Урок 5

Параметрические критерии

Этот урок посвящен параметрическим критериям проверки гипотез. Эти критерии называются параметрическими потому, что в проверяемых ими гипотезах высказывается предположение о значении параметра распределений, из которых предположительно взята выборка.

5.1. Одновыборочные критерии Стьюдента

Семейство критериев Стьюдента позволяет проверять гипотезы о математических ожиданиях нормальных распределений.

Пример: средний вес детей при рождении. Средний вес детей при рождении составляет 3300 г. В то же время, если мать ребёнка живёт за чертой бедности, то средний вес таких детей — 2800 г. Вес при рождении — это очень важный показатель здоровья ребенка. Так, только 7% детей рождаются с весом меньше 2.5 кг, однако на них приходится 70% детских смертей.

С целью увеличить вес тех детей, чьи матери живут за чертой бедности, разработана экспериментальная программа ведения беременности. Чтобы проверить ее эффективность, проводится эксперимент. В нем принимают участие 25 женщин, живущих за чертой бедности. У всех них рождаются дети, и их средний вес составляет 3075 г.

Для того, чтобы ответить на вопрос, эффективна ли программа, используется критерий Стьюдента.

5.1.1. Z-критерий

Информация критерии суммирована в таблице 5.1, нулевое распределение показано на рисунке 5.1. Этот критерий называется Z-критерием (как и большинство критериев, статистики которых имеют стандартное нормальное нулевое распределение).

выборка:	$X^n = (X_1, \ldots, X_n)$, $X \sim N(\mu, \sigma^2)$, σ известна;
нулевая гипотеза:	$H_0 \colon \mu = \mu_0$;
альтернатива:	$H_1 \colon \mu <\neq> \mu_0$;
статистика:	$Z(X^n) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$;
нулевое распределение:	$Z(X^n) \sim N(0,1)$.

Таблица 5.1: Описание Z-критерия

Способ подсчёта достигаемого уровня значимости такого критерия зависит от используемого типа альтернативы. Если альтернатива односторонняя:

$$H_1 \colon \mu < \mu_0,$$

то, если она справедлива, более вероятными являются маленькие значения Z-статистики, то есть, левый хвост нулевого распределения (рисунок 5.2). Таким образом, чтобы посчитать достигаемый уровень значимости, нужно взять интеграл плотности стандартного нормального распределения от $-\infty$ до значения статистики Z,

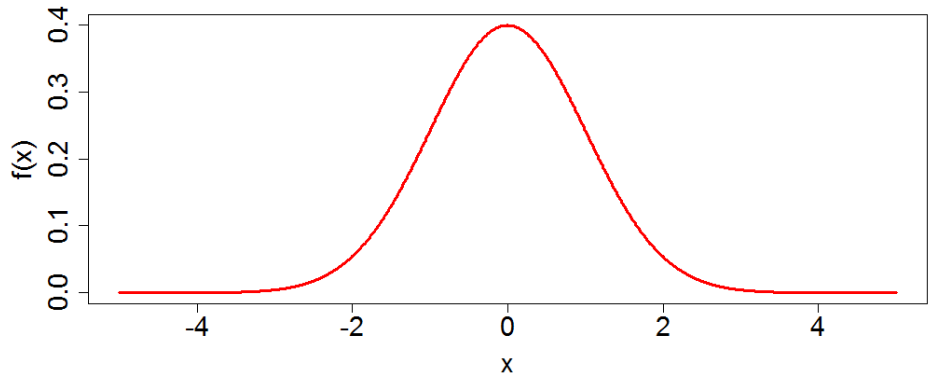

Рис. 5.1: Стандартное нормальное распределение

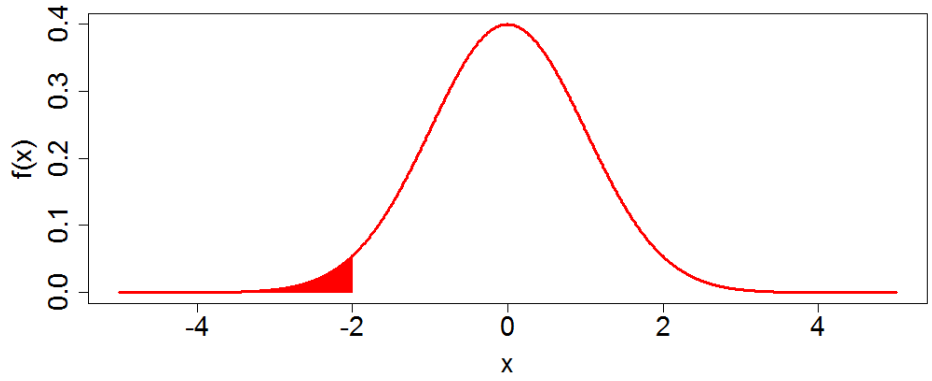

Рис. 5.2: Более вероятные значения статистики при использовании альтернативы $H_1 \colon \mu < \mu_0$

реализовавшегося в эксперименте. Например, если в эксперименте получено значение $Z = -2$, то достигаемый уровень значимости — это интеграл плотности стандартного нормального распределения от $-\infty$ до -2. На самом деле, значение интеграла вычислять не нужно, потому что оно в точности равно значению функции стандартного нормального распределения в точке Z:

$$p = F_{N(0,1)}\left(z\right).$$

Если используется противоположная односторонняя альтернатива вида

$$H_1 \colon \mu > \mu_0,$$

то более вероятными являются большие значения статистики (рис. 5.3), и, чтобы посчитать достигаемый уровень значимости, нужно взять интеграл по правому хвосту плотности нулевого распределения. Этот интеграл, в свою очередь, равен

$$p = 1 - F_{N(0,1)}\left(z\right).$$

Если же используется двусторонняя альтернатива

$$H_1 \colon \mu \neq \mu_0,$$

то при ее справедливости более вероятными будут большие по модулю значения статистики Z (рис. 5.4). Поэтому при подсчете достигаемого уровня значимости представляют интерес и левый, и правый хвосты нулевого распределения. Если получено значение статистики $Z = -2$, то достигаемый уровень значимости равен сумме интегралов от $-\infty$ до -2 и от 2 до ∞. Чтобы не считать эти два интеграла, можно снова использовать функцию стандартного нормального распределения:

$$p = 2\left(1 - F_{N(0,1)}\left(|z|\right)\right).$$

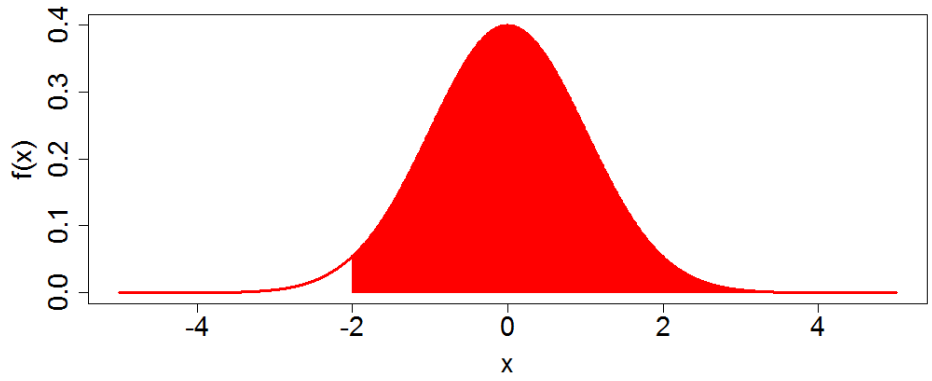

Рис. 5.3: Более вероятные значения статистики при использовании альтернативы $H_1 : \mu > \mu_0$

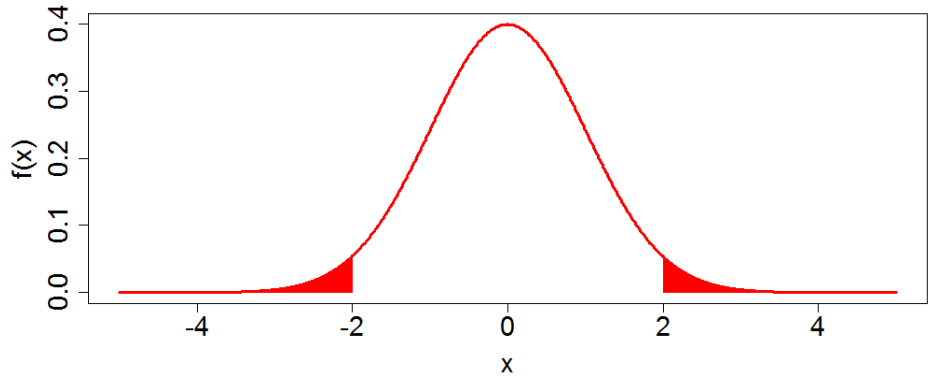

Рис. 5.4: Более вероятные значения статистики при использовании альтернативы $H_1 : \mu \neq \mu_0$

5.1.2. t-критерий

Если дисперсия выборки неизвестна, вместо Z-критерия Стьюдента нужно применять t-критерий Стьюдента (таблица 5.2). Он основан на следующей идее: поскольку σ неизвестна, то нужно там, где используется σ (в формуле статистики), заменить σ на S (выборочное стандартное отклонение). Такая статистика имеет уже не стандартное нормальное нулевое распределение, а распределение Стьюдента с числом степеней свободы $n-1$ (рис. 5.5).

выборка:	$X^n = (X_1, \ldots, X_n)$, $X \sim N\left(\mu, \sigma^2\right)$, σ неизвестна;
нулевая гипотеза:	$H_0 : \mu = \mu_0$;
альтернатива:	$H_1 : \mu <\neq> \mu_0$;
статистика:	$T\left(X^n\right) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$;
нулевое распределение:	$T\left(X^n\right) \sim St(n-1)$.

Таблица 5.2: Описание t-критерия

Достигаемый уровень значимости для t-критерия Стьюдента считается абсолютно так же, как и для Z-критерия. В зависимости от типа альтернативы нужно выбрать одно из трех выражений для достигаемого уровня значимости:

$$p = \begin{cases} F_{St(n-1)}\left(t\right), & H_1 : \mu < \mu_0, \\ 1 - F_{St(n-1)}\left(t\right), & H_1 : \mu > \mu_0, \\ 2\left(1 - F_{St(n-1)}\left(|t|\right)\right), & H_1 : \mu \neq \mu_0. \end{cases}$$

Единственное отличие от Z-критерия заключается в том, что вместо функции стандартного нормального распределения используется функция распределения Стьюдента с числом степеней свободы $n-1$.

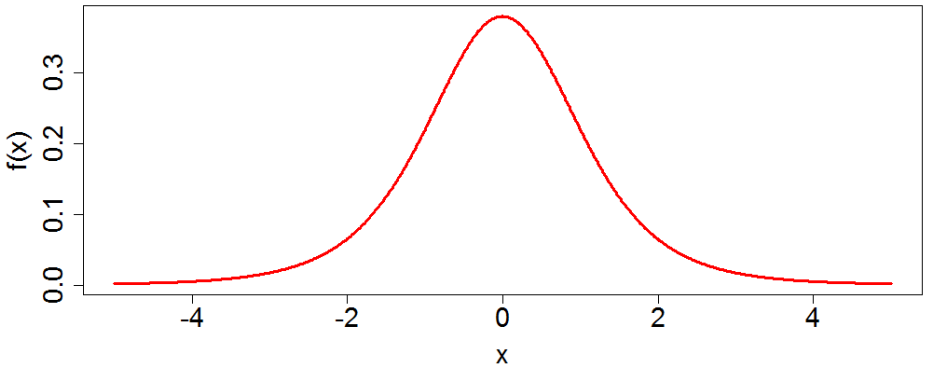

Рис. 5.5: Распределение Стьюдента

Чем больше объем выборки в задаче, тем меньше различий между t-критерием и Z-критерием. Это происходит по двум причинам: во-первых, чем больше n, тем точнее выборочная дисперсия S^2 оценивает теоретическую дисперсию σ^2. Во-вторых, с ростом n увеличивается число степеней свободы у нулевого распределения t-критерия, а чем больше степеней свободы у распределения Стьюдента, тем больше оно похоже на стандартное нормальное. Ранее уже упоминалось, что, начиная с 30 степеней свободы, распределение Стьюдента визуально практически неотличимо от стандартного нормального. Благодаря этим двум фактам для достаточно больших выборок знание истинного значения дисперсии не оказывает большое влияние на результат.

5.1.3. Математическая формулировка задачи о весе детей при рождении

Пример: вес детей при рождении (продолжение) С использованием описанных критериев можно математически сформулировать задачу оценки эффективности экспериментальной программы, о которой шла речь ранее.

Выдвигается нулевая гипотеза о том, что программа неэффективна:

$$H_0 \colon \mu = 2800,$$

то есть средний вес детей, прошедших экспериментальную программу, такой же, как и в целом у детей, живущих за чертой бедности. Эту нулевую гипотезу необходимо проверить против двусторонней альтернативы (программа как-то влияет на вес детей):

$$H_0 \colon \mu \neq 2800.$$

Поскольку теоретическая дисперсия σ^2 неизвестна, а известна только ее выборочная оценка S^2, то нужно использовать t-критерий Стьюдента. С его помощью для такой пары гипотеза-альтернатива получается достигаемый уровень значимости $p = 0.0111$, то есть нулевая гипотеза отклоняется. Точечная оценка для прироста среднего веса детей в результате экспериментальной программы — это $\mu - \mu_0 = 275$ г. 95% доверительный интервал для этой величины получается из применённого критерия Стьюдента, и он составляет $[233.7, 316.3]$ г.

Вообще говоря, в этой задаче нулевую гипотезу можно проверять против не двусторонней, а односторонней альтернативы. Тогда нулевая гипотеза остаётся той же (программа неэффективна):

$$H_0 \colon \mu = 2800,$$

а альтернатива — «программа эффективна», то есть средний вес детей в результате программы повышается:

$$H_0 \colon \mu > 2800.$$

Для такой пары гипотеза-альтернатива t-критерий дает достигаемый уровень значимости ровно в 2 раза меньше: $p = 0.0056$. Точечная оценка для прироста среднего веса не меняется и составляет все еще 375 граммов. А вот доверительный интервал становится односторонним, то есть утверждается, что на уровне доверия 95% средний вес детей увеличивается не меньше, чем на 241 грамм.

5.1.4. Выбор альтернативы

Рассмотренный пример показывает, что, используя одностороннюю альтернативу вместо двусторонней, можно получить достигаемый уровень значимости в два раза меньше. В данной задаче это было не критично, поскольку оба достигаемых уровня значимости маленькие. Но иногда может оказаться, что p-value при двусторонней альтернативе больше магического порога в 0.05, а при односторонней альтернативе — меньше. То есть, используя одностороннюю альтернативу вместо двухсторонней, можно отвергнуть нулевую гипотезу. В таком случае кажется, что можно всегда использовать одностороннюю альтернативу, однако это нечестно. Альтернатива может быть односторонней только в некоторых случаях. Во-первых, если среднее должно изменится в какую-то определенную сторону и изменение в противоположную сторону невероятно. Во-вторых, направление изменения нужно определить до получения данных. Если односторонняя альтернатива выбирается после того, как данные получены, и она выбирается так, что ее знак соответствует знаку изменения выборочного среднего относительно μ_0, то это — переобучение, и так делать нельзя.

5.2. Двухвыборочные критерии Стьюдента, независимые выборки

С помощью двухвыборочных критериев Стьюдента можно сравнивать среднее значение двух выборок из нормального распределения. Использование этих критериев будет продемонстрировано на данных General Social Survey. Это социологический опрос, который проводится на достаточно больших выборках в США уже больше 40 лет. В этом опросе очень много вопросов, которые задают респондентам, здесь будет рассматриваться только один.

В 1974 году число респондентов, работающих неполный рабочий день, составляло 108. В 2014 году — 196. Для каждого из опрошенных известно количество рабочих часов за неделю, предшествующую опросу. Используя эти данные, требуется понять, изменилось ли за прошедшие 40 лет среднее время работы у работающих неполный день.

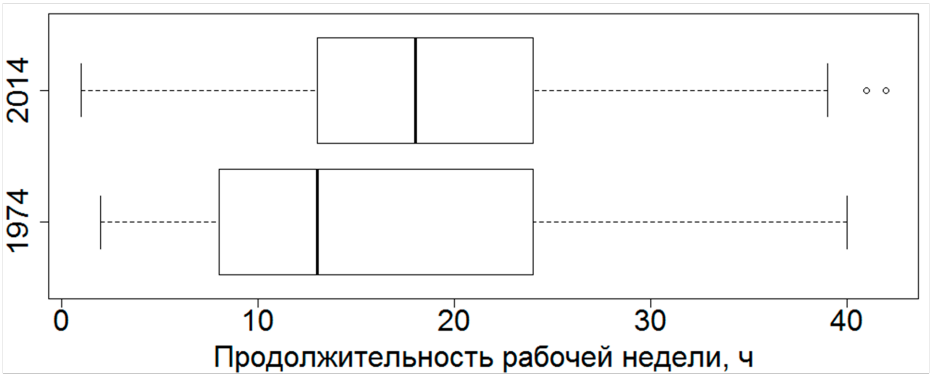

Рис. 5.6: Количество рабочих часов у работающих неполный день в 1974 году и в 2014 году

Данные, которые требуется проанализировать, показаны на рисунке 5.6. Изображённый график называется boxplot, или ящик с усами. Это способ визуализации основных характеристик распределения, принцип построения такого графика показан на рисунке 5.7. Boxplot состоит из прямоугольника — ящика — и торчащих из него усов. Черта в середине прямоугольника соответствует выборочной медиане выборки. Ширина ящика равна интерквартильному размаху, то есть, его нижняя граница — это 25%-й квантиль, а верхняя — 75%-й квантиль. Длина усов составляет 1.5 интерквартильных размаха, однако в разных реализациях кончик уса может рисоваться в разных местах. Так, на рисунке 5.7 усы обрезаются так, что их конец соответствует последнему элементу выборки в этом направлении. Два кружочка на верхнем графике на рисунке 5.6 — это объекты выборки, не попавшие в диапазон 1.5 интерквартильных размаха.

Из рисунка 5.6 видно, что выборочные медианы выборок, соответствующих 1974 и 2014 году, отличаются. В 2014 году люди работали в среднем больше. Для того, чтобы проверить, значимо ли это различие,

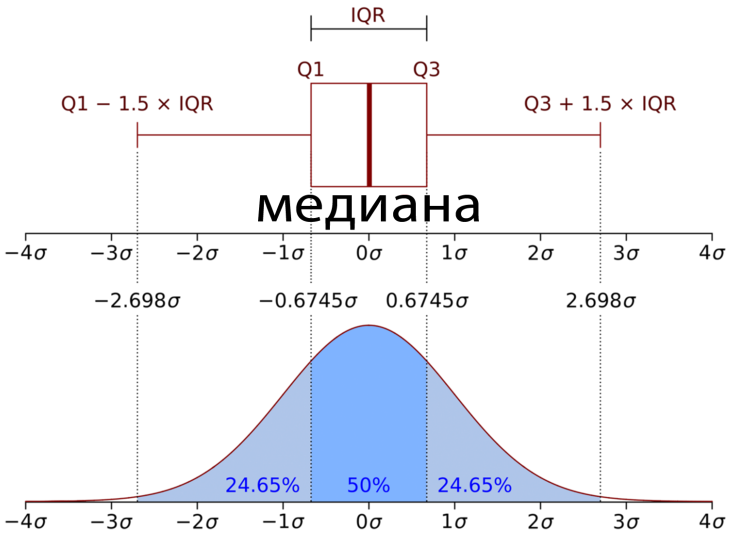

Рис. 5.7: Структура "ящика с усами"

необходимо использовать статистический критерий.

5.2.1. Z-критерий

Если имеются две выборки из нормального распределения с различными параметрами, дисперсии для которых известны, то используется Z-критерий (таблица 5.3). Статистика этого критерия имеет стандартное нормальное распределение (рисунок 5.1).

выборки:	$X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$, $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$, $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$, σ_1, σ_2 известны;
нулевая гипотеза:	$H_0 \colon \mu_1 = \mu_2$;
альтернатива:	$H_1 \colon \mu_1 <\neq> \mu_2$;
статистика:	$Z(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$;
нулевое распределение:	$Z(X_1^{n_1}, X_2^{n_2}) \sim N(0,1)$.

Таблица 5.3: Описание двухвыборочного Z-критерия

5.2.2. t-критерий

В более сложном случае дисперсии выборок неизвестны. Можно действовать по аналогии с одновыборочным случаем: в формуле для Z-критерия заменить все неизвестные σ на их выборочные оценки S_1 и S_2. Получится t-статистика (таблица 5.4). При выполнении нулевой гипотезы она будет распределена по Стьюденту (рисунок 5.5).

У этой задачи есть две проблемы. Во-первых, число степеней свободы ν у этого нулевого распределения Стьюдента вычисляется по достаточно сложной формуле:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}.$$

Во-вторых, нулевое распределение t-статистики не точное, а приближенное. Точного решения (то есть точного нулевого распределения для такой статистики) не существует. Эта проблема называется проблемой Беренца-Фишера: невозможно точно сравнить средние значение в двух выборках, дисперсии которых неизвестны.

выборки:	$X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$, $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$, $X_1 \sim N\left(\mu_1, \sigma_1^2\right), X_2 \sim N\left(\mu_2, \sigma_2^2\right)$, σ_1, σ_2 неизвестны;
нулевая гипотеза:	$H_0 \colon \mu_1 = \mu_2$;
альтернатива:	$H_1 \colon \mu_1 <\neq> \mu_2$;
статистика:	$T\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$,
нулевое распределение:	$T\left(X_1^{n_1}, X_2^{n_2}\right) \approx\sim St(\nu)$.

Таблица 5.4: Двухвыборочный t-критерий

Однако рассмотренная аппроксимация достаточно точна в двух ситуациях. Во-первых, если выборки X_1 и X_2 одинакового объема, то есть $n_1 = n_2$. Во-вторых, если знак неравенства между n_1 и n_2 такой же, как между σ_1 и σ_2, то есть выборка с большей дисперсией по размеру не может быть меньше другой выборки. Если это условие выполнено, то можно использовать t-критерий Стьюдента и не переживать о точности аппроксимации. Если это не так, то возникает проблема: критерий Стьюдента перестает правильно работать, и вероятность ошибок первого рода начинает превышать уровень значимости α. Это проблема не только критерия Стьюдента, она возникает при проверке любым способом гипотезы о равенстве средних в двух выборках с разной дисперсией. Поэтому, сравнивая средние значения в двух выборках, важно всегда следить за тем, чтобы выборка с большей дисперсией всегда была не меньшего объема, чем вторая выборка.

Итак, необходимо проверить нулевую гипотезу о том, что средняя продолжительность рабочей недели у людей, которые работают не полный рабочий день, не изменилась за прошедшие 40 лет:

$$H_0 \colon \mu_1 = \mu_2.$$

Альтернативная гипотеза двусторонняя, среднее время работы изменилось:

$$H_0 \colon \mu_1 \neq \mu_2.$$

Можно было бы использовать и одностороннюю альтернативу. Однако сложно заранее предугадать знак сравнения, и данные уже известны, так что выбирать одностороннюю альтернативу нечестно.

Критерий Стьюдента в этой задаче дает достигаемый уровень значимости $p = 0.02707$. То есть, гипотеза о равенстве средних отвергается на уровне значимости 0.05. Точечная оценка для прироста средней продолжительности рабочей недели составляет 2.57 часов. 95%-й доверительный интервал для нее: $[0.29, 4.85]$ ч. То есть, люди в среднем стали работать больше, и доверительный интервал прироста этого времени составляет от получаса до 5 часов.

5.3. Двухвыборочные критерии Стьюдента, связанные выборки

5.3.1. Лечение СДВГ

Проводится исследование метода лечения синдрома дефицита внимания и гиперактивности (СДВГ) у умственно отсталых детей. В эксперименте участвуют 24 ребенка. Каждый из них неделю принимает плацебо, а неделю препарат метилфенидат. По окончании каждой недели каждый ребенок проходит тест на способность к подавлению импульсивных поведенческих реакций. Анализируемые данные показаны на диаграмме рассеяния, (см. рисунок 5.8). По горизонтальной оси отложена способность к подавлению импульсивных поведенческих реакций после недели приема плацебо, по вертикальной — после недели приема препарата. Каждая точка соответсвует одному ребенку. Таким образом, несмотря на то, что имеются две выборки, они не являются независимыми, поскольку значения здесь измерены на одних и тех же объектах. Такие выборки называются связанными.

Хочется понять, эффективно ли лечение с помощью метилфенидата. Большая часть точек на этом графике лежит выше диагонали. Это значит, что после приема метилфенидата у большинства детей способность к подавлению импульсивных поведенческих реакций увеличилась. Для того, чтобы определить, значимо ли это изменение, необходимо использовать статистический критерий.

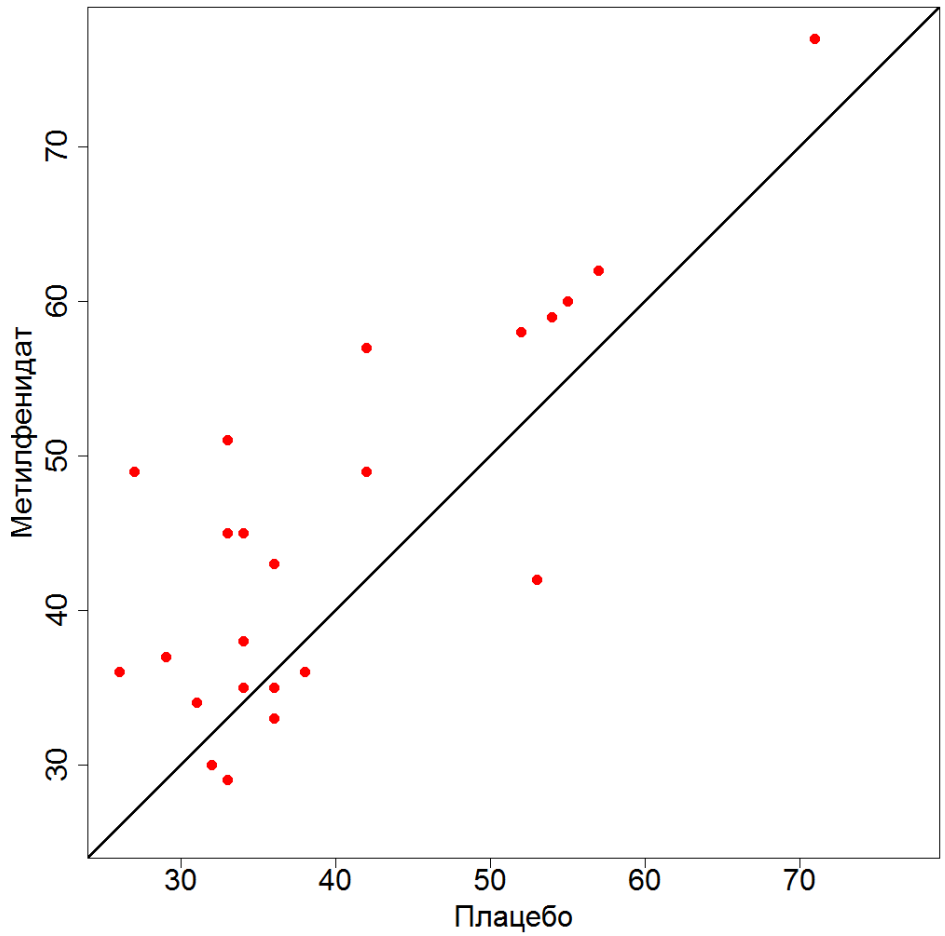

Рис. 5.8: Результаты эксперимента по сравнению действия плацебо и препарата метилфенидат на умственно-отсталых детей с синдромом дефицита внимания и гиперактивности

5.3.2. t-критерий Стьюдента для связанных выборок

Для проверки равенства математических ожиданий двух выборок одинакового объёма из нормальных распределений используется t-критерий Стьюдента (таблица 5.5).

выборки:	$X_1^n = (X_{11}, \ldots, X_{1n}), X_1 \sim N\left(\mu_1, \sigma_1^2\right),$ $X_2^n = (X_{21}, \ldots, X_{2n}), X_2 \sim N\left(\mu_2, \sigma_2^2\right),$
нулевая гипотеза:	$H_0 \colon \mu_1 = \mu_2;$
альтернатива:	$H_1 \colon \mu_1 <\neq> \mu_2;$
статистика:	$T\left(X_1^n, X_2^n\right) = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}},$ $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2, D_i = X_{1i} - X_{2i};$
нулевое распределение:	$T\left(X_1^n, X_2^n\right) \sim St(n-1).$

Таблица 5.5: Описание t-критерия Стьюдента для связанных выборок

В числителе T-статистики стоит разность $\bar{X}_1 - \bar{X}_2$. Это то же самое, что $\overline{X_1 - X_2}$. Таким образом, t-критерий для двух связанных выборок эквивалентен одновыборочному t-критерию, примененному к выборке попарных разностей.

5.3.3. Применение t-критерия Стьюдента к задаче лечения СДВГ

В описанной ранее задаче нулевая гипотеза — это неэффективность лечения (способность к подавлению импульсивных поведенческих реакций не изменилась):

$$H_0 \colon \mu_1 = \mu_2.$$

Проверить эту гипотезу нужно против двусторонней альтернативы поскольку нельзя исключать, что способность к подавлению импульсивных поведенческих реакций в результате применения препарата может уменьшится:

$$H_1 \colon \mu_1 \neq \mu_2.$$

t-критерий Стьюдента для связанных выборок дает значение достигаемого уровня значимости $p = 0.00377$. Нулевая гипотеза о том, что средняя способность к подавлению импульсивных поведенческих реакций не изменилась, отвергается на уровне значимости 0.05. Точечная оценка изменения признака в результате применения препарата (разность выборочных средних) — 4.95 пунктов. 95% доверительный интервал для этой величины, построенный с помощью распределения Стьюдента: $[1.78, 8.14]$ пунктов.

5.4. Нормальность выборок

5.4.1. Критерий хи-квадрат

Критерии Стьюдента проверяют гипотезы о средних значениях выборок в предположении, что эти выборки взяты из нормального распределения. Нормальность можно проверять с помощью критерия согласия Пирсона, или критерия хи-квадрат (таблица 5.6).

выборка:	$X^n = (X_1, \ldots, X_n)$;
нулевая гипотеза:	$H_0 \colon X \sim N\left(\mu, \sigma^2\right)$;
альтернатива:	$H_1 \colon H_0$ неверна;
статистика:	$\chi^2\left(X^n\right) = \sum\limits_{i=1}^{K} \frac{(n_i - np_i)^2}{np_i}$;
нулевое распределение:	$\chi^2\left(X^n\right) \sim \begin{cases} \chi^2_{K-1}, & \mu, \sigma \text{ заданы}, \\ \chi^2_{K-3}, & \mu, \sigma \text{ оцениваются}; \end{cases}$
	n_i — число элементов выборки в $[a_i, a_{i+1}]$,
	$p_i = F_{N(\mu,\sigma^2)}\left(a_{i+1}\right) - F_{N(\mu,\sigma^2)}\left(a_i\right)$.

Таблица 5.6: Описание критерия хи-квадрат

Статистика критерия конструируется следующим образом: область изменения случайной величины разбивается на K интервалов (карманов). Границы этих интервалов задаются величинами a_i. Для каждого интервала $[a_i, a_{i+1}]$ вычисляются две величины. Во-первых, n_i — число элементов выборки, которое попало в интервал. Во-вторых, p_i — теоретическая вероятность попадания в этот интервал при условии справедливости нулевой гипотезы. В данном случае это разность функций нормального распределения в точках a_{i+1} и a_i:

$$p_i = F_{N(\mu,\sigma^2)}\left(a_{i+1}\right) - F_{N(\mu,\sigma^2)}\left(a_i\right).$$

Значение статистики выглядит следующим образом:

$$\chi^2\left(X^n\right) = \sum_{i=1}^{K} \frac{(n_i - np_i)^2}{np_i}.$$

Если нулевая гипотеза справедлива, то такая статистика имеет распределение хи-квадрат (рисунок 5.9).

Чтобы вычислить достигаемый уровень значимости, необходимо взять интеграл от распределения хи-квадрат, начиная от значения статистики, которое реализуется в данных, до бесконечности.

Критерий хи-квадрат обладает несколькими очевидными недостатками. Во-первых, разбиение на интервалы в нём никак не зафиксировано, и, выбирая различные интервалы, можно получать разные результаты. Кроме того, для того, чтобы его использовать, необходимо иметь достаточно большую выборку: ожидаемое количество объектов выборки np_i в каждом интервале должно превышать 5 как минимум для 80% ячеек.

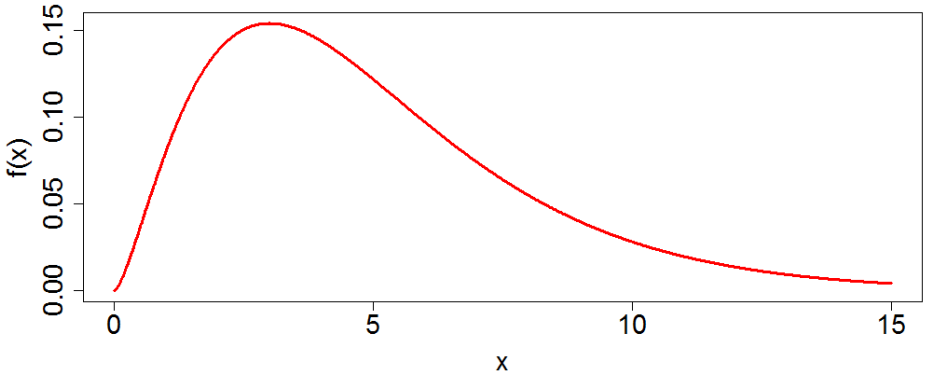

Рис. 5.9: Распределение хи-квадрат

5.4.2. Q-Q график

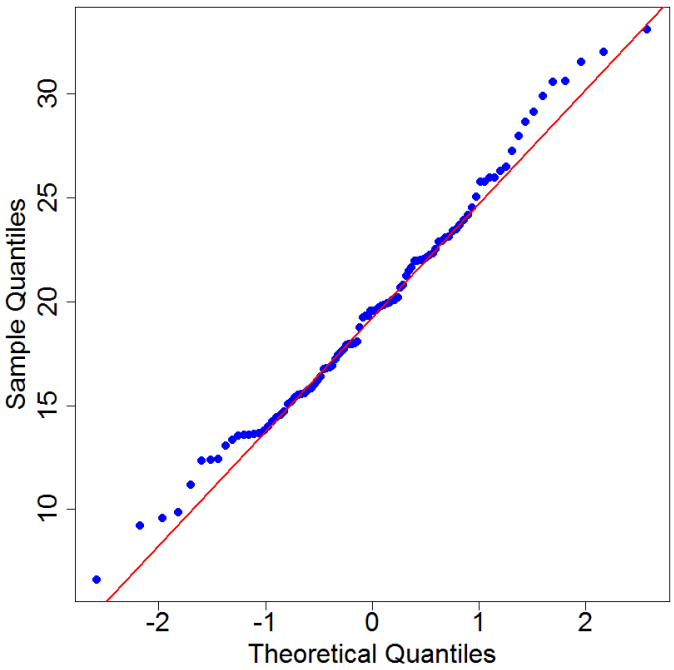

Рис. 5.10: Q-Q график

Очень удобный способ визуальной проверки предположения нормальности — Q-Q график (рисунок 5.10). Чтобы построить такой график, выборку нужно превратить в вариационный ряд, то есть отсортировать по неубыванию, а дальше каждому объекту выборки сопоставить точку на графике. Значение по вертикальной оси соответствует значению X, а значение по горизонтальной оси — математическому ожиданию квантиля стандартного нормального распределения, посчитанного по выборке такого объема.

Чтобы это лучше понять, можно посмотреть на точку в нижнем левом углу. Эта точка соответствует наименьшему значению в выборке. Пусть объём выборки — 100. Таким образом, эта точка — это минимум из всех 100 элементов. Значение этого минимума отложено по вертикальной оси. По горизонтальной оси отложено математическое ожидание минимума из 100 независимых одинаково распределенных случайных величин из стандартного нормального распределения. Если выборка взята из нормального распределения, точки на Q-Q графике должны лежать примерно на прямой. Если точки лучше описываются нелинейной кривой или какие-то из точек лежат от прямой очень далеко, скорее всего, распределение отличается от нормального.

5.4.3. Критерий Шапиро-Уилка

Критерий Шапиро-Уилка (таблица 5.7) — это ещё один способ формально проверить соответствие распределения выборки нормальному. Этот критерий основан на Q-Q графике. Фактически он проверяет, насколько сильно точки на Q-Q графике отклоняются от прямой.

выборка:	$X^n = (X_1, \ldots, X_n)$;
нулевая гипотеза:	$H_0 \colon X \sim N\left(\mu, \sigma^2\right)$;
альтернатива:	$H_1 \colon H_0$ неверна;
статистика:	$W\left(X^n\right) = \frac{\left(\sum\limits_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum\limits_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$,
нулевое распределение:	табличное.

Таблица 5.7: Описание критерия Шапиро-Уилка

Статистика W рассчитывается на основании вариационного ряда, полученного из выборки, и некоторых величин a. Эти величины основаны на математических ожиданиях порядковых статистик из стандартного нормального распределения, они табулированы, для них не существует аналитических выражений. Кроме того, табулировано и нулевое распределение статистики критерия Шапиро-Уилка, то есть его невозможно записать аналитически. Используя таблицы этих величин, можно вычислить достигаемый уровень значимости.

5.4.4. Зачем проверять нормальность?

Для проверки гипотезы нормальности существуют еще десятки других критериев: критерий Харке-Бера, Колмогорова, он же Лиллиефорса, Крамера-фон Мизеса, Андерсона-Дарлинга, и т. д. Чтобы понять, какие из этих критериев лучше использовать, необходимо вернуться на шаг назад и вспомнить, зачем нужно формально проверять нормальность.

Дело в том, что проверка гипотезы нормальности наследует плохие свойства всего аппарата проверки гипотез: на маленьких выборках нулевая гипотеза, как правило, не отклоняется, а на выборках огромного размера — практически наверняка отклоняется. То есть, если выборка маленькая, то, формально проверяя гипотезу о нормальности, её не получается отклонить, а если выборка огромна, то гипотеза отклоняется, даже если распределение отличается от нормального совсем чуть-чуть.

Многие методы, предполагающие нормальность, в том числе критерии Стьюдента, нечувствительны к небольшим отклонениям от нормальности, то есть истинное распределение выборки может слегка отличаться от нормального, и t-критерий будет всё еще правильно работать. Нормальное распределение — это математический конструкт. Никаких нормальных выборок в природе не существует. Однако, как говорил Джордж Бокс: «Все модели неверны, а некоторые полезны» — а нормальные модели очень полезны, поэтому их имеет смысл использовать.

5.4.5. Как проверять нормальность?

В итоге предлагается использовать следующий алгоритм. Если анализируемые данные имеют распределение, явно отличающееся от нормального (например, выборка бинарна или измеряемый признак — категориальный), не нужно применять метод, предполагающий нормальность. Лучше использовать метод, специально разработанный для такого распределения. Если исследуемый признак, по крайней мере, измерен в непрерывной шкале, можно построить Q-Q график. Если на этом графике не видно существенных отклонений от нормальности (точки лежат примерно на прямой), можно использовать методы, устойчивые к небольшим отклонениям от нормальности, например, критерии Стьюдента. Если используемый метод чувствителен к отклонениям от нормальности, необходимо формально проверить нормальность, и рекомендуется это делать с помощью метода Шапиро-Уилка. Показано, что критерий Шапиро-Уилка обладает достаточно хорошей мощностью для разных классов альтернатив. Если критерий Шапиро-Уилка отвергает нормальность, не нужно использовать методы, чувствительные к отклонениям от нормальности.

5.5. Гипотезы о долях

Ещё одно семейство параметрических критериев — это критерии, которые работают с распределениями Бернулли. Они принимают на вход выборки из нулей и единиц и проверяют гипотезы о параметрах p этих распределений (вероятность появления единицы в выборке). С распределением Бернулли работать удобно потому, что, в отличие от нормального распределения, не нужно применять никаких методов, чтобы доказать, что выборка взята именно из этого распределения. Если в выборке присутствуют только 2 значения, то она взята из распределения Бернулли.

Далее будут рассмотрены критерии, решающие три задачи: одновыборочную, двухвыборочную с независимыми выборками и двухвыборочную со связанными выборками.

5.5.1. Задача о присяжных

В 70-х годах известный педиатр и автор книг по воспитанию детей Бенджамин Спок был арестован за участие в антивоенной демонстрации в Бостоне. Его дело должен был рассматривать суд присяжных. Отбор присяжных — это сложная многоступенчатая процедура. На очередном этапе остаётся 300 человек, из которых отбираются финальные 12. В процессе Бенджамина Спока среди этих 300 только 90 были женщинами, и адвокаты подали протест. Поскольку в те времена воспитанием детей занимались в основном женщины, Бенджамин Спок среди них был более популярен, поэтому адвокаты заподозрили, что обвинение специально пытается сделать финальный состав присяжных менее благосклонным к подсудимому.

5.5.2. Z-критерий для доли

Чтобы по описанным выше данным проверить, был ли отбор беспристрастным, нужно использовать статистический критерий, например, Z-критерий для доли (таблица 5.8).

выборка:	$X^n = (X_1, \ldots, X_n)$, $X \sim Ber(p)$;
нулевая гипотеза:	$H_0 \colon p = p_0$;
альтернатива:	$H_1 \colon p <\neq> p_0$;
статистика:	$Z(X^n) = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, $\hat{p} = \bar{X}_n$;
нулевое распределение:	$Z(X^n) \sim N(0,1)$.

Таблица 5.8: Описание Z-критерия для доли

В задаче про отбор присяжных нулевая гипотеза состоит в том, что процедура отбора беспристрастна, женщины попадают в выборку с вероятностью 0.5; альтернатива — двусторонняя. Эта нулевая гипотеза отвергается: достигаемый уровень значимости $p = 4.6 \times 10^{-12}$. Точечная оценка вероятности попадания женщин в выборку составляет 0.3. 95% интервал для этой вероятности: $[0.248, 0.352]$. Выборка достаточно большая, поэтому неизвестную долю можно оценить с погрешностью порядка 10 %.

5.5.3. Рейтинг премьер-министра

1600 гражданам Великобритании с правом голоса задают вопрос: одобряют ли они деятельность премьер-министра. 944 человека говорят, что одобряют. Через 6 месяцев опрос повторяется. На этот раз из 1600 опрошенных 880 говорят, что поддерживают премьер-министра. Чтобы понять, изменился ли рейтинг премьер-министра, нужно использовать статистический критерий.

5.5.4. Z-критерий для доли для двух независимых выборок

Для решения предыдущей задачи можно использовать Z-критерий для двух долей (таблица 5.9).

Данные в подобных задачах можно записать при помощи таблицы сопряженности 2×2 (таблица 5.10). В ней в столбцах расположены выборки, а в строках — исходы.

выборки:	$X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_1 \sim Ber(p_1)$; $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_2 \sim Ber(p_2)$, выборки независимы;
нулевая гипотеза:	$H_0 \colon p_1 = p_2$;
альтернатива:	$H_1 \colon p_1 <\neq> p_2$;
статистика:	$Z(X_1^{n_1}, X_2^{n_2}) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, $P = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$;
нулевое распределение:	$Z(X_1^{n_1}, X_2^{n_2}) \sim N(0,1)$.

Таблица 5.9: Описание Z-критерия для доли для двух независимых выборок

Исход \ Выборка	X_1	X_2
1	a	b
0	c	d
$\sum$	n_1	n_2

Таблица 5.10: Таблица сопряжённости

Если выборки независимы, то из четырёх чисел a, b, c, d в таблице сопряжённости Z-критерий использует только два, стоящие в первой строчке (количество единиц в первой и во второй выборках):

$$\hat{p}_1 = \frac{a}{n_1},\ \hat{p}_2 = \frac{b}{n_2}.$$

Результат \ Опрос	I	II
+	$a = 944$	$b = 880$
-	$c = 656$	$d = 720$
$\sum$	$n_1 = 1600$	$n_2 = 1600$

Таблица 5.11: Таблица сопряжённости для задачи о рейтинге премьер-министра

В задаче оценки изменения рейтинга премьер-министра (таблица 5.11) нулевая гипотеза о том, что рейтинг не изменился против двусторонней альтернативы отвергается Z-критерием с достигаемым уровнем значимости $p = 0.022$. Рейтинг упал на 4 %, 95% доверительный интервал — $[0.6, 7.4]\%$

5.5.5. Z-критерий для доли для двух связанных выборок

На самом деле, в двух рассматриваемых опросах участвовали одни и те же люди, то есть, выборки являются связанными, поскольку значения признаков измерены на одних и тех же объектах. Таблица 2×2, с помощью которой записываются данные, слегка меняет свой вид (таблица 5.12).

I \ II	+	-	$\sum$
+	794	150	944
-	86	570	656
$\sum$	880	720	1600

Таблица 5.12: Таблица сопряженности для случая двух связанных выборок в задаче о рейтинге премьер-министра

Теперь в строках таблицы находятся результаты первого опроса, в столбцах – результаты второго, а в каждой ячейке — количество людей, которые в первом и втором опросе ответили именно так (например, 794 человека поддерживали премьер-министра в обоих опросах, 150 — только в первом и т.д.).

Чтобы использовать новые данные для проверки гипотезы о том, что рейтинг не изменился, нужно применить модифицированную версию Z-критерия для связанных выборок (таблица 5.13). Новые обозначения приведены в таблице 5.14.

выборки:	$X_1^n = (X_{11}, \ldots, X_{1n}), X_1 \sim Ber(p_1)$; $X_2^n = (X_{21}, \ldots, X_{2n}), X_2 \sim Ber(p_2)$, выборки связанные;
нулевая гипотеза:	$H_0 \colon p_1 = p_2$;
альтернатива:	$H_1 \colon p_1 <\neq> p_2$;
статистика:	$Z(X_1^n, X_2^n) = \frac{f-g}{\sqrt{f+g-\frac{(f-g)^2}{n}}}$;
нулевое распределение:	$Z(X_1^n, X_2^n) \sim N(0,1)$.

Таблица 5.13: Описание Z-критерия для связанных выборок

X_1 \ X_2	1	0	$\sum$
1	e	f	$e+f$
0	g	h	$g+h$
$\sum$	$e+g$	$f+h$	n

Таблица 5.14: Таблица сопряженности для случая двух связанных выборок

В Z-критерии для связанных выборок используется статистика, в которую входят только недиагональные элементы f, g таблицы 2×2, то есть только те объекты, на которых значения двух признаков отличаются. Объекты e и h, на которых значения признаков совпадают, в критерии не используются.

В задаче о рейтинге премьер-министра Z-критерий для связанных выборок уверенно отвергает нулевую гипотезу о том, что рейтинг не изменился, против двусторонней альтернативы. Достигаемый уровень значимости $p = 2.8 \times 10^{-5}$, что существенно меньше, чем без учёта связанности выборок. Точечная оценка не меняется: рейтинг упал на 4%. А вот 95-% доверительный интервал для изменения уже другой: $[2.1, 5.8]\%$. Этот интервал уже и его край дальше отстоит от 0, значения, соответствующего нулевой гипотезе.