А. Е. ФИЛАТОВА

А. Е. Филатова, доцент кафедры вычислительной техники и программирования Национального технического университета «Харьковский политехнический институт», кандидат технических наук, доцент

Основные принципы обработки медицинских и медико-биологических данных в научных исследованиях

Введение

Ни одно серьезное медицинское или медико-биологическое научное исследование не обходится без статистического анализа экспериментальных данных. При этом перед исследователем сразу возникает много вопросов. Какой должен быть объем экспериментальной выборки? С чего начать обработку экспериментальных данных? Как описать данные? Какие методы использовать для подтверждения выводов или найденных зависимостей? И это далеко не полный список вопросов. На первый взгляд можно потеряться в многообразии статистических методов и математических понятий. К сожалению, это часто приводит к некорректному применению статистических методов или к ошибочным выводам по полученным с помощью статистических пакетов данным. Исследователю надо помнить, что статистические пакеты обрабатывают просто цифры, и только исследователь знает, какие физические единицы измерения и смысл стоят за этими цифрами. Однако достаточно понять некоторые принципы статистической обработки экспериментальных данных, и выбор метода для обработки не будет уже вызывать столько вопросов у исследователя.

Типы данных, используемые в исследованиях

В первую очередь для адекватного выбора статистических методов обработки экспериментальных данных необходимо понять, в какой шкале измерен тот или иной признак, включенный в исследование.

Исходные данные по возможностям их дальнейшей обработки и применяемым методам делят на три типа шкал признаков [2]:

- количественные шкалы, когда каждый признак определен числовым эквивалентом (например, температура — градусами Цельсия (°С), данные о весе — единицами измерения массы (кг, г, мг), данные о размерах — единицами измерения длины (мм, см, м) и т. п.);
- ординальные или порядковые шкалы, когда в соответствие каждому признаку может быть поставлено каким-то условным образом определенное число, характеризующее степень проявления изучаемого свойства. Например, признак оценивается врачом с помощью специально разработанных шкал типа «норма», «плохо», «очень плохо». То есть существует несколько градаций, но между ними нельзя установить никаких

количественных отношений, можно говорить только о порядке или ранге. Порядковая шкала может изменяться: а) от одной противоположности до другой и тогда допустимые значения располагаются по обе стороны относительно нуля — точки неопределенности; б) от точки отсутствия качества до точки наивысшего его проявления — и тогда признаку придаются только положительные значения;

- номинальные признаки, когда между возможными значениями признаков нельзя априорно установить ни количественных отношений, ни отношений порядка. То есть при сравнении двух значений признака можно говорить только о совпадении или несовпадении этих значений. Единственной их количественной оценкой служит частота встречаемости (число мужчин и женщин, число лиц с голубыми или зелеными глазами, курильщиков и не курильщиков, утомленных и отдохнувших, сильных и слабых и т.п.). Частным случаем номинальной шкалы являются дихотомические или бинарные признаки, когда каждый признак оценивается одним из двух состояний (ответ на вопрос: «да» или «нет»; состояние «0» или «1»).

Ординальные и номинальные признаки еще называют качественными или категориальными признаками. Для качественных признаков нельзя определить расстояния между значениями, так как для них не существуют интервальные измерительные шкалы.

Кроме того, количественные признаки подразделяются на непрерывные и дискретные.

Непрерывные признаки — это количественные признаки, которые получают при измерении на непрерывной шкале, т. е. теоретически они могут иметь дробную часть (например, данные о температуре, артериальном давлении, весе, росте пациента и т. д.). Для непрерывных количественных признаков определены все арифметические операции, и к их значениям применимы любые статистические процедуры.

Дискретные признаки — это количественные признаки, которые не могут иметь дробную часть (например, количество детей, лейкоцитов и т. д.).

В свою очередь непрерывная шкала может быть интервальной или шкалой отношений.

Интервальные признаки — вид непрерывных количественных признаков, которые измеряются в абсолютных величинах, имеющих физический смысл.

Шкала отношений — количественная шкала, позволяющая отразить долю изменения (увеличения или уменьшения) значения признака по отношению к исходному или какому-либо другому значению этого признака. Признаки, выраженные в шкале отношений, являются безразмерными величинами или выражаются в процентах.

Приведенную выше классификацию типов признаков в зависимости от метода их дальнейшей обработки можно представить в виде, показанном на рис.

Описательная статистика

В научных исследованиях практически никогда нет возможности получить данные обо всех объектах совокупности, подлежащих изучению. Такую совокупность объектов называют генеральной

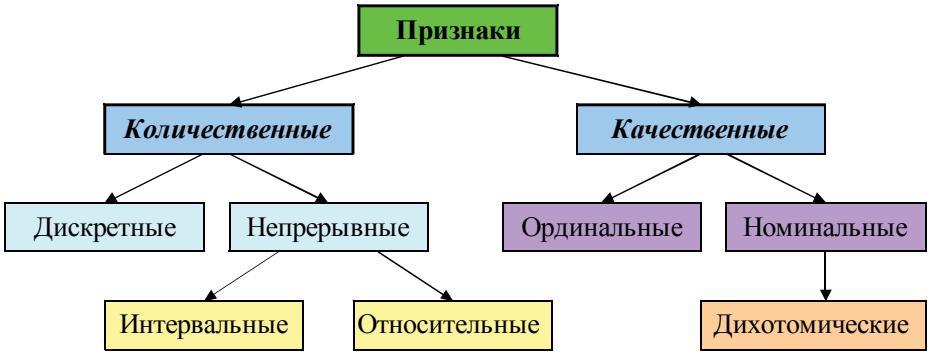

Рис. Классификация типов признаков

совокупностью. Исследователь практически всегда имеет дело с ограниченным набором объектов, т. е. с выборкой из генеральной совокупности, полагая, что эта выборка отражает свойства всей совокупности объектов. Поэтому все статистические характеристики, полученные на основании изучения выборки, называются оценками, т. е. исследователь имеет дело не с точными значениями среднего или стандартного отклонения, а с их оценками [1]. Так, например, оценка среднего, полученная по выборке, называется выборочным средним, оценка стандартного отклонения — выборочным стандартным отклонением и т. д. Но раз исследователь не может получить точные значения статистических параметров, а разные выборки из одной генеральной совокупности дадут разные оценки, то для характеристики точности выборочных оценок используется стандартные ошибки, которые можно рассчитать для любых статистических показателей.

Надо помнить, что такие статистические параметры, как среднее, дисперсия, стандартное отклонение, медиана и другие, можно рассчитывать только для количественных признаков. Очевидно, что глупо рассчитывать среднее значение пола пациентов выборки или степени проявления болезни. Если есть предположение, что исходные данные распределены по нормальному закону распределения, то для описания всей совокупности часто бывает достаточно указать выборочное среднее и выборочное стандартное отклонение, которое показывает разброс значений относительно среднего. При этом примерно 68 % значений отличается от среднего (в большую или меньшую сторону) не более чем на одно стандартное отклонение и примерно 95 % — на два стандартных отклонения [1, 2]. Надо отметить, что нормальный закон распределения полностью определяется двумя параметрами — средним m и стандартным отклонением s. Поэтому при описании выборки и используется запись $M \pm s$, где M — выборочное среднее; s — выборочное стандартное отклонение.

И вот тут кроется одна из распространенных ошибок, встречаемых в публикациях. Часто исследователи путают стандартное отклонение со стандартной ошибкой среднего, указывая в статье $M \pm m$, где m — стандартная ошибка среднего, которая в $\sqrt{n}$ раз меньше выборочного стандартного отклонения (n — длина или объем выборки). Поскольку стандартная ошибка всегда меньше выборочного стандартного отклонения, то исследователю кажется более привлекательным указывать не разброс значений, а точность вычисления оценки.

Для описания данных, которые распределены по закону, сильно отличающемуся от нормального, необходимо использовать медиану и процентили (или квантили). Это особенно важно при описании данных с асимметричными законами распределения [1]. Процентиль — это значение случайной величины, характеризующее площадь функции плотности вероятности, находящуюся слева от этого значения и выраженную в процентах. Квантиль — это такая же случайная величина, как и процентиль, только площадь выражается в частях. Так, например, значение, не выше которого оказалось 25% результатов измерений, называют 25-м процентилем (или квантилем 0,25), а значение, не выше которого оказалось 75 % результатов измерений — 75-м процентилем (или квантилем 0,75). Отсюда вытекает, что медиана — это 50-й процентиль или квантиль 0,5 (т. е. значение, которое делит распределение пополам). Квантили 0,25 и 0,75 еще называют нижним и верхним квартилями.

Надо помнить, что медиана и процентили (квантили) не дают полного описания распределения (в отличие от среднего и стандартного отклонения), но при этом между 25-м и 75-м процентилями (т. е. между нижним и верхним квартилями) находится половина всех значений случайной величины. Для того чтобы найти оценки процентилей (т. е. вычислить их по выборке), необходимо отсортировать выборку по возрастанию значений. Тогда медиана — это среднее по номеру значение в отсортированной выборке, т. е. значение с номером $(n + 1)/2$; 25-й процентиль (нижний квартиль) — это значение с номером $(n + 1)/4$, а 75-й процентиль (верхний квартиль) — это значение с номером $3(n + 1)/4$. Если объем выборки таков, что целочисленное деление невозможно, то в качестве искомого процентиля (квантиля) принимается аппроксимированное значение, находящееся между двух элементов с целочисленными номерами справа и слева от полученного

дробного значения. Например, n=29, тогда при делении на 4 получаем значение 7,5, т. е. нижний квартиль равен среднему значению 7-го и 8-го элемента отсортированной выборки.

Проверка статистических гипотез

Кроме описания исходных выборок часто исследователю нужно доказать статистическую значимость результатов проведенных опытов. Тут на помощь приходят методы, связанные с проверкой статистических гипотез.

Статистическая гипотеза — это утверждение относительно одного или нескольких параметров генеральной совокупности или о форме распределения генеральной совокупности, сделанное на основании выборки.

Методы, позволяющие оценить статистическую значимость отличий, называют критериями значимости или просто критериями [1, 2]. Существует большое многообразие критериев, но в основу всех критериев положены одни и те же принципы. В первую очередь формулируется, так называемая, нулевая гипотеза H_o, смысл которой заключается в том, что полученные отличия случайны. Например, параметры распределений выборок (средние, дисперсии) равны, связи между выборками отсутствуют, частота встречаемости исходов качественного признака не зависит от выборки и т. д. Соответственно обратное утверждение называется альтернативной гипотезой и обозначается H_1.

При проверке статистических гипотез допускаются ошибки, которые бывают двух типов. Ошибка 1-го рода (обозначается a) — это вероятность того, что гипотеза H_o отвергнута, хотя на самом деле она верна. Ошибка 2-го рода (обозначается b) — это вероятность принятия гипотезы H_o, хотя на деле она не верна. Другими словами, ошибка 1-го рода характеризует вероятность «пропуска цели», а ошибка 2-го рода — «ложную цель». Поскольку всегда проверяется нулевая гипотеза, то и для проверки задается лишь желаемая ошибка 1-го рода. Поэтому величина a называется уровнем значимости критерия и обычно задается равной 0,1 (10 %), 0,05 (5 %) или 0,01 (1 %). Чем меньше значение уровня значимости, тем меньше вероятность найти отличия там, где их нет.

Статистические критерии

Для проверки нулевой гипотезы используют специально подобранную случайную величину, точное или приближенное распределение которой известно. Например, если такая величина распределена по нормальному закону, то ее обозначают через z, если по закону Стьюдента — t, если по закону Фишера — F, если по закону «хи-квадрат» — χ^2 и т. д. Часто по названию закона распределения этой случайной величины называют критерий, например, критерий Стьюдента, критерий Фишера, критерий «хи-квадрат» и т.д. Все эти случайные величины обозначим через K. Таким образом, статистический критерий — это случайная величина K (распределенная по соответствующему закону), которая служит для проверки нулевой гипотезы [1–3].

Множество всех возможных значений статистического критерия делится на два непересекающихся подмножества. Первое подмножество включает в себя те значения критерия, при которых основная гипотеза отвергается. Это подмножество называется критической областью. Второе подмножество включает в себя те значения критерия, при которых основная гипотеза принимается. Значение случайной величины, которое делит все возможные значения статистического критерия на указанные подмножества, называется критическим значением $K_{кр}$. Критическое значение $K_{кр}$ равно процентилю соответствующего закона распределения, который определяется исходя из уровня значимости.

По виду критической области критерии бывают односторонние и двусторонние [2]. Вид критической области определяется видом альтернативной гипотезы. Как было отмечено выше, если нулевая гипотеза отвергается, то принимается альтернативная. При этом ее тоже нужно формулировать. Для наглядности рассмотрим пример.

Пусть в результате эксперимента была получена оценка среднего M. Из теоретических данных известно, что это значение должно быть равно M_0. Формулируем нулевую гипотезу как утверждение о том, что нет отличий между полученной оценкой и теоретическим значением среднего, т. е. $H_0: M = M_0$. А что же принимается в случае, когда H_0 отвергается? Очевидно, что есть существенные

отличия между оценкой и теоретическим значением среднего, но эти отличия могут быть разными: можно утверждать, что оценка и истинное значение среднего просто не равны, что оценка существенно больше истинного значения среднего, и, наконец, что оценка существенно меньше истинного значения среднего. Таким образом, есть 3 варианта альтернативной гипотезы:

1) $H_1 : M \neq M_0$;
2) $H_1 : M > M_0$;
3) $H_1 : M < M_0$.

В первом случае речь идет о так называемой двусторонней гипотезе. Т. е. исследователю не важно, в какую сторону от истинного значения среднего отклонена оценка (в большую или меньшую), поэтому величина ошибки 1-го рода делится на 2 равные части, следовательно, и критическая область делится на 2 части. В этом случае и говорят о двустороннем критерии. Тогда критическая область определяется парой неравенств: $K < K_{кр1}$, $K > K_{кр2}$, где $K_{кр1}=K_{\alpha/2}$, $K_{кр2}= K_{1-\alpha/2}$ — левое и правое критическое значение (соответствующие процентили). Если критические точки симметричны относительно нуля, т. е. $K_{\alpha/2}=K_{1-\alpha/2}$, то пара неравенств объединяется в одно неравенство $|K| > K_{кр}$, где $K_{кр} = K_{1-\alpha/2}$.

В остальных двух случаях говорят об односторонней гипотезе, и, соответственно, об одностороннем критерии. При этом различают правосторонний и левосторонний критерии, и соответствующие им критические области. Если критическая область определяется неравенством $K > K_{кр} = K_{1-\alpha}$, то критерий правосторонний, а если $K < K_{кр} = K_{\alpha}$, то критерий левосторонний.

Вероятность справедливости нулевой гипотезы

Как было отмечено выше, для проверки статистической гипотезы необходимо сформулировать нулевую и альтернативную гипотезы и определиться с критерием. При этом для проверки справедливости нулевой гипотезы необходимо знать критическое значение $K_{кр}$, которое выбирается по специальным таблицам и с которым сравнивается расчетное значение критерия K. Для исследователя это не очень удобно, поэтому все статистические пакеты показывают не только значение K, но и величину p. Дадим упрощенное определение p.

P — это вероятность ошибочно опровергнуть нулевую гипотезу. Другими словами, по наблюдаемому значению K определяется как бы расчетная ошибка 1-го рода, т. е. сумма вероятностей всех случайных величин, которые попадают в расчетную критическую область [1]. Если для одностороннего критерия $p>\alpha$, то нулевая гипотеза справедлива, т. е. между значениями или явлениями нет статистических различий. В противном случае нулевая гипотеза отвергается. Например, если необходимо доказать однородность выборок, то должно быть $p>\alpha$.

Выбор статистического критерия

В зависимости от типа обрабатываемых данных и предположений о законах распределения исследуемых выборок критерии делятся на параметрические и непараметрические [1–3].

Параметрические критерии — это статистические критерии, для вычисления которых необходимо использовать параметры распределения (например, средние, стандартные отклонения и т. д.). То есть такие критерии можно использовать только для количественных признаков. При этом большинство параметрических критериев основывается на том, что случайная величина распределена по нормальному закону распределения.

Если же признаки качественные или не подтверждается гипотеза о том, что количественный признак распределен по нормальному закону распределения, то необходимо использовать непараметрические методы.

Непараметрические критерии — это статистические критерии, для вычисления которые не нужно знать параметры распределения; эти критерии оперируют частотами или рангами.

Очень важно запомнить, что нельзя использовать параметрические методы для качественных признаков, в том числе и для ординальных. Нужно хорошо понимать, что если признак выражен в ординальной шкале в виде чисел (например, степень проявления болезни), то с помощью статистического пакета можно использовать как непараметрический, так и параметрический

критерий, хотя использование последнего является некорректным и ошибочным. Дело в том, что пакет оперирует числами, не учитывая физического смысла этих чисел, поэтому выбор метода полностью лежит на ответственности исследователя.

Надо отметить, что всегда можно количественную шкалу представить в виде ординальной, ординальную шкалу в виде номинальной. Но нельзя перевести номинальную шкалу в ординальную, а ординальную шкалу в количественную. То есть переходить от более «сильных» шкал к менее «сильным» можно, а наоборот нельзя.

Такой переход необходим, когда, например, нужно проверить наличие связи между количественным и качественным признаком. В этом случае количественный признак приходится переводить в качественную шкалу и применять для обработки непараметрические методы.

Заключение

В заключение хотелось бы отметить, что в данной работе рассмотрена лишь малая часть вопросов, которые затрагивает математическая статистика. К сожалению, в рамках статьи невозможно рассмотреть все нюансы статистической обработки данных в научных исследованиях. Но рассмотренные вопросы очень важны для понимания принципов, логики статистической обработки медицинских и медико-биологических данных. Более детальное изучение вопросов медико-биологической статистики рекомендуем начать с книг и пособий, приведенных в списке литературы к данной статье.

Список литературы

1. Гланц С. Медико-биологическая статистика / С. Гланц. — М. : Практика, 1999. — 459 с.
2. Петри А. Наглядная статистика в медицине / А. Петри, К. Сэбин. — М. : Гэотар-Мед, 2003. — 144 с.
3. Реброва О. Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О. Ю. Реброва. — М. : Медиа Сфера, 2006. — 312 с.

Резюме

Основные принципы обработки медицинских и медико-биологических данных в научных исследованиях

А. Е. Филатова

В данной статье рассматриваются основные понятия теории математической статистики, необходимые для уяснения логики статистического анализа медицинских и медико-биологических данных в научных исследованиях. В работе рассмотрены типы данных, используемые в исследованиях, основные понятия описательной статистики, проверка статистических гипотез, а также показаны типичные ошибки статистической обработки и ее описания, встречающиеся в научных медицинских и медико-биологических статьях.

Ключевые слова: математическая статистика, типы данных, описательная статистика, статистическая гипотеза.

Summary

Key Principles for Processing of Medical and Biomedical Data in Research Studies

A. Ye. Filatova

In this paper we review basic concepts of the mathematical statistics theory which are essential for understanding the logic of the statistical analysis of medical and biological research data. We discuss frequently used data types, the basic concepts of descriptive statistics, and statistical hypothesis testing. In addition we describe typical errors of statistical analysis and its description occurring in scientific medical and biological papers.

Key words: mathematical statistics, data types, descriptive statistics, statistical hypothesis.

Основні принципи обробки медичних та медико-біологічних даних у наукових дослідженнях

Г. Є. Філатова

У даній статті розглядаються основні поняття теорії математичної статистики, необхідні для з'ясування логіки статистичного аналізу медичної та медико-біологічних даних в наукових дослідженнях. У роботі розглянуті типи даних, що використовувані в дослідженнях, основні поняття описової статистики, перевірка статистичних гіпотез, а також показані типові помилки статистичної обробки та її опису, що зустрічаються в наукових медичних та медико-біологічних статтях.

Ключові слова: математична статистика, типи даних, описова статистика, статистична гіпотеза.