

Процедуры вычисления минимального необходимого объёма выборки для планируемых исследований и их программная реализация

М. Н. Нессонова, Е. Ю. Литвиненко

Национальный фармацевтический университет, Харьков, Украина

Резюме

Введение. Расчёт необходимого размера выборки является неотъемлемой частью планируемых исследований, однако большая часть рассматриваемых в литературе методов и реализованных в статистическом ПО процедур предназначены для определения размера выборки для анализа количественных показателей, в то время как задачи анализа качественных признаков остаются недостаточно исследованными.

Цель работы. Обзор методов расчёта минимально необходимого объёма выборки для отдельных задач анализа качественных признаков и разработка программного средства, позволяющего осуществлять указанные расчёты.

Объект и методы. Системный анализ специализированной литературы позволил определить наиболее адекватные методы расчёта объёма выборки с учётом специфики задач исследования и его исходных данных. Программное обеспечение разрабатывалось средствами IDLE Python 3.4.4.

Результаты. Рассмотрены методы расчёта необходимого размера выборки для задач, связанных с анализом частоты встречаемости качественных дихотомических признаков в генеральной совокупности. Представлено разработанное авторами программное обеспечение, позволяющее автоматизировать решение указанных задач, которое в частности будет полезно при планировании и анализе данных клинических и доклинических испытаний.

Ключевые слова: расчёт размера выборки; качественные признаки; анализ частот.

Клін. інформат. і Телемед. 2017. Т.12. Вип.13. с.14–19

1. Введение

Определение необходимого размера выборки — важная часть планирования исследований, направленных на проверку и подтверждение формальных гипотез. Особенно актуален этот вопрос для обоснования экономической эффективности исследования и планирования его финансирования, что является неотъемлемой частью государственных и выполняемых по грантам НИР, маркетинговых исследований, клинических и доклинических испытаний, обеспечения стандартов контроля качества. Большая часть методических рекомендаций, соответствующих инструкций и предлагаемых в литературе подходов, а также процедур, реализованных в специализированных компьютерных программах, концентрируется на методах планирования объёма выборки для задач сравнения двух групп по некоторому показателю, который, как правило, является количественным. В то же время, задачи анализа частоты качественного (номинального) признака в одной выборке рассматриваются в литературе и реализуются в специализированном программном обеспечении (ПО) недостаточно полно.

Целью работы является обзор методов расчёта минимально необходимого объёма выборки для задач по оценке частоты встречаемости дихотомического признака в генеральной совокупности, а также представление разработанного авторами программного средства, позволяющего осуществлять эти расчёты.

2. Материалы и методы

При обзоре методов анализировалась специализированная литература [1–8], проведён системный анализ доступной информации по исследуемой тематике. Программное обеспечение для реализации процедур расчёта размера выборки написано средствами интегрированной среды разработки (Integrated Development Environment, IDLE) Python 3.4.4 с использованием методов GUI, реализованных в библиотеке Tkinter.

3. Результаты и обсуждение

3.1. Анализ и выбор адекватных процедур расчёта размера выборки

Наиболее часто в доступной литературе для вычисления объёма выборки, необходимого, чтобы оценить ожидаемую частоту номинального бинарного признака в генеральной совокупности с заданной точностью и при заданной доверительной вероятности, предлагается формула:

$$n = \left[Z_{\alpha/2}^2 \cdot \frac{p \cdot (1-p)}{I^2} \right], \quad (1)$$

где n — искомый объём выборки; p — ожидаемая частота встречаемости признака в генеральной совокупности; I — половина

ширины доверительного интервала (точность оценки частоты); $Z_{\alpha/2}$ – квантиль стандартного нормального распределения порядка $\alpha/2$, где α – предельно допустимый уровень ошибки первого рода (определяется как единица минус доверительная вероятность); [...] – операция округления до ближайшего большего целого числа.

Данная формула, использующая нормальную аппроксимацию биномиального распределения, в большинстве источников приводится как единственный вариант для расчёта необходимого количества наблюдений при оценке ожидаемой частоты встречаемости дихотомического признака в одной выборке. В большинстве версий специализированного статистического ПО это также единственный способ рассчитать необходимый размер выборки для оценки одной частоты. При этом предполагается, что ожидаемая частота (p) известна исследователю заранее, а при отсутствии информации о ней рекомендуется использовать формулу (1) с параметром $p=0.5$. Следует однако отметить, что нормальная аппроксимация, на которой основан данный способ вычисления, является удовлетворительной при $n \cdot p \cdot (1-p) > 5$ и $p \in [0.1; 0.9]$, а наилучшая сходимость обеспечивается при ожидаемой частоте, равной 0.5. Т. е. для обеспечения надёжности оценок, полученных по формуле (1), необходим достаточно большой объём наблюдений, кроме того, ожидаемая частота не должна быть слишком маленькой или слишком большой. Таким образом, соотношение (1) даёт наиболее адекватную оценку необходимого размера выборки в случае очень большого размера генеральной совокупности или неизвестного её размера, когда он предполагается достаточно большим, и неизвестной ожидаемой частоте. Оно также применимо для значений ожидаемых частот из диапазона от 0.1 до 0.9, но не является адекватным в случаях, когда необходимо оценить ожидаемые частоты, близкие к 0 или к 1, а именно при $p < 1/(n+1)$ или $p > n/(n+1)$.

При выполнении условий применимости аппроксимации биномиального распределения нормальным и известном размере генеральной совокупности необходимое количество наблюдений может быть гораздо меньше, чем найденное по формуле (1), если для его определения использовать соотношение:

$$n = \left\lceil \left(\frac{1}{N} + \frac{4 \cdot I^2}{Z_{\alpha/2}^2 \cdot p \cdot (1-p)} \right)^{-1} \right\rceil \quad (2)$$

где N – известный размер генеральной совокупности.

Если задача исследователя состоит не в получении доверительного интервала для частоты, а в подтверждении гипотезы о том, что частота встречаемости интересующего его признака в генеральной совокупности не больше ожидаемой, то для вычисления минимального необходимого объёма выборки при неизвестном или очень большом размере генеральной совокупности следует воспользоваться формулой:

$$n = \left\lceil \frac{\ln \alpha}{\ln(1-p)} \right\rceil \quad (3)$$

а при известном размере генеральной совокупности использовать формулу:

$$n = \left\lceil \left(1 - N^p \sqrt[p]{\alpha} \right) \cdot \left\{ \frac{1 + N \cdot (1-p)}{2} \right\} \right\rceil \quad (4)$$

где $\{...\}$ – целая часть числа.

Формулы (3) и (4) также считаются более предпочтительными в случаях, когда необходимо оценить ожидаемую частоту, которая близка к 0 или к 1 (100%).

3.2. Алгоритм программной реализации

В разработанном ПО рассмотренные выше формулы расчёта размера выборки используются в зависимости от параметров решаемой задачи. Выбор способа расчёта осуществляется по алгоритму, показанному на рис. 1, на основании диалога с пользователем, определяющим цели и задачи своего эксперимента и задающего известные параметры.

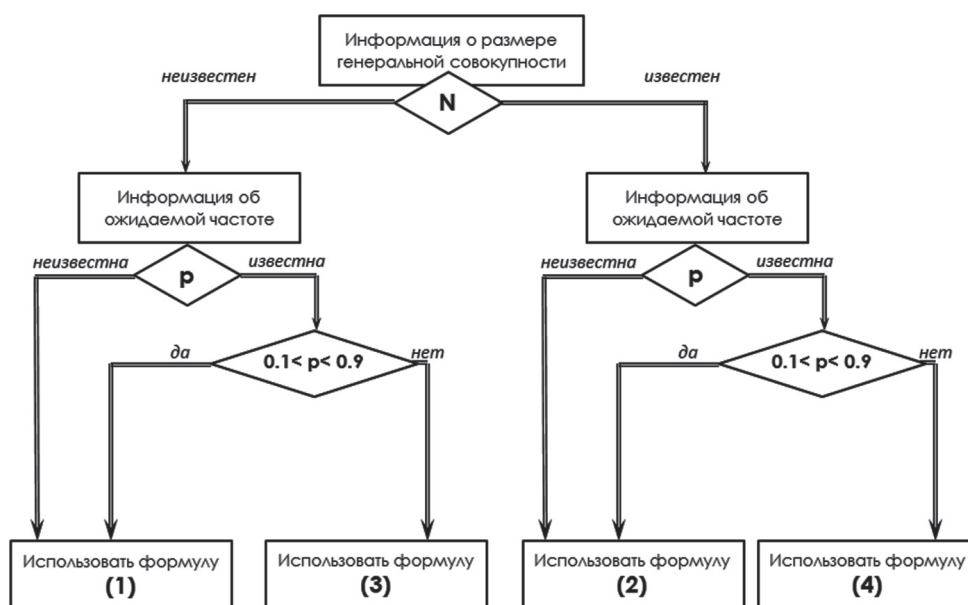


Рис. 1. Алгоритм выбора способа расчёта минимально необходимого размера выборки для оценки встречаемости дихотомического признака в одной выборке.

3.3. Результаты применения разработанного программного средства

Нижче обговорюється застосування запропонованого ПО на прикладах рішення деяких завдань, характерних для медичних досліджень.

Приклад 3.3.1. Нехай необхідно оцінити поширеність деякого захворювання (наприклад, вірусу герпесу) серед жителів певного регіону (наприклад, Харківської області). За даними ВОЗ відомо, що поширеність цього захворювання в світі становить 65%. Скільки

необхідно спостережень, щоб оцінити цю частоту з точністю $\pm 7\%$ при довірчій ймовірності 95%? Дане завдання відповідає випадку невідомого або дуже великого об'єму генеральної сукупності і відомої очікуваної частоти. Діалогові вікна з встановленими параметрами умов завдання показані на рис. 2 (а, б). Розрахунок необхідного розміру вибірки в даному випадку проводиться за формулою (1). Як бачимо, для того, щоб при довірчій ймовірності 95% оцінити вказану очікувану частоту з точністю $\pm 7\%$, необхідно провести дослідження як мінімум 179 пацієнтів (рис. 3).

Приклад 3.3.2. Припустимо, що в деякому клінічному дослідженні необхідно оцінити об'єм вибірки,

a)

б)

Рис. 2. Вхідні параметри задачі об оцінці поширеності захворювання (приклад 3.3.1): а) вибір довірчій ймовірності, вказання свідчень про наявності інформації про розмір генеральної сукупності і очікуваної частоти; б) введення відомого значення очікуваної частоти і вимоги точності її оцінки.

Summary of the task :

So, you want to estimate the proportion of a dichotomous feature occurrence in the population of unknown or extremely large size at 95 % confidence probability level. The proportion is expected to be equal to 65.0 %, and the desired precision of estimate is 7.0 %.

Start Again

=>

Recommendations :

The number of observations you need to estimate the expected proportion with desired precision :

n = 179

Finish

Рис. 3. Вікно виводу для задачі об оцінці поширеності захворювання (приклад 3.3.1).

необходимый для того, чтобы с доверительной вероятностью 99% подтвердить гипотезу о том, что частота побочных реакций при применении исследуемого лекарственного средства не превышает 2.5%. Если использовать терминологию математической статистики, задача состоит в расчёте размера выборки при неизвестном размере генеральной совокупности и известной ожидаемой частоте. Так как значение ожидаемой частоты в данном примере очень мало, для расчёта необходимого размера выборки необходимо использовать формулу (3), в соответствии с которой имеем, что необходимо наблюдать как минимум 182 случая отсутствия побочных реакций, чтобы подтвердить проверяемую гипотезу (рис. 4).

Пример 3.3.3. Проводится исследование иммунологической эффективности некоторой вакцины против вируса гриппа на 200 добровольцах. При доверительной вероятности 90% необходимо подтвердить эффективность вакцинации на уровне 91 ÷ 95%. Какое количество заболевших гриппом допустимо среди привитых добровольцев при вспышке заболевания? В терминах статистики задача состоит в расчёте размера выборки при известном размере генеральной совокупности (200 наблюдений) и известной ожидаемой частоте (93%), которую в данном примере необходимо оценить с точностью $\pm 2\%$. Так как значение ожидаемой частоты в данном примере достаточно велико, для расчёта необходимого размера выборки необходимо использовать формулу (4), в соответствии с которой имеем, что максимальное количество привитых добровольцев, заболевших гриппом, не опровергающее гипотезу о требуемом уровне эффективности вакцинации, составляет 20 человек (рис. 5).

Рассмотренный ряд примеров, не являясь исчерпывающим, иллюстрирует круг задач, возможности и сферу применения разработанного ПО для оценки размера выборки при анализе частоты встречаемости дихотомических номинальных признаков. Задачи, аналогичные примеру 3.3.1, часто решаются на этапе планирования клинических испытаний лекарственных средств, при планировании социологических опросов и маркетинговых исследований. Ситуация из примера 3.3.3 аналогична стандартной задаче промышленного контроля качества, когда необходимо оценить долю годных/бракованных изделий в партии. Кроме того подобные расчёты, как и аналогичные примеру 3.3.2, необходимы при проведении доклинических испытаний и других вариантов пилотных исследований.

4. Заключение

Предлагаемое ПО для определения необходимого размера выборки при различных вариантах исследований, связанных с оценкой частоты встречаемости качественного бинарного признака в генеральной совокупности, будет полезно при планировании клинических и доклинических испытаний, исследовании рынков, для медицинских и инженерных исследований, специалистов по контролю качества в их профессиональной деятельности. Простота созданного графического интерфейса позволяет использовать программу даже самым неискущённым в области ИТ пользователям. В качестве одного из перспективных направлений наших разработок следует обозначить расширение возможностей программы для других

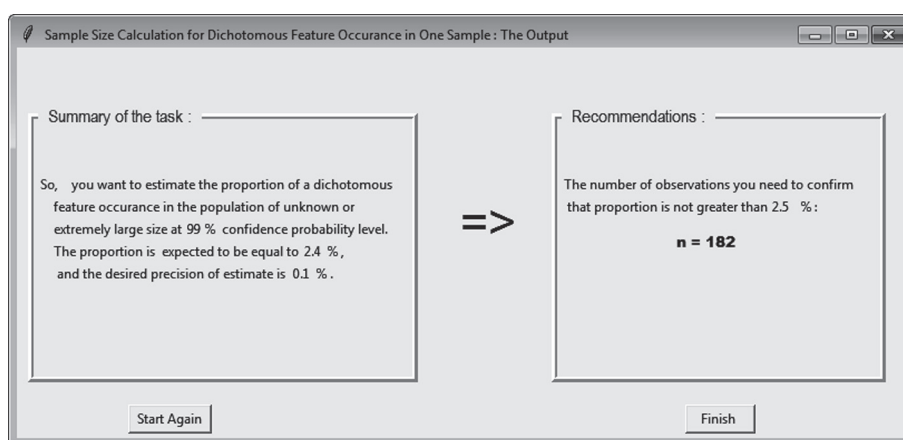


Рис. 4. Окно вывода для задачи об оценке доли побочных реакций (пример 3.3.2).

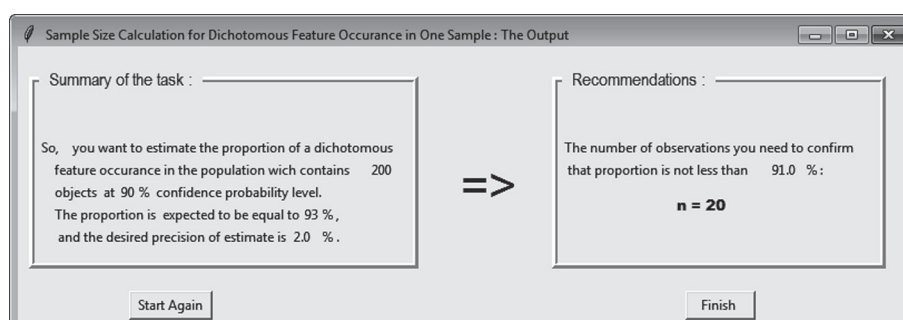


Рис. 5. Окно вывода для задачи об оценке иммунологической эффективности вакцины (пример 3.3.3).

задач планирования объёма выборочных исследований, связанных с анализом качественных признаков (показателей, принимающих более двух возможных значений, межгрупповых сравнений качественных признаков, оценки значимости и силы их взаимосвязи).

Исследования проводились с соблюдением национальных норм биоэтики и положений Хельсинкской декларации (в редакции 2013 г.). Авторы статьи — М. Н. Нессонова, Е. Ю. Литвиненко — подтверждают, что у них нет конфликта интересов.

Литература

1. Кобзарь А. И. Прикладная математическая статистика. М. ФИЗМАТЛИТ, 2012. 816 с.
2. Dell R. B., Holleran S., Ramakrishnan R. Sample Size Determination. *ILAR Journal*, 2002, vol. 43(4), pp. 123–131.
3. Bland M. Расчет размера выборки, доступ: <http://pubhealth.spb.ru/COPC/STATSH/ssize.htm>.
4. Елисеева И. И., Юзбашев М. М. Общая теория статистики. М. Финансы и статистика, 2001. 480 с.
5. Blyth C. R. Approximate binomial confidence limits. *JASA*, 1986, vol. 81, pp. 843–855.
6. Green J. Asymptotic sample size for given confidence intervals length. *JRSS, Sec. C*, 1982, vol. 31, pp. 298–300.
7. Румянцев П. О., Саенко В. А., Румянцева У. В., Чекин С. Ю. Статистические методы анализа в клинической практике, доступ: <https://www.kantiana.ru/medicinal/help/StatMethodsInClinics.pdf>.
8. Койчубеков Б.К., Сорокина М.А., Мхитарян К.Э. Определение размера выборки при планировании научного исследования. *Международный журнал прикладных и фундаментальных исследований*, 2014, №4, сс. 71–74.

Процедури обчислення мінімального необхідного обсягу вибірки для планованих досліджень та їх програмна реалізація

М. М. Нессонова, Є. Ю. Литвиненко

Національний фармацевтичний університет, Харків, Україна

Резюме

Вступ. Розрахунок необхідного обсягу вибірки є невід'ємною частиною досліджень, що плануються, однак більша частина методів, які розглядаються в літературі, і процедур, що реалізовані у статистичному ПЗ, призначена для визначення розміру вибірки для аналізу кількісних показників, у той час як задачі аналізу якісних ознак лишаються недостатньо дослідженими.

Мета роботи. Огляд методів розрахунку мінімально необхідного обсягу вибірки для окремих задач аналізу якісних ознак і розробка програмного засобу, що дозволяє здійснювати вказані розрахунки.

Об'єкт і методи. Системний аналіз спеціалізованої літератури дозволив визначити найбільш адекватні методи розрахунку обсягу вибірки з урахуванням специфіки задач дослідження і його вхідних даних. Програмне забезпечення розроблялося засобами IDLE Python 3.4.4.

Результати. Розглянуто методи розрахунку необхідного розміру вибірки для задач, пов'язаних з аналізом частоти зустрічальності якісних дихотомічних ознак у генеральній сукупності. Представлено розроблене авторами програмне забезпечення, яке дозволяє автоматизувати розв'язання вказаних задач, яке зокрема буде корисним при плануванні та аналізі даних клінічних і доклінічних випробувань.

Ключові слова: розрахунок розміру вибірки; якісні ознаки; аналіз частот.

Calculation procedures for sample size determination for planned studies, and their program implementation

M. N. Nessonova, Y. Y. Lytvynenko
National University of Pharmacy, Kharkiv, Ukraine
e-mail: saddymn@gmail.com

Abstract

Introduction. Sample size determination is an inherent integral part of planned studies, but the majority of methods considered in literature and realized in statistical software is devoted to evaluation of sample size in a case of quantitative feature under analysis, while the tasks of qualitative features analysis are studied insufficiently.

The purpose of the work is to give a survey of minimal necessary sample size determination methods for certain tasks of qualitative variables analysis; and to develop the software for automation of calculations under consideration.

Object and Methodology. System analysis of specialized literary and internet sources gives ability to determine proper and relevant procedures, which are acceptable for sample size evaluation subject to the traits of the research and its input data. The software was developed in IDLE Python 3.4.4.

Results. Methods of sample size calculation in tasks, connected with analysis of dichotomous feature occurrence in population, are studied. The software developed by the authors is presented, which allows automation of the tasks under consideration and can be particularly useful in planning and analysis of clinical and preclinical trials.

Key words: Sample size calculation; Qualitative features; Analysis of frequencies.

©2017 Institute Medical Informatics and Telemedicine Ltd, ©2017 Ukrainian Association of Computer Medicine, ©2017 Kharkiv medical Academy of Postgraduate Education. Published by Institute of Medical Informatics and Telemedicine Ltd. All rights reserved.

ISSN 1812-7231 *Klin.inform.telemed.* Volume 12, Issue 13, 2017, Pages 14–19

http://kit-journal.com.ua/en/index_en.html

References (8)

References

1. Kobzar' A. I. *Applied mathematical statistics* [Prikladnaya matematicheskaya statistika]. *FIZMATHLIT* [Physical and Mathematical Literature], Moscow, 2012. 816 p. (In Russ.).
2. Dell R. B., Holleran S., Ramakrishnan R. Sample Size Determination. *ILAR Journal*, 2002, vol. 43(4), pp. 123–131.
3. Bland M. Calculation of Sample Size, available at: <http://pub-health.spb.ru/COPC/STATSH/ssize.htm>. (In Russ.).
4. Yeliseeva I. I., Yuzbashev M. M. *General statistical theory* [Obshaya teoriya statistiki]. *Financy i statistika* [Finance and Statistics], Moscow, 2001. 480 p. (In Russ.).
5. Blyth C. R. Approximate binomial confidence limits. *JASA*, 1986, vol. 81, pp. 843–855.
6. Green J. Asymptotic sample size for given confidence intervals length. *JRSS, Sec. C*, 1982, vol. 31, pp. 298–300.
7. Rumiantsev P. O., Saenko V. A., Rumiantseva U. V., Chekin S. Yu. Statistical methods of analysis in clinical practice, available at: <https://www.kantiana.ru/medicinal/help/Stat-MethodsInClinics.pdf>. (In Russ.).
8. Koichubekov B. K., Sorokina M. A., Mhitaryan K. E. Sample size determination in planning scientific research. *Mezhdunarodnyi zhurnal prikladnyh i fundamental'nyh issledovaniy* [International journal of applied and fundamental studies], 2014, no. 4, pp. 71–74. (In Russ.).

Переписка

К.Т.Н. **М. Н. Нессонова**
Национальный фармацевтический университет
кафедра фармакоинформатики
ул. А. Невского, 18, Харьков, 61140, Украина
тел.: +380 (57) 771 81 52
эл. почта: saddymn@gmail.com