DOI: https://doi.org/10.17816/humeco569406

Расчёт объёма выборки при планировании поперечных исследований

Н.А. Митькин[1], С.Н. Драчев[1], Е.А. Кригер[1], В.А. Постоев[1], А.М. Гржибовский[1, 2]

[1] Северный государственный медицинский университет, Архангельск, Российская Федерация;
[2] Северный (Арктический) федеральный университет имени М.В. Ломоносова, Архангельск, Российская Федерация

АННОТАЦИЯ

Поперечные исследования наиболее широко распространены в отечественной медицинской литературе. Однако в подавляющем их большинстве не проводится расчёт размера выборки на этапе планирования, а анализ выполняется с помощью простейших методов статистики. Это не только ограничивает возможности использования данных, но и может привести к ошибочным выводам.

Качество научного исследования определяется грамотным планированием, чёткой постановкой задач и формулировкой статистических гипотез, которые будут проверяться наиболее подходящими для них методами. Одно из центральных мест в этом процессе занимает определение необходимого объёма выборки. В данной статье мы представляем пошаговый алгоритм расчёта объёма выборки, который может применяться для планирования поперечных исследований с различными научными задачами и типами данных. Доступным языком описывается применение самых популярных в биомедицинской литературе методов многомерного анализа данных: логистической регрессии для изучения бинарных исходов и их предикторов и линейной регрессии для оценки независимого влияния нескольких факторов на количественные исходы.

Несмотря на наличие большого числа программ для расчёта объёма выборки, в данной публикации мы демонстрируем применение свободно распространяемой программы G*Power. Программа имеет интуитивно-понятный интерфейс, может применяться для различных статистических тестов и использоваться для расчёта величины эффекта и графического отображения результатов анализа мощности. Каждый этап сопровождается примерами и скриншотами с пошаговым разбором, что делает материал удобным для восприятия и практического применения.

Мы надеемся, что статья станет полезным практическим руководством на этапе планирования исследований и поможет учёным решать большее число задач и оценивать влияние факторов риска на изучаемые исходы с достаточной статистической мощностью.

Ключевые слова: поперечное исследование; объём выборки; регрессионный анализ; G*Power.

Как цитировать:
Митькин Н.А., Драчев С.Н., Кригер Е.А., Постоев В.А., Гржибовский А.М. Расчёт объёма выборки при планировании поперечных исследований // Экология человека. 2023. Т. 30, № 7. С. 509–522. DOI: https://doi.org/10.17816/humeco569406

Рукопись получена: 14.09.2023 **Рукопись одобрена:** 28.09.2023 **Опубликована online:** 18.11.2023

DOI: https://doi.org/10.17816/humeco569406

Sample size calculation for cross-sectional studies

Nikita A. Mitkin[1], Sergei N. Drachev[1], Ekaterina A. Krieger[1], Vitaly A. Postoev[1], Andrej M. Grjibovski[1, 2]

[1] Northern State Medical University, Arkhangelsk, Russian Federation;
[2] M.V. Lomonosov Northern (Arctic) Federal University, Arkhangelsk, Russian Federation

ABSTRACT

The cross-sectional study design is widely prevalent in Russian medical literature. However, a significant number of these studies neglect to calculate the sample size during the planning phase, and the analysis often relies solely on basic bivariate statistics. This compromises the validity of the findings and increases the risk of drawing inaccurate conclusions.

The scientific rigor of a study depends on a quality of planning, a clear problem statement, and precise formulation of statistical hypotheses, which are then tested using the most appropriate analytical methods. At the core of this process lies the determination of the appropriate sample size. The primary objective of this article is to provide a comprehensive, step-by-step guide for the sample size calculation process. By adhering to our guidelines, researchers can ensure that their cross-sectional studies possess sufficient statistical power to generate meaningful results. We acknowledge the significance of tailoring sample size calculations to the specific objectives and data characteristics of each study. Therefore, our approach is designed to be flexible and adaptable, accommodating the unique requirements of diverse research endeavors.

There are several software options available for sample size calculation; however, we use the G*Power software for all the examples presented in this paper. Our guide is designed to provide practical understanding of the topic, with each step being accompanied by illustrative examples and detailed screenshots. This approach ensures that the material is not only understandable but also applicable in real-world scenarios. Furthermore, we take the extra step of interpreting every dialog box and screenshot, aiming to create a comfortable user experience with the software. We hope that this paper will serve as a valuable guide in the planning stage of a study, helping researchers to address a wider range of issues and reliably estimate the associations between selected exposures and the outcomes of interest with sufficient statistical power.

Keywords: cross-sectional studies; sample size; regression analysis; G*Power.

To cite this article:
Mitkin NA, Drachev SN, Krieger EA, Postoev VA, Grjibovski AM. Sample size calculation for cross-sectional studies. *Ekologiya cheloveka (Human Ecology).* 2023;30(7):509–522. DOI: https://doi.org/10.17816/humeco569406

Received: 14.09.2023 **Accepted:** 28.09.2023 **Published online:** 18.11.2023

ВВЕДЕНИЕ

Поперечные (одномоментные) исследования широко применяются в науках о здоровье в России и странах ближнего зарубежья. В таких исследованиях одновременно собираются данные о воздействии (как правило, факторе риска) и исходе (заболевании или состоянии), что обеспечивает быстрый и экономически эффективный сбор данных. Они позволяют определить распространённость состояния или заболевания, а также найти связи между исследуемыми явлениями. Это открывает путь к построению новых гипотез и создаёт основу для последующих продольных обсервационных и экспериментальных исследований [1].

Однако исследователи при планировании работы часто допускают две существенные ошибки. Во-первых, расчёт необходимого объёма выборки на этапе планирования проводится лишь в небольшой части исследований. Этот показатель определяет количество участников, которых необходимо включить в выборочную совокупность для обнаружения статистически значимых различий или связей, если они существуют. При недостаточном объёме выборки можно сделать вывод об отсутствии связи, в то время как она есть, но не могла быть выявлена из-за недостаточной мощности исследования. При избыточном объёме выборки легко можно выявить результаты, которые будут статистически значимы, но не будут иметь важного клинического значения, не говоря уже о завышенной стоимости такого исследования.

Во-вторых, в отечественной литературе данные анализируются в основном бивариантными методами статистического анализа, т.е. рассматривается связь между одной зависимой и одной независимой переменной [2, 3]. Хотя бивариантный анализ позволяет получить предварительные сведения о наличии связи между переменными, он не учитывает конфаундеры — вмешивающиеся факторы, которые связаны как с воздействием, так и с исходом, и могут ослабить, усилить, а в некоторых случаях даже развернуть связь в противоположном направлении, приводя к ошибочным выводам. Рассмотрим связь между потреблением алкоголя и риском развития рака лёгкого. Бивариантный анализ может показать наличие сильной прямой корреляции. Однако включение в модель курения снижает силу связи до незначимых уровней. Поэтому для установления независимой связи между переменными необходимо использовать математические модели, учитывающие эти факторы [4, 5].

Многомерный регрессионный анализ решает эту проблему. Данный метод помогает понять, как несколько факторных признаков (например, возраст, питание или физическая активность) влияют на один исход (например, масса тела или артериальное давление) [6]. Основная ценность многомерного регрессионного анализа заключается в том, что он позволяет выявить независимое влияние каждого признака на исход с учётом коррекции на другие факторы, включённые в модель. Выбор вида регрессионного анализа зависит от типа зависимой переменной (результативного признака). Линейная регрессия подходит для количественных исходов, а логистическая регрессия — для категориальных [7–9]. Для дискретных данных может использоваться регрессия Пуассона, которая не будет рассматриваться в данной работе (хотя и заслуживает отдельного подробного описания), так как этот метод анализа в отечественной медицинской науке применяется достаточно редко.

Ранее мы представили пошаговый расчёт объёма выборки для исследований, в которых применяются наиболее популярные критерии для бивариантного анализа данных — критерий Стьюдента, однофакторный дисперсионный анализ, ранговый критерий Вилкоксона и U-критерий Манна–Уитни [10]. В данной работе мы переходим к расчёту размера выборки для исследований более высокого уровня, в которых планируется оценка независимых от конфаундеров связей между факторами риска и исходами с помощью логистической и линейной регрессии.

Несмотря на возможное применение формул для расчёта объёма выборки, всё большее распространение получают автоматизированные инструменты. В связи с тем, что большинство статистических пакетов для расчёта объёма выборки являются платными, а бесплатные калькуляторы редко позволяют делать расчёты для многомерных методов, мы рассмотрим возможности программного обеспечения G*Power. Основным преимуществом этой программы является интуитивно-понятный интерфейс и гибкий функционал, который позволяет применять различные семейства статистических тестов, вычислять величину эффекта и визуализировать результаты анализа мощности. Программа доступна для бесплатного скачивания с официального сайта https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower и работает в операционных системах Windows и MacOS. В данной статье использована программа G*Power версии 3.1.9.7 для Windows.

РАСЧЁТ ОБЪЁМА ВЫБОРКИ ПРИ ИСПОЛЬЗОВАНИИ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Логистическая регрессия является одним из самых распространённых видов анализа в биомедицинских исследованиях. Она помогает понять, как независимые переменные (факторные признаки, факторы риска, предикторы), выраженные и категориальными, и количественными переменными, способны влиять на зависимую переменную — бинарный исход, который может иметь только два состояния (например, 1 или 0, да или нет) [11].

К примеру, с помощью логистической регрессии можно ответить на следующие биомедицинские вопросы:

DOI: https://doi.org/10.17816/humeco569406

1. Связана ли диета с низким содержанием жиров с риском развития инфаркта миокарда (исход: наличие или отсутствие инфаркта миокарда)?
2. Как связан уровень холестерина с пищевыми предпочтениями пациентов (исход: вегетарианство или невегетарианство)?
3. Влияет ли индекс массы тела на скорость восстановления после эндопротезирования тазобедренного сустава (исход: срок реабилитации до 3 мес или более 3 мес)?
4. Как связан уровень физической активности (низкий, средний, высокий) и риск развития диабета 2-го типа (исход: наличие или отсутствие диабета) с поправкой на возраст и индекс массы тела?
5. Связан ли уровень тревожности по шкале Бека со статусом вакцинации против COVID-19 (исход: вакцинирован или не вакцинирован) с поправкой на пол и возраст участников?

Логистическая регрессия рассчитывает коэффициенты регрессии для каждого предиктора. Однако в исходном виде эти коэффициенты могут быть трудными для интерпретации. Чтобы сделать их более понятными, коэффициенты преобразуют в отношения шансов (ОШ). Для более глубокого погружения в теоретические основы метода читателям рекомендуется обратиться к специализированной литературе [8].

Отношение шансов представляет собой количественную оценку шансов наступления события в основной группе (подвергающейся воздействию фактора риска) по отношению к шансам в контрольной группе (не подвергающейся воздействию). Приведём его интерпретацию:

1. Если ОШ равно 1, это означает отсутствие разницы в шансах между двумя группами. Другими словами, воздействие не влияет на вероятность исхода.
2. ОШ больше 1 указывает на более высокие шансы наступления события в основной группе. Например, если ОШ равно 3,5, это означает, что шансы наступления события в основной группе в 3,5 раза выше по сравнению с контрольной группой. Часто это формулируется как «шансы наступления события в 3,5 раза выше в основной группе».
3. ОШ меньше 1 предполагает сниженные шансы в основной группе. Так, значение ОШ, равное 0,5, может быть интерпретировано двумя распространёнными способами:

   1) «шансы наступления события в основной группе в два раза ниже, чем в контрольной»;

   2) «шансы наступления события в контрольной группе в два раза выше, чем в основной».

Например, при изучении влияния определённой диеты на риск развития диабета ОШ больше 1 будет означать, что шансы развития диабета у тех, кто придерживается такой диеты, выше. И наоборот, ОШ меньше 1 указывает на защитное действие диеты против диабета.

Чтобы рассчитать необходимый объём выборки для исследования с применением логистического анализа, исследователям необходимо предварительно определить несколько аспектов:

1. Тип проверки гипотезы: односторонний или двусторонний тест. Этот выбор основан на предполагаемом направлении связи. Например, если мы твёрдо уверены, что диета с высоким содержанием сахара может только повысить риск развития деменции и не может его снизить, то мы выберем односторонний тест. Если мы предполагаем, что диета может как повысить, так и понизить риск развития деменции, следует выбрать двусторонний тест.
2. Минимальное ОШ. Представляет собой наименьшую силу связи, которая имеет клиническое или исследовательское значение. Например, мы хотим исследовать связь между объёмом потребления рыбы в год (воздействие) и низкой плотностью костной ткани (исход). Из опубликованных исследований мы узнали, что у людей с низким потреблением рыбной продукции ОШ было равно 2,0. Это означает, что шанс наличия низкой плотности костной ткани у них был в 2,0 раза выше, чем у тех, кто потреблял нормальное количество рыбы. Поэтому для нашего исследования мы можем выбрать ОШ равным 2,0.
3. Распространённость исхода. Это доля участников, у которых предположительно проявляется изучаемый исход. Например, мы изучаем связь между малоподвижным образом жизни (воздействие) и ожирением (исход). Под распространённостью исхода будет пониматься предполагаемая доля населения, страдающая ожирением. Эта оценка может быть основана на результатах ранее проведённых исследований.
4. Уровень α-ошибки (ошибка 1-го рода), т.е. вероятность предположить наличие эффекта, когда его нет. Например, вывод о том, что лекарство эффективно, когда в действительности оно не оказывает эффекта (ложноположительный результат). Обычно уровень α-ошибки задаётся как 0,05 или меньше.
5. Уровень β-ошибки (ошибка 2-го рода), т.е. вероятность упустить существующий эффект. Например, когда не удаётся обнаружить эффект лекарства, при том что в действительности оно эффективно (ложноотрицательный результат).
6. Статистическая мощность, которая рассчитывается как 1–β и представляет собой вероятность правильно выявить истинную связь. В биомедицинских исследованиях мощность традиционно устанавливается как 80% или выше.

Рассмотрим конкретный пример. Предположим, мы планируем исследовать, как стоматологическая тревожность (страх посещения стоматолога) связана с социально-экономическим статусом. Прежде чем рассчитать,

сколько участников нам необходимо включить в исследование, нужно выполнить три шага:

1. Определить переменные.
   1.1. Воздействие: социально-экономический статус. Мы будем использовать шкалу MacArthur, которая позволяет участникам субъективно оценить своё социальное положение по шкале от 1 (наименее благоприятное) до 10 (наиболее благоприятное) [12].
   1.2. Исход: стоматологическая тревожность. Исследования показали, что стоматологическая тревожность может быть оценена с помощью всего одного вопроса: «Вы боитесь ходить к стоматологу?» и данная оценка является надёжной и достоверной [13]. Поэтому в зависимости от ответа участникам будет присвоен статус наличия («Да») или отсутствия тревоги при посещении стоматолога («Нет»).
2. Сформулировать гипотезы.
   2.1. H0 (нулевая гипотеза): стоматологическая тревожность не связана с социально-экономическим статусом. В статистическом выражении отношение шансов равно 1, а коэффициент регрессии для социально-экономического статуса равен 0.
   2.2. H1 (альтернативная гипотеза): стоматологическая тревожность связана с социально-экономическим статусом. Это означает, что отношение шансов не равно 1, а коэффициент регрессии для социально-экономического статуса не равен 0.
3. Установить входные параметры для исследования.
   3.1. Установим ОШ. Представим, что по результатам предыдущих исследований распространённость стоматологической тревожности среди всего населения составляет 50%, или 0,5, а среди людей с низким социально-экономическим статусом — 60%, или 0,6. Рассчитаем шансы стоматологической тревожности для первой группы — 0,5/(1–0,5)=1 и для второй группы — 0,6/(1–0,6)=1,5 соответственно. Затем найдём ОШ, разделив шансы для основной группы на шансы для контрольной группы: 1,5/1=1,5. Полученный результат говорит о том, что шансы встретить человека с низким социально-экономическим статусом среди испытывающих стоматологическую тревожность в 1,5 раза выше, чем среди населения без стоматологической тревожности.
   3.2. Статистическую мощность исследования установим 0,8, т.е. вероятность обнаружения истинной связи составит 80%.
   3.3. Уровень α-ошибки укажем 0,05, что является стандартным порогом для определения статистической значимости.
   3.4. Применим двусторонний тест, поскольку мы можем обнаружить связь в любом направлении (более высокий или более низкий социально-экономический статус влияет на стоматологическую тревожность).
   3.5. Наконец, мы предполагаем, что субъективные оценки участниками своего социально-экономического статуса будут иметь нормальное распределение, что является обычным допущением во многих биомедицинских исследованиях.

С учётом приведённых выше характеристик мы можем рассчитать необходимый объём выборки для нашего исследования. Для этого устанавливаем и запускаем приложение G*Power, после чего откроется главное диалоговое окно программы (рис. 1).

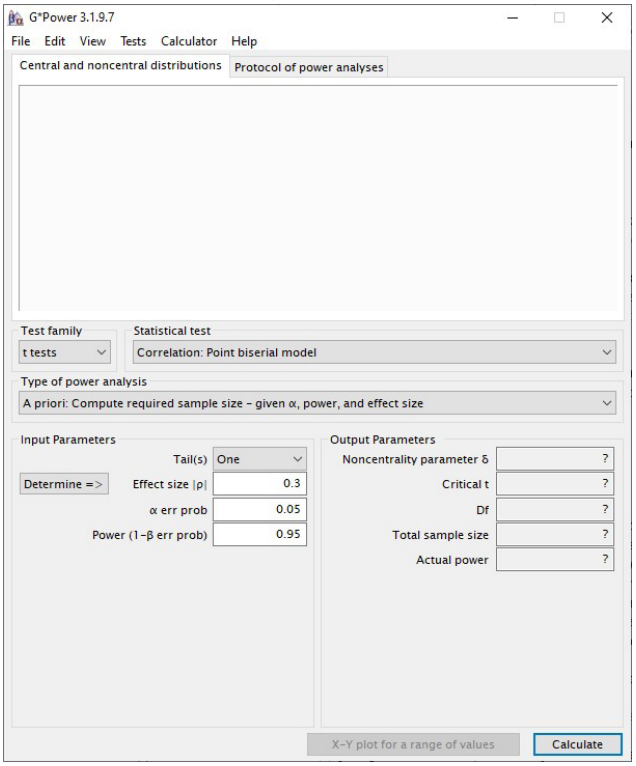

Рис. 1. Главное диалоговое окно программы G*Power.
Fig. 1. G*Power main dialog box.

Проведём расчёты объёма выборки для нашего исследования следующим образом:

1. Сперва в графе «Test family/Семейство тестов» выбираем нужную категорию статистического теста из выпадающего списка (в нашем примере — это «z tests/z-тесты») и сам статистический тест «Logistic regression/Логистическая регрессия».
2. Далее в графе «Type of power analysis/Тип анализа мощности» следует выбрать «A priori: compute required sample size — given α, power, and effect size/Априори: вычислить необходимый размер выборки — учитывая величину α-ошибки, мощность и величину эффекта», поскольку мы проводим расчёт на этапе планирования данного исследования.
3. Далее в разделе «Input parameters/Входные параметры» нужно задать параметры для расчёта:
   3.1. «Tail(s)/Вид теста»: «two/двусторонний тест»;
   3.2. «Odds ratio/Отношение шансов» — 1,5;

3.4. pr (Y=1|X=1) H0: вероятность проявления стоматологической тревожности в исследуемой популяции — 0,5;
3.5. «α err prob/Уровень α-ошибки» — 0,05;
3.6. «Power (1–β err prob)/Статистическая мощность» — 0,8;
3.7. «R^2 other X/Коэффициент детерминации ковариат». Этот параметр лежит в интервале [0; 1] и показывает вариабельность исхода, которая обусловлена другими предикторами (ковариатами), т.е. без учёта основного предиктора. Если ковариаты отсутствуют (как в нашем случае с одним предиктором), указываем значение 0. В противном случае необходимо рассчитать квадрат коэффициента множественной корреляции других ковариат.

4. «X distribution/Распределение независимой переменной»: «Normal/Нормальное распределение» с параметрами основного предиктора μ (среднее арифметическое) и σ (стандартное отклонение), равными 0 и 1 соответственно.

После ввода вышеуказанных параметров в соответствующие поля диалогового окна программы необходимо нажать «Calculate/Рассчитать».

В строке «Total sample size/Общий размер выборки» мы получаем интересующий нас результат (рис. 2). Для нашего гипотетического исследования необходимо включить в анализ минимум 208 человек. Однако, учитывая непредвиденные обстоятельства, такие как отказ от участия в исследовании или ошибки при сборе данных, целесообразно повысить этот показатель примерно на 15–25%.

В данном примере независимая переменная (социально-экономический статус) была количественной. Однако расчёт может быть выполнен и для ситуации, когда независимая переменная будет категориальной.

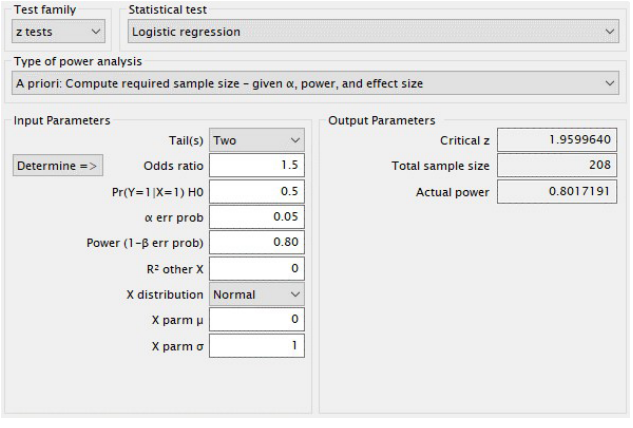

Рис. 2. Диалоговое окно программы G*Power с введёнными параметрами и рассчитанным объёмом выборки для логистической регрессионной модели с одной независимой количественной переменной.
Fig. 2. G*Power dialog box with entered parameters and calculated sample size for logistic regression model with one independent numeric variable.

Представим, что социально-экономический статус был классифицирован как низкий или высокий, и распространённость низкого статуса в популяции составила 60%. Не меняя входные параметры, мы должны сперва заменить «X distribution/Распределение независимой переменной» на «Binomial/Биноминальное» и затем указать значение 0,6 в графе в поле «X parm π». В этом случае необходимый объём выборки составит 809 человек (рис. 3). Учитывая процент отклика, итоговое количество участников исследования рекомендуется увеличить по крайней мере на 15% (в нашем примере оно составит 931 человек).

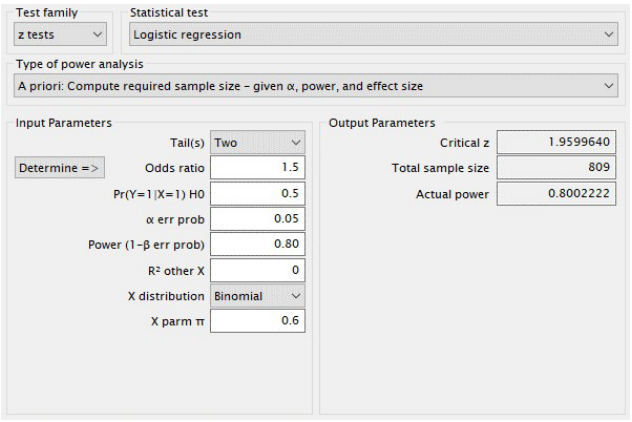

Рис. 3. Диалоговое окно программы G*Power с введёнными параметрами и рассчитанным объёмом выборки для логистической регрессионной модели с одной независимой бинарной переменной.
Fig. 3. G*Power dialog box with entered parameters and calculated sample size for logistic regression model with one independent binary variable.

Мы также можем рассчитать объём выборки N', когда в анализ будут включены несколько независимых переменных. Для этого используется формула $N'=N/(1–R^2)$, где N — расчётный объём выборки для одной независимой переменной, а R^2 — коэффициент детерминации ковариат [14].

Допустим, исследователи планируют оценить связь стоматологической тревожности с субъективным социально-экономическим статусом, одновременно корректируя её на конфаундеры, такие как частота посещения врача-стоматолога и уровень образования. При расчёте будет важно не количество этих факторов, а их совокупное влияние на зависимую переменную, исключая основную переменную воздействия.

Допустим, величина R^2 в нашем примере будет достигать 20% без учёта основного предиктора. Следовательно, искомый объем выборки N' составит 809/(1–0,2), или 1012 человек. Это же значение (с погрешностью на округление) может быть получено, если в программе мы введём значение 0,2 в поле «R^2 other X/Коэффициент детерминации для других переменных-предикторов» и нажмём «Calculate/Рассчитать» (рис. 4).

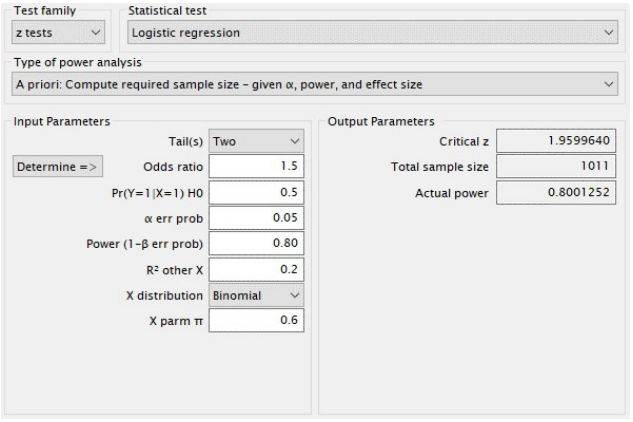

Рис. 4. Диалоговое окно программы G*Power с введёнными параметрами и рассчитанным объёмом выборки для логистической регрессионной модели с несколькими независимыми переменными.
Fig. 4. G*Power dialog box with entered parameters and calculated sample size for logistic regression model with several independent variables.

При увеличении значения R^2 объём выборки увеличивается. Например, если R^2 будет 30% (или 0,3), то нам потребуется уже минимум 1156 участников. При этом следует предостеречь начинающих исследователей от использования в расчётах коэффициента детерминации, близкого к 100%, поскольку ни одна модель, построенная на данных биомедицинских исследований, не является идеально точной [15].

Визуальное представление позволяет наглядно понять эти расчёты. Для этого необходимо перейти на вкладку «X-Y plot for a range of values/График X-Y для диапазона значений», которая находится в нижней части основного диалогового окна. Программа предлагает нам построить графики зависимости между объёмом выборки, статистической мощностью, α-ошибкой и ОШ.

В качестве демонстрации построим график зависимости объёма выборки от статистической мощности. Для этого в нижней части окна «Plot Parameters/Параметры графика» выберем «Total sample size/Общий объём выборки» по оси Y (Plot (on y axis)) как функцию от статистической мощности (as a function of Power (1–β err prob)). Кроме того, на данном графике мы можем отразить также зависимость не только для ОШ 1,5 и выше, но и, скажем, для 2,0 и выше, выбрав в выпадающем меню «Plot/График» цифру 2 и в меню «Odds ratio from 1,5 in steps of 0,5/Отношения шансов 1,5 с шагом 0,5». После этого нажимаем на «Draw plot/Нарисовать график» (рис. 5).

Из графика видно, что при увеличении статистической мощности объём выборки увеличивается. Более того, для выявления более сильного эффекта (ОШ 2,0) требуется меньший объем выборки по сравнению с более слабым эффектом (ОШ 1,5). Полученные значения могут быть представлены в виде таблицы при нажатии на вкладку «Table/Таблица».

Описанная выше процедура расчёта объёма выборки при выполнении регрессионного анализа для категориальных бинарных исходов проводится на этапе планирования исследования, до сбора данных. Однако, если исследование основывается на уже собранных данных, можно оценить его мощность — определить, соответствует ли имеющийся объём выборки поставленным научным задачам.

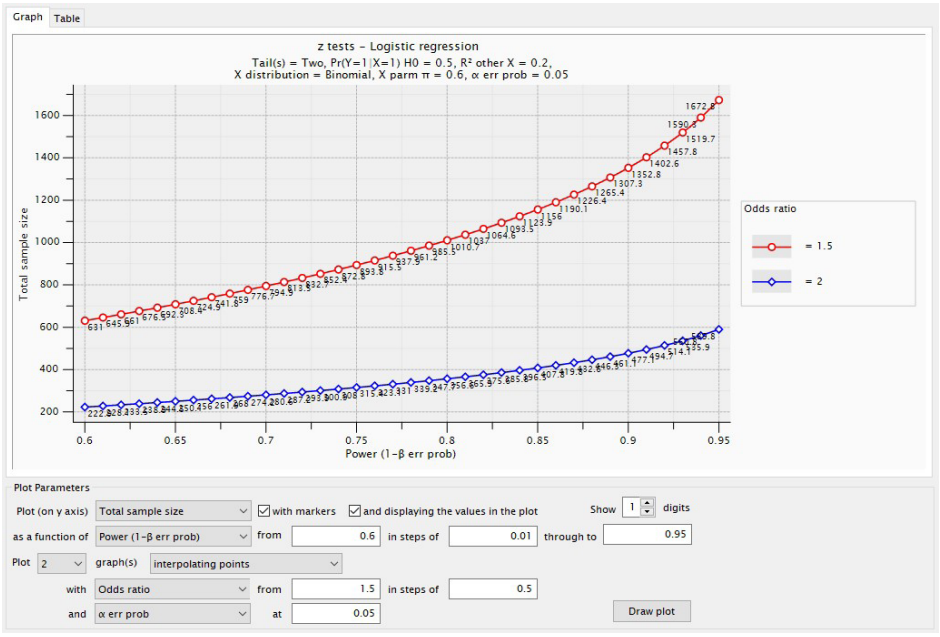

Рис. 5. График зависимости объёма выборки от статистической мощности для отношения шансов 1,5 и 2,0, двустороннего теста, распространённости исхода 50%, уровня α-ошибки 0,05, распространённости фактора риска 60% и коэффициента детерминации многомерной модели 0,2.
Fig. 5. Relationship between sample size and statistical power for odds ratios of 1.5 and 2.0, a two-tailed test, an outcome prevalence of 50%, an α-error level of 0.05, a risk factor prevalence of 60%, and a multivariate model determination coefficient of 0.2.

DOI: https://doi.org/10.17816/humeco569406

Для этого в основном диалоговом окне программы, в графе «Type of power analysis/Тип анализа мощности» следует выбрать «Post hoc: compute achieved power — given α, sample size, and effect size/Апостериорный анализ: вычислить достигнутый уровень мощности, учитывая уровень α-ошибки, объём выборки, величину эффекта». При объёме выборки 1011 человек и неизменности других параметров, которые использовались в вышеприведённом примере, не удивительно, что мы получим достигнутый уровень статистической мощности 0,8, или 80% (рис. 6).

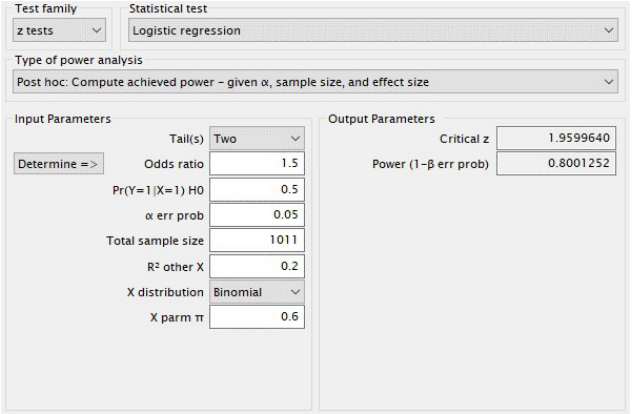

Рис. 6. Диалоговое окно программы G*Power для проведения апостериорного анализа с целью оценки статистической мощности для логистического регрессионного анализа.
Fig. 6. G*Power dialog box for conducting post hoc analysis aimed at assessing statistical power for a logistic regression.

По сути, вычислительные основы остаются неизменными, меняются лишь входные параметры — объём выборки и мощность исследования.

Следует обратить внимание начинающих исследователей на такой аспект: если приступать к решению научных проблем, опираясь только на уже имеющиеся данные, без предварительной оценки статистической мощности таких исследований, это может привести к необоснованным или ошибочным выводам.

Допустим, исследователь собрал базу данных в размере 1011 человек и решил проверить новую гипотезу: связан ли низкий уровень самооценки состояния полости рта с проживанием в сельской местности? Допустим, входные параметры для расчёта статистической мощности будут следующие: распространённость низкой самооценки полости рта среди населения составляет 25% (или 0,25); минимальное ОШ — 1,5; двусторонний тест; уровень α-ошибки — 0,05; распространённость фактора риска (проживание в сельской местности) — 15%; коэффициент детерминации — 0,2. Полученная статистическая мощность такого исследования составит 48% (рис. 7), что существенно ниже 80%, традиционно принятого в биомедицинских исследованиях, и данный факт существенно увеличит вероятность не обнаружить имеющиеся в реальности взаимосвязи.

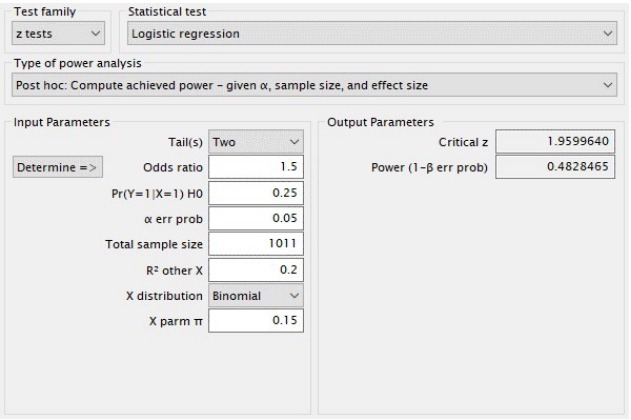

Рис. 7. Диалоговое окно программы G*Power для оценки статистической мощности исследования при следующих входных параметрах: двусторонний тест; отношение шансов — 1,5 и выше; распространённость исхода — 25%; уровень α-ошибки — 0,05; объём выборки — 1011; коэффициент детерминации — 0,2; распространённость фактора риска — 15%.
Fig. 7. G*Power dialog box evaluating statistical power of a study with the following input parameters: two-tailed test; odds ratios of 1.5 or greater; outcome prevalence of 25%; α-error level of 0.05; sample size of 1011; determination coefficient of 0.2; and risk factor prevalence of 15%.

Этот наглядный пример ещё раз подчеркивает необходимость и важность тщательного планирования исследования, чтобы размер выборки соответствовал тому набору гипотез, которые будут впоследствии проверяться на собранных данных.

РАСЧЁТ ОБЪЕМА ВЫБОРКИ ПРИ ИСПОЛЬЗОВАНИИ ЛИНЕЙНОЙ РЕГРЕССИИ

Линейная регрессия является эффективным статистическим инструментом, позволяющим исследователям выявлять и количественно оценивать линейную связь между исходом (зависимой количественной переменной) и одним или несколькими предикторами (независимыми переменными). Этот метод помогает исследователям понимать независимое влияние множества факторов на исход и строить прогностические модели.

Например, применение линейной регрессии в биомедицинских исследованиях позволяет ответить на следующие вопросы:

1. Как уровень потребления соли (воздействие) предсказывает артериальное давление (исход) с поправкой на возраст и физическую активность?
2. Как связан уровень физической активности (воздействие) с индексом массы тела (исход) с учётом диеты и генетических факторов?
3. Какова связь различных компонентов рациона питания (воздействие) и уровня холестерина (исход) с поправкой на возраст и приём лекарств?

DOI: https://doi.org/10.17816/humeco569406

4. Как изменение количества часов сна в сутки (воздействие) влияет на оценку психического здоровья (исход) с учётом уровня стресса и приёма лекарств?
5. Связан ли объём потребления кальция (воздействие) с плотностью костной ткани (исход) с поправкой на пол и возраст?

Несмотря на широкие возможности линейной регрессии, исследователи должны тщательно изучить интересующие данные и знать несколько основных допущений для её применения, которые будут обеспечивать обоснованность полученных результатов.

При использовании линейной регрессии в качестве главного условия рассматривается нормальное распределение остатков (разности между наблюдаемыми и прогнозируемыми значениями исхода). Нормальное распределение остатков обеспечивает надёжность используемых статистических тестов и доверительных интервалов. Что касается предикторов, то они необязательно должны быть нормально распределены. Однако их связь с исходом должна быть линейной, что можно визуально проверить с помощью диаграмм рассеяния (скатерограмм). Кроме того, данные предикторы не должны быть слишком тесно связаны друг с другом (условие отсутствия мультиколлинеарности), поскольку это может ослабить независимый эффект каждого предиктора. Наконец, разброс остатков должен оставаться постоянным при всех значениях предикторов, что гарантирует одинаковую точность предсказаний модели.

При планировании исследования нередки случаи, когда некоторые допущения линейной регрессии оказываются неопределёнными или невыполнимыми. В таких случаях исследователи могут предпринять несколько упреждающих шагов. Во-первых, можно использовать опубликованные данные или пилотные исследования для оценки характера данных и их соответствия допущениям регрессии. Если окажется, что допущения нарушены, то преобразование данных (например, логарифмическое преобразование) часто помогает в выполнении требований. Во-вторых, в качестве альтернативы исследователи могут рассмотреть другие статистические методы, более подходящие к характеристикам данных. Например, если связь между переменными не является линейной, то более подходящими могут оказаться нелинейная регрессия или другие непараметрические методы. В-третьих, консультация со статистиком или опытным исследователем на этапе планирования может дать ценные идеи и рекомендации по выбору оптимального аналитического подхода.

Более подробные сведения о линейном регрессионном анализе и принципы интерпретации результатов подробно представлены в специализированной литературе по статистическому анализу [16–18]. Для более глубокого усвоения материала читателям рекомендуется обратиться к этим источникам.

Определение необходимого объёма выборки в линейных регрессионных моделях опирается на ряд подходов, которые подбираются под задачи исследователя. Один из таких подходов основан на проверке нулевой гипотезы, согласно которой коэффициент регрессии равен нулю, что означает отсутствие взаимосвязи между переменными.

В качестве примера рассмотрим простую линейную регрессию, которая используется для изучения связи между зависимой количественной переменной и одной независимой переменной. Предположим, что исследователь хочет понять, как продолжительность беременности, или срок гестации, влияет на вес ребёнка при рождении. Чтобы определить необходимый объём выборки для ответа на этот исследовательский вопрос, необходимо осуществить следующие шаги:

1. Определить переменные.
   1.1. Воздействие: срок гестации (недели).
   1.2. Исход: масса тела новорождённого (граммы).
2. Сформулировать гипотезы.
   2.1. H0 (нулевая гипотеза): срок гестации не влияет на массу тела новорождённого. В статистическом смысле это означает, что коэффициент регрессии для срока гестации равен нулю.
   2.2. H1 (альтернативная гипотеза): срок гестации влияет на массу тела новорождённого. С точки зрения статистики это означает, что коэффициент регрессии для срока гестации не равен нулю.
3. Установить входные параметры для исследования.
   3.1. Уровень α-ошибки (ошибка 1-го рода).
   3.2. Статистическая мощность.
   3.3. Число независимых переменных.
   3.4. Коэффициент корреляции между зависимой и независимыми переменными.
   3.5. Размер эффекта (effect size f^2). Эта величина количественно характеризует силу связи или различий данных. В контексте сравнения двух групп данный размер часто рассчитывается как разница между средними значениями групп по отношению к вариабельности данных (часто это суммарное стандартное отклонение). По сути, он определяет, насколько велика наблюдаемая разница с учётом разброса или дисперсии точек данных [19]. Если мощность исследования и уровень α-ошибки обычно устанавливаются независимо, то размер эффекта определяется строго в соответствии с характером исследования и данных. Для оценки размера эффекта учёные часто опираются на результаты предыдущих исследований либо экспериментальные данные для определения ожидаемой величины связей или различий. Для обнаружения большего размера эффекта требуется меньшая выборка, и наоборот. Значения 0,02; 0,15; 0,35 и выше считаются малым, средним и большим размером эффекта соответственно [20, 21].

Предположим, результаты пилотного исследования показали, что корреляция между сроком гестации и массой тела составляет 0,4. Перед исследователем встаёт

задача: рассчитать, сколько новорождённых необходимо включить в исследование, если известно, что корреляция между сроком гестации и массой тела составляет 0,4. Статистическую мощность установим на уровне 80%, уровень α-ошибки — 0,05 при двустороннем тесте.

Для расчёта объёма выборки в программе G*power нам необходимо выполнить следующие действия:

1. В графе «Test family/Семейство тестов» выбираем нужную категорию статистического теста из выпадающего списка: в данном примере это «F tests/F-тесты».
2. После чего указываем статистический тест «Linear multiple regression: fixed model, R^2 deviation from zero/Множественная линейная регрессия: фиксированная модель, отклонение R^2 от нуля».
3. Далее в графе «Type of power analysis/Тип анализа мощности» следует выбрать «A priori: compute required sample size — given α, power, and effect size/Априори: вычислить необходимый размер выборки — учитывая величину α-ошибки, мощность и величину эффекта».
4. Далее в разделе «Input parameters/Входные параметры» необходимо задать следующие параметры для расчёта:

4.1. «α err prob/уровень α-ошибки» — 0,05;

4.2. «Power (1–β err prob)/Статистическая мощность» — 0,8;

4.3. «Number of predictors/Число предикторов» — 1, исходя из условий примера.

5. Последний входной параметр «Effect size f^2/Размер эффекта» требует вычисления, если заранее неизвестен. Для этого:

5.1. Рядом с «Effect size f^2» выбираем «Determine/Определить».

5.2. В появившемся окне необходимо выбрать способ расчёта «From correlation coefficient/С помощью коэффициента корреляции». В нашем примере мы обладаем информацией о коэффициенте корреляции 0,4 (исходя из результатов пилотного исследования).

5.3. Вводим квадрат коэффициента корреляции (в нашем случае это 0,16) в квадрат в графу «Squared multiple correlation/Квадрат множественного коэффициента корреляции» и нажимаем «Calculate and transfer to main window/Рассчитать и перенести в главное окно». В результате расчёта получается значение effect size f^2, равное 0,19, которое автоматически переносится в главное диалоговое окно (рис. 8).

6. После этого остается нажать только «Calculate/Рассчитать».

В строке «Total sample size/Общий размер выборки» получаем результат, который показывает, что необходимо включить в исследование минимум 44 новорождённых (рис. 9). Однако, учитывая возможные отказы от участия

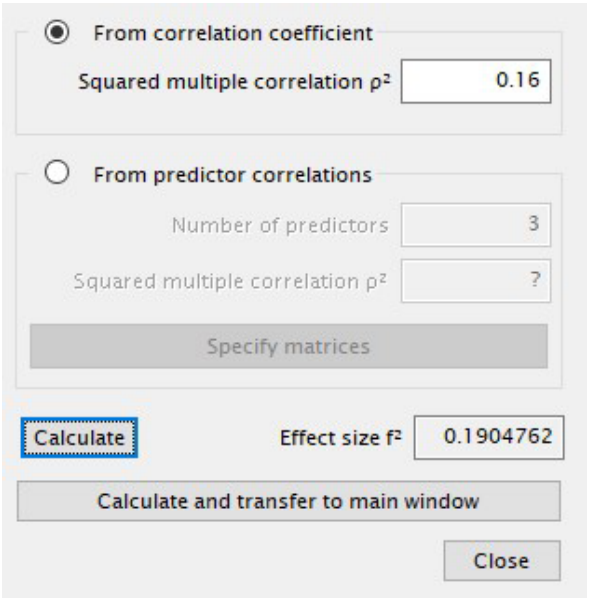

Рис. 8. Диалоговое окно программы G*Power для расчёта «Effect size f^2/Размер эффекта» в линейном регрессионном анализе.
Fig. 8. G*Power dialog box for calculating the "Effect size f^2" in a linear regression analysis.

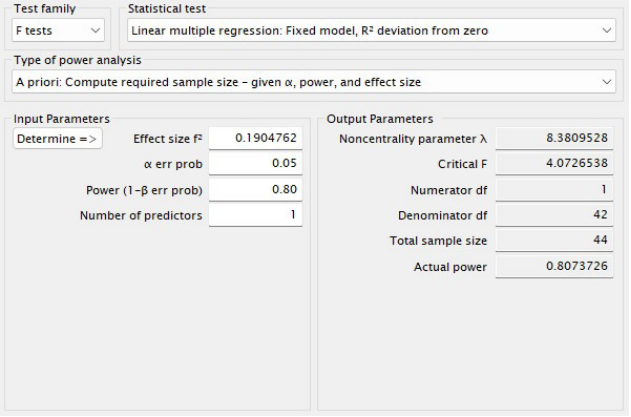

Рис. 9. Диалоговое окно программы G*Power с введёнными параметрами расчёта и полученным результатом для простой линейной регрессионной модели.
Fig. 9. G*Power dialog box presenting entered calculation parameters and an output for a simple linear regression model.

в исследовании, целесообразно увеличить это число на 15–25%.

Далее рассмотрим пример с несколькими предикторами. Допустим, исследователь стремится определить связь не только между сроком гестации и массой тела новорождённого, но и рядом таких факторов, как уровень образования матери и отца, индекс массы тела матери, которые также будут введены в модель в качестве независимых переменных.

Рассчитаем, сколько новорождённых требуется включить в исследование, чтобы определить средний размер эффекта (0,15) независимых переменных на зависимую переменную при статистической мощности 0,8 и уровне α-ошибки 0,05.

Для этого:

1. В графе «Test family/Семейство тестов» выбираем категорию «F tests/F-тесты» и статистический тест «Linear multiple regression: fixed model, R^2 deviation from zero/Множественная линейная регрессия: фиксированная модель, отклонение R^2 от нуля».
2. Далее в графе «Type of power analysis/Тип анализа мощности» следует выбрать «A priori: compute required sample size — given α, power, and effect size/Априори: вычислить необходимый размер выборки — учитывая величину α-ошибки, мощность и величину эффекта».
3. В разделе «Input parameters/Входные параметры» необходимо задать параметры для расчёта:
   3.1. «Effect size/Размер эффекта» — 0,15;
   3.2. «α err prob/уровень α-ошибки» — 0,05;
   3.3. «Power (1–β err prob)/Статистическая мощность» — 0,8;
   3.4. «Number of predictors/Число предикторов» — 4.
4. В завершение нажимаем «Calculate/Рассчитать».

Можно сделать вывод, что для ответа на поставленный исследовательский вопрос необходимо включить в исследование минимум 85 человек (рис. 10). С учётом дополнительных 15% на возможный отказ необходимый размер выборки составит 98 человек.

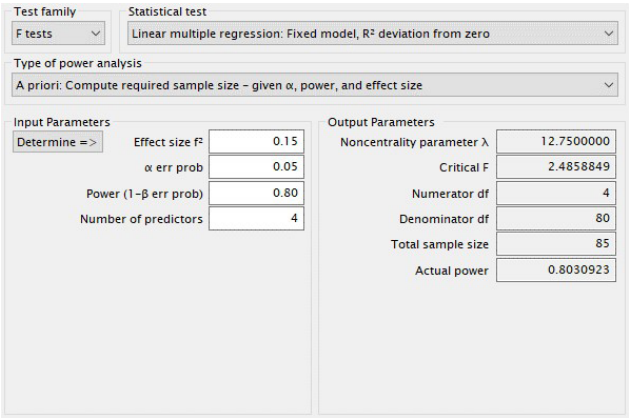

Рис. 10. Диалоговое окно программы G*Power с введёнными параметрами расчёта и результатом для множественной линейной регрессионной модели с несколькими независимыми переменными.
Fig. 10. G*Power dialog box displaying entered calculation parameters and results for a multiple linear regression model with several independent variables.

Если размер эффекта неизвестен, но исследователь может предположить коэффициент корреляции между переменными, программа сама рассчитает размер эффекта.

Для этого:

1. Напротив графы «Effect size f^2/Размер эффекта» необходимо нажать «Determine/Определить», затем активировать поле «From correlation coefficient/Из коэффициента корреляции».
2. В появившемся окне нужно указать соответствующий коэффициент в графе «Squared multiple correlation ρ^2/Квадрат множественного коэффициента корреляции», если он известен.
3. Если квадрат множественного коэффициента корреляции неизвестен, его можно рассчитать с помощью программы. Для этого выбираем флажок «From predictor correlations/Из коэффициентов корреляции предикторов» и указываем число предикторов в окне «Number of predictors/Количество предикторов» (в нашем случае 4). Нажимаем «Specify metrics/Указать показатели». В открывшемся окне в таблице в строке «Corr with outcome Y/Корреляция с зависимой переменной» указываем коэффициенты корреляции исхода с каждым из предикторов в модели. Допустим, по результатам пилотного исследования мы получили следующие коэффициенты корреляции между массой тела новорождённого (исход) и изучаемыми предикторами: 0,4; 0,15; 0,24 и 0,31. Вводим данные и нажимаем «Accept values/Принять значения» (рис. 11).

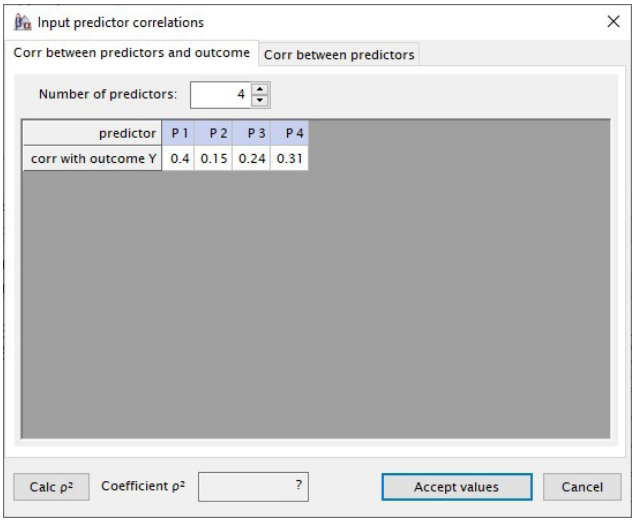

Рис. 11. Диалоговое окно программы G*Power с введёнными параметрами расчёта квадрата множественного коэффициента корреляции для множественной линейной регрессионной модели с несколькими независимыми переменными.
Fig. 11. G*Power dialog box displaying entered calculation parameters for squared multiple correlation coefficient for a multiple linear regression model with several independent variables.

4. Затем необходимо нажать «Calculate and transfer to main window/Рассчитать и перенести в главное окно».
5. Полученное значение автоматически перенесется в графу «Effect size/Размер эффекта», после чего остаётся только нажать «Calculate/Рассчитать» и получить готовый результат.

Рассмотрим ещё один пример расчёта объёма выборки для линейных регрессионных моделей. Предположим,

исследователи хотят отдельно оценить эффект срока гестации на массу тела новорождённого с коррекцией на длину тела новорождённого и уровень образования матери и отца.

Исследователи могут рассчитать, сколько новорождённых необходимо включить в исследование, чтобы определить даже малый размер эффекта (0,02) срока гестации на массу тела новорождённого при статистической мощности 0,8 и уровне α-ошибки 0,05. В данном примере расчёт будет проводиться для тестирования влияния одного из четырёх предикторов на зависимую переменную с коррекцией на три оставшихся предиктора.

Для этого:

1. В графе «Test family/Семейство тестов» выбираем «F tests/F-тесты» и статистический тест «Linear multiple regression: fixed model, R^2 increase/Множественная линейная регрессия: фиксированная модель, увеличение R^2».
2. Далее в графе «Type of power analysis/Тип анализа мощности» выбираем «A priori: compute required sample size — given α, power, and effect size/Априори: вычислить необходимый размер выборки — учитывая величину α-ошибки, мощность и величину эффекта».
3. В разделе «Input parameters/Входные параметры» необходимо задать параметры для расчёта:
   3.1. «Effect size/Размер эффекта» — 0,02;
   3.2. «α err prob/Уровень α-ошибки» — 0,05;
   3.3. «Power (1–β err prob)/Статистическая мощность» — 0,8;
   3.4. «Number of tested predictors/Количество тестируемых предикторов» — 1 (срок гестации);
   3.5. «Total number of predictors/Общее число предикторов» — 4.
4. Для получения результата необходимо нажать на «Calculate/Рассчитать».

Программа рассчитала, что для ответа на поставленный вопрос необходимо включить в исследование минимум 395 человек (рис. 12). С учётом увеличения объема на 15% из-за возможного отказа от исследования размер выборки увеличится до 455 человек.

Важно подчеркнуть, что подобно моделям логистической регрессии линейный анализ в G*Power также имеет наглядную визуализацию с помощью вкладки «X-Y plot for a range of values/График X-Y для диапазона значений».

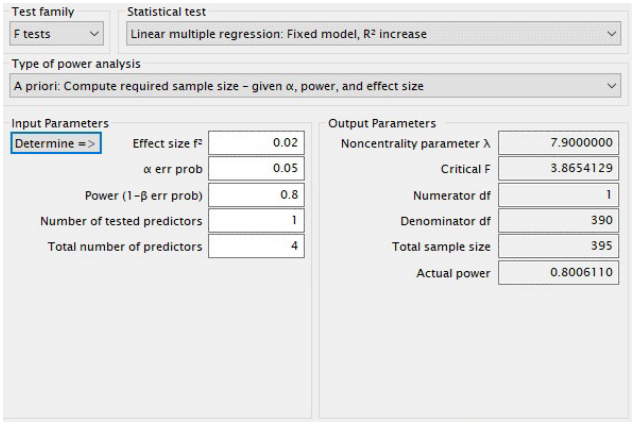

Рис 12. Диалоговое окно программы G*Power с введёнными параметрами расчёта и результатом для множественной линейной регрессионной модели.
Fig. 12. G*Power dialog box presenting entered calculation parameters and an outcome for a multiple linear regression model.

ЗАКЛЮЧЕНИЕ

В данном обзоре мы рассмотрели расчёт необходимого объёма выборки на примере поперечных исследований с использованием регрессионного анализа. Регрессионный анализ позволяет исследователям изучать независимый, или свободный от воздействия учтённых конфаундеров, эффект факторных признаков на исход. Хотя мы сделали акцент на его применении в поперечных исследованиях, линейная и логистическая регрессии применяются и в когортных, и в экспериментальных, и в исследованиях типа случай-контроль. Мы надеемся, что данная статья станет практическим руководством для студентов, аспирантов, молодых учёных и всех, кто планирует своё первое исследование в области наук о здоровье.

ДОПОЛНИТЕЛЬНО

Вклад авторов. Вклад распределён следующим образом: Н.А. Митькин — обзор литературы, сбор и анализ литературных источников, написание текста и редактирование статьи; С.Н. Драчев — обзор литературы, сбор и анализ литературных источников, написание текста и редактирование статьи; Е.А. Кригер — написание текста и редактирование статьи; В.А. Постоев — написание текста и редактирование статьи; А.М. Гржибовский — идея, написание текста, редактирование статьи и научное руководство. Все авторы подтверждают соответствие своего авторства международным критериям ICMJE. Все авторы внесли существенный вклад в разработку концепции и подготовку статьи, прочли и одобрили финальную версию перед публикацией.

Источник финансирования. Авторы заявляют об отсутствии внешнего финансирования при проведении исследования.

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

ADDITIONAL INFORMATION

Authors' contribution. All authors confirm that their authorship meets the international ICMJE criteria. All authors have made a significant contribution to the development of the concept and preparation of the article, read and approved the final version before publication. N.A. Mitkin — literature review, collection and analysis of literary sources, writing the text and editing the article; S.N. Drachev — literature review, collection and analysis of literary sources, writing the text and editing the article; E.A. Krieger — writing

the text and editing the article; V.A. Postoev — writing the text and editing the article; A.M. Grjibovski — idea, writing the text, editing the article and general scientific supervision.

Funding source. No funding.
Competing interests. The authors declare that they have no competing interests.

СПИСОК ЛИТЕРАТУРЫ

1. Холматова К.К., Горбатова М.А., Харькова О.А., Гржибовский А.М. Поперечные исследования: планирование, размер выборки, анализ данных // Экология человека. 2016. Т. 23, № 2. С. 49–56. doi: 10.33396/1728-0869-2016-2-49-56
2. Chan Y.H. Biostatistics 102: quantitative data — parametric & non-parametric tests // Singapore Med J. 2003. Vol. 44, N 8. P. 391–396.
3. Kim H.Y. Analysis of variance (ANOVA) comparing means of more than two groups // Restor Dent Endod. 2014. Vol. 39, N 1. P. 74–77. doi: 10.5395/rde.2014.39.1.74
4. Rothman K.J., Greenland S., Lash T.L. Modern epidemiology. 3rd ed. Lippincott Williams & Wilkins, 2008. 758 p.
5. Groenwold R.H., Klungel O.H., Grobbee D.E., Hoes A.W. Selection of confounding variables should not be based on observed associations with exposure // Eur J Epidemiol. 2011. Vol. 26, N 8. P. 589–593. doi: 10.1007/s10654-011-9606-1
6. Duleba A.J., Olive D.L. Regression analysis and multivariate analysis // Semin Reprod Endocrinol. 1996. Vol. 14, N 2, P. 139–153. doi: 10.1055/s-2007-1016322
7. Шарашова Е.Е., Холматова К.К., Горбатова М.А., Гржибовский А.М. Применение множественного логистического регрессионного анализа в здравоохранении с использованием пакета статистических программ SPSS // Наука и Здравоохранение. 2017. № 4. С. 5–26.
8. Agresti A. An introduction to categorical data analysis. 3rd ed. John Wiley & Sons, 2019. 400 c.
9. Cameron A., Pravin K. Regression analysis of count data. 2nd ed. 1999. doi: 10.1017/CBO9780511814365
10. Кригер Е.А., Драчев С.Н., Митькин Н.А., и др. Расчет необходимого объема выборки с использованием программы G*Power // Морская медицина. 2023. Т. 9, № 2. С. 111–125. doi: 10.22328/2413-5747-2023-9-2-111-125
11. Bewick V., Cheek, L., Ball J. Statistics review 14: logistic regression // Critical care. 2005. Vol. 9, N 1. P. 112–118. doi: 10.1186/cc3045
12. Adler N.E., Epel E.S., Castellazzo G., Ickovics J.R. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy white women // Health Psychol. 2000. Vol. 19, N 6. P. 586–592. doi: 10.1037//0278-6133.19.6.586
13. Neverlien P.O. Assessment of a single-item dental anxiety question // Acta Odontol Scand. 1990. Vol. 48, N 6. P. 365–369. doi: 10.3109/00016359009029067
14. Hsieh F.Y., Bloch D.A., Larsen M.D. A simple method of sample size calculation for linear and logistic regression // Stat Med. 1998. Vol. 17, N 14. P. 1623–1634. doi: 10.1002/(sici)1097-0258(19980730)17:14<1623::aid-sim871>3.0.co;2-s
15. Steyerberg E.W., Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation // Eur Heart J. 2014. Vol. 35, N 29. P. 1925–1931. doi: 10.1093/eurheartj/ehu207
16. Гржибовский А.М., Иванов С.В., Горбатова М.А. Однофакторный линейный регрессионный анализ с использованием программного обеспечения Statistica и SPSS // Наука и Здравоохранение. 2017. № 2. С. 5–33.
17. Ziegel E.R., Neter J., Kutner M., et al. Applied linear statistical models // Technometrics. 1997. Vol. 39, N 3. P. 342. doi: 10.2307/1271154
18. Novotny J., Bilokon P., Galiotos A., Délèze F. Machine learning and big data with kdb+/q. 2019. doi: 10.1002/9781119404729
19. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale : Lawrence Erlbaum Associates, 1998.
20. Kang H. Sample size determination and power analysis using the G*Power software // J Educ Eval Health Prof. 2021. Vol. 18. P. 17. doi: 10.3352/jeehp.2021.18.17
21. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals // JAMA. 1997. Vol. 277, N 11. P. 927–934.

REFERENCES

1. Kholmatova KK, Gorbatova MA, Kharkova OA, Grjibovski AM. Cross-sectional studies: planning, sample size, data analysis. *Ekologiya cheloveka (Human Ecology)*. 2016;23(2):49–56. (In Russ). doi: 10.33396/1728-0869-2016-2-49-56
2. Chan YH. Biostatistics 102: quantitative data — parametric & non-parametric tests. *Singapore Med J*. 2003;44(8):391–396.
3. Kim HY. Analysis of variance (ANOVA) comparing means of more than two groups. *Restor Dent Endod*. 2014;39(1):74–77. doi: 10.5395/rde.2014.39.1.74.
4. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology. 3rd ed.* Lippincott Williams & Wilkins; 2008. 758 p.
5. Groenwold RH, Klungel OH, Grobbee DE, Hoes AW. Selection of confounding variables should not be based on observed associations with exposure. *Eur J Epidemiol*. 2011;26(8):589–593. doi: 10.1007/s10654-011-9606-1
6. Duleba AJ, Olive DL. Regression analysis and multivariate analysis. *Semin Reprod Endocrinol*. 1996;14(2):139–153. doi: 10.1055/s-2007-1016322
7. Sharashova EE, Kholmatova KK, Gorbatova MA, Grjibovski AM. Multivariable logistic regression using SPSS in health research. *Science & Healthcare*. 2017;(4):5–26. (In Russ).
8. Agresti A. *An introduction to categorical data analysis. 3rd ed.* John Wiley & Sons; 2019. 400 p.
9. Cameron A, Pravin K. *Regression analysis of count data. 2nd ed.* 1999. doi: 10.1017/CBO9780511814365
10. Krieger EA, Drachev SN, Mitkin NA, et al. Sample size calculation using G*Power software. *Marine Medicine*. 2023;9(2):111–125. (In Russ). doi: 10.22328/2413-5747-2023-9-2-111-125
11. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. *Crit Care*. 2005;9(1):112–118. doi: 10.1186/cc3045

12. Adler NE, Epel ES, Castellazzo G, Ickovics JR. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy white women. *Health Psychol*. 2000;19(6):586–592. doi: 10.1037//0278-6133.19.6.586
13. Neverlien PO. Assessment of a single-item dental anxiety question. *Acta Odontol Scand.* 1990;48(6):365–369. doi: 10.3109/00016359009029067
14. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med*. 1998;17(14):1623–1634. doi: 10.1002/(sici)1097-0258(19980730)17:14<1623::aid-sim871>3.0.co;2-s
15. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925–1931. doi: 10.1093/eurheartj/ehu207
16. Grjibovski AM, Ivanov SV, Gorbatova MA. Univariate regression analysis using Statistica and SPSS software. *Science & Healthcare*. 2017;(2):5–33. (In Russ).
17. Ziegel ER, Neter J, Kutner M, et al. Applied linear statistical models. *Technometrics*. 1997;39(3):342. doi: 10.2307/1271154
18. Novotny J, Bilokon P, Galiotos A, Délèze F. *Machine learning and big data with kdb+/q*. 2019. doi: 10.1002/9781119404729
19. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates; 1998.
20. Kang H. Sample size determination and power analysis using the G*Power software. *J Educ Eval Health Prof*. 2021;18:17. doi: 10.3352/jeehp.2021.18.17
21. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *JAMA*. 1997;277(11):927–934.

ОБ АВТОРАХ

*** Митькин Никита Андреевич;**
адрес: Российская Федерация, 163061, Архангельск, Троицкий проспект, д. 51;
ORCID: 0000-0002-0027-8155;
eLibrary SPIN: 8865-1820;
e-mail: n.a.mitkin@gmail.com

Драчев Сергей Николаевич, к.м.н., PhD, доцент;
ORCID: 0000-0002-1548-690X;
eLibrary SPIN: 3879-8612;
e-mail: drachevsn@mail.ru

Кригер Екатерина Анатольевна, к.м.н., доцент;
ORCID: 0000-0001-5179-5737;
eLibrary SPIN: 2686-7226;
e-mail: kate-krieger@mail.ru

Постоев Виталий Александрович, к.м.н., PhD, доцент;
ORCID: 0000-0003-4982-4169;
eLibrary SPIN: 6070-2486;
e-mail: ispha@nsmu.ru

Гржибовский Андрей Мечиславович, PhD;
ORCID: 0000-0002-5464-0498;
eLibrary SPIN: 5118-0081;
e-mail: a.grjibovski@yandex.ru

* Corresponding author / Автор, ответственный за переписку

AUTHORS' INFO

*** Nikita A. Mitkin;**
address: 51 Troickij avenue, 163061 Arhangel'sk, Russian Federation;
ORCID: 0000-0002-0027-8155;
eLibrary SPIN: 8865-1820;
e-mail: n.a.mitkin@gmail.com

Sergei N. Drachev, MD, Cand. Sci. (Med.), MPH, PhD, Associate Professor;
ORCID: 0000-0002-1548-690X;
eLibrary SPIN: 3879-8612;
e-mail: drachevsn@mail.ru

Ekaterina A. Krieger, MD, Cand. Sci. (Med.), MPH, Associate Professor;
ORCID: 0000-0001-5179-5737;
eLibrary SPIN: 2686-7226;
e-mail: kate-krieger@mail.ru

Vitaly A. Postoev, MD, Cand. Sci. (Med.), MPH, PhD, Associate Professor;
ORCID: 0000-0003-4982-4169;
eLibrary SPIN: 6070-2486;
e-mail: ispha@nsmu.ru

Andrej M. Grjibovski, MD, MPhil, PhD;
ORCID: 0000-0002-5464-0498;
eLibrary SPIN: 5118-0081;
e-mail: a.grjibovski@yandex.ru