УДК 612

Н.В. АКУЛИЧ

ИСПОЛЬЗОВАНИЕ БАЙЕСОВСКОГО КРИТЕРИЯ ДЛЯ ПРОВЕРКИ ГИПОТЕЗ В БИОЛОГИИ

В статье рассматривается проблема анализа гипотез в биологических исследованиях. Проанализированы условия использования теста Стьюдента. Обоснован и предложен метод Байеса для проверки гипотез в биологических исследованиях. Проведена проверка наших гипотез с применением минимального Байесовского критерия, показавшего целесообразность его использования.

Делается вывод, что биологии необходимо использовать те же принципы, которые присутствуют в доказательной медицине, в противном случае дальнейшее исследования направляются в ложное русло, что влечет за собой неэффективное использование финансовых ресурсов, времени исследователей, а участники исследований подвергаются неоправданному риску. Предлагается представление вместе с рукописью статьи протоколов экспериментов для проведения экспертизы и введение должности статистика в редакционную коллегию.

Современная биология без опоры на математический аппарат не может обеспечить точный и логически строгий анализ экспериментальных данных. Только с помощью правильно выбранного статистического метода можно описать, объяснить и углубленно исследовать всю совокупность результатов измерений [1–8, 16–18], и, по мнению О.Ю. Ребровой, "... в биологическом эксперименте для анализа данных необходимо применение статистики, в противном случае выводы нельзя считать научно обоснованными" [17].

В настоящей статье мы ставим перед собой следующие задачи: проанализировать материал по обоснованию применения параметрических методов статистического анализа, в частности, теста Стьюдента и обосновать необходимость применения в биологических исследованиях альтернативного ему метода.

Известно, что тест Стьюдента с использованием *величины p* – показателя статистической значимости (достоверности) данных [8, 9–13], предложенной Фишером, является практически общепринятым методом статистического анализа при проверке гипотез в биологии. Популярность этого теста настолько велика, что во многих случаях он используется неправомерно; это лишает выводы какой-либо научной ценности. Так, в работе В.П. Леонова [16] приведены итоги анализа более 1800 журнальных статей и 160 диссертаций. Вывод, сделанный автором по итогам анализа этих работ, таков: использование параметрических методов анализа является неправомерным, поскольку – в 50-70% случаев биомедицинские количественные показатели **не подчиняются нормальному распределению**, то в исследованиях часто игнорируются остальные параметры распределения такие, как меры рассеяния, меры формы, корреляции между признаками, структура наблюдений и переменных и т.д.

В.П. Леонов сообщает также и о том, что в отечественных публикациях крайне редко используются проверка нормальности распределения, многомерные методы, анализ таблиц сопряженности, непараметрические критерии и анализ выживаемости. Следовательно, налицо стремление исследователей использовать (даже пусть и не обосновано) тест Стьюдента, словно этот метод является единственным претендующим на строгость научного вывода.

Представим пример гипотетического исследования о влиянии ***условия x*** (например, температуры) на ***переменную y*** (например, параметры внешнего дыхания), где достигнут уровень статистической значимости полученных результатов p = 0,06. В разделе "Результаты и их обсуждение" обычно читаем вывод: ...выявлены достоверные ($p < 0,05$) изменения между контрольной и опытной группами, следовательно температура оказывает влияние на дыхание.

Это наиболее типичный пример рассуждения: в начале раздела "Обсуждение результатов" приводится вывод, основанный на полученных результатах, и лишь потом приводится спекуляция о предполагаемом биологическом механизме. Создается впечатление, что **вывод автоматически вытекает из результатов и представляет собой не более чем словесное выражение утверждения "p = 0,06", не нуждаясь в предварительном обсуждении**. Таковы последствия применения статистического метода, который почти полностью лишает читателя возможности различать математическую обработку результатов

исследования (например, определение средней величины, среднеквадратического отклонения и др.) и научное мышление, основанное на проверке гипотез с помощью статистических методов.

Президент издательства "Медиа-Сфера" С.Е. Бащинский пишет: "По сути, эти клише представляют собой магические заклинания, служащие, по мысли авторов, "пропуском в науку", а статистике в биомедицинских исследованиях отводится довольно своеобразная роль: звучные и непонятные авторам термины нужны для достижения основной цели – придания работе научного "веса", достаточного для опубликования в журнале или для защиты диссертации" [16].

Величина p

Статистический анализ на последнем этапе анализа экспериментальных данных предполагает определение величины p, которая представляет собой вероятность получить эффект, равный наблюдаемому или превосходящий последний, при условии справедливости нулевой гипотезы (гипотезы об отсутствии различий между сравниваемыми группами) [8, 15]. Р. Фишер предложил *величину p* в качестве меры соответствия полученных данных нулевой гипотезе[1]. Он рекомендовал использовать этот показатель не как составную часть формального построения статистических умозаключений, а как компонент предварительных, качественных выводов, основывающихся на данных наблюдений и учитывающих прошлый опыт [17].

В пособие П.Ф. Рокицкого [18], которое до сих пор является единственным в своем роде учебником по биологической статистике, приводится другое обоснование применения *величины p*: "*Статистические показатели и различия между ними характеризуются определенными уровнями значимости, и отбрасывание нулевой гипотезы должно быть связано с принятием определенного уровня значимости.*

Так, если признан необходимым уровень значимости 0,01 (величины p – примечание Н.В. Акулича) *и если вероятность достоверности данного статистического показателя или разницы между показателями не удовлетворяет этому условию, т. е. она ниже 0,99 (например, 0,97, 0,91, 0,88), то нет оснований для отбрасывания нулевой гипотезы.*

Таким образом, если полученные данные характеризуются уровнем значимости p > 0,05, то нет оснований отклонять нулевую гипотезу. Если p < 0,01, то для отбрасывания нулевой гипотезы основания достаточные".

На наш взгляд, это не совсем верная интерпретация *величины p*, поскольку ее **рассчитывают исходя из предположения, что нулевая гипотеза верна**. Следовательно, ***величина p* не может служить мерилом вероятности того, что необходимо выдвигать альтернативную гипотезу и тем более, на основании того, что p < 0,05, считать ее обоснованной**. Из этой логической ошибки вытекает неверное представление о том, что об истинности гипотезы можно судить исключительно по результатам исследования.

Еще одним фактором сомнительности применения предложенной Р. Фишером *величины p* является то, что статистическая значимость наблюдаемого эффекта оценивается без учета его величины, поскольку **небольшой эффект при**

[1] Согласно этой гипотезе, первоначально принимается, что между данными показателями (или группами, на основе которых они получены) достоверного различия нет, т.е. что обе группы вместе составляют один и тот же однородный материал, одну совокупность. Статистический анализ должен привести или к отклонению нулевой гипотезы, если доказана достоверность полученных различий, или к ее сохранению, если достоверность различий не доказана, т. е. различия признаны случайными [18].

большом размере выборки может характеризоваться такой же *величиной p*, как и значительный эффект в исследовании с малым размером выборки!

Эта критика привела к появлению дополнительных критериев оценки статистических гипотез [11–12], но (так часто случается в истории) *величина p* была увековечена (см. далее) в том методе, который должен был заменить ее – в методе проверки статистических гипотез Неймана и Пирсона.

Дж. Нейман и Э. Пирсон [15] разработали **подход, при котором выдвигаются две гипотезы о природе явления**: нулевая, обычно подразумевающая отсутствие эффекта, и альтернативная, обычно противоположная нулевой (например, предполагающая, что эффект имеется).

Результатом проверки гипотез становится решение исследователя об отказе от одной и о принятии другой гипотезы только на основании полученных данных. При этом возникают ошибки двух типов: ложное заключение о различии в эффективности между двумя методами лечения, когда в действительности такие различия отсутствуют (ложноположительный результат – ошибка I рода) и ложное заключение об одинаковой эффективности двух методов лечения, когда в действительности между ними существуют различия (ложноотрицательный результат – ошибка II рода).

Подобный подход предпочтителен тем, что, приняв гипотезу, можно рассчитать вероятность этих ошибок. Подразумевалось, что при проверке гипотез должна учитываться и другая информация (например, при определении величины той и другой ошибки должны учитываться последствия ложноположительного и ложноотрицательного заключения) [10–12, 14]. Этот подход является прогрессивным, но его использование в качестве научно-практической модели имеет ограничения, обусловленное тем, что любой элемент индукции рано или поздно приводит к теореме Байеса, которую Дж. Нейман и Э. Пирсон пытались обойти. Поэтому они предложили отказаться от индуктивных рассуждений при анализе результатов отдельных исследований и использовать дедуктивные методы для уменьшения количества ошибок, допущенных в серии экспериментов [13].

В работе [14] показано, что: ... ни один критерий, опирающийся на теорию вероятностей, не может обеспечить абсолютной уверенности в правильности или ложности гипотезы. Но, не рассчитывая узнать, какая из гипотез, выдвинутых в отдельном исследовании, верна, а какая ложна, можно разработать правила, следуя которым возможно минимизировать ошибки в серии экспериментов".

Ключевое в этой фразе то, что авторы назвали вещи своими именами. **Если исследователь намерен быть объективным, то он должен отказаться от возможности оценивать достоверность (истинность) результатов отдельного эксперимента и ограничится лишь выводом о статистической значимости (или незначимости) результатов**.

Поскольку проверка гипотез по вышеизложенному методу лишает ученого права определять ценность отдельного опыта и познавать скрытую в нем истину (но во всех опытах исключены ошибки!), то со временем появилось (к сожалению) дополнение, благодаря которому, казалось бы, удалось избежать этой жертвы. **Таким дополнением стала величина p**!

Будучи несовместимыми, эти статистические приемы **слились воедино** и **ошибочно воспринимаются многими как неотъемлемые составляющие единого метода построения логических умозаключений в статистике** [9–14].

Еще одно из распространенных заблуждений при проведении медико-биологических исследований заключается в том, что статистические методы, основанные на использовании *величины p*, позволяют получить количественный

показатель, который отражает вероятность ошибочных выводов без учета природы изучаемого явления и результатов предыдущих исследований.

Это обстоятельство извращает логику умозаключений и сами выводы, затрудняя понимание связи между доказательностью результатов отдельного исследования и убедительностью других доказательств (данные клинических и экспериментальных исследований), поэтому результаты многих исследований не выдерживают проверки временем [17]. Кроме того, в статьях биологического профиля довольно часто происходит выхолащивание раздела "Результаты и их обсуждение", отодвигая на второй план или вообще вытесняя теоретическое обоснование наблюдаемых явлений и накопленный практический опыт.

Практически общепринятым является правило, когда авторы в разделе актуальность лишь ограничиваются перечислением ученых и научных школ, работающих в данном направлении, но очень немногие из них упоминают о предыдущих исследованиях по той же тематике исследования, с обязательным учетом доказательности полученных экспериментальных данных [16]. Это – закономерное следствие использования методологии, подразумевающей самодостаточность отдельного опыта, на основании результатов которого можно сделать выводы с определенной вероятностью ошибки, не учитывая информацию из других источников.

Альтернативой этому подходу, по-нашему мнению, является более точный статистический инструмент – Байесовский критерий (БК), позволяющий отделить доказательность результатов отдельного эксперимента от общих тенденций, отмеченных в ряде исследований, которые могут формироваться на основе данных хорошо проведенного эксперимента, но математический аппарат их не совершенен [13].

БК представляют в виде коэффициента, отражающего степень изменения вероятности того, что гипотеза верна после полученных данных опыта, т.е. БК равен отношению вероятности получения данных при условии справедливости нулевой гипотезы к вероятности получения данных при условии справедливости альтернативной гипотезы [11, 12].

Понятие о байесовском критерии

Байесовский критерий используют для проверки научных представлений с помощью вновь полученных данных [12], в простейшей его форме называют также **отношением правдоподобия**. Минимальный БК представляет собой объективный показатель, способный заменить *величину p*. В отличие от *величины p* теоретическое обоснование и интерпретация БК позволяют использовать его как в процессе проверки гипотез, так и в процессе принятия решений.

БК показывает, что оценка данных с помощью *величины p* преувеличивает доказательства, опровергающие нулевую гипотезу [13]. И самое главное, **БК подразумевает включение в анализ прошлого опыта в виде вероятности истинности вывода**.

БК отделяет логическое умозаключение от данных опыта и в то же время дает исследователю возможность комбинировать старую и новую информацию, то есть такой подход обычно рассматривается как **способ переоценки наших представлений с помощью данных опыта**. Используется этот метод также и для расчета доказательности данных.

Формула Байеса имеет две составляющие: показатель, характеризующий данные опыта, и показатель, характеризующий степень нашей уверенности в истинности гипотезы. Этот критерий называют также относительными шансами, а в логарифмической форме – весом доказательства [6, 7]. Следовательно, различие между доказательностью данных и вероятностью ошибки становится оче-

видным, если БК (доказательство) представить в виде коэффициента, отражающего степень изменения вероятности того, что гипотеза верна, после получения данных опыта.

Формула Байеса выглядит следующим образом:

Априорные шансы справедливости нулевой гипотезы × **БК** = **Апостериорные шансы справедливости нулевой гипотезы**,

где

$$\text{БК} = \frac{\text{вероятность получения данных при условии соблюдения нулевой гипотезы}}{\text{вероятность получения данных при условии соблюдения альтернативной гипотезы}}$$

БК показывает, насколько каждая из двух гипотез соответствует полученным данным. Та из них, которая лучше описывает данные, имеет больше доказательств в свою пользу. В отличие от *величины p*, использование БК теоретически обосновано и допустимо как при проверке гипотез, так и в процессе принятия решений. БК позволяет связать объективную вероятность с доказательством и субъективной вероятностью и может рассматриваться со всех трех точек зрения.

Если БК равен 1/2, то:

1. С точки зрения объективной вероятности: вероятность получить наблюдаемые результаты при условии справедливости нулевой гипотезы в 2 раза меньше, чем вероятность получить их при условии справедливости альтернативной гипотезы.

2. С точки зрения индуктивного доказательства: доказательство в 2 раза слабее поддерживает нулевую гипотезу, чем альтернативную.

3. С точки зрения субъективной вероятности: шансы того, что нулевая гипотеза верна по отношению к шансам того, что верна альтернативная гипотеза, после получения результатов опыта уменьшились в 2 раза.

Отметим несколько различий между БК и *величиной p*:

1) БК отражает не вероятность, а отношение вероятностей, и его значение колеблется от нуля до бесконечности;

2) БК подразумевает наличие двух гипотез, следовательно, опровергая нулевую гипотезу, доказательство должно свидетельствовать в пользу альтернативной;

3) БК зависит только от вероятности получения результатов конкретного опыта, поэтому на него не влияют факторы, не связанные непосредственно с полученными данными и от которых зависит *величина p*.

Нами проведена проверка наших гипотез с применением минимального БК, но полученные результаты выходят за рамки данной статьи и будут опубликованы позднее.

Подавляющее большинство авторов статей медико-биологического профиля не видят разницы между убедительностью доказательства и вероятностью ошибки, и оперируя *величиной p*, рассуждают о "достоверности" данных. Без использования БК нельзя оперировать мерой убедительности доказательства; БК изменяет априорные вероятности, и по степени их изменения мы видим, какова степень доказательности данных – высокая или низкая.

Разные значения БК по-разному влияют на априорные вероятности нулевой гипотезы. Например, при априорной вероятности справедливости нулевой гипотезы равной 90% и БК, равном 0.1, мы получаем апостериорную вероятность

справедливости нулевой гипотезы, равной всего 47%, что приведет к появлению сомнения в истинности нулевой гипотезы (т.е. это тот самый случай, когда ученый, потирая руки, пишет $p < 0,05$ и ...), а если БК будет равен 0.01, то вероятность справедливости нулевой гипотезы снизится, с 90% (перед проведением опыта) до 8% (после завершения опыта), соответственно.

Таким образом, **чем весомее имеющееся доказательство, тем меньше новой внешней информации нужно для подтверждения гипотезы. И наоборот, чем меньше такой информации свидетельствует об истинности гипотезы, тем убедительнее доказательство того, что гипотеза выглядит правдоподобной**.

Выводы

Представленные в статье сведения и сделанные на их основе обобщения позволяют сделать несколько выводов:

1. На мой взгляд, в биологии необходимо использовать те же принципы, что присутствуют в доказательной медицине, одним из которых является следующий: **вес каждого факта тем больше, чем строже методика научного исследования, в ходе которого он получен**. В противном случае дальнейшее исследование направляется в ложное русло, что влечет за собой неэффективное использование финансовых ресурсов, времени исследователей, а участники таких "исследований" подвергаются неоправданному риску.

2. Применение статистики для обработки результатов биомедицинских исследований в ряде случае неправомерно, поскольку у читателя нет возможности самостоятельно проверить результаты, то целесообразно представление вместе с рукописью статьи протоколов экспериментов для проведения экспертизы, что делает желательным введение должности статистика в редакционной коллегии. В некоторых авторитетных журналах статьи, различающиеся по качеству статистического анализа, помечаются особыми индексами.

Сказываются до сих пор и последствия периода "лысенковщины", когда математика, и особенно статистика, активно изгонялись из биологии и медицины. Т.Д. Лысенко, выступая с заключительным словом на августовской сессии 1948 г. ВАСХНИЛ, окончательно сформулировал тезис о том, что теория вероятностей и статистика нужны только менделистам-морганистам, а "мичуринской биологии" эти науки не нужны. Более того, известны случаи, когда в те годы ВАК СССР отказывал в присвоении ученых степеней из-за того, что медики использовали в своих диссертациях статистику [16].

ЛИТЕРАТУРА

1. ***Ahman, D.G.*** Confidence intervals in research evaluation / D.G. Ahman // ACP J. Club. – 1992. – Suppl 2. – A. 28-97.
2. ***Berry, G.*** Statistical significance and confidence intervals / G. Berry // Med J Aust. – 1986. – Vol. 144. – № 61. – P. 8-9.
3. ***Brophy, J.M.*** Placing trials in context using Bayesian analysis / J.M. Brophy, L. Joseph // GUSTO revisited by Reverend Bayes. JAMA. – 1995. – № 273.– P. 87.
4. ***Browner, W.*** Are all significant p values created equal? The analogy between diagnostic tests and clinical research / W. Browner, T. Newman // JAMA. – 1987. – Vol. 257. – № 2459. – P. 163.
5. ***Cotton, T.*** Statistics in Medicine / T. Cotton. – Boston: Little, Brown; 1974. – 214 p.
6. ***Cox, D.*** Theoretical Statistics / D. Cox, D. Ilinckley. New York: Chapman and Hall; – 1974. – 113 p.
7. ***Dupont, W.D.*** Sequential stopping rules and sequentially adjusted p values: does one require the other? / W.D. Dupont // Controlled Clin Trials. – 1983. – № 4. – P.31.

8. ***Fisher, R.*** Statistical Methods and Scientific Inference, 3d ed. / R. Fisher. – New York: Macmillan; 1973. – 98 p.
9. ***Freeman, P.R.*** The role of p-values in analysing trial results / P.R. Freeman // Stat Med. – 1993 – Vol.12. – № 1442. – P. 552.
10. ***Gigerenzer, C.*** The Empire of Chance / C. Gigerenzer [and etc.]. – Cambridge, UK: Cambridge Univ Pr; 1989. – 118 p.
11. ***Goodman, S.*** Multiple comparisons, explained / S. Goodman // Am. J. Epidemiol. – 1998. – № 147. – P. 815.
12. ***Goodman, S.N.*** p-Values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate / S.N. Goodman // Am. J. Epidemiol. – 1993. – № 137. – P. 485.
13. ***Greenland, S.*** Empirical-Bayes adjustments for multiple comparisons are sometimes useful / S. Greenland, J.M. Robins // Epidemiology. – 1991. – № 2. – P. 295.
14. ***Ludbrook, J.*** Issues in biomedical statistics: statistical inference / J. Ludbrook, H. Dudley // Aust. N. Z. J. Surg. – 1994. – № 64. – P. 636.
15. ***Pearson, E.*** Some thoughts on statistical inference / E. Pearson // Annals of Mathematical Statistics. – 1962. – № 33. – P. 403.
16. ***Леонов, В.П.*** Наукометрика статистической парадигмы экспериментальной биомедицины / В.П. Леонов // Вестник ТГУ. – 2002. – № 275. – С. 17-24.
17. ***Реброва, О.Ю.*** Статистический анализ медицинских данных. Применение пакета прикладных Программ SIATISTICA / О.Ю. Реброва. – М.: МедиаСфера, 2002. – 312 с.
18. ***Рокицкий, П.Ф.*** Биологическая статистика. Изд-е- 3-е, испр. / П.Ф. Рокицкий. – Минск: Вышэйш. школа, 1973. – 320 с.

Поступила в редакцию 14.09.2006 г.