Описание статистического анализа данных в оригинальных статьях. Типичные ошибки

О.Ю. РЕБРОВА

Description of statistical analysis of data in original articles. Typical errors

O.YU. REBROVA

Межрегиональная общественная организация «Общество специалистов доказательной медицины»

Статистический анализ традиционно применяется для обработки собранной в ходе исследования информации, однако с современных позиций это нельзя признать оптимальным. Статистическое «сопровождение» работы должно осуществляться с ее первого этапа (формулировка задач) и до последнего (подготовка публикации).

Важнейшая цель статистического анализа — сделать вывод о существовании некой общей закономерности на основании анализа ограниченного числа наблюдений. К сожалению, это осознается не всеми исследователями (даже выполняющими диссертационные исследования на присвоение ученой степени доктора наук), и обращение к статистическому анализу поначалу бывает вызвано тем, что «так делают все».

Описание методов статистического анализа. Первичный и вторичный анализ данных

Раздел оригинальных статей «Материал и методы» — важнейший для оценки научной обоснованности результатов работы.

Среди описания многих методов — клинических, лабораторных, инструментальных — в данном разделе статьи следует давать описание и математических методов анализа полученных в ходе исследования данных. Следует указать, какой программный пакет использовался (упоминание использования конкретного программного пакета подразумевает легальность его приобретения), затем описать форматы представления описательной статистики, назвать методы сравнения групп, анализа взаимосвязей признаков и т.д., а также указать пороговый уровень статистической значимости p_o (если в работе проводится проверка статистических гипотез).

Во многих публикациях до настоящего времени описание методов статистического анализа отсутствует. Часто исследователь (или его научный руководитель) имеет сильную содержательную гипотезу, основанную на их клиническом опыте. В этом случае цель статистического анализа — проверить, верна ли эта гипотеза. Если в процессе анализа данных выясняется, что исходную гипотезу обосновать не удалось, то типичное поведение исследователя — поиск эффектов в подгруппах наблюдений (почти всегда обнаруживаемых, однако, чаще всего, ложноположительных, о чем обычно никто не задумывается, потому что «поджимают» сроки защиты диссертации и нужно выдать «положительный» результат) и последующая переформулировка (подгонка) задач под полученные результаты.

Хуже того, авторы часто «перелопачивают» весь массив данных в поисках хоть каких-нибудь закономерностей («data dredging»). Находя таковые, авторы относятся к ним некритично, не осознавая очень высокую вероятность статистических ошибок I рода — обнаружение случайных, несуществующих в реальности закономерностей, а также неизбежные систематические смещения (неэквивалентность подгрупп и пр.). Такой анализ в подгруппах (вторичный анализ данных) имеет право на существование, однако интерпретация его результатов и формулировка выводов должны быть весьма осторожными. Полученные в ходе вторичного анализа выводы могут рассматриваться лишь как предварительные, дающие повод к дальнейшим исследованиям.

Как известно, в научной периодике наблюдается так называемые публикационные смещения, или систематические ошибки, обусловленные опубликованием только положительных результатов исследования [2]. По величине этого смещения можно судить о степени предпочтений авторов исследований публиковать лишь положительные (т.е. доказывающие существование изучаемого эффекта) результаты. Это связано с тем, что авторы большинства исследований бессознательно следуют стратегии «Reject-Support», т.е. стремятся отвергнуть нулевую гипотезу, приняв альтернативную (соответствующую собственной содержательной гипотезе). В этом случае психологически негативный результат (неотклонение нулевой гипотезы) воспринимается как неудача. Работы, в которых не выявлен искомый эффект, остаются втуне. Невыявление статистически значимого эффекта может быть вызвано двумя причинами:

1) отсутствием этого эффекта в объективной реальности;

2) недостаточным объемом выборок.

Конечно, авторы обычно не без основания надеются, что в их случае имеется вторая ситуация и продолжают набор наблюдений. Если же это не помогает, и остается только первый вариант, такое исследование чаще всего не публикуется, и опыт исследователей остается скрытым от

научного сообщества, что неправильно. Более того, такая ситуация признается неэтичной, поскольку приводит к бессмысленной трате финансовых и человеческих ресурсов на напрасное прохождение уже пройденной кем-то дороги, ведущей в тупик.

Описательная статистика

Описательная статистика нужна для обобщенного числового представления результатов и может решать следующие 3 задачи:

1. Описывать центральную тенденцию и вариабельность количественных данных.

2. Описывать частоты (абсолютные и относительные) качественных данных.

3. Описывать величину изучаемого эффекта.

Чаще всего исследователи хотят решить первую задачу, имея нечеткие представления о тех мерах центральной тенденции и вариабельности, которые существуют. В качестве меры центральной тенденции обычно используется среднее арифметическое (M) — параметр, известный из школьного курса математики и интуитивно понятный. Однако среднее не всегда может правильно характеризовать выборку: в случае распределения, отличного от нормального (Гауссова), среднее арифметическое неприменимо. В этом случае следует пользоваться медианой (Me). Казалось бы, при наличии современных пакетов анализа данных проблем с вычислением медиан не должно возникать, однако препятствием является то, что появляется необходимость понять, каково же распределение — нормальное или нет? А решение этой задачи представляет определенные трудности — ситуация усугубляется отсутствием знаний о критериях выбора и, следовательно, предпочтением легкого решения — выбором среднего арифметического. Однако выбор легкого пути — не всегда выбор правильного пути.

К счастью, медиана может применяться в качестве меры центральной тенденции не только в случае распределений, отличных от нормального, но и для нормальных распределений (в этом случае среднее и медиана попросту совпадают). Таким образом, проблема выбора между средним и медианой снимается (в пользу последней). А с учетом того, что нормальные распределения встречаются не более чем в 20% случаев [3] — во многом потому, что выборки в большинстве работ являются небольшими — можно смело рекомендовать всегда использовать медиану. В этом случае ошибка исключена.

Ситуация с мерами вариабельности аналогична. При нормальном распределении (и только в этом случае) можно использовать среднеквадратическое отклонение (s). В то же время для любого распределения (нормального либо нет) можно применять нижний и верхний квартили (lower quartile, LQ, upper quartile, UQ) — значения признака, отсекающие по 25% объектов выборки в левом и правом «хвостах» распределения количественного признака.

Отдельная проблема — порочная традиция использовать в качестве меры вариабельности стандартную ошибку среднего (m) — величину, характеризующую лишь точность оценки среднего арифметического на основании данных выборки. Использование этого параметра позволяет маскировать бессмысленные ситуации, когда для признаков, имеющих по определению только положительные значения (например, концентрации) величина параметра вариабельности превышает величину параметра центральной тенденции (3 ± 5), что происходит часто в распределениях, не являющихся нормальными. Дело в том, что стандартная ошибка среднего обычно в несколько раз меньше среднеквадратического отклонения, и запись 3 ± 1 выглядит намного приятнее. Однако, к нашему удивлению, многих авторов возникающая ситуация не смущает (или они ее вовсе не замечают?), и мы видим в свежайших журналах не последнего ряда такие вопиющие ошибки.

С описательной статистикой качественных признаков проще: подсчитать число случаев с тем или иным значением качественного признака (абсолютные частоты) и вычислить проценты относительные частоты — задача несложная. Отметим только, что вычисление процентов корректно, если число наблюдений не слишком мало. Иначе может получиться, что 1 больной — это 20%.

Еще хуже положение с описанием величины эффекта. Большинство российских исследователей не использует такие общепринятые меры, как абсолютный риск (absolute risk, AR) и относительный риск (ОР; relative risk, RR), отношение шансов (ОШ; odds ratio, OR) и др., для описания величины изучаемого ими эффекта. Подробно об этих показателях можно прочитать в специальной литературе [1, 4, 5].

Выявление эффектов

Следующий этап научной работы обычно формулируется исследователями как «сравнить группы» и «посчитать корреляции». Если первая из этих задач очевидна и отражает цели исследования, то вторая — обычно абсолютно ритуальная («так сказал научный руководитель»). Часто авторы не имеют представления о том, что корреляционный анализ может выполняться только для количественных данных (для качественных данных аналогом является анализ ассоциаций, при котором используются иные статистические критерии). Кроме того, интерпретация результатов корреляционного анализа должна проводиться очень осторожно: нельзя допускать однозначной интерпретации корреляционной связи как причинно-следственной (о критериях причинно-следственной связи см. [4]).

Задача «сравнить группы» может быть решена двумя способами:

1) проверкой нулевой статистической гипотезы об отсутствии различий групп;

2) построением доверительных интервалов для параметров центральной тенденции распределения либо для показателя величины эффекта.

Остановимся сначала на первом способе (подходе) — проверке гипотез.

Проверка гипотез применяется для сравнения групп традиционно широко. В статистике разработано множество методов для решения таких задач, что приводит к необходимости выбора статистических критериев. Правильный выбор может быть сделан после ответа на несколько вопросов:

1. Какие признаки сравниваются в группах — количественные или качественные (порядковые, номинальные, бинарные)?

2. В случае количественных признаков: каковы распределения этих признаков в каждой из сравниваемых групп — нормальные или нет? Равны ли дисперсии в группах?

3. Являются ли выборки независимыми (несвязанными) или зависимыми (связанными)?

4. Какое количество групп сравнивается — 2, 3 или более?

Поскольку ответить на некоторые из этих вопросов типичному аспиранту затруднительно, выбор обычно делается в пользу *t*-критерия Стьюдента, традиционно применяемого в течение многих десятилетий для сравнения групп по количественным признакам. Такая традиция связана с тем, что значение критерия Стьюдента в классическом варианте относительно легко рассчитать «на бумажке» (в отличие от многих других тестов), т.е. и в докомпьютерную эпоху это было несложно, однако он широко используется и в настоящее время. Проблема же состоит в том, что корректное применение этого теста возможно лишь в небольшой доле случаев — при наличии нормальных распределений признаков в каждой из сравниваемых групп. В случае если данное условие выполняется, авторы должны выбрать адекватный вариант критерия Стьюдента — для независимых либо зависимых групп, для групп с равными либо с неравными дисперсиями. При изложении результатов следует упомянуть, какой из вариантов теста использовался. К сожалению, это почти никогда не делается.

В настоящее время при доступности компьютерных программных средств анализа данных рекомендуется максимально широко пользоваться непараметрическими методами сравнения групп. Эти методы не налагают никаких ограничительных условий на данные, т.е. могут применяться в случае любых распределений количественных или порядковых признаков. Вторым преимуществом этих методов является то, что они устойчивы к высокой вариабельности данных: на их результаты мало влияют «выбросы» (крайне малые или крайне большие значения). Следствием этих обстоятельств является их применимость к анализу так называемых малых (менее 30 случаев) выборок, характерных для медицинских и биологических исследований.

Приведем упрощенную таблицу по выбору адекватных методов сравнения групп (см. таблицу). Известно, что использование неподходящих методов проверки гипотез приводит к искажению результатов и формулировке неверных выводов [3], что для клинической медицины является неприемлемым ни с научной, ни с этической точки зрения.

В процессе проверки гипотезы вычисляют значение статистического критерия и уровень статистической значимости p, который сравнивают с заданным априори пороговым уровнем значимости p_o. Ранее (в докомпьютерную эпоху) было принято обозначать лишь интервал, в который попадает эта величина (например, $p<0,02$). В настоящее время она вычисляется компьютерами с большой точностью, в связи с чем возникла другая тенденция — приводить вычисленное значение с точностью до тысячных долей (например, $p=0,012$). Такой подход более прозрачен — позволяет читателю самостоятельно оценить, велика ли ошибка I рода по отношению к пороговому уровню p_o. Раньше приходилось лишь слепо доверять выводу автора.

При $p<p_o$ нулевая гипотеза (об отсутствии различий групп и т.п.) отклоняется, и принимается альтернативная гипотеза — о том, что различия групп существуют и являются статистически значимыми. Подчеркнем, что, если по результатам теста нулевая гипотеза не отклоняется, это не означает, что различия групп отсутствуют. Причин может быть две: недостаточные объемы выборок и/или отсутствие эффекта.

При проверке гипотез нельзя забывать о так называемой проблеме множественных сравнений (ПМС). Она заключается в следующем: чем больше статистических гипотез проверяется на одних и тех же данных, тем более вероятно заключение о наличии различий между группами (либо наличии статистической связи признаков), в то время как на самом деле верна нулевая гипотеза об отсутствии различий/связей. Так, если за пороговый уровень значимости принято значение $p_o=0,05$, то 5 из 100 вычисленных значений p в силу случайности (по теории вероятности) окажется меньше 0,05 (хотя на самом деле верна нулевая гипотеза об отсутствии различий). На практике принято считать, что учет ПМС следует начинать в тех случаях, когда число рассчитываемых значений p (и соответственно публикуемых в статьях и диссертациях) превышает 10. Для преодоления проблемы множественных сравнений может применяться поправка Бонферрони (вычисление порогового уровня статистической значимости с учетом числа проверенных гипотез), специальные методы множественного сравнения групп, дисперсионный анализ.

Построение доверительных интервалов (ДИ) — второй способ выявления эффектов (различий групп, связей признаков).

ДИ — интервал значений признака, рассчитанный по выборке для какого-либо параметра распределения или показателя величины эффекта, и с определенной вероятностью (например, 95% в случае 95% ДИ) включающий истинное значение этого параметра/показателя. Ширина ДИ зависит от объема выборки и вариабельности в ней. Чем шире ДИ, тем менее точной является выборочная оценка. При увеличении числа наблюдений (объектов исследования) ДИ сужается и точность оценки увеличивается.

Наиболее распространенные методы сравнения групп

Тип признака	2 независимые группы	2 зависимые группы	3 независимые группы и более
Количественный признак: нормальные распределения	*t*-критерий Стьюдента для независимых групп	*t*-критерий Стьюдента для зависимых групп	ANOVA по Пирсону
Количественный признак: любые распределения	Критерий Манна—Уитни	Критерий Вилкоксона	ANOVA по Крускалу—Уоллису
Качественный порядковый признак	Критерий Манна—Уитни (при числе значений признака более 5), χ^2	Критерий Вилкоксона	ANOVA по Крускалу—Уоллису (при числе значений признака более 5), χ^2
Качественный номинальный признак	χ^2	Критерий Кокрана	χ^2
Бинарный признак	Точный критерий Фишера	Критерий МакНемара	χ^2

В настоящее время в зарубежной научной медицинской литературе для представления результатов исследования ДИ используются очень широко, а в ряде изданий это является обязательным требованием. Это обусловлено следующим:

1) ДИ наглядно представляет спектр значений признака/показателя;

2) ДИ позволяет оценить степень различий групп при заданном доверительном коэффициенте (обычно равном 95%). Применение ДИ основано на простых правилах:

— если два ДИ для параметров (например, средних или медиан) двух групп пересекаются, то статистически значимых различий этих групп нет;

— если ДИ для коэффициента корреляции включает ноль, то статистической связи признаков нет;

— если ДИ для ОР или ОШ включает 1, то статистически значимого эффекта нет.

Отметим различия двух указанных подходов: проверка гипотез позволяет оценить вероятность неверного отклонения нулевой гипотезы (однако ничего не говорит о вероятности альтернативной гипотезы), а ДИ позволяет при фиксированном уровне доверия (т.е. фиксированной допускаемой ошибке) оценить размер эффекта и его точность.

Интерпретация и обсуждение результатов

Независимо от того, какой подход использовался при анализе данных — проверка гипотез или построение ДИ — существенное внимание в публикации должно быть уделено корректной интерпретации и обсуждению результатов. Обязательно следует обсудить клиническую значимость результатов, которая не всегда совпадает с их статистической значимостью. Так, статистически значимый результат, полученный на большой выборке, может незначительно влиять на клиническую практику, в то время как статистически незначимый эффект (например, высокая эффективность нового хирургического метода), изучавшийся на малой выборке, может быть клинически значимым, что дает основание для дальнейших исследований.

Как мы уже упоминали выше, важнейшая цель статистического анализа — сделать вывод о существовании некой общей закономерности на основании анализа ограниченного числа наблюдений. Например, вывод о том, что такие же пациенты, как обследованные, будут реагировать на новое лекарство таким же образом. В связи с этим при обсуждении результатов главным должно быть формулирование их практической значимости. Последняя определяется так называемой обобщаемостью исследования. Чем строже критерии включения и исключения объектов исследования, тем уже популяция, для которой полученные результаты могут быть полезными: результаты исследования применимы только к таким же (по возрасту, полу, форме и стадии болезни и т.д.) больным, что и изученные.

К сожалению, подготовка медиков в российских вузах не включает дисциплины, которые могли бы привить знания и навыки исследователя, и врачи, заинтересовавшиеся научной работой, вынуждены осваивать методологию медицинских исследований, в том числе статистический анализ данных, самостоятельно.

В то же время за последние годы выпущено много литературы на русском языке как для исследователей, так и для врачей. Это позволяет осуществлять непрерывное самообразование, что важно для каждого человека как в профессиональном, так и в личностном контексте.

ЛИТЕРАТУРА

1. *Власов В.В.* Эпидемиология. М: Гэотар-Мед 2004; 68, 250—253, 332.
2. *Власов В.В.* Значение научных публикаций в специализированных журналах. Рос вестн акуш-гин 2010; 10: 4: 72—75.
3. *Леонов В.П.* Ошибки статистического анализа биомедицинских данных. Междунар журн мед практики 2007; 2: 19—35; http://www.biometrica.tomsk.ru/error.htm
4. *Реброва О.Ю.* Статистический анализ медицинских данных: Применение пакета прикладных программ Statistica. М: Медиа Сфера 2006; 166—176.
5. *Флетчер Р., Флетчер С., Вагнер Э.* Клиническая эпидемиология: Основы доказательной медицины. М: Медиа Сфера 2004; 297.