

А.М. Носовский^{1*}, А.Э. Пихлак², В.А. Логачев², И.И. Чурсинова³, Н.А. Мутьева²

СТАТИСТИКА МАЛЫХ ВЫБОРОК В МЕДИЦИНСКИХ ИССЛЕДОВАНИЯХ

¹Государственный научный центр Российской Федерации — Институт медико-биологических проблем Российской академии наук, 123007, Москва, Россия; ²ГБОУ ВПО «Московский государственный медико-стоматологический университет имени А.И.Евдокимова» Минздрава России, 127473, Москва, Россия; ³АНО «Артрологическая больница НПО СКАЛ», 109044, Москва, Россия

*Носовский Андрей Максимович, E-mail: collega@pisem.net

♦ Экспериментально найдены характеристики статистических критериев. В результате вычисляли значение статистик W. Ансари-Бредли (Ansari-Bradly) и К. Клотца (Klotz). Для каждой исходной статистики вычисляется нормальная аппроксимация (Z-статистика) и уровень значимости p нулевой гипотезы об отсутствии различий в разбросе значений двух выборок. Если $p > 0.05$, нулевая гипотеза может быть принята.

Предлагаемые методы математической статистики позволяют подтверждать достоверность различий полученных результатов даже в небольших группах наблюдений, если различия достаточно значимы.

Иллюстрацией служили клинические примеры пациентов с костно-суставной патологией.

Ключевые слова: малая выборка, мощность критерия, коксартроз, подагрический полиартрит

A.M. Nosovskiy¹, A.E. Pikhla², V.A. Logachev², I.I. Chursinova³, N.A. Muteva²

SMALL-DATA STATISTICS ANALYSIS IN MEDICAL STUDIES

¹The state research center-institute of medical biological problems of the Russia academy of medical sciences, 123007 Moscow, Russia; ²Moscow State University of Medicine and Dentistry named after A.I. Evdokimov, 127473 Moscow, Russia; ³Arthrology hospital of scientific and practical association SKAL, 109044 Moscow, Russia

♦ The experimentally was found characteristics of statistical criteria. As a result, calculated the value of the statistics by W. Ansari-Bradly and K. Klotz. For each source of statistics calculated normal approximation (Z-statistics) and the significance level of p of the null hypothesis of no difference in the spread of the values of the two samples. At $p > 0.05$ the null hypothesis can be accepted. Suggested methods of mathematical statistics can be confirming the accuracy of the differences of the results, even in small groups of observations, if the differences are significant enough.

We used medical cases of patients with joint and bone pathology.

Key words: small data analysis, power of criteria, coxarthrosis, gouty arthritis

Принципы доказательной медицины предъявляют высокие требования к достоверности сравнительной оценки полученных результатов исследований. Это становится тем более важным, что большинство врачей имеет весьма поверхностное представление о методиках статистической обработки, ограничиваясь в своих публикациях помимо вычисления процентов, в лучшем случае t -критерием Стьюдента.

Однако для проведения полноценного анализа результатов исследования в ряде случаев этого бывает недостаточно. Не вызывает обычно сомнений достоверность выявленных закономерностей, когда число наблюдений составляет несколько тысяч или даже сотен. А если это — несколько десятков? А если мы имеем лишь несколько случаев? Ведь в медицине встречаются достаточно редкие заболевания, хирурги порой выполняют уникальные операции, когда количество наблюдений совсем невелико. Где та грань, тот необходимый и достаточный объем исследований, позволяющий утверждать о несомненном наличии той или иной закономерности?

Этот вопрос имеет важнейшее значение не только при оценке уже проведенных исследований, но и при планировании научной работы. Достаточно ли провести наблюдение за 20 пациентами или необходимо минимум 40? А может быть, хватит и 10 случаев? От своевременного и правильного ответа на этот вопрос зависит не только достоверность сделанных выводов, но и сроки проведения исследований, их стоимость, потребность в кадрах, оснащении и т.д.

Современная статистика знает довольно много приемов, с помощью которых можно определять достоверность результатов даже при небольшом числе наблюдений. Это — методы «малой выборки». Принято считать, что начало статистике малых выборок было положено в первом десятилетии XX века публикацией работы У. Гос-

сета, где он под псевдонимом «Стьюдент» (студент) постулировал так называемое t -распределение. В отличие от теории нормального распределения, теория t -распределения для малых выборок не требует априорного знания или точных оценок математического ожидания и дисперсии генеральной совокупности, а также не требует допущений относительно параметров. В t -распределении одно из отклонений от выборочного среднего всегда фиксировано, так как сумма всех таких отклонений должна равняться нулю. Это сказывается на сумме квадратов при вычислении выборочной дисперсии как несмещённой оценки дисперсии генеральной совокупности и ведёт к тому, что число степеней свободы df получается равным числу измерений минус единица для каждой выборки. Отсюда, в формулах и процедурах вычисления t -статистики для проверки нулевой гипотезы $df = n - 1$. Известны также классические работы крупнейшего английского статистика Р.А. Фишера (в честь которого получило свое название F -распределение) по дисперсионному анализу — статистическому методу, явно ориентированному на анализ малых выборок. Из многочисленных статистик, которые можно обоснованно применять к малым выборкам, можно упомянуть: критерий точной вероятности Фишера; двухфакторный непараметрический (ранговый) дисперсионный анализ Фридмана; коэффициент ранговой корреляции t Кендалла; коэффициент конкордации (W) Кендалла; H -критерий Краскала—Уоллеса для непараметрического (рангового) однофакторного дисперсионного анализа; U -критерий Манна—Уитни; медианный критерий; критерий знаков; коэффициент ранговой корреляции r Спирмена; t -критерий Уилкоксона.

Определённого ответа на вопрос, какой объем должен иметь выборка, чтобы её можно было считать малой, не существует. Однако условной границей между малой и большой выборкой принято считать $df = 30$. Основанием

Коэффициент доверия t и соответствующие уровни доверительной вероятности

t	1,00	1,96	2,00	2,58	3,00
F(t)	0,683	0,950	0,954	0,990	0,997

для этого в какой-то мере произвольного решения служит результат сравнения t -распределения (для малых выборок) с нормальным распределением (z). Расхождение значений t и z имеет тенденцию возрастать с уменьшением и снижаться с увеличением df . Фактически, t начинает тесно приближаться к z задолго до предельного случая, когда $t=z$. Простое визуальное изучение табличных значений t позволяет увидеть, что это приближение становится довольно быстрым, начиная с $df=30$ и выше. Сравнительные величины t (при $df=30$) и z равны соответственно: 2,04 и 1,96 для $p=0,05$; 2,75 и 2,58 для $p=0,01$; 3,65 и 3,29 для $p=0,001$.

В математической статистике употребляют коэффициент доверия t , значения функции $F(t)$ табулированы при разных его значениях, при этом получают соответствующие уровни доверительной вероятности (табл. 1) [1].

Коэффициент доверия позволяет вычислить предельную ошибку выборки $\Delta X_{cp.}$, вычисляемую по формуле $\Delta X_{cp.} = t_{cp.}$, т.е. предельная ошибка выборки равна t -кратному числу средних ошибок выборки [2].

Таким образом, величина предельной ошибки выборки может быть установлена с определённой вероятностью. Как видно из последней графы таблицы 1, вероятность появления ошибки равной или большей утроенной средней ошибки выборки, т.е. $\Delta X_{cp.} = 3\mu_{cp.}$ крайне мала и равна 0,003 (1—0,997). Такие маловероятные события считаются практически невозможными, а потому величину $\Delta X_{cp.} = 3\mu_{cp.}$ можно принять за предел возможной ошибки выборки [3].

Интервал, в который с данной степенью вероятности будет заключена неизвестная величина оцениваемого параметра, называют доверительным, а вероятность P — доверительной вероятностью [4]. Чаще всего доверительную вероятность принимают равной 0,95 или 0,99, тогда коэффициент доверия t равен соответственно 1,96 и 2,58.

Это означает, что доверительный интервал с заданной вероятностью заключает в себе генеральную среднюю.

Чем больше величина предельной ошибки выборки, тем больше величина доверительного интервала и тем, следовательно, ниже точность оценки [4].

Применение данного подхода может быть проиллюстрировано наблюдением за 20 пациентами с коксартрозом, находившихся на лечении в Артрологической больнице НПО «СКАЛ» (Научно-производственное объединение «Специализированное курсовое амбулаторное лечение») г. Москвы.

При проверке статистической гипотезы возможны ошибки. Есть два рода ошибок. Ошибка первого рода заключается в том, что отвергают нулевую гипотезу, в то время как в действительности эта гипотеза верна. Ошибка второго рода состоит в том, что принимают нулевую гипотезу, в то время как в действительности эта гипотеза неверна.

Вероятность ошибки первого рода называется уровнем значимости и обозначается α . Таким образом, $\alpha = P\{U \in \Psi | H_0\}$, т.е. уровень значимости α — это вероятность события $\{U \in \Psi\}$, вычисленная в предположении, что верна нулевая гипотеза H_0 .

Уровень значимости и мощности критерия объединяются в понятие функции мощности критерия — функции, определяющей вероятность того, что нулевая гипотеза будет отвергнута. Функция мощности зависит от критической области Ψ и действительного распределения результатов наблюдений. В параметрической

задаче проверки гипотез распределение результатов наблюдений задается параметром θ . В этом случае функция мощности обозначается $M(\Psi, \theta)$ и зависит от критической области Ψ и действительного значения исследуемого параметра θ . Если $H_0: \theta = \theta_0$, $H_1: \theta = \theta_1$, то $M(\Psi, \theta_0) = \alpha$, $M(\Psi, \theta_1) = 1 - \beta$, где α — вероятность ошибки первого рода, β — вероятность ошибки второго рода. Тогда, мощность критерия — это вероятность того, что нулевая гипотеза будет отвергнута, когда альтернативная гипотеза верна.

Функция мощности $M(\Psi, \theta)$ в случае одномерного параметра θ обычно достигает минимума, равного α , при $\theta = \theta_0$, монотонно возрастает при удалении от θ_0 и приближается к 1 при $|\theta - \theta_0| \rightarrow \infty$.

Оценим необходимую мощность статистических критериев (рис. 1), которые могли бы быть применены для анализа лечения 20 пациентов с коксартрозом.

Как видим, при среднеквадратическом отклонении равном 3,0, что бывает крайне редко, будут получены результаты с высокой степенью надёжности $p \leq 0,05$, если разность между средними будет превышать 8. Но уже при среднеквадратическом отклонении равном 1,5, эта разность должна превышать всего 4.

Для определения уровня значимости p обычно используется приближенная нормальная Z -аппроксимация соответствующей статистики. Такая аппроксимация дает хорошее приближение при достаточно больших размерах выборок. При малом объеме выборки и значениях p , близких к 0,05, мы проверяли вывод о нулевой гипотезе срав-

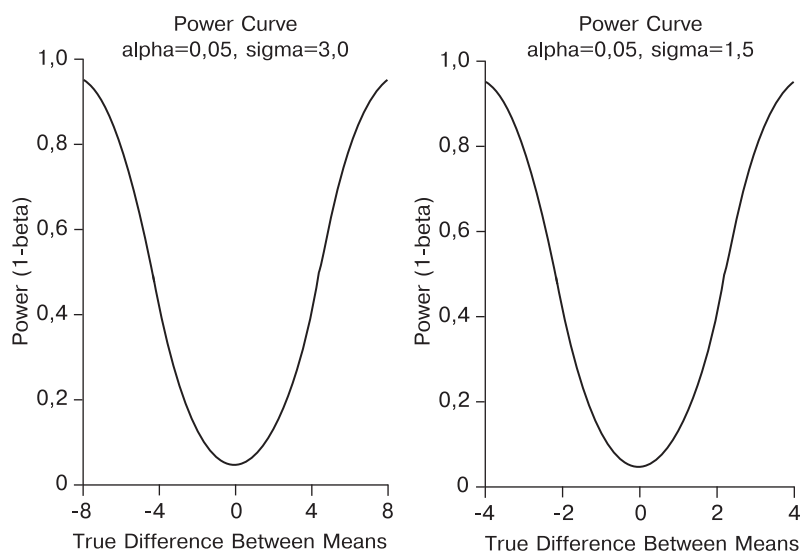


Рис. 1. Экспериментально найденные характеристики статистических критериев.

Таблица 2.

Группы наблюдения

Содержание лечебного комплекса	Группы наблюдения			Всего наблюдений
	Группа 1 Комплекс № 1 (n=5)	Группа 2 Комплекс № 2 (n=7)	Группа 3 Комплекс № 3 (n=8)	
Нимесулид, витамины, хондропротекторы, лечебная физкультура	+	+	+	20
+ физиотерапия	---	+	+	15
+ массаж	---	---	+	8

нением вычисленного значения статистики с критическим значением в таблице соответствующего распределения из статистического справочника.

Критерии различия сдвига (положения). Мы использовали эти критерии для проверки следующих гипотез:

- ♦ отсутствие различий во взаимном положении (медианах) двух исследованных выборок;
- ♦ сдвиг выборок друг относительно друга равен некоторому значению d ;
- ♦ медиана одной анализируемой выборки равна значению d .

В случае б) необходимо было предварительно все значения второй выборки уменьшить на величину d : $y_i = y_i - d$.

В случае в) необходимо подготовить вспомогательную парную выборку, все элементы которой равны d .

В результате вычисляли:

- ♦ значение статистики W. Вилкоксона (Wilcoxon) — сумма рангов R_{xi} элементов одной из выборок в объединенной ранжированной выборке;
- ♦ значение статистики V. Ван дер Вардена (van der Varden), основанную на использовании метода «произвольных меток».

Для каждой статистики вычислялась нормальная аппроксимация (Z-статистика) и уровень значимости Р нулевой гипотезы об отсутствии различий в сдвиге по отношению друг к другу. Если $p > 0,05$ нулевая гипотеза может быть принята.

Некоторые пакеты и авторы предлагают использовать U-критерий Манна—Уитни (Mann—Whitney) и критерий Вальда-Вольфовица (Vald-Wolfowitz). Однако давно уже доказано [4—7], что критерий Манна—Уитни эквивалентен, т.е. обладает теми же возможностями, что и крите-

Средние показатели интенсивности боли (в баллах по ВАШ)

Показатель	Группа 1 (n=5)			Группа 2 (n=7)			Группа 3 (n=8)		
	Начало наблюдения	Конец наблюдения	Снижение боли	Начало наблюдения	Конец наблюдения	Снижение боли	Начало наблюдения	Конец наблюдения	Снижение боли
Боль при движении	70 ± 8	39 ± 19*	-31 (44,3%)	77 ± 5	31 ± 9*	-46 (59,7%)	76 ± 5	19 ± 7*	-57 (75,0%)
Боль в покое	43 ± 13	27 ± 17	-16 (37,2%)	44 ± 9	13 ± 5*	-31 (70,5%)	45 ± 8	6 ± 1*	-39 (86,7%)

* $p \leq 0.05$.

Таблица 4.

Данные лабораторного обследования больного Б.

№	Показатель	Норма	Результат предпоследнего посещения	Результат последнего посещения
1	Гематокрит, %	40—48	38,7	43,5
2	Лимфоциты, %	19—37	42	39
3	СОЭ, мм/час	2—10	39	10
4	Мочевая кислота, мкмоль/л	200—416	504	489
5	Креатинин, мкмоль/л	44—106	238	202
6	Паратиреоидный гормон, пг/мл	7—53	76,8	101
7	Фибриноген, г/л	1,69—3,92	5,7	3
8	Белок в моче, г/л	0—0,1	1	0,3

рий Вилкоксона, а критерий Вальда-Вольфовица страдает сравнительно малой чувствительностью.

Критерии различия масштаба (рассеяния). Мы использовали эти критерии для проверки следующих гипотез:

- ♦ гипотеза об отсутствии различий в масштабах (в разбросе или рассеянии значений) исследуемых выборок;
- ♦ гипотеза о том, что отношение масштабов выборок равно заданной величине g .

В последнем случае необходимо предварительно изменить значения второй выборки $y_i = (y_i - m_0)/g$, где m_0 —общая медиана двух исследуемых спектров.

Если медианы генеральных совокупностей, из которых извлечены выборки, не равны по величине, но их значения известны; m_1, m_0 , то настоящий метод можно применить, предварительно модифицировав одну из выборок, например, в выборку $y_i = y_i - m_2 + m_1$.

Если же медианы не равны и не известны, то следует подтвердить гипотезу об отсутствии различий сдвига или же использовать метод для обнаружения произвольных альтернатив.

В результате вычисляли значение статистик W. Ансари-Бредли (Ansari-Bradley) и К. Клотца (Klotz), которые являются концептуальными аналогами статистик Вилкоксона и Ван дер Вардена.

Для каждой исходной статистики вычисляется нормальная аппроксимация (Z-статистика) и уровень значимости Р нулевой гипотезы об отсутствии различий в разбросе значений двух выборок. Если $p > 0.05$, нулевая гипотеза может быть принята.

Таким образом, предлагаемые выше методы математической статистики позволяют подтверждать достоверность различий

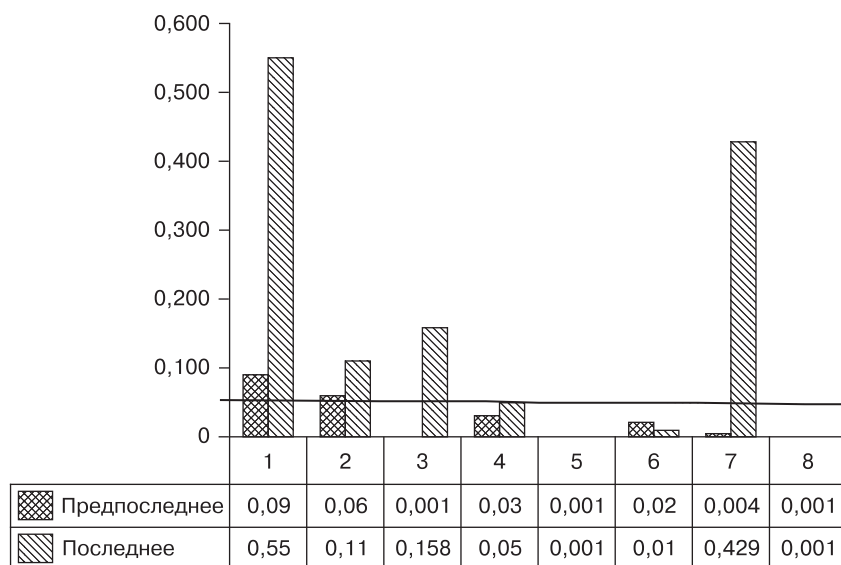


Рис. 2. p -значения клинических показателей больного Б. при предпоследнем и последнем обследовании.

полученных результатов даже в небольших группах наблюдений, если различия достаточно значимы.

Иллюстрацией могут служить два клинических примера пациентов с костно-суставной патологией.

Клинический пример № 1. У 20 пациентов с коксартрозом применяли базовый лечебный комплекс, включающий пероральный прием нимесулида, хондропротекторов, внутримышечные инъекции витаминов и лечебную физкультуру. Кроме этого у 15 из них применяли физиотерапевтическое лечение, а у 6 пациентов — массаж. Таким образом, образовалось 3 группы пациентов с небольшим (от 5 до 8) числом наблюдений (табл. 2).

Среди прочих параметров перед началом лечения и после завершения курса (21 ± 2 дня) оценивали интенсивность боли при движении и в покое по 100-бальной визуальной аналоговой шкале (ВАШ).

Использовались следующие методы статистик W. Ансари-Бредли (Ansari-Bradly) и К. Клотца (Klotz) (табл. 3).

Согласно полученным данным (табл. 3) было отмечено, что снижение боли в покое в группе 1 в конце наблюдения не являлось достоверным. Однако по всем другим изучаемым параметрам выявлены достоверные значения. Рассматриваемый клинический пример свидетельствует о возможности получения достоверных результатов на малом количестве выборки.

В клиническом примере № 2 рассматриваются в динамике лабораторные данные больного Б., страдающего хроническим подагрическим полиартритом, подагрической нефропатией с явлениями ХПН, которые находились за пределами референсных значений (табл. 4).

Рассчитаем вероятность того, что результаты анализа статистически достоверно выходят за границы клинической нормы. Для этого используем вероятностный калькулятор статистического пакета «STATISTICA 6.0». В данном случае р-значение характеризует ошибку первого рода: вероятность отклонить правильную гипотезу, когда на самом деле она верна. В большинстве случаев результаты предпоследнего посещения статистически достоверно отличаются от нормы (рис. 2). Поскольку пороговый уровень значимости в данном случае мы принимаем равным 0,05, то результаты гематокрита, лимфоцитов, СОЭ, фибриногена статистически значимо улучшились при последнем посещении. Соответственно, клинические показатели мочевой кислоты, креатинина, паратиреоидного гормона и белка в моче, с точки зрения математической статистики, не улучшились.

Таким образом, при планировании исследования важно учитывать мощность применяемых статистических критериев, которые определяются вариабельностью выборки и заданным уровнем значимости.

Предлагаемый подход может быть интересен специалистам в области персонифицированной медицины для

анализа в динамике применяемых методов лечения и лекарственных средств, при контроле за проводимыми лечебными и диагностическими мероприятиями.

ЛИТЕРАТУРА

1. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. М.: Наука; 1995.
2. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. М.: Наука; 2003.
3. *Кобзарь А.И.* Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ; 2006.
4. *Правецкий Н.В., Носовский А.М., Матросова М.А., Холин С.Ф., Шакин В.В.* Математическое обоснование достаточного количества измерений для достоверной оценки регистрируемых параметров в космической биологии и медицине. Космическая биология и авиакосмическая медицина. М.: Медицина; 1990; 5: 53-6.
5. *Холлендер М., Вульф Д.А.* Непараметрические методы статистики. М.: Финансы и статистика; 1983.
6. *Носовский А.М.* Применение вероятностных моделей на окружности в медико-биологических исследованиях. Космическая биология и авиакосмическая медицина. Тезисы докладов IX Всесоюзная конференция. Калуга, 19-21 июня 1990.
7. *Носовский А.М., Правецкий Н.В., Холин С.Ф.* Математический подход к оценке точности измерений физиологического параметра различными методами. Космическая биология и авиакосмическая медицина. М.: Медицина; 1991; 6: 53-5.

REFERENCES

1. *Bol'shev L.N., Smirnov N.V.* Tables of Mathematical Statistics. Moscow: Nauka; 1995 (in Russian).
2. *Korn G., Korn T.* Mathematical Handbook for Scientists and Engineers. Moscow: Nauka; 2003 (in Russian).
3. *Kobzar' A.I.* Applied Mathematical Statistics. For engineers and scientists. Moscow: FIZMATLIT; 2006 (in Russian).
4. *Pravetskiy N.V., Nosovskiy A.M., Matrosova M.A., Kholin S.F., Shakin V.V.* Mathematical justification of a sufficient number of measurements for reliable evaluation of recorded parameters in space biology and medicine. Space Biology and Aerospace Medicine. Moscow: Meditsina; 1990; 5: 53-6 (in Russian).
5. *Khollender M., Vulf D.A.* Non-parametric statistical methods. Moscow: Finansy i statistika; 1983 (in Russian).
6. *Nosovskiy A.M.* The use of probabilistic models on the circle in biomedical research. Space Biology and Aerospace Medicine. Abstracts of the IX All-Union Conference. Kaluga, June 19-21, 1990 (in Russian).
7. *Nosovskiy A.M., Pravetskiy N.V., Kholin S.F.* Mathematical approach to estimation accuracy of the physiological parameter by different methods. Space Biology and Aerospace Medicine. Moscow: Meditsina; 1991; 6: 53-5 (in Russian).