

АКАДЕМИЧЕСКОЕ ПИСЬМО / ACADEMIC WRITING

Дискуссионная статья / Discussion paper

<https://doi.org/10.24069/SEP-24-01>

Диалоги о статистике в научных публикациях

Н. Н. Хромов-Борисов  

Комиссия по борьбе с лженаукой Экспертного совета РАН,
г. Санкт-Петербург, Российская Федерация

 Nikita.KhromovBorisov@gmail.com

Резюме. Проведен критический разбор статьи С.В. Мыльникова (*Научный редактор и издатель*. 2023;8(1 Suppl.):S52–S61) в форме заочного диалога. Обсуждены такие аспекты, как частотнические и байесовские точечные и интервальные оценки параметров биномиального и полиномиального распределений; коэффициента линейной корреляции (r); коэффициента детерминации (R^2); параметров линейной регрессии; разности (PD) и отношения долей (RR); отношения шансов (OR); оценки минимально необходимых объемов выборок; преимущества статистического оценивания на основе бутстрепа; генератор случайных чисел; графические методы статистического анализа: коробки с декольте и с усами, вероятностные графики, гистограммы, графики Гарднера-Олтмена, дождевые облака; поправки на множественность сравнений; сложности корреляционного анализа; размер эффекта; вероятности воспроизведения (p_{rep} , p_{srep} и p_{rerep}); значения $mid-p$; предсказательные интервалы для r -значений, для размеров эффекта и для суммарных эффектов в метаанализе. Представлено описание программ, рекомендуемых для решения обсуждаемых статистических задач.

Ключевые слова: статистика, доверительные интервалы, предсказательные интервалы, программное обеспечение статистического анализа, бутстреп, графические методы

Благодарности. Автор благодарит Сергея Владимировича Мыльникова, Руслана Темирсултановича Сайгитова, рецензентов, редакторов и корректоров за ценные советы и замечания.

Для цитирования: Хромов-Борисов Н.Н. Диалоги о статистике в научных публикациях. *Научный редактор и издатель*. 2024;9(1 Suppl. 1):1S5–1S32. <https://doi.org/10.24069/SEP-24-01>

Dialogues on statistics in scientific publications

N.N. Khromov-Borisov  

Commission for Combating Pseudoscience of the Expert Council of the Russian Academy of Sciences,
St. Petersburg, Russian Federation

 Nikita.KhromovBorisov@gmail.com

Abstract. A critical analysis of the article by S.V. Mylnikov (*Scientific Editor and Publisher*. 2023;8(1 Suppl.):S52–S61) in the form of a correspondence dialogue is conducted. Aspects such as frequentist and Bayesian point and interval estimation of the parameters of binomial and polynomial distributions, linear correlation coefficient (r), coefficient of determination (R^2), linear regression parameters, difference (PD) and proportion ratio (RR), odds ratio (OR), estimation of minimum required

sample sizes, advantages of statistical estimation based on bootstrap, random number generator, graphical methods of statistical analysis: *décolleté* and whisker boxes, probability plots, histograms, Gardner-Altman plots, rain clouds, corrections for multiple comparisons, complexities of correlation analysis, effect size, replication probabilities (p_{rep} , p_{srep} and p_{rerep}), mid- p values, prediction intervals for p values, for effect sizes, and for combined effects in meta-analysis. A description of the programs recommended for solving the discussed statistical problems is presented.

Keywords: statistics, confidence intervals, predictive intervals, statistical analysis software, bootstrap, graphical methods

Acknowledgments. The author thanks, Sergei V. Mylnikov, Ruslan T. Saigitov, reviewers, editors, and correctors for valuable advice and comments.

For citation: Khromov-Borisov N.N. Dialogues on statistics in scientific publications. *Science Editor and Publisher*. 2024;9(1 Suppl. 1):1S5–1S32. <https://doi.org/10.24069/SEP-24-01>

Статистика (точнее, статистический анализ данных) давно стала неотъемлемой и обязательной частью научных исследований и публикаций и, по сути, является самостоятельным научным исследованием. Это живая развивающаяся наука, и в последние десятилетия ее идеология и методология претерпевают существенные изменения. Сейчас говорят о ее реформе, о возникновении «новой статистики» [1; 2]. Недавно журнал «Научный редактор и издатель» опубликовал статью С.В. Мыльникова, содержащую рекомендации по описанию статистики в медико-биологических журналах [3]. Статья содержит много спорных положений, поэтому здесь представлен их критический разбор в формате заочного диалога С.В. Мыльникова (СМ) и Н.Н. Хромова-Борисова (НХ).

НХ. Автор (СМ) начинает с цитирования рекомендаций Международного комитета редакторов биомедицинских журналов (ICMJ): «По возможности, подвергайте полученные данные количественной оценке и представляйте их с соответствующими показателями ошибок измерения и неопределенности или варьирования измерений (такими как доверительные интервалы)». Этими рекомендациями руководствуются тысячи зарубежных и многие отечественные журналы. Наверное, стоило упомянуть, что они переведены на русский язык¹.

Наиболее четко требование вычислять доверительные интервалы (ДИ) сформулировано в «Хартии качественной редакционной практики», которую разработал Совет кардиологических

журналов (СКЖ): «Представляются доверительные интервалы для всех основных результатов, независимо от того, были ли они статистически значимыми или нет»².

СМ. Построение интервальной оценки (доверительного интервала) параметра распределения признака в генеральной совокупности/популяции (Universe/Total population) похоже на бросание колечка на невидимый штырек. При грамотном формировании репрезентативной выборки большинство колечек будет попадать на штырек, но часть будет падать в стороне. Такая ситуация называется смещенной оценкой (bias). Мы не можем узнать, смещена наша оценка или нет, но можем узнать диаметр колечка, построив доверительный интервал.

НХ. Аналогия с набрасыванием кольца (обруча) на штырь, да еще почему-то на «невидимый», довольно сомнительна. И непонятно, при чем тут смещение? Смещение – это отклонение математического ожидания оценщика (статистики, использованной для оценки параметра) от его истинного значения. Аналогия состоит скорее в том, что попадание зависит от ширины кольца. Чем шире кольцо, тем больше вероятность набросить кольцо на штырь.

СМ. Биномиальное распределение

Пример. В захоронении индейцев «доколумбовой» Америки найдены останки 15 человек. Молекулярно-генетическими методами определены их группы крови (см. табл. 1).

¹ Рекомендации по проведению, описанию, редактированию и публикации результатов научной работы в медицинских журналах. Режим доступа: <https://www.icmje.org/recommendations/translations/russian2020.pdf> (дата обращения: 27.01.2024).

² Хартия качественной редакционной практики. Режим доступа: https://scardio.ru/content/education/courses/docs/Charter_Buklet_150x150_v3.pdf (дата обращения: 27.01.2024).

Таблица 1. Точечная и интервальная оценка частот групп крови в доколумбовой Северной Америке [4]
Table 1. Point and interval estimation of blood group frequencies in pre-Columbian North America [4]

Группа крови системы ABO	Число останков	Оценки биномиальных частот					Оценки полиномиальных частот
		частотнические			бейзовские		
		Точечные	Интервальные	Правильные границы 95%-х ДИ по Клопперу-Пирсону	Точечные (медианы)	95%-е ДИ	Границы точных 95%-х ДИ
I (O)	14	0.93±0.066	0.660–0.997	0,681; 0,998	0,897	0,698; 0,984	0,618; 0,999
II (A)	1	0.07±0.066	0.040–0.340	0,002; 0,319	0,103	0,016; 0,302	0,008; 0,382
III (B)	0	0	0–0.254	0,000; 0,218;	0,042	0,002; 0,206	0,000; 0,264
IV (AB)	0	0	0–0.254	0,000; 0,218;	0,042	0,002; 0,206	0,000; 0,264

НХ. Первые три столбца приведены как у СМ.

Замечание: в отечественной литературе пока что все еще принято использовать запятую (а не точку) в качестве десятичного разделителя.

Точечной оценкой является одно значение (точка): 0,93, а не 0.93 ± 0.066. Что указано после оператора ±? В любом случае, такая запись допустима (имеет смысл) только для симметричных распределений. Биномиальное распределение симметрично только при значении его параметра $\varphi = 0,5$.

Если 0,066 – стандартное отклонение, то оно вычислено неверно, должно быть 0,0644. В таком случае принята запись: 0,933 (0,0644). Недопустимо также использовать знак «минус» (или дефис) между границами ДИ. Следует использовать точку с запятой, см. пятый, седьмой и восьмой столбцы в табл. 1.

Как вычислено?

СМ. Для расчета использован точный метод построения, известный также как метод Клоппера-Пирсона (Clopper-Pearson exact method) [5].

НХ. Ссылаться на исходную статью Клоппера и Пирсона некорректно, потому что в ней не приведены указанные формулы в явном виде. Во избежание путаницы здесь и далее авторское обозначение P для доли заменено на φ . Представленные в статье формулы для границ ДИ, по-видимому, заимствованы из учебного пособия [6], которое СМ упоминает в списке литературы.

$$\varphi_{\text{up}} = \frac{v_1 F_{\alpha/2}}{v_2 + v_1 F_{\alpha/2}},$$

где α – выбранный уровень значимости, $v_1 = 2(k+1)$ и $v_2 = 2(n-k)$
и

$$\varphi_{\text{low}} = \frac{v_1}{v_2 + v_1 F_{\alpha/2}},$$

где $v_1 = 2(n-k+1)$ и $v_2 = 2k$.

Однако во второй формуле для нижней границы ДИ для доли по Клопперу-Пирсону вкралась непозволительная опечатка: в числителе дроби должно быть v_2 , а не v_1 :

$$\varphi_{\text{low}} = \frac{v_2}{v_2 + v_1 F_{\alpha/2}},$$

Значения правильно вычисленных ДИ по методу Клоппера-Пирсона представлены в пятом столбце табл. 1.

СМ. Для расчета значения F можно использовать MS Excel. Например, для значений $\alpha = 0,05$, $v_1 = 30$, $v_2 = 6$ в ячейку нужно ввести выражение:

$$= F.ОБР.ПХ(0,025; 30; 6) = 5,07.$$

НХ. Предлагать пользователям вычислять вручную промежуточные значения F явно нерационально, и вряд ли кто-либо воспользуется такой рекомендацией. Но уж если такая потребность возникнет, проще использовать соответствующие калькуляторы (см. рис. 1).

Для пользователей, имеющих опыт работы в среде R , программа приводит простой скрипт. Вычисления сопровождаются соответствующим графиком плотности F -распределения. Программа обеспечивает также аналогичные высокоточные вычисления для трех других основных распределений: нормального, Стьюдента и хи-квадрат. Существует сайт для сверхточных вычислений для 26 распределений, но, к сожалению, его английская версия была закрыта 20 сентября 2023 года, осталась лишь версия на японском языке, но с помощью Google-переводчика им можно пользоваться. Для вычисления статистических функций это страница по адресу <https://keisan.casio.jp/menu/system/000000000850>, для F -распределения – страница <https://keisan.casio.jp/exec/system/1161228871>.

P-Value Calculator for F Distribution

Degree of freedom 1 (numerator) :

Degree of freedom 2 (denominator) :

F-value:

p-value:

p-value type:
☒ right tail
☐ left tail

Right-tail p-value is 0.025, df1 = 30, df2 = 6

F-value is 5.065

R command: `qf(0.025, 30, 6, lower.tail=FALSE)` or `qf(1-0.025, 30, 6)`

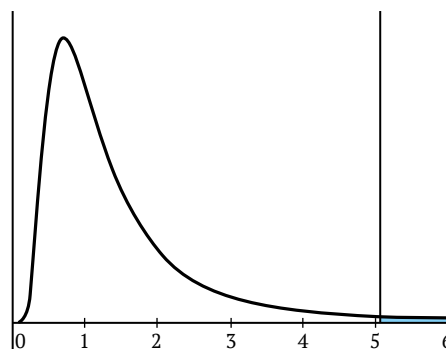


Рис. 1. Вычисление значения F для $\alpha/2 = 0.025$ при $df_1 = 30$ и $df_2 = 6$ на калькуляторе <http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pf.html>

Fig. 1. Calculation of F value for $\alpha/2 = 0.025$ with $df_1 = 30$ and $df_2 = 6$ in <http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pf.html>

Биномиальные доли (частоты) как случайные переменные имеют бета-распределение, которое однозначно связано с F -распределением, и в терминах бета-распределений формулы для границ ДИ для долей по Клопперу-Пирсону выглядят более компактно:

$$\varphi_{\text{low}} = B\left(\frac{\alpha}{2}; k, n - k + 1\right)$$

и

$$\varphi_{\text{up}} = B\left(1 - \frac{\alpha}{2}; k + 1, n - k\right).$$

В Excel их можно вычислять напрямую, без промежуточных вычислений. Например, для данных в первой строке табл. 1:

$$P_{\text{low}} = \text{БЕТА.ОБР}(0.025; 14; 2) = 0,681$$

и

$$P_{\text{up}} = \text{БЕТА.ОБР}(0.975; 15; 1) = 0,998.$$

Интервал Клоппера-Пирсона – это один из первых и широко распространенных методов расчета биномиальных доверительных интервалов. Его часто называют «точным» методом. Однако термин «точный» неудачен и вводит в заблуждение: он отражает лишь тот факт, что для нахождения интервальных оценок используют истинное распределение, а не какое-то его приближение, он не означает «идеально точный». О методе Клоппера-Пирсона известно, что он, как и другие «точные» методы, является излишне консервативным.

Библиография публикаций об интервальных оценках параметра биномиального распределения (доли, частоты) насчитывает более 150 (по сведениям на 2010 год) [7].

Следует также отметить, что, строго говоря, ДИ для долей (частот) по Клопперу-Пирсону являются фидуциальными, что отражено в заголовке их статьи [5].

Итак, биномиальная доля (как случайная переменная) имеет бета-распределение. Поэтому его квантильную функцию можно использовать для построения ДИ для долей. По Клопперу-Пирсону это будет:

$$B\left(\frac{\alpha}{2}; k, n - k + 1\right) < \tilde{\varphi} < B\left(1 - \frac{\alpha}{2}; k + 1, n - k\right).$$

Оптимальным вариантом, однако, являются байесовские точечные оценки долей и ДИ для них.

Бейзовская³ точечная оценка неизвестной вероятности (доли)

Руководствуясь принципом недостаточного основания Лапласа (принцип безразличия, индифферентности)⁴, резонно предположить, что неизвестная нам оцениваемая доля φ_{unkn} *a priori* имеет равномерное распределение на отрезке $[0; 1]$, или (что то же самое) бета-распределение с параметрами: $a_0 = 1$ и $b_0 = 1$ (см., например: Wikipedia⁵).

³ В русском языке принята фонетическая транскрипция иностранных имен и фамилий («как слышится, так и пишется»). Фамилия Bayes произносится как Бейз (/beiz/). Прослушать можно здесь: https://upload.wikimedia.org/wikipedia/commons/0/02/De-Thomas_Bayes.ogg. В отечественной научной литературе нет единообразия в написании этой фамилии. В нематематических текстах это Байес, в математических энциклопедиях – Бейес. Очевидно, следует руководствоваться рекомендациям лингвистов.

⁴ Principle of indifference. Available at: https://en.wikipedia.org/wiki/Principle_of_indifference (accessed: 27.01.2024).

⁵ Binomial distribution. Available at: https://en.wikipedia.org/wiki/Binomial_distribution (accessed: 27.01.2024).

Тогда, согласно правилу следования Лапласа⁶, апостериорной точечной байзовской оценкой доли является:

$$\varphi = \frac{k+1}{n+2},$$

здесь k – число «успехов», и n – число «испытаний» (объем выборки) в классической схеме Бернулли.

Бейзовские интервальные оценки долей

Соответствующей апостериорной байзовской оценкой нижней границы $100(1 - \alpha)\%$ -го ДИ для доли является $\alpha/2$ -квантиль бета-распределения, т.е. функция, обратная к функции его распределения с параметрами $a = k + 1$ и $b = n - k + 1$:

$$l_{\alpha/2} = \text{BetaInv}\left[\frac{\alpha}{2}, k+1, n-1+1\right],$$

а апостериорной байзовской оценкой верхней границы ее $100(1 - \alpha)\%$ -го ДИ является $(1 - \alpha/2)$ -квантиль бета-распределения с теми же параметрами:

$$u_{(1-\alpha/2)} = \text{BetaInv}\left[1 - \frac{\alpha}{2}, k+1, n-1+1\right].$$

Конечно, нет необходимости вычислять квантили бета-распределения. Получаемые байзовские точечные и интервальные оценки представлены в предпоследнем столбце табл. 1.

Для вычислений обоих вариантов ДИ – фи-дуциальных по Клопперу-Пирсону и байзов-

ских – удобно использовать программу LePAC⁷ (см. рис. 2).

В этой же программе можно ознакомиться с результатами 32 методов вычисления ДИ для долей (они скрыты в желтом поле с ромбом) (см. рис. 3).

Можно видеть, что некоторые приближенные методы дают абсурдные результаты: либо верхние границы ДИ превышают значения 1, либо нижние границы оказываются отрицательными. Очевидно, что таких методов следует избегать.

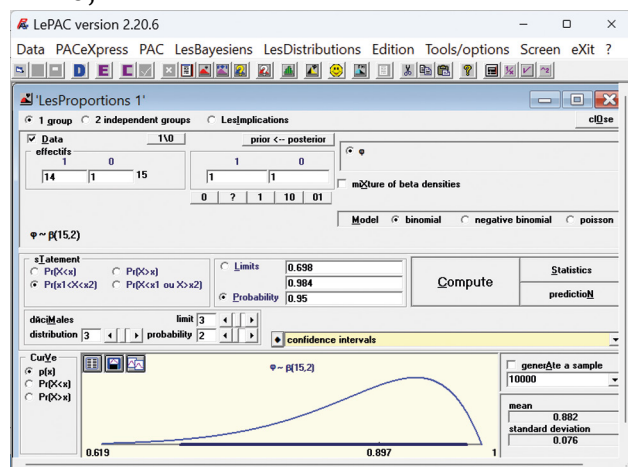
СМ. Таким образом, исходя из точечной оценки, равной нулю, можно заключить, что частота группы крови АВ в «доколумбовой» Америке не превышала 25%.

НХ. Корректнее было бы интерпретировать результат примерно такими словами: «Итак, в выборке объемом $n = 15$ авторы исследования не нашли останков с группами крови В и АВ, но это не означает, что их частоты в изучаемой популяции равны нулю. С вероятностью 95 % при повторении наблюдений эти частоты могут оказаться в интервале от 0 до 26 %. Если удастся найти еще 15 таких останков, то с вероятностью 95 % от 0 до 4 из них могут оказаться с группами крови В и/или АВ».

⁶ Rule of succession. Available at: https://en.wikipedia.org/wiki/Rule_of_succession (accessed: 27.01.2024).

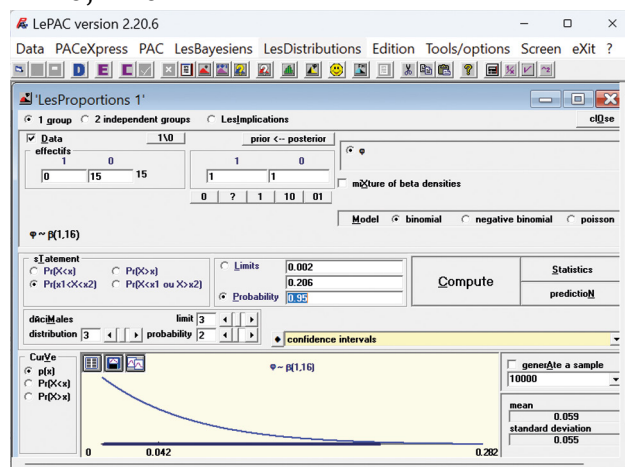
⁷ Télécharger LePAC [Windows]. Available at: <https://eris62.eu/ErisLePAC.html> (accessed: 27.01.2024).

$n = 15, k = 14$



$\varphi_{Me} = 0,90$; 95%ДИ = 0,70; 0,98

$n = 15, k = 0$



$\varphi_{Me} = 0,04$; 95%ДИ = 0,002; 0,21

Рис. 2. Вычисление 95%-х ДИ для долей, байзовский подход
Fig. 2. Calculation of 95% CI for proportions, Bayesian approach

Полиномиальное распределение

Однако все эти рассуждения и вычисления справедливы исключительно для **биномиального** распределения. В данном случае с четырьмя группами крови вероятностной моделью для их численностей и частот является **полиномиальное** (тетраномиальное) распределение. Оценивать эти частоты как параметры четырех **независимых** биномиальных распределений недопустимо.

Вычисление точных ДИ для них реализовано в известном коммерческом пакете StatXact⁸ (см. рис. 4).

Для сравнения с биномиальными оценками результаты представлены в последнем столбце табл. 1.

⁸ StatXact® 12. Available at: <https://www.cytel.com/software/statxact/> (accessed: 27.01.2024).

confidence intervals	
two-sided 0.95	
[0.728,0.993]	Bayes prior Beta[1/2,1/2] [Jeffreys]
[0.728,1]	Bayes prior Beta[1/2,1/2] modified [Brown-Cai-Das Gupta]
[0.728,0.990]	Bayes prior de Cristofaro-Lecoutre
[0.712,0.989]	Bayes prior 'data dependent' [Bolboaca]
[0.698,0.984]	Bayes prior Beta[1,1] uniform
[0.681,0.983]	Bayes prior Beta[0,1]
[0.782,0.998]	Bayes prior Beta[1,0]
[0.681,0.998]	Binomial including 'exact' [Clopper-Pearson]
[0.782,0.983]	Binomial excluding
[0.713,0.997]	Binomial 'mid-p'
[0.758,1.000]	Arc sine
[0.704,0.995]	Arc sine corrected
[0.728,0.997]	Arc sine modified [Anscombe]
[0.680,1.000]	Arc sine modified [Anscombe] corrected
[0.698,0.997]	Blaker (exact)
[0.698,0.997]	Blyth-Still-Casella ('exact')
[0.648,0.991]	Logit
[0.678,0.978]	Logit modified [Anscombe]
[0.702,0.988]	score [Wilson]
[0.702,0.997]	score [Wilson] modified [Brown, Cai, DasGupta]
[0.660,0.997]	score [Wilson] corrected
[0.726,1.001]	Second order corrected
[0.807,1.060]	Wald
[0.774,1.093]	Wald corrected
[0.719,0.971]	Wald/score interval estimate -> center
[0.685,1.005]	Wald/score interval estimate -> center corrected
[0.682,1.008]	Wald/score interval estimate -> center + variance
[0.678,1.006]	Wald/Add 2+2' [Agresti-Coull]
[0.708,1.060]	Wald/SAIFS [Borkowf]
[0.692,1.072]	Wald/SAIFS 'Student t' [Borkowf]
[0.721,1.146]	Wald/'Student t' [Pan]
[0.659,1.025]	Wald/Add 2+2/Student t' [Pan]

confidence intervals	
two-sided 0.95	
[3.2194E-05,0.152]	Bayes prior Beta[1/2,1/2] [Jeffreys]
[0,0.218]	Bayes prior Beta[1/2,1/2] modified [Brown-Cai-Das Gupta]
[8.9100E-04,0.164]	Bayes prior de Cristofaro-Lecoutre
[0.002,0.206]	Bayes prior 'data dependent' [Bolboaca]
[0.002,0.206]	Bayes prior Beta[1,1] uniform
[0.002,0.206]	Bayes prior Beta[0,1] (corrected for 0)
[0.002,0.218]	Bayes prior Beta[1,0] (corrected for 0)
[0,0.218]	Binomial including 'exact' [Clopper-Pearson]
[0,0]	Binomial excluding
[0,0.181]	Binomial 'mid-p' (corrected for 0)
[0,0.063]	Arc sine
[0,0.179]	Arc sine corrected
[0,0.154]	Arc sine modified [Anscombe]
[0,0.219]	Arc sine modified [Anscombe] corrected
[0,0.215]	Blaker (exact)
[0,0.213]	Blyth-Still-Casella ('exact')
[0,0]	Logit
[0.004,0.206]	Logit modified [Anscombe]
[0,0.204]	score [Wilson]
[0,0.204]	score [Wilson] modified [Brown, Cai, DasGupta]
[0.006,0.253]	score [Wilson] corrected
[-0.003,0.164]	Second order corrected
[0,0]	Wald
[-0.033,0.033]	Wald corrected
[0.102,0.102]	Wald/score interval estimate -> center
[0.069,0.135]	Wald/score interval estimate -> center corrected
[-0.035,0.239]	Wald/score interval estimate -> center + variance
[-0.033,0.243]	Wald/Add 2+2' [Agresti-Coull]
[0,0.185]	Wald/SAIFS [Borkowf]
[0,0.197]	Wald/SAIFS 'Student t' [Borkowf]
[0,0]	Wald/'Student t' [Pan]
[-0.065,0.276]	Wald/Add 2+2/Student t' [Pan]

Рис. 3. Результаты вычисления 95%-х ДИ для долей 32 методами

Fig. 3. Results of calculating 95% CI for proportions using 32 methods

Estimation of Multinomial Probability Parameters (PI1, PI2... PIk)

mltmnl (category = var1, frequency = var2, method = exact, time_limit = none);

Data File:

Column Variable: Var1
Frequency Variable: Var2
Number of Categories: 4
Number of Trials: 15

Results:

Category	Count	Prob (PI)	Asymptotic		Exact	
			Lower	Upper	Lower	Upper
1	14	0.9333	0.6204	0.9917	0.6179	0.9992
2	1	0.06667	0.008268	0.3796	0.008382	0.3821
3	0	0	0	0.2937	0	0.2642
4	0	0	0	0.2937	0	0.2642

Рис. 4. Вычисление асимптотических и точных 95%-х ДИ для полиномиальных долей

Fig. 4. Calculation of asymptotic and exact 95% CIs for polynomial fractions

СМ. Построение доверительного интервала для доли также позволяет оценить необходимый объем выборки и потенциальные затраты. Так, из табл. 2 видно, что увеличение выборки до объема более 500 является излишним процессом, поскольку величина доверительного интервала уже практически не меняется.

Таблица 2. Точечные и интервальные оценки «частоты успеха» для разных объемов выборок [8]

Table 2. Point and interval estimates of the “success probability” for different sample sizes [8]

Число успехов	Число попыток	Точечная оценка	Нижняя граница 95 % ДИ	Верхняя граница 95 % ДИ	Ширина 95% ДИ (%)
5	10	50±15,8	19	81	62
10	20	50±11,2	28	72	44
25	50	50±7,1	36	64	28
50	100	50±5,0	40	60	20
100	200	50±3,5	43	57	14
200	400	50±2,5	45	55	10
300	600	50±2,0	46	54	8
2	20	10±6,7	0 1,3	23 32	31
4	40	10±4,7	1 2,8	19 24	21
20	200	10±2,1	6 6,2	14 15	9

НХ. Здесь правильный десятичный разделитель – запятая.

Во-первых, автору (СМ) следовало указать, что доли в этой таблице выражены в процентах. И это диссонирует с данными в табл. 1, где доли выражены в долях от единицы. Во-вторых, как и в предыдущем случае, использовать оператор ± в представленных данных типа 50±15,8 нерационально и непродуктивно. В-третьих, границы ДИ в трех последних строках вычислены неправильно, для вычислений автор (СМ) использовал асимптотические формулы, малоприменимые для малых долей [8]. В табл. 2 они зачеркнуты, и правильные («точные») значения выделены жирным шрифтом.

Выражение «величина доверительного интервала» – неудачное, у любого интервала две величины – его границы. Утверждение автора (СМ), что «увеличение выборки до объема более 500 является излишним процессом, поскольку величина доверительного интервала уже практически не меняется», спорно. Полезным показателем точности интервальных оценок является давно известная, но подзабытая величина – ширина интервала, т.е. разность между его верхней и нижней границами. В данной таблице значения

ширины 95%-го ДИ представлены в последнем столбце. Опять-таки для оценки объема выборки, необходимого для оценки долей с заданной точностью, существуют соответствующие программы. Например, presize [9].

Минимальные объемы выборок, необходимые для оценки двух значений долей 0,5 и 0,1 при трех значениях ширины 95%-х ДИ, представлены в табл. 3.

Таблица 3. Минимальные объемы выборок, необходимые для оценки долей с ожидаемыми значениями ширины 95%-го ДИ [9]

Table 3. Minimum sample sizes required to estimate proportions with expected 95% CI widths [9]

Ожидаемая доля	Ожидаемый точный 95%-й ДИ		Минимально необходимый объем выборки (n)
	ширина	границы	
0,5	0,20	0,40; 0,60	103
0,5	0,10	0,45; 0,55	401
0,5	0,05	0,48; 0,53	1570
0,1	0,10	0,06; 0,16	157
0,1	0,05	0,08; 0,13	865
0,1	0,02	0,09; 0,11	3458

Предположим, что в изучаемой популяции 50% субъектов имеют интересующий нас признак. Тогда для оценки ожидаемой доли с шириной 95%-го ДИ в 10% потребуется объем выборки не менее 401 (см. рис. 5).

Другими словами, если из такой популяции мы извлечем случайную выборку объемом не менее 401 человека, то с вероятностью 95% доля субъектов с данным признаком будет находиться в интервале от 45 до 55%.

Программа presize позволяет решать и обратную задачу – оценивать границы и ширину ДИ при заданных уровне доверия и объеме выборки.

Помимо оценки объема выборки для долей, программа позволяет осуществлять подобные вычисления для примерно 20 других показателей: для средних и их разностей, для различных мер корреляции, для показателей качества диагностических тестов, для отношения рисков и шансов и др.

СМ. Доверительные интервалы для нормально-го распределения

НХ. Такой заголовок наводит на мысль о ДИ для функции плотности нормального распределения (Н.Р.) или его кумулятивной функции. На самом деле автор (СМ) имеет в виду ДИ для параметров Н.Р.

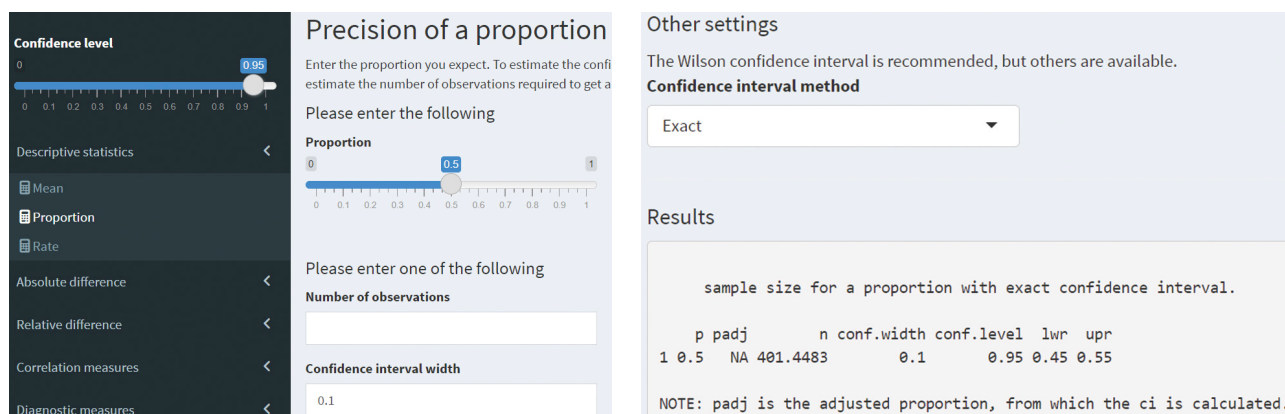


Рис. 5. Пример вычисления минимально объема выборки, необходимого для оценки доли $\phi = 0,50$ с шириной 95%-го ДИ равной 10% (по <https://shiny.ctu.unibe.ch/app/presize/>)

Fig. 5. An example of calculating the minimum sample size required to estimate a proportion $\phi = 0.50$ with a 95% CI width of 10% (on <https://shiny.ctu.unibe.ch/app/presize/>)

При этом автор (СМ) небрежно обращается с формулами и пренебрегает разъяснениями использованных символов. В формуле ДИ для среднего (математического ожидания) нормального распределения

$$m - t_{\alpha;v} S_{\bar{X}} < \mu < m + t_{\alpha;v} S_{\bar{X}}$$

следовало разъяснить, что μ – оцениваемое среднее. Если обозначение m для выборочного среднего еще более-менее узнаваемо, хотя сейчас его чаще обозначают прописным M , то обозначение $S_{\bar{X}}$ для стандартной ошибки среднего далеко не общепринято, чаще сейчас используют двухбуквенное обозначение SE . При этом наблюдается смешение обозначений: либо m следовало заменить на $\bar{X}$, либо $S_{\bar{X}}$ заменить на S_m . Для числа степеней свободы теперь тоже чаще используют не v , а двухбуквенное обозначение df .

Принципиальной ошибкой является использование α -квантиля t -распределения Стьюдента для вычисления ДИ с доверительной вероятностью $(1 - \alpha)$. Корректным должно быть выражение:

$$M - t_{1-\alpha/2; df} SE < \mu < M + t_{1-\alpha/2; df} SE.$$

Как и в случае с построением ДИ для доли, автор (СМ) предлагает пользователям находить необходимые для этого промежуточные значения квантилей распределения $\chi^2_{\alpha/2; df}$ и $\chi^2_{1-\alpha/2; df}$ в Excel. Опять-таки, вряд ли кто-либо из пользователей воспользуется такой рекомендацией.

Бутстреп

Обсуждая и рекомендуя процедуры интервальных статистических оценок, нельзя обойти молчанием их алгоритмически-вычислительные

аспекты. Благодаря революционному прогрессу вычислительной техники и технологии, теперь стали доступны для использования в «домашних условиях» более универсальные, чем традиционные, методы статистического оценивания на основе бутстрепа. Они не «вгоняют» изучаемые статистические данные в прокрустово ложе гипотетической вероятностной модели (в первую очередь, модель нормального распределения), а реконструируют исходное распределение («генеральную совокупность») посредством «ресемплинга» (формирования, генерации, порождения повторных выборок из имеющихся данных).

Вот цитата⁹: «На протяжении более столетия сложность, присущая статистике, основанной на формулах, сбивала с толку ученых, вызывала ошибки в исследованиях и заставляла миллионы студентов ненавидеть этот предмет. Сложность – это болезнь. Ресемплинг (формирование повторных выборок из имеющихся данных или совокупности, предполагаемой этими данными) является проверенным методом лечения этой болезни. Бутстреп, перестановки, рандомизации и другие трудоемкие компьютерные процедуры произвели революцию в статистике. Ресемплинг теперь является предпочтительным методом для построения доверительных пределов, проверки гипотез и других повседневных задач, связанных со статистическими выводами.

Вместо сложных формул и загадочных таблиц параметрических и непараметрических тестов, основанных на сложной математике и загадочных приближениях, основными инструментами

⁹ About Resampling. Available at: <https://resample.com/> (accessed: 27.01.2024).

ресемплинга (бутстрепа) являются «симуляции» (имитационное моделирование), созданные специально для поставленной задачи практиками, которые полностью понимают, что они делают и почему они делают это. Ресемплинг позволяет проверять гипотезы и рассчитывать доверительные интервалы для большинства типов данных, даже тех, которые невозможно проанализировать с помощью формул.

Усиливающийся поток научных работ с использованием техники ресемплинга как в качестве основного инструмента, так и для сложных приложений, свидетельствует о его несомненной полезности. Современная литература по математической статистике демонстрирует признание его теоретической основы после долгого блуждания впотьмах.

Тщательные исследования показывают, что пользователи, которых обучают идеям и методам ресемплинга, лучше усваивают статистику и им действительно нравятся такие вводные курсы. Они признают, что ресемплинг – это понятный и продуктивный способ преподавания статистики».

СМ. В табл. 4 представлены результаты оценивания основных параметров модельного распределения.

Таблица 4. Доверительные интервалы для основных параметров модельного Гауссова распределения объемом $n = 100$

Table 4. Confidence intervals for the main parameters of the model Gaussian distribution with size $n = 100$

Параметр	Точечная оценка	Границы 95 % доверительного интервала	
		нижняя	верхняя
Математическое ожидание	67,6	65,20	69,95
Стандартное отклонение	11,96	10,39	13,42
Асимметрия	0,12	–0,12	0,37
Экссесс	–0,38	–0,84	0,08
Коэффициент вариации	0,18	0,16	0,20

НХ. Иллюстрировать практические рекомендации искусственно смоделированными бессодержательными примерами – не лучший способ. Желательно, чтобы это были данные из реальной жизни, из реальных биомедицинских исследований.

Автор (СМ) не указал, какой генератор случайных значений для Н.Р. и какие программы для вычислений ДИ он использовал и рекомендует пользователям. Ниже приведены результаты оценки параметров методом бутстрепа для выборки объ-

емом $n = 100$ из Н.Р. с параметрами, указанными автором: среднее $M = 67,6$ и стандартным отклонением $SD = 11,96$.

Использован генератор истинно случайных чисел, снованный на атмосферном шуме¹⁰.

Для вычислений методом бутстрепа можно рекомендовать пакет PAST [10]¹¹. Пользователей не должно смущать, что этот пакет создан для палеонтологов. Статистические методы довольно универсальны и не имеют узкой специализации (табл. 5).

Таблица 5. Пример точечных и интервальных оценок основных параметров методом бутстрепа. Программа PAST [10]

Table 5. Example of point and interval estimates of the main parameters using the bootstrap method. Program PAST [10]

		Lower conf.	Upper conf.	Точечная оценка	Границы 95%-го ДИ	
					нижняя	верхняя
N	100					
Min	28,9					
Max	105					
Sum	6932	6686	7177			
Mean	69,3	66,9	71,8	67,6	65,2	70,0
Std. error	1,26	1,06	1,45			
Variance	159,0	112,1	211,7			
Stand. dev	12,61	10,58	14,54	11,96	10,39	13,42
Median	69,6	65,9	72,8			
25 prcntil	61,5	58,3	64,0			
75 prcntil	78,4	73,8	81,7			
Mode	NA					
Skewness	–0,25	–0,81	0,41	0,12	–0,12	0,37
Kurtosis	0,70	–0,41	1,90	–0,38	–0,84	0,08
Geom. mean	68,1	65,5	70,7			
Coeff. var	18,2	15,1	21,2	0,18	0,16	0,20

Справа – значения из табл. 3 обсуждаемой статьи автора (СМ) [3, S54].

На основании того, что ДИ для асимметрии и эксцесса накрывают нуль, автор (СМ) справедливо приходит к выводу: у нас появились серьезные основания считать, что распределение изучаемого признака в популяции является Гауссовым. Было бы странно, если распределение, сгенерированное как нормальное (Гауссово), существенно отличалось бы от него.

¹⁰ Gaussian Random Number Generator. Available at: <https://www.random.org/gaussian-distributions/> (accessed: 27.01.2024).

¹¹ На настоящий момент (январь 2024 г.) статья цитируется 29498 раз.

СМ. Для редких любителей оценивать еще и коэффициент вариации отметим, что у него тоже есть стандартная ошибка и доверительный интервал.

НХ. Оставим на совести автора (СМ) ничем не обоснованную идиосинкразию к коэффициенту вариации (CV). Напротив, CV применяется довольно часто при решении разнообразных задач. Он полезен хотя бы потому, что в отличие от других мер изменчивости его фактическое значение не зависит от единицы измерения, он является безразмерным числом (обычно выражается в процентах). Поэтому его следует, например, использовать при сравнении вариативности наборов данных с различными единицами измерения или сильно отличающимися средними значениями.

СМ. Попутно отмечу, что одним из признаков того, что выборка, взятая из нормального распределения, имеет достаточный объем, является величина коэффициента вариации $< 33\%$ [11, с. 81].

НХ. В оригинале это утверждение выглядит иначе:

В не слишком маленькой выборке из нормальной генеральной совокупности величина V [CV в современных обозначениях – Н.Х.] должна быть не больше 33 %.

К чести автора (Закса), или переводчиков в английском издании 1984 года цитируемого энциклопедического руководства по методам

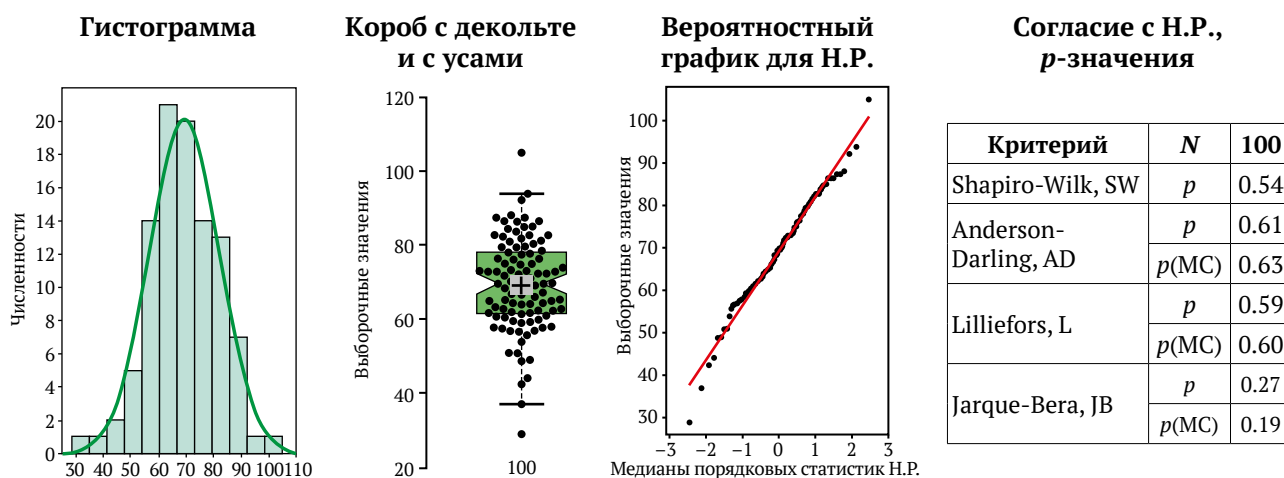
прикладной статистики этой загадочной фразы нет [12, р. 78]. К сожалению, автор, Лотар Закс, скончался 24 июня 2019 года, и уже не узнать, что он имел в виду.

Для построения гистограммы, вероятностного графика и проверки согласия данных с Н.Р. можно рекомендовать программу PAST [10], а для построения «коробов с декольте (с вырезами) и с усами» – программу VoxPlotR [13] (см. рис. 6). Они очень информативны: перетяжка короба – медиана, крест – среднее, высота серого прямоугольника – 95%-е ДИ для среднего, ширина вырезов – 95%-е ДИ для медианы. Красная прямая на вероятностном графике – теоретически ожидаемая спрямленная кумулятивная функция Н.Р.

СМ. Интервальная оценка медианы

В случае, когда результаты интервального оценивания асимметрии и эксцесса оказываются неутешительными (есть серьезные основания сомневаться в нормальности распределения), общепринятым является построение графика с «коробками» и «усами».

НХ. Построение графиков типа «коробы с декольте и с усами» отнюдь не ограничивается случаями «неутешительных» результатов проверки согласия данных с Н.Р. Оно пригодно во многих случаях. Это очевидно из представленного графика (рис. 6). Вообще не стоит привносить субъективную эмоциональную окраску (утешительно-неутешительно и т.п.) в обсуждение результатов статистического анализа.



Обозначения: Н.Р. – нормальное распределение, $p(MC)$ – оценка p -значения методом Монте-Карло.

Рис. 6. Пример графических методов описания статистических данных и проверки их согласия с нормальным распределением (Н.Р.)

Fig. 6. Example of graphical methods for describing statistical data and testing their agreement with the normal distribution (N.D.)

Автор (СМ) иллюстрирует свое утверждение рис. 7 и 8 (рис. 2 и 3 в статье СМ – НХ).

Рисунок 8 вряд ли приемлем, он противоречит рис. 7: закрашенные столбцы упираются в ось 0 и создают ложную иллюзию наличия нулевых значений в анализируемых данных.

В этом исследовании проверяли скорость и степень агрегации тромбоцитов до и после воздействия на сыворотку крови аспирином и двумя препаратами экстрактов простаты, т.е. это были сопряженные парные наблюдения.

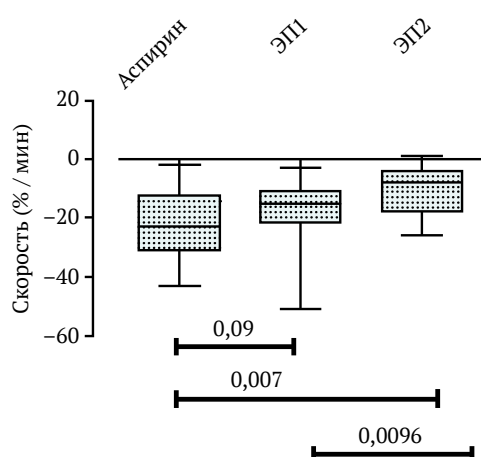


Рис. 7. Пример графика с «коробками» и «усами»

Fig. 7. Example of Box and Whisker plot

Для парных наблюдений (измерений) следует использовать графики совершенно иного вида – графики Гарднера-Олтмена (Gardner-Altman estimation plot) (рис. 9). Для них создан пакет соответствующих программ. Они описаны в статье [14] и в Википедии¹².

Суть в том, что с помощью бутстрепа моделируют (реконструируют) распределение для **разности** средних и вычисляют ДИ для этих разностей.

¹² Estimation statistics. Available at: https://en.wikipedia.org/wiki/Estimation_statistics (accessed: 27.01.2024).

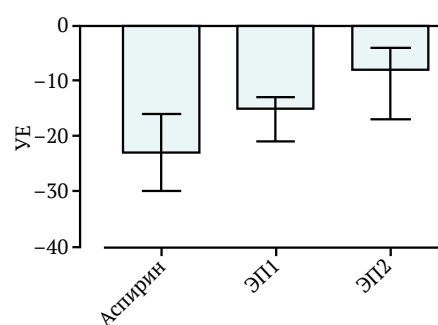
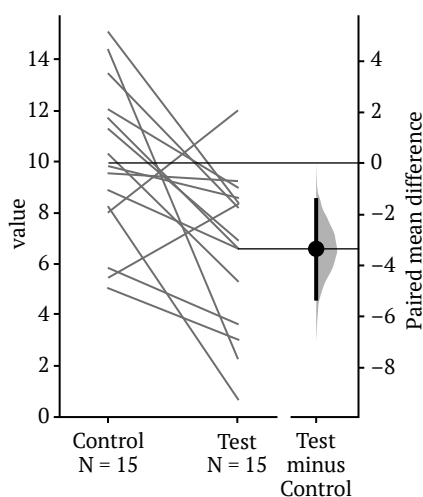
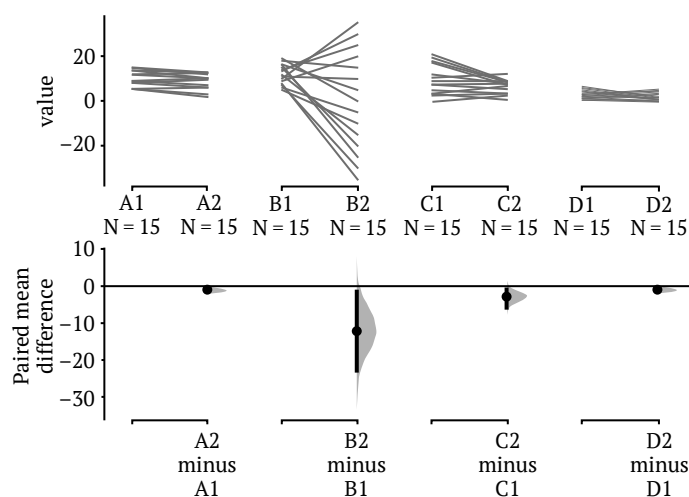


Рис. 8. Медианы для парных сравнений скорости агрегации тромбоцитов до и после воздействия препаратов. Плашками погрешностей отложены 95% доверительные интервалы

Fig. 8. Medians for paired comparisons of platelet aggregation rates before and after drug exposure. Error bars demonstrate 95% confidence intervals



A



B

Рис. 9. Графики Гарднера-Олтмена (Gardner-Altman) для визуализации сравнений парных наблюдений:

A – парные сравнения (<https://www.estimationstats.com/#/analyze/paired>);

B – множественные парные сравнения (<https://www.estimationstats.com/#/analyze/multi-paired>)

Fig. 9. Gardner-Altman plots for visualizing comparisons of paired observations

A – paired (<https://www.estimationstats.com/#/analyze/paired>);

B – multiple paired (<https://www.estimationstats.com/#/analyze/multi-paired>)

Там же можно строить ДИ для разностей средних независимых выборок (рис. 10).

Еще один вид полезных графиков – «дождевые облака» – представлен на рис. 11 [15].

Их построение реализовано, например, в пакете JASP [16].

СМ. Для медианы можно также построить доверительный интервал. Способ его построения изложен, в частности, в Национальном стандарте Российской Федерации¹³.

¹³ Национальный стандарт Российской Федерации. Статистические методы. Статистическое представление данных. Оценка медианы. ГОСТ Р ИСО 16269-7-2004. Дата введения: 01.06.2004.

НХ. В этом стандарте предлагается вычислять ДИ для медианы фактически вручную и с использованием таблиц. Более компактно процедура описана в цитируемом автором учебном пособии [6]. Однако в настоящее время есть программы, которые вычисляют границы ДИ для медианы: AtteStat, StatAnt, PAST др.

СМ. Результаты интервального оценивания можно представлять по-разному. Автору представляется наиболее симпатичным вариант, когда границы доверительного интервала представлены в виде нижних индексов. Аспирин – $_{30}^{23}$ ₁₇, ЭП1 – $_{21}^{15}$ ₁₃, ЭП2 – $_{48}^{17}$.

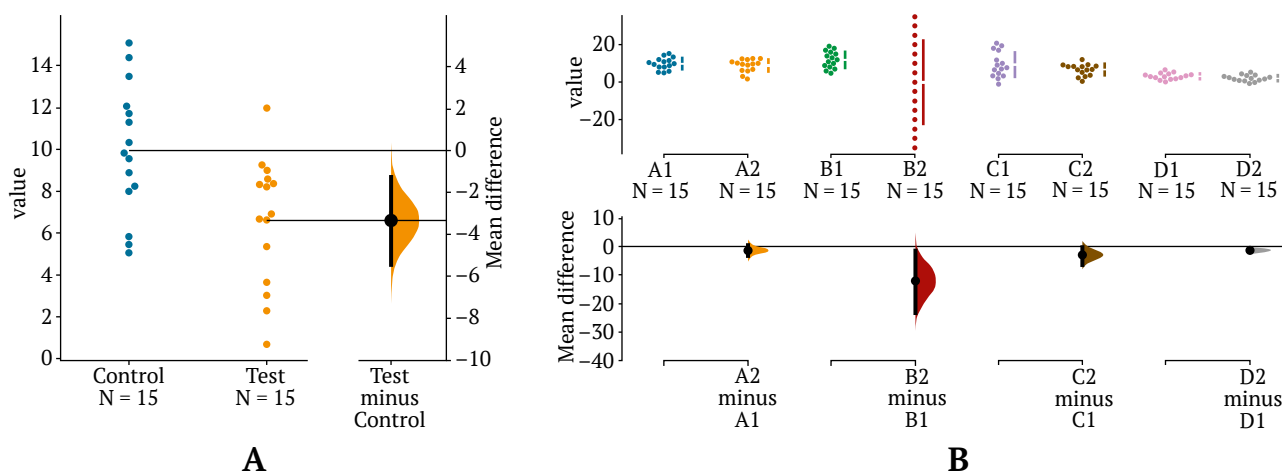


Рис. 10. Графики Гарднера-Олтмена (Gardner-Altman) для визуализации сравнений независимых выборок: **A** – сравнение двух независимых выборок (<https://www.estimationstats.com/#/analyze/two-independent-groups>); **B** – множественные попарные сравнения (<https://www.estimationstats.com/#/analyze/multi>)

Fig. 10. Gardner-Altman plots for visualizing independent sample comparisons **A** – two groups (<https://www.estimationstats.com/#/analyze/two-independent-groups>); **B** – multiple two-groups (<https://www.estimationstats.com/#/analyze/multi>)

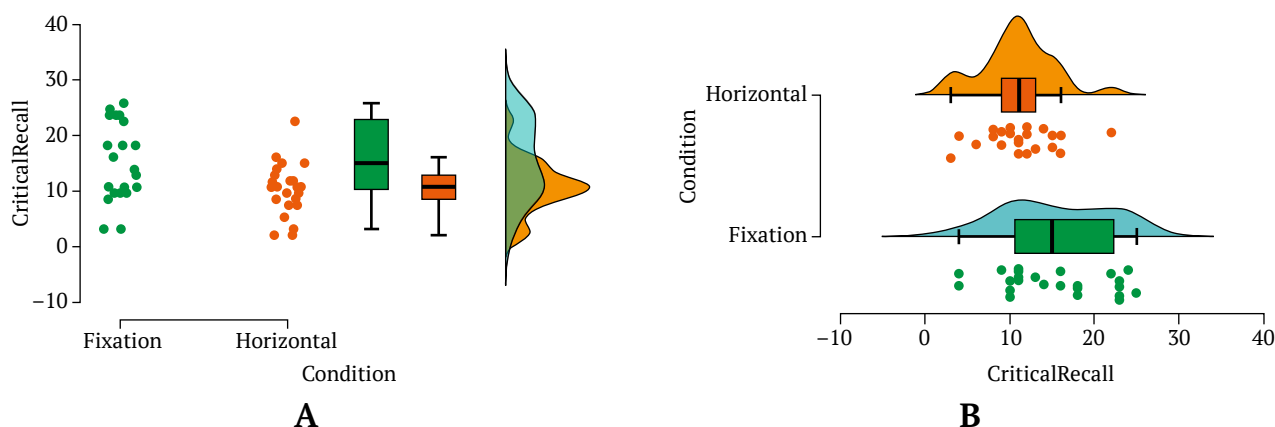


Рис. 11. Графики «дождевые облака»: **A** – вертикальные; **B** – горизонтальные

Fig. 11. Rain cloud plots: **A** – vertical; **B** – horizontal

НХ. Тут явно стоило дать ссылку на первоисточник [17].

СМ. Таблицы сопряженности

НХ. В этом разделе обсуждаются только таблицы 2×2 .

СМ. Для оценки эффективности вакцины используют формулу, в которую входит соотношение шансов (Odds Ratio). Результаты ее оценки представлены в табл. 6. Если автор следует рекомендациям ICMJE, он представит и доверительный интервал (в данном случае – 99%).

НХ. Допущена небрежность: очевидно, имеется в виду доверительная вероятность (уровень уверенности), а не ДИ.

СМ.

Таблица 6. Результаты фазы III клинического испытания вакцины Спутник V

Table 6. Results of Phase III clinical trial of the Sputnik V vaccine

Группа	Здоровы	Заболели	Всего
Вакцина	14948	16	14964
Плацебо	4840	62	4902
Всего	19788	78	19866

Таблица 7. Результаты статистического анализа таблицы 6, полученные с помощью статистического пакета GraphPad Instat

Table 7. The results of the statistical analysis of Table 6 obtained by the GraphPad Instat statistical package

Chi-square, df	126.6, 1	Точные оценки
P value	$P < 0.0001$	$2,3 \cdot 10^{-29}$ *
Odds ratio	11,97	$3,3 \cdot 10^{-24}$ **
99% confidence interval	5.803 to 24.68	6,9 to 20,8***
Positive Predictive Value	0,9989	
99% confidence interval	0.9983 to 0.9994	

Использованы программы:

* <http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pchisq.html>

** <https://www.cog-genomics.org/software/stats>

*** LePAC

НХ. Тут автор (СМ) указал, какую программу он использовал (табл. 7). Он ратует за применение точных методов статистики, но применил асимптотический критерий хиквадрат и продолжает рекомендовать вычислять p -значения в Excel.

В результате получаются «точные» p -значения для «неточного» асимптотического критерия. К тому же p -значение приведено как неравенство $P < 0.0001$, а не реальное значение $p = 2,3 \cdot 10^{-29}$ (для $\chi^2 = 126,6$ и $df = 1$)¹⁴. В остальных таблицах обсуждаемой статьи реальные p -значения указаны. Однако для таблиц сопряженности 2×2 существует точный критерий Фишера, вычисления которого давно уже не имеют ограничений на численности в таких таблицах. Если довольствоваться вычислениями p -значений, то удобно использовать надежную онлайн-программу¹⁵. Ее особенность состоит в том, что можно вычислять mid- p – значения p с поправкой на консервативность точного метода (рис. 12).

OR (которое лучше называть отношением шансов за–против) является мерой связи (ассоциации) между двумя событиями – вакцинацией и ее результатом. Поэтому целесообразно оценивать еще два других важных показателя: разность долей (PD) и их отношение (называемое также отношением рисков RR).

Для вычисления точечных и интервальных оценок для всех трех удобно использовать уже упомянутую выше программу LePAC, в ней реализованы бейзовские подходы. Точные частотные ДИ для перечисленных параметров вычисляет также коммерческий пакет StatXact (рис. 13).

Statistical tests

Fisher's exact test p-value calculator, 2x2 and 2x3

62	4840	
16	14948	
clear	Alternative hypothesis: Two-sided	
☐ mid-p adjustment		
P-value: 3.29458425254e-24		

A

Statistical tests

Fisher's exact test p-value calculator, 2x2 and 2x3

62	4840	
16	14948	
clear	Alternative hypothesis: Two-sided	
☒ mid-p adjustment		
P-value: 1.78182028181e-24		

B

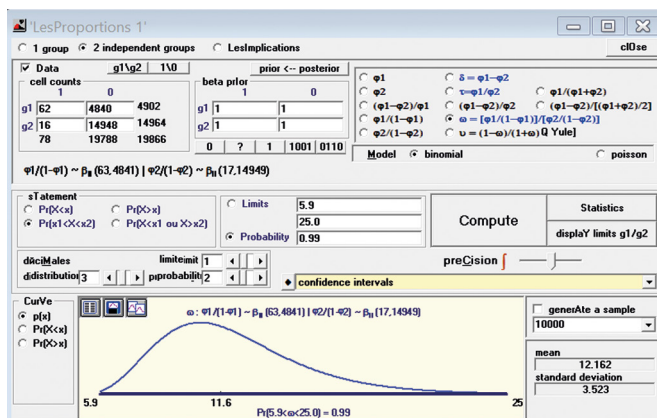
Рис. 12. Вычисление точных значений p (A) и mid- p (B)

Fig. 12. Calculation of exact p (A) and mid- p (B) values

¹⁴ Вычислено с помощью упомянутого выше калькулятора: <http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pchisq.html>

¹⁵ Statistical tests. Available at: <https://www.cog-genomics.org/software/stats> (accessed: 27.01.2024).

Оценка отношения шансов за-против



Odds Ratio of Two Binomial Proportions

binomial (type = independent, test_type = oddsratio, compute = ci, method = asymp, popl = col 0.99, time_limit = none);

Data File:

Number of Rows: 2

Number of Columns: 2

Summary of the Test Statistic:

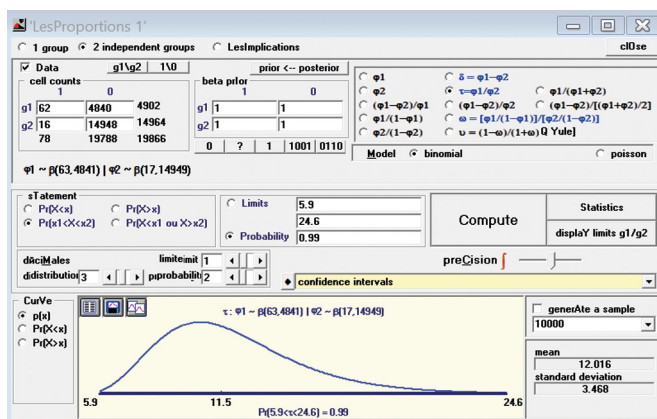
Binomial Proportions for [col1]: pi_1	0.2051
Binomial Proportions for [col2]: pi_2	0.7554
Odds Ratio:	11.97

Inference:

Type	P-Value 2-Sided	99.00% CI Limits	
		Lower	Upper
Asymptotic	7.16e-010	5.805	24.67

$$OR = 12,0 (5,8; 24,7)$$

Отношение долей (рисков)



Confidence Interval on Ratio of Two Binomial Proportions Based on the Standardized Statistic and Inverting Two 1-Sided Tests

binomial (type = independent, test_type = ratio, test_statistic = std1, compute = ci, method = asymp, gamma time_limit = none);

Data File:

Number of Rows: 2

Number of Columns: 2

Summary of the Test Statistic:

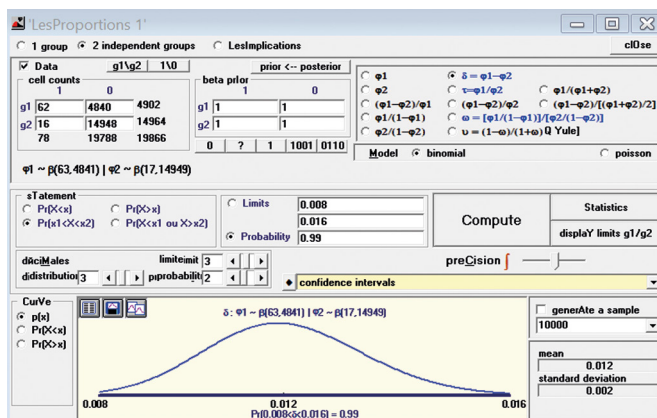
Binomial Proportions [col1]: piHat_1	0.001069
Binomial Proportions [col2]: piHat_2	0.01265
Ratio of Proportions: piHat_2/piHat_1	11.83
Std. Error: (pooled estimate of stdev of piHat_2 - piHat_1)	0.001029
Standardized Difference: (piHat_2 - piHat_1)/Std. Error	11.25

Inference:

Type	Tail	P-Value		95.00% Conf. Interval for pi_2/pi_1	
		1-Sided	2*1-Sided	Lower Limit	Upper Limit
Asymptotic	.GE.	1.15e-029	2.3e-029	6.88	20.34

$$RR = 11,3 (6,9; 20,3)$$

Разность долей



Confidence Interval on Difference of Two Binomial Proportions Based on the Standardized Statistic and Inverting Two 1-Sided Tests

binomial (type = independent, test_type = difference, test_statistic = std1, compute = ci, method = asymp, popl = col gamma = 1e-006, conf = 0.99, time_limit = none);

Data File:

Number of Rows: 2

Number of Columns: 2

Summary of the Test Statistic:

Binomial Proportions [col1]: piHat_1	0.0008021
Binomial Proportions [col2]: piHat_2	0.01265
Difference of Proportions: piHat_2 - piHat_1	0.01185
Std. Error: (pooled estimate of stdev of piHat_2 - piHat_1)	0.001003
Standardized Difference: (piHat_2 - piHat_1)/Stderr	11.81

Inference:

Type	Tail	P-Value		99.00% Conf. Interval for (pi_2 - pi_1)	
		1-Sided	2*1-Sided	Lower Limit	Upper Limit
Asymptotic	.GE.	0	0	0.008277	0.01669

$$PD = 0,012 (0,008; 0,017)$$

A

B

Рис. 13. Точечные и интервальные оценки отношения шансов (OR), отношения долей (RR) и разности долей (PD) с 99%-ми ДИ в программах LePAC (A) и StatXact (B). Бейзовский подход

Fig. 13. Point and interval estimates of odds ratio (OR), proportion ratio (RR), and proportion difference (PD) with 99% CI in LePAC (A) and StatXact (B) programs. Bayesian approach

Здесь также можно ознакомиться с результатами применения других методов оценивания (рис. 14).

ДИ для RR и OR одновременно можно вычислить (несколькими методами) также с помощью следующего калькулятора: <https://effect-size-calculator.herokuapp.com/> (рис. 15).

Совершенно непонятно, зачем автор (СМ) приводит точечную и интервальную оценки PPV – предсказательной вероятности положительных результатов диагностического теста (индексного теста) при оценке его информативности, точности и надежности путем сравнения с результатами референсного теста («золотого стандарта»). Никакого отношения к оценке эффективности вакцины эта величина не имеет.

СМ. Доверительные интервалы и точное значение p в дисперсионном анализе

Сначала был рассчитан внутриклассовый коэффициент вариации, представляющий собой точечную оценку силы влияния фактора. Он оказался равным 77. Затем был построен доверительный интервал. Способ построения описан в одном из лучших, на взгляд автора, учебников по вариационной статистике [6, с. 81].

НХ. Это не «внутриклассовый коэффициент вариации», а **коэффициент внутриклассовой корреляции**. ДИ для него можно вычислить с помощью пакетов 'irrICC' и 'ICC' для системы R.

Внутриклассовый коэффициент корреляции по определению является долей влияний изучаемого фактора в общей дисперсии, соответственно,

Разность двух независимых долей	Отношение долей (рисков)	Отношение шансов за–против
confidence intervals	confidence intervals	confidence intervals
two-sided 0.99	two-sided 0.99	two-sided 0.99
[0.008,0.016] Bayes prior Beta(1/2,1/2) - Beta(1/2,1/2) [Jeffreys]	[5.981,25.502] Bayes prior Beta(1/2,1/2) - Beta(1/2,1/2) [Jeffreys]	[6.035,25.859] Bayes prior Beta(1/2,1/2) - Beta(1/2,1/2) [Jeffreys]
[0.008,0.016] Bayes prior Beta(1,1) - Beta(1,1)	[5.838,23.971] 'score-test' - Koopman	[5.840,27.197] conditional including 'exact' [Cornfield]
[0.008,0.016] Bayes prior Beta(1,0) - Beta(0,1)		[6.237,24.615] conditional excluding
[0.008,0.016] Bayes prior Beta(0,1) - Beta(1,0)		[6.001,26.109] conditional 'mid-p'
[0.007,0.016] pseudo Jeffreys (Brown, Li)		[5.805,24.674] Woolf
[0.007,0.016] simple asymptotic (Wald)		[5.859,24.445] Gart
[0.007,0.016] simple asymptotic with continuity correction		[5.727,23.893] Gart2
[0.007,0.016] simple asymptotic recentered (Brown, Li)		[5.891,24.310] 'score-test'
[0.008,0.016] Beal: Haldane		[6.091,25.494] Killier
[0.008,0.016] Beal: Jeffreys-Perks		
[0.008,0.016] Beal revisited [Roths and Tebbs]: 'Maximum likelihood'		
[0.008,0.016] Beal revisited [Roths and Tebbs]: 'Method of moments'		
[0.008,0.017] 'score-test' - Mee		
[0.008,0.017] 'score-test' - Mee with continuity correction		
[0.008,0.017] 'score-test' - Miettinen-Nurminen		
[0.008,0.016] 'true profile likelihood' - Miettinen-Nurminen		
/ 'exact'		
/ 'mid-p'		
[0.008,0.016] 'score' (Newcombe)		
[0.008,0.017] 'score' with continuity correction (Fleiss)		
[-0.007,0.030] Yule		
[-0.007,0.030] modified Yule		
[0.008,0.016] Agresti-Caffo (Wald '+2')		
[0.008,0.016] Agresti-Caffo (Wald '+z^2_alpha/2')		
[0.008,0.016] Zhou-Qin - 'direct Edgeworth expansion'		
[0.007,0.016] Zhou-Qin - 'transformation'		
[0.008,0.017] Herranz-Martin		
[0.008,0.017] Herranz-Martin with continuity correction		

Рис. 14. Вычисление 99%-х ДИ для OR, RR и PD разными методами

Fig. 14. Calculation of 99% CI for OR, RR and PD by different methods

Odds/risk/absolute ratios & Number needed to treat

Inputs		Results	
	Yes	Outcome Frequency	
		No	
Treatment	62	4840	Odds ratio: 11.9654
Control	16	14948	Risk ratio/Relative risk: 11.8289
Method (Odds-ratio):	Conditional maximum likelihood estimation (Fisher)		Lower limit on odds ratio: 5.8402
Method (Relative-risk):	Bootstrap estimation (boot)		Lower limit on risk ratio: 6.1053
Compute relative risk reduction in place of relative risk?:	No		Upper limit on odds ratio: 27.2004
Confidence Interval:	99		Upper limit on risk ratio: 29.6541
			Number needed to treat: 86.3657351
			Absolute risk: 0.0115787

Calculate Clear

$$OR = 12,0 (5,8; 27,0); RR = 11,8 (6,1; 29,7)$$

Рис. 15. Одновременная оценка 99%-х ДИ для OR и RR

Fig. 15. Simultaneous estimation of 99% CI for OR and RR

область допустимых значений для него заключена в границах от 0 до +1, он никак не может быть равным 77.

Словосочетание «вариационная статистика» – устаревшее.

Номер страницы в ссылке указан неверно, правильно: с. 181.

СМ. В табл. 8 (табл. 6 в статье СМ – НХ) представлены результаты анализа частоты микроядер в лимфоцитах периферической крови мышей, проведенные методом двухфакторного дисперсионного анализа [18]. Сразу обращает на себя внимание тот факт, что примененная программа не рассчитала точных значений p . Это резко снизило шансы данной работы попасть в метаанализ.

Таблица 8. Результаты двухфакторного дисперсионного анализа частоты микроядер в лимфоцитах периферической крови мышей

Table 8. Results of two-way ANOVA of micronucleus frequency in peripheral blood lymphocytes of mice

Фактор	SS	d.f.	MS	F	Значение p (стат-пакет)	Значение p (точное)
Эпиталон	58,54	1	58,54	74,23	0,00	$1,6 \cdot 10^{-12}$
Циклофосфан	32,00	1	32,00	40,57	0,00	0,0017
Взаимодействие факторов	28,50	1	28,50	36,13	0,00	$8,1 \cdot 10^{-8}$
Неконтролируемые факторы	53,63	68	0,79	–	–	–
Общая дисперсия	191,375	71			–	

НХ. Это не единственный недостаток цитируемой работы: в разделе «Материалы и методы» отсутствует обязательное описание использованных статистических методов и программ. «Попадание в метаанализ» при наличии точных p -значений мы обсудим далее в разделе «Метаанализ».

СМ. Точные значения были получены с помощью MS Excel в течение примерно пяти минут (включая запуск программы). Для этого нужно было ввести в ячейку выражение вида

$$=F.PAC.PX(74,23; 1; 68)$$

НХ. Автор (СМ) почему-то не привел результат вычисления.

В одной из дружественных уже упомянутых программ¹⁶, которая производит вычисления точ-

ных p -значений, такая процедура занимает секунды (рис. 16).

P-Value Calculator for F Distribution

Degree of freedom 1 (numerator) :

Degree of freedom 2 (denominator) :

F-value:

p-value:

p-value type:
☒ right tail
☐ left tail

F = 74.23, df1 = 1, df2 = 68

Right-tail p-value is 1.673e-12

R command: `pf(74.23, 1, 68, lower.tail=FALSE)` or `1-pf(74.23, 1, 68)`

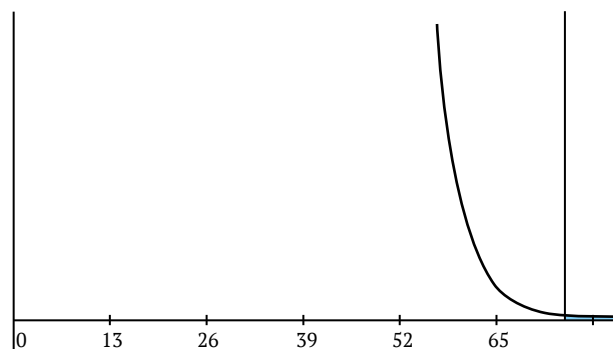


Рис. 16. Вычисление p -значения для $F = 74,23$ при $df_1 = 1$ и $df_2 = 68$

Fig. 16. Calculation of p -value for $F = 74.23$ with $df_1 = 1$ and $df_2 = 68$

СМ. Доверительные интервалы в регрессионном анализе

Результаты регрессионного анализа обычно представляют так, как на рис. 17.

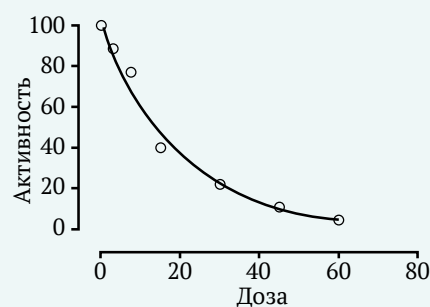
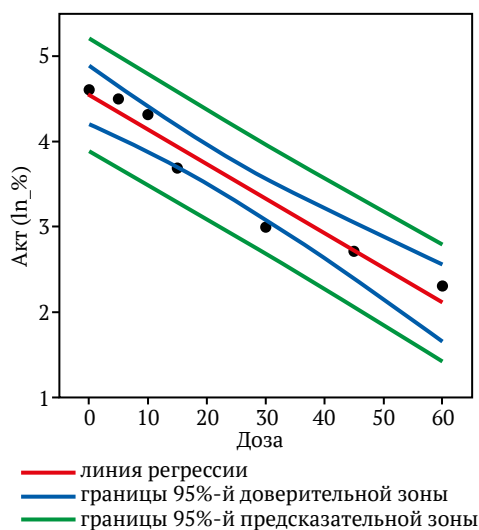


Рис. 17. Регрессия активности фермента по дозе ионизирующего излучения (модельная ситуация из источника [19])

Fig. 17. Regression of enzyme activity by ionizing radiation dose (model situation from [19])

¹⁶ P-Value Calculator for F Distribution. Available at: <http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pf.html> (accessed: 27.01.2024).



$$Y = {}_{4,3}4,5_{4,8} + ({}_{-0,046}-0,041_{-0,020})X;$$

$$R^2 = {}_{0,65}0,95_{0,99}; \quad p = 2 \cdot 10^{-4}$$

Рис. 18. Регрессионный анализ данных из рис. 4 в статье (график после логарифмического преобразования (нормальный логарифм))

Fig. 18. Regression analysis of data from Fig. 4 in the article (graph after logarithmic transformation (normal logarithm))

НХ. Строго говоря, это не результат регрессионного анализа, а всего лишь дозовая зависимость. В источнике [19] самой этой кривой нет, есть только исходные данные. Результаты ее регрессионного анализа представлены на рис. 18.

СМ. В табл. 9 представлены результаты регрессионного анализа зависимости, представленной на рис. 8.

Таблица 9. Стандартная форма представления результатов регрессионного анализа

Table 9. Standard form of presentation of regression analysis results

Параметр	Коэффициент регрессии	Стандартная ошибка	Значение t	Значение p (стат-пакет)	Значение p (MS Excel)
Свободный член	4,62	0,052	87,53	0,0000	$5,10 \cdot 10^{-08}$
Множитель	-0,05	0,002	-30,52	0,0000	$3,01 \cdot 10^{-06}$

НХ. Никаких объяснений этим значениям не дано. Какая зависимость (модель) предложена для этих данных, непонятно, ДИ, за которые ратует автор (СМ), не представлены в явном виде. Нисходящая плавная кривая на рис. 17 выглядит

довольно типичной для дозовых зависимостей, и ее можно спрямить, для чего чаще всего подходит логарифмическое преобразование. Пришлось поломать голову, чтобы догадаться, что автор (СМ) применил преобразование натурального логарифма.

Логарифмированные данные можно анализировать как линейную зависимость вида:

$$Y = a + bX,$$

для которой программ PAST [10] (и некоторые другие, например, JASP [16]) выдает уравнение линейной регрессии:

$$Y = {}_{4,3}4,5_{4,8} + ({}_{-0,046}-0,041_{-0,020})X.$$

Здесь подстрочные индексы слева и справа от точечных оценок коэффициентов линейной регрессии суть нижние и верхние границы 95%-х ДИ [17].

Непонятно, как автор (СМ) получил значения в табл. 9 и какую программу он использовал. Правильные значения указаны в табл. 10. Значение t -критерия Стьюдента и соответствующее p -значение для свободного члена не имеют практического смысла, так как их вычисляют для проверки равенства нулю этого члена.

Таблица 10. Результаты регрессионного анализа

Table 10. Regression analysis results

Параметр	Значения параметров	Стандартная ошибка (SE)	Значение t	p -значение
Свободный член	4,54	0,13	—	—
Угловой коэффициент (наклон)	-0,041	0,0042	-9,56	$2 \cdot 10^{-4}$

СМ. Следует помнить, что:

– у обоих коэффициентов регрессии есть стандартные ошибки, а значит, и доверительные интервалы, следовательно, правильный график регрессии должен представлять собой не линию, а полосу, так, как это сделано на рис. 19;
– если используемая программа не смогла вычислить точное значение p , как в табл. 6 (нумерация как в статье – НХ), это должны сделать авторы, например, с помощью MS Excel.

НХ. На рис. 19 представлена некая зависимость, которая кардинальным образом отличается от обсуждаемой.

Опять автор (СМ) предлагает вычислять реальные p -значения «вручную».

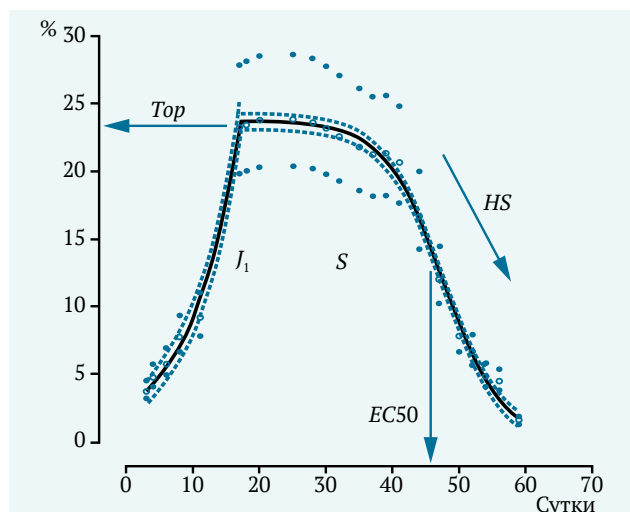


Рис. 5. Зависимость дисперсии выживаемости от возраста когорты у *Drosophila melanogaster*. По оси X – возраст когорты; по оси Y – стандартное отклонение доли живых особей; пунктирная линия – 95% ДИ регрессии; светлые кружки – точечные оценки; темные кружки – границы 99% ДИ для стандартного отклонения [20]

Fig. 5. Regression of survival variance by cohort age in *Drosophila melanogaster*. On the X-axis – the age of the cohort; on the Y-axis – the standard deviation of the proportion of living individuals; dotted line – 95% CI regression; light circles – point estimates; dark circles – the boundaries of 99% CI for the standard deviation [20]

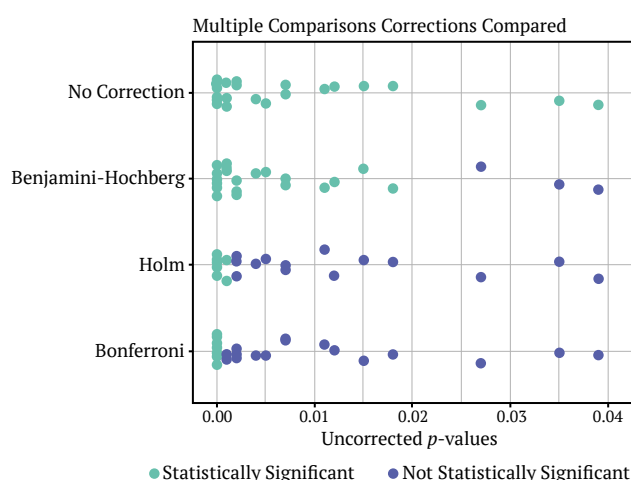
Полезным показателем качества выбранной модели (в данном случае линейной) является коэффициент детерминации R^2 . Он показывает, какая доля изменчивости в данных объясняется выбранной моделью. Но чтобы особо не обольщаться высоким значением его точечной оценки $R^2 = 0,95$, для него также следует оценить ДИ. Для этого можно использовать программу <https://effect-size-calculator.herokuapp.com/>

В данном примере $R^2 = {}_{0,65}^{0,95} {}_{0,99}^{0,99}$, ДИ излишне широк, что объясняется малым количеством точек ($n = 7$) на графике.

Зависимость на рис. 19 также никак не объяснена: каким уравнением она описывается, как получены границы ДИ, неясно.

СМ. Метод Холма (поправка Холма – Бонферрони)

НХ. СМ доходчиво излагает процедуру вычисления p -значений, подправленных на множественность сравнений по Холму, однако не сообщает, с помощью каких программ их (и другие поправки) можно производить.



A

Bonferroni		More false negatives
Bonferroni Step-Down		
Westfall and Young Permutation		
Benjamini and Hochberg False Discovery Rate		
None		More false positives

B

Рис. 20. Визуализация различий между основными способами поправки на множественность сравнений: A – <https://egap.org/resource/10-things-to-know-about-multiple-comparisons/> ;

B – https://physiology.med.cornell.edu/people/banfelter/qbio/resources_2008/1.5_GenespringMTC.pdf

Fig. 20. Visualization of the differences between the basic multiple comparison correction methods:

A – <https://egap.org/resource/10-things-to-know-about-multiple-comparisons/> ;

B – https://physiology.med.cornell.edu/people/banfelter/qbio/resources_2008/1.5_GenespringMTC.pdf

Предложено множество (несколько десятков) подходов к проблеме множественных сравнений. Иллюстрация того, как могут выглядеть соотношения между результатами популярных поправок, представлена на рис. 20.

Какой бы метод ни был использован, одних подправленных p -значений недостаточно, нужны интервальные оценки наблюдаемых эффектов.

В этом отношении для попарных сравнений после проведенного дисперсионного анализа наиболее популярным и надежным признан *post hoc* HSD критерий Тьюки, его бутстреп-версии реализованы в программах PAST [10], JASP [16].

Обсуждаемая статья предназначена как авторам, так и редакторам. Поэтому уместно воспроизвести мнение редакторов одного из медицинских журналов [21]:

«1. Запланируйте сравнения с самого начала. Если вы этого не сделали и изучили данные, прежде чем решить, какой анализ провести, рассмотрите эти апостериорные анализы как все возможные сравнения.

2. Решите, независимы ли ваши сравнения друг от друга или некоторые из них могут быть связаны друг с другом.

3. Рассмотрите возможность корректировки альфа-уровня (уровня значимости) или p -значений для нескольких сравнений.

4. Четко изложите свои решения в разделе “Методы”.

5. Обсудите, как ваши методы могли повлиять на ваши результаты в разделе “Обсуждение”.

6. Если у вас есть сомнения, обратитесь к дружелюбному специалисту по биостатистике» [21].

Корреляция

Конечно, в одной статье невозможно осветить все аспекты построения интервальных оценок, но про коэффициент корреляции следовало упомянуть. Большие значения его точечных оценок могут неоправданно обнадеживать исследователей. Простой (реальный) пример представлен на рис. 21.

Результаты оценки коэффициента линейной корреляции Пирсона r с использованием пяти программ представлены в табл. 11.

Можно видеть, что в данном примере значения нижних границ ДИ для линейного коэффициента корреляции Пирсона r резко различаются. Это обусловлено тем, что его распределение очень сложное и не имеет простого аналитического решения, и разные программы используют для него различные приближения. В недавней статье Gnambs [23] приведены 10 формул разной сложности для стандартной ошибки для r . Очевидно, математики и программисты должны обратить на них пристальное внимание и прийти к консенсусу, а пользователи вынуждены знать об этих тонкостях и, по возможности, вникать в них.

Доверяя, повторяй.

Повторение – мать познания

Мы должны задумываться над тем, не является ли полученный результат артефактом вследствие изменчивости выборки, которую нам удалось обнаружить в результате наивного однократного применения классической проверки гипотез, и поэтому лучший способ решить этот вопрос – **провести эксперимент еще раз**.

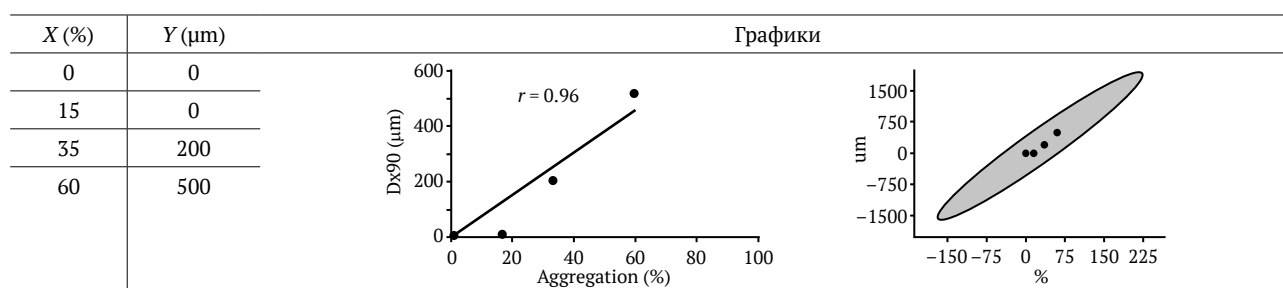


Рис. 21. Пример корреляции

Fig. 21. Example of correlation

Таблица 11. Точечные и интервальные оценки коэффициента линейной корреляции Пирсона r с применением семи методов

Table 11. Point and interval estimates of Pearson's linear correlation coefficient r using seven methods

Программа	Алгоритм	Точечная оценка	95% ДИ	p	BF10
Vassar*	МП	0,964	0,04; 1,00	–	–
StatAnt**	МП	0,964	0,037; 0,999	0,018	–
PAST [10]	МП	0,964	–	0,081 (MC)	–
Psychometrica***	[22; 23]	0,964	0,825; 1,000	0,036	–
JASP [16]	МП	0,964	0,037; 0,999	0,036	–
	Бутстреп	0,964	0,870; 1,000	0,036	–
	Бейз	0,754 (Me)	0,198; 0,996	–	2,69

Обозначения: МП – максимум правдоподобия; MC – Монте Карло; Me – медиана

* <http://www.vassarstats.net/>

** <https://sourceforge.net/projects/statant/>

*** <https://www.psychometrica.de/correlation.html#confidence>

СМ. Не только значение p

В рекомендациях ICMJE есть следующая фраза: «Не следует полагаться исключительно на проверку статистических гипотез, например, на использование значений p , которые не содержат важную информацию о размере эффекта (effect size)».

Полагаем, в нее следует добавить упоминание о двух вероятностях:

– вероятности получить в следующей повторности разницу между опытом и контролем с тем же знаком P_{rep} ;

– вероятности статистически значимого воспроизведения P_{srep} , то есть вероятности того, что в следующей повторности вы получите разницу между опытом и контролем с тем же знаком, и эта разница будет статистически значима.

Приведенная ниже табл. 12 была составлена на компьютере для учебно-методического пособия [24].

Таблица 12. Результаты расчетов P_{rep} и P_{srep} для нескольких модельных ситуаций [24, с. 14]

Table 12. Results of P_{rep} and P_{srep} calculations for several model situations

N_1	N_2	Значение t	Значение p	Значение P_{rep}	Значение P_{srep}
50	50	2,00	$p = 0,05$	0,920	0,504
50	50	2,65	$p = 0,01$	0,968	0,678
50	50	3,41	$p = 0,001$	0,991	0,838
150	150	2,00	$p = 0,05$	0,921	0,511
150	150	2,60	$p = 0,01$	0,967	0,674
150	150	3,35	$p = 0,001$	0,991	0,837

НХ. Действительно, p -значение ничего не говорит ни о вероятности нулевой гипотезы, ни о размере эффекта. С точки зрения байесовской статистики половина p -значения есть всего-навсего вероятность того, что истинное направление эффекта противоположно наблюдаемому в проведенном исследовании. Не более того.

Рекомендуемые СМ вероятности воспроизведения P_{rep} и P_{srep} предложил в 2005 г. Killeen [25]. Статья сразу вызвала много споров, резкую критику, противоречивые мнения и интерпретации, выявила терминологическую путаницу¹⁷.

Поскольку предлагаемые вероятности воспроизведения далеко не общеприняты, СМ обязан был сообщить, какие программы их вычисляют. Очевидно, он использовал программу LePrep¹⁸. Вот результаты для данных в первой и последней строках табл. 12 (см. рис. 22).

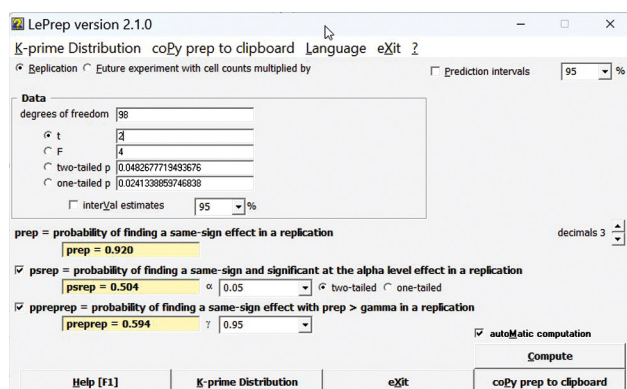
Программа LePrep позволяет оценивать вероятности воспроизведения наблюдаемых p -значений, см. табл. 13.

Жирным шрифтом выделены значения вероятностей, которые можно считать убедительно высокими. Закрашены равновероятные значения.

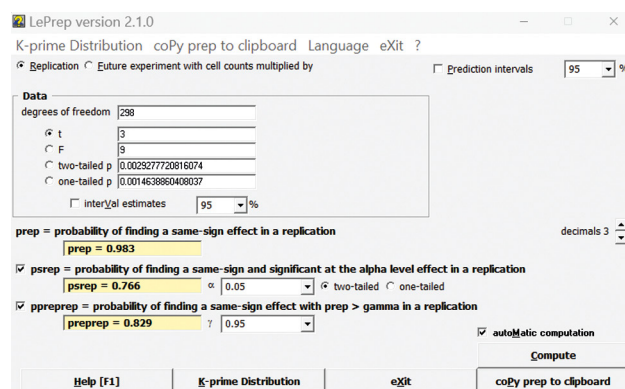
Можно видеть, что вероятность того, что при повторении эксперимента воспроизведется полученное p -значение, равна 50%. Чтобы достигнуть высокой вероятности для желаемого p -значения, исходное наблюдаемое p -значение должно быть существенно (фактически на порядок) меньше.

¹⁷ P-rep. Available at: <https://en.wikipedia.org/wiki/P-rep> (accessed: 27.01.2024).

¹⁸ Télécharger LePAC [Windows]. Available at: <https://eris62.eu/ErisLePAC.html> (accessed: 27.01.2024).



A



B

Рис. 22. Вычисление трех вероятностей воспроизводимости p_{rep} , p_{srep} и p_{reprep} :

A – $n_1 = n_2 = 50$; **B** – $n_1 = n_2 = 150$

Fig. 22. Calculation of three reproducibility probabilities p_{rep} , p_{srep} and p_{reprep} :

A – $n_1 = n_2 = 50$; **B** – $n_1 = n_2 = 150$

Например, можно ожидать, что при наблюдаемом $p_{\text{obs}} = 0,001$ значение $p_{\text{exp}} = 0,05$ будет получено с вероятностью 83 %.

Таблица 13. Вероятности воспроизведения наблюдаемых p -значений

Table 13. Probabilities of reproducing the observed p -values

p -значение:	p_{exp} – ожидаемое (желаемое)				
	0,05	0,01	0,005	0,001	10^{-4}
p_{obs} – наблюдаемое	$P_{\text{сrep}}$ – вероятность воспроизведения p_{obs}				
0,05	0,50	0,33	0,27	0,17	0,086
0,01	0,67	0,50	0,44	0,31	0,18
0,005	0,73	0,57	0,50	0,37	0,22
0,001	0,83	0,69	0,63	0,50	0,34
10^{-4}	0,91	0,82	0,78	0,66	0,50

Предсказательные интервалы

Предсказательные интервалы давно используются в регрессионном анализе, начинают использоваться в метаанализе. Но их редко вычисляют при оценках «обычных» параметров, в частности для p -значений.

Многочисленным приверженцам p -значений стоит напомнить, что вычисляемые значения суть, как принято говорить, одиночные реализации (выборочные точечные оценки) соответствующей случайной переменной $\tilde{p}$, поскольку оно основано на анализе случайных переменных. При отсутствии эффекта, т.е. когда разность параметров, корреляция, ассоциация, наклон линии регрессии равны нулю, или отношение параметров равно 1, или площадь под ROC-кривой равна 0,5 и т.п., величина $\tilde{p}$ имеет (стандартное) равномерное распределение на отрезке от 0 до 1. При отсутствии эффекта это распределение не зависит от объема выборки n , но при наличии эффекта величина $\tilde{p}$ становится («сложной») функцией как от n , так и от размера эффекта ES (разность средних MD , коэффициент корреляции и пр.), т.е. зависит от статистики используемого критерия проверки отсутствия эффекта.

Интуитивно наиболее понятной является бейзовская интерпретация p -значения, согласно которой половина p -значения ($p/2$) есть доля повторных экспериментов, в которых будет наблюдаться направление эффекта, противоположное наблюдаемому в первом эксперименте (и не более того). Или $(1 - p/2)$ есть всего лишь предсказательная вероятность получить эффект с тем же знаком (в том же направлении) при повторении эксперимента с бесконечно большим набором данных. Можно понять, что p -значение говорит

нам лишь о знаке эффекта, о его направлении, но не о размере эффекта. Когда эффект отсутствует, то при многократных повторениях эксперимента отклонения от ожидаемого (безразличного, неинформативного) значения будут происходить равновероятно в противоположных направлениях. В этом случае $p/2 = 0,5$.

Доверительные и предсказательные интервалы для p -значений вычисляет программа Statistical Intervals for P-value Evidence and Power (version 2.0) [26]. Как это выглядит для p -значений от 0,05 до 10^{-6} показано в табл. 14.

Таблица 14. Доверительные и предсказательные интервалы для p -значений

Table 14. Confidence and prediction intervals for p values

p -значение	Интервалы	
	Доверительный, ДИ	Предсказательный, ПИ
0,05	$9 \cdot 10^{-5}$; 1,00	$2 \cdot 10^{-6}$; 1,00
0,01	$6 \cdot 10^{-6}$; 0,54	$9 \cdot 10^{-8}$; 1,00
0,001	10^{-7} ; 0,18	10^{-9} ; 0,60
10^{-4}	$5 \cdot 10^{-9}$; 0,054	$3 \cdot 10^{-11}$; 0,26
10^{-5}	$2 \cdot 10^{-10}$; 0,014	$7 \cdot 10^{-13}$; 0,10
10^{-6}	$7 \cdot 10^{-12}$; 0,0034	10^{-14} ; 0,034

Из этих данных понятно, почему статистики настойчиво призывают отказаться от традиционных пороговых значений p : 0,05; 0,01 и даже 0,001.

Вот впечатляющие цитаты ведущих статистиков:

« p -значения подобны мосkitам (комарам). Они нашли свою экологическую нишу и никакими шлепками, расчесыванием и распылителями (спреями) от них не избавиться» [27];

«Мы создали монстра. И мы продолжаем подкармливать его, надеясь, что он перестанет пакостить. Это гиблое дело. Никакая плоть (и/или клеть) не сможет сдержать этого монстра. Единственный разумный путь впредь – покончить с ним» [28].

Предсказательный интервал для нового (случайного) одиночного наблюдения медикам известен как референсный интервал. Для вычислений можно рекомендовать программу: Reference Value Advisor¹⁹ [29].

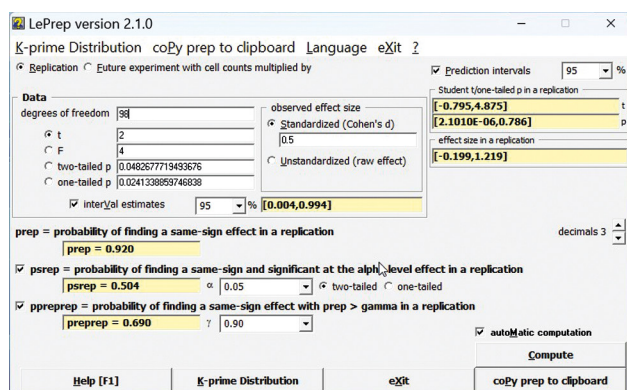
Предсказательные интервалы для разности средних и соответствующей стандартизированной величины (размер эффекта по Коуэну), а также для p -значений вычисляет программа LePrep (рис. 23).

¹⁹ Reference Value Advisor. Available at: <https://www.biostat.enscm.fr/reference-value-advisor/> (accessed: 27.01.2024).

Исходно статистически значимые различия ($p = 0.048$ и $p = 0.0029$) при повторении исследования в том же объеме ($n_1 = n_2 = 50$ и $n_1 = n_2 = 150$) могут оказаться статистически незначимыми (верхние границы 95%-х ПИ могут достигнуть значений 0,79 и 0,42, соответственно). С помощью этой программы можно спланировать, при каких объемах выборок результаты будут более надеж-

ными. Для второго примера увеличим объемы обеих выборок в 5 раз, получаем ДИ и ПИ, представленные на рис. 24.

При повторении исследования с такими объемами выборок с вероятностью 95 % гарантировано p не превысит значения 0,031. Можно также видеть, что в данном случае ПИ для d (от 0,56 до 3,45) приближается к ДИ (от 0,68 до 3,32).



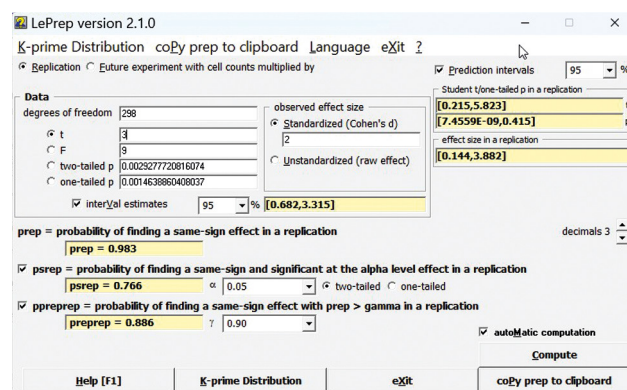
$$n_1 = n_2 = 50; t = 2$$

95%-й ДИ для d : 0,004; 0,994

95%-й ПИ для d : -0,20; 1,22

95%-й ПИ для p : $2 \cdot 10^{-6}$; 0,79

A



$$n_1 = n_2 = 150; t = 3$$

95%-й ДИ для d : 0,68; 3,32

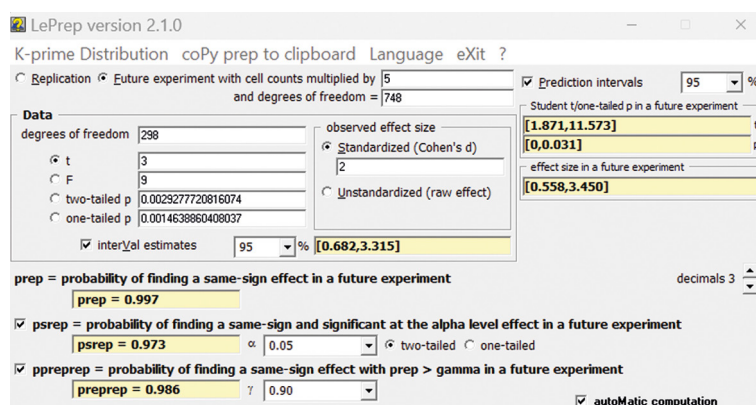
95%-й ПИ для d : 0,14; 3,88

95%-й ПИ для p : $7 \cdot 10^{-9}$; 0,42

B

Рис. 23. Сравнение доверительных (ДИ) и предсказательных (ПИ) интервалов для двух стандартизованных разностей средних $d = 0,5$ (A) и $d = 2,0$ (B) и соответствующих значений

Fig. 23. Comparison of confidence (CI) and predictive intervals (PI) for two standardized mean differences $d = 0.5$ (A) and $d = 2.0$ (B)



$$n_1 = n_2 = 750; t = 3; d = 2,0$$

95%-й ДИ для d : 0,68; 3,32

95%-й ПИ для d : 0,56; 3,45

95%-й ПИ для p : $< 10^{-16}$; 0,031

Рис. 24. Доверительный интервал для стандартизованного размера эффекта d и предсказательные интервалы для него и для p -значения после 5-кратного увеличения объема сравниваемых выборок

Fig. 24. Confidence interval for the standardized effect size d and prediction intervals for it and for p values after a 5-fold increase in the size of the compared samples

Отсюда следует интерпретация: ДИ есть интервал, до которого сужается ПИ после многократных повторений исследования с большими объемами выборок. Другими словами, ДИ является статистической оценкой гипотетического интервала, к которому сходится ПИ при бесконечном повторении исследования, или ДИ есть асимптотический ПИ. Таким образом, близость ДИ и ПИ может служить показателем достаточности объемов выборок.

СМ. Приведенная выше табл. 12 была составлена на компьютере для учебно-методического пособия [24]. Когда автор показал ее своей ученице и соавтору, она воскликнула: «Черт побери, Фишер это знал. Вот почему он предложил в качестве одного из пороговых значений $p = 0,05$ ». Предоставляю читателям самим оценить справедливость этого восклицания.

НХ. Совершенно непонятно, как к этому выводу пришла ученица-соавтор.
Вместо этой таблицы более информативным будет применение следующей программы: <https://replication.shinyapps.io/dvalue/> (см. рис. 25).
95%-е ПИ для четырех d -значений при некоторых объемах исходной и повторной выборок представлены в табл. 15.

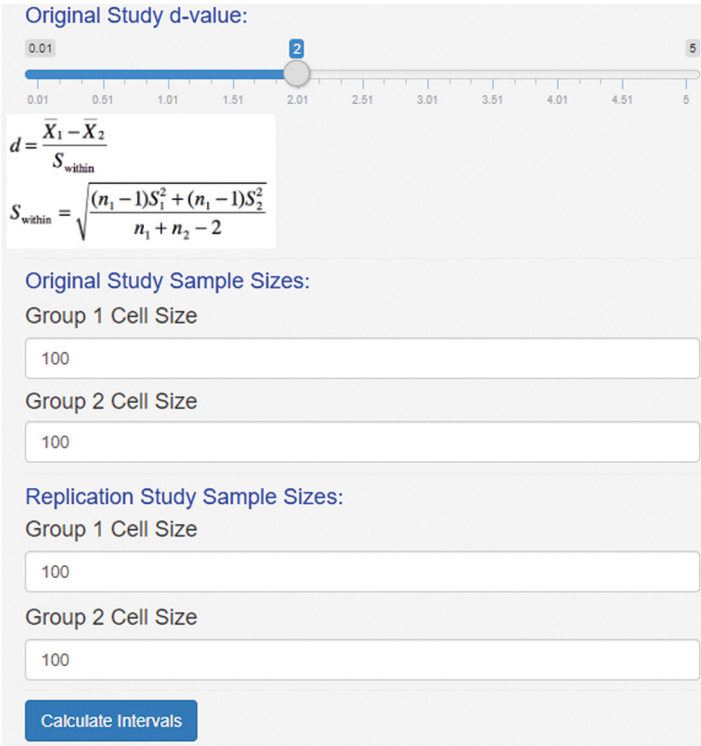


Рис. 25. Вычисление предсказательного интервала для стандартизированной разности средних (т.е. для d -значение)

Fig. 25. Calculation of the prediction interval for the standardized mean difference (i.e. for d -value)

Таблица 15. 95% предсказательные интервалы для четырех d -значений при некоторых объемах исходной и повторной выборок

Table 15. 95% prediction intervals for four d values at some sizes of the initial and repeated samples

Объемы выборок				95%е ПИ для d значений:			
Первых	n_{11}	n_{12}	Вторых, повторных n_{21} n_{22}	0,5	1,0	2,0	3,0
25	25	25	25	-0,30; 1,30	0,17; 1,83	1,03; 2,97	1,85; 4,15
25	25	50	50	-0,19; 1,19	0,28; 1,72	1,16; 2,83	2,00; 3,99
50	50	50	50	-0,05; 1,06	0,41; 1,59	1,32; 2,68	2,19; 3,81
100	100	50	50	0,01; 0,99	0,49; 1,51	1,41; 2,59	2,30; 3,70
100	100	100	100	0,10; 0,90	0,58; 1,42	1,52; 2,48	2,43; 0,89

Можно убедиться, что при умеренных объемах выборок значение стандартизированного размера эффекта $d = 0,5$ может оказаться статистически незначимым при повторении, и по этой причине его вряд ли можно признать практически важным. Наиболее надежными и воспроизводимыми будут различия с $d = 3,0$ и более. Тем самым мы возвращаемся к старому доброму «правилу трех сигм».

Исходное значение d = равно 2,00.
Объемы исходных выборок
Объемы повторных выборок
Предсказательный интервал составляет 95%
PI[1,52; 2,48].

Если в повторном исследовании значение d незначительно отличается от исходного значения d (в пределах допустимой случайной ошибки), то с вероятностью 95% оно попадет в этот интервал.

Если повторное значение d выходит за допустимые пределы, то это означает, что на разницу, вероятно, влияют дополнительные факторы.

Данный веб-калькулятор частично основан на работах:
Stanley D.J., Spence J.R. Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*. 2014;9:305–318.

Cumming G., Finch S. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*. 2001;61(4):532–574.

Cumming G., Maillardet S. Confidence intervals and replications: Where will the next mean fall? *Psychological Methods*. 2006;11(3):217–227.

Вычисление предсказательных интервалов для коэффициента корреляции реализовано в пакете `predictionInterval` для среды R. Например:

`pi.r(r=.74,n=46,rep.n=50)`

Результат:

Original study: $r = 0.74$, $N = 46$, 95% CI[0.57, 0.85]

Replication study: $N = 50$

Prediction interval: 95% PI[0.54, 0.93].

Метаанализ

СМ неоднократно утверждает:

«... сегодня все более популярными становятся метаанализы. Для проведения анализа требуется знание точного значения p » (S56).

«... обращает на себя внимание тот факт, что примененная программа не рассчитала точных значений p . Это резко снизило шансы данной работы попасть в метаанализ» (S57).

«**Необходимость указания точных значений p .** Помимо повышения вероятности попадания статьи в метаанализ автору приходит в голову еще один пример ...» (S58).

НХ. Считая это утверждение важным, автор (СМ) даже вынес его в резюме своей статьи: «Указание точного значения p , полученного в ходе исследования, повышает вероятность попадания опубликованной статьи в метаанализ» (S52).

Действительно, известны попытки использовать в метаанализе p -значения в качестве показателя неоднородности данных из различных источников, однако ведущие специалисты признали такой подход ошибкой [30]²⁰.

Наиболее популярной и многоцелевой некоммерческой программой для метаанализа является *Meta-Essentials*²¹ [31]. Кроме универсальности, важной ее особенностью является вычисление предсказательных интервалов для суммарных эффектов. Для вычисления предсказательных интервалов в метаанализе существуют также самостоятельные программы: *Prediction Interval Calculator for Random effects meta-analysis*²² и *СМАРП*²³ [32].

²⁰ Книга свободно доступна по запросу на сайте: <https://meta-analysis.com/#downloadBookAndPaper> (дата обращения: 20.10.2023).

²¹ Meta-Essentials: workbooks for meta-analysis. Available at: <https://www.erim.eur.nl/research-support/meta-essentials/> (accessed: 27.01.2024).

²² Prediction Interval Calculator for Random effects meta-analysis. Available at: <https://medstats.github.io/ranefpredict.html> (accessed: 27.01.2024).

²³ Prediction Intervals Program Download. Available at: <https://www.meta-analysis-workshops.com/pages/predictionintervalsdownload.php?&email=nikita.khromovborisov%40gmail.com&valid=s306314> (accessed: 27.01.2024).

Заключение

СМ уверяет, что якобы:

В большинстве случаев ситуация в проводимом эксперименте выглядит следующим образом:

1. Наблюдаемое явление: 18 августа 1913 г. в казино Монте-Карло шарик 28 раз подряд угодил на «черное»!

2. Гипотеза: что-то с колесом 18 августа 1913 г.

НХ. Согласно Википедии и другим источникам, в тот день в Монте-Карло черное выпало 26, а не 28 раз подряд, в связи с чем игроки ставили на красное, надеясь, что последовательность выпадения черного прервется, и проигрывали. Эту историю часто приводят исследователи, занимающиеся психологией азартных игр.

СМ предлагает:

– проводим точечное и интервальное оценивание параметров;

– получаем решение в виде одного или нескольких значений вероятности.

Чем больше таких вероятностей вы представите читателям, тем убедительнее для них будут ваши выводы.

НХ. В данном случае параметр всего один – частота (вероятность) выпадения черного, точечная бейсова оценка ее медианы $\phi = 0,97$ с 95%-м ДИ от 0,87 до 1,00, с 99%-м ДИ от 0,82 до 1,00 и с 99,9%-м ДИ от 0,75 до 1,00. Они идентичны ДИ по Клопперу-Пирсону.

Вероятность тут тоже всего одна – p -значение, которое равно $p = 3 \cdot 10^{-8}$ с предсказательным интервалом от 10^{-16} до 0,0056. Так что, скорее всего, в этот день колесо рулетки было неисправно. Либо произошло редчайшее событие, о чем предупреждал Р.А. Фишер: «Никакой отдельно взятый эксперимент, как бы значим он ни был сам по себе, не может считаться достаточным для доказательства какого-либо реального факта – “один шанс на миллион” все равно будет обязательно осуществляться не реже и не чаще, чем ему положено, как бы мы ни удивлялись, произойди такое с нами. Чтобы утверждать, что какое-либо явление природы экспериментально проверяемо, необходимо не изолированное наблюдение, необходимый надежный метод экспериментирования» [33].

СМ заключает:

Чем больше таких вероятностей вы представите читателям, тем убедительнее для них будут ваши выводы.

НХ. В действительности исследователям нужны не столько разного рода вероятности, сколько интервальные (доверительные и предсказательные) оценки тех или иных показателей (параметров, мер эффекта).

«Основным продуктом научного исследования является один или несколько показателей размера эффекта, а не p -значения» (Cohen) [34²⁴].

²⁴ Статья цитируется около 3500 раз.

Как уже было сказано, статистический анализ научных данных является самостоятельным кропотливым научным исследованием. Он должен быть комплексным, гармонизирующим статистические описания, оценки, сравнения и предсказания. Поскольку зачастую разные пакеты статистических программ используют разные алгоритмы, то желательно для одной и той же задачи использовать несколько программ и проверять согласованность получаемых результатов (табл. 16).

Таблица 16. Описание использованных статистических методов и программ*

Table 16. Description of the statistical methods and programs used

Программа	Версия и/или дата	Использованные процедуры и методы, и их назначение	Ссылка
LePAC	2.20.062 08.06.2020	Точные байесовские (credible) доверительные интервалы (БДИ) для долей, их разностей и отношений, и отношений шансов, и их сравнение с результатами 32 других методов	https://eris62.eu/ErisLePAC.html
LePrep	2.1.0 01.04.2011	Предсказательные вероятности p -гер, p -стер, p -прегрер и предсказательные интервалы (ПИ) для разности средних, для размера эффекта и для p -значений	
Statistical Intervals for Pvalue Evidence and Power	2.0 2016	Доверительные (ДИ) и предсказательные интервалы (ПИ) для p -значений в последующих повторных исследованиях	[26]
StatXact	12 2023	Точные ДИ для параметров полиномиального распределения и коэффициента сопряженности Крамера, V	https://www.cytel.com/software/statxact/
resize	2021	Вычисление объема выборки для оценивания доли с заданной шириной ДИ	[9]
PAST	4.14 10.2023	Интервальные оценки и сравнения на основе алгоритмов бутстрепа и Монте-Карло. Гистограммы и вероятностные графики для Н.Р. 4 критерия проверки согласия данных с Н.Р. Корреляционные эллипсы и регрессионные линии с доверительными и предсказательными зонами	[10]
BoxPlotR	17.01.2021	Коробки с декольте и с усами	[13]
Estimation Statistics	2017-2021	Анализ данных с помощью размеров эффекта. Графики Гарднера-Олтмена	[14]
JASP	0.18.1 29.09.2023	ДИ для размера эффекта и коэффициентов корреляции. Байесовы факторы, BF . Графики «дождевые облака»	[16]
P-Value Calculator for F Distribution	2015	Вычисление квантилей и/или p -значений для четырех распределений: нормального, t , F и хи-квадрат	http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pt.html
Keisan		Библиотека сверхточных вычислений, в частности, 20 распределений	https://keisan.casio.jp/exec/system/1161228871
Reference Value Advisor	2.1 09.12.2015	Референсные интервалы. Шесть методов	[29]
Effect Size Calculator	2017	Интервальные оценки различных мер эффекта (OR , NNT , ω^2 , η^2 , R^2 , ICC и др.)	https://effect-size-calculator.herokuapp.com/
Statistical tests	2020	Калькулятор p -значений и значений $mid-p$ для точного критерия Фишера для таблиц 2×2 и 2×3	https://www.cog-genomics.org/software/stats
Prediction Interval	–	Предсказательный интервал для стандартизированной разности средних (d -значение)	https://replication.shinyapps.io/dvalue/
Meta-Essentials	1.15 2023	Многоцелевой метаанализ с вычислением предсказательных интервалов	[31]
CMA Prediction Intervals			[32]
Prediction Interval Calculator for Random effects meta-analysis		Предсказательные интервалы в метаанализе	https://medstats.github.io/ranefpredict.html

* Использована предложенная ранее табличная форма описания использованных программ в научных публикациях [35].

ФИНАНСИРОВАНИЕ

Работа выполнена в порядке личной инициативы.

FUNDING

The work was done on a personal initiative.

КОНФЛИКТ ИНТЕРЕСОВ

Автор заявляет об отсутствии потенциального конфликта интересов.

CONFLICT OF INTERESTS

The author declares no conflict of interests.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Cumming G., Calin-Jageman R. *Introduction to the New Statistics Estimation, Open Science, and Beyond*. 2nd ed. Routledge; 2024. 610 p. <https://doi.org/10.4324/9781032689470>
2. Hopkins W. *A new view of statistics*. 2014. Available at: <https://complementarytraining.net/wp-content/uploads/2013/10/Will-Hopkins-A-New-View-of-Statistics.pdf> (accessed: 27.01.2024).
3. Мыльников С.В. Доверительные интервалы и точное значение p в научной статье в медико-биологическом журнале. *Научный редактор и издатель*. 2023;8(1 Suppl):S52–S61. <https://doi.org/10.24069/SEP-23-13>
Mylnikov S.V. Confidence intervals and precise p -value in a scientific paper in a biomedical journal. *Science Editor and Publisher*. 2023;8(1 Suppl):S52–S61. (In Russ.) <https://doi.org/10.24069/SEP-23-13>
4. Halverson M.S., Bolnick D.A. An ancient DNA test of a founder effect in Native American ABO blood group frequencies. *American Journal of Physical Anthropology*. 2008;137(3):342–347. <https://doi.org/10.1002/ajpa.20887>
5. Clopper C.J., Pearson E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26(4):404–413. <https://doi.org/10.1093/biomet/26.4.404>
6. Глотов Н.В., Животовский Л.А., Хованов Н.В., Хромов-Борисов Н.Н. *Биометрия*. Л.: Изд-во ЛГУ; 1982. 263 с.
7. Khurshid A., Ageel M.I. Binomial and Poisson confidence intervals and its variants: A bibliography. *Pakistan Journal of Statistics and Operation Research*. 2010;6(1):75–100. <https://doi.org/10.18187/pjsor.v6i1.139>
8. Мыльников С.В. *Азы биометрии*. СПб.: Издательство Н-Л; 2007. 56 с.
9. Haynes A.G., Lenz A., Stalder O., Limacher A. presize: An R-package for precision-based sample size calculation in clinical research. *Journal of Open Source Software*. 2021;6(60):3118. <https://doi.org/10.21105/joss.03118>
10. Hammer Ø., Harper D.A.T., Ryan P.D. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*. 2001;4(1):4. Available at: http://palaeo-electronica.org/2001_1/past/issue1_01.htm (accessed: 27.01.2024).
11. Закс Л. *Статистическое оценивание* [Ю.П. Адлер, В.Г. Горкий (ред.); В.Н. Варыгин (пер.)]. М.: Статистика; 1976. 598 с.
12. Sachs L. *Applied Statistics. A Handbook of Techniques*. 2nd ed. New York: Springer; 1984. 707 p.
13. Spitzer M., Wildenhain J., Rappsilber J., Tyers M. BoxPlotR: a web tool for generation of box plots. *Nature Methods*. 2014;11(2):121–122. <https://doi.org/10.1038/nmeth.2811>
14. Ho J., Tumkaya T., Aryal S., Choi H., Claridge-Chang A. Moving beyond P values: Data analysis with estimation graphics. *Nature Methods*. 2019;16(7):565–566. <https://doi.org/10.1038/s41592-019-0470-3>

15. Allen M., Poggiali D., Whitaker K., Marshall T.R., van Langen J., Kievit R.A. Raincloud plots: a multi-platform tool for robust data visualization [version 2; peer review: 2 approved] *Wellcome Open Research*. 2021;4:63. <https://doi.org/10.12688/wellcomeopenres.15191.2>
16. JASP Team 2023. JASP (version 0.18.1). [Computes software]
17. Louis T.A., Zeger S.L. Effective communication of standard errors and confidence intervals. *Biostatistics*. 2009;10(1):1–2. <https://doi.org/10.1093/biostatistics/kxn014>
18. Мыльников С.В., Павлова Н.В., Барабанова Л.В. Антимутагенные эффекты синтетического пептида эпیتالона у мышей, *Mus musculus*, различающихся по окраске шерсти. *Экологическая генетика*. 2011;9(3):75–88. <https://doi.org/10.17816/ecogen9375-88>
Mylnikov S.V., Pavlova N.V., Barabanova L.V. Anti-mutagenic action of synthetic “epitalon” peptide in mice with different body colour. *Ecological Genetics*. 2011;9(3):75–88. (In Russ.) <https://doi.org/10.17816/ecogen9375-88>
19. Глотов Н.В., Филатов А.А., Хромов-Борисов Н.Н. *Сборник задач по биометрии*. Л.: Изд-во ЛГУ; 1985. 81 с.
20. Мыльников С.В. Зависимость дисперсии выживаемости от возраста когорты у *Drosophila melanogaster*. Сообщение 1: Норма реакции выживаемости и ее наследование. *Успехи геронтологии*. 2018;31(3):323–329.
Mylnikov S.V. Regression of survivability variance on the cohort age in *Drosophila melanogaster*. I. survivability: reaction norm and its inheritance. *Advances in Gerontology*. 2018;8(4):277–283. <https://doi.org/10.1134/S2079057018040112>
21. Sullivan G.M., Feinn R.S. Facts and fictions about handling multiple comparisons. *Journal of Graduate Medical Education*. 2021;13(4):457–460. <https://doi.org/10.4300/JGME-D-21-00599.1>
22. Bonett D.G., Wright T.A. Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. *Psychometrika*. 2000;65(1):23–28. <https://doi.org/10.1007/BF02294183>
23. Gnambs T. A brief note on the standard error of the Pearson correlation. *Collabra: Psychology*. 2023;9(1):87615. <https://doi.org/10.1525/collabra.87615>
24. Федорова Е.М., Мыльников С.В. *Нелегкий путь к диплому*. СПб.: Изд-во Н-Л; 2012. 100 с.
25. Killeen P.R. An alternative to null hypothesis significance tests. *Psychological Science*. 2005;16(5):345–353. <https://doi.org/10.1111/j.0956-7976.2005.01538.x>
26. Lazzeroni L.C., Lu Y., Belitskaya-Lévy I. Solutions for quantifying P-value uncertainty and replication power. *Nature Methods*. 2016;13(2):107–108. <https://doi.org/10.1038/nmeth.3741>
27. Berry D. A p-value to die for. *Journal of the American Statistical Association*. 2017;112(519):895–897. <https://doi.org/10.1080/01621459.2017.1316279>
28. Campbell J.P. Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology*. 1982;67(6):691–700. <https://doi.org/10.1037/h0077946>
29. Geffré A., Concordet D., Braun J.P., Trumel C. Reference Value Advisor: A new freeware set of macroinstructions to calculate reference intervals with Microsoft Excel. *Veterinary Clinical Pathology*. 2011;40(1):107–112. <https://doi.org/10.1111/j.1939-165X.2011.00287.x>
30. Bornstein M. *Common mistakes in meta-analysis and how to avoid them*. Biostat, Inc. 2019. 409 p. Available at: <https://meta-analysis.com/#downloadBookAndPaper> (accessed: 27.01.2024).
31. Suurmond R., van Rhee H., Hak T. Introduction, comparison, and validation of *Meta-Essentials*: A free and simple tool for meta-analysis. *Research Synthesis Methods*. 2017;8(4):537–553. <https://doi.org/10.1002/jrsm.1260>
32. Borenstein M., Higgins J.P., Hedges L.V., Rothstein H.R. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*. 2017;8(1):5–18. <https://doi.org/10.1002/jrsm.1230>
33. Фишер Р.А. Математика дамы, дегустирующей чай. В кн.: *Современные проблемы математики: сб. ст.* М.: Знание; 1981. С. 46–58.
34. Cohen J. Things I have learned (so far). *American Psychologist*. 1990;45(12):1304–1312. <https://doi.org/10.1037/0003-066X.45.12.1304>

35. Хромов-Борисов Н. Н. Табличная форма описания статистических методов и программ в научных публикациях. *Научный редактор и издатель*. 2022;7(2)6:182–184. <https://doi.org/10.24069/SEP-22-40>
Khromov-Borisov N.N. Tabular form of description of statistical methods and programs in scientific publications. *Science Editor and Publisher*. 2022;7(2):182–184. (In Russ.) <https://doi.org/10.24069/SEP-22-40>

ИНФОРМАЦИЯ ОБ АВТОРЕ

Никита Николаевич Хромов-Борисов, кандидат биологических наук, старший научный сотрудник, член Комиссии по борьбе с лженаукой Экспертного совета РАН, г. Санкт-Петербург, Российская Федерация; <https://orcid.org/0000-0001-6435-7218>; e-mail: Nikita.KhromovBorisov@gmail.com

INFORMATION ABOUT THE AUTHOR

Nikita N. Khromov-Borisov, Cand. Sci. (Biol.), member of the Commission for Combating Pseudoscience of the Expert Council of the Russian Academy of Sciences, St. Petersburg, Russian Federation; <https://orcid.org/0000-0001-6435-7218>; e-mail: Nikita.KhromovBorisov@gmail.com

Поступила в редакцию / Received 12.12.2023

Поступила после рецензирования / Revised 06.02.2024

Принята к публикации / Accepted 15.03.2024