УДК 311
http://dx.doi.org/10.22328/2413-5747-2020-6-1-101-106

ОБЪЕМ ВЫБОРКИ ДЛЯ КОРРЕЛЯЦИОННОГО АНАЛИЗА

[1,2]А. М. Гржибовский[*], [1]М. А. Горбатова, [3]А. Н. Наркевич, [3]К. А. Виноградов
[1]Северный государственный медицинский университет, г. Архангельск, Россия
[2]Северо-Восточный федеральный университет имени М. К. Аммосова, г. Якутск, Россия
[3]Красноярский государственный медицинский университет имени профессора В. Ф. Войно-Ясенецкого, г. Красноярск, Россия

В отечественных биомедицинских исследованиях недопустимо редко рассчитываются необходимые объемы выборки при планировании научных проектов, что часто приводит к ситуации, когда статистической мощности недостаточно для ответа на поставленные задачи. Это встречается как при оценке параметров, так и для проверки статистических гипотез. Кроме того, даже в тех случаях, когда расчеты проводятся, они осуществляются с помощью упрощенных формул, подразумевающих нормальное распределение признака, что не всегда корректно в биомедицинских исследованиях. Корреляционный анализ является одним из наиболее часто встречающихся видов анализа в отечественной биомедицинской литературе для оценки связи между двумя признаками, причем параметрический коэффициент Пирсона, несмотря на известные ограничения, встречается в литературе гораздо чаще его непараметрических альтернатив. Мы рассчитали минимальные размеры выборки, необходимые для применения коэффициента корреляции Пирсона и непараметрических коэффициентов Спирмена и Кендалла, что позволит начинающим исследователям оценить необходимый объем в зависимости от поставленной задачи, ожидаемой величины связи, типа и распределения данных. Результаты расчетов представлены в виде готовых для практического использования таблиц с минимально необходимым размеров выборки для получения статистически значимых коэффициентов корреляции от 0,10 до 0,90 с шагом 0,05 для мощности 0,8 и 0,9 на уровне альфа-ошибки 5%, а также для их определения с точностью на уровне ширины доверительного интервала 0,1 и 0,2 при доверительной вероятности 95%.

Ключевые слова: морская медицина, размер выборки, доверительный интервал, корреляционный анализ, коэффициент Пирсона, коэффициент Спирмена, коэффициент Кендалла

Конфликт интересов: авторы заявили об отсутствии конфликта интересов.

Для цитирования: Гржибовский А.М., Горбатова М.А., Наркевич А.Н., Виноградов К.А. Объем выборки для корреляционного анализа // *Морская медицина.* 2020. Т. 6, № 1. С. 101–106, http://dx.doi.org/10.22328/2413-5747-2020-6-1-101-106.

Контакт: *Гржибовский Андрей Мечиславович, Andrej.Grjibovski@gmail.com*

REQUIRED SAMPLE SIZE FOR CORRELATION ANALYSIS

[1,2]Andrej M. Grjibovski[*], [1]Maria A. Gorbatova, [3]Artem N. Narkevich, [3]Konstantin A. Vinogradov
[1]Northern State Medical University, Arkhangelsk, Russia
[2]North-Eastern Federal University named after M.K. Ammosov, Yakutsk, Russia
[3]Professor V. F. Voyno-Yasenetsky Krasnoyarsk State Medical University, Krasnoyarsk, Russia

Sample size calculation prior to data collection is still relatively rare in Russian research practice. This situation threatens validity of the conclusion of many projects due to insufficient statistical power to estimate the parameters of interest with desired precision or to detect the differences of interest. Moreover, in a substantial proportion of cases where sample size calculations are performed simplified formulas with assumption of a normal distribution of the studied vari-

ables are used in spite of the fact that this assumption does not hold for many research questions in biomedical research. Correlation analysis is still one of the most commonly used methods of statistical analysis used in Russia. Pearson's correlation coefficient despite its well-known limitations appears in a greater proportion of publications that non-parametric coefficients. We calculated minimal sample sizes for the parametric Pearson's coefficient as well its non-parametric alternatives — Spearman's rho and Kendall's tau-b correlation coefficients to assist junior researchers with the tool to be able to plan data collection and analysis for several types of data, various expected strengths of associations and research questions. The results are presented in ready-for-use tables with required sample size for the three abovementioned coefficients within the range from 0,10 through 0,90 by 0,05 for statistical power 0,8 and 0,9 and alpha-error or 5% as well as for estimation of the same correlation coefficients with the 95% confidence intervals width equal to 0,1 and 0,2.
Key words: marine medicine, Sample size, statistical power, correlation analysis, Pearson's coefficient, Spearman's coefficient, Kendall's coefficient

Conflict of interest: the authors stated that there is no potential conflict of interest.
For citation: Grjibovski A.M., Gorbatova M.A., Narkevich A.N., Vinogradov K.A. Required sample size for correlation analysis // *Marine medicine*. 2020. Vol. 6, No. 1. P. 101–106. http://dx.doi.org/10.22328/2413-5747-2020-6-1-101-106.
Contact: *Grzhibovskiy Andrey, Andrej.Grjibovski@gmail.com*

Корреляционный анализ, несмотря на ограниченность его возможностей, по-прежнему остается одним из наиболее популярных статистических методов в отечественной биомедицинской литературе. В наших более ранних публикациях мы детально описывали пошаговое выполнение корреляционного анализа с помощью пакета статистических программ SPSS [1, с. 50–60], причем, поскольку анализ сам по себе достаточно несложен и обычно не вызывает затруднений, мы уделяли большое внимание ограничениям корреляционного анализа, а также призывали к осторожности при интерпретации коэффициентов корреляции. Также описание корреляционного анализа является неотъемлемой частью большинства учебных пособий по основам биостатистики. Тем не менее попрежнему продолжают публиковаться работы, где коэффициенты корреляции (особенно коэффициент Пирсона) применяются без проверки необходимых условий, а также декларируется наличие причинно-следственных связей при получении статистически значимых коэффициентов. Также нам встречались работы, где далеко идущие выводы делались на основании корреляционного анализа, проведенного на крайне ограниченном числе наблюдений. Поскольку нашей задачей не было выявление статей, содержащих такие ошибки или авторов, делающих такие ошибочные выводы, мы намеренно не приводим ссылки на такие публикации.

К нам часто обращаются начинающие исследователи с вопросами относительно необходимого размера выборки для проведения того или иного исследования, в том числе и для экологических (корреляционных) исследований, в которых единицей наблюдения выступает группа людей, часто объединенная по территориальному признаку, например, район, область или страна [2, с. 57–64]. Экологические (корреляционные) исследования требуют минимум времени, поскольку для них часто используются данные официальной статистики, поэтому они очень популярны среди начинающих исследователей. Однако ни ретроспективных расчетов статистической мощности, ни определения необходимого размера выборки для проведения экологических (корреляционных) исследований, нам за последние годы не встретилось, что, впрочем, не означает, что их нет, просто, вероятно, редакторы журналов не требуют таких расчетов. Несмотря на то, что существует большое количество бесплатных онлайн-калькуляторов и опций бесплатного программного обеспечения для расчета необходимого размера выборки для корреляционного анализа, эти возможности крайне редко используются отечественными исследователями.

Для того чтобы помочь начинающим исследователям оценить необходимый объем выборки при планировании экологических (корреляционных) исследований, а также для изучения силы и направления связи между двумя переменными с помощью корреляционного анализа, мы рассчитали и представили в таблицах минимальные размеры выборки, необходимые для применения коэффициента корреляции Пирсона и непараметрических коэффициентов Спирмена и Кендалла.

Расчет минимальных размеров выборки производили отдельно для коэффициентов корреляции r Пирсона, rho Спирмена и tau-b Кендалла, так как первый предназначен только для определения линейной связи между двумя непрерывными переменными при нормальном распределении данных, а второй и третий можно применять как для количественных переменных, имеющих распределение, отличающееся от нормального, так и для порядковых (ранговых) переменных. Результаты расчетов представлены в виде готовых для практического использования таблиц с минимально необходимым размеров выборки. На первом этапе рассчитывали размер выборки для проверки статистической гипотезы об отличии коэффициентов корреляции от 0,10 до 0,90 с шагом 0,05 для мощности 0,8 (80%) и 0,9 (90%) на уровне альфа-ошибки 5%[1] [3, с. 45–48; 5]. Поскольку, начиная с середины 1980-х годов международным статистическим сообществом активно продвигается идея о необходимости интервальной оценки параметров и соответствующего представления результатов [5, с. 746–750; 6, с. 305–307], которая, правда, пока не является доминирующей в русскоязычной научной печати, мы также рассчитали размер выборки необходимый для определения коэффициентов корреляции от 0,10 до 0,90 с шагом 0,05 для ширины доверительного интервала 0,1 и 0,2 на уровне доверительной вероятности 95% [7, с. 23–28; 8, с. 240–245; 9]. Все расчеты проводили с помощью программы PASS-2019 (NCSS, LLC. Kaysville, Utah, USA).

Результаты расчетов минимального размера выборки, достаточного для того, чтобы можно было отклонить нулевую гипотезу о равенстве коэффициентов корреляции нулю для различных альтернативных гипотез на наиболее часто встречающихся уровнях статистической мощности 80 и 90% при уровне доверительной вероятности 95% представлены в табл. 1. Проще говоря, представленные размеры выборок позволяют выявить статистически значимые коэффициенты корреляции.

Таблица 1

Минимальный размер выборки (N) для выявления статистически значимых коэффициентов корреляции r Пирсона, rho Спирмена и tau-b Кендалла для достижения статистической мощности 0,8 и 0,9 на уровне доверительной вероятности 95%

Table 1

Minimum sample size (N) to identify statistically significant correlation coefficients R Pearson, Rho Spearman and tau-b Kendall to achieve statistical power of 0,8 and 0,9 at the confidence level of 95%

Значение коэффициента корреляции	N для коэффициента корреляции Пирсона при статистической мощности		N для коэффициента корреляции Спирмена при статистической мощности		N для коэффициента корреляции Кендалла при статистической мощности	
	0,80	0,90	0,80	0,90	0,80	0,90
0,10	736	994	837	1093	586	1660
0,15	359	438	446	519	393	490
0,20	190	202	191	256	198	313
0,25	123	199	137	172	137	194
0,30	90	118	85	132	97	138
0,35	62	82	68	78	71	78
0,40	44	59	53	62	55	68
0,45	37	47	39	54	37	57
0,50	30	38	32	44	33	43
0,55	23	33	29	34	27	34
0,60	19	25	20	30	22	30
0,65	16	21	20	22	18	25
0,70	13	17	15	20	15	20
0,75	11	14	13	16	14	17
0,80	9	11	11	14	12	14
0,85	8	9	9	11	10	12
0,90	6	7	8	9	9	10

[1] Kendall M., Gibbons J.D. Rank Correlation Methods. 5th ed. New York: Oxford University Press, 1990.

Результаты расчета минимального размера выборки, достаточного для определения коэффициентов корреляции с заданной точностью (шириной 95% доверительных интервалов), представлены в табл. 2.

Таблица 2

Минимальный размер выборки для расчета коэффициентов корреляции r Пирсона, rho Спирмена и tau-b Кендалла для ширины 95%-доверительного интервала (ДИ) 0,1 и 0,2

Table 2

The minimum sample size for calculating the correlation coefficients R Pearson, Rho Spearman and tau-b Kendall for the width of the 95% confidence interval (CI) 0,1 and 0,2

Значение коэффициента корреляции	Ширина доверительного интервала для коэффициента корреляции Пирсона		Ширина доверительного интервала для коэффициента корреляции Спирмена		Ширина доверительного интервала для коэффициента корреляции Кендалла	
	0,10	0,20	0,10	0,20	0,10	0,20
0,10	1507	378	1515	379	662	168
0,15	1469	368	1486	372	645	164
0,20	1417	355	1446	362	622	158
0,25	1352	339	1394	350	594	151
0,30	1274	320	1331	334	560	143
0,35	1185	298	1257	316	521	133
0,40	1086	273	1173	295	478	122
0,45	980	247	1079	271	431	111
0,50	867	219	975	246	382	99
0,55	751	190	864	218	331	86
0,60	633	161	746	189	280	73
0,65	517	132	625	159	229	61
0,70	404	105	503	129	180	49
0,75	299	79	383	100	134	37
0,80	205	56	269	72	93	27
0,85	125	36	169	47	57	19
0,90	62	20	86	27	30	12

Результаты расчетов показывают, что для определения небольших коэффициентов корреляции требуется значительно более объемные выборки, чем для больших коэффициентов, причем коэффициент корреляции Пирсона требует меньшего количества наблюдений (обладает большей чувствительностью), чем непараметрические коэффициенты Спирмена и Кендалла. Учитывая, что до начала исследования мы можем лишь предполагать, какой коэффициент корреляции будет получен при анализе собранных данных, то целесообразно определиться, какой коэффициент будет считаться достаточным для того, чтобы считать корреляционную связь важной. Например, во многих учебных пособиях отмечается, что корреляционную связь можно считать сильной при значении коэффициента корреляции от 0,7 и выше, средней — от 0,3 до 0,7 и т.д. Из табл. 1 следует, что для обнаружения сильной корреляционной связи между переменными, которые являются количественными и имеют нормальное распределение, необходимо минимум 13 наблюдений для статистической мощности 80% и 17 наблюдений для статистической мощности 90%. В то же время для обнаружения статистически значимой корреляционной связи средней силы необходимо не менее 90 и 118 наблюдений для уровней статистической мощности 80 и 90% соответственно, при соблюдении условий, необходимых для применения коэффициента корреляции Пирсона. Если у нас нет уверенности в том, что эти условия выполняются, то оценку связи между переменными лучше проводить с помощью коэффициента корреляции Спирмена, а для этого необходимо уже 85 наблюдений для достижения мощности 80% и 132 наблюдения для статистической мощности 90%.

Таким образом, проведение корреляционных исследований в пределах одной области, в которой 10–12 районов, вряд ли можно считать целесообразным, особенно если требуются финансовые затраты, поскольку такой размер вы-

борки не дает возможности обнаружить даже сильную корреляционную связь, не говоря уже о связи средней силы. В этом случае необходимо рассматривать возможность оценки данных на уровне муниципальных образований. Приведенными таблицами также можно пользоваться при проведении корреляционного анализа для оценки связи между признаками, измеренными на индивидуальном уровне.

Для интервальной оценки коэффициентов корреляции с заданной шириной 95% доверительного интервала можно воспользоваться таблицей 2. Например, для того, чтобы обеспечить точность коэффициента корреляции Пирсона 0,5 на уровне ширины доверительного интервала 0,1 нам необходимо 867 наблюдений! В дополнение к таблицам мы построили графики для всех трех коэффициентов корреляции, демонстрирующие связь между размером выборки и точностью определения коэффициентов, из которых видно, что чем точнее мы хотим определить коэффициент корреляции, тем больший объем выборки нам необходим (рис. 1–3).

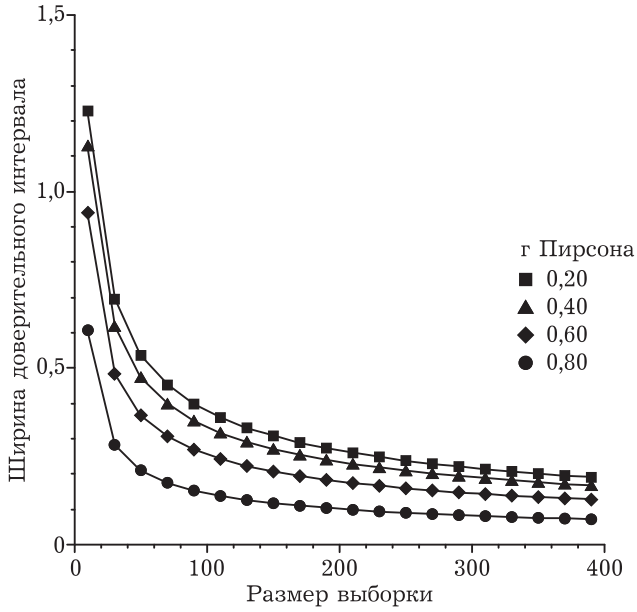

Рис. 1. Связь между точностью оценки коэффициента корреляции Пирсона и размером выборки

Fig. 1. Relationship between the accuracy of the Pearson correlation coefficient estimate and the sample size

Мы надеемся, что наши расчеты помогут начинающим исследователям в планировании исследований, в которых основной мерой оценки связи между признаками являются коэффициенты корреляции. Следует помнить: если нет

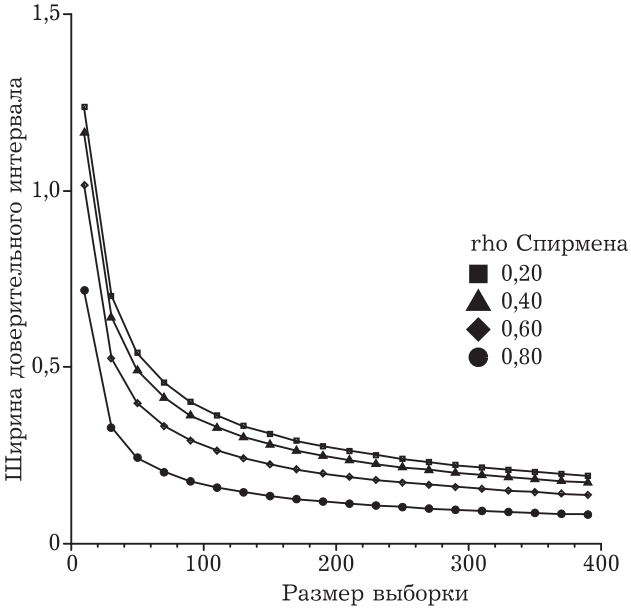

Рис. 2. Связь между точностью оценки коэффициента корреляции (rho) Спирмена и размером выборки

Fig. 2. Relationship between the accuracy of Spearman's correlation coefficient (rho) estimate and sample size

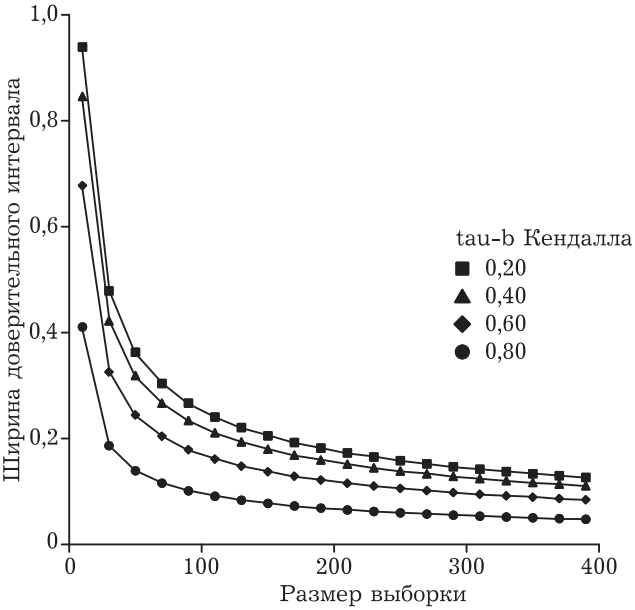

Рис. 3. Связь между точностью оценки коэффициента корреляции tau-b Кендалла и размером выборки

Fig. 3. Relationship between the accuracy of the Kendall Tau-b correlation coefficient estimation and the sample size

уверенности в том, что будут выполняться условия для применения коэффициента корреляции Пирсона, целесообразно изначально при планировании исследования ориентироваться

на непараметрические коэффициенты. Также всегда надо помнить о вероятности наличия пропущенных значений по причине несобранных данных или данных с низкой валидностью по тем или иным причинам. В исследованиях, где данные собираются на индивидуальном уровне, делается поправка на отказы от участия в исследовании. Например, если наши расчеты показывают, что для проведения анализа необходимо 132 человека, но мы предполагаем, что каждый пятый откажется от участия в исследовании, то пригласить в исследование надо на 20% больше человек, то есть минимум 159. Кроме того, следует учитывать, что наши расчеты показывают минимальный размер выборки для проведения корреляционного анализа, который является довольно чувствительным инструментом, то есть если планируется какой-то менее чувствительный анализ (например, критерий χ^2 Пирсона), то и размер выборки считается для наименее чувствительного критерия. О расчетах минимального необходимого размера выборки для других часто применяющихся статистических критериев мы расскажем в последующих выпусках журнала «Морская медицина».

Литература/References

1. Гржибовский А.М. Корреляционный анализ // *Экология человека.* 2008. № 9. С. 50–60. [Grjibovski A.M. Correlation analysis. *Human Ecology,* 2008, No. 9, pp. 50–60 (In Russ.)].
2. Холматова К.К., Горбатова М.А., Гржибовский А.М. Применение экологических исследований в медицине и общественном здравоохранении // *Экология человека.* 2016. № 9. С. 57–64. [Kholmatova K.K., Grjibovski A.M. Ecological Studies in Medicine and Public Health. *Human Ecology,* 2016, No. 9, pp. 57–64 (In Russ.)].
3. Guenther W.C. Desk Calculation of Probabilities for the Distribution of the Sample Correlation Coefficient // *The American Statistician.* 1977. Vol. 31 (1). P. 45–48.
4. Devroye L. *Non-Uniform Random Variate Generation.* New York: Springer-Verlag, 1986.
5. Gardner M.J., Altman D.G. Confidence intervals rather than P values: estimation rather than hypothesis testing // *Brit. Med. J. (Clin. Res. Ed.).* 1986. Mar. 15; Vol. 292 (6522). P. 746–750.
6. Amrhein V., Greenland S., McShane B. Scientists rise up against statistical significance // *Nature.* 2019. Vol. 567 (7748). P. 305–307.
7. Bonett D.G., Wright T.A. Sample Size Requirements for Estimating Pearson, Kendall and Spearman Correlations // *Psychometrika.* 2000. Vol. 65 (1). P. 23–28.
8. Looney S.W. Sample size determination for correlation coefficient inference: Practical problems and practical solutions // *American Statistical Association.* 1996. Proceedings of the Section on Statistical Education. P. 240–245.
9. Cook R.D., Weisburg S. *Applied Regression Including Computing and Graphics.* John Wiley and Sons Inc., 1999.

Поступила в редакцию/Received by the Editor: *26.02.2020 г.*

Авторство:

Все авторы внесли существенный вклад в планирование работы, проведение анализа и представление результатов, равнозначно участвовали в подготовке первого варианта статьи, а также на всех этапах ее доработки. Все авторы утвердили окончательную версию рукописи.

Сведения об авторах:

Гржибовский Андрей Мечиславович — доктор медицины, заведующий Центральной научно-исследовательской лабораторией Федерального государственного бюджетного образовательного учреждения высшего образования «Северный государственный медицинский университет» Министерства здравоохранения Российской Федерации; профессор кафедры общественного здоровья, здравоохранения, общей гигиены и биоэтики федерального государственного автономного образовательного учреждения высшего образования «Северо-Восточный федеральный университет имени М. К. Аммосова», г. Якутск; 163000, г. Архангельск, Троицкий проспект, д. 51; e-mail: Andrej.Grjibovski@gmail.com; ORCID: 0000–0002–5464–0498, SPIN: 5118–0081;

Горбатова Мария Александровна — кандидат медицинских наук, доцент, магистр общественного здоровья, доцент кафедры стоматологии детского возраста Федерального государственного бюджетного образовательного учреждения высшего образования «Северный государственный медицинский университет» Министерства здравоохранения Российской Федерации; 163000, г. Архангельск, Троицкий проспект, д. 51; e-mail: marigora@mail.ru; ORCID: 0000–0002–6363–9595, SPIN: 7732–0755;

Наркевич Артем Николаевич — кандидат медицинских наук, доцент, заведующий научно-исследовательской лабораторией медицинской кибернетики и управления в здравоохранении, доцент кафедры медицинской кибернетики и информатики Федерального государственного бюджетного образовательного учреждения высшего образования «Красноярский государственный медицинский университет им. проф. В. Ф. Войно-Ясенецкого» Министерства здравоохранения Российской Федерации; 660022, г. Красноярск, ул. Партизана Железняка, д. 1; e-mail: narkevichart@gmail.com; ORCID: 0000–0002–1489–5058, SPIN: 9030–1493;

Виноградов Константин Анатольевич — доктор медицинских наук, профессор, заведующий кафедрой медицинской кибернетики и информатики Федерального государственного бюджетного образовательного учреждения высшего образования «Красноярский государственный медицинский университет им. проф. В. Ф. Войно-Ясенецкого» Министерства здравоохранения Российской Федерации; 660022, г. Красноярск, ул. Партизана Железняка, д. 1; e-mail: vinogradov16@yandex.ru. ORCID: 0000–0001–6224–5618, SPIN: 6924–0110.