**Сара Бослаф. Статистика для всех**

Статистика для всех – ясное и краткое введение и руководство для всех новичков. Тщательно переработанное, расширенное, это издание поможет вам глубоко понять статистику, избегая ошеломляющей сложности многих университетских учебников. Эта книга – руководство, которое можно приспосабливать к имеющимся знаниям и нуждам отдельных читателей. Некоторые главы посвящены темам, которые часто отсутствуют в вводных книгах по статистике. Каждая глава представляет собой простые для понимания объяснения, дополненные диаграммами, формулами, задачами с решениями и взятыми из практики заданиями.

Сара Бослаф. Статистика для всех. – М.: ДМК Пресс, 2017. – 586



Купить книгу в [издательстве](https://dmkpress.com/catalog/computer/statistics/978-5-94074-969-1/), цифровую книгу в [ЛитРес](https://www.litres.ru/sara-boslaf/statistika-dlya-vseh-22968987/?lfrom=13042861), бумажную книгу в [Ozon](https://www.ozon.ru/context/detail/id/140646186/?partner=baguzin) или [Лабиринте](https://www.labirint.ru/books/450022/?p=13320)

### Глава 1. Основные понятия, связанные с измерениями

Различают *номинальные данные*, где числа выступают в виде имени или ярлыка и не имеют смысла как числа. Например, вы можете создать переменную для пола, которая принимает значение 1 для мужчин и 0 для женщин. Когда данные принимают только два значения, их называют бинарными.

*Порядковые данные* – это данные, которые можно расположить в каком-либо осмысленном порядке, так что большие значения соответствуют большему проявлению какого-либо признака. Например, ожоги часто характеризуются их степенью. Многие порядковые шкалы используют ранжирование.

Данные, характеризующие отношения, имеют осмысленный порядок, равные интервалы и естественный ноль. Их также называют интервальными данными.

*Операционализация* означает процесс определения способа описания и измерения признаков. Операционализация всегда необходима, когда интересующий нас признак не может быть измерен напрямую. Например, интеллект. Не существует способа прямого измерения интеллекта, так что вместо этого мы должны предложить какую-то величину, которую мы можем измерить, такую как баллы теста на IQ (см. также [Определение – ключ к овладению понятием](http://baguzin.ru/wp/?p=448)).

Классическая теория измерений рассматривает каждое измерение или наблюдаемое значение, Х как сумму двух составляющих: истинного значения, T и ошибки, E (от англ. True, Error): *X = T + E*. В реальном мире мы редко точно знаем истинное значение и, следовательно, также не можем знать точное значение ошибки. Например, если вы делаете ряд измерений веса одного и того же человека в течение короткого промежутка времени (так что его истинный вес можно считать постоянным), используя недавно откалиброванные весы, вы можете использовать среднее арифметическое всех этих измерений как хорошую оценку истинного веса этого человека. Затем вы можете трактовать различия между отдельным измерением и средним значением как ошибку измерений.

Не все ошибки имеют одинаковое происхождение, и мы можем научиться жить со *случайными* ошибками, но любыми способами должны избегать *систематических* ошибок.

Для оценки способов измерений есть два параметра – надежность и валидность. *Надежность* характеризует воспроизводимость наблюдений. *Валидность* характеризует, насколько хорошо тест или балльная шкала измеряют то, что планировалось измерить. Поскольку каждая система измерений имеет свои недостатки, исследователи часто используют несколько подходов к измерению одной и той же величины. Процесс объединения информации из многих источников называется *триангуляцией.*

Смещение измерений может произойти на двух основных этапах: во время отбора объектов для исследования или во время сбора информации об этих объектах. В любом случае ключевой признак смещения – то, что его источником служит скорее систематическая, а не случайная ошибка.

### Глава 2. Теория вероятности

Статистика основана на теории вероятности. Вероятность связана с результатом испытаний, которые также называются экспериментами или наблюдениями. Обычный способ изображения вероятности событий и комбинаций событий – это диаграммы Венна, в которых круги изображают определенные события (рис. 1–3).



Рис. 1. Объединение событий *Е* и *F* (заштрихованная область), обозначается *Е* U *F*

Предположим, что событие – это бросок игральной кости с шестью гранями и что *Е* = {1, 3}, *F* = {1, 2}. Событие *Е* U *F* происходит при выпадении 1, 2 или 3; также можно сказать, что *Е* U *F* = {1, 2, 3}. Событие Е ∩ F происходит, только если выпадает 1, поскольку это значение входит в оба набора элементарных событий, так что Е ∩ F = {1}.



Рис. 2. Пересечение событий *E* и *F* (заштрихованная область), обозначается *Е* ∩ *F*

Если событие заключается в бросании игральной кости с шестью гранями и *F* = {1, 3}, то *~F* = {2, 4, 5, 6}. Дополнение *F* соответствует заштрихованной области на диаграмме Венна (рис. 3). Вероятность события и его дополнения всегда равна 1: P(E) + P(~E) = 1



Рис. 3. Дополнение *F* (заштрихованная область), обозначается *~F*

Если два испытания независимы, то исход одного из них не влияет на исход другого. Иначе говоря, если испытания независимы, то информация об исходе одного из них не дает никакой информации об исходе другого. Классический пример независимости - это подбрасывание обычной монетки; если вы подбросили монетку дважды, результат первого испытания никак не влияет на результат второго испытания.

Часто мы хотим знать вероятность некоторого события, при условии, что произошло другое событие. Это записывается как P(E|F) и читается как «вероятность Е при условии F». Например, тот факт, что вероятность рака легких (исход) выше у курильщиков (фактор), чем у тех, кто не курит, можно выразить при помощи символов следующим образом:

Р(рак легких|курильщик) > Р(рак легких|некурящий).

Говорят, что две переменные независимы, если выполняется следующее равенство:

P(E|F) = P(E).

Если объединяются взаимно исключающие события, то P(*E* U *F*) = P(E) + P(F). Если объединяются не взаимно исключающие события, то P(*E* U *F*) = P(E) + P(F) – P(E ∩ F). Если события E и F независимы, то вероятность их совместного наступления P(*E* ∩ *F*) = P(E)\*P(F). Если события зависимы, то P(*E* ∩ *F*) = P(E) \* P(F|E).

Теорема Байеса (подробнее см. [Дэн Моррис. Теорема Байеса: визуальное введение для начинающих](http://baguzin.ru/wp/?p=18576)):

### Глава 3. Статистический вывод

Статистический вывод – это методология, которая позволяет охарактеризовать генеральную совокупность или сформировать суждения о ней на основании информации о выборке, извлеченной из этой генеральной совокупности. В словаре Мерриам-Вебстер есть два определения термина «вывод (рассуждение)» ([inference](https://www.merriam-webster.com/dictionary/inference)):

* Переход от одного предположения, утверждения или суждения, считаемого верным, к другому, истинность которого следует из истинности первого.
* Переход от данных о статистической выборке к обобщениям (в виде значений параметров генеральной совокупности), как правило, с вычислением степени уверенности.

Для генеральной совокупности среднее обозначается греческой буквой μ, которую правильно называть параметром, поскольку это число характеризует генеральную совокупность, тогда как для обозначения выборочного среднего вы используете латинскую букву х̅, которую правильно называть статистикой, поскольку это число характеризует выборку. В тех случаях, когда вы изучаете составляющие генеральную совокупность случаев и не хотите выходить за их рамки, вам следует использовать *описательную статистику* (см. следующую главу). В тех случаях, когда изучаемые вами случаи не составляют всей генеральной совокупности, и вы хотите сделать обобщения, выходящие за рамки этих случаев, вам следует использовать *статистический вывод*.

На практике статистические заключения часто опираются на допущения о том, как распределены данные. [Нормальное распределение](http://baguzin.ru/wp/?p=1170) – наиболее часто используемый тип непрерывного распределения в статистике. Существует бесконечное множество нормальных распределений, все из которых в целом имеют одну и ту же форму, но различаются из-за их среднего μ и стандартного отклонения σ. Все они путем преобразований могут быть сведены к *стандартному нормальному распределению* (или Z-распределению) со средним арифметическим = 0, и стандартным отклонением = 1.

Для любого нормального распределения (рис. 4):

* около 68% данных находятся в интервале ± одно стандартное отклонение от среднего;
* около 95% данных находятся в интервале ± два стандартных отклонения от среднего;
* около 99% данных находятся в интервале ± три стандартных отклонения от среднего.



Рис. 4. Доля данных, которые попадают в определенные интервалы нормального распределения

Знание этих свойств нормального распределения предоставляет способ решить, насколько типично конкретное значение данных для генеральной совокупности.

*Z-значение* (или нормализованное значение) – это разница между заданным числом и средним арифметическим, выраженная в единицах стандартного отклонения.

Если переменная *x* имеет нормальное распределение со средним арифметическим 100 и стандартным отклонением 5, что можно записать как x ~N(100, 5), то число 105 имеет Z-значение = (105 – 100) / 5 = 1.

*Z-статистика* оценивает вероятность определенного среднего значения для выборки (а не одного конкретного значения, как в случае Z-значения):

где х̅ – среднее значение для выборки; μ – среднее значение для генеральной совокупности; σ – стандартное отклонение для генеральной совокупности; n – размер выборки.

События дискретного биномиального распределения происходят в результате процесса Бернулли. Биномиальное распределение описывает число положительных исходов в *n* испытаниях процесса Бернулли. Формула для вычисления вероятности (*Р*) определенного числа успехов при данном числе испытаний:

где число сочетаний

где *n* – это число испытаний, *k* – число успехов, *p* – вероятность успеха в одном испытании. Сокращенный способ записать биномиальную вероятность – это b(k;n;p) или P(k = k;n;p). Каждая комбинация *p* и *n* даст свое распределение (рис. 5). Если одновременно *np* и *n(1 – p)* больше 5, то биномиальное распределение может быть хорошо описано нормальным распределением.



Рис. 5. Три биномиальных распределения

Мы будем использовать термин *независимая переменная* для обозначения переменных, которые отображают результат исследования, и *зависимая переменная* для переменных, которые, согласно ожиданиям, влияют на результат.

*Центральная предельная теорема.* Распределение значений выборочных средних близко к нормальному вне зависимости от распределения значений генеральной совокупности при условии, что выборки достаточно велики.

В качестве примера рассмотрим случайную равномерную генеральную совокупность из 1000 чисел. Каждое число с равной вероятностью может принимать значение от 1 до 100 (рис. 6). Видно, что числа приблизительно равномерно распределены по десятичным диапазонам. Далее сформируем большое число выборок по 2, 5, 10, 50 чисел, и найдем среднее каждой выборки. Распределение этих средних значений всё ближе к нормальному с ростом числа элементов в выборке (рис. 7). Вы можете открыть приложенный Excel-файл, лист ЦПТ, и нажимая F9, посмотреть, как изменяется картинка.



Рис. 6. Равномерное распределение



Рис. 7. Распределение средних значений выборок в зависимости от числа элементов в выборке

*Проверка гипотез* составляет основу статистического вывода. Она состоит из нескольких основных этапов:

1. Формулировка рабочей гипотезы, которая может быть проверена статистическими методами.
2. Формальное описание нулевой и альтернативной гипотез.
3. Выбор подходящего статистического теста, сбор данных, проведение вычислений.
4. Выработка решения на основании полученных результатов.

Отклонение нулевой гипотезы иногда называется «нахождением статистически значимого результата». Процесс определения того, значимы ли результаты, включает не только статистические расчеты, но и применение основанных на традициях правил, которые могут различаться в зависимости от области исследований или других факторов. Часто вероятность, при которой нулевая гипотеза будет отклонена, принимается p < 0,05.

Статистический вывод – это мощное средство, которое позволяет формулировать вероятностные суждения о данных. Однако поскольку эти суждения вероятностные, а не абсолютно верные, нельзя исключить возможность ошибки.

### Глава 4. Описательная статистика и графическое представление данных

Числа, которые характеризуют генеральную совокупность, называют параметрами и обозначают греческими буквами μ (для среднего) и σ (для стандартного отклонения); числа, которые описывают выборку, называются статистиками и обозначаются латинскими буквами x̅ (выборочное среднее) и s (выборочное стандартное отклонение).

Меры центральной тенденции дают представление о типичном или часто встречающемся значении в данной переменной. Три самые часто применяемые из них – это среднее, медиана и мода (подробнее см. [Определение среднего значения, вариации и формы распределения. Описательные статистики](http://baguzin.ru/wp/?p=5381)).

Разброс говорит о том, насколько сильно рассеяны значения в данных. Из-за этого меры рассеяния часто называют мерами разброса. Меры разброса: размах и межквартильный размах, дисперсия и стандартное отклонение. Дисперсия – это средний квадрат отклонения от среднего, а стандартное отклонение – это квадратный корень из дисперсии. Дисперсия генеральной совокупности:

Дисперсия для выборки

*Таблицы частот* являются очень эффективным способом представления больших объемов данных и являются чем-то средним между текстом (абзацами с описаниями значений данных) и чистой графикой (такой как гистограмма). Подробнее о графическом представлении данных см. [Джин Желязны. Говори на языке диаграмм](http://baguzin.ru/wp/?p=190), [Новые диаграммы в Excel 2016](http://baguzin.ru/wp/?p=16080), [Визуализация статистических данных с помощью диаграммы ящик с усами](http://baguzin.ru/wp/?p=17422), [Джон Тьюки. Анализ результатов наблюдений. Разведочный анализ](http://baguzin.ru/wp/?p=15897), [Нейтан Яу. Искусство визуализации в бизнесе](http://baguzin.ru/wp/?p=14725).

### Глава 5. Категориальные данные

В случае, когда анализ касается исследования связи между двумя категориальными переменными, их распределение в данных часто показывают с помощью RxC-таблиц, которые чаще называют таблицами сопряженности. Мера согласия двух категорий рассчитывается с помощью коэффициента *каппа* (в Excel такой функции нет).

При проверке гипотез о категориальных данных часто лучшим вариантом статистики является один из тестов *хи-квадрат* (χ2). В распределении хи-квадрат есть только положительные значения, поскольку оно основано на сумме квадратов:

где *Хi* – независимые переменные, распределенные по стандартному нормальному закону с μ = 0 и σ = 1, *k* – число степеней свободы (рис. 8; модель см. в файле Excel).



Рис. 8. Функция плотности распределения хи-квадрат при различном числе степеней свободы

Критерий хи-квадрат – это один из наиболее распространенных способов изучения связей между двумя и более категориальными переменными (подробнее см. [Применение χ2-критерия для проверки гипотезы о равенстве двух или нескольких долей](http://baguzin.ru/wp/?p=5971) и [Применение χ2-критерия независимости](http://baguzin.ru/wp/?p=5998)). В Excel есть 4 основные функции, работающие со статистикой хи-квадрат (первые в списке) и 4 устаревшие (рис. 9).



Рис. 9. Функции Excel для работы с распределением хи-квадрат

Критерий хи-квадрат Пирсона подходит для данных, в которых все наблюдения независимы. Кроме того, предполагается, что ни в одной из ячеек ожидаемое значение не меньше 1, и не более чем у 20% ячеек ожидаемое значение меньше 5. Причина возникновения двух последних требований связана с тем, что критерий хи-квадрат асимптотический, и его некорректно применять к разреженным данным (то есть таким, в которых у одной или нескольких ячеек маленькая ожидаемая частота).

Если данных мало или в случае разреженного распределения данных, рекомендуется точный тест Фишера. Когда данные получены из связанных выборок, или в случае парных данных рекомендуется использовать тест МакНемара. Например, для анализа результатов опроса общественного мнения до и после просмотра испытуемыми политической рекламы.

### Глава 6. t-критерий

t-распределение Стьюдента используется вместо нормального, когда выборки слишком малы. t-распределение имеет более тяжелые хвосты, сигнализируя, что в малых выборках отклонения от среднего более вероятны (рис. 10).



Рис. 10. t-распределение с 4 и 18 степенями свободы в сравнении с нормальным

где х̅ – среднее значение для выборки; μ – среднее значение для генеральной совокупности; s – стандартное отклонение для генеральной совокупности; n – размер выборки. Эта формула очень напоминает формулу Z-статистики (3); отличие заключается в том, что при вычислении t-статистики используется стандартное отклонение выборки (s), тогда как при вычислении Z-статистики – отклонение генеральной совокупности (σ).



Рис. 11. Семейство функций в Excel, работающих с t-статистикой (последние три – устаревшие)

Для двухстороннего теста с использованием t-распределения с α = 0,05 критическое значение зависит от числа степеней свободы. Для df = 1 оно составляет 12,7; для df = 10 – 2,2; для df = 30 – 2,04; для df = 50 – 2,01; для df = 100 – 1,98; для бесконечного числа степеней свободы – 1,96 (и, естественно, равно критерию для стандартного нормального распределения).

*Одновыборочный t-критерий* применяется для сравнения среднего выборки и генеральной совокупности. с известным средним. Для расчета t-статистики одновыборочного критерия используется формула (9). Предположим, что выборочное среднее = 90, стандартное отклонение = 10, объем выборки = 15, а среднее генеральной совокупности = 100. Тогда

Число степеней свободы для одновыборочного t-критерия равно n – 1 (одна степень свободы «съедается» средним); df = 14. t-значение для уровня значимости 5% можно рассчитать в Excel c помощью формулы СТЬЮДЕНТ.ОБР(0,975;14) = 2,145. Видно, что значение (10) по абсолютному значению больше. Это позволяет отклонить нулевую гипотезу. С другой стороны, вероятность значения |-3,87| можно определить по формуле р = СТЬЮДЕНТ.РАСП(-3,87;14;ИСТИНА) = 0,00085. Т.е., вероятность того, что выборка с такими параметрами происходит из указанной генеральной совокупности, не превышает 0,085%.

Доверительный интервал (CI, confidence interval) – это некий диапазон значений вокруг среднего. Например, 95%-ный доверительный интервал (самый часто применимый) задает границы вокруг среднего, в которые с вероятностью 95% попадет среднее генеральной совокупности. Поскольку это среднее часто не известно, можно утверждать, что с вероятностью 95% в него попадут иные выборочные средние, если мы будем делать много выборок.

В нашем примере, α = 0,05; x = 90; n = 15; df = n – 1 = 14; s = 10; t0,025;14 = СТЬЮДЕНТ.ОБР(0,975;14) = 2,145.

Т.е., если мы возьмем 100 выборок, то среднее 95 из них попадет в интервал от 84,46 до 95,54.

*t-критерий для независимых выборок*, также называемый двухвыборочным t-критерием, сравнивает средние двух выборок. Задача этого теста состоит в проверке, равны ли средние генеральных совокупностей, из которых были взяты выборки:

В этой формуле x̅1 и х̅2 – это средние двух выборок, μ1 и μ2 – это средние двух генеральных совокупностей, s2p – это объединенная дисперсия, n1 и n2 – это объемы двух выборок, а s21 и s22 – это дисперсии двух выборок.

Формула кажется сложнее, чем есть на самом деле. Поскольку нулевая гипотеза часто состоит в том, что разница между истинными средними равна 0, выражение (μ1 – μ2) можно опустить. Если n1 = n2 = n, то

Иногда данные в выборках – это измерения, сделанные дважды на одних и тех же людях, например, кровяное давление до и после приема лекарства. Формула для расчета t-статистики для парного t-критерия основана на разностях, рассчитанных для каждой пары элементов выборок:

В этой формуле d̅ – средняя разница, и μ1 и μ2 – это средние двух совокупностей, sd – стандартное отклонение разниц, а n – число пар.

Нулевая гипотеза для парного t-критерия обычно состоит в том, что средняя разница d̅ равна 0, тогда как альтернативная гипотеза говорит, что она отлична от 0. При этом величина (μ1 – μ2) предполагается равной 0, и в таком случае ее можно опустить.

Предположим, мы хотим проверить эффективность программы диеты с физическими упражнениями в снижении общего уровня холестерина у мужчин среднего возраста.



Рис. 12. Уровень холестерина до и после диеты и упражнений

Абсолютное значение нашей t-статистики превосходит t-критерий для α = 0?05, поэтому мы отвергаем нулевую гипотезу и заключаем, что упражнения и диета оказали значимый эффект на общий уровень холестерина.

### Глава 7. Коэффициент корреляции Пирсона

Самое главное в корреляции – то, что она никак не может выявить причину. Поэтому даже самая сильная корреляция сама по себе не может свидетельствовать о причинно-следственной связи; она может быть подтверждена только с помощью постановки эксперимента.

Диаграмма рассеяния – это полезное средство для изучения взаимоотношений между переменными, и обычно создание таких диаграмм для непрерывных переменных проводится на [разведочной стадии](http://baguzin.ru/wp/?p=15897) работы с данными.

Коэффициент корреляции Пирсона:

где SSx – это сумма квадратов отклонений x, SSy – это сумма квадратов отклонений у, SSху – это сумма произведений отклонений x и у:

В Excel можно использовать функции КОРРЕЛ(), СУММКВ() и СУММПРОИЗВ().

Предположим, мы получили выборку 10 американских старшеклассников и анализируем результаты выполнения ими разделов Академического оценочного теста (Scholastic Aptitude Test, SAT), направленных на проверку вербальных и математических умений (рис. 13).



Рис. 13. Баллы SAT

Значима ли данная корреляция? Нулевая гипотеза (Н0) для корреляционного анализа состоит в том, что переменные не связаны, то есть r = 0. Альтернативная гипотеза (Н1) состоит в том, что r ≠ 0. t-статистика для коэффициента корреляции рассчитывается с помощью формулы:

где r – это коэффициент корреляции Пирсона для выборки, n – это ее объем.

Критическое t при α = 0,05 и df = n – 2 (*n* уменьшается на 2 из-за фиксированного среднего и коэффициента корреляции) при двустороннем тесте можно рассчитать в Excel по формуле =СТЬЮДЕНТ.ОБР(1-α/2;n-2)=2,306

Мы отвергаем нулевую гипотезу о том, что результаты в математической и вербальных частях не связаны. Мы также рассчитали точное p-значение для этих данных с помощью формулы =СТЬЮДЕНТ.РАСП.2Х(r;n-2) =СТЬЮДЕНТ.РАСП.2Х(4,997;8) и получить двустороннее р-значение, равное 0,0011, что также показывает, что наши результаты очень маловероятны, если на самом деле эти переменные не связаны в генеральной совокупности, из которой мы брали выборку.

*Коэффициент детерминации* (r2) показывает, какую долю дисперсии одной переменной можно связать с другой переменной.

### Глава 8. Введение в регрессию и дисперсионный анализ

Регрессия и дисперсионный анализ (ANOVA, от ANalysis Of Variance) – два метода, использующие общую линейную модель. Зависимая переменная *у* может быть выражена через предикторную переменную *х* линейным уравнением *у = ах + b*.

Регрессионный анализ можно представить, как проведение прямой линии (регрессионной прямой), изображающей взаимосвязь между двумя переменными; эту линию часто накладывают на диаграмму рассеяния для дальнейшего уточнения связи (рис. 14).



Рис. 14. Диаграмма рассеяния роста в метрах и массы тела в килограммах для 436 взрослых американцев

Коэффициент корреляции (r) и коэффициент детерминации (r2) между ростом и массой тела равны соответственно 0,47 и 0,22. Это значит, что около 22% дисперсии массы тела может быть связано с ростом, совсем не идеальное предсказание или объяснение, но значительно более хорошее, чем 0. Уравнение регрессии для этих данных: y = 91x - 74.

Когда статистический пакет выдает линию регрессии для набора данных, он подбирает уравнение методом наименьших квадратов. Разницу между каждой точкой и линией регрессии называют остатком, потому что она представляет изменчивость в данных, не объясняемую уравнением прямой.

Как в случае большинства статистических процедур, в ходе линейной регрессии приходится делать некоторые предположения об анализируемых данных; если они нарушаются, то результат анализа может быть неверным. Важнейшие условия для проведения простой линейной регрессии включают:

* *Подходящий тип данных.* Результирующая переменная должна быть непрерывной, представленной интервальными или характеризующими отношения данными и неограниченной (или хотя бы варьирующей в широких пределах)
* *Независимость*. Каждое значение зависимой переменной независимо от других значений. Это допущение нарушается, если есть некоторая зависимость от времени, или если независимые переменные относятся к объектам, объединенным в группы (члены одной семьи, ученики класса).
* *Линейность*. Отношения между независимой и зависимой переменными напоминают прямую линию.
* *Распределение*. Непрерывные переменные приблизительно нормально распределены и не имеют выбросов. Распределение непрерывных переменных можно проверить «на глаз» с помощью построения гистограммы и с помощью статистических тестов на нормальность, таких как тест Колмогорова-Смирнова.
* *Гомоскедастичность*. Ошибки предсказания *у* постоянны на всем промежутке данных *х*.
* *Независимость и нормальность ошибок.* Ошибка предсказания для каждой точки не должна зависеть от ошибки предсказания других точек, и остатки должны быть нормально распределены. Предположение о независимости проверяется тестом Дарбина-Уотсона, а предположение о нормальности – с помощью построения графика остатков (ошибок).

Предположим, нас интересует уровень рождаемости среди подростков (отношение числа родов молодых женщин 15-19 лет к их общему числу) и то, какие факторы на уровне страны влияют на него. Наше первое предположение состоит в том, что подростковая рождаемость может быть связана с равноправием полов, и мы предполагаем, что она ниже в тех странах, где у женщин больше прав. Мы проведем регрессионный анализ для проверки этой гипотезы, используя данные, скачанные с сайта проекта ООН по развитию человечества (<http://hdr.undp.org/en/data>). Мы будем использовать индекс неравенства полов (ИНП) как независимую переменную (меньшие значения соответствуют большему равенству).

К сожалению, одно из важных допущений регрессионного анализа (линейность), не выполняется (рис. 15). Однако, если сделать логарифмическое преобразование уровня подростковой рождаемости, то переменные будут связаны гораздо более линейно (рис. 16).



Рис. 15. Диаграмма рассеяния подростковой рождаемости и индекса неравенства полов



Рис. 16. Диаграмма рассеяния для натурального логарифма подростковой рождаемости и индекса неравенства полов

Регрессионная модель: ln(подростковая рождаемость) = β0 + β1(ИНП) + ε. Качество модели обычно выражается в терминах F-статистики и оценивает, лучше ли модель, чем ее отсутствие. Другими словами, значения F-статистики позволяют сравнить нашу модель с моделью, в которой все коэффициенты равны 0. Нас также интересует, какую долю дисперсии зависимой переменной может объяснить наша модель; возможна ситуация, особенно с большими наборами данных, когда модель с достоверными коэффициентами объясняет лишь малую долю всей дисперсии.

F-статистика для модели составляет 190,9 и p-значение менtе 0,001; таким образом, мы заключаем, что лучше такая модель, чем никакой. r-значение, или корреляция, составляет 0,714, а коэффициент детерминации (R2), составляет 0,509; это значит, что индекс неравенства полов может объяснить больше 50% дисперсии подростковой рождаемости. Коэффициенты регрессии модели приведены на рис. 17.



Рис. 17. Коэффициенты регрессионного анализа

Столбец B показывает коэффициенты для уравнения регрессии: ln(подростковая рождаемость) = 1,798 + 4,446(ИНП). Т.е., на каждую единицу ИНП натуральный логарифм подростковой рождаемости увеличивается примерно на 4,4 единицы; эта связь положительна, что подтверждает нашу догадку о том, что более сильное неравенство приводит к более высокой подростковой рождаемости. Столбец «Станд. ошибка» показывает стандартные ошибки для оценок коэффициентов. Столбец «Бета» содержит стандартизованные коэффициенты регрессии; они полезны в регрессионном анализе со множеством независимых переменных, измеряемых в различных масштабах. Столбец t показывает t-статистику для каждого коэффициента и рассчитывается как отношение B к соответствующей стандартной ошибке. Последняя колонка показывает значимость t-статистики.

Последним шагом должна быть проверка наших предположений, чтобы быть уверенными в адекватности полученных результатов. Мы можем проверить гомоскедастичность с помощью построения графика остатков (рис. 18). Это классическое облако данных, не дающее никаких оснований предполагать непостоянство ошибки предсказания, так что условие гомоскедастичности выполнено.



Рис. 18. Диаграмма рассеяния стандартизованных остатков

**Дисперсионный анализ (ANOVA)**

Дисперсионный анализ используется для сравнения средних значений определенной переменной в двух и более независимых группах. Основная статистика для ANOVA – это F-отношение, которое может быть использовано для определения статистической значимости различий между группами. Простейший вариант дисперсионного анализа – это однофакторный дисперсионный анализ, в котором при формировании групп для сравнения используется только одна переменная. Данную переменную часто называют «фактором».

Предположим, нас интересует эффективность нового лекарства, которое должно снижать сахар в крови у больных диабетом второго типа; мы можем проверить ее с помощью дисперсионного анализа, сравнив новое лекарство с другим уже используемым препаратом. Фактором в данном исследовании служит используемое лекарство, и у него есть два уровня: новый и старый препараты. Фактор в однофакторном дисперсионном анализе может иметь и более двух уровней.

Однофакторный дисперсионный анализ с двумя уровнями аналогичен t-критерию. Наша нулевая гипотеза в таком анализе обычно гласит о равенстве средних двух групп, тогда как альтернативная говорит о том, что средние различны (двусторонний тест) или различаются в определенном направлении (односторонний тест). Даже если есть значимое отличие в средних между двумя группами, мы не можем ожидать, что значения в двух группах не будут перекрываться; на самом деле отсутствие такого перекрытия очень необычно. Также мы ожидаем, что внутри каждой группы будет наблюдаться изменчивость, и однофакторный дисперсионный анализ принимает в расчет изменчивость внутри групп (например, изменчивость в уровне сахара среди пациентов, принимающих новое лекарство) и изменчивость между группами (разницу между пациентами, принимающими исследуемый и стандартный препараты). Линейная регрессия и ANOVA – это на самом деле два способа исследовать данные, используя общую линейную модель.

Предположим, мы сравниваем два метода тренировок по подъему веса, и наши измерения показывают увеличение поднятого веса после трех месяцев тренировок одним или другим методом. Наша нулевая гипотеза состоит в том, что средние в обеих группах после тренировок равны; другими словами, в среднем никакой из методов не лучше другого. В начале эксперимента мы случайным образом выбираем, каким способом каждый из испытуемых будет тренироваться, и измеряем средний вес, который они могут поднять; средние были приблизительно равны. Диаграмма на рис. 19 показывает улучшение «грузоподъемности» после трех месяцев. Испытуемые, тренировавшиеся по первому методу, оказались успешнее, поскольку первая группа имеет большую медиану, представленную в виде черной линии в середине «ящика», и весь интервал расположен выше. Тем не менее видно, что в обеих группах есть изменчивость, и имеется значительное перекрытие выборок. Не все члены первой группы улучшили свой результат больше, чем все члены второй группы, только в среднем первая группа показала более хороший результат.



Рис. 19. Увеличение поднятого веса после трех месяцев тренировок одним из двух методов

На самом деле в группе 1 среднее улучшение составляет 34,21 фунта, а в группе 2 – 26,42 фунта. Значима ли эта разница статистически? Для ответа на этот вопрос мы проведем однофакторный дисперсионный анализ. Для начала рассчитаем описательную статистику (рис. 20).



Рис. 20. Описательная статистика по данным о подъеме веса (два метода тренировок)

Размеры групп одинаковы, в группах приблизительно одинаковы дисперсии, и 95%-ные доверительные интервалы для средних двух групп не перекрываются (хотя они почти соприкасаются). Результаты дисперсионного анализа обычно представляют в виде таблицы (рис.21).



Рис. 21. Результаты однофакторного дисперсионного анализа данных о подъеме веса (два метода тренировки)

Руководствуясь стандартом α < 0,05, эти результаты можно считать значимыми, так что мы можем отвергнуть нулевую гипотезу о том, что средние двух групп равны; на самом деле метод 1 привел к достоверно более хорошим результатам, чем метод 2. В таблице есть три строки: одна содержит данные для межгрупповой дисперсии, одна – для внутригрупповой, и одна – для общей. Суммируя внутри- и межгрупповые суммы квадратов и степени свободы (df), мы получаем значения для всей выборки. Дисперсия между группами – это дисперсия, связанная с групповой принадлежностью, то есть с той, которая определяется методом тренировок. Внутригрупповая дисперсия – это дисперсия значений в каждой группе. Число степеней свободы отражает то, сколько параметров могут изменяться при вычислении каждой части статистики; общее число степеней свободы равно n – 1 (на один меньше, чем число испытуемых), число степеней свободы между группами равно k – 1 (на один меньше числа групп), внутригрупповое число степеней свободы равно n – k. Сумма квадратов (SS) – это сумма возведенных в квадрат отклонений между группами, внутри групп и в целом, тогда как средний квадрат (MS) – это сумма квадратов, разделенная на число степеней свободы.

F-статистика - это отношение меж- и внутригрупповой сумм квадратов: F = 455,86/46,23 = 9,86.

Наш статистический пакет автоматически рассчитал значимость F-статистики (0,004). Это значение показывает вероятность того, что различие является случайным. Т.е., вероятность, что выборки сделаны из одной генеральной совокупности не превышает 0,4%.

Дальнейшее изложение, на мой взгляд, представляет интерес для относительно небольшой аудитории, интересующейся более глубокими аспектами статистики. Поэтому я приведу лишь отдельные замечания.

### Глава 9. Многофакторный дисперсионный анализ и ковариационный анализ

См. [Двухфакторный дисперсионный анализ](http://baguzin.ru/wp/?p=5919)

### Глава 10. Множественная линейная регрессия

См. [Введение в множественную регрессию](http://baguzin.ru/wp/?p=6136), [Построение модели множественной регрессии](http://baguzin.ru/wp/?p=6198)

### Глава 11. Логистическая, мультиномиальная и полиномиальная регрессия

Логистическая регрессия подходит для таких ситуаций, когда зависимая переменная дихотомическая, а не непрерывная, мультиномиальная – работает в случае категориальных зависимых переменных (с более чем двумя категориями), а полиномиальная - больше подходит для случаев, когда связь между независимой и зависимой переменными описывается с помощью уравнения, включающего многочлен (например, с х2 или х3).

Зависимая переменная в логистической регрессии – это *логит*, т.е., преобразованная вероятность значения исследуемой характеристики. Если p – это вероятность того, что объект исследования будет обладать данным свойством, то логит для этого объекта

Одна из самых поразительных возможностей современных статистических пакетов состоит в том, что вы можете автоматически выбрать и провести любое число сложных статистических тестов по нажатию одной кнопки. Тем не менее многие статистики хмурятся при построении моделей исключительно на основании ваших данных и сравнивают это с «выуживанием» закономерностей наугад, а если используется нелинейная регрессия, называют это произвольной подгонкой под кривые, или *переподгонкой*.

Это означает, что ваша модель слишком хорошо аппроксимирует данные и объясняет не только достоверные зависимости, но и случайные отклонения. Поскольку задачей статистического анализа являются обобщение результатов и перенос их на другие выборки из той же генеральной совокупности, переподгонка мешает достижению этой цели. Вы можете получить модель, которая замечательно описывает ваши данные, но она совсем не обязательно подойдет для каких-то других данных, так что она не привносит новых полезных знаний в вашу область. Лучшая защита от переподгонки – построение моделей на основании теории.

### Глава 12. Факторный, кластерный и дискриминантный анализы

См. [Однофакторный дисперсионный анализ](http://baguzin.ru/wp/?p=5884), [Двухфакторный дисперсионный анализ](http://baguzin.ru/wp/?p=5919), [Кластерный анализ: сетевые графы и определение сообществ](http://baguzin.ru/wp/?p=13926).

### Глава 13. Непараметрическая статистика

Основа статистического анализа - оценка параметров распределения, то есть оценка свойств генеральной совокупности по информации, полученной из выборки, взятой из этой совокупности. Многие из самых обычных статистических методов полагаются на то, что исследуемое распределение принадлежит к какому-то известному типу, например, оно нормальное. Но что же делать, если вы знаете или подозреваете, что генеральная совокупность отнюдь не подходит под требования определенного статистического теста? В таких ситуациях используют другой набор статистических методов, называемых непараметрическими. Они не зависят от распределения, то есть делают мало или не делают вовсе никаких предположений о свойствах распределения данных.

Непараметрические статистики часто применяют при исследовании данных, если их получали скорее, как ранги, а не как чистые значения. Ранговые данные являются порядковыми, и их нельзя анализировать методами, предназначенными для интервальных или характеризующих отношения данных.

Рассмотрим следующий пример. Отборочный комитет Олимпийских игр должен выбрать лучшую команду по тхэквондо из двух штатов (Калифорния и Невада). Поскольку, кроме индивидуальных зачетов, будут и групповые, к которым члены команд готовились вместе, команды нельзя перемешивать, чтобы получить составную команду из самых лучших спортсменов; вместо этого необходимо выбрать одну или другую команду как целое. Каждый член команд получил общий балл за свое выступление (рис. 22).



Рис. 22. Результаты членов команд по тхэквондо

Более высокий балл указывает на более хорошие навыки. Самым подходящим способом описания таких данных являются ранги, а не значения. Мы припишем ранг каждому испытуемому. Для равных значений следует подсчитать средний ранг (рис. 23).



Рис. 23. Ранги для оценок выступления борцов тхэквондо с учетом равных значений

r̅ – средний ранг для группы. Команда Калифорнии выступила лучше, чем команда из Невады. Как проверить, достоверно ли отличие между командами? Мы можем использовать z-критерий для определения, является ли различие между двумя группами достоверным на стандартном уровне значимости α = 0,05. По нулевой гипотезе у этих двух групп средние ранги равны, так что мы можем рассчитать ожидаемую сумму рангов – μW. Общая сумма рангов μW = 210. Так что на каждую команду должно приходиться по 105 баллов.

Z-критерий:

где W – сумма рангов одной из команд, μW – ожидаемая сумма рангов, σW – стандартное отклонение:

В этой формуле n1 и n2 – это объемы выборок в группах, а 12 – это константа. Обратите внимание, что стандартное отклонение рангов зависит только от объемов выборок, но не от значений их элементов. z = (97,5 – 105)/13,23 = –0,57. z-значение лежит ближе к нулю (внутри диапазона), чем критическое z-значение для α = 0,05. zα = ±1,96. Вероятность нашего z-значения равного ±0,57 составляет 43%, таким образом, мы не можем отвергнуть нулевую гипотезу. Различие между двумя командами статистически не значимо.

Далее рассмотрен Н-критерий Краскела-Уоллиса, парный критерий Вилкоксона, медианный критерий, критерий знаков, тест Фридмана.

### Глава 14. Статистика для бизнеса и контроля качества

Рассмотрены индексы (в частности индекс потребительских цен США.

*Временн****ы****е ряды* часто используются в бизнес-статистике для отображения изменения величин во времени. Одно из свойств временных рядов заключается в том, что данные во временной последовательности не независимы друг от друга, как это ожидается в стандартной обобщенной линейной модели, для них характерна *автокорреляция*. Это значит, что значение величины в данный момент времени связано со значениями, которые идут перед и после нее, а возможно, и с более удаленными значениями этого временного ряда (см. [Колебания курса рубля по отношению к доллару не подчиняются нормальному распределению](http://baguzin.ru/wp/?p=1760), [Выявление долгосрочной зависимости изменения курса доллара на основе R/S-анализа](http://baguzin.ru/wp/?p=1929)).

Для описания составляющих временн***о***го ряда часто используются аддитивные модели, такие как

Yt = Tt + Ct + St + Rt

В этой модели к составляющим тренда Yt относятся: Tt – долговременный тренд, общий тренд за все время исследований; Ct – циклический эффект, колебания вокруг долговременного тренда из-за состояния бизнеса или экономики, такие как периоды общей рецессии или экспансии экономики; St – сезонный эффект, колебания из-за времени года (например, различия между зимними и летними месяцами); Rt – остаточный, или ошибочный эффект, который остается после того, как учтены долговременный, циклический и сезонный эффекты; может отражать как случайные события, так и редкие, такие как ураганы или эпидемии.

Точные значения регистрируемых на протяжении отрезка времени переменных, также известные как нескорректированные («сырые») временные ряды, почти всегда характеризуются значительной изменчивостью. Она может скрыть основные тренды, которые используются для объяснения закономерностей и точного прогнозирования. Для борьбы с этой проблемой разработаны разные методы сглаживания. Они могут быть разделены на две основные группы: *методы скользящего среднего*, при которых для ряда последовательных точек вычисляется определенный тип среднего, и это среднее вычитается из исходных значений, и *экспоненцирование*, при котором для взвешивания исходных значений используются показательные ряды.

Рассмотрены методы анализа решений (минимакс, максимин и максимакс), принятие решений в условиях риска, деревья решений (подробнее см. [Авинаш Диксит, Барри Нейлбафф. Теория игр](http://baguzin.ru/wp/?p=12139)).

Раздел, посвященный контрольным картам, содержит как смысловые ошибки автора, так и особенности перевода. Так автор сначала задает параметры процесса (в том числе среднее и границы), а затем наносит на такую карту реальные данные. На самом деле на контрольные карты сначала наносят реальные значения, и по ним вычисляют среднее и границы. Переводчик же превратил [Уолтера Шухарта](https://ru.wikipedia.org/wiki/%D0%A8%D1%83%D1%85%D0%B0%D1%80%D1%82%2C_%D0%A3%D0%BE%D0%BB%D1%82%D0%B5%D1%80) в Вальтера Шеварта… Рекомендую [Дональд Уилер, Дэвид Чамберс. Статистическое управление процессами](http://baguzin.ru/wp/?p=15577), [Пример построения контрольной карты Шухарта в Excel](http://baguzin.ru/wp/?p=1479).

### Глава 15. Статистика в медицине и эпидемиологии

В эпидемиологии и здравоохранении обычно используют два типа отношений – это *отношения рисков* и *отношения благоприятных исходов*.

Отделяют вновь возникшие случаи заболевания от уже существующих, измеряя два показателя: распространенность заболевания и заболеваемость. *Распространенность заболевания* характеризует число случаев заболевания, которые отмечены в данной популяции в определенный момент времени. *Заболеваемость* характеризует число новых случаев заболевания или состояния, развившиеся в группе риска за определенный интервал времени.

Также рассмотрены отношения шансов. Шансы некоторого события – это просто другой способ выражения его правдоподобности, сходный с вероятностью; разница заключается в том, что вероятность вычисляется при помощи деления числа событий на общее число испытаний, а шансы вычисляются как отношение числа событий к числу не-событий.

Рассмотрены искажения, послойный анализ и коэффициент Мантеля-Гензеля; анализ мощности; вычисление размера выборки.

### Глава 16. Статистика в образовании и психологии

Рассмотрены перцентили, стандартизованные баллы, классическая теория тестов, надежность тестов, «натаскивание» на тесты, современная теория тестирования (о последней теме см. [Использование модели Раша в пересчете баллов ЕГЭ](http://baguzin.ru/wp/?p=3370)).

### Глава 17. Управление данными

Использование статистики обычно подразумевает анализ данных, а надежность статистических результатов во многом зависит от надежности проанализированных данных, так что, если вы будете использовать статистику, вам нужно знать что-то об управлении данными вне зависимости от того, будете ли вы заниматься этим сами или поручите кому-то другому. Качество анализа частично зависит от качества данных, факт, блестяще сформулированный программистами: «мусор на входе – мусор на выходе» (garbage in, garbage out, GIGO).

Рассмотрена иерархическая структура, обеспечивающая сбор данных, кодификатор информации, электронные таблицы и реляционные базы данных, текстовые и числовые данные, пропущенные данные.

### Глава 18. Планирование исследования

Рассмотрены основы планирования эксперимента, разведочный анализ, слепой метод, объединение в блоки и латинский квадрат.

### Глава 19. Представление статистических материалов

### Глава 20. Оценка работ по статистике других авторов

### Приложения

Включают обзор основных математических понятий, краткий обзор статистических пакетов (Minitab, SPSS, SAS, язык программирования R, Microsoft Excel), ссылки на литературу по разделам (к сожалению, не указаны книги, переведенные на русский язык), таблицы вероятностей для распространенных типов распределений (с очень полезными пояснениями; хотя сами таблицы не так интересны, учитывая наличие функций в Excel), Интернет-ресурсы, словарь статистических терминов.