Письма в ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ
Letters to Vavilov Journal of Genetics and Breeding
2025 • 11 • 1 • 43-50

pismavavilov.ru
doi 10.18699/letvjgb-2025-11-07

Дискуссия

Как считать критерий Стьюдента статистической значимости разности средних двух выборок. I. Проверка нормальности

В.М. Ефимов [1, 2, 3, 4]✉

Аннотация: Критерий Стьюдента статистической значимости разности средних двух выборок, предложенный Р. Фишером в 1925 г., до сих пор является одним из самых востребованных методов статистического анализа. За почти столетие его использования сложилась устойчивая система рекомендаций и условий его применения, которая изложена в учебниках и руководствах по статистической обработке данных. Как правило, настоятельно требуется предварительная проверка нормальности распределения исходных выборок и равенства их дисперсий. В случае ненормальности рекомендуется использовать непараметрические методы, например критерий Манна–Уитни. В работе представлена более современная точка зрения на эту проблему, обусловленная несколькими взаимосвязанными причинами. Во-первых, за прошедшее столетие накоплен значительный практический опыт применения *t*-критерия Стьюдента, который заставляет сильно сомневаться в обязательности проверки нормальности и равенства дисперсий, а также применения ранговых критериев в случае отсутствия нормальности. Во-вторых, теория тоже не стояла на месте. Появились расчет критерия Стьюдента через точечно-бисериальный коэффициент корреляции и альтернативы методам «нормальной теории» в виде свободных от распределения процедур. В-третьих, кардинально выросли вычислительные возможности, позволяющие без дополнительных предположений моделировать в компьютере генеральные распределения исходных выборок и по ним оценивать требуемые *p*-value.

Ключевые слова: нормальное распределение; точечно-бисериальный коэффициент корреляции; свободные от распределения процедуры; *p*-value; бутстреп

Для цитирования: Ефимов В.М. Как считать критерий Стьюдента статистической значимости разности средних двух выборок. I. Проверка нормальности. *Письма в Вавиловский журнал генетики и селекции*. 2025;11(1):43-50. doi 10.18699/letvjgb-2025-11-07

Финансирование: Работа поддержана бюджетным проектом № FWNR-2022-0019.

Discussion

How to calculate the Student's test for the statistical significance of the difference between the means of two samples. I. Testing for normality

V.M. Efimov [1, 2, 3, 4]✉

Abstract: The Student's *t*-test for the statistical significance of the difference in the means of two samples, proposed by R. Fisher in 1925, is still one of the most popular methods of statistical analysis. Over almost a century of its use, a fairly stable system of recommendations and conditions for its application has developed, which is set out in textbooks and manuals on statistical data processing. As a rule, a preliminary check of the normality of the original sample distributions and the equality of their variances is urgently required. In case of abnormality, it is recommended to use nonparametric methods, for example, the Mann–Whitney test. The paper presents a more modern point of view on this problem, caused by several interrelated reasons. Firstly, a century of practical experience has accumulated in using the Student's *t*-test, which makes one strongly doubt the necessity of checking normality and equality of variances, as well as the use of rank criteria in the absence of normality. Secondly, the theory has not stood still either. An alternative to the methods of

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Институт систематики и экологии животных Сибирского отделения Российской академии наук, Новосибирск, Россия
Institute of Systematics and Ecology of Animals of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
Novosibirsk State University, Novosibirsk, Russia
[4] Национальный исследовательский Томский государственный университет, Томск, Россия
Tomsk State University, Tomsk, Russia

✉ vmefimov@gmail.com

the "normal theory" has appeared in the form of distribution-free procedures. Thirdly, over the past century, computing capabilities have increased dramatically, making it possible to model general distributions of initial samples on a computer without additional assumptions and use them to estimate the required *p*-values.

Key words: normal distribution; point-biserial correlation coefficient; free-distribution procedures; *p*-value; bootstrap

For citation: Efimov V.M. How to calculate the Student's test for the statistical significance of the difference between the means of two samples. I. Testing for normality. *Pisma v Vavilovskii Zhurnal Genetiki i Selektsii = Letters to Vavilov Journal of Genetics and Breeding.* 2025;11(1):43-50. doi 10.18699/letvjgb-2025-11-07 (in Russian)

Funding: The work was supported by BP #FWNR-2022-0019.

Введение

Критериев Стьюдента много. Самым востребованным является критерий Стьюдента статистической значимости разности средних двух выборок (далее *t*-критерий Стьюдента для двух средних), и именно его мы будем рассматривать. Задача заключается в следующем. Даны две совокупности объектов, у которых измерен некоторый количественный признак. Для каждой совокупности вычислены средние значения этого признака. Как правило, они несколько отличаются друг от друга. Требуется принять решение: считаем ли мы разницу между этими средними случайной или нет. Алгоритм должен быть пригодным для использования на практике, т. е. простым и надежным.

В рамках статистической науки проблема формулируется так: имеется нулевая гипотеза (H_0) (обычно отсутствия различий), а наблюдаемые отклонения от нее возникли по случайным причинам. Необходимо принять или отклонить нулевую гипотезу. Для этого надо оценить суммарную вероятность (*p*-value) наблюдаемых и всех менее вероятных отклонений от H_0. Если *p*-value меньше некоторого заранее установленного порога – уровня значимости (α), то нулевая гипотеза отвергается (Fisher, 1925b). На сегодняшний день приняты и повсеместно используются три стандартных уровня статистической значимости: 0.05, 0.01, 0.001.

Трехсотлетняя история возникновения и развития идеи "$p < 0.05$" обстоятельно изложена в работе (Kennedy-Shaffer, 2019). Однако в наше время этот подход вызывает очень сильный протест прикладных статистиков из-за преимущественного использования только *p*-value, и цитируемая работа является частью их мощного движения против мира "$p < 0.05$". Протестующие не призывают отменить *p*-value, но предлагают наряду с ним в обязательном порядке вычислять размер эффекта и доверительные интервалы, а также приводить научные и практические обоснования, например экономическую значимость, принимаемого решения (Wasserstein et al., 2019). Некоторые просто предлагают заменить $p < 0.05$ на $p < 0.005$ (Benjamin et al., 2018). Но редакторы и рецензенты большинства научных журналов еще не в курсе предстоящих перемен и требуют от авторов работать по привычным для всех шаблонам, следуя заветам Фишера столетней давности.

Для средних двух выборок нулевая гипотеза формулируется просто: ($\bar{x}_1 = \bar{x}_2$), или $\bar{x}_1 - \bar{x}_2 = 0$. Для каждой совокупности ($k = 1, 2$) достаточно знать число объектов (N_k), среднее ($\bar{x}_k$) и сумму квадратов отклонения от среднего (SS_k). Тогда *t*-критерий Стьюдента для двух средних вычисляется по формуле, предложенной Фишером (Fisher, 1925a):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) \cdot \sqrt{N_1 + N_2 - 2}}{\sqrt{SS_1 + SS_2}} \cdot \sqrt{\frac{N_1 N_2}{N_1 + N_2}}, \quad (1)$$

$$\bar{x}_k = \frac{\sum \bar{x}_i}{N_k}, \quad SS_k = \sum (x_i - \bar{x}_k)^2.$$

Статистика t подчиняется *t*-распределению Стьюдента с $df = N_1 + N_2 - 2$ степенями свободы. Само распределение и первый *t*-критерий Стьюдента были найдены ранее при решении более простой задачи Уильямом Сили Госсетом. 'Student' – псевдоним Госсета. Задача заключалась в оценке статистической значимости отклонения средней одной выборки от константы (Student, 1908). Позже оказалось, что этому распределению подчиняется множество критериев, полученных другими авторами (в частности, *t*-критерии статистической значимости линейного коэффициента корреляции Пирсона между двумя количественными признаками и точечно-бисериального коэффициента корреляции между количественным признаком и двоичным) (Кендалл, Стюарт, 1973), и многие из них тоже стали называться *t*-критериями Стьюдента. (Возможно, во имя справедливости и во избежание путаницы стоило бы называть критерий (1) *t*-критерием Фишера или критерием Стьюдента–Фишера.)

Проверка нормальности и равенства дисперсий

До появления компьютеров исследователи были вынуждены пользоваться печатными таблицами, в которых приводились значения нормального, χ^2, t и некоторых других распределений для разных уровней значимости α и степеней свободы *df*. Госсет лично вычислил и напечатал первые таблицы своего распределения, назвав его *z*-распределением и придавая из практических соображений особое значение малым степеням свободы, начиная с $N = 4$ (Student, 1917). Его таблицы не пользовались популярностью. Фишер переформатировал таблицу Госсета под свою концепцию оценки значимости, которая на практике почти полностью свелась к использованию "$p < 0.05$", переименовал *z*- в *t*-распределение (поскольку у него было свое *z*-распределение) и включил в свой знаменитый учебник, выдержавший около полутора десятков изданий (Fisher, 1925b). Свой выбор $\alpha = 0.05$ он обосновывал тем, что $t(0, N) = 1.96 \approx 2$ при больших N, и тогда 95 % доверительный интервал для среднего $\bar{x}$ практически равен $\pm 2\sigma_{\bar{x}}$. Фишер считал это разумным и весьма удобным для исследователей. Исследователи действительно приняли это на ура, так же как и идею "$p < 0.05$". Кроме того, Фишер придавал большое значение проверке нормальности выборок, а также равенства их дисперсий, поскольку все теоретические результаты были получены им именно при этих

предположениях, а практики применения критерия, естественно, еще не было и не могло быть. К сожалению, несмотря на уже столетие практического применения *t*-критерия Стьюдента, этого и сегодня все еще строго требуют авторы многих статистических учебников и руководств, а также редакторы и рецензенты большинства научных журналов, добавляя, что в случае ненормальности надо использовать непараметрические критерии.

Устаревшая рекомендация проверять нормальность обеих выборок – чисто математическая (Лойко и др., 2019), поскольку *t*-критерий Стьюдента для двух средних выведен Фишером именно при этом предположении. Причем это предположение нужно только для того, чтобы получить нормальность распределения разности средних. В формуле (1) вообще не используются значения самих выборок, есть только параметры, полученные из этих значений: объемы, средние и суммы квадратов отклонений от средней. Поэтому не очень-то и важно, как распределены сами значения. Но Фишер всегда стремился к математической строгости, в отличие от Госсета, которому было достаточно, чтобы формула хорошо работала на практике, даже если она просто угадана. И предложение перейти в случае ненормальности выборок к непараметрическим критериям (т. е. заменить значения их рангами) – тоже чисто математическое, распределение рангов всегда известно заранее, поэтому вопрос всегда решается до конца (Conover, 2012). Математически очень удобно. А то, что при этом происходит подмена понятий на предметном уровне и изучается уже не совсем та задача, которая на самом деле интересует практиков, математики просто не замечают, поскольку это не их проблема. Хотя разность средних рангов, вообще говоря, совсем не то же самое, что разность самих средних, например, в случае, когда в выборке много малых значений, близких друг к другу, и мало больших.

Вопрос о том, надо ли все-таки проверять нормальность, возникал много раз. Достаточно указать на 31-ю главу второго тома знаменитого трехтомника Кендалла–Стюарта, полностью посвященную этому вопросу. Авторы приходят к четкому мнению: критерии типа критерия Стьюдента, касающиеся генеральных средних, довольно нечувствительны к отклонениям от нормальности. Особенно выделяются два случая: «…если объемы выборок равны, то даже асимметрия исходного распределения вызывает малое отклонение от нормальной теории. Если исходное распределение симметрично, то критерий будет устойчив даже при неодинаковых объемах выборок» (Кендалл, Стюарт, 1973).

На практике критерий Стьюдента достаточно хорошо работает для любых реально встречающихся непрерывных распределений независимо от того, нормальны они или нет. А также для распределений, которые по определению не могут быть нормальными, например ранговых или двоичных, но, заметим, их средние являются количественными признаками. Причина проста: распределение любых средних всегда приближенно нормально в силу центральной предельной теоремы. Соответственно, разность средних тоже будет всегда распределена приближенно нормально. Поэтому даже в случае отклонения от нормальности распределения исходных выборок больших погрешностей не ожидается. Более того, еще сам Госсет указывал, что в этом случае (формула (1)) гораздо более вероятно занижение статистической значимости, а не ее завышение, так как большие отклонения в данных, увеличивая числитель, еще больше увеличивают знаменатель в силу его квадратичности. Некоторое завышение статистической значимости происходит только в случае, когда более пологие хвосты асимметричных распределений («треугольники») ориентированы навстречу друг другу, но такая ситуация крайне редко встречается в реальных данных. А отношение дисперсий в *F*-тесте Фишера, наоборот, очень неустойчиво по той же причине, и его вообще не стоит применять для проверки равенства дисперсий. Любопытно, что для того, чтобы прийти к этим выводам без строгой математики, Госсет использовал численное моделирование распределений на основе случайного выбора чисел, написанных на отдельных листочках бумаги, – по сути метод Монте-Карло, причем задолго до появления компьютеров (Lehmann, 1999). Современная наука приходит к тем же выводам (Кендалл, Стюарт, 1973; Lumley et al., 2002; Лемешко Б.Ю., Лемешко С.Б., 2008; Орлов, 2020).

Как практик, Госсет считал, что гораздо важнее оценивать размер эффекта, чем его статистическую значимость. А насчет проверки нормальности писал, обращаясь к Фишеру (через *Nature*): «…вопрос о применимости нормальной теории к ненормальному материалу имеет большое значение и заслуживает внимания как со стороны математика, так и со стороны тех из нас, в чьей компетенции лежит применение результатов его трудов в практической работе. Лично я всегда считал, возможно, без каких-либо вполне определенных оснований для этой веры, что на самом деле на распределение Стьюдента будут очень мало влиять те небольшие отклонения от нормальности, которые имеют место в большинстве биологических и экспериментальных исследований, и недавние работы с небольшими выборками подтверждают меня в этом убеждении» (Student, 1929).

Да и точка зрения Фишера со временем стала более реалистичной: «В основе статистической оценки средних лежит следующее основное положение: если некоторая величина распределена нормально с дисперсией σ^2, то средняя случайной выборки, состоящей из n наблюдений, распределена нормально с дисперсией σ^2/n. Практическое значение этого положения отчасти усиливается еще тем фактом, что даже если исходное распределение не будет точно нормальным, все же распределение средней стремится к нормальной форме при возрастании выборки. Таким образом, это положение имеет широкую область применения и остается правомерным и в тех случаях, когда мы не имеем достаточных оснований считать, что исходное распределение нормально, лишь бы при этом у нас была уверенность в том, что оно не принадлежит к тому исключительному классу распределений, при которых распределение средней не стремится к нормальному» (Фишер, 1958).

Современная Интернет-энциклопедия уже не считает, что должны быть распределены нормально сами выборки: «Для применения данного критерия необходимо, чтобы выборочные средние имели нормальное распределение». Но дальше все по-старому: «При маленьких выборках это

означает требование нормальности исходных значений. В случае применения двухвыборочного критерия для независимых выборок также необходимо соблюдение условия равенства дисперсий. Существуют, однако, альтернативы критерию Стьюдента для ситуации с неравными дисперсиями. Также не вполне корректно применять *t*-критерий Стьюдента при наличии в данных значительного числа выбросов. При несоблюдении этих условий при сравнении выборочных средних должны использоваться аналогичные методы непараметрической статистики, среди которых наиболее известными являются U-критерий Манна–Уитни (в качестве двухвыборочного критерия для независимых выборок…» (https://руни.рф/T-критерий_Стьюдента).

Как видно из этого текста, несмотря на некоторые подвижки, многие догматические требования, к сожалению, остались.

Дихотомия (разделение на две части)

Вычислять *t*-критерий Стьюдента для двух средних можно по-разному. Во втором томе трехтомника Кендалла–Стюарта приведена формула 26.80 (исправлена опечатка):

$$\frac{r_{pb}^2}{1-r_{pb}^2}=\frac{(\bar{x}_1-\bar{x}_2)^2}{SS_1+SS_2}\cdot\frac{N_1N_2}{N_1+N_2}=\frac{t^2}{N_1+N_2-2}, \qquad (2a)$$

где r_{pb} – точечно-бисериальный коэффициент корреляции между количественным признаком и двоичным(!). Она же (в словесной формулировке) приведена в Википедии (https://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient).

Из формулы (2a) следует

$$t^2=\frac{(N_1+N_2-2)\cdot r_{pb}^2}{1-r_{pb}^2}. \qquad (2b)$$

Это означает, что вычислять *t*-критерий Стьюдента для двух средних можно и через точечно-бисериальный коэффициент корреляции признаков x и b. Количественный признак (x) получается, если объединить значения обеих выборок в одну. Двоичный (b) принимает значение 1, если значение x_i объединенной выборки относится к первой из выборок, и 0, если ко второй. Между количественным признаком (x) и двоичным (b) можно вычислить точечно-бисериальный коэффициент корреляции r_{pb} либо по формуле, выведенной исходя из биномиального распределения двоичного признака b (Кендалл, Стюарт, 1973):

$$r_{pb}=\frac{(\bar{x}_1-\bar{x}_2)\sqrt{\bar{b}(1-\bar{b})}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{N}}}=\frac{(\bar{x}_1-\bar{x}_2)\sqrt{\left[\frac{N_1N_2}{N}\right]}}{\sqrt{\sum x_i^2-N\bar{x}^2}}, \qquad (3a)$$

либо по обычной формуле Пирсона для двух количественных признаков:

$$r_{pb}=\frac{\sum(x_i-\bar{x})(b_i-\bar{b})}{\sqrt{\sum(x_i-\bar{x})^2}\cdot\sqrt{\sum(b_i-\bar{b})^2}}=$$

$$=\frac{\sum x_ib_i-N\bar{x}\bar{b}}{\sqrt{\sum x_i^2-N\bar{x}^2}\cdot\sqrt{\sum b_i^2-N\bar{b}^2}}, \qquad (3b)$$

которые эквивалентны друг другу.

Тогда, используя

$$N=N_1+N_2;\quad \bar{x}=\frac{N_1\bar{x}_1+N_2\bar{x}_2}{N};\quad \bar{b}=\frac{N_1\bar{b}_1+N_2\bar{b}_2}{N}=\frac{N_1}{N},$$

$$\sum x_ib_i=N_1\bar{x}_1;\quad \sum x_i^2=SS_1+N_1\bar{x}_1^2+SS_2+N_2\bar{x}_2^2,$$

получим

$$\frac{r_{pb}^2}{1-r_{pb}^2}=\frac{(\sum x_ib_i-N\bar{x}\bar{b})^2}{(\sum x_i^2-N\bar{x}^2)\cdot(\sum b_i^2-N\bar{b}^2)-(\sum x_ib_i-N\bar{x}\bar{b})^2}=$$

$$=\frac{\left(N_1\bar{x}_1-N\left(\frac{N_1\bar{x}_1+N_2\bar{x}_2}{N}\right)\left(\frac{N_1}{N}\right)\right)^2}{(SS_1+N_1\bar{x}_1^2+SS_2+N_2\bar{x}_2^2-N\bar{x}^2)\left(N_1-N\left(\frac{N_1}{N}\right)^2\right)-\left[\frac{N_1N_2}{N}(\bar{x}_1-\bar{x}_2)\right]^2}=$$

$$=\frac{\left[\frac{N_1N_2}{N}(\bar{x}_1-\bar{x}_2)\right]^2}{\left(SS_1+SS_2+\frac{N_1N_2}{N}(\bar{x}_1-\bar{x}_2)^2\right)\frac{N_1N_2}{N}-\left[\frac{N_1N_2}{N}(\bar{x}_1-\bar{x}_2)\right]^2}=$$

$$=\frac{\frac{N_1N_2}{N}(\bar{x}_1-\bar{x}_2)^2}{SS_1+SS_2}=\frac{(\bar{x}_1-\bar{x}_2)^2}{SS_1+SS_2}\cdot\frac{N_1N_2}{N}. \qquad (4)$$

Из формулы (4) следует, что формулы (1) и (2a) математически эквивалентны друг другу. Следовательно, оценивать достоверность разности средних двух выборок можно по любой из них. Все это отмечено также в трехтомнике Кендалла, Стюарта (разделы 26.34, 26.35). Более того, формула (2b) статистической значимости точечно-бисериального коэффициента корреляции выведена исходя из биномиального распределения двоичного признака b. Но заметим, что именно по этой формуле вычисляется *t*-критерий Стьюдента статистической значимости линейного коэффициента корреляции Пирсона между двумя количественными признаками, подчиняющимися двумерному нормальному распределению (Крамер, 1975, разд. 29.7).

Таким образом, и для двух количественных признаков, подчиняющихся двумерному нормальному распределению, и для ситуации, когда один из признаков является количественным и разбит на две части, по отдельности подчиняющиеся нормальному распределению, а второй – двоичным, который подчиняется биномиальному, а не нормальному распределению, профессионалы-статистики утверждают, что применим один и тот же *t*-критерий Стьюдента статистической значимости коэффициента корреляции. Заметим, что в случае действительного различия средних распределение объединенного количественного признака становится бимодальным (двугорбым), т. е. явно не нормальным. А поскольку второй признак двоичный, он тоже никак не может иметь нормальное распределение. Следовательно, *t*-критерий Стьюдента статистической значимости линейного коэффициента корреляции Пирсона между двумя количественными признаками вполне законно применим даже для двух заведомо ненормальных признаков (один двоичный, другой количественный, но бимодальный). Применять, действительно, можно и в теории, и на практике, но почему тогда от нас при расчете корреляций всегда требуют проверять нормальность исходных признаков?

И этот же *t*-критерий математически эквивалентен *t*-критерию Стьюдента статистической значимости разности средних. Стоит ли после этого удивляться, что такой универсальный критерий хорошо работает и во многих других, отклоняющихся от нормальности ситуациях?

Таким образом, нормальность проверять, как правило, незачем. Применяя *t*-критерий Стьюдента даже для явно ненормально распределенных данных, мы практически никогда не завысим достоверность наших результатов. По крайней мере, не завысим сильно.

С равенством дисперсий дело обстоит хуже. Хорошо бы его все-таки проверять, но все критерии проверки равенства дисперсий сами по себе крайне неустойчивы, в частности обычно рекомендуемый *F*-критерий отношения дисперсий Фишера, и требуют большого числа наблюдений (Орлов, 2020). По этой причине многие авторы рекомендуют вместо *t*-критерия Стьюдента для двух средних использовать *t*-критерий Уэлча (Welch, 1938), который тоже выведен из предположения нормальности распределения выборок (и тоже в нем не нуждается по той же причине):

$$t_W = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{SEM_1^2 + SEM_2^2}}, \quad (5)$$

$$s_k^2 = \frac{SS_k}{N_k - 1}, \quad SEM_k^2 = \frac{s_k^2}{N_k} = \frac{SS_i}{N_k(N_k - 1)},$$

где s_k^2 – выборочная дисперсия, SEM (standard error of mean) – стандартная ошибка среднего.

Но *t*-критерий Уэлча удобен тем, что изначально не требует равенства дисперсий. Несколько раньше Уэлча этот же критерий предложил аспирант Фишера Смит (Smith, 1936a; см. Davenport, Webster, 1975). (Возможно, селекционерам интересно, что этот же Смит является автором знаменитого селекционного индекса Смита–Хазеля (Smith, 1936b; Hazel, 1943).)

Некоторое неудобство доставляет расчет степеней свободы для *t*-критерия Уэлча:

$$df = \frac{\left(SEM_1^2 + SEM_2^2\right)^2}{\frac{\left(SEM_1^2\right)^2}{N_1 - 1} + \frac{\left(SEM_2^2\right)^2}{N_2 - 1}}. \quad (6)$$

К счастью, при $N_1 = N_2$ оба критерия совпадают, так что в наиболее важном для практики случае – контроль и опыт, когда размеры выборок, как правило, выбираются равными, – можно спокойно использовать обычный *t*-критерий Стьюдента, не обращая внимания на любую разницу дисперсий. А также, как мы уже знаем, когда распределения исходных выборок симметричны.

Любопытно, что советские биологи, когда им разрешили использовать «буржуазную статистику» (см. предисловие к (Фишер, 1958)), применяли при счете вручную именно *t*-критерий Уэлча (формула (5)), но с $df = N_1 + N_2 - 2$ степенями свободы, искренне считая, что это и есть «настоящий» *t*-критерий Стьюдента. Причина кроется в отечественных учебниках биометрии того времени. В них задача сравнения средних вполне логично сводилась к отношению разности средних и стандартного отклонения этой разности. В случае нормального распределения разности это отношение, конечно, подчиняется *t*-распределению и, следовательно, может служить статистическим *t*-критерием. Упускался маленький математический нюанс: для того чтобы быть *t*-критерием Стьюдента именно с $df = N_1 + N_2 - 2$ степенями свободы, нужно было дополнительно предположить равенство дисперсий двух выборок, как это сделал Фишер, и ввести его в формулу (1). А без этого предположения такое отношение как раз является *t*-критерием Уэлча (5), но тогда число степеней свободы должно определяться по формуле (6). Считать вручную формулу (6) было довольно затруднительно, да ее и не знали. Большой беды от этого не было, всего лишь чуть-чуть завышалось число степеней свободы, а следовательно, и статистическая значимость разности средних. При равенстве объемов выборок все просто совпадало.

Когда появились большие компьютеры (типа БЭСМ-6), ситуация изменилась не сильно. Ввиду полного отсутствия готового программного обеспечения каждый исследователь писал себе программы сам. А это еще надо было уметь. Да и вычислительный ресурс был крайне ограничен. Например, три БЭСМ-6 с объемом ОЗУ каждой менее 800 килобайт(!) обслуживали почти весь Академгородок, включая все институты биологического профиля. И только с переходом в начале 1990-х на персональные компьютеры из Юго-Восточной Азии и западные статистические пакеты возникла возможность более-менее адекватно обрабатывать реальные биологические данные. К сожалению, своих (отечественных) персональных компьютеров и статистических пакетов с того времени и до сих практически так и не появилось, поскольку в этом вроде бы и не было необходимости. Но сейчас, в изменившихся международных условиях, это может стать большой проблемой.

Статпакеты за редким исключением пишут не статистики, а программисты, иногда биологи (например, PAST (Hammer et al., 2001)), добросовестно копируя статистические учебники. А учебники всегда отстают от современного уровня, отраженного в статьях, на одно, а то и два поколения ученых. И конечно, в этих статпакетах запрограммирован более старый *t*-критерий Стьюдента (формула (1)), а критерий Уэлча (формула (5)) либо не назван прямо (а назван, например, *t*-критерием с неравными дисперсиями), либо не запрограммирован вообще. Вместе с пакетами появились и руководства к ним, в которых, конечно, требуются ненужные проверки нормальности и равенства дисперсий. И конечно, во всех пакетах, если у вас вычислен выборочный коэффициент корреляции *r* между двумя признаками вместе с его достоверностью *p*-value, то можете не сомневаться, что он вычислен с использованием формулы (2b):

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}, \quad (2c)$$

хотя еще тот же Фишер более века назад нашел более точное нормализующее *z*-преобразование для коэффициента корреляции (Fisher, 1915):

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right); \quad s_z = \frac{1}{\sqrt{N-3}}, \tag{7}$$

где s_z – стандартное отклонение z, а z распределено нормально. Соответственно, отношение $t = z/s_z$ подчиняется t-распределению с $N-3$ степенями свободы. z-преобразование Фишера тоже есть во всех пакетах, но понятно, что если r и p-value для него уже посчитаны компьютером по формуле (2c), то исследователь-практик вряд ли будет прилагать дополнительные усилия, чтобы вычислить достоверность чуточку точнее. По-хорошему, это должно делаться прямо в пакете при вычислении корреляций.

Что действительно надо проверять, так это наличие выбросов. Статистический анализ всегда проводится в предположении некоторого единства или контролируемой исследователем неоднородности данных, например разбиения на группы по их биологическому смыслу. Но даже одна точка, по каким-то случайным или неслучайным причинам далеко отлетевшая от своей группы, может сильно исказить ситуацию. Эту возможность надо проверять, лучше всего визуально, через построение графиков (например, бокс-плота), и до статистической обработки удалять такие точки на том основании, что они не принадлежат основной совокупности. Причем это необязательно будут ошибки измерения или учета, это могут быть и реальные объекты (например, экзотические животные, сбежавшие от своих хозяев, при изучении диких животных в природе). Можно и даже желательно описать их отдельно. Правда, со временем такие объекты могут стать частью местной фауны, как, например, верблюды в Австралии, когда-то завезенные для перевозки грузов (https://ru.wikipedia.org/wiki/Дикие_верблюды_Австралии).

Еще пример из экспедиционного опыта автора: в зоологических отловах обнаружилась трехногая водяная полевка, самка, судя по плацентарным пятнам в матке, рожавшая. Полностью отсутствовала передняя лапа. Никаких повреждений, даже их следов. Нормальная ровная шкурка на том месте, где должна быть лапа. Но описания нового таксона в систематике не случилось. При вскрытии оказалось, что четвертая лапа все же была, но в виде бледной прозрачной полоски, протянувшейся под шкуркой вдоль тела. Однако трехногая жизнь заметно отразилась на параметрах внутренних органов, особенно их симметрии, и поэтому весь набор ее показателей пришлось исключить из дальнейшего статистического анализа.

Для выявления выбросов можно применять и статистические методы. Для этого тоже существуют аналоги t-критериев, например: «Таблицы 4.8. Критерии исключения резко выделяющихся наблюдений» (Большев, Смирнов, 1983). Современная литература на эту тему весьма обширна (Wang et al., 2019; Boukerche et al., 2020).

При ручном счете, который применялся в докомпьютерную эпоху, если все-таки для двух средних считать именно t-критерий Стьюдента, формула (1) была предпочтительнее формулы (2b) из-за удобства вычислений. Появление персональных компьютеров резко изменило ситуацию с расчетами. Во-первых, они стали доступны всем исследователям и сложность формул перестала быть серьезным препятствием. Соответственно, отпала необходимость пользоваться статистическими таблицами, все необходимые расчеты можно делать в электронных таблицах типа Excel или в специализированных статистических пакетах. Во-вторых, формула (2b) оказалась более удобной при машинном представлении данных в виде таблиц, особенно если измерено несколько признаков у одних и тех же объектов, разбитых на две группы. Тогда один двоичный признак отражает это разбиение сразу для всех признаков и рассчитать статистические значимости разностей сразу для всех признаков можно и удобнее через корреляции (не забывая про необходимость учета множественных сравнений).

Свободные от распределения процедуры

Почему мы всегда исходим из гипотезы нормальности? Почему мы так за нее держимся? В реальности мы, как правило, не знаем не только параметров, но даже формы распределения, которому подчиняется наша исходная выборка. Можем только более или менее уверенно предполагать, что это гипотетическое распределение непрерывно. Можем ли мы подобрать критерий для решения нашей задачи, который не зависит от формы этого распределения? Оказывается, можем, по крайней мере для некоторых задач. Например, для задачи независимости двух количественных признаков. В случае независимости генеральный коэффициент корреляции равен нулю. Поэтому в качестве критерия отклонения от независимости можно взять выборочный коэффициент корреляции r. Можно даже вычислить его распределение, причем не делая никаких предположений относительно формы распределения, которому подчиняется исходная выборка. Профессионалам-статистикам достаточно его непрерывности. И удивительнейшим образом оказывается, что наилучшим распределением выборочного коэффициента корреляции r в общем случае является так называемое перестановочное распределение r, которое с очень хорошим приближением даже для малых N совпадает с t-распределением Стьюдента с $N-2$ степенями свободы для коэффициента корреляции r в «нормальной теории». Первые два момента распределений совпадают полностью (Кендалл, Стюарт, 1973, п. 31.19).

«Но если фактическое совпадение перестановочного распределения с распределением нормальной теории не является полной неожиданностью, то оно во всяком случае очень удобно и приятно, так как мы можем по-прежнему пользоваться таблицами нормальной теории (в данном случае t-распределения Стьюдента) для свободного от распределения критерия непараметрической гипотезы независимости» (Кендалл, Стюарт, 1973, п. 31.20).

Теперь становится понятным, почему и в случае двумерного нормального распределения (формула (3b)), и в случае пары бинарного и двумодального распределений (формула (3a)) t-распределение для выборочного коэффициента корреляции r (и одновременно для t-критерия разности двух средних) вычисляется по одной и той же формуле. А оно и должно всегда по ней вычисляться, если мы не знаем настоящего генерального распределения. Так говорит теория. Разница только в том, что если мы исходим из гипотезы нормального распределения, то мы должны ее

проверять и только в случае удачи использовать *t*-критерий Стьюдента, а иначе уходить на ранги и критерии типа Манна–Уитни. Если же мы исходим из гипотезы неизвестного непрерывного распределения, то ничего проверять не должны, а должны использовать тот же самый *t*-критерий Стьюдента, причем в любой удобной для нас форме. Исследователь-практик вправе сам выбирать, из какой гипотезы ему исходить.

Бутстреп для расчета статистической значимости *t*-критерия Стьюдента разности средних двух выборок

А можно ли проверить теорию? То есть можно ли что-то сказать относительно статистической значимости, например, различия средних, имея только сами данные и совсем не выдвигая никаких теоретических гипотез насчет того, как именно они должны себя вести и какому распределению обязаны подчиняться? У нас же теперь компьютеры есть!

Американский статистик Б. Эфрон нашел такой путь и назвал его бутстрепом (Efron, 1979). Он предложил размножать исходную выборку прямо в компьютере. Пусть она состоит из N элементов. Новую выборку получим следующим образом. С помощью датчика случайных чисел выберем с равными вероятностями любой элемент исходной выборки и включим его копию в новую таблицу. Повторим процесс N раз. Бутстреп-копия исходной выборки сформирована.

Повторим весь процесс много раз (K) и получим K бутстреп-копий исходной выборки. Поскольку все элементы в них равновероятно выбраны из исходной выборки, они подчиняются ее эмпирическому распределению. А с ростом K оно сходится к генеральному. То, что оно нам неизвестно, не имеет значения. Главное – что оно одно и то же для исходной выборки и для всех ее бутстреп-копий. По сути, мы получили некую искусственную модель генерального распределения, которому заведомо подчиняется наша выборка, поскольку именно ее эмпирическое распределение и взято за основу (Rochowicz, 2010; Hesterberg, 2015; Rousselet et al., 2023). Поскольку это все же «не настоящее» генеральное распределение, то у него есть и свои преимущества, и неизбежные недостатки. Начнем с преимуществ: 1. Проверка предположений о нормальности и равенстве дисперсий не требуется. И вообще никаких предположений о распределениях выборок не требуется. 2. Выводы делаются на более солидном, хотя и насчитанном искусственно материале. Недостатки: 1. Необходимы более мощные компьютеры, специализированное программное обеспечение и много машинного времени. 2. Не для всех задач его можно применять. Например, бутстреп в принципе не позволяет оценить среднее «настоящего» генерального распределения, а следовательно, и его доверительные интервалы. Чем больше бутстреп-копий выборки мы сделаем, тем ближе среднее их средних приблизится к среднему самой выборки, а не к среднему генерального распределения.

В случае средних двух выборок бутстреп можно применять по-разному, хотя нулевая гипотеза всегда одинакова: средние обеих выборок равны. Прямой способ изложен в (Hesterberg, 2015). Вычисляем для выборок стандартный *t*-критерий разности средних, для которого нам и нужно получить *p*-value. Центрируем обе выборки, каждую своим средним. Теперь для них нулевая гипотеза заведомо выполняется, средние значения каждой выборки равны нулю. Остальные важные параметры каждой выборки (объем, дисперсия, асимметрия, эксцесс) при центрировании не меняются. Делаем для каждой центрированной выборки бутстреп-копию. Естественно, все параметры обеих выборок по случайным причинам отклоняются от исходных. Вычисляем *t*-критерий разности между ними. Повторяем много раз (10^3, 10^4, 10^5 – чем больше, тем лучше). Получаем бутстреп-распределение *t*-критерия при гарантированном выполнении нулевой гипотезы. Заметим, что это бутстреп-распределение совсем не обязано быть табличным *t*-распределением Стьюдента. Оно может быть сильно искажено особенностями обеих выборок – неравенством дисперсий, объемов, асимметрией и т. д. Но именно это нам и надо – учесть особенности самих выборок. А для этого достаточно посмотреть, какова доля насчитанных *t*-критериев в бутстреп-распределении, которые превысили (по модулю) *t*-критерий для исходных нецентрированных выборок. Это и будет бутстреп-оценкой *p*-value. Никаких дополнительных предположений и проверок не потребовалось.

Список литературы / References

Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1983
[Bolshev L.N., Smirnov N.V. Tables of Mathematical Statistics. Moscow: Nauka Publ., 1983 (in Russian)]

Кендалл М., Стюарт А. Статистические выводы и связи. Т. 2. М.: Наука, 1973
[Kendall M., Stewart A. The Advanced Theory of Statistics. Vol. 2. Moscow: Nauka Publ., 1973 (in Russian)]

Крамер Г. Математические методы статистики. М.: Мир, 1975
[Cramer H. Mathematical Methods of Statistics. Moscow: Mir Publ., 1975 (in Russian)]

Лемешко Б.Ю., Лемешко С.Б. Об устойчивости и мощности критериев проверки однородности средних. *Измерительная техника*. 2008;9:23-28
[Lemeshko B.Y., Lemeshko S.B. Power and robustness of criteria used to verify the homogeneity of means. *Meas Tech*. 2008;51(9):950-959. doi 10.1007/s11018-008-9157-3]

Лойко В.И., Луценко Е.В., Орлов А.И. Высокие статистические технологии и системно-когнитивное моделирование в экологии. Краснодар: КубГАУ, 2019
[Loiko V.I., Lutsenko E.V., Orlov A.I. High Statistical Technologies and System-cognitive Modeling in Ecology. Krasnodar: KubSAU Publ., 2019 (in Russian)]

Орлов А.И. О методах проверки однородности двух независимых выборок. *Заводская лаборатория. Диагностика материалов*. 2020;86(3):67-76. doi 10.26896/1028-6861-2020-86-3-67-76
[Orlov A.I. On methods of testing the homogeneity of two independent samples. *Zavodskaya Laboratoriya. Diagnostika Materialov* = *Industrial Laboratory. Diagnostics of Materials*. 2020;86(3):67-76. doi 10.26896/1028-6861-2020-86-3-67-76 (in Russian)]

Фишер Р.А. Статистические методы для исследователей. М.: Госстатиздат, 1958
[Fisher R.A. Statistical Methods for Research Workers. Moscow: Gosstatizdat Publ.,1958 (in Russian)]

Benjamin D.J., Berger J.O., Johannesson M., Nosek B., Wagenmakers E.J., Berk R., Bollen K.A., … Wolpert R., Xie Y., Young C., Zinman J., Johnson V.E. Redefine statistical significance. *Nat Hum Behav*. 2018;2(1):6-10. doi 10.1038/s41562-017-0189-z

Boukerche A., Zheng L., Alfandi O. Outlier detection: methods, models, and classification. *ACM Computing Surveys* (*CSUR*). 2020;53(3):1-37. doi 10.1145/3381028

Conover W.J. The rank transformation – an easy and intuitive way to connect many nonparametric methods to their parametric counter-

parts for seamless teaching introductory statistics courses. *Wiley Interdiscip Rev Comput Stat*. 2012;4(5):432-438. doi 10.1002/wics.1216

Davenport J.M., Webster J.T. The Behrens–Fisher problem, an old solution revisited. *Metrika*. 1975;22(1):47-54

Efron B. Computers and the theory of statistics: thinking the unthinkable. *SIAM Rev*. 1979;21(4):460-480. doi 10.1137/1021092

Fisher R.A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*. 1915;10(4):507-521

Fisher R.A. Applications of "Student's" distribution. *Metron*. 1925a;5: 90-104

Fisher R.A. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd, 1925b

Hammer Ø. Harper D.A.T., Ryan P.D. PAST: PAleontological STatistics software package for education and data analysis. *Palaeontol Electronica*. 2001;4(1):9

Hazel L.N. The genetic basis for constructing selection indexes. *Genetics*. 1943;28(6):476-490

Hesterberg T.C. What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. *Am Stat*. 2015; 69(4):371-386. doi 10.1080/00031305.2015.1089789

Kennedy-Shaffer L. Before $p < 0.05$ to beyond $p < 0.05$: using history to contextualize *p*-values and significance testing. *Am Stat*. 2019; 73(sup1):82-90. doi 10.1080/00031305.2018.1537891

Lehmann E.L. "Student" and small-sample theory. *Stat Sci.* 1999;14(4): 418-426

Lumley T., Diehr P., Emerson S., Chen L. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health*. 2002;23(1):151-169. doi 10.1146/annurev.publhealth.23.100901.140546

Rochowicz J.A., Jr. Bootstrapping analysis, inferential statistics and EXCEL. *Spreadsheets in Education* (*SiE*). 2010;4(3). http://epublications. bond.edu.au/ejsie/vol4/iss3/4

Rousselet G., Pernet C.R., Wilcox R.R. An introduction to the bootstrap: a versatile method to make inferences by using data-driven simulations. *Meta-Psychology.* 2023;7:2058. doi 10.15626/MP.2019.2058

Smith H.F. The problem of comparing the results of two experiments with unequal errors. *J Council Sci Industrial Res.* 1936a;9:211-212

Smith H.F. A discriminant function for plant selection. *Annals of Eugenics*. 1936b;7(3):240-250.

Student. The probable error of a mean. *Biometrika*. 1908;6:1-25. doi 10.2307/2331554

Student. Tables for estimating the probability that the mean of a unique sample of observations lies between $-\infty$ and any given distance of the mean of the population from which the sample is drawn. *Biometrika*. 1917;XV:414-417

Student. Statistics in biological research. *Nature*. 1929;124:93. doi 10.1038/124093b0

Student's *t*-test (*t*-критерий Стьюдента) https://en.wikipedia.org/wiki/Student%27s_t-test [обновлено 10 января 2023; процитировано 30 июня 2024. доступно: https://руни.рф]

Wang H., Bah M.J., Hammad M. Progress in outlier detection techniques: a survey. *IEEE Access*. 2019;7;107964-108000. doi 10.1109/ACCESS.2019.2932769

Wasserstein R.L., Schirm A.L., Lazar N.A. Moving to a world beyond "*p*<0.05". *The American Statistician*. 2019;73(sup.1):1-19. https://www.jstor.org/stable/48783683

Welch B.L. The significance of the difference between two means when the population variances are unequal. *Biometrika.* 1938; 29(3/4):350-362

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию 01.09.2024. После доработки 13.11.2024. Принята к публикации 09.12.2024.