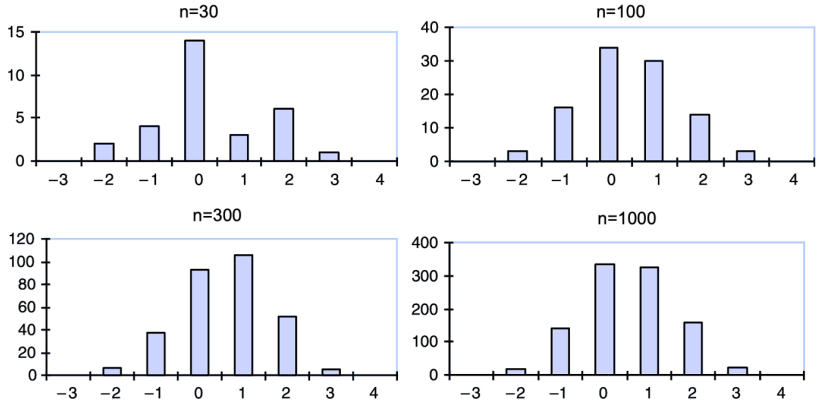


Рис. 15. Гистограммы при увеличении выборки

какова была форма распределения исходной совокупности (рисунок 15).

По ***мере увеличения количества наблюдений*** параметры нового «колокола» ведут себя по следующим правилам:

1. ***Центр колокола***: среднее значение всех возможных выборочных средних будет приближаться к истинному среднему значению генеральной совокупности (μ).

2. ***Разброс колокола*** (стандартная ошибка): новый колокол из средних значений всегда будет значительно уже, чем исходное распределение. Это логично: скорость пульсовой волны может быть экстремально большой (у очень пожилых) или маленькой (у молодых), но среднее значение по 30 или 50 разным пациентам неизбежно сглаживает крайности. Стандартное отклонение распределения средних называют стандартной ошибкой среднего (Standard Error of the Mean — SEM) и рассчитывают по формуле (немного позже мы к ней вернёмся):

$$SEM = \frac{\sigma}{\sqrt{n}}$$

Чем больше объем выборки (n), тем сильнее «сжимается» колокол средних и тем точнее мы знаем истину.

3. ***Форма колокола***: приближение к идеальной нормальности улучшается с ростом n.

Какой объем выборки (n) считается достаточным? В учебниках по статистике принято эмпирическое правило: Если исход-

ное распределение идеально симметрично (это бывает редко), уже при выборках размера $n > 5$ … 10 можно делать какие-то дальнейшие статистические исследования. А если исходное распределение «изломано», скошено или имеет выбросы, для гарантированной нормальности средних требуется выборка объемом $n \geq 30$ (иногда $n \geq 50$). Именно поэтому число «30» в прикладной статистике часто является магическим водоразделом между «малой» и «большой» выборками.

Значение этой теоремы невозможно переоценить. Она является фундаментом, на котором построена вся классическая статистика, так как освобождает от ограничений: исследователю в медицине, биологии или маркетинге не нужно молиться на то, чтобы природа распределила его показатели идеально «по Гауссу». Благодаря ЦПТ мы можем работать со средними значениями любых процессов. ЦПТ – это основа для доверительных интервалов и тестов. Поскольку мы точно знаем, что выборочные средние распределены нормально, мы имеем право использовать универсальные Z-оценки и критические значения (вроде знаменитого шага в 1,96 σ). На этом строятся t-критерий Стьюдента, дисперсионный анализ (ANOVA) и расчет доверительных интервалов.

Почему рост людей, вес плодов деревьев или ошибки прицеливания при стрельбе в тире распределены нормально? Потому что каждый из этих параметров — это результат суммирования огромного количества мелких, независимых случайных факторов (генетика, питание, экология, ветер, дрожание рук). Сумма множества случайных величин, согласно ЦПТ, рождает нормальный колокол.

4.8. Доверительные интервалы и стандартная ошибка средней

До этого момента мы использовали так называемые точечные оценки. Мы брали одно число (например, средний рост 100 случайно протестированных студентов, равный 175 см) и заявляли: «По-видимому, средний рост всех студентов университета тоже равен 175 см».

Однако в глубине души каждый исследователь понимает: вероятность того, что среднее значение в генеральной совокупности (μ) абсолютно точно совпадет с выборочным средним ($\bar{x}$) до

десятого знака после запятой, практически равна нулю. Стоит нам изучить другую сотню студентов, и мы получим среднее 174,3 см или 176,1 см. Из-за случайной изменчивости (ошибки выборки) точечная оценка всегда несет в себе долю ложной уверенности.

Вместо того чтобы давать одно неточное число, гораздо честнее и профессиональнее указать интервал (диапазон) значений и сопроводить его определенной долей уверенности. Например: «Мы не знаем точный средний рост всех студентов, но мы уверены, что он находится в пределах от 173 до 177 см».

Такие диапазоны в статистике называют доверительными интервалами.

Давайте дадим определение: ***доверительный интервал*** (CI) — это математически рассчитанный диапазон значений, который с заданной высокой вероятностью (называемой уровнем доверия) накрывает истинное, но неизвестное значение параметра генеральной совокупности (например, математическое ожидание μ).

Любой доверительный интервал строится вокруг нашей точечной оценки по единому шаблону: доверительный интервал = точечная оценка ± предельная ошибка выборки (ME, Margin of Error):

$$CI = \bar{x} \pm ME$$

Величина ME (предельная ошибка) определяет ширину этого интервала. Она зависит от трех факторов:

1. Разброса данных. Чем сильнее хаос внутри выборки (больше стандартное отклонение s), тем шире и неопределеннее будет наш интервал.

2. Объема выборки (n). Чем больше людей мы опросили, тем выше точность нашей оценки. Большая выборка сжимает интервал, делая его более узким.

3. Желаемого уровня доверия. Чем сильнее мы хотим быть застрахованы от ошибки, тем шире нам придется «распахнуть» границы интервала.

Прежде чем перейти к расчетам, разберемся с терминами «уверенности».

Уровень доверия (1-α) — это вероятность того, что наш интервал окажется «удачным» и действительно накроет истинный параметр популяции. В науке принято выбирать стандарты: 95% (наиболее частый), 99% или 90%.

Уровень значимости (α, альфа) 3 это вероятность нашей ошибки, то есть того, что истинный параметр останется за бортом интервала. Если уровень доверия равен 95%, то $\alpha = 1 - 0{,}95 = 0{,}05$ (или 5%).

Представьте себе рыбака, который забрасывает сеть в мутную воду, пытаясь поймать одну конкретную золотую рыбку (μ).

- Рыбак не видит рыбку (параметр скрыт).
- Размер сети — это ширина доверительного интервала.
- Если рыбак использует 95%-ю сеть, это значит, что при 100 забросах в аналогичных условиях сеть в 95 случаях накроет рыбку, а в 5 — промахнется.

Как рассчитать границы интервала на практике? Здесь сценарий разделяется на два пути в зависимости от того, знаем ли мы точный разброс в популяции или вынуждены оценивать его по выборке.

Сценарий А: стандартное отклонение популяции (σ) известно. В реальной жизни это редкость (например, если мы годами изучаем станок и точно знаем его погрешность σ). В этом случае мы опираемся на свойства стандартизированного нормального распределения и Z-оценки. Формула выглядит так:

$$CI = \bar{x} \pm Z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Величина называется ***стандартной ошибкой среднего*** (Standard Error of the Mean — SEM). Она показывает, насколько выборочные средние в принципе склонны гулять вокруг центра под действием «центральной предельной теоремы».

Значение $Z_{1-\alpha/2}$ — это критическое значение Z-оценки, отсекающее хвосты ошибки.

Для 95%-го уровня доверия ($\alpha = 0{,}05$, по 2,5% на каждый хвост) «магическое» число Z всегда строго равно 1,96, а для 99%-го уровня доверия число $Z = 2{,}58$.

Сценарий Б: Популяционная σ неизвестна, как в реальном мире. В 99% исследований мы не знаем истинную σ. У нас есть только выборочное стандартное отклонение s. Из-за этого в наши расчеты проникает дополнительная неопределенность.

Чтобы компенсировать её, в 1908 году Уильям Госсет (писавший под псевдонимом Стьюдент) разработал t-распределение Стьюдента. Оно похоже на нормальное, но имеет более «тяже-

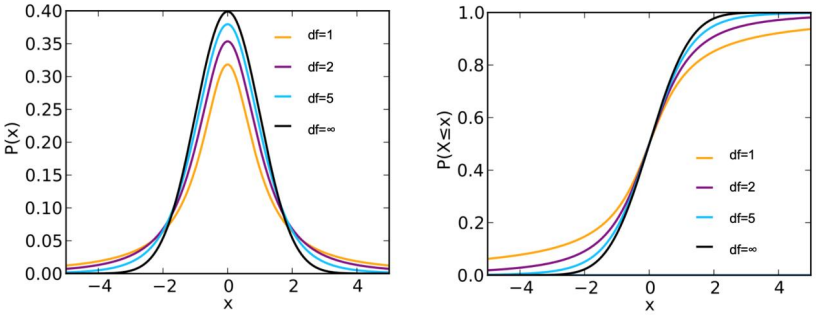


Рис. 15. Распределения Стьюдента при различных степенях свободы (df=∞ — это нормальное распределение).

лые» хвосты, которые защищают исследователя от ложных выводов на малых выборках.

Формула трансформируется:

$$CI = \bar{x} \pm t_{\frac{\alpha}{2},df} \times \frac{s}{\sqrt{n}}$$

Где df = n – 1 — число степеней свободы. По мере роста объема выборки t-распределение плавно превращается в идеальное нормальное Z-распределение, и критические коэффициенты сближаются.

Давайте попрактикуемся. Представим, что фармацевтическая компания тестирует новый препарат для снижения артериального давления. У выборки из n = 25 пациентов среднее снижение давления составило $\bar{x}$ = 12 мм рт. ст., с выборочным стандартным отклонением s = 5 мм рт. ст. Требуется построить 95%-й доверительный интервал для истинного среднего эффекта препарата.

Решение:

1. Выборка малая (n = 25), истинная σ неизвестна. Используем t-распределение Стьюдента.
2. Число степеней свободы: df = 25 – 1 = 24.
3. По таблицам Стьюдента (или через Excel =СТЬЮДЕНТ.ОБР.2Х(0,05; 24)) находим критическое значение t для 95% доверия: $t \approx 2{,}064$. (Обратите внимание, оно чуть больше, чем Z-стандарт 1,96, интервал будет шире ради безопасности).

4. Рассчитаем стандартную ошибку:

$$SEM = \frac{s}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1 \text{ мм рт. ст.}$$

5. Находим предельную ошибку:

ME = 2,064×1 = 2,064 мм рт. ст.

6. Вычисляем границы интервала:
Нижняя граница = 12 – 2,064 = 9.936
Верхняя граница = 12 + 2.064 = 14.064

Результат: 95%CI = [9.94; 14.06] мм рт. ст.

Как НЕ надо понимать CI? Огромное число студентов, аспирантов и даже практикующих врачей допускают грубейшую ошибку при чтении результатов. Получив интервал [9.94; 14.06], они говорят: «Вероятность того, что истинное среднее μ лежит внутри этого интервала, равна 95%».

Это неверно! В классической (частотной) статистике истинное среднее μ — это фиксированная, константная величина. Она не является случайной переменной. Она либо лежит внутри этого интервала, либо нет. Вероятность здесь может быть либо 0 (не лежит), либо 1 (лежит). Она не может быть равна 0,95 для уже построенного интервала.

Случайной величиной является сам наш интервал! Его границы (9,94 и 14,06) — это функции от случайной выборки. Если мы завтра соберем новую выборку, интервал сдвинется. Поэтому корректная формулировка звучит так:

«Если мы проведем это исследование 100 раз и каждый раз по аналогичной методике построим доверительный интервал, то в 95 случаях из 100 наши интервалы накроют истинное среднее популяции μ, а в 5 случаях промахнутся. Мы заявляем, что наш конкретный интервал [9,94; 14,06] получен в результате такого процесса, надежность которого составляет 95%».

Доверительные интервалы — это не просто способ описания точности. Это полноценный инструмент проверки статистических гипотез. Представьте, что конкуренты утверждают: «Ваш препарат пустышка, его реальный эффект равен нулю ($\mu_0 = 0$)». Вам не нужно проводить дополнительные тесты. Достаточно посмотреть на ваш 95%-й доверительный интервал: [9.94; 14.06].

Поскольку число «0» абсолютно не попадает в этот диапазон (оно лежит далеко слева), вы можете мгновенно заявить: «Мы отвергаем нулевую гипотезу об отсутствии эффекта на уровне значимости $p < 0,05$, так как ноль не входит в наш доверительный диапазон».

4.9. Логнормальное распределение

Мы очень много «хвалили» нормальное распределение. Мы говорили, что оно симметрично, красиво и описывает массу явлений в природе. Напомню, что нормальное распределение возникает тогда, когда на итоговую величину влияет огромное множество мелких, независимых факторов, которые складываются (аддитивный процесс). Однако реальный мир не всегда устроен аддитивно. Представьте себе следующие показатели:

Доходы или богатство людей в стране;

Размер частиц золота в жиле руды или крупинок песка на пляже;

Длительность интернет-сессий пользователей;

Концентрация загрязняющих веществ или лекарств в крови.

Если мы построим графики этих величин, мы никогда не увидим симметричный колокол. Мы увидим распределение с резким пиком слева, около нуля, и бесконечно длинным, плавно затухающим «хвостом» справа. Величины этих показателей иногда принимают колоссально большие значения.

Почему так происходит? Потому что в этих процессах случайные факторы не складываются, а перемножаются (мультипликативный процесс). И главным математическим инструментом для описания таких явлений служит логнормальное распределение.

Давайте дадим логнормальному распределению максимально простое и красивое определение через его связь с уже знакомым нам распределением Гаусса: «случайная величина X имеет логнормальное распределение, если натуральный логарифм этой величины ($Y = lnX$) распределен нормально». Логика работает и в обратную сторону: «если у вас есть нормально распределенная величина Y, то её экспонента $X = e^Y$ будет иметь логнормальное распределение.

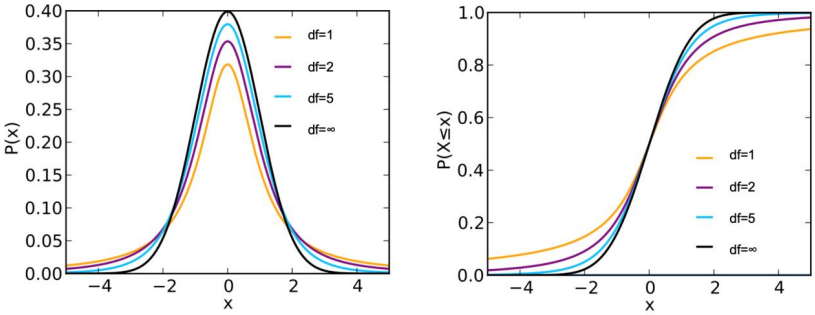


Рис. 16. Функции нормального и логнормального распределения

Логнормальное распределение, как и нормальное, полностью задается двумя параметрами. Но здесь скрыта важная ловушка для новичков! Эти параметры обозначаются как μ (мю) и σ (сигма), но они являются средним значением и стандартным отклонением не самой величины X, а её логарифма (lnX).

μ (масштабный параметр) — среднее арифметическое логарифмов данных.

σ (параметр формы) — стандартное отклонение логарифмов данных.

Поскольку $X = e^Y$, а экспонента от любого числа всегда строго больше нуля, логнормальная случайная величина никогда не может принимать отрицательные значения или быть равной нулю. Её область определения: $X \in (0; +\infty)$. Это свойство делает логнормальную модель незаменимой для описания цен, физических размеров и времени, которые физически не могут быть меньше нуля.

Из-за сильной правосторонней асимметрии логнормального распределения три главные характеристики центра — мода, медиана и среднее арифметическое — больше не совпадают в одной точке, как это было в нормальном колоколе. Они строго разделяются и всегда следуют друг за другом в определенном порядке.

1. Мода (самая высокая точка графика): это наиболее часто встречающееся значение. Оно находится ближе всего к нулю.

2. Медиана (точка, делящая площадь графика ровно пополам): 50% наблюдений лежат левее медианы, 50% — правее. Медиана логнормального распределения рассчитывается очень легко: это просто экспонента от среднего логарифмов: Медиана = e^{μ}.

3. Среднее арифметическое (математическое ожидание): длинный правый хвост из редких, но гигантских значений (например, миллионеров в анализе доходов) работает как мощный рычаг и утягивает среднее значение далеко вправо. Среднее всегда больше медианы и моды.

При изменении параметра σ (разброса логарифмов) меняется сама форма графика. Если σ очень мала (например, 0.1), логнормальное распределение визуально становится почти симметричным и очень похожим на нормальное. Но чем больше растет σ, тем сильнее график прижимается к нулю слева и тем более экстремальным, и длинным становится его правый хвост.

Почему логнормальность возникает в жизни? Давайте поймем физику процесса. Центральная предельная теорема (ЦПТ) гласит: если процесс формируется суммой мелких факторов, итогом будет нормальное распределение. Но что, если мы применим ЦПТ к логарифмам? Из школьной математики мы помним, что сумма логарифмов равна логарифму произведения:

$$ln(x_1) + ln(x_2) + \ldots + ln(x_n) = ln(x_1\, x_2\, \ldots\, x_n)$$

Это значит, что если факторы внутри процесса перемножаются, то логарифм итоговой величины будет являться суммой, а значит, распределится нормально. Следовательно, сама итоговая величина распределится логнормально.

Логнормальное распределение — один из главных инструментов инженеров, занимающихся расчетом надежности техники. Им часто описывают «время до отказа» (срок службы) различных узлов и механизмов. Например, износ металлических деталей под воздействием усталости или трения. Сначала микротрещина растет очень медленно, но каждая новая стадия деструкции ускоряет процесс, «размножая» дефекты мультипликативно. В итоге большая часть деталей ломается примерно в один гарантированный период (пик распределения), но отдельные сверхотработанные экземпляры благодаря уникальному стечению обстоятельств живут аномально долго, формируя тот самый длинный правый хвост логнормальной кривой.

4.10. Нормальное распределение: краткие итоги

- Нормальное распределение описывается двумя параметрами: центром μ и разбросом σ.
- Вероятности рассчитываются через площади под кривой с помощью функции распределения $F(x)$.
- У непрерывной случайной величины вероятность встретить точное конкретное значение равна нулю. Поэтому мы всегда работаем с вероятностью попадания в интервал.
- Функция распределения $F(x)$ отвечает на вопрос: «Какова вероятность, что наша величина окажется меньше или равна x?». Она всегда плавно растет от 0 до 1. Чтобы найти вероятность интервала, нужно просто вычесть $F(a)$ из $F(b)$.
- Функция плотности $f(x)$ показывает форму распределения, демонстрируя, где данные «скучкованы» сильнее всего. Вероятность интервала для плотности — это площадь под графиком на этом отрезке.
- Плотность — это производная от функции распределения, а функция распределения — это интеграл от функции плотности.
- Благодаря центральной предельной теореме, выборочные средние значения почти всегда распределены нормально, что открывает двери для проведения точных статистических тестов и построения доверительных интервалов.
- Z-оценка — это инструмент, переводящий любые сырые данные в количество стандартных отклонений от среднего значения. Стандартизация превращает любое нормальное распределение в эталонное распределение со средним $\mu = 0$ и стандартным отклонением $\sigma = 1$.
- Положительная Z-оценка говорит о том, что значение находится выше среднего, отрицательная — ниже среднего. Z-оценка, равная нулю, в точности соответствует среднему значению группы.
- Благодаря Z-оценкам мы можем объективно сопоставлять между собой параметры, измеренные в совершенно разных шкалах и единицах.
- Квантили и процентили позволяют переходить от заданной вероятности (доли данных) к конкретным физическим границам признака на оси X.

- Процентили — это просто квантили, переведенные в процентную шкалу. *k*-й процентиль оставляет позади себя k% распределения.
- Благодаря симметрии нормального распределения, его среднее значение μ одновременно является медианой, 5-м децилем и 50-м процентилем.
- Проверка на нормальность — обязательный «зеленый коридор» перед использованием параметрической статистики.
- Лучший инструмент визуального контроля нормальности — график *Q*-*Q* Plot. Данные нормальны, если точки лежат строго на диагонали.
- Главный тест на нормальность для клинициста и исследователя — критерий Шапиро-Уилка. Нам нужно получить $p > 0{,}05$, чтобы подтвердить нормальность.
- Если нормальность грубо нарушена, у нас есть три пути: трансформировать данные (например, логарифмировать их), удалить легитимные выбросы или перейти к непараметрическим критериям.
- Точечная оценка $\bar{x}$ указывает лишь примерное направление. Доверительный интервал очерчивает зону обоснованной уверенности.
- Если вы работаете с реальными выборками и не знаете популяционный разброс, вашим проводником становится *t*-критерий Стьюдента, а не *Z*-критерий.
- 95% доверия — это характеристика метода генерации интервалов, а не вероятность нахождения фиксированной константы внутри конкретных двух чисел.
- Логнормальное распределение описывает положительные, асимметричные случайные величины с длинным правым хвостом. Если из каждого значения логнормального массива данных извлечь натуральный логарифм, мы получим идеальное нормальное распределение.
- Параметры логнормального распределения μ и σ — это среднее и стандартное отклонение логарифмов этих данных, а не самих сырых чисел.

Глава 5. Анализ качественных данных

5.1. Биномиальное распределение

В повседневной жизни и в научно-практических задачах мы постоянно сталкиваемся с процессами, которые имеют всего два возможных исхода:

- Монета падает либо орлом, либо решкой.
- Новорожденный ребенок может быть либо мальчиком, либо девочкой.
- Деталь, сошедшая с конвейера, оказывается либо качественной, либо бракованной.
- Пациент после приема лекарства либо выздоравливает, либо нет.

В статистике такие испытания с двумя исходами называются испытаниями Бернулли (в честь швейцарского математика Якоба Бернулли). Один из исходов принято условно называть «успехом» (то событие, которое нас интересует), а второй «неудачей».

Если мы проведем не одно такое испытание, а целую серию (например, проверим 10 деталей или опросим 50 покупателей), нам захочется узнать: «Какова вероятность того, что среди них будет ровно X успехов?». Закономерность, которая управляет этим процессом, называется биномиальным распределением.

Не каждый процесс с двумя исходами можно описать биномиальной моделью. Чтобы распределение считалось биномиальным, должны строго выполняться четыре условия:

1. Фиксированное число испытаний (n): мы точно знаем, сколько раз повторяется эксперимент (например, мы подбрасываем монету строго $n = 10$ раз).

2. Всего два исхода: в каждом отдельно взятом испытании возможен только «успех» или «неудача».

3. Постоянная вероятность успеха (p): вероятность получить «успех» во 2-м, 5-м или 10-м испытании остается абсолютно одинаковой. Она не меняется от шага к шагу. Вероятность неудачи обозначается буквой q и равна $1 – p$.

4. Независимость испытаний: результат одного испытания никак не влияет на результаты других. Если первая монета упала орлом, это не увеличивает и не уменьшает шансы второй монеты упасть орлом.

Пример нарушения условий: если мы вытаскиваем карты из колоды одна за другой без возвращения, вероятность вытащить

туза меняется с каждым шагом. Такое распределение не будет биномиальным (оно называется гипергеометрическим). Но если мы возвращаем карту обратно и перемешиваем колоду — условия биномиальности соблюдены.

Давайте поговорим о формуле биномиального распределения. Итак, пусть мы проводим n независимых испытаний, в каждом из которых вероятность успеха равна p. Мы хотим найти вероятность того, что количество успехов будет равно конкретному числу k (где k может быть от 0 до n).

Формула биномиального распределения выглядит следующим образом:

$$P(X = k) = C_n^k \times p^k \times q^{n-k}$$

Давайте разберем эту формулу по винтикам, в ней скрыта железная логика:

1. p^k — это вероятность того, что мы получили ровно k успехов подряд.

2. q^{n-k} — это вероятность того, что оставшиеся $(n - k)$ испытаний завершились неудачами.

3. C_n^k (Число сочетаний из n по k) — это биномиальный коэффициент. Зачем он нужен? Дело в том, что успехи и неудачи могут перемешиваться в разном порядке. Например, если мы бросаем монету 3 раза и хотим получить 2 орла, это может случиться в разной последовательности: (Орел, Орел, Решка), (Орел, Решка, Орел) или (Решка, Орел, Орел). Коэффициент C_n^k как раз рассчитывает количество таких возможных комбинаций:

$$C_n^k = \frac{n!}{k!\,(n-k)!}$$

Биномиальное распределение является дискретным. Его график представляет собой не плавную линию, а набор столбиков (гистограмму), где высота каждого столбика соответствует вероятности получить ровно k успехов. Форма этого распределения сильно зависит от вероятности p:

Если $p = 0{,}5$ (как при подбрасывании идеальной монеты), распределение будет абсолютно симметричным. Самые высокие столбики будут в центре, а к краям (0 успехов или n успехов) вероятность будет плавно падать.

Если $p < 0{,}5$ (успех случается редко, как например, при подбрасывании игрального кубика), распределение будет скошено вправо. Пик будет смещен к левой стороне (рисунок 17).

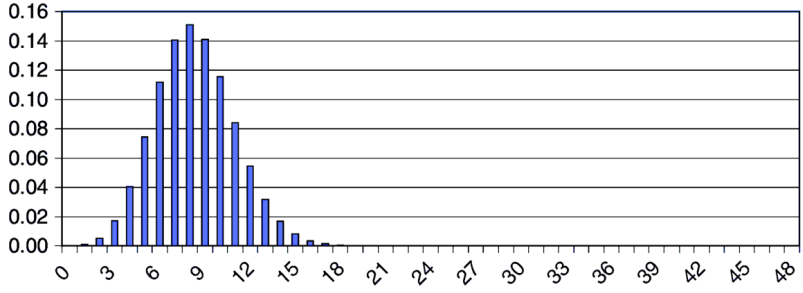


Рис. 17. Пример биноминального распределения с $n = 48$ и $p = 1/6$

Если $p > 0{,}5$ (успех случается часто), распределение будет скошено влево, а пик сместится к максимуму.

Вам не нужно собирать выборки, чтобы узнать среднее значение или разброс биномиального процесса. Их можно мгновенно рассчитать теоретически: во-первых, математическое ожидание (среднее количество успехов, μ):

$$\mu = n\,p$$

Логика: если мы знаем, что всхожесть семян составляет 80% ($p = 0{,}8$), и мы посадили $n = 100$ семян, то в среднем мы ожидаем, что взойдет $100\ 0{,}8 = 80$ штук.

Во-вторых, дисперсия (σ^2) и стандартное отклонение (σ):

$$\sigma^2 = n \times p \times q$$

$$\sigma = \sqrt{n \times p \times q}$$

Давайте попрактикуемся, порешаем задачи.

Условие: не имеющий аналогов в мире производитель изготавливает «сфигмографы» с пневматическими манжетами. Оборудование нельзя назвать экспертным, да и в среднем 5% манжет для этих «сфигмографов» выходят с дефектом: на них расходятся швы ($p = 0{,}05$). Комплект, формально требующийся для оснащения кабинета функциональной диагностики, состоит из аппарата с 4 манжетами. Какова вероятность того, больница получит «сфигмограф» с одной (не двумя!) бракованной манжетой?

Решение:

1. Проверим условия: количество испытаний фиксировано ($n = 4$), исходов два (брак/норма), детали независимы, вероятность брака неизменна ($p = 0{,}05$). Вероятность исправной детали $q = 1 - 0{,}05 = 0{,}95$.

2. Применим формулу для $k = 1$:

$$P(X = 1) = C_4^1 \times (0{,}05)^1 \times (0{,}95)^{4-1}$$

3. Посчитаем число сочетаний (сколькими способами среди 4 манжет может встретиться 1 бракованная):

$$C_4^1 = \frac{4!}{1!\,(4-1)!} = 4$$

4. Подставим числа в итоговый расчет:

$$SP(X = 1) = 4 \times 0{,}05 \times 0{,}95^3 = 0{,}1715$$

Ответ: вероятность получить не имеющий аналогов «сфигмограф» с одной бракованной манжетой составляет примерно 17,15%. Заметьте, если бы вопрос был сформулирован «Какова вероятность получить «сфигмограф» ***как минимум*** с одной бракованной манжетой?», то сочетаний, а следовательно, и конечной вероятности было бы больше.

Что будет, если число испытаний n станет очень большим? Представьте, что банк проверяет кредитные истории не 4 и не 20 клиентов, а $n = 1000$ человек. Считать число сочетаний C_{1000}^{250} по формуле вручную или даже на калькуляторе станет невозможно — факториалы превратятся в астрономические числа. Абрахам де Муавр и Пьер-Симон Лаплас доказали замечательную теорему: «если n достаточно велико, биномиальное распределение начинает стремиться к непрерывному нормальному распределению».

Этим приближением можно смело пользоваться, если выполняется условие: $n \times p \geq 5$ и $n \times q \geq 5$. Вместо сложных расчетов с факториалами аналитики просто берут среднее $\mu = np$, стандартное отклонение $\sigma = \sqrt{npq}$ и используют привычный нам «гауссовский колокол» для быстрого нахождения вероятностей.

5.2. Стандартная ошибка доли и доверительные интервалы

Ранее мы подробно разбирали, как измеряется неопределенность для числовых данных (роста, веса, прибыли). Там все казалось интуитивно понятным: у нас есть среднее значение, и вокруг него «гуляют» данные со своим стандартным отклонением. Но как быть, если наши данные неколичественные? В анализе качественных (категориальных) данных мы оперируем не числами, а фактами принадлежности к категории «успех» (или

«неудача»). В таких задачах главным итоговым показателем становится доля (процент) «успехов» в выборке, которую мы обозначаем как $\hat{p}$ (читается «п-хэт» или «п с крышкой»).

Если мы опросили 1000 человек, и 450 из них сказали, что поддерживают новый законопроект госдумы о запрете чтения иностранных медицинских журналов, то выборочная доля составит $\hat{p}$ = 450 / 1000 = 0,45 (или 45%). Здесь и возникает статистическая неопределенность. Означает ли это, что ровно 45% всего населения страны (генеральной совокупности) поддерживают этот закон? Конечно, нет. Если мы завтра случайным образом выберем другую тысячу человек, мы получим долю 43% или 47%. Из-за случайности формирования выборки (ошибки выборки) наша точечная оценка $\hat{p}$ неизбежно содержит в себе неопределенность. Сегодня мы научимся строго измерять и описывать эту неопределенность.

Чтобы измерить неопределенность, нам нужно понять, как вели бы себя выборочные доли, если бы мы повторяли опрос тысячи человек бесконечное количество раз. Представьте, что истинная доля сторонников закона в стране равна p (без крышки). Каждый отдельный опрос — это серия из n независимых испытаний Бернулли. Количество успехов в нем подчиняется биномиальному распределению. Но когда мы делим количество успехов на n, мы получаем распределение доли.

Если объем выборки n достаточно велик (эмпирически: $np > 5$ и $n(1 - p) > 5$), то распределение выборочных долей будет стремиться к нормальному распределению. Параметры этого распределения долей строго фиксированы:

1. Математическое ожидание: центр этого распределения совпадает с истинной долей в генеральной совокупности ($M(\hat{p}) = p$). В среднем наши опросы бьют точно в цель.

2. Стандартная ошибка доли (Standard Error, SE) – это и есть мера нашей статистической неопределенности. Она показывает, насколько сильно выборочные доли склонны хаотично колебаться вокруг истинного центра. Формула стандартной ошибки выглядит так:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Рассмотрим логику формулы. Влияние объема выборки (n): чем больше людей мы опросили, тем больше знаменатель, и тем меньше ошибка. Большая выборка эффективно подавляет нео-

пределенность. Влияние самой доли (p): максимальная неопределенность достигается при $p = 0,5$ (когда мнения разделились ровно 50/50). В этой точке числитель $0,5 \times 0,5 = 0,25$ максимален. Если же признак редкий (например, редкая болезнь, $p = 0,01$), неопределенность исходов становится меньше.

В реальном исследовании мы сталкиваемся с парадоксом: чтобы посчитать истинную стандартную ошибку SE, нам нужно знать истинную популяционную долю p. Но ведь именно её мы и пытаемся выяснить с помощью опроса! Выход из этого замкнутого круга прост: мы заменяем неизвестную константу p на её выборочную оценку $\hat{p}$, полученную из нашего единственного опроса. В таком случае мы получаем выборочную стандартную ошибку (SEM):

$$SEM = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Вернемся к нашему примеру: $n = 1000$ человек, доля поддержки $\hat{p} = 0,45$. Вероятность противоположного исхода $\hat{q} = 1 - 0,45 = 0,55$.

Рассчитаем меру неопределенности:

$$SEM = \sqrt{\frac{0,45 \times 0,55}{1000}} = \sqrt{\frac{0,2475}{1000}} \approx 0,0157 \text{ (или } 1,57\%)$$

Физический смысл этой меры: в среднем наши выборочные опросы такого объема ошибаются в ту или иную сторону примерно на 1,57%.

Но просто заявить об ошибке в 1,57% — мало. Нам нужно очертить диапазон и сопроводить его строгой гарантией (уровнем доверия). Для этого строится доверительный интервал для доли.

Поскольку распределение долей нормальное, мы можем использовать критические значения Z-оценки. Для самого популярного в науке 95%-го уровня доверия критическое значение Z равно 1,96 (оно означает, что в пределах ±1,96 стандартных ошибок от центра сосредоточено 95% всех возможных исходов).

Формула предельной погрешности опроса (Margin of Error — ME):

$$ME = Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1,96 \times SEM$$

Для нашего примера:

$$ME = 1{,}96\ 0{,}0157 = 0{,}0307 \text{ или } 3{,}07\%$$

Обратите внимание, «три процента статистической погрешности» — это часто можно увидеть в отчетах профессиональных социологических служб при выборке около 1000 человек!

Нижняя граница = $\hat{p}$ – ME = 0,45 – 0,0307 = 0,4193 (41,93%)

Верхняя граница = $\hat{p}$ + ME = 0,45 + 0,0307 = 0,4807 (48,07%)

Итоговый результат: 95%CI = [41,93%; 48,07%]. Корректная интерпретация:

Мы не имеем права утверждать, что «вероятность нахождения истинной доли внутри этих границ равна 95%». Истинная доля — фиксированная константа страны. Правильно говорить так: «Мы применили метод построения интервалов, который в 95% случаев аналогичных случайных опросов успешно перекрывает истинное значение генеральной совокупности. Основываясь на наших данных, мы заявляем, что истинная поддержка нового важного и меняющего жизнь многих учёных-медиков законопроекта госдумы с высокой долей уверенности лежит в диапазоне от 41,93% до 48,07%».

Важное ограничение! Все формулы, которые мы рассмотрели выше, предполагают, что наша генеральная совокупность бесконечна или огромна (например, все жители страны или все отчаянные пользователи интернета в ней). Но что делать, если мы исследуем качественные данные внутри закрытого, относительно небольшого сообщества? Например, мы спрашиваем о необходимости массажного кабинета непосредственно на рабочем месте у сотрудников огромного научного центра, где всего трудится $N = 500$ человек, а наша выборка составила $n = 150$ сотрудников. Если ваша выборка составляет существенную часть от генеральной совокупности (более 5–10%), неопределенность начинает стремительно падать. Ведь вы уже опросили значительную часть системы! В этом случае в формулу стандартной ошибки вводится поправка на конечную популяцию (Finite Population Correction — FPC):

$$SEM_{\text{конеч}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Множитель $\sqrt{\frac{N-n}{N-1}}$ всегда меньше единицы. Он выступает в роли «сжимателя» неопределенности, делая ваш доверительный интервал более узким и точным.

5.3. Репрезентативность выборки

Итак, мы научились рассчитывать статистическую неопределенность — мы строили доверительные интервалы для долей и вычисляли предельную погрешность. Однако у всех этих красивых математических формул есть одна скрытая, но жесткая оговорка: они работают только в том случае, если выборка является репрезентативной.

Что же это значит? ***Репрезентативность*** (представительность) выборки — это свойство выборки корректно и в правильных пропорциях воспроизводить ключевые характеристики той генеральной совокупности (популяции), из которой она была извлечена. Когда мы анализируем количественные данные (например, средний доход), нерепрезентативная выборка просто исказит среднее арифметическое. Но когда мы анализируем качественные данные (доли, проценты, мнения, категории), потеря репрезентативности полностью искажает структуру распределения ответов.

Представьте, что вы хотите узнать долю жителей города, поддерживающих строительство новой велодорожки (качественный признак: «да» или «нет»). Если вы выйдете на улицу и опросите 1000 случайных людей в парке в воскресенье утром, вы получите огромную выборку. Математическая погрешность для n=1000 будет очень маленькой (около 3%). Но будет ли этот результат отражать мнение всего города?

Нет, потому что в парке утром гуляет специфическая группа людей (спортсмены, молодые родители), чье качественное мнение сильно отличается от мнения автомобилистов или пенсионеров, оставшихся дома. Выборка может быть огромной, но абсолютно бесполезной, если она смещена.

В истории статистики есть хрестоматийный пример того, как погоня за объемом выборки при полном игнорировании репрезентативности привела к грандиозному провалу. В 1936 году в США проходили президентские выборы: Альфред Лэндон против Франклина Рузвельта. Популярный журнал «The Literary Digest» решил провести масштабный опрос общественного мнения (анализ качественных данных: «За кого вы проголосуете?»). Они разослали более 10 миллионов открыток-анкет и получили обратно колоссальное число ответов —2,3 миллиона человек!

На основе этого гигантского массива данных журнал уверенно спрогнозировал: Лэндон победит с огромным отрывом (57% против 43%).

Что произошло в реальности? Рузвельт разгромил Лэндона, набрав 62% голосов. Ошибка журнала составила безумные 19%. В то же время молодой ученый Джордж Гэллап опросил всего 50 тысяч человек (в 46 раз меньше!), но его выборка была строго репрезентативной. Гэллап предсказал победу Рузвельта с точностью до пары процентов.

Почему так вышло? Выборка журнала была глубоко смещенной по двум качественным критериям:

1. Смещение базы контактов: журнал брал адреса из телефонных справочников и списков владельцев автомобилей. В 1936 году (разгар Великой депрессии) телефоны и машины были только у зажиточных граждан, которые традиционно голосовали за республиканцев (Лэндона). Бедные слои населения, голосовавшие за Рузвельта, просто не попали в радар исследователей.

2. Смещение из-за добровольного ответа: из 10 миллионов анкет вернулись только 2,3 миллиона. Ответили те, у кого было много свободного времени или сильная эмоциональная позиция, что также нарушило репрезентативность.

Чтобы ваша выборка не повторила судьбу легендарного провала 1936 года, необходимо знать главных «врагов» репрезентативности — системные ошибки (смещения).

1. ***Ошибка охвата*** (Selection / Coverage Bias). Возникает тогда, когда определенные качественные группы генеральной совокупности физически не имеют шанса попасть в выборку. Пример: проведение интернет-опроса о качестве пенсионной реформы. Пожилые люди, не пользующиеся интернетом, выпадают из анализа. Доля недовольных в выборке окажется заниженной.

2. ***Ошибка самоотбора*** (Self-Selection / Volunteer Bias). Проявляется, когда люди сами решают, участвовать им в исследовании или нет. Пример: Опрос на сайте ресторана: «Оцените наше меню: вкусно / невкусно». В таком опросе чаще всего участвуют либо экстремально недовольные клиенты (чтобы выплеснуть гнев), либо восторженные фанаты. Мнение умеренной, спокойной и самой массовой категории клиентов останется неуслышанным.

3. ***Ошибка выжившего*** (Survivorship Bias). Явление, при котором анализируется только та группа («выжившие»), по которой

есть большой объем данных, а качественные характеристики «погибших» игнорируются. Пример: опрос успешных предпринимателей на тему «Каковы главные качества для старта бизнеса?». Мы опрашиваем только тех, кто удержался на плаву. Мы не знаем качественных характеристик 90% предпринимателей, чьи бизнесы разорились, хотя они могли обладать точно такими же качествами.

Как добиться репрезентативности? Единственный надежный способ гарантировать репрезентативность качественных данных — использовать вероятностные (случайные) методы формирования выборки. В этом случае каждый элемент генеральной совокупности имеет известную и ненулевую вероятность быть выбранным.

1. ***Простая случайная выборка***. Каждый объект извлекается наугад (как шары из лототрона). Если у нас есть полная база клиентов банка (N=100 000), генератор случайных чисел выбирает из нее n=1000 человек. Это идеальный метод, но он требует наличия исчерпывающей базы данных обо всех элементах.

2. ***Стратифицированная случайная выборка***. Этот метод — настоящее спасение для анализа качественных данных. Если мы знаем, что наша популяция делится на важные качественные слои (страты), мы обязаны воспроизвести эту структуру в выборке принудительно. Пример: мы исследуем отношение студентов вуза к онлайн-обучению. В вузе учится 70% девушек и 30% юношей. При простой случайной выборке из-за капризов судьбы к нам может попасть 90% девушек, и данные сместятся. При стратификации мы делим базу на два списка (юноши и девушки) и случайно извлекаем из первого списка строго 300 человек, а из второго — строго 700 человек. Структура выборки идеально повторяет структуру вуза.

3. ***Кластерная выборка*** применяется, когда генеральная совокупность географически или организационно разбросана. Популяция делится на естественные группы — кластеры (например, школы, районы, заводы). Сначала случайно выбираются несколько кластеров, а затем внутри этих кластеров проводится сплошной или случайный опрос.

Что делать, если случайная выборка невозможна? В коммерческих маркетинговых исследованиях или социологии собрать строго случайную выборку бывает слишком дорого или технически нереализуемо. Тогда аналитики используют не-

вероятностные методы, пытаясь искусственно сымитировать репрезентативность.

Главный инструмент здесь — ***Квотная выборка***. Интервьюерам выдаются жесткие квоты (задания). Например: «Вам нужно опросить на улицах города ровно 40 мужчин в возрасте 20–30 лет, 60 мужчин в возрасте 40–50 лет, 50 женщин в возрасте 20–30 лет...». Эти квоты рассчитываются на основе официальных данных переписи населения. Хотя этот метод уязвим для личных предпочтений интервьюера (он подсознательно выбирает людей с более дружелюбными лицами), квоты позволяют удержать качественную структуру выборки в рамках разумной репрезентативности.

5.4. Таблицы частот и критерий Хи-квадрат

Качественные данные описывают свойства, состояния или категории объектов. Их нельзя измерить линейкой, у них нет физического числового значения. Поскольку мы не можем сложить, например, свойства «мелкопузырчатые хрипы» и «налёт на языке» и посчитать среднее арифметическое, базовым инструментом для описания таких данных становятся таблицы частот.

Таблица частот — это таблица, которая показывает, как именно распределились объекты выборки по существующим качественным категориям. В ней указываются абсолютные частоты (прямой подсчет количества объектов) и/или относительные частоты (доли, проценты от общего числа).

Если мы анализируем один признак (например, функциональный класс по NYHA), нам достаточно одномерной таблицы. Но если мы хотим узнать, связаны ли два качественных признака между собой (например, зависит ли выбор цвета автомобиля от пола покупателя), нам нужна двумерная таблица, которую называют ***таблицей сопряженности***.

Давайте разберем структуру таблицы сопряженности на конкретном практическом примере. Представьте, что медицинская клиника тестирует новый метод профилактики простудных заболеваний. В эксперименте приняли участие $n = 100$ человек. Одни из них прошли курс профилактики, другие — нет. В конце сезона зафиксировали, заболел человек или остался здоров.

Мы имеем два качественных бинарных (дихотомических) признака:

1. Профилактика: пройдена / не пройдена (строки).

2. Результат: заболел / здоров (столбцы).

Мы сводим реальные, фактически подсчитанные данные в таблицу. Эти числа в статистике называются ***наблюдаемыми частотами*** и обозначаются буквой *O*.

Давайте проанализируем структуру: числа 10, 40, 20, 30 внутри таблицы — это наши наблюдаемые частоты (*O*). Числа по краям («Итого») называются ***маргинальными частотами***. Они показывают распределение каждого признака по отдельности. Например, мы видим, что всего заболело 30 человек (30% выборки), а здоровыми остались 70 человек (70% выборки).

Таблица 8

Наблюдаемые частоты (*O*)

Группа	Заболел	Здоров	Итого (Маргинальные частоты)
Прошел профилактику	10	40	50
Не проходил	20	30	50
Итого (Маргинальные частоты)	30	70	100 Общий итог, *n*

Переведя данные в проценты внутри групп, мы заметим: среди тех, кто прошел профилактику, заболели всего 20% (10 / 50). А среди тех, кто не проходил — целых 40% (20 / 50).

Разница кажется очевидной. Но достаточно ли этого, чтобы заявить, что профилактика официально работает? Вдруг нам просто попалась такая случайная выборка из 100 человек, и при повторном тесте всё изменится? Нам нужен строгий математический судья — критерий Хи-квадрат (χ^2) Пирсона.

Критерий Хи-квадрат проверяет стандартную для статистики пару гипотез:

Нулевая гипотеза (H_0): признаки независимы. Между ними нет связи. (Профилактика никак не влияет на вероятность заболеть, а разница в процентах — чистая случайность).

Альтернативная гипотеза (H_1): признаки зависимы. Между ними существует статистически значимая связь.

Идея критерия гениальна. Математика моделирует идеальный мир, в котором нулевая гипотеза H_0 абсолютно верна. Если профилактика вообще никак не влияет на здоровье, какими должны были бы быть числа внутри ячеек нашей таблицы? Такие теоретические числа называются ***ожидаемыми частотами*** и обозначаются буквой *E*.

Как их рассчитать? Если связи нет, то в обеих группах (и среди тех, кто прошел курс, и среди тех, кто проигнорировал) доля заболевших должна в точности соответствовать среднему показателю по больнице. А средний показатель нам известен из маргинальных частот: заболели 30% от всех участников (30 / 100 = 0,3).

Значит, в группе профилактики (50 человек) должны были заболеть 50×0,3 = 15 человек. Для автоматизации этого процесса существует простая универсальная формула расчета ожидаемой частоты для любой ячейки таблицы:

$$SE = \frac{\text{Итого по строке} \times \text{Итого по столбцу}}{\text{Общий итог } (n)}$$

Давайте рассчитаем ожидаемые частоты (E) для всех четырех ячеек нашей таблицы:

1. Ячейка (Прошел / Заболел): E = 50×30/100 = 15;
2. Ячейка (Прошел / Здоров): E = 50×70/100 = 35;
3. Ячейка (Не проходил / Заболел): E = 50×30/100 = 15;
4. Ячейка (Не проходил / Здоров): E = 50×70/100 = 35.

Построим теоретическую таблицу ожидаемых частот (табл. 9).

Таблица 9

Ожидаемые частоты (E)

Группа	Заболел	Здоров	Итого
Прошел профилактику	15	35	50
Не проходил	15	35	50
Итого (Маргинальные частоты)	30	70	100

Критически важное правило применимости теста: критерий Хи-квадрат работает на основе математических приближений. Он будет давать точные результаты только тогда, когда объем данных достаточен. Главное правило гласит: все рассчитанные ожидаемые частоты (E) должны быть больше или равны 5. В нашем случае минимальное значение равно 15, условие идеально выполняется. Если бы в какой-то ячейке E оказалось меньше 5, тест Хи-квадрат применять было бы нельзя (пришлось бы использовать точный критерий Фишера).

Теперь у нас есть две таблицы: наблюдаемая (реальность O) и ожидаемая (теория E). Наша задача — измерить, насколько сильно реальность отклонилась от теории. Формула критерия Хи-квадрат выглядит так:

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

Мы должны проделать это несложное действие для каждой из ячеек нашей таблицы, а результаты сложить:

1. Ячейка (Прошел / Заболел): $(10 – 15)^2/15 \approx 1{,}667$
2. Ячейка (Прошел / Здоров): $(40 – 35)^2/35 \approx 0{,}714$
3. Ячейка (Не проходил / Заболел): $(20 – 15)^2/15 \approx 1.667$
4. Ячейка (Не проходил / Здоров): $(30 – 35)^2/35 \approx 0.714$

Суммируем все четыре полученных значения отклонений:

$$\chi^2 = 1.667 + 0.714 + 1.667 + 0.714 = 4.762$$

Мы получили итоговое расчетное значение критерия: $\chi^2 = 4{,}76$. Осталось понять, является ли это отклонение критическим.

Само по себе число 4,76 нам ни о чем не говорит. Масштаб распределения χ^2 зависит от количества степеней свободы (df). Для таблиц сопряженности количество степеней свободы показывает, сколько ячеек таблицы могут заполняться абсолютно свободно, прежде чем маргинальные итоги жестко зафиксируют оставшиеся значения.

Формула расчета степеней свободы:

$$df = (R – 1)\,(C – 1)$$

Где R — количество строк таблицы, а C — количество столбцов (итоговые строки и столбцы не считаются). У нас таблица 2×2, следовательно:

$$df = (2 – 1)\times(2 – 1) = 1$$

Математическое распределение Хи-квадрат для 1 степени свободы имеет вполне конкретную форму. Существуют специальные таблицы критических значений (или встроенные функции в Excel/Python), которые говорят: при df = 1, чтобы опровергнуть случайность с надежностью 95% (уровень значимости $\alpha = 0{,}05$), наше значение Хи-квадрат должно преодолеть критический порог, равный 3,84.

Поскольку наш реальный χ^2 оказался больше критического ($4{,}76 > 3{,}84$), он попадает в так называемую зону отклонения нулевой гипотезы. Соответствующее значение p-value в данном случае составит примерно 0,029 (или 2,9%), что значительно меньше стандартного порога в 0,05 (5%).

Мы с уверенностью отвергаем нулевую гипотезу H_0 и принимаем альтернативную гипотезу H_1. Различия в заболеваемости между группами не могут быть объяснены простой случайностью выборки. Мы доказали, что прохождение курса профилактики статистически значимо связано со снижением риска заболеть.

5.5. Анализ качественных данных: краткие итоги

- Качественные данные анализируются через подсчет частот встречаемости и долей.
- Базовым кирпичиком для дихотомических (да/нет) данных является биномиальное распределение. При больших объемах выборки оно аппроксимируется нормальным.
- Лучший инструмент для защиты качественных данных от перекосов — стратифицированная случайная выборка
- Для оценки долей в опросах всегда рассчитывайте доверительный интервал, чтобы понимать истинную точность вашего исследования.
- Стандартная ошибка доли напрямую зависит от баланса мнений (максимальна при 50/50) и обратно пропорциональна квадратному корню из объема выборки. Чтобы уменьшить неопределенность в 2 раза, вам придется увеличить выборку в 4 раза.
- Доверительный интервал — это официальный способ зафиксировать масштаб неопределенности и ограничить её рамками математической погрешности.
- Чтобы проверить взаимосвязь двух качественных признаков, сводите их в таблицы частот и применяйте тест Хи-квадрат.
- Всегда проверяйте требование теста Хи-квадрат: ни одна из ожидаемых частот в расчетной таблице не должна быть меньше 5.
- Критерий Хи-квадрат доказывает факт существования связи, но помните, что он не измеряет её силу и не доказывает прямую биологическую или психологическую причину — поиск первопричин лежит на плечах самого исследователя.

Глава 6. Оценка взаимосвязей

6.1. Диаграмма рассеяния, корреляция

До этого момента в наших лекциях мы изучали переменные изолированно друг от друга. Это называется одномерным анализом. Однако реальный мир устроен гораздо интереснее. В нем все процессы переплетены между собой. Мы пытаемся понять: «Связана ли дозировка лекарства со скоростью выздоровления пациента?» «Одинаковы ли значения начальных и повторных измерений при изучении их воспроизводимости?» «зависит ли длина 6-минутной ходьбы от массы тела пациента?». Когда мы начинаем изучать две количественные переменные одновременно, мы вступаем в область анализа взаимосвязей.

Любой серьезный анализ взаимосвязей всегда должен начинаться с визуализации. Главным инструментом здесь выступает диаграмма рассеяния — это двумерный график, где:

1. По горизонтальной оси (X) откладываются значения независимой переменной (её также называют объясняющей переменной или предиктором).

2. По вертикальной оси (Y) откладываются значения зависимой переменной («отклика» или результирующего признака).

Каждое наблюдение из нашей выборки (например, конкретный магазин с его затратами на рекламу и выручкой) наносится на этот график в виде точки с координатами (x, y).

Когда мы смотрим на получившееся «облако точек», мы можем мгновенно определить характер взаимосвязи:

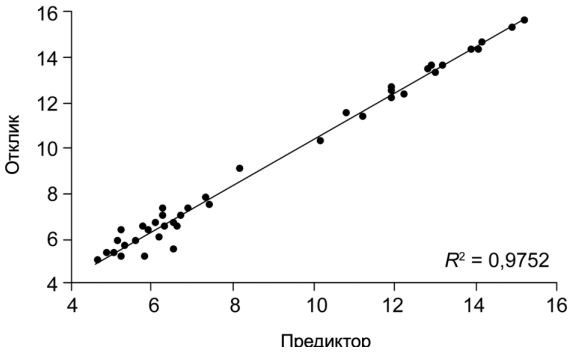


Рис. 18. Пример диаграммы рассеяния

1. Положительная связь: точки устремляются из левого нижнего угла в правый верхний. Это значит, что с ростом X значения Y тоже имеют тенденцию увеличиваться (например, рост расходов на рекламу ведет к росту продаж).

2. Отрицательная связь: точки идут из левого верхнего угла в правый нижний. С ростом X значения Y уменьшаются (например, чем выше цена товара, тем ниже спрос на него).

3. Отсутствие связи: точки хаотично разбросаны по графику в виде бесформенного круглого облака. Изменение X никак не влияет на Y.

4. Нелинейная связь: точки выстраиваются в форме дуги, параболы или волны. Связь есть, но она более сложная, чем простая прямая линия.

Диаграмма рассеяния дает нам качественное понимание, но языку науки нужны точные цифры. Нам необходим показатель, который ответит на вопрос: «Насколько эта связь сильная и насколько она линейная?».

Для этого великий статистик Карл Пирсон разработал коэффициент линейной корреляции (r). Ключевые свойства коэффициента корреляции Пирсона:

- Он строго ограничен диапазоном от –1 до +1.
- Знак коэффициента показывает направление связи. Если $r > 0$, связь положительная. Если $r < 0$, связь отрицательная.
- Абсолютное значение $|r|$ показывает силу связи:
- $r = 0$ — линейная связь полностью отсутствует.
- $|r|$ от 0,1 до 0,3 — связь слабая;
- $|r|$ от 0,3 до 0,7 — связь умеренная;
- $|r|$ от 0,7 до 0,99 — связь сильная (высокая);
- $r = +1$ или $r = –1$ — идеальная (строго функциональная) связь. Все точки на графике лежат идеально на одной прямой линии.

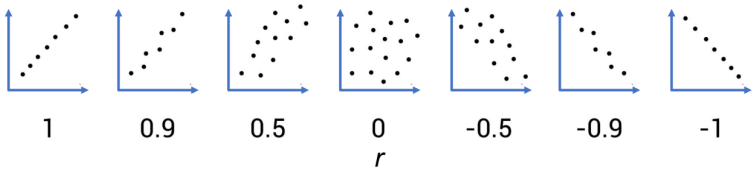


Рис. 19. Диаграммы рассеяния

Формула коэффициента корреляции соотносит совместную изменчивость (ковариацию) двух переменных с их индивидуальным разбросом (стандартными отклонениями):

$$r = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Если точки отклоняются от своих средних значений синхронно (когда *X* больше своего среднего, *Y* тоже больше своего среднего), числитель становится большим и положительным — это рождает сильную положительную корреляцию.

6.2. Линейная регрессия и коэффициент детерминации

Корреляция отвечает на вопрос «Есть ли связь?». Но она не позволяет нам строить прогнозы. Если мы знаем, что связь сильная, мы хотим получить математическое уравнение, в которое можно подставить будущее значение *X* и рассчитать ожидаемое значение *Y*.

Процесс поиска такой оптимальной математической линии называется линейной регрессией. Уравнение прямой линии выглядит как в школьной геометрии:

$$\hat{y} = a + b \times x$$

где: $\hat{y}$ (игрек с крышкой) — предсказанное (теоретическое) значение зависимой переменной; *a* — свободный член (интерцепт), точка, в которой наша линия пересекает ось *Y* при $x = 0$; *b* — угловой коэффициент (коэффициент наклона или коэффициент регрессии). Он имеет важнейший физический смысл: показывает, на сколько единиц изменится *Y*, если *X* увеличится ровно на одну свою единицу.

Если мы попытаемся провести линию через облако точек «на глаз», у каждого исследователя получится свой вариант. Чтобы найти единственную, математически идеальную прямую, используется метод наименьших квадратов.

Для каждой точки на графике существует вертикальное расстояние между её реальным значением (y) и нашей теоретической линией ($\hat{y}$). Это расстояние называется остатком или ошибкой прогноза ($e = y - \hat{y}$).

Метод наименьших квадратов ищет такие коэффициенты *a* и *b*, при которых сумма квадратов всех этих остатков будет минимально возможной:

$$\sum e^2 = \sum (\hat{y} - \bar{y})^2 \to min$$

Формулы для расчета коэффициентов регрессии напрямую завязаны на уже посчитанный нами коэффициент корреляции Пирсона (*r*):

$$b = r \times \frac{s_y}{s_x} \text{ и } a = \bar{y} - b \times \bar{x}$$

где s_x и s_y — выборочные стандартные отклонения переменных *X* и *Y*.

Построить уравнение регрессии можно для любых данных, даже если они слабо связаны. Как нам понять, насколько хорошо наша получившаяся линия описывает реальность? Насколько точны наши прогнозы?

Метрикой качества выступает ***коэффициент детерминации*** (R^2). Для парной линейной регрессии он рассчитывается очень просто — это квадрат коэффициента корреляции Пирсона: $R^2 = (r)^2$.

Коэффициент детерминации измеряется от 0 до 1 (или от 0% до 100%) и имеет блестящую логическую интерпретацию:

R^2 показывает долю дисперсии (разброса) зависимой переменной *Y*, которая успешно объясняется построенной математической моделью регрессии. Пример: если мы изучаем зависимость веса человека от его роста и получили $r = 0{,}8$, то $R^2 = 0{,}8^2 = 0{,}64$ (или 64%). Это значит, что 64% всех различий в весе людей в нашей выборке объясняются их ростом. А оставшиеся 36% разброса вызваны другими факторами, которые наша модель не учла (генетика, диета, спорт и т.д.).

Необходимо хорошо уяснить, что корреляция не означает причинность! Наличие сильной корреляции между *X* и *Y* само по себе никогда не доказывает, что *X* является причиной возникновения *Y*. Высокий коэффициент корреляции лишь констатирует факт: «Эти две цифры меняются синхронно». Но эта синхронность может быть вызвана тремя совершенно разными причинами:

1. Прямая причинность: *X* действительно порождает *Y* (Увеличение осадков увеличивает урожай).

2. Обратная причинность: на самом деле *Y* порождает *X* (Мы видим корреляцию между количеством полицейских на улице и уровнем преступности. Но не полицейские вызывают преступ-

ность, а высокая преступность заставляет город нанимать больше полиции).

3. Третий скрытый фактор (ложная корреляция): существует скрытая переменная Z, которая одновременно управляет и переменной X, и переменной Y. Классический пример: статистика показывает сильную положительную корреляцию между продажами мороженого (X) и количеством лесных пожаров (Y). Означает ли это, что поедание мороженого поджигает леса? Нет. Существует скрытый фактор Z — летняя жара. Когда наступает жара, люди покупают больше мороженого, и одновременно с этим из-за суши вспыхивают леса.

6.3. Множественная линейная регрессия

Мы только что изучили простую линейную регрессию, которая связывала между собой две переменные. Например, мы могли попытаться предсказать стоимость квартиры (Y), зная только её площадь (X). Но давайте будем честны: устроен ли наш мир так просто? Зависит ли цена недвижимости только от площади? Конечно, нет. На нее одновременно влияет огромное количество факторов: количество комнат, удаленность от центра города или метро, год постройки здания, этаж и наличие парковки. Если мы будем игнорировать эти факторы и смотреть только на площадь, наш прогноз будет очень неточным. Реальный мир многомерен. Чтобы прогнозировать сложные процессы, нам нужен инструмент, способный объединить в одной математической модели сразу несколько объясняющих переменных. Этот инструмент — ***множественная линейная регрессия***.

В случае простой регрессии мы строили прямую линию на плоскости. В множественной регрессии мы строим математическую плоскость (или гиперплоскость, если факторов больше двух) в многомерном пространстве. Уравнение множественной линейной регрессии в общем виде выглядит так:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$$

Давайте разберем каждый элемент этого уравнения, так как в них скрыта вся суть анализа: $\hat{y}$ — зависимая переменная (результат, который мы прогнозируем); $\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$ — независимые переменные (наши факторы, предикторы); β_0 (бета-нулевое) — свободный член (интерцепт). Это теоретическое значение Y, если абсолютно все факторы x будут равны

нулю; ε (эпсилон) — случайная ошибка модели. Это тот хаос реальной жизни, который наша модель принципиально не может учесть (например, каприз конкретного покупателя).

Коэффициенты $\beta_1, \beta_2, \ldots \beta_k$ называют частичными (парциальными) коэффициентами регрессии. Каждый из них показывает, на сколько единиц изменится зависимая переменная *Y*, если соответствующий фактор x_i вырастет ровно на одну свою единицу, при условии, что все остальные факторы в модели останутся неизменными.

Слова «***при прочих в моделях равных условиях***» — это самое важное в множественной регрессии. Модель математически «очищает» влияние каждого фактора от примесей других переменных.

Как и в простой регрессии, для оценки того, насколько хорошо наша модель описывает реальность, используется коэффициент детерминации (R^2). Он показывает долю объясненной дисперсии. Однако в множественной регрессии с R^2 связана одна серьезная математическая ловушка. Дело в том, что R^2 упрямо растет (или как минимум не уменьшается) каждый раз, когда вы добавляете в модель любой новый фактор, даже если этот фактор — абсолютный бред (например, цвет носков продавца квартиры).

Если бездумно добавлять переменные, обычный R^2 покажет ложное улучшение модели. Это приводит к так называемому переобучению (overfitting). Чтобы решить эту проблему, используют скорректированный коэффициент детерминации (adjusted R^2). В формулу скорректированного R^2 встроен «штраф» за количество переменных:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n = 1}{n - k - 1}$$

где *n* — объем выборки, а *k* — количество факторов. Если вы добавляете в модель бесполезный фактор, скорректированный R^2 отреагирует на это и «упадет», сигнализируя о том, что модель ухудшилась. При сравнении многофакторных моделей всегда нужно ориентироваться именно на Adjusted R^2.

Чтобы прогнозы множественной регрессии были несмещенными и надежными (чтобы модель была «лучшей»), ваши данные должны строго соответствовать четырем главным условиям:

1. ***Линейность***: взаимосвязь между зависимой переменной и каждым из предикторов должна быть в целом линейной.

2. ***Гомоскедастичность***: разброс (дисперсия) случайных ошибок ε должен быть постоянным для любых значений факторов. Если с ростом масштаба данных ошибки начинают «разлетаться» сильнее (например, при прогнозировании доходов богатых людей ошибки прогноза гораздо больше, чем для бедных), возникает гетероскедастичность, которая искажает точность тестов значимости.

3. ***Нормальность остатков***: случайные ошибки (остатки) модели должны быть распределены по закону нормального распределения с математическим ожиданием, равным нулю.

4. ***Независимость остатков*** (отсутствие автокорреляции): ошибка одного наблюдения не должна быть связана с ошибкой другого. Это допущение часто нарушается в временных рядах (например, ошибка прогноза курса акций сегодня почти наверняка связана с ошибкой прогноза вчера).

При построении многофакторной модели аналитики часто совершают ошибку, включая в нее переменные, которые практически дублируют друг друга. Например: «Общая площадь квартиры» и «Жилая площадь квартиры». Или «Возраст человека» и «Стаж работы». Ситуация, когда между собой сильно коррелируют сами независимые переменные (x_i), называется ***мультиколлинеарностью***.

К чему приводит мультиколлинеарность? Она буквально сводит математический алгоритм с ума. Компьютер не может понять, какому именно из двух взаимосвязанных факторов нужно приписать заслугу в изменении *Y*. В результате:

- Стандартные ошибки коэффициентов резко возрастают.
- Коэффициенты β становятся крайне нестабильными: если вы измените всего одну точку в данных, коэффициент может поменять знак с плюса на минус.
- Тесты значимости (*t*-тесты) начинают утверждать, что ни один фактор не важен по отдельности, хотя модель в целом (по F-критерию) работает отлично.

Как с ней бороться? Для поиска мультиколлинеарности рассчитывают показатель *VIF* (Variance Inflation Factor) — коэффициент «вздутия» дисперсии:

Если $VIF > 5$, между факторами есть связь, следует насторожиться.

Если $VIF > 10$, мультиколлинеарность критическая. В этом случае один из дублирующих факторов необходимо безжалостно удалить из модели.

Давайте потренируемся, на примере. Представьте, что мы построили модель предсказания веса топ-менеджеров (Y, кг) на основе двух факторов: x_1 — количество потребляемых калорий в день, x_2 — количество часов спорта в неделю.

Программа выдала нам следующее уравнение:

$$\hat{y} = 45 + 0{,}02\, x_1 - 1{,}5\, x_2$$

Интерпретируем:

1. Интерцепт (45): если человек вообще не ест (калории = 0) и не занимается спортом (часы = 0), его теоретический базовый вес составит 45 кг. (В данном контексте значение имеет скорее математический смысл, так как жить с 0 калорий физически невозможно).

2. Коэффициент при x_1 (+0,02): при неизменном количестве занятий спортом, увеличение ежедневного рациона на каждые 100 калорий ведет к увеличению веса менеджера в среднем на $100 \times 0{,}02 = 2$ кг.

3. Коэффициент при x_2 (–1,5): при строго фиксированном рационе питания, каждый дополнительный час занятий спортом в неделю снижает вес менеджера в среднем на 1,5 кг.

6.4. Оценка взаимосвязей: краткие итоги

- Оценка взаимосвязей начинается с визуализации данных на диаграмме рассеяния.
- Направление и силу линейной связи оценивает коэффициент корреляции Пирсона (r) в диапазоне от –1 до +1.
- Линейная регрессия ($\hat{y} = a + bx$) позволяет составить математическое уравнение прогноза, где коэффициент b показывает точный шаг изменения Y при росте X на единицу.
- Линия регрессии строится по методу наименьших квадратов, минимизируя квадраты ошибок прогноза.
- Насколько модель хороша, оценивает коэффициент детерминации (R^2), показывающий процент объясненной изменчивости.
- Корреляция — это лишь синхронность числовых рядов, истинную причинно-следственную связь можно доказать

только глубоким знанием сути процесса или контролируемым экспериментом.

- Множественная линейная регрессия позволяет заглянуть в реальный многомерный мир, оценивая влияние нескольких факторов на один результат одновременно.
- Коэффициенты модели показывают изолированное влияние фактора при условии, что все остальные переменные «заморожены».
- Для оценки качества многофакторных моделей всегда смотрите на скорректированный R^2, так как обычный R^2 склонен к самообману при раздувании модели.
- Опасайтесь мультиколлинеарности — не включайте в модель предикторы, которые сильно коррелируют друг с другом. Контролируйте их через показатель *VIF*.

Глава 7. Сравнение двух и более групп

7.1. Критерий Стьюдента: сравнение независимых групп

В прикладной аналитике и научных исследованиях одна из самых частых задач — это сравнение двух групп между собой. Нам нужно отвечать на вопросы вида: снижает ли новый препарат среднее артериальное давление эффективнее, чем старый стандарт лечения (группа А против группы Б). Речь идёт о независимых (несвязанных) выборках. Это значит, что выбор объектов в одну группу никак не влияет на то, какие объекты попадут во вторую группу. Это разные люди.

Когда мы берем две выборки и считаем в них средние арифметические ($\bar{x}_1$ и $\bar{x}_2$), мы почти никогда не получим два абсолютно одинаковых числа. Между ними всегда будет какая-то разница. Главный вопрос статистического анализа: является ли эта разница признаком реального эффекта (статистически значимой) или это просто случайное колебание выборки? Для ответа на этот вопрос мы выдвигаем пару гипотез:

Нулевая гипотеза (H_0): истинные средние значения в генеральных совокупностях равны ($\mu_1 = \mu_2$). Различие между выборочными средними случайно.

Альтернативная гипотеза (H_1): Истинные средние значения не равны ($\mu_1 \neq \mu_2$). Различие значимо и вызвано реальным фактором.

Главным инструментом для решения этой задачи выступает ***двухвыборочный t-критерий Стьюдента***. Его логика невероятно элегантна. Чтобы понять, насколько силен обнаруженный нами эффект, критерий соотносит величину этого эффекта с уровнем естественного «шума» (вариативности) в данных.

Архитектура формулы *t*-статистики выглядит как дробь:

$$t = \frac{\text{Эффект}}{\text{Шум}} = \frac{\bar{x}_1 - \bar{x}_2}{\text{Стандартная ошибка разности средних}}$$

Давайте вдумаемся в эту структуру:

1. Числитель: чем дальше друг от друга находятся средние значения двух групп, тем больше числитель, тем больше итоговое значение *t*, и тем легче нам заявить, что группы действительно разные.

2. Знаменатель (шум): это погрешность, вызванная тем, что данные внутри групп имеют разброс. Если внутри групп люди

очень разные (высокая дисперсия), то «облака данных» будут сильно пересекаться, и доказать различие средних будет трудно. Если же данные внутри групп однородны (низкая дисперсия), даже небольшая разница средних станет очевидной.

В зависимости от того, равны ли дисперсии (разброс данных) в двух группах, математическая формула знаменателя будет немного отличаться. Классический вариант (при равных дисперсиях) использует объединенную стандартную ошибку:

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

где s_p^2 — это объединенная дисперсия, которая рассчитывается как средневзвешенное значение дисперсий обеих групп:

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

n_1, n_2 — объемы выборок в первой и второй группах (они могут быть разными).

s_1^1, s_2^2 — выборочные дисперсии групп.

Количество степеней свободы (*df*), которое определит форму распределения Стьюдента для нашего случая, рассчитывается как:

$$df = n_1 + n_2 - 2$$

Если же дисперсии в группах существенно различаются, применяется модификация — *t*-критерий Уэлча (Welch's *t*-test). В нем формула ошибки корректируется, а число степеней свободы рассчитывается по более сложной формуле, защищая исследователя от ложноположительных результатов.

t-критерий Стьюдента — это параметрический тест. Вы не можете просто запустить его на любых данных. Сначала нужно проверить три строгих условия:

1. ***Независимость наблюдений***. Объекты внутри групп и между группами не должны быть связаны. Если один человек участвовал в обоих тестах (дизайн «До — После»), этот критерий использовать нельзя (нужен парный *t*-тест).

2. ***Нормальность распределения***. Данные в каждой из двух групп должны быть распределены приблизительно нормально.

Как проверять: визуально (через гистограммы или график *Q-Q* plot) либо с помощью статистических тестов (Критерий Шапиро-Уилка).

Но есть послабление: если ваши выборки большие ($n_1 > 30$ и $n_2 > 30$), критерий Стьюдента становится относительно устойчивым к нарушениям нормальности.

Как проверить, что разброс данных в первой группе должен быть примерно равен разбросу данных во второй группе? Для этого есть тест Левена (Levene's test) или тест Бартлетта. Если тест Левена показывает $p < 0{,}05$, значит, дисперсии не равны. В таком случае в программе нужно поставить галочку «использовать поправку Уэлча».

Давайте попрактикуемся. Задача: профессор проводит дистанционные лекции и тестирует два подхода: в одном он демонстрирует обычные статичные слайды PowerPoint, в другом материал богато иллюстрирован видео и анимацией.

Группа 1 ($n_1 = 15$ клинических ординаторов) слушала лекции в старой версии. Их среднее время удержания в приложении для дистанционного обучения составило $\bar{x}_1 = 40$ минут ($s_1^2 = 25$).

Группа 2 ($n_2 = 15$ клинических ординаторов) слушала и смотрела лекции в новой версии. Их среднее время составило $\bar{x}_2 = 45$ минут ($s_2^2 = 20$).

Выясним, действительно ли видео и анимация удерживают пользователей лучше, или разница в 5 минут случайна. Дисперсии считаем примерно равными.

Шаг 1. Считаем объединенную дисперсию

Поскольку объемы выборок равны (15 и 15), объединенная дисперсия — это просто среднее между 25 и 20:

$$s_p^2 = \frac{(15-1)\times 25 + (15-1)\times 20}{15+15-2} = 22{,}5$$

Шаг 2. Вычисляем *t*-статистику

$$t_{\text{расч}} = \frac{40-45}{\sqrt{\frac{22{,}5}{15}+\frac{22{,}5}{15}}} \approx -2{,}89$$

Знак минус просто говорит о том, что второе среднее больше первого. Для проверки гипотезы мы смотрим на модуль: $|t| = 2{,}89$).

Шаг 3. Находим критическое значение и принимаем решение. Количество степеней свободы: $df = 15 + 15 - 2 = 28$. По таблице распределения Стьюдента для двустороннего теста при уровне значимости $\alpha = 0{,}05$ и $df = 28$ находим критическое значение:

$$t_{\text{крит}} = 2{,}048$$

Сравнивая расчетное (модуль) и критическое значения, видим, что эмпирическое значение ушло далеко в критическую зону (соответствующее значение *p*-value составит примерно 0,007, что значительно меньше порога в 0,05).

Итоговый статистический вывод: мы уверенно отвергаем нулевую гипотезу H_0. Разница в 5 минут не является случайной. Использование видео и анимации в лекции статистически значимо увеличивает среднее время удержания ординаторов.

7.2. Парный *t*-тест: сравнение зависимых групп

Однако в науке, медицине и бизнесе существует огромный пласт задач, где выборки жестко связаны друг с другом. Такие выборки называются ***зависимыми или парными***.

Самый классический дизайн исследования, приводящий к парным выборкам — это схема «До — После». Например, измерение артериального давления у одних и тех же пациентов до приема лекарства и после него. Или тестирование знаний студентов в начале семестра (пре-тест) и в конце семестра (пост-тест).

Биологические или социальные близнецы, а также пары, подобранные по принципу копий (мэтчинг), тоже образуют парные выборки. Главная особенность здесь в том, что каждому наблюдению из первой выборки строго соответствует ровно одно конкретное наблюдение из второй выборки. Из-за этого объемы обеих выборок всегда абсолютно одинаковы ($n_1 = n_2$).

Зачем нам нужен отдельный критерий именно для пар? Почему нельзя применить обычный двухвыборочный *t*-тест для независимых групп? Дело в том, что в независимом тесте случайный разброс между людьми (индивидуальные различия) сильно «зашумляет» данные. Например, у одного пациента базовое давление 140 мм рт. ст., а у другого — 110. Этот индивидуальный шум мешает увидеть реальный эффект лекарства.

Парный *t*-критерий совершает изящный математический маневр: он полностью устраняет межиндивидуальную изменчивость. Вместо того чтобы анализировать два разных массива данных ($X_{\text{до}}$ и $X_{\text{после}}$), критерий рассчитывает попарную разность для каждого участника:

$$d_i = X_{\text{после},i} - X_{\text{до},i}$$

Таким образом, задача сравнения двух выборок мгновенно превращается в задачу анализа одной-единственной выборки разностей (d).

Формулировка гипотез:

- Нулевая гипотеза (H0): истинное среднее значение разностей в популяции равно нулю ($\mu_d = 0$). Это означает, что никакого эффекта нет: показатели «до» и «после» в среднем одинаковы, а любые колебания случайны.
- Альтернативная гипотеза (H1): истинное среднее значение разностей не равно нулю ($\mu_d \neq 0$). Эффект существует (показатели изменились).

Поскольку мы свели задачу к одной выборке разностей, формула парного t-критерия становится очень похожей на классический одновыборочный критерий Стьюдента. Формула расчетного значения t-статистики:

$$t = \frac{\bar{d}}{SEM_d} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

Давайте разберем компоненты: $\bar{d}$ — среднее арифметическое всех получившихся разностей в нашей выборке; s_d — стандартное отклонение этих разностей (показывает, насколько сильно индивидуальные сдвиги «гуляют» вокруг среднего сдвига); n — количество пар наблюдений (а не общее количество измерений); SEM_d — стандартная ошибка среднего для разностей.

Логика формулы: чем больше средний сдвиг $\bar{d}$ и чем больше объем выборки (n), тем выше будет значение t-статистики, и тем легче нам будет опровергнуть нулевую гипотезу о неэффективности. Напротив, если индивидуальные разности слишком хаотичны (большое s_d), разглядеть истинный эффект будет тяжело.

Число степеней свободы (df) для парного критерия рассчитывается на основе количества пар:

$$df = n - 1$$

Парный t-критерий Стьюдента относится к группе параметрических тестов. Это значит, что он выдвигает определенные требования к структуре ваших данных. Условий всего два, но они критичны:

1. Зависимость внутри пар и независимость между парами: измерения внутри одной пары (например, «до» и «после» одного человека) обязаны быть связанными. Но при этом сами пары

(разные люди) должны быть полностью независимы друг от друга.

2. Нормальное распределение разностей: Внимание! Нормально распределенными должны быть не исходные данные ($X_{до}$ или $X_{после}$), а именно получившиеся разности (d_i).

Если объем выборки мал ($n < 30$), а распределение разностей сильно скошено или имеет тяжелые хвосты (выбросы), парный t-тест использовать нельзя — он начнет безбожно ошибаться. В таких ситуациях нужно переходить на непараметрический аналог — критерий знаковых рангов Уилкоксона. Однако, если выборка большая ($n > 50$), благодаря центральной предельной теореме t-критерий становится устойчивым к небольшим отклонениям от нормальности.

7.3. Дисперсионный анализ, ANOVA

Мы детально разобрали, как сравнивать средние значения в двух независимых или зависимых группах. Но что делать, если групп в нашем исследовании три, четыре или больше? Первое интуитивное решение, которое приходит в голову — просто взять и сравнить все группы попарно с помощью обычного t-критерия (по очереди группу 1 с группой 2, группу 2 с группой 3, и группу 1 с группой 3). Для трех групп нам понадобится 3 сравнения. А если групп будет 5, то количество парных сравнений возрастет до 10!

В статистике такой подход не используется из-за проблемы множественных сравнений. Каждое применение t-критерия несет в себе 5%-й риск совершить ошибку первого рода (найти различия там, где их нет, — ложная тревога). Если мы сделаем 10 независимых сравнений подряд, общая вероятность ошибиться и сделать ложное заявление вырастет почти до 40%. Чтобы избежать этой катастрофы и проверить различия между всеми группами одновременно за один шаг, известный статистик Рональд Фишер разработал метод, который называется ***ANOVA (ANalysis Of Variance) — Дисперсионный анализ***.

Название метода часто вводит новичков в ступор. Наша цель — узнать, равны ли между собой средние значения нескольких групп. Но почему тогда анализ называется дисперсионным (анализом разброса)? В этом скрыта главная изящная идея Фишера. Чтобы понять, отличаются ли группы друг от

друга, ANOVA разделяет общий разброс (дисперсию) всех собранных нами данных на две изолированные части:

1. ***Межгрупповая дисперсия***: это разброс между групповыми средними значениями. Она показывает, насколько далеко центры наших групп ушли друг от друга. Если эта дисперсия велика, это сигнализирует о том, что наш изучаемый фактор действительно оказывает сильное влияние.

2. ***Внутригрупповая дисперсия***: это разброс данных внутри каждой отдельной группы относительно её собственного среднего. Это естественный, случайный «шум» данных — ведь, например, даже при одинаковом удобрении каждый конкретный колосок пшеницы вырастает немного разным из-за микроклимата, генетики или капризов природы.

Логика ANOVA проста: мы сравниваем эти два вида разброса между собой. Если межгрупповой разброс (расстояние между группами) существенно превышает внутригрупповой разброс (случайный шум вокруг средних), мы делаем вывод: группы действительно значимо отличаются друг от друга, и наш фактор работает. Если же группы расположены близко друг к другу, а внутри групп данные хаотично и сильно разбросаны, то межгрупповые различия просто утонут в шуме. Значит, эффект фактора не доказан.

Давайте переведем эту логику на язык строгих математических формул. Весь расчет ANOVA сводится к заполнению так называемой «таблицы дисперсионного анализа».

Пусть k — количество групп, а n — общее число наблюдений во всех группах вместе взятых.

Шаг 1. Расчет сумм квадратов отклонений (SS — Sum of Squares). Сначала мы считаем объемы изменчивости:

$SS_{\text{между}}$ (Межгрупповая сумма квадратов) — измеряет изменчивость, обусловленную различием между группами.

$SS_{\text{внутри}}$ (Внутригрупповая сумма квадратов, или сумма квадратов ошибок) — измеряет случайный разброс внутри групп.

$SS_{\text{общ}}$ (Общая сумма квадратов) — их сумма.

Шаг 2. Расчет степеней свободы (df — Degrees of Freedom)

Для каждой суммы квадратов рассчитывается свое число степеней свободы:

$df_{\text{между}} = k - 1$ (Количество групп минус 1).

$df_{\text{внутри}} = n - k$ (Общее число наблюдений минус количество групп).

Шаг 3. Расчет средних квадратов (MS — Mean Squares)

Средние квадраты — это и есть те самые очищенные показатели дисперсий (разброса), приведенные к сопоставимому масштабу. Чтобы их получить, мы просто делим суммы квадратов на соответствующие степени свободы:

$$MS_{\text{между}} = \frac{SS_{\text{между}}}{df_{\text{между}}}$$

$$MS_{\text{внутри}} = \frac{SS_{\text{внутри}}}{df_{\text{внутри}}}$$

Шаг 4. Вычисление *F*-статистики Фишера

Финальный аккорд — нахождение отношения межгрупповой дисперсии к внутригрупповой:

$$F = \frac{MS_{\text{между}}}{MS_{\text{внутри}}}$$

Полученное расчетное значение F мы сопоставляем с теоретическим (табличным) распределением Фишера для заданных степеней свободы ($df_1 = k - 1$, $df_2 = n - k$). Если наш расчетный коэффициент F больше критического (порог для уровня значимости $\alpha = 0{,}05$), а значение p-value $< 0{,}05$, то мы имеем право праздновать победу.

Как и любой серьезный параметрический критерий, ANOVA выдвигает жесткие требования к структуре данных, нарушение которых превратит результаты в бессмыслицу:

1. Независимость наблюдений: все наблюдения должны быть независимы как внутри групп, так и между ними (для связанных групп, например, замеры одних и тех же людей в 3 разных промежутка времени, используется специальная модификация — Repeated Measures ANOVA).

2. Нормальность распределения: данные внутри каждой из сравниваемых групп должны быть распределены приблизительно нормально. Проверить это можно визуально (гистограммы, *Q-Q* plots) или тестом Шапиро-Уилка. При больших выборках критерий устойчив к небольшим отклонениям от нормальности.

3. Гомогенность (равенство) дисперсий: разброс данных внутри каждой группы должен быть примерно одинаковым. Это допущение проверяется тестом Левена (Levene's test). Если тест Левена показывает $p < 0{,}05$, это значит, что дисперсии в группах значимо различаются. В таком случае классический ANOVA

проводить нельзя, и нужно использовать альтернативные поправки (например, тест Уэлча для ANOVA).

Что дальше? Представьте, что мы успешно провели ANOVA для изучения теста 6-минутной ходьбы у пациентов с четырьмя ФК по NYHA (I, II, III, IV) и получили заветный результат: $F = 5{,}67$, p-value = 0,003.

Мы радостно формулируем вывод: «Нулевая гипотеза отвергнута! Различия между средними значениями групп статистически значимы!».

Но тут к нам подходит научный руководитель и задает логичный вопрос: «Отлично, различия есть. Но какой именно класс отличается? IV от III? Или II лучше, чем III? Или они все одинаковые, но только вариант IV провалился?».

И вот здесь обнаруживается ограничение метода: ANOVA — это «омнибусный» (общий) критерий. Он лишь констатирует факт: «Среди ваших групп есть как минимум одна пара, средние значения которой значимо различаются». Но он не указывает, какая именно это пара.

Чтобы ответить на этот вопрос, после значимого результата ANOVA мы обязаны запустить пост-хок тесты (Post-hoc или тесты множественных сравнений).

Эти тесты устроены так, что они проводят попарные сравнения между всеми группами, но при этом автоматически корректируют (ужимают) уровень значимости, чтобы защитить нас от накопления ошибки ложноположительного результата. Самые популярные Пост-хок тесты:

Критерий Тьюки (Tukey's HSD): самый сбалансированный и популярный тест для сравнения всех групп со всеми. Используется, когда объемы выборок в группах примерно равны.

Поправка Бонферрони (Bonferroni correction): очень строгий (консервативный) метод, который просто делит требуемый уровень α (0,05) на количество сравнений. Из-за своей жесткости может упустить реальные мелкие различия.

Критерий Даннета (Dunnett's test): Применяется в специфических задачах, когда нам не нужно сравнивать все группы между собой, а нужно сравнить несколько экспериментальных групп строго с одной — контрольной (базовой).

7.4. Непараметрические тесты в контексте сравнения групп

Ранее мы изучили мощные инструменты сравнения групп: *t*-критерий Стьюдента (для двух групп) и дисперсионный анализ ANOVA (для трех и более групп). Эти тесты называются параметрическими, потому что они опираются на расчет параметров распределения (среднее арифметическое, дисперсия) и требуют, чтобы данные подчинялись строгой математической модели — закону нормального распределения. Однако в реальной исследовательской практике (особенно в медицине, биологии, социологии и психологии) мы постоянно сталкиваемся со сценариями, когда эти требования нарушаются, и наблюдается сильное отклонение от нормальности (асимметрия), когда ваши данные могут иметь тяжелые «хвосты», несколько пиков или экстремальные выбросы, которые невозможно удалить без потери смысла исследования, либо малый объем выборки, если у вас в группе всего 5–10 человек (например, редкое заболевание или дорогостоящий эксперимент) и проверить данные на нормальность физически невозможно — тесты вроде Шапиро-Уилка просто не обладают достаточной мощностью. Часто результаты измеряются не в физических величинах, а в шкалах порядка. Например: стадия рака (I, II, III, IV), степень тяжести симптомов (нет / легкая / средняя / тяжелая), оценки по шкале Лайкерта в анкетах. Считать «средний балл» для таких данных математически некорректно.

В этих ситуациях на арену выходят непараметрические тесты (тесты, свободные от распределения — distribution-free tests). Они не делают никаких предположений о форме распределения данных в генеральной совокупности и не пытаются сравнивать средние значения.

Как непараметрические критерии умудряются анализировать данные без расчета средних и дисперсий? Они используют концепцию ранжирования.

Ранжирование — это процесс замены реальных числовых или порядковых значений их позициями (рангами) в упорядоченном по возрастанию ряду данных.

Представьте, что у нас есть выборка из 5 значений: 12, 105, 15, 11, 14.

Среди них есть явный выброс — число 105. Если мы посчитаем среднее арифметическое, этот выброс исказит результат. Но давайте заменим числа их порядковыми номерами (рангами) от меньшего к большему:

11 → ранг 1
12 → ранг 2
14 → ранг 3
15 → ранг 4
105 → ранг 5

Наш новый ранговый ряд: 2, 5, 4, 1, 3. Обратите внимание: экстремальное число 105 превратилось в обычный ранг 5. Оно больше не имеет разрушительной силы для математического анализа!

Переходя от сырых данных к рангам, мы теряем часть информации (мы больше не знаем точное расстояние между объектами), но взамен получаем абсолютную устойчивость к выбросам и независимость от формы распределения. Вместо гипотез о равенстве средних значений непараметрические тесты проверяют гипотезы о равенстве медиан или о равенстве распределений групп.

Давайте структурируем основные непараметрические тесты в зависимости от дизайна вашего исследования. Для каждого классического параметрического теста существует свой непараметрический аналог (табл. 10).

Таблица 10

Параметрические и непараметрические тесты в зависимости от дизайна исследования

Дизайн исследования	Параметрический тест	Непараметрический аналог
Одна группа (сравнение с константой)	Одновыборочный *t*-критерий	Критерий знаков, Критерий знаковых рангов Уилкоксона
Две независимые группы	Двухвыборочный *t*-критерий	Критерий Манна-Уитни (U-тест)
Две зависимые группы (До — После)	Парный *t*-критерий	Критерий знаковых рангов Уилкоксона
Три и более независимых групп	Однофакторный ANOVA	Критерий Краскела-Уоллиса (H-тест)
Три и более зависимых групп	Repeated Measures ANOVA	Критерий Фридмана

Разберем логику главных критериев подробнее.

1. ***Две независимые группы: Критерий Манна-Уитни*** (U-тест). Используется, когда мы сравниваем две изолированные группы (например, время заживления раны при использовании мази А и мази Б).

Логика теста: Все данные из обеих групп сливаются в один общий массив и ранжируются от 1 до $n_1 + n_2$. Затем ранги разделяются обратно по группам, и подсчитывается сумма рангов в каждой группе. Если распределения групп одинаковы (верна H_0), то высокие и низкие ранги перемешаются равномерно, и суммы рангов в группах будут примерно равны. Если же в одной группе скопились преимущественно высокие ранги, а в другой — низкие, критерий зафиксирует статистически значимые различия.

2. ***Две зависимые группы: Критерий знаковых рангов Уилкоксона***. Применяется для дизайна «До — После» на одних и тех же объектах (например, уровень тревожности участников до тренинга и после него).

Логика теста: для каждого участника вычисляется разность показателей ($d = X_{\text{после}} - X_{\text{до}}$). Нулевые разности отбрасываются. Затем все получившиеся разности ранжируются по их абсолютному значению (без учета знака). После этого рангам присваиваются знаки исходных разностей (плюс или минус). Критерий сравнивает сумму положительных рангов со суммой отрицательных рангов. Если показатели значимо улучшились, то положительные сдвиги получат гораздо более высокие ранги, чем редкие отрицательные флуктуации.

3. ***Три и более независимых групп: Критерий Краскела–Уоллиса*** (*H*-тест). Это непараметрический аналог дисперсионного анализа (ANOVA). Он проверяет нулевую гипотезу о том, что все сравниваемые группы происходят из одного и того же распределения.

Логика теста: расширяет идею теста Манна-Уитни. Данные всех k групп объединяются, ранжируются, а затем вычисляется статистика H, оценивающая степень расхождения между средними рангами групп. Если значение H превышает критический порог (который аппроксимируется распределением Хи-квадрат), нулевая гипотеза отвергается. Важно: как и ANOVA, критерий Краскела-Уоллиса — это общий (омнибусный) тест. Он говорит лишь о том, что какая-то из групп отличается от остальных.

Чтобы понять, какие именно группы различаются, после него необходимо проводить непараметрические пост-хок тесты (например, критерий Данна с поправкой Бонферрони).

Подводя итог, давайте честно взвесим все «за» и «против», чтобы вы могли осознанно выбирать метод в своих будущих проектах.

Плюсы (Преимущества):

- Универсальность: работают на данных любого распределения.
- Устойчивость: на них не влияют экстремальные выбросы.
- Применимость к качественным данным: идеальны для ранговых, порядковых шкал и балльных оценок.
- Эффективность на малых выборках: Незаменимы, когда данных катастрофически мало.

Минусы (Ограничения):

- Меньшая статистическая мощность: если ваши данные на самом деле распределены нормально, а вы применили непараметрический тест, ему понадобится больший объем выборки, чтобы обнаружить тот же самый эффект, который *t*-критерий Стьюдента заметил бы сразу. Непараметрика более консервативна.
- Смена фокуса гипотез: они не позволяют делать выводы о средних величинах. Вы докажете, что «распределения значимо различаются», но не сможете красиво заявить инвесторам или руководству, что «наш новый продукт увеличил среднюю прибыль именно на 15.4%».
- Сложность при многофакторном анализе: Непараметрические аналоги для сложных многофакторных планов (например, двухфакторного ANOVA с взаимодействием факторов) математически очень громоздки и имеют жесткие ограничения.

7.5. Сравнение групп: краткие итоги

- Сравнение средних в двух независимых группах — базовая задача проверки гипотез, которая решается с помощью *t*-критерия Стьюдента.
- Критерий работает как весы: он сопоставляет чистый эффект (разницу средних) с уровнем случайного шума в данных (стандартной ошибкой).

- Перед запуском теста обязательно проверяйте данные на нормальность и равенство дисперсий (тест Левена). При нарушении равенства дисперсий всегда переключайтесь на тест Уэлча.
- Статистическая значимость ($p < 0{,}05$) доказывает, что разница существует, но не забывайте оценивать практическую значимость (величину эффекта, например, через коэффициент *d* Коэна), чтобы понять масштаб этой разницы для вашего бизнеса или науки.
- Парный *t*-критерий Стьюдента используется, когда два измерения проводятся на одних и тех же объектах (дизайн «До — После») или на тесно связанных парах. Главная «фишка» критерия — он уничтожает межиндивидуальный шум, сжимая две выборки в одну выборку индивидуальных разностей. Тест проверяет, отличается ли среднее значение этих разностей от нуля.
- Главное допущение, которое необходимо проверять перед запуском теста — это нормальность распределения получившихся разностей, а не исходных массивов.
- Дисперсионный анализ (ANOVA) применяется, когда нам необходимо сравнить средние значения в трех и более независимых группах одновременно.
- Использовать серию обычных *t*-тестов Стьюдента напрямую нельзя из-за лавинообразного роста риска совершить ложноположительную ошибку (проблема множественных сравнений).
- Логика ANOVA заключается в сравнении межгрупповой дисперсии (сигнал эффекта) и внутригрупповой дисперсии (случайный шум) с помощью *F*-критерия Фишера.
- Обязательными условиями для проведения ANOVA являются независимость данных, их нормальность внутри групп и равенство дисперсий (гомоскедастичность).
- Если ANOVA показал значимый результат, исследование не закончено. Чтобы понять, какие именно группы отличаются друг от друга, необходимо провести дополнительное попарное сравнение с помощью Пост-хок тестов (например, теста Тьюки).

- Непараметрические тесты — это надежная «подушка безопасности» исследователя, когда нарушены условия нормальности данных или выборки слишком малы.
- В основе большинства таких методов лежит замена реальных значений их рангами (позициями в общем упорядоченном ряду).
- При сравнении двух независимых групп используется критерий Манна-Уитни, для зависимых — критерий Уилкоксона, а для множества групп — критерий Краскела-Уоллиса.
- Помните правило: если данные хорошие и нормальные — используйте параметрику (она мощнее). Если данные «грязные», порядковые или выборки крошечные — смело переходите на непараметрические критерии.

Глава 8. Анализ выживаемости, рисков и шансов

8.1. Цензурированные наблюдения

Существует огромный класс задач, где исследователя интересует не просто вопрос «Произойдет ли событие?», а вопрос «Когда именно оно произойдет?». Исторически этот подход зародился в медицине и демографии, где изучалось время от момента постановки диагноза или операции до смерти пациента. Отсюда и название — анализ выживаемости (Survival Analysis). Сегодня сфера применения этого метода простирается далеко за пределы медицины: медицина и биология – время от начала лечения до рецидива болезни; время до заживления раны; бизнес и маркетинг (Анализ оттока / Churn Analysis – время от момента регистрации пользователя в приложении до расторжения контракта (жизненный цикл клиента — Customer Lifetime); управление кадрами (HR) – время от приема сотрудника на работу до его увольнения; техника и надежность – время от запуска двигателя до его первой поломки (наработка на отказ).

Почему мы не можем применить обычную линейную регрессию, чтобы предсказать время до поломки или ухода клиента? Нам мешают две фундаментальные проблемы:

1. Распределение времени обычно сильно скошено (оно не нормальное), имеет длинный правый хвост и не может быть отрицательным.

2. Проблема цензурирования данных. К моменту окончания нашего наблюдения далеко не у всех объектов событие успевает произойти.

Главная отличительная черта анализа выживаемости — способность корректно работать с «незавершенными» наблюдениями. Такие данные называются ***цензурированными*** (censored data).

Существует несколько видов цензурирования, но самым частым в практике является ***правостороннее цензурирование***. Представьте, что мы проводим 5-летнее исследование эффективности нового лекарства. В нашей выборке есть три типа пациентов:

Пациент А: заболел, лечился, и, к сожалению, умер на 3-й год исследования. Здесь мы знаем точное время до события ($t = 3$). Это нецензурированное наблюдение.

Пациент Б: лечился, все было хорошо, но на 4-й год он переехал в другую страну и связь с ним прервалась. Мы знаем, что он прожил минимум 4 года, но точное время его смерти нам неизвестно.

Пациент В: прошел все 5 лет исследования и остался жив на момент закрытия проекта. Мы знаем, что он выжил в течение минимум 5 лет.

Пациенты Б и В являются цензурированными справа.

Если мы просто выбросим пациентов Б и В из анализа, мы сильно занизим реальные показатели выживаемости (ведь мы избавимся от тех, у кого все было хорошо). Если мы зафиксируем их время как ровно 4 и 5 лет, мы совершим ошибку, приняв, что событие произошло именно в день окончания наблюдения. Анализ выживаемости учитывает их особым математическим способом: он говорит модели, что данные объекты «внесли вклад в безопасный стаж длиною в T лет, после чего исчезли из поля зрения».

8.2. Функции, изучаемые в анализе выживаемости

Математический аппарат анализа выживаемости базируется на двух взаимосвязанных функциях:

А. Функция выживаемости — Survival Function ($S(t)$). Показывает вероятность того, что объект «выживет» (событие не наступит) дольше определенного момента времени t. В начальный момент времени ($t = 0$) все объекты живы, поэтому $S(0) = 1$ (или 100%).

Со временем событие постепенно происходит у разных объектов, поэтому функция $S(t)$ монотонно убывает и стремится к 0 при $t) \rightarrow \infty$.

Б. Функция риска — Hazard Function ($h(t)$). Показывает мгновенную интенсивность наступления события в конкретный момент времени t при условии, что до этого момента объект успешно выжил.

Пример: для прыжков с парашютом функция риска $h(t)$ максимальна в моменты взлета и приземления. В страховании жизни функция риска человека растет с его возрастом. В технике функция риска часто имеет «U-образную» форму (высокий риск из-за заводского брака в начале, затем долгий период стабильности и резкий рост риска из-за износа в конце).

8.3. Метод Каплана-Майера

Для оценки функции выживаемости $S(t)$ на основе реальных данных без каких-либо предположений о форме распределения используется непараметрический метод Каплана-Майера.

Построение кривой выживаемости основано на правиле умножения вероятностей. Мы рассчитываем вероятность выжить в каждый конкретный шаг времени, когда происходят события, учитывая только тех, кто реально дожил до этого момента (рискует прямо сейчас). Формула Каплана-Майера выглядит так:

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i})$$

где: t_i — момент времени, в который произошло хотя бы одно событие; d_i — количество событий (смертей / уходов), произошедших в момент t_i; n_i — количество объектов, находившихся под риском непосредственно перед моментом t_i (то есть те, кто еще жив и не успел отцензурироваться до этого шага).

График Каплана-Майера представляет собой ступенчатую линию, идущую вниз от 100%. Каждый шаг вниз — это момент наступления события. Маленькие вертикальные штрихи или точки на ступенях обычно обозначают моменты, когда пациенты были отцензурированы.

Важной метрикой, извлекаемой из этого графика, является медиана выживаемости — это точка на оси времени, в которой

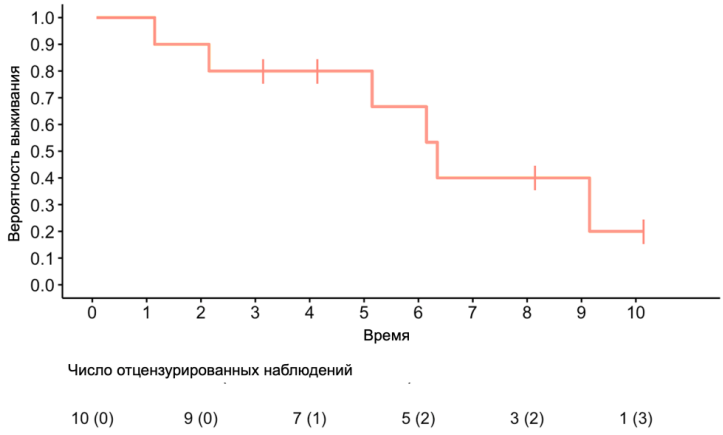


Рис. 20. График Каплана-Майера

кривая пересекает уровень 0,5 (50%). Она показывает время, к которому ровно половина объектов в выборке столкнулась с событием.

8.4. Сравнение выживаемости в группах

Очень часто перед аналитиком стоит задача сравнить выживаемость в двух или более группах. Например: «Правда ли, что пациенты с новым типом терапии (Группа А) живут дольше, чем пациенты на старой терапии (Группа Б)?» или «Увольняются ли участковые терапевты быстрее, чем нефрологи?».

Визуально на графике мы увидим две разные ступенчатые кривые Каплана-Майера. Но как доказать, что расстояние между ними статистически значимо, а не вызвано случайностью выборки? Для этого применяется ***Лог-ранговый критерий (Log-Rank Test)***.

Логика критерия: это непараметрический тест, который проверяет нулевую гипотезу (H_0) о том, что между кривыми выживаемости в популяциях нет никаких различий. Тест анализирует каждый временной шаг, где произошло хотя бы одно событие. Для каждого шага, зная пропорции участников в обеих группах под риском, математика рассчитывает ожидаемое число смертей в каждой группе, если бы гипотеза об их равенстве была верна.

Затем тест суммирует отклонения реальных наблюдаемых смертей (O) от теоретических ожидаемых (E) по всей временной шкале. Статистика критерия аппроксимируется распределением Хи-квадрат. Если полученное p-value < 0,05, мы отвергаем нулевую гипотезу и заявляем: кривые выживаемости групп статистически значимо различаются.

8.5. Модель пропорциональных рисков Кокса

Лог-ранговый критерий прекрасен, но, как и ANOVA, он позволяет сравнивать группы только по одному категориальному признаку. А что делать, если мы хотим оценить влияние сразу множества факторов на время выживаемости? Например, учесть не только тип терапии, но и возраст пациента, его вес, стаж болезни и результаты анализов?

Для многофакторного анализа выживаемости используется ***регрессионная модель Кокса*** (Cox Proportional Hazards Model).

Это полупараметрический метод, который моделирует функцию риска $h(t)$ следующим образом:

$$h(t, X) = h_0(t) \times exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

где:

$h_0(t)$ — базовый риск (baseline hazard). Это риск объекта, у которого все факторы x равны нулю. Кокс изящно разрешил не специфицировать форму этой функции, что сделало модель невероятно гибкой.

$exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$— экспоненциальная мультипликативная часть, которая увеличивает или уменьшает базовый риск в зависимости от характеристик объекта.

8.6. Hazard Ratio и Odds Ratio

Необходимо научиться четко разделять и корректно интерпретировать два важнейших статистических показателя — отношение рисков (Hazard Ratio, HR) и отношение шансов (Odds Ratio, OR), понять их математическую природу и области применения.

В доказательной медицине, эпидемиологии, маркетинге и биостатистике исследователи постоянно ищут способы измерить, насколько сильнее один фактор влияет на исход по сравнению с другим. Для этого человечество придумало целый арсенал относительных показателей. Два самых популярных, но при этом наиболее часто путаемых коэффициента — это Odds Ratio (OR — ***отношение шансов***) и Hazard Ratio (HR — ***отношение рисков***).

В научной литературе, а уж тем более в статьях бизнес-аналитиков, можно встретить катастрофическую ошибку, когда и OR, и HR переводят на обывательский язык одинаково: «Шансы/риски группы А во столько-то раз выше, чем группы Б». Однако с математической точки зрения это совершенно разные метрики. Ошибка в их интерпретации может привести к ложным выводам — например, к сильному завышению эффективности лекарства или к неверному прогнозу оттока клиентов.

Начнем с Odds Ratio (OR). Этот показатель неразрывно связан с классической логистической регрессией и анализом таблиц сопряженности. Но, прежде чем говорить об отношении шансов, нужно понять, что такое сам шанс (Odds). В обычном языке мы

часто используем слова «вероятность» и «шанс» как синонимы. В статистике это табу.

Вероятность (p) — это отношение количества благоприятных исходов к общему количеству возможных исходов. Измеряется от 0 до 1.

Шанс (Odds) — это отношение количества благоприятных исходов к количеству неблагоприятных исходов. Или, выражаясь через вероятность:

$$Odds = \frac{p}{1-p}$$

Пример: если из 100 пациентов 20 человек столкнулись с побочным эффектом, то:

Вероятность побочного эффекта: p = 20 / 100 = 0,2 (или 20%).

Шанс получить побочный эффект: Odds = 20 / 80 = 0,25 (то есть 1 к 4).

Отношение шансов (OR) — это просто шанс из первой группы (например, экспериментальной), деленный на шанс из второй группы (контрольной):

$$OR = \frac{Odds_1}{Odds_2}$$

Посмотрите внимательно на формулу. В ней есть только количество случившихся событий и общее число участников. Отношение шансов абсолютно «слепо» ко времени. Оно отвечает на вопрос: «Каковы были шансы получить исход к моменту фиксации результатов исследования?».

Именно поэтому OR является главным критерием в исследованиях типа «случай-контроль» (case-control studies), когда событие уже произошло в прошлом, и мы просто смотрим ретроспективно, накоплен ли у пациентов определенный признак.

Теперь перейдем к ***отношению рисков (HR).*** Этот коэффициент живет исключительно в мире анализа выживаемости и модели регрессии пропорциональных рисков Кокса.

В отличие от OR, базовым понятием здесь является не шанс, а функция риска (Hazard function, $h(t)$). Риск в анализе выживаемости — это не просто вероятность. Это мгновенная скорость (интенсивность) наступления события в конкретный узкий момент времени t, при условии, что до этого момента объект успешно «выжил».

Hazard Ratio (HR) — это отношение интенсивности наступления события в первой группе к интенсивности во второй группе на протяжении всего времени наблюдения:

$$HR = \frac{h_1(t)}{h_2(t)}$$

В модели Кокса коэффициенты β_i интерпретируются через экспоненту —$e^{\beta i}$, которая называется отношением рисков (Hazard Ratio, HR). Если HR = 1: фактор никак не влияет на риск наступления события.

Если HR > 1: фактор является вредным, он увеличивает риск наступления события (и, соответственно, уменьшает время выживаемости). Например, если для фактора «Курение» HR = 1,8, это значит, что у курящих пациентов интенсивность наступления негативного события в любой момент времени в 1,8 раза выше, чем у некурящих.

Если HR < 1: фактор является защитным, он снижает риск. Например, если для нового лекарства HR = 0,6, это означает, что препарат снижает риск наступления исхода на 40% (1 – 0,6 = 0,4) в каждый момент времени.

HR учитывает не только то, сколько человек столкнулось с событием, но и то, когда именно это произошло. Он анализирует всю динамику процесса.

Рассмотрим пример. Допустим, мы регулярно наблюдаем пациентов с артериальной гипертензией и рекомендуем им самостоятельно измерять АД дома, строго придерживаясь методики «722» (семь дней по два измерения 2 раза в день) и тестируем два подхода: в группе А даём рекомендацию записывать свои данные в свой блокнот, группу Б снабжаем бланками для записи, снабжёнными QR-кодом, в котором дополнительно содержатся обучающие материалы. Через 6 мес в обеих группах перестали соблюдать правила измерения АД по 50% пользователей. Если мы посчитаем Odds Ratio (OR) в конце года, то шансы окажутся равны, и OR = 1. Мы решим, что подходы одинаковы.

Но что, если в группе А все 50% пользователей перестали соблюдать методику в первую же неделю, а в группе Б пользователи уходили равномерно в течение всего периода? Для врача группа Б намного ценнее, ведь пациенты соблюдали рекомендации дольше! Hazard Ratio (HR) зафиксирует эту разницу. В первые недели риск ухода в группе А будет колоссальным, и

итоговый HR будет значительно больше 1, показывая, что подход в группе А хуже.

Чтобы окончательно зафиксировать материал, давайте сопоставим OR и HR по ключевым критериям.

1. ***Фактор времени***:

OR: статичен. Фиксирует срез данных на определенный момент времени. Ему неважно, произошло событие на 1-й день или на 300-й день исследования.

HR: динамичен. Учитывает точное время до наступления события и статус цензурированных данных (тех, у кого событие еще не произошло).

2. ***Тип исследования***:

OR: идеален для ретроспективных исследований («случай-контроль»), где временная шкала уже недоступна.

HR: применяется только в проспективных (когортных) исследованиях и клинических испытаниях, где за объектами следят во времени.

3. ***Математическая взаимосвязь***:

Если событие в популяции происходит очень редко (менее 5–10%), значения OR и HR будут практически одинаковыми.

Если событие частое, OR имеет свойство сильно преувеличивать (раздувать) реальный эффект по сравнению с HR.

8.7. Метаанализ и форест-плот

Представьте себе классическую научную ситуацию. Одно клиническое исследование в Германии утверждает: «Новое лекарство от давления снижает риск инфаркта на 10%». Второе исследование, проведенное в США на меньшей выборке, говорит: «Эффект не обнаружен, лекарство бесполезно». Третье исследование в Японии заявляет: «Риск снижается, но на 25%». Кому верить практикующему врачу или министру здравоохранения? Каждое отдельное исследование имеет свои ограничения: малый объем выборки, специфику популяции или случайные погрешности.

Чтобы решить это противоречие, в 1970-х годах ученые разработали метод, ставший сегодня золотым стандартом доказательной науки —метаанализ.

Метаанализ — это статистический метод, который объединяет результаты нескольких независимых исследований, посвя-

щенных одному и тому же вопросу, для получения обобщенной, гораздо более точной и математически обоснованной оценки эффекта.

Почему метаанализ сильнее одиночного исследования?

1. Рост статистической мощности: объединяя 10 исследований по 100 человек, мы получаем виртуальную выборку в 1000 человек. На таком массиве гораздо проще заметить истинную закономерность и отсечь случайный шум.

2. Разрешение конфликтов: метаанализ математически примиряет исследования с противоположными результатами.

3. Объективность: вместо субъективного обзора литературы («автор А считает так, а автор Б — иначе») метаанализ дает точную цифру с доверительным интервалом.

Метаанализ обычно проводится в рамках систематического обзора — строгого, задокументированного поиска абсолютно всех существующих статей по теме, выполненного по правилам PRISMA, чтобы исключить предвзятость автора.

Просто сложить результаты всех исследований и посчитать среднее арифметическое нельзя. Исследования имеют разный вес: крупная работа на 5000 пациентов с жестким контролем качества заслуживает большего доверия, чем маленькая статья на 20 человек.

Поэтому метаанализ рассчитывает взвешенное среднее. Вес каждого исследования чаще всего обратно пропорционален его дисперсии (разбросу данных). Проще говоря: чем больше выборка и чем точнее результат исследования, тем больший вес оно получает в финальном расчете.

При объединении данных ученые выбирают одну из двух математических моделей:

1. ***Модель фиксированных эффектов*** предполагает, что во всех исследованиях существует один-единственный истинный размер эффекта, а все различия между результатами статей вызваны исключительно случайной ошибкой выборки. Эта модель используется редко, только если все эксперименты проводились в абсолютно идентичных условиях.

2. ***Модель случайных эффектов*** более реалистична. Она допускает, что истинный эффект может слегка отличаться от исследования к исследованию (например, из-за разного возраста пациентов, дозировок или климата). Модель случайных эффектов распределяет веса более равномерно, не позволяя одному

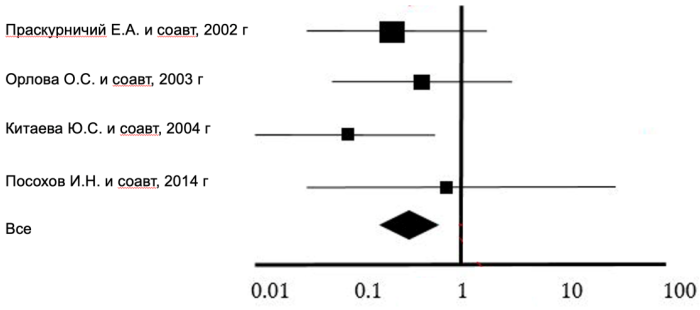


Рис. 21. Форест-плот

гигантскому исследованию полностью подавить результаты мелких, но важных работ.

Результаты метаанализа визуализируются на одном из самых красивых и информативных графиков в статистике — ***форест-плоте*** (или лесной диаграмме). Свое название он получил потому, что обилие горизонтальных линий напоминает чащу деревьев (рисунок 21).

Давайте пошагово разберем анатомию этого графика.

Элемент 1. ***Список исследований*** (левая колонка). Здесь перечислены авторы и год публикации каждого исследования, вошедшего в метаанализ (например, Праскурничий Е.А. et al., 2002, Орлова О.С. et al., 2021). Обычно они расположены в хронологическом порядке.

Элемент 2. ***Вертикальная линия*** («линия отсутствия эффекта») — это важнейший ориентир графика, проходящий через нулевую точку эффекта. Если анализируется разница средних, линия стоит на отметке 0. Если анализируются относительные метрики (Odds Ratio, Hazard Ratio, Relative Risk), линия стоит на отметке 1 (так как отношение 1 означает, что риски в двух группах абсолютно одинаковы).

Элемент 3. ***Квадраты*** (точечные оценки эффекта). Напротив каждого исследования нарисован квадрат. Центр квадрата — это точный результат, который получился в данной конкретной статье (размер эффекта). Внимание на размер квадрата! Площадь квадрата прямо пропорциональна весу этого исследования в метаанализе. Чем больше и надежнее было исследование, тем крупнее квадрат.

Элемент 4. ***Горизонтальные усы*** (доверительные интервалы). Через каждый квадрат проходит горизонтальная линия. Это 95%-й доверительный интервал (CI). Он показывает границы, в которых с вероятностью 95% лежит истинное значение эффекта для данного исследования. Чем длиннее «усы», тем меньше была выборка и тем менее точен результат исследования.

Золотое правило чтения: Если горизонтальный «ус» исследования пересекает центральную вертикальную линию отсутствия эффекта (0 или 1), это значит, что результаты данной конкретной статьи статистически незначимы ($p > 0,05$).

Элемент 5. ***Ромб*** (финальный итог). В самом низу графика, под всеми исследованиями, располагается ромб (иногда его называют алмазом — summary diamond). Это и есть главный результат всего метаанализа. Центр ромба (его самая широкая часть по вертикали) показывает объединенный размер эффекта. Ширина ромба по горизонтали — это объединенный 95%-й доверительный интервал для всего метаанализа.

Главный вердикт: если ромб целиком лежит по одну сторону от вертикальной линии и не касается её своими острыми боковыми углами, метаанализ доказал общую статистическую значимость эффекта! Если ромб пересекает линию — объединение данных показало, что исследуемый фактор не работает.

Перед тем как безоговорочно поверить финальному ромбу, исследователь обязан задать вопрос: «А можно ли было вообще объединять эти статьи? Насколько они похожи друг на друга?». Если мы попытаемся объединить в один метаанализ «яблоки и апельсины» (например, принципиально разные методики лечения), финальный результат будет ложным.

Разброс результатов между исследованиями называется гетерогенностью (неоднородностью). Внизу форест-плота всегда пишется текстовая строчка с оценкой гетерогенности. Мы смотрим на два показателя:

1. ***Q-критерий Кокрейна*** (p-value): проверяет нулевую гипотезу о том, что все исследования однородны. Если $p < 0,05$, мы отвергаем однородность — данные значимо различаются между собой.

2. ***Индекс I^2*** (Индекс неоднородности): измеряется в процентах и показывает, какая доля вариации между статьями вызвана реальными различиями, а не случайным шумом.

$I^2 < 25\%$ — неоднородность слабая (исследования очень похожи, объединение идеально).

I^2 от 25% до 50% — умеренная неоднородность.

$I^2 > 50\%$ — высокая неоднородность (исследования сильно противоречат друг другу). В этом случае ученые обязаны провести анализ в подгруппах (subgroup analysis), чтобы понять, почему результаты так сильно «гуляют» (например, разделить пациентов по возрасту или полу).

Давайте закрепим материал. Представим, что перед нами форест-плот метаанализа эффективности нового антибиотика против старого стандарта. Исход — отношение шансов выздороветь (Odds Ratio). Линия отсутствия эффекта проходит через 1. Все, что правее 1 — говорит в пользу нового антибиотика.

Проводим анализ:

1. Исследование Smith (2015): Квадрат маленький, «усы» очень длинные и пересекают линию 1. Вывод: исследование маленькое, результат незначим.

2. Исследование Green (2018): Квадрат огромный (большой вес), доверительный интервал узкий и находится целиком правее линии 1 (от 1,2 до 1.8). Вывод: крупное надежное исследование, доказавшее высокую эффективность нового препарата.

3. Финальный ромб внизу: находится в диапазоне от 1,3 до 1,6. Он не пересекает линию 1.

4. Строка гетерогенности: $I^2 = 12\%$.

Общий вывод: Метаанализ успешно объединил данные. Неоднородность исследований крайне низкая ($I^2 = 12\%$), данные согласуются между собой. Объединенный результат статистически значим. Новый антибиотик доказал свою эффективность: шансы на выздоровление при его использовании в среднем в 1,45 раза выше по сравнению со старым стандартом.

8.8. Анализ выживаемости, рисков и шансов: краткие итоги

- Анализ выживаемости работает со специфическими данными — временем до наступления события, где ключевой проблемой является правостороннее цензурирование (когда событие у объекта еще не произошло на момент окончания наблюдения).

- Непараметрический метод Каплана-Майера строит ступенчатую кривую выживаемости и позволяет легко находить медиану времени до события.
- Для попарного сравнения нескольких кривых выживаемости используется лог-ранговый критерий.
- Если необходимо оценить одновременное влияние количественных и качественных факторов на выживаемость, применяется регрессия Кокса, результаты которой интерпретируются через Отношение Рисков (Hazard Ratio).
- Odds Ratio (OR) — это отношение шансов. Метрика «слепа» ко времени, используется для фиксированных исходов (логистическая регрессия).
- Hazard Ratio (HR) — это отношение рисков (интенсивностей). Метрика полностью завязана на время и скорость наступления события (анализ выживаемости).
- Никогда не заменяйте HR на OR, если у вас есть данные о времени до наступления события, иначе вы потеряете огромный пласт ценной информации о динамике процесса.
- Метаанализ — это форма аналитики, объединяющая результаты множества независимых исследований для поиска объективной истины.
- Главный инструмент визуализации метаанализа — форест-плот. При чтении форест-плота всегда ищите вертикальную линию отсутствия эффекта (0 для разностей, 1 для отношений). Положение и ширина финального ромба внизу графика мгновенно говорят о том, доказан ли общий эффект и насколько точен наш финальный вывод.
- Всегда проверяйте индекс неоднородности I^2: если он выше 50%, объединять данные «в лоб» нельзя, требуется глубокий поиск причин различий.

Глава 9. Диагностическая точность

9.1. Чувствительность и специфичность

Каждый день мы сталкиваемся с необходимостью принимать бинарные решения: «да» или «нет». Например, болен пациент конкретным заболеванием или здоров? Чтобы принять такое решение, мы используем различные инструменты проверки: медицинские анализы, ПЦР-тесты, алгоритмы машинного обучения, обобщить такие инструменты можно словом «тест».

Идеальный тест в идеальном мире никогда не ошибается. Однако в реальности ***любой тест несовершенен***. Больного человека он может назвать здоровым, а у здорового — обнаружить несуществующую патологию. Чтобы измерить цену ошибки и понять, насколько можно доверять результатам проверки, в статистике разработана система метрик, главными из которых являются чувствительность и специфичность.

Чтобы рассчитать характеристики любого теста, нам необходим ориентир — «золотой стандарт» (наиболее точный, экспертный метод подтверждения реальности, например, биопсия или долгое консилиумное наблюдение). Мы сопоставляем результаты нашего быстрого теста с истинным статусом из «золотого стандарта» и сводим данные в таблицу сопряженности 2 2, которую в анализе данных называют матрицей ошибок.

Таблица 11

Матрица ошибок

Результат теста	**Есть заболевание (Истина +)**	**Нет заболевания (Истина -)**
Тест Положительный (+)	Истинно положительный (*TP*)	Ложноположительный (Ошибка I рода) (*FP*)
Тест Отрицательный (-)	Ложноотрицательный (Ошибка II рода) (*FP*)	Истинно отрицательный (*TN*)

Давайте разберем эти четыре исхода на простом примере металлоискателя на рамке безопасности:

1. TP (Истинно положительный). У человека есть нож, рамка зазвенела. Тест сработал верно.

2. TN (Истинно отрицательный). У человека нет запрещенных предметов, рамка молчит. Тест сработал верно.

3. FP (Ложноположительный). У человека нет оружия (например, просто крупная бляха на ремне), но рамка зазвенела. Это ложная тревога.

4. FN (Ложноотрицательный). У человека есть нож, но он его спрятал, и рамка промолчала. Это самая опасная ошибка — пропуск цели.

Чувствительность и специфичность оценивают тест «вертикально» по столбцам таблицы. Они показывают, как тест ведет себя на заведомо известных группах больных и здоровых людей.

Чувствительность — это доля истинно положительных результатов теста среди всех тех, у кого действительно есть заболевание. Формула:

$$\text{Чувствительность} = \frac{TP}{TP + FN}$$

Чувствительность отвечает на вопрос: «Какова вероятность того, что тест обнаружит болезнь, если человек действительно болен?». Тест с чувствительностью 95% обнаружит 95 больных из 100, а 5 человек пропустит (выдаст им ложноотрицательный результат).

Когда нужен высокочувствительный тест? На этапе скрининга (первичного отбора). Например, при донорстве крови тест на ВИЧ должен обладать почти 100%-й чувствительностью. Пусть лучше он выдаст несколько ложных тревог, но мы гарантированно не пропустим ни одной капли зараженной крови.

Специфичность — это доля истинно отрицательных результатов теста среди всех тех, у кого заболевания гарантированно нет.

$$\text{Специфичность} = \frac{TN}{TN + FP}$$

Специфичность отвечает на вопрос: «Какова вероятность того, что тест промолчит, если человек абсолютно здоров?». Тест со специфичностью 90% правильно назовет здоровыми 90 человек из 100, а 10 здоровым людям ошибочно припишет диагноз (ложноположительный результат). Когда нужен высокоспецифичный тест? На этапе подтверждения диагноза, особенно если последующее лечение связано с тяжелыми рисками (например, химиотерапия или сложная операция). Нам нужна абсолютная уверенность, что мы не лечим здорового человека.

Представьте себя на месте пациента. Вы сдали анализ, и вам пришел бумажный бланк, на котором написано: «Тест положительный». Вам в этот момент не так важна абстрактная чувствительность теста на этапе его разработки. Вас интересует один шкурный вопрос: «Какова вероятность, что я действительно

болен, раз уж тест выдал плюс?». На этот вопрос отвечают метрики прогностической ценности. Они анализируют таблицу «горизонтально» по строкам.

Прогностическая ценность положительного результата (PPV — Positive Predictive Value / Precision) — это доля истинно больных людей среди всех тех, у кого тест оказался положительным. Формула:

$$PVV = \frac{TP}{TP + FP}$$

Прогностическая ценность отрицательного результата (NPV — Negative Predictive Value) — это доля истинно здоровых людей среди всех тех, у кого тест оказался отрицательным. Формула:

$$NPV = \frac{TN}{TN + FN}$$

Главным парадоксом диагностики является влияние распространенности. Чувствительность и специфичность — это постоянные характеристики самого лабораторного оборудования. Но прогностическая ценность (PPV) драматически зависит от того, насколько часто заболевание вообще встречается в обществе (от распространенности).

Если заболевание очень редкое, то даже при использовании великолепного теста с чувствительностью 99% и специфичностью 99%, большинство положительных результатов будут ложными тревогами. Пример: Мы тестируем 10 000 человек на суперредкую болезнь, которая встречается у 1 человека из 1000 (всего 10 больных в нашей выборке). Наш тест идеален: чувствительность 99%, специфичность 99%. Тест найдет почти всех больных: TP = 10. Но из 9990 здоровых людей тест из-за 1% погрешности выдаст ложную тревогу для 1% участников: FP =100 человек!

Считаем PPV: 10/(10 + 100) = 9%.

Посмотрите на этот шокирующий результат! Пациент получил пугающий «положительный» результат теста с надежностью 99%, но реальная вероятность того, что он действительно болен, составляет всего 9%! Оставшиеся 91% — это здоровые люди, ставшие жертвами ложной тревоги. Именно поэтому массовый скрининг редких болезней повально среди всех людей без симптомов проводить нельзя.

Если нам нужно оценить качество теста одним числом, используется комбинированные показатели:

1. ***Общая точность*** (Accuracy): Доля всех правильных ответов теста (и плюсов, и минусов) ко всем проведенным тестам:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Критический минус: Accuracy бесполезна на несбалансированных выборках (если у нас 99% здоровых, то алгоритм, который всегда говорит «здоров», будет иметь Accuracy 99%, но пропустит 100% больных).

2. F1-мера (F1-score), которая используется в машинном обучении. Это среднее гармоническое между чувствительностью (Recall) и точностью (Precision):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

9.2. ROC-кривая и показатель AUC

Мы разобрали такие метрики, как чувствительность и специфичность. Мы научились рассчитывать их на основе матрицы ошибок. Однако у матрицы ошибок есть один фундаментальный минус: она строится для уже принятого жесткого решения. Большинство современных алгоритмов классификации (будь то логистическая регрессия, нейросети или скоринговые системы банков) на выходе выдают не просто бинарный ответ «Да/Нет», а вероятность (число от 0 до 1). Например, алгоритм говорит: «Вероятность того, что эта транзакция является мошеннической, составляет 0,72». Чтобы превратить эту вероятность в финальное действие (заблокировать карту или одобрить операцию), аналитик должен задать порог отсечения (cut-off / threshold).

По умолчанию в программах часто стоит порог 0,5. Если вероятность больше 0,5, мы говорим: «Да», если меньше — «Нет». Но справедлив ли порог 0,5 для всех задач? Конечно, нет. Если мы оцениваем риск редкого, но смертельного заболевания, мы захотим снизить порог до 0,1, чтобы не пропустить ни одного больного. Если мы настраиваем спам-фильтр, мы поднимем порог до 0.9, чтобы случайно не удалить важное письмо по работе.

Как же оценить качество самой математической модели в целом, независимо от выбранного порога? Для этого в 1940-х годах во время Второй мировой войны американские радиолокационные инженеры разработали метод, который сегодня стал стандартом в Data Science и аналитике — ROC-кри-

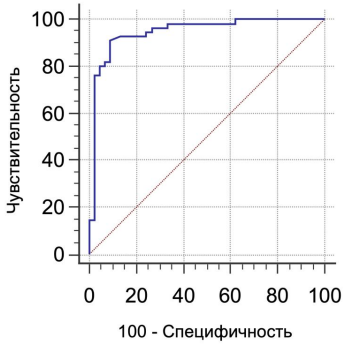


Рис. 22. ROC-кривая

вую и метрику AUC. Аббревиатура ROC расшифровывается как Receiver Operating Characteristic (Рабочая характеристика приемника). Название осталось со времен радиолокации, когда операторы радаров пытались отличить летящий японский самолет от шума птичьих стай.

ROC-кривая строится в двумерной системе координат. Давайте разберем, что откладывается по осям.

Вертикальная ось (Ось *Y*): True Positive Rate (TPR). Это не что иное, как знакомая нам чувствительность (или полнота / Recall). Она показывает долю правильно предсказанных положительных исходов среди всех реальных положительных исходов.

$$TPR = \frac{TP}{TP + FN}$$

Горизонтальная ось (Ось *X*): False Positive Rate (FPR). Это доля ложных тревог. Она показывает, какую часть здоровых людей (или безопасных транзакций) наша модель ошибочно назвала больными (опасными). FPR математически связана со специфичностью:

$$FPR = 1 - \text{Специфичность} = \frac{FP}{TN + FP}$$

Каждый конкретный порог отсечения (например, 0,1, 0,5, 0,9) дает нам свою матрицу ошибок, а значит — одну точку на графике с координатами (FPR, TPR).

Точка (0, 0): мы выставили порог отсечения равным 1,0 (100%). Модель ни для кого не прогнозирует положительный исход. Мы получили 0 ложных тревог (FPR = 0), но и не нашли ни одного больного (TPR = 0).

Точка (1, 1): мы опустили порог до 0.0. Модель абсолютно всех называет больными. Мы нашли 100% больных (TPR = 1), но при этом получили 100% ложных тревог (FPR = 1).

Точка (0, 1) [Левый верхний угол] – это идеальный классификатор. Здесь мы поймали всех больных (TPR = 1) и не совершили ни одной ложной ошибки (FPR = 0). Цель любого аналитика — приблизить кривую модели как можно ближе к этой точке.

Чтобы построить ROC-кривую, алгоритм выполняет простую последовательность действий:

1. Берется тестовая выборка объектов, для которых нам известны реальные ответы.

2. Модель рассчитывает вероятности для каждого объекта.

3. Все объекты сортируются в таблицу по убыванию предсказанной вероятности.

4. Алгоритм начинает двигать порог отсечения от 1,0 до 0,0 шаг за шагом. На каждом шагу рассчитываются текущие значения TPR и FPR, и на график наносится точка.

5. Соединив эти точки, мы получаем плавную (или ступенчатую, если выборка мала) линию, которая идет из левого нижнего угла (0,0) в правый верхний угол (1,1).

Построить кривую — это полдела. Нам нужен численный показатель, который позволит мгновенно сравнить модель А с моделью Б. Этот показатель называется AUC (Area Under Curve), или точнее ROC-AUC — ***площадь под ROC-кривой***. Поскольку весь график ограничен квадратом со сторонами 11, максимальная площадь под кривой может быть равна 1,0, а минимальная — 0,0.

Интерпретация значений AUC:

AUC = 1,0 — Идеальная модель. Модель безошибочно разделяет классы. Существует такой порог, при котором TPR равен 1, а FPR равен 0. Распределения вероятностей для больных и здоровых вообще не пересекаются.

0,9 < AUC < 1,0 — Отличное качество классификации.

0,8 < AUC < 0,9 — Очень хорошее качество.

0,7 < AUC < 0,8 — Среднее (приемлемое) качество.

AUC = 0,5 — Случайное угадывание. На графике это прямая диагональная линия, идущая из угла в угол. Такое качество выдает обычная монетка. Если ваша сложная нейросеть показала AUC = 0,5, значит, она абсолютно бесполезна.

AUC < 0,5 — Модель работает хуже, чем случайность. Однако на практике это означает, что модель просто нужно инвертировать (поменять ответы «Да» и «Нет» местами), и вы получите хорошую модель с AUC > 0,5.

У метрики AUC есть потрясающее математическое определение, которое важно понимать: AUC равна вероятности того, что если мы случайно выберем одного истинно положительного объекта (больного) и одного истинно отрицательного (здорового), то предсказанная моделью вероятность для больного объекта окажется выше, чем для здорового. Таким образом, AUC оценивает способность модели к ранжированию объектов.

Итак, мы построили отличную модель, например, с AUC = 0,88. Кривая построена. Как теперь аналитику выбрать ту самую единственную точку (порог), по которой система будет работать? Выбор порога зависит от контекста и «стоимости» ошибок. Однако существует классический математический способ поиска идеального баланса — ***Индекс Юдена*** (Youden's J statistic). Индекс Юдена максимизирует разницу между чувствительностью и долей ложных тревог для каждой точки на кривой:

$$J = TPR - FPR$$

Геометрически индекс Юдена — это точка на ROC-кривой, которая максимально удалена по вертикали вверх от диагонали случайного угадывания (AUC=0,5).

Когда ROC-AUC начинает лгать? ROC-кривая устойчива к изменению соотношения классов, но только до определенного предела. Если ваши данные критически несбалансированы (например, при поиске редкого мошенничества, где на 1 000 000 безопасных транзакций приходится всего 100 фактов фрода), метрика ROC-AUC покажет обманчиво высокий результат.

Почему так происходит? Из-за того, что в знаменателе формулы FPR стоит гигантское число истинно отрицательных исходов (TN = 1 000 000). Даже если модель совершит 5 000 ложных тревог (что ужасно для бизнеса — мы заблокируем 5 000 честных клиентов), показатель FPR составит всего 0,005 (близко к нулю). ROC-кривая останется идеальной, а значение AUC будет стремиться к 0,99.

Запомните правило: если дисбаланс классов в ваших данных превышает соотношение 1:10, откажитесь от ROC-кривой. Вместо нее стройте PR-кривую (Precision-Recall Curve) и оценивайте площадь под ней (PR-AUC или Average Precision). PR-кривая не использует показатель TN и сфокусирована только на качестве предсказания редкого класса.

9.3. Метод Бланда-Альтмана

В клинической медицине и лабораторной диагностике постоянно возникает задача: сравнить два метода измерения одной и той же непрерывной величины. Представьте сценарии:

У нас есть старый, точный, но очень дорогой и медленный «золотой стандарт» измерения артериального давления (например, ртутный сфигмоманометр) и новый — дешевый, быстрый, автоматический электронный тонометр. Можно ли заменить старый метод новым? Два разных лаборанта на одном и том же оборудовании оценивают уровень гемоглобина в крови у одних и тех же пациентов. Насколько согласуются их результаты? Умные часы считают пульс пользователя. Мы хотим сравнить их показания с данными профессионального медицинского ЭКГ-монитора.

До середины 1980-х годов исследователи массово совершали грубейшую статистическую ошибку. Чтобы доказать, что новый прибор работает так же хорошо, как старый, они рассчитывали коэффициент корреляции Пирсона (r). Если r оказывался близко к 1 (например, $r = 0{,}95$), ученые радостно заявляли: «Методы идеально согласуются, новый прибор можно пускать в производство!».

В 1983 и 1986 годах два британских биостатистика, Мартин Бланд (Martin Bland) и Дуглас Альтман (Douglas Altman), опубликовали в журнале The Lancet статьи, которые совершили революцию. Они доказали, что корреляция измеряет силу линейной связи, но абсолютно не доказывает согласие или взаимозаменяемость методов.

Почему корреляция лжет при сравнении приборов?

1. Коэффициент корреляции оценивает, меняются ли величины синхронно. Если второй прибор всегда, у каждого пациента, завышает давление ровно на 30 мм рт. ст., корреляция между приборами будет всё равно идеальной ($r = 1{,}0$). Но при этом при-

бор абсолютно непригоден для медицины, так как он безбожно врет.

2. Корреляция жестко зависит от диапазона измеряемой величины. Если мы возьмем выборку людей с гигантским разбросом показателей (от новорожденных до тяжелых гипертоников), корреляция будет высокой просто за счет широты диапазона, даже если приборы имеют огромную погрешность.

Поняв это, Бланд и Альтман предложили простой, элегантный графический и математический метод (Bland-Altman Analysis), который сегодня является мировым стандартом.

Логика метода Бланда-Альтмана: вместо того, чтобы строить стандартную диаграмму рассеяния, где по осям откладываются показания первого и второго приборов, Бланд и Альтман предложили перейти к двум другим метрикам для каждого наблюдения (пациента):

1. ***Разность измерений*** (D — Difference):

$$D_i = X_{1,i} - X_{2,i}$$

Она показывает, насколько и в какую сторону разошлись показания приборов у конкретного человека.

2. ***Среднее значение*** (A — Average):

$$A_i = \frac{X_{1,i} + X_{2,i}}{2}$$

Поскольку мы не знаем, какой из двух приборов абсолютно точен, наилучшей оценкой истинного значения измеряемой величины является их среднее арифметическое.

Координатная сетка графика Бланда-Альтмана:

1. По горизонтальной оси (X) откладывается среднее значение двух методов: ($X_1 + X_2$) / 2.

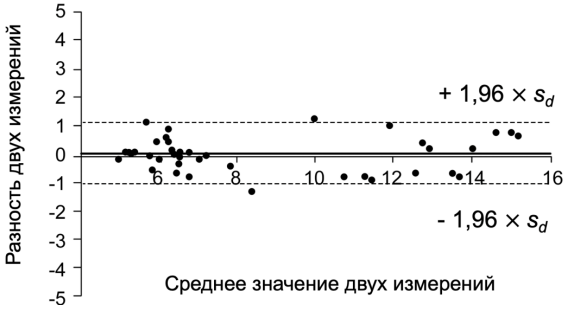


Рис. 23. Диаграмма Бланда-Альтмана

2. По вертикальной оси (Y) откладывается разность между методами: $X_1 - X_2$.

Каждый пациент наносится на график как точка с координатами (A_i, D_i). Если приборы идеально согласуются, все точки должны лечь строго на горизонтальную линию $Y = 0$.

График Бланда-Альтмана — это не просто визуальное облако точек. На него наносятся три важнейшие математические линии, которые рассчитываются по всей выборке.

Линия 1. ***Средняя разность*** (Mean Difference / Bias). Это среднее арифметическое всех индивидуальных разностей:

$$\bar{d} = \frac{\sum d_i}{n}$$

Эта линия называется также ***системным сдвигом*** (bias).

Если $\bar{d} = 0$, это значит, что приборы в среднем не завышают и не занижают показатели относительно друг друга.

Если $\bar{d} = +5$, это говорит о том, что первый метод систематически выдает значения в среднем на 5 единиц выше, чем Второй метод.

Линии 2 и 3. ***Границы согласия*** (Limits of Agreement — LoA). Точки на графике неизбежно будут «гулять» вокруг линии системного сдвига из-за случайной погрешности. Нам нужно измерить этот разброс. Для этого мы рассчитываем стандартное отклонение разностей (s_d).

Бланд и Альтман предложили провести верхнюю и нижнюю границы, внутри которых (при условии нормального распределения разностей) будет лежать 95% всех индивидуальных расхождений:

$$\text{Верхняя граница (Upper LoA)} = \bar{d} + 1{,}96 \times s_d$$

$$\text{Нижняя граница (Lower LoA)\}} = \bar{d} - 1{,}96 \times s_d$$

Иногда, в некоторых исследованиях указывают границы согласия в $\pm 2\sigma_d$. Эти границы показывают, насколько сильно приборы могут «разойтись во мнениях» в самом худшем (но статистически ожидаемом) случае для конкретного пациента.

Как принять решение о взаимозаменяемости методов? Самый частый вопрос новичков: «Программа посчитала границы согласия. Какое значение p-value доказывает, что приборы взаимозаменяемы?». Метод Бланда-Альтмана не рассчитывает финальное значение p. Он не дает готового ответа «да» или «нет». Решение принимает сам исследователь, исходя из клинической или прак-

тической допустимости. Перед тем как строить график, ученый (или врач) должен сесть и зафиксировать для себя медицинский порог: «Какое расхождение приборов я считаю безопасным для жизни пациента?».

Пример с тонометрами: Врач решает, что погрешность в пределах ±10 мм рт. ст. допустима и не повлияет на схему лечения. Мы построили график Бланда-Альтмана и получили: системный сдвиг $\bar{d} = +1$ мм рт. ст., а границы согласия (LoA) составили от -8 до +9 мм рт. ст.

Решение: так как получившиеся границы согласия (от -8 до +9) полностью уложились в наш клинический допуск (±10), методы признаются взаимозаменяемыми. Врач решает использовать электронный тонометр. Для справки: стандарты ISO гораздо жестче, по ним такой тонометр не проходит, и согласно им, новый прибор использовать опасно.

Внимательный взгляд на облако точек Бланда-Альтмана может рассказать о приборах гораздо больше, чем просто цифры границ. Мы всегда ищем графические аномалии:

А. ***Пропорциональная ошибка*** (Тренды). Если облако точек не лежит горизонтально, а устремляется по диагонали вверх или вниз, это означает, что расхождение между приборами зависит от масштаба самой величины. Например, при низком пульсе (в покое) умные часы и ЭКГ работают одинаково, но при высоком пульсе (на тренировке) часы начинают катастрофически занижать показатели. На графике мы увидим, как с ростом среднего значения (X) разность (Y) уходит глубоко в минус. Такую проблему лечат калибровкой или переходом к относительным разностям (в процентах).

Б. ***Гетероскедастичность разностей.*** Иногда облако точек напоминает «мегафон» или воронку: при малых значениях точки сидят кучно, а при больших — разлетаются во все стороны. Это значи,т, что прибор теряет точность и начинает хаотично «гулять» только при больших масштабах измерений.

9.4. Диагностическая точность: краткие итоги

- Качество любого диагностического теста оценивается через матрицу ошибок, разделяющую результаты на истинные и ложные.

- Чувствительность — это способность теста не пропустить больного (минимизация ложноотрицательных результатов).
- Специфичность — это способность теста не брать здорового на ложную тревогу (минимизация ложноположительных результатов).
- Положительный результат теста не означает автоматический диагноз. Насколько ему можно верить, жестко определяется распространенностью болезни в популяции. Чем реже болезнь, тем ниже прогностическая ценность «плюса».
- ROC-кривая — это графический инструмент, оценивающий качество модели классификации на всем диапазоне возможных порогов отсечения. График сопоставляет долю правильных ответов (TPR / Чувствительность) и долю ложных тревог (FPR / 1 – Специфичность).
- AUC — площадь под этой кривой. Метрика показывает способность модели разделять и ранжировать классы (от 0,5 — случайность, до 1,0 — идеал). Для поиска математически сбалансированного порога используют индекс Юдена.
- При экстремальном дисбалансе классов ROC-AUC склонен завышать качество модели, в таких задачах следует переходить на PR-кривую.
- Коэффициент корреляции нельзя использовать для оценки согласия или взаимозаменяемости приборов — он оценивает связь, но игнорирует системные ошибки.
- Метод Бланда-Альтмана переводит анализ в пространство «Среднее значение двух методов» против «Разность методов».
- Метод рассчитывает системный сдвиг (bias) и 95%-е границы согласия (Limits of Agreement).
- Взаимозаменяемость доказывается только тогда, когда полученные границы согласия полностью укладываются в заранее утвержденный клинический/практический допуск.

Заключение

Изучив материалы данного курса, мы прошли путь, который проходит любое качественное клиническое исследование. Теперь вы обладаете системным пониманием того, что статистика — это не абстрактные цифры, а строгий язык доказательной медицины.

Главные выводы и принципы, которые необходимо забрать в практику:

1. Качество входа определяет качество выхода (GIGO — Garbage In, Garbage Out): никакие самые современные и сложные методы анализа (будь то множественная регрессия или модель Кокса) не способны исправить базу данных, собранную с ошибками. Аккуратность на этапе заполнения ИРК и верификации данных — залог успеха всей научной работы.

2. Всегда необходимо принимать во внимание тип данных и распределение. Выбор статистического критерия никогда не должен быть случайным. Прежде чем рассчитать среднее значение или применить критерий Стьюдента, исследователь обязан проверить форму распределения и тип данных. Там, где нормальное распределение нарушено, нас всегда застрахуют устойчивые медианы, IQR и непараметрические тесты.

3. Клиническая сущность выше статистической. Подходить к анализу нужно осознанно. Границы согласия Бланда-Альтмана, пороговые точки ROC-кривых и значимость факторов в регрессионных моделях обретают смысл только тогда, когда они интерпретируются с точки зрения биологической логики и безопасности для жизни пациента.

Помните: биостатистика — это не барьер на пути к защите диссертации или публикации статьи, а ваш главный союзник. Она защищает вас от ошибочных выводов, а ваших будущих пациентов — от неэффективных методов лечения. Успехов в ваших научных изысканиях и практической деятельности!

Глоссарий

А

Альтернативная гипотеза (Alternative hypothesis) — гипотеза, противоположная нулевой гипотезе. Принимается истинной, когда нулевая гипотеза ложна

Анализ в подгруппах (Subgroup analysis) — разделение выборки в метаанализе или эксперименте на более мелкие однородные группы (например, по полу или возрасту) для поиска скрытых закономерностей или причин высокой гетерогенности.

Анализ выживаемости (Survival Analysis) — ветвь статистики, изучающая данные типа «время до наступления события» (смерть пациента, уход клиента, поломка оборудования) с учетом цензурирования данных.

Асимметрии коэффициент (асимметрия, Skewness) — мера отклонения распределения от симметрии. Положительное значение указывает на правостороннюю асимметрию (длинный «хвост» распределения уходит вправо). Негативное значение указывает на левостороннюю асимметрию. Значение около 0 говорит о симметричности распределения.

Б

Базовый риск (Baseline hazard) — функция риска в модели Кокса для объекта, у которого значения всех объясняющих переменных (предикторов) равны нулю.

Биномиальное распределение (Binomial distribution) — статистическое распределение, используемое для описания вероятности того, что x объектов в выборке размера n обладают определенным признаком. Вероятность обладания признаком для каждого отдельного объекта постоянна. Наблюдения за объектами независимы друг от друга.

Бланда-Альтмана метод — графический метод оценки согласия и взаимозаменяемости двух методов непрерывного измерения, строящий график в осях «Среднее значение» и «Разность значений».

Бонферрони критерий — очень строгий (консервативный) метод корректировки уровня значимости при множественных сравнениях.

В

Вариации коэффициент (Coefficient of variation, CV) — стандартное отклонение, выраженное в процентах от выборочного среднего. Также называется относительным стандартным отклонением (RSD)

Вероятность события (Probability of event) — число от 0 до 1, указывающее на степень возможности (правдоподобность) наступления события.

Верхняя граница допуска (Upper Specification Limit, USL) — максимально допустимое значение характеристики качества.

Верхняя контрольная граница (Upper Control Limit, UCL) — на простых контрольных картах рассчитывается как: среднее + 3 стандартных отклонений

Вес исследования (Study weight) — показатель значимости конкретной статьи в метаанализе. Чем больше объем выборки и чем меньше дисперсия в исследовании, тем выше его вес при расчете объединенного эффекта.

Внутригрупповая дисперсия (Within-group variance / Error) — показатель разброса данных внутри каждой из сравниваемых групп относительно их собственного группового среднего. Отражает естественный случайный «шум» данных.

Выборка (Sample) — некоторое количество объектов из генеральной совокупности, отобранных (случайным образом) для получения информации обо всей совокупности.

Выборка (процесс отбора, Sampling) — процесс отбора (извлечения) выборки из генеральной совокупности.

Выборочная доля (доля отбора, Sampling fraction) — отношение количества объектов в выборке к общему количеству объектов в генеральной совокупности.

Выборочная статистика (Sample statistic) — функция от значений выборочных данных, например, выборочное среднее.

Выборочное среднее (Average) — рассчитывается как сумма всех значений данных, деленная на их количество; вычисляется для значений в выборке или эксперименте. Среднее значение для всей генеральной совокупности называется математическим ожиданием (истинным средним).

Выживаемости функция (S(t)) — функция, возвращающая вероятность того, что исследуемый объект успешно «выживет» (не столкнется с целевым негативным событием) дольше заданного момента времени t.

Г

Генеральная совокупность (Population) — полный набор всех объектов или индивидов, подлежащих рассмотрению.

Гетерогенность / Неоднородность (Heterogeneity) — степень различия результатов (размеров эффекта) между независимыми исследованиями, включенными в один метаанализ. Оценивается индексами Q и I^2.

Гетероскедастичность — изменчивость дисперсии случайных ошибок модели (остатков) при изменении значений независимых переменных. Нарушает допущения линейной регрессии и ANOVA.

Гистограмма (Histogram) — столбчатая диаграмма, показывающая частоту сгруппированных интервалов значений количественной переменной.

Гомоскедастичность (Гомогенность дисперсий) — допущение о том, что дисперсия случайных ошибок в регрессии или разброс данных внутри сравниваемых групп (в ANOVA и *t*-тестах) являются постоянными и равными.

Границы допуска (технические условия, Specification limits) — предельные значения характеристики качества, которые признаются допустимыми. Обычно существуют как верхняя, так и нижняя границы допуска.

Границы согласия (Limits of Agreement, LoA) — интервал на графике Бланда-Альтмана, внутри которого ожидаемо лежит 95% всех индивидуальных различий между двумя методами измерения.

Д

Двусторонний критерий (тест) (Two-sided test) — тест статистической гипотезы, при котором нулевая гипотеза отклоняется как при слишком малых, так и при слишком больших значениях выборочной статистики. Это наиболее стандартная ситуация в исследованиях.

Детерминации коэффициент (R^2) — метрика качества регрессионной модели, показывающая долю дисперсии зависимой переменной, объясненную включенными в модель факторами.

Дисперсионный анализ (Analysis of variance, ANOVA) — статистический метод, разделяющий общую дисперсию данных на компоненты, обусловленные влиянием одного или нескольких факторов. Может использоваться, например, для проверки равенства средних значений в нескольких группах.

Дисперсия (Variance) — мера, выражающая средний квадрат расстояний от значений данных до их выборочного среднего.

Доверительный интервал (Confidence interval) — интервал, который с заданной вероятностью (например, 95% или 99%) содержит истинное значение параметра генеральной совокупности (например, математическое ожидание).

Е

Единица отбора (элемент выборки, Sampling unit) — отдельный объект генеральной совокупности, который может быть отобран в выборку.

И

Интерцепт (Intercept) — свободный член в уравнении регрессии, показывающий теоретическое значение зависимой переменной *Y*, когда все независимые факторы *X* равны нулю.

Исход / Результат (Outcome) — результат наблюдения или измерения.

К

Каплана-Майера метод — непараметрический способ расчета и графического построения ступенчатой кривой функции выживаемости S(t) на основе данных с цензурированием.

Квантиль (Fractile / Quantile) — в распределении p-квантиль — это значение, которое отделяет долю p наименьших значений данных от наибольших.

Квартиль (Quartile) — нижний (1-й) квартиль: 25%-й процентиль распределения. Верхний (3-й) квартиль: 75%-й процентиль распределения.

Кластерная (гнездовая) выборка (Cluster sampling) — метод, при котором генеральная совокупность делится на кластеры (группы), каждый из которых состоит из нескольких единиц отбора. Случайным образом выбирается несколько кластеров. Крупные кластеры могут выбираться с большей вероятностью, чем мелкие. Внутри каждого выбранного кластера отбираются одна, несколько или все единицы.

Кокса модель пропорциональных рисков — полупараметрическая регрессионная модель в анализе выживаемости, оценивающая одновременное влияние множества факторов на функцию риска наступления события.

Контрольная карта (Control chart) — график, на котором данные откладываются в хронологической последовательности с целью отличить случайные (естественные) колебания от систематических изменений.

Контрольные границы (Control limits) — линии на контрольной карте, показывающие естественную изменчивость процесса, когда он подвержен исключительно случайным колебаниям.

Корреляции коэффициент/ Корреляция (Coefficient of correlation/ Correlation) — Мера силы линейной зависимости между двумя переменными. Выражается числом от –1 до 1. Значения –1 и 1 соответствуют строгой линейной зависимости (с отрицательным и положительным наклоном соответственно). Значение 0 указывает на отсутствие линейной связи.

Краскела-Уоллиса критерий (H-тест) — непараметрический аналог однофакторного дисперсионного анализа (ANOVA), используемый для одновременного сравнения распределений трех и более независимых групп на основе рангов.

Критическое значение (Critical value) — квантиль (обычно 95% или 97,5%) распределения (например, распределения хи-квадрат или *t*-распределения), с которым сравнивается выборочная статистика в статистическом тесте для принятия или отклонения нулевой гипотезы.

Л

Линия регрессии (Regression line) — статистическая модель, в которой математическое ожидание переменной *Y* линейно зависит от *X*. На графике представляет собой прямую линию, построенную методом наименьших квадратов.

Лог-ранговый критерий (Log-Rank Test) — непараметрический метод пошагового сопоставления кривых выживаемости Каплана-Майера для проверки гипотезы о равенстве выживаемости в двух или более группах.

Логнормальное распределение (Lognormal distribution) — если логарифм исходных данных может быть описан нормальным распределением, то говорят, что исходные значения данных следуют логнормальному распределению.

Ложноотрицательный результат (False Negative, FN) — ошибка II рода в диагностике: ситуация, когда тест ошибочно называет больного человека (целевое событие) здоровым.

Ложноположительный результат (False Positive, FP) — ошибка I рода в диагностике: ложная тревога, когда тест обнаруживает патологию или событие у здорового человека (объекта).

М

Манна-Уитни критерий (U-тест) — непараметрический статистический тест для сравнения распределений (медиан) двух независимых выборок на основе сумм их рангов.

Математическое ожидание / Истинное среднее (Mean / Expectation) — среднее значение генеральной совокупности. Обычно неизвестно. Его оценка вычисляется как выборочное среднее по выборке.

Матрица ошибок (Confusion Matrix) — таблица сопряженности 2x2, сопоставляющая бинарные предсказания теста (модели) с реальными данными «золотого стандарта» (классами TP, TN, FP, FN).

Медиана (Median) — число, которое делит упорядоченные данные на две равные части. Значение, находящееся «посередине», или 50%-й фрактиль (2-й квартиль).

Медиана выживаемости — показатель времени, к которому ровно 50% объектов из исследуемой выборки сталкиваются со своим целевым событием (умирают, увольняются, оттекают).

Межквартильный размах (IQR, Inter-Quartile Range) — разность между верхним (3-м) и нижним (1-м) квартилями. (В некоторых книгах определяется как половина этой разности)

Метаанализ — количественный статистический метод объединения результатов нескольких независимых научных исследований для получения обобщенной, высокоточной оценки размера эффекта.

Метод наименьших квадратов (МНК, Method of least squares) — метод поиска наилучшей модели для описания данных (например, прямой линии). Идея заключается в выборе модели, минимизирующей сумму квадратов остатков (т. е. расстояний между реальными значениями данных и значениями, предсказанными моделью).

Множественная линейная регрессия — метод моделирования линейной зависимости между одной непрерывной результирующей переменной (*Y*) и несколькими независимыми факторами-предикторами (X1, X2 …).

Множественные сравнения — лавинообразный рост вероятности совершить ошибку первого рода (ложноположительный вывод) при проведении нескольких попарных статистических тестов подряд на одних и тех же данных.

Мода (Mode) — значение данных, встречающееся с наибольшей частотой.

Мультиколлинеарность — состояние в множественной регрессии, при котором две или более независимые переменные (Xi) сильно коррелируют между собой, искажая расчет коэффициентов и завышая их стандартные ошибки.

Н

Независимые события (Independent events) — два события называются независимыми, если вероятность их совместного наступления (пересечения) в точности равна произведению вероятностей этих событий по отдельности.

Непараметрические тесты (Distribution-free tests) — группа статистических методов проверки гипотез, не требующих подчинения данных закону нормального распределения и работающих преимущественно со значениями рангов.

Нижняя граница допуска (LSL, Lower Specification Limit) — минимально допустимое значение характеристики качества.

Нижняя контрольная граница (LCL, Lower Control Limit) — на простых контрольных картах рассчитывается как: Среднее - 3 станд. откл.

Нормальное распределение (Normal distribution) — симметричное распределение, используемое для описания количественных (непрерывных) данных.

Нулевая гипотеза (Null hypothesis) — статистическая гипотеза (предположение) о параметре генеральной совокупности (например, о среднем). Нулевая гипотеза считается истинной, пока выборочные данные не докажут обратное (т. е. пока не будет принята альтернативная гипотеза).

О

Общая точность (Accuracy) — базовая метрика классификации, отражающая долю всех правильных ответов модели ко всему объему выборки. Неэффективна при сильном дисбалансе классов.

Односторонний критерий (тест) (One-sided test) — критерий проверки гипотезы, при котором нулевая гипотеза отклоняется только при очень малых либо только при очень больших значениях выборочной статистики на основе априорных содержательных знаний.

Омнибусный критерий — общий тест (например, глобальный ANOVA или тест Краскела-Уоллиса), который проверяет гипотезу целиком для всей системы данных, но не указывает, где конкретно внутри групп лежат точечные различия.

Опрос / Обследование (Survey) — сплошное исследование всей генеральной совокупности или изучение репрезентативной выборки

Относительная частота (Relative frequency) — частота, выраженная в виде доли или процента от общего числа наблюдений (суммарной частоты).

Отношение рисков (Hazard Ratio, HR) — метрика модели Кокса, показывающая соотношение мгновенных интенсивностей (скоростей) наступления события в двух сравниваемых группах на протяжении времени.

Отношение шансов (Odds Ratio, OR) — статическая относительная метрика (результат логистической регрессии), показывающая отношение шанса наступления исхода в первой группе к шансу этого же исхода во второй группе без учета фактора времени.

Отсутствие ответов (неответы, Nonresponse) — факт неучастия некоторых респондентов в выборочном опросе. Может быть вызван проблемами на этапе сбора данных.

Оценка (выборочная оценка) (Estimate) — значение параметра генеральной совокупности (например, среднего значения), рассчитанное по выборке

Ошибка I рода (Type I error) — ошибка, заключающаяся в отклонении нулевой гипотезы, когда она на самом деле верна (ложная тревога).

Ошибка II рода (Type II error) — ошибка, заключающаяся в принятии нулевой гипотезы, когда она на самом деле ложна (пропуск цели).

П

Параметрические тесты — статистические методы (*t*-критерий, ANOVA, линейная регрессия), которые опираются на расчет параметров распределения и требуют строгого соблюдения допущения о нормальности данных.

Парные (зависимые) выборки — выборки, в которых каждому наблюдению из первой группы строго соответствует одно конкретное наблюдение из второй группы (дизайн «До — После», измерения на близнецах).

Площадь под ROC-кривой (ROC-AUC) — интегральная метрика (от 0,5 до 1,0), оценивающая общую способность модели разделять и ранжировать классы.

Положение (центр распределения, Location) — центр значений данных распределения или генеральной совокупности.

Порог отсечения (Threshold / Cu*t*-off point) — граничное значение вероятности (число от 0 до 1), при превышении которого непрерывный прогноз модели преобразуется в жесткое бинарное решение «Да», а при недостижении — в «Нет».

Пост-хок тесты (Pos*t*-hoc) — дополнительные попарные критерии, запускаемые только после обнаружения значимого эффекта общим омнибусным критерием (ANOVA) для выявления конкретных различающихся групп с защитой от накопления ошибки.

Прогностическая ценность отрицательного результата (NPV) — доля истинно здоровых людей среди всех тех объектов, для которых диагностический тест выдал отрицательный знак («минус»).

Прогностическая ценность положительного результата (PPV / Precision) — точность теста; доля действительно больных людей среди всех тех объектов, для которых тест сработал положительно («плюс»).

Простая случайная выборка (Simple random sampling) — Метод отбора выборки, при котором все объекты генеральной совокупности извлекаются случайно и имеют абсолютно одинаковую вероятность быть задействованными.

P

Размах (вариационный размах) (Range) — разность между наибольшим и наименьшим значениями данных.

Размер эффекта (Effect size) — количественная мера силы взаимосвязи между переменными или масштаба различий между группами (например, коэффициент d Коэна), независимая от объема выборки.

Рандомизация (Randomization) — перемешивание объектов в случайном порядке. Используется для: (простого) случайного отбора единиц в выборку; проведения эксперимента со сложными комбинациями нескольких факторов.

Ранжирование — математическая замена реальных количественных значений данных их позициями (порядковыми номерами / рангами) в общем отсортированном по возрастанию ряду.

Распределение хи-квадрат (Chi-squared distribution) — статистическое распределение, принимающее только положительные значения. Требует указания числа степеней свободы. Используется, например, для: тестов на независимость в таблицах сопряженности (частот) или построения доверительного интервала для дисперсии.

Распространенность (Prevalence) — доля лиц с изучаемым заболеванием (или признаком) во всей обследуемой популяции в определенный момент времени. Критически влияет на метрику PPV.

Рассеяние / Разброс данных (Dispersion) — степень разброса значений данных в распределении или в генеральной совокупности.

Регрессионный анализ (Regression analysis) — статистический метод оценки связи (например, линейной) между зависимой переменной (Y) и одной или несколькими независимыми переменными (X).

Риска функция (Hazard Function) — показатель мгновенной интенсивности наступления целевого события в момент времени t для объекта, успешно дожившего до этого рубежа.

С

Систематический обзор (Systematic review) — жестко задокументированное исследование, выполняющее структурированный поиск и критическую оценку всех существующих научных публикаций по строго заданной теме для минимизации предвзятости.

Системный сдвиг (Bias) — средняя разность между измерениями на графике Бланда-Альтмана. Показывает степень постоянного завышения или занижения результатов одним прибором относительно другого.

Скорректированный R^2 (Adjusted R^2) — модификация коэффициента детерминации, содержащая «штраф» за добавление в множественную регрессию новых объясняющих факторов. Используется для предотвращения переобучения модели.

Случайная величина (Random variable) — математическая функция, заданная на пространстве элементарных исходов.

Случайная ошибка (Random error) — разность между выборочным средним и истинным средним генеральной совокупности, вызванная естественной вариабельностью между объектами и случайным характером формирования выборки.

Смещение/Систематическая ошибка (Bias / Systematic error) — разность между истинным значением и средним значением, обусловленная определенными (известными или неизвестными) причинами, например, отсутствием ответов респондентов при выборочном опросе.

Специфичность (Specificity) — способность теста правильно определять здоровых людей. Доля истинно отрицательных ответов (TN) среди всех реально здоровых участников исследования.

Стандартная ошибка (Standard error) — стандартное отклонение выборочного среднего. Рассчитывается как стандартное отклонение данных, деленное на n.

Стандартное отклонение (Standard deviation) — квадратный корень из дисперсии. Наиболее часто используемая мера разброса (рассеяния) данных.

Статистическая неопределенность (Statistical uncertainty) — половина длины (обычно 95%-го) доверительного интервала (например, для математического ожидания). Это число, которое записывается в отчетах после знака ±.

Степени свободы (df, Degrees of freedom) — параметр *t*-распределения или распределения хи-квадрат. Примеры: для одной выборки — количество значений данных минус один; для таблицы сопряженности частот — (число строк – 1) (число столбцов – 1).

Стратифицированная (расслоенная) выборка (Stratified sampling) — метод, при котором генеральная совокупность делится на однородные группы (страты/слои). Внутри каждой группы применяется простой случайный отбор. Доля отбора в разных стратах может различаться.

Стьюдента распределение (*t*-distribution / Student's *t*-distribution) — симметричное распределение с более «тяжелыми хвостами», чем у нормального распределения. Требует указания числа степеней свободы. Применяется, например, для построения доверительного интервала среднего или теста равенства средних двух групп.

Т

Тьюки критерий (Tukey's HSD) — популярный пост-хок тест для попарного сравнения средних всех групп со всеми после получения значимого результата в ANOVA.

У

Уилкоксона критерий (знаковых рангов) — непараметрический тест для сравнения двух связанных (зависимых) выборок (дизайн «До — После»).

Уровень значимости (Significance level) — вероятность совершить ошибку I рода (ложно отклонить верную нулевую гипотезу). Часто уровень значимости выбирается равным 5% или (в некоторых случаях) 1%.

Уэлча критерий (Welch's *t*-test) — модификация *t*-критерия Стьюдента для независимых групп, применяемая в условиях неравенства дисперсий в сравниваемых

Ф

Фишера критерий (F-критерий) — статистический критерий, оценивающий отношение двух дисперсий. Используется в ANOVA для определения значимости модели.

Форест-плот (Forest Plot) — графическое отображение результатов метаанализа, собирающее на одной плоскости точечные оценки размера эффекта исследований (квадраты), их доверительные интервалы (усы) и финальный обобщенный итог (ромб).

Функция плотности (плотность распределения) (Density function) — может рассматриваться как идеализированная гистограмма (возможно, вымышленной) генеральной совокупности. Площадь под кривой плотности равна 1, что соответствует 100% (всем значениям данных).

Функция распределения (Distribution function) — площадь под кривой плотности для значений данных вплоть до заданного значения x. Соответствует вероятности того, что случайная величина примет значение меньше x ($P(X < x)$).

Ц

Цензурирование данных справа (Right censoring) — ситуация в анализе выживаемости, когда точное время наступления целевого события для объекта неизвестно, так как исследование завершилось раньше или объект выбыл из наблюдения, находясь в «безопасном» статусе.

Центральная линия (Centre line) — линия на контрольной карте, представляющая собой (историческое) среднее значение процесса

Ч

Частичный коэффициент регрессии — коэффициент множественной регрессии, показывающий чистый масштаб изменения переменной *Y* при увеличении фактора Xi на одну единицу при условии, что все прочие факторы в модели остаются неизменными.

Частота (Frequency) — число появлений конкретного значения в распределении. Используется для качественных (категориальных) данных или сгруппированных значений количественных данных.

Чувствительность (Sensitivity / Recall / TPR) — способность теста находить больных. Доля истинно положительных срабатываний (TP) среди всех реально больных участников выборки.

Ш

Шанс (Odds) — отношение вероятности того, что событие произойдет (p), к вероятности того, что оно не произойдет (1-p).

Э

Эксперимент (Experiment) — систематическое исследование с целью определить, какие факторы влияют на продукт или процесс. Различные комбинации факторов тестируются на некотором количестве единиц (объектов), после чего измеряется результат (отклик).

Эксцесса коэффициент (эксцесс, Kurtosis) — параметр, показывающий, насколько «тяжелые хвосты» имеет распределение по сравнению с нормальным распределением. Положительные значения указывают на «тяжелые хвосты», отрицательные — на «легкие хвосты». Значения около 0 характерны для распределения с хвостами, как у нормального.

Ю

Юдена индекс (Youden's J statistic) — интегральный показатель качества теста на ROC-кривой, используемый для математического поиска оптимального порога классификации.

I^2 индекс — метрика метаанализа, показывающая в процентах долю вариации между результатами исследований, обусловленную реальной неоднородностью, а не случайной ошибкой выборки.

P-значение (P-value) — вероятность получить столь же или еще более экстремальные значения выборочной статистики, чем наблюдаемые в опыте.

Рассчитывается для обеих сторон распределения (двусторонний тест) или только для одной стороны (односторонний тест).

ROC-кривая (Receiver Operating Characteristic) — график, визуализирующий качество бинарного классификатора на всех возможных порогах отсечения в осях (Чувствительность) и (1 – Специфичность).

VIF (Variance Inflation Factor) — коэффициент вздутия дисперсии, используемый для количественной оценки уровня мультиколлинеарности для каждого предиктора в множественной регрессии. Значения VIF > 10 говорят о критической проблеме.

Полезные ссылки

Электронный учебник статистики (есть версия pdf, на английском языке) https://en.wikibooks.org/wiki/Statistics

Статистический вычислительный ресурс http://www.socr.ucla.edu

Проект R для статистических вычислений https://www.r-project.org

Гарвардский он-лайн курс по статистике и теории вероятности, https://pll.harvard.edu/course/data-analysis-basic-probability-and-statistics

Медицинские статистические калькуляторы MedCalc https://www.medcalc.org/en/calc/

Рекомендуемая литература

Гланц, С.. Медико-биологическая статистика. Пер. с англ. — М., Практика, 1998. — 459 с.

Гоник, Л. Статистика. Краткий курс в комиксах / Ларри Гоник, Вуллкотт Смит ; пер. с англ. В. Кадученко. — Москва : КоЛибри, 2017. — 152 с.

Уилан, Ч. Голая статистика. Самая интересная книга о самой скучной науке / Чарльз Уилан ; пер. с англ. — М. : Манн, Иванов и Фербер, 2013. — 320 с.

Шпигельхалтер, Д. Искусство статистики. Как находить ответы в данных / Дэвид Шпигельхалтер ; пер. с англ. Евгения Поникарова. — М. : Манн, Иванов и Фербер, 2020. — 448 с.

Райс, Д. Математическая статистика и анализ данных / Д. Райс ; пер. с англ. — Москва : ДМК Пресс, 2012. — 704 с.

Ганичева, А. В. Прикладная статистика: учебное пособие / А. В. Ганичева. — Санкт-Петербург: Лань, 2021. — 172 с.

Формат 60x90/16, объем 9.125 усл. печ. л.
Бумага 80 г/м^2. Офсетная. Гарнитура Times New Roman.
Тираж 1000 Заказ № 2026-466-мбу
Отпечатано в типографии ФГБУ ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России
123098, Москва, ул. Живописная, 46
Тел. +7 (499) 190-93-90
rcdm@mail.ru, lochin59@mail.ru
www.fmbafmbc.ru