

УДК 519.254

А.Н. Герасимов,

д-р физ.-мат. наук, доц., зав. кафедрой медицинской информатики и статистики Первого МГМУ имени И.М. Сеченова Минздрава России

М.Ю. Сердюкова,

ст. преподаватель кафедры медицинской информатики и статистики Первого МГМУ имени И.М. Сеченова Минздрава России

A.N. Gerasimov,

Dr. Phys.-math. Sciences, Assoc. prof., head. the Department of medical Informatics and statistics of the First Sechenov Moscow State Medical University

M.Yu. Serdyukova,

Senior lecturer of the Department of medical Informatics and statistics of the state budget educational institution of higher professional education the First Sechenov Moscow State Medical University

ПРОВЕРКА ПРИМЕНИМОСТИ МЕТОДОВ ПАРАМЕТРИЧЕСКОЙ СТАТИСТИКИ ДЛЯ АНАЛИЗА МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ

PARAMETRIC STATISTICS APLICABILITY TO BIO-MEDICAL DATA ANALYSIS

КОНТАКТНАЯ ИНФОРМАЦИЯ:

Герасимов Андрей Николаевич, д-р физ.-мат. наук, доц., зав. кафедрой медицинской информатики и статистики Первого МГМУ имени И.М. Сеченова Минздрава России

Адрес: 119991, Москва, ул. Трубецкая, д. 8, стр. 2

Телефон: + 7 905 550-50-84

e-mail: andr-gerasim@yandex.ru

Статья поступила в редакцию: 29.02.2016 г.

Статья принята к печати: 25.04.2016 г.

CONTACT INFORMATION:

Andrey Gerasimov, Dr. Phys.-math. Sciences, Assoc. Prof., head. the Department of medical Informatics and statistics of the First Sechenov Moscow State Medical University

Address: p. 8, bld. 2, Trubetskaya str., Moscow, Russia, 119991

e-mail: andr-gerasim@yandex.ru

The article received: February 29, 2016.

The article approved for publication: April 25, 2016.

Аннотация. Методы параметрической статистики рассчитывают статистическую значимость различий в предположении о нормальности изучаемых случайных величин. Так как все реально используемые в медико-биологических исследованиях случайные величины не являются нормально распределенными, то необходимо проверять корректность использования методов параметрической статистики, т. е. оценивать величину погрешности при вычислении статистической значимости различий. К сожалению, сложилась плохая практика, при которой данную проверку производят неправильно. В статье изложены причины, приводящие к погрешностям при использовании параметрических методов исследования, и предложена программа, рассчитывающая величину погрешности, полученную вследствие ненормальности эмпирического распределения среднего из набора наблюдений.

Abstract. Parametric statistics calculates the statistical significance of differences assuming normality of the random variables in question. Since all random variables used in real biomedical research are not normally distributed, it is necessary to validate parametric statistical methods, i.e. evaluate the errors in the calculation of the statistical significance of differences. Unfortunately this test is performed incorrectly. The article describes the causes of errors in parametric statistics and presents a software program which calculates the error value resulting from the abnormal empirical distribution of the mean from a set of observations.

Ключевые слова. Статистический анализ данных, параметрические методы статистики, критерий Стьюдента, дисперсионный анализ.

Keywords. Statistical analysis of data, parametric methods of statistics, Student's t test, analysis of variance.

За последние годы наблюдается повышение культуры проведения статистического анализа медико-биологических данных [1]. Помимо ужесточения

требований к дизайну исследований и расширению списка используемых методов в практику вошла и проверка применимости методов параметрической

статистики. К сожалению, проверка эта обычно проводится совершенно неправильно [2].

Многие авторы, обратившись к книгам по математике, выясняют, что параметрические методы можно применять, строго говоря, только если наблюдаемые случайные величины распределены нормально. Из этого они делают вывод, что для проверки применимости методов параметрической статистики надо проверить имеющийся набор наблюдаемых значений на нормальность распределения. Если статистически достоверных отличий эмпирического распределения от нормального получено не будет, то методы параметрической статистики использовать можно, а если отличия будут достоверны, то нельзя.

При кажущейся разумности подобных рассуждений они неверны. Дело в том, что проверить имеющиеся эмпирические распределения на нормальность нельзя. Можно проверить их на ненормальность. Из того, что статистически достоверных различий получить не удалось, не следует, что их нет. Любой реально наблюдаемый медико-биологический показатель распределен не в точности нормально, и «проверка на нормальность» говорит лишь об одном — достаточно ли количества наблюдений, чтобы эти отличия зафиксировать. При увеличении числа наблюдений эти различия обязательно станут статистически достоверными.

Чтобы доказать, что наблюдаемые показатели не распределены нормально, вообще не надо проводить какие-то статистические проверки. Дело в том, что любая нормально распределенная случайная величина с ненулевой вероятностью принимает значения из любого диапазона. Однако большинство актуальных для медико-биологических исследований показателей не могут принимать отрицательные значения. Следовательно, они распределены ненормально.

Основой для практического применения методов параметрической статистики служит то, что «основной» частью оценки большинства параметров является сумма отдельных наблюдений и «центральная предельная теорема» [3], говорящая о том, что по мере увеличения числа независимых наблюдений распределение их суммы стремится к нормальному распределению. А если сумма наблюдений вне зависимости от того, как была распределена исходная случайная величина, распределена практически нормально, то мы можем работать с ней так, как будто исходная случайная величина тоже распределена нормально. В этом и состоит суть «параметрических методов статистики»: если все равно, каково было исходное распределение, то мы можем делать вид, что оно — нормально [2].

Из-за этого все методы параметрической статистики являются приближенными, статистическую достоверность различий они вычисляют с неко-

торой погрешностью, которая уменьшается при увеличении числа наблюдений. И нужно проверять их корректность — достаточно ли наблюдений для того, чтобы эта погрешность была достаточно мала.

К сожалению, для большинства работающих в медицине «центральная предельная теорема» известна в упрощенной формулировке: «если число независимых наблюдений — 30 или более, то погрешность вычисления достоверности различий мала и параметрическими методами статистики пользоваться можно».

В подобной формулировке есть сразу два изъяна. Во-первых, не уточняется, какова величина этой самой погрешности. Во-вторых, постулируется число 30 как необходимое и достаточное число наблюдений, тогда как это неверно. Да, действительно, для любой случайной величины с любыми конечными средним арифметическим и дисперсией распределение суммы независимых наблюдений сходится к нормальному распределению при увеличении числа наблюдений, но никаких общих оценок скорости сходимости сделать нельзя. Скорость сходимости индивидуальна, зависит от формы распределения. Когда-нибудь сойдутся все распределения, но для любого числа наблюдений найдутся те, у которых сумма наблюдений еще далека от нормального распределения, будь это сто наблюдений, тысяча, миллион или миллиард.

Есть оценки скорости сходимости, основанные на величине эксцесса [4], так что оценкой в 30 наблюдений можно пользоваться для тех случайных величин, у которых небольшой коэффициент эксцесса, однако и для них величина погрешности оценки достоверности различий остается открытой.

Проверка корректности используемых методов особенно актуальна при анализе медицинских данных по двум причинам:

Многие показатели (в частности, биохимические) распределены крайне асимметрично и некомпактно.

Увеличение объема наблюдений затруднено или невозможно.

Для экспериментальных исследований увеличение количества испытуемых или животных сверх минимума неприемлемо по этическим соображениям. При изучении редких заболеваний приходится анализировать малые группы, потому что увеличить их нельзя в принципе. Даже при больших по объему исследованиях обязательно появляются небольшие группы (например, умершие или больные с определенным набором признаков), которые надо сравнивать с остальными. Поэтому стандартный рецепт «увеличивайте объем наблюдений» в этих случаях неприменим.

Однако совершенствование вычислительной техники, позволяющее рассчитать эмпирическое рас-

пределение суммы наблюдений, позволяет получить необходимые оценки.

Ошибки расчета достоверности различий при использовании методов параметрической статистики могут быть следствиями следующих факторов:

- отличия распределения эмпирического среднего от нормального;
- отличия эмпирического распределения от истинного;
- невыполнения требований центральной предельной теоремы.

Для проверки величины отличия эмпирического среднего от нормального среднего была создана программа, выложенная по адресу <http://1mgmu.com/progil/www-norma.aspx>.

Рассмотрим пример (рис. 1). Пусть у нас есть данные по длительности пребывания в РАО после операции на дуге аорты у 67 пациентов [5]. Введем их в поле для ввода указанной программы, в качестве разделителя будем использовать знак пробела. В результате программа рассчитает статистические параметры введенного эмпирического распределения (среднее арифметическое, моду, медиану, квартили, среднее квадратичное отклонение, статистическую погрешность среднего, коэффициенты асимметрии и эксцесса), а также построит частотную гистограмму и гистограмму частоты нарастающим итогом.

Справа на диаграмме частоты нарастающим итогом также приведена функция распределения для самой близкой нормально распределенной случайной величины. Максимальная разность между эмпирической функцией распределения и нормальным распределением составляет около 0,28.

Средняя длительность пребывания составила 7,55 дня при разбросе от 0 до 76 дней. Так как коэффициент эксцесса для изучаемого распределения составил 18,4, то даже при числе наблюдений, более чем вдвое превышающем стандартный минимум в 30, мы не можем быть уверены в коррект-

ности использования методов параметрической статистики.

При расчете распределения суммы 67 наблюдений, из полученного эмпирического наблюдения, получим, что распределение по-прежнему достаточно далеко от нормального — максимальное различие составляет около 0,036.

Подобная величина погрешности при расчете достоверности различий p недопустимо велика даже при работе с уровнем значимости $p = 0,05$. Однако максимальное различие между фактическим распределением суммы и его приближением к нормальному распределению может быть вблизи середины распределения, тогда как для определения достоверности различий используются вероятности больших отклонений. В связи с этим программа рассчитывает также максимальные различия между фактическим распределением и нормальным распределением суммы для отклонений от среднего арифметического: больше среднего квадратичного отклонения (σ), двух, трех, четырех и пяти среднего квадратичных отклонений.

Для нормального распределения вероятность отклонения более чем на 2σ равна 0,0455..., т. е. рассчитанная величина погрешности вычисления p актуальна для выводов о статистически значимых различиях при $p < 0,05$. Вероятность отклонения нормального распределения более чем на 3σ равна 0,0027..., т. е. рассчитанная величина погрешности вычисления p актуальна для выводов о статистически значимых различиях при $p < 0,001$.

Так, для данного примера для отклонения на два среднего квадратичных отклонения максимальная разность между фактическим и нормальным распределениями суммы составила 0,016. Отклонение на два среднего квадратичных отклонения или более соответствует проверке статистических гипотез с $p = 0,05$, следовательно, если при проверке достоверности отличий мы получили достоверность отличий $p = 0,021$, то реально мы имеем статистическую достоверность отличий $p < 0,037$.

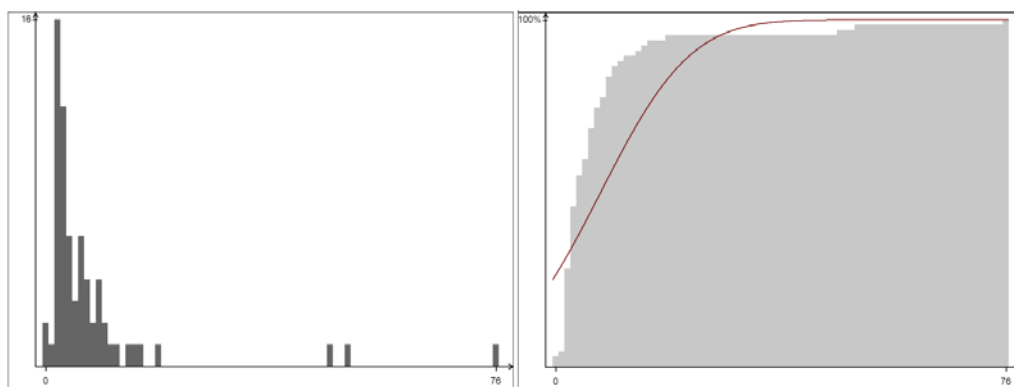


Рис.1. Частотная диаграмма и диаграмма частоты нарастающим итогом для эмпирического распределения длительности пребывания пациентов в РАО

Для отклонений на три среднеквадратичных отклонения или более мы имеем разность между эмпирическим и нормальным распределениями суммы, равную 0,0035. Однако разность в три среднеквадратичных отклонения соответствует $p = 0,001$, следовательно, если мы получаем достоверность различия в $p < 0,001$, то верить мы ей не можем, так как погрешности расчета достоверности различий в несколько раз больше полученной достоверности.

Программа рассчитывает таблицу максимального различия для всех величин суммирования от 1 до числа наблюдений. Приведем ее в сокращенном варианте, исключив часть строк (табл. 1).

Таблица 1

Максимальная величина отклонения эмпирической функции распределения суммы нескольких наблюдений длительности пребывания в РАО от нормального распределения

Число суммирования	Отклонения					
	Все	> s	> 2s	> 3s	> 4s	> 5s
1	0,2876	0,1040	0,0444	0,0444	0,0149	0,0149
5	0,2242	0,1492	0,0430	0,0124	0,0049	0,0011
10	0,1547	0,1025	0,0260	0,0112	0,00272	0,00049
15	0,1003	0,0777	0,0226	0,0088	0,00185	0,00030
20	0,0656	0,0618	0,0226	0,0074	0,00147	0,00020
25	0,0577	0,0506	0,0225	0,0065	0,00120	0,00015
30	0,0547	0,0424	0,0221	0,0060	0,00101	0,00012
35	0,0507	0,0363	0,0216	0,0054	0,00088	0,00010
40	0,0468	0,0316	0,0211	0,0051	0,00079	0,00008
45	0,0438	0,0280	0,0206	0,0047	0,00071	0,00007
50	0,0415	0,0253	0,0196	0,0044	0,00065	0,00006
55	0,0395	0,0232	0,0188	0,0042	0,00061	0,00005
60	0,0378	0,0217	0,0176	0,0040	0,00056	0,00005
67	0,0357	0,0201	0,0165	0,0038	0,00050	0,00004

Из таблицы следует, что при определении достоверности различий при $p = 0,05$ расчет достоверности отличий будет достаточно корректен и при уменьшении размера группы до стандартных 30. Тогда как для достоверности различий с $p = 0,01$ и менее, 67 наблюдений не хватает и группа должна быть большего размера.

Так как реализованная в программе методика требует достаточно большого количества вычислений, а сама программа для удобства использования размещена на сервере, то есть некоторые техни-

ческие ограничения: в качестве исходных данных надо вводить только целые неотрицательные числа. Подробная инструкция по использованию этой программы, включая указания по тому, как ускорить проведение анализа, выложена в Интернет, ссылка на инструкцию открывается после запуска программы.

Вторая причина, приводящая к неточности при расчете p , — отличие эмпирической функции распределения от истинной.

Рассмотрим пример. Пусть 97,5% наблюдаемых пациентов — лица с умеренно повышенным артериальным давлением, со средним систолическим артериальным давлением 140 мм рт. ст. и компактным распределением, а 2,5% — лица со злокачественной гипертензией и с систолическим артериальным давлением 280 мм рт. ст. В этом случае среднее систолическое артериальное давление составит 143,5 мм рт. ст.

Пусть, к примеру, мы изучаем длительность пребывания больных после операции в отделении реанимации [5]. Тогда, если перевод больных из реанимации принимается, в том числе и с учетом имеющихся возможностей отделения, то длительности пребывания разных больных в реанимации будут коррелировать друг с другом отрицательно, так как длительное пребывание одного больного вынуждает сократить длительность пребывания других. Если же увеличение пребывания в реанимации связано с наличием внутрибольничных инфекций, то заболевание одних пациентов будет увеличивать вероятность заболевания других, и, следовательно, длительности пребывания в реанимации разных больных будут коррелировать положительно.

Выводы:

1. Параметрические методы статистики — приближенные, они определяют достоверность (статистическую значимость) различий p с погрешностью, величина которой зависит как от числа наблюдений, так и от характера распределения наблюдаемой случайной величины.

2. Для использования методов параметрической статистики в медицине нужно проверять их корректность.

3. Проверка исходного распределения на нормальность не имеет отношения к проверке корректности использования методов параметрической статистики.

4. Предложена программа, с помощью которой можно точно рассчитать величину погрешности при вычислении p , возникающую из-за ненормальности распределения суммы наблюдений эмпирического распределения.

Статья подготовлена при поддержке гранта РФФИ 15-07-06947.

Список литературы

1. Брико Н.И., Покровский В.И. Эпидемиологические исследования, клиническая эпидемиология и доказательная медицина // Сеченовский вестник. 2010; 2(2): 106-111.
[Briko N. And., Pokrovskiy V.I. Epidemiological studies, clinical epidemiology and evidence-based medicine // sechenovskiy Gazette. 2010; 2(2): 106-111.]
2. Герасимов А.Н. Параметрические и непараметрические методы в медицинской статистике // Эпидемиология и Вакцинопрофилактика. 2015; 5(84): 6-12.
[Gerasimov A.N. Parametric and non-parametric methods in medical statistics / Epidemiology and vaccinal prevention. 2015; 5(84): 6-12.]
3. Герасимов А.Н. Медицинская статистика. М. МИА. 2007. 480 с.
4. Рид М., Саймон Б. Методы современной математической физики. Т. 1. Функциональный анализ. М. Мир. 1977. 355 с.
[Reed M., Simon B. Methods of modern mathematical physics. Vol. 1. Functional analysis. M. The World. 1977. 355 S.]
5. Белов Ю.В., Герасимов А.Н., Комаров Р.Н. и др. EuroSCORE model ARCH: Прогнозирование периперационного риска в хирургии дуги аорты с защитой головного мозга (ГМ) // Кардиология и сердечно-сосудистая хирургия. 2015; № 4, том 8: 17-25.
[Belov Yu.V., Gerasimov A.N., Komarov R.N. etc. ARCH EuroSCORE model: Prediction of perioperative risk in surgery of the aortic arch with the protection of the brain (GM) // Cardiology and cardiovascular surgery. 2015; No. 4, volume 8: 17-25.]