УДК 519.2:303.732.4

01.00.00 Физико-математические науки

РАСПРЕДЕЛЕНИЯ РЕАЛЬНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ НЕ ЯВЛЯЮТСЯ НОРМАЛЬНЫМИ

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н., профессор
РИНЦ SPIN-код: 4342-4994
Московский государственный технический университет им. Н.Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., 5, prof-orlov@mail.ru

В учебных курсах по теории вероятностей и математической статистике рассматривают различные параметрические семейства распределений числовых случайных величин. А именно, изучают семейства нормальных распределений, логарифмически нормальных, экспоненциальных, гамма-распределений, распределений Вейбулла - Гнеденко и др. Все они зависят от одного, двух или трех параметров. Поэтому для полного описания распределения достаточно знать или оценить одно, два или три числа. Широко развита параметрическая теория математической статистики, в которой предполагается, что распределения результатов наблюдений принадлежат тем или иным параметрическим семействам. Эта традиция идет от Карла Пирсона, который в начале XX в. предложил использовать четырехпараметрическое семейство распределений. Перечисленные выше семейства распределений - это подмножества четырехпараметрического семейства Пирсона. К сожалению, параметрические семейства существуют лишь в головах авторов учебников по теории вероятностей и математической статистике. В реальной жизни их нет. Поэтому современная прикладная статистика и эконометрика используют в основном непараметрические методы, в которых распределения результатов наблюдений могут иметь произвольный вид. Сначала на примере нормального распределения обсуждаем невозможность практического использования параметрических семейств для описания распределений конкретных экономических данных. Приводим результаты исследований метрологов и оценки сходимости в предельных теоремах. Затем разбираем параметрические методы отбраковки резко выделяющихся наблюдений. Весьма неустойчивы как уровни значимости при фиксированном правиле отбраковки, так и параметр правила отбраковки при фиксированном уровне значимости. Следовательно, отбраковка по классическим правилам математической статистики не является научно

UDC 519.2:303.732.4

Physics and mathematical sciences

DISTRIBUTIONS OF REAL STATISTICAL DATA ARE NOT NORMAL

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci., professor
Bauman Moscow State Technical University, Moscow, Russia

In the training courses on the theory of probability and mathematical statistics there are various parametric families of distributions of numerical random variables considered. Namely, we have been studying the families of normal distributions, log-normal distributions, exponential distributions, gamma distributions, Weibull-Gnedenko distributions, etc. All of them depend on one, two or three parameters. Therefore, for a complete description of the distribution it is sufficient to know or estimate one, two or three numbers. Parametric theory of mathematical statistics is widely developed, where it is assumed that the distribution of observations belong to one or another parametric family of distributions. This tradition comes from Karl Pearson, who in the early twentieth century proposed the use of four parametric family of distributions. The above families of distributions - are the subsets of a four-parametric family of Pearson. Unfortunately, parametric families exist only in the minds of the authors of textbooks on probability theory and mathematical statistics. In real life, they are not. Therefore, modern applied statistics and econometrics mainly use non-parametric methods, in which the distribution of observations can have arbitrary form. First, on an example of a normal distribution, we are discussing the impossibility of practical use of parametric families of distributions to describe specific statistical data. We give the results of research of metrologists and estimation of convergence in limit theorems. Then we discuss how the parametric methods can use for reject outlying observations. It is very unstable the significance levels for a fixed rejection rule and the parameter of the rejection rules for a fixed level of significance. Consequently, the rejection of the classic rules of mathematical statistics is not science-based

обоснованной

Ключевые слова: МАТЕМАТИЧЕСКАЯ СТАТИСТИКА, ПРИКЛАДНАЯ СТАТИСТИКА, СТАТИСТИЧЕСКИЕ МЕТОДЫ, НЕПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА, ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ, НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ, СТАТИСТИЧЕСКИЕ ДАННЫЕ, ПРОВЕРКА СОГЛАСИЯ РАСПРЕДЕЛЕНИЯ, ВЫБРОСЫ, ОТБРАКОВКА, УСТОЙЧИВОСТЬ

Keywords: MATHEMATICAL STATISTICS, APPLIED STATISTICS, STATISTICAL METHODS, NONPARAMETRIC STATISTICS, STATISTICAL HYPOTHESIS TESTING, NORMAL DISTRIBUTION, STATISTICAL DATA, GOODNESS-OF-FIT TEST, OUTLIER, REJECTION, STABILITY

1. Введение

В учебных курсах по теории вероятностей и математической статистике рассматривают различные параметрические семейства распределений числовых случайных величин. А именно, изучают семейства нормальных распределений, логарифмически нормальных, экспоненциальных, гамма-распределений, распределений Вейбулла - Гнеденко и др. Все они зависят от одного, двух или трех параметров. Поэтому для полного описания распределения достаточно знать или оценить одно, два или три числа. Очень удобно. Поэтому широко развита параметрическая теория математической статистики, в которой предполагается, что распределения результатов наблюдений принадлежат тем или иным параметрическим семействам. Эта традиция идет от Карла Пирсона, который в начале XX в. предложил использовать четырехпараметрическое семейство распределений [1]. Перечисленные выше семейства распределений - это подмножества четырехпараметрического семейства Пирсона.

К сожалению, параметрические семейства существуют лишь в головах авторов учебников по теории вероятностей и математической статистике. В реальной жизни их нет. Поэтому современная прикладная статистика [2 - 4] и эконометрика [5] используют в основном непараметрические методы [6, 7], в которых распределения результатов наблюдений могут иметь произвольный вид.

Сначала на примере нормального распределения достаточно подробно обсудим невозможность практического использования параметрических семейств для описания распределений конкретных статистических данных. Затем разберем параметрические методы отбраковки резко выделяющихся наблюдений и продемонстрируем невозможность практического использования ряда методов параметрической статистики, покажем ошибочность выводов, к которым они приводят.

2. Часто ли распределение результатов наблюдений является нормальным?

В эконометрических и экономико-математических моделях, применяемых, в частности, при изучении и оптимизации процессов маркетинга и менеджмента, управления предприятием и регионом, точности и стабильности технологических процессов, в задачах надежности, обеспечения безопасности, в том числе экологической, функционирования технических устройств и объектов, разработки организационных схем часто применяют понятия и результаты теории вероятностей и математической статистики. При этом зачастую используют те или иные параметрические семейства распределений вероятностей. Как уже отмечалось, наиболее популярно нормальное распределение. Используют также логарифмически нормальное распределение, экспоненциальное распределение, гамма-распределение, распределение Вейбулла-Гнеденко и т.д.

Очевидно, всегда необходимо проверять соответствие моделей реальности. Возникают два вопроса. Отличаются ли реальные распределения от используемых в модели? Насколько это отличие влияет на выводы?

Ниже на примере нормального распределения и основанных на нем методов отбраковки резко отличающихся наблюдений (выбросов) показано,

что реальные распределения практически всегда отличаются от включенных в классические параметрические семейства, а имеющиеся отклонения от заданных семейств делают неверными выводы, в рассматриваемом случае, об отбраковке, основанные на использовании этих семейств.

Есть ли основания априори предполагать нормальность результатов измерений?

Иногда утверждают, что в случае, когда погрешность измерения (или иная случайная величина) определяется в результате совокупного действия многих малых факторов, то в силу Центральной Предельной Теоремы (ЦПТ) теории вероятностей эта величина хорошо приближается (по распределению) нормальной случайной величиной. Это утверждение, вообще говоря, неверно.

Точнее, такое утверждение справедливо, если малые факторы действуют аддитивно и независимо друг от друга. Если же они действуют мультипликативно (и независимо друг от друга), то в силу той же ЦПТ аппроксимировать распределение рассматриваемой величины надо логарифмически нормальным распределением. В прикладных задачах обосновать аддитивность, а не мультипликативность действия малых факторов обычно не удается.

Если же зависимость имеет общий характер, не приводится к аддитивному или мультипликативному виду, а также нет оснований принимать модели, дающие экспоненциальное, Вейбулла-Гнеденко, гамма или иные распределения, то о распределении итоговой случайной величины практически ничего не известно, кроме внутриматематических свойств типа регулярности.

При обработке конкретных данных иногда считают, что погрешности измерений имеют нормальное распределение. На предположении нормальности построены классические модели регрессионного,

дисперсионного, факторного анализов, метрологические модели, которые еще продолжают встречаться как в отечественной нормативно-технической документации, так и в международных стандартах. На то же предположение опираются модели расчетов максимально достигаемых уровней тех или иных характеристик, применяемые при проектировании систем обеспечения безопасности функционирования экономических структур, технических устройств и объектов. Однако теоретических оснований для такого предположения нет. Необходимо экспериментально изучать распределения погрешностей.

3. Результаты экспериментов метрологов

Что же показывают результаты экспериментов? В классической монографии В.В. Налимова 1960 г. [8], посвященной применению математической статистики при анализе вещества, рассматриваемой проблеме посвящен специальный раздел "Отклонения от нормального распределения в аналитической работе" (гл.IV, параграф 4, с.122-134). Разбирается распространенное утверждение (со ссылкой на ЦПТ), что "истинное" распределение погрешностей - нормальное, а отклонения от нормальности - результат смешивания (разных генеральных совокупностей, например, серий измерений, проведенных в различных условиях). Вместе с тем приведены следующие экспериментальные данные: "В работе Клэнси [9] было изучено 250 распределений для различных аналитических методов, включающих в общей сложности 50 000 отдельных определений, и показано, что с практической точки зрения только в 10 - 15% случаев имеет место нормальное распределение" [8, с.122 - 123].

Развернутые исследования распределений погрешностей измерений проведены проф. П.В. Новицким и его научной школой. Сводка, данная в

монографии [10], позволяет утверждать, что в большинстве случаев распределение погрешностей измерений отличается от нормального. Так, в Машинно-электротехническом институте (г. Варна в Болгарии) было исследовано распределение погрешностей градуировки шкал аналоговых электроизмерительных приборов. Изучались приборы, изготовленные в Чехословакии, СССР и Болгарии. Согласно [10], закон распределения погрешностей оказался одним и тем же. Он имеет плотность

$$f(x) = 0{,}534\exp(1-|x|^7).$$

Были проанализированы данные о параметрах 219 фактических распределениях погрешностей, исследованных разными авторами, при измерении как электрических, так и не электрических величин самыми разнообразными (электрическими) приборами. В результате этого исследования оказалось, что 111 распределений, т.е. примерно 50%, принадлежат классу распределений с плотностью

$$f(x;\alpha,b.\sigma) = \frac{\alpha}{2\lambda\sigma\Gamma(1/\alpha)}\exp\left(-\left|\frac{x-b}{\lambda\sigma}\right|^{\alpha}\right),$$

где α - параметр степени; b - параметр сдвига; σ - параметр масштаба; $\Gamma(\beta)$ - гамма-функция от аргумента β ;

$$\lambda = \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$$

(см. [10, с. 56]); 63 распределения, т.е. 30%, имеют плотности с плоской вершиной и пологими длинными спадами и не могут быть описаны как нормальные или, например, экспоненциальные. Оставшиеся 45 распределений оказались двухмодальными.

В другой книге известного метролога проф. П. В. Новицкого [11] приведены результаты исследования законов распределения различного рода погрешностей измерения. Он изучил распределения погрешностей

электромеханических приборов на кернах, электронных приборов для измерения температур и усилий, цифровых приборов с ручным уравновешиванием. Объем выборок экспериментальных данных для каждого экземпляра составлял 100 - 400 отсчетов. Оказалось, что 46 из 47 распределений значимо отличались от нормального. Исследована форма распределения погрешностей у 25 экземпляров цифровых вольтметров Щ-1411 в 10 точках диапазона. Результаты аналогичны. Дальнейшие сведения содержатся в монографии [10].

В лаборатории прикладной математики Тартуского государственного университета проанализировано 2500 выборок из архива реальных статистических данных [12]. В 92% гипотезу нормальности пришлось отвергнуть.

Приведенные описания экспериментальных данных показывают, что погрешности измерений в большинстве случаев имеют распределения, отличные от нормальных. Это означает, в частности, что большинство применений критерия Стьюдента, классического регрессионного анализа и других статистических методов, основанных на нормальной теории, строго говоря, не является обоснованным, поскольку неверна лежащая в их основе аксиома нормальности распределений соответствующих случайных величин.

Очевидно, для оправдания или обоснованного изменения существующей практики анализа статистических данных требуется изучить свойства процедур анализа данных при "незаконном" применении. Изучение процедур отбраковки показало, что они крайне неустойчивы к отклонениям от нормальности, а потому применять их для обработки реальных данных нецелесообразно (см. ниже); поэтому нельзя утверждать, что произвольно взятая процедура устойчива к отклонениям от нормальности.

Иногда предлагают перед применением, например, критерия Стьюдента

однородности двух выборок проверять нормальность. Хотя для этого имеется много критериев, но проверка нормальности - более сложная и трудоемкая статистическая процедура, чем проверка однородности (как с помощью статистик типа Стьюдента, так и с помощью непараметрических критериев). Для достаточно надежного установления нормальности требуется весьма большое число наблюдений. Так, чтобы гарантировать, что функция распределения результатов наблюдений отличается от некоторой нормальной не более, чем на 0,01 (при любом значении аргумента), требуется порядка 2500 наблюдений [3]. В большинстве экономических, технических, медико-биологических и других прикладных исследований число наблюдений существенно меньше. Особенно это справедливо для данных, используемых при изучении проблем, связанных с обеспечением безопасности функционирования экономических структур и технических объектов.

4. Скорость сходимости в Центральной предельной теореме

Иногда пытаются использовать ЦПТ для приближения распределения погрешности к нормальному, включая в технологическую схему измерительного прибора специальные сумматоры. Оценим полезность этой меры.

Пусть Z_1 , Z_2 ,…, Z_k - независимые одинаково распределенные случайные величины с функцией распределения $H = H(x)$ такие, что $M(Z_1) = 0,\ D(Z_1) = 1,\ M\,|Z_1|^3 = \rho < +\infty.$ Рассмотрим

$$w = \frac{Z_1 + Z_2 + ... + Z_k}{\sqrt{k}}.$$

Показателем обеспечиваемой сумматором близости к нормальности является

$$C = \sup_H \sup_x |\,P(w < x) - \Phi(x)\,|.$$

Тогда

$$0{,}3989\frac{\rho}{\sqrt{k}} \le C \le 0{,}7975\frac{\rho}{\sqrt{k}}.$$

Правое неравенство в последнем соотношении вытекает из оценок константы в неравенстве Берри - Эссеена, полученном в книге [13, с.172], а левое - из примера в монографии [14, с.140-141]. Для нормального закона ρ =1,6, для равномерного ρ = 1,3, для двухточечного ρ =1 (это - нижняя граница для ρ). Следовательно, для обеспечения расстояния (в метрике Колмогорова) до нормального распределения не более 0,01 для "неудачных" распределений необходимо не менее k_0 слагаемых, где

$$0{,}4\sqrt{k_0} < 0{,}01, \;\; k_0 > 1600.$$

В обычно используемых сумматорах слагаемых значительно меньше. Сужая класс возможных распределений *H*, можно получить, как показано в монографии [15], более быструю сходимость, но теория здесь еще не смыкается с практикой. Кроме того, не ясно, обеспечивает ли близость распределения к нормальному (в определенной метрике) также и близость распределения статистики, построенной по случайным величинам с этим распределением, к распределению статистики, соответствующей нормальным результатам наблюдений. Видимо, для каждой конкретной статистики необходимы специальные теоретические исследования, Именно к такому выводу приходит автор монографии [15]. В задачах отбраковки выбросов ответ: "Не обеспечивает" (см. ниже).

Отметим, что результат любого реального измерения записывается с помощью конечного числа десятичных знаков, обычно небольшого (2-5), так что любые реальные данные целесообразно моделировать лишь с помощью дискретных случайных величин, принимающих конечное число значений. Нормальное распределение - лишь аппроксимация реального распределения. Так, например, данные конкретного исследования, приведенные в работе [16],

принимают значения от 1,0 до 2,2, т.е. всего 13 возможных значений. Из принципа Дирихле следует, что в какой-то точке построенная по данным работы [16] функция распределения отличается от ближайшей функции нормального распределения не менее чем на 1/26, т.е. на 0,04. Кроме того, очевидно, что для нормального распределения случайной величины вероятность попасть в дискретное множество десятичных чисел с заданным числом знаков после запятой равна 0.

Из сказанного выше следует, что результаты измерений и вообще статистические данные имеют свойства, приводящие к тому, что моделировать их следует случайными величинами с распределениями, более или менее отличными от нормальных. В большинстве случаев распределения существенно отличаются от нормальных, в других нормальные распределения могут, видимо, рассматриваться как некоторая аппроксимация, но никогда нет полного совпадения. Отсюда вытекает как необходимость изучения свойств классических статистических процедур в неклассических вероятностных моделях (подобно тому, как это сделано в [17] для критерия Стьюдента), так и необходимость разработки устойчивых (учитывающих наличие отклонений от нормальности) и непараметрических, в том числе свободных от распределения процедур, их широкого внедрения в практику статистической обработки данных.

Опущенные здесь рассмотрения для других параметрических семейств приводят к аналогичным выводам. Итог можно сформулировать так. Распределения реальных данных практически никогда не входят в какое-либо конкретное параметрическое семейство. Реальные распределения всегда отличаются от тех, что включены в параметрические семейства. Отличия могут быть большие или маленькие, но они всегда есть. Попробуем понять, насколько важны эти различия для проведения статистического анализа

данных.

5. Неустойчивость параметрических методов отбраковки резко выделяющихся результатов наблюдений

При обработке реальных статистических данных, полученных в процессе наблюдений, измерений, расчетов, иногда один или несколько результатов наблюдений резко выделяются, т.е. далеко отстоят от основной массы данных. Такие резко выделяющиеся результаты наблюдений часто считают содержащими грубые погрешности, соответственно называют промахами или выбросами. В рассматриваемых случаях возникает естественная мысль о том, что подобные наблюдения не относятся к изучаемой совокупности, поскольку содержат грубую погрешность, а получены в результате ошибки, промаха. В метрологии об этом явлении говорят так: "Грубые погрешности и промахи возникают из-за ошибок или неправильных действий оператора (его психофизиологического состояния, неверного отсчета, ошибок в записях или вычислениях, неправильного включения приборов и т.п.), а также при кратковременных резких изменений проведения измерений (вибрации, поступления холодного воздуха, толчка прибора оператором и т.п.). Если грубые погрешности и промахи обнаруживают в процессе измерений, то результаты, содержащие их, отбрасывают. Однако чаще всего их выявляют только при окончательной обработке результатов измерений с помощью специальных критериев оценки грубых погрешностей" [18, с.46-47].

Есть два подхода к обработке данных, которые могут быть искажены грубыми погрешностями и промахами:

1) отбраковка резко выделяющихся результатов наблюдений, т.е. обнаружение наблюдений, искаженных грубыми погрешностями и промахами, и исключение их из дальнейшей статистической обработки;

2) применение устойчивых (робастных) методов обработки данных, на результаты работы которых мало влияет наличие небольшого числа грубо искаженных наблюдений (см. [19 - 22] и др.).

В настоящей статье обсуждаются методы отбраковки.

Наиболее изучена ситуация, когда результаты наблюдений - числа x_1, x_2,…, x_n, резко выделяется один результат наблюдения, для определенности, максимальный $x_{\max}$.

Простейшая вероятностно-статистическая модель такова [23]. При нулевой гипотезе H_0 результаты наблюдения x_1, x_2,…, x_n рассматриваются как реализация независимых одинаково распределенных случайных величин числа X_1, X_2,…, X_n с функцией распределения $F(x)$. При альтернативной гипотезе H_1 случайные величины X_1, X_2,…, X_n также независимы, X_1, X_2,…, X_{n-1} имеют распределение $F(x)$, а X_n - распределение $G(x)$, оно "существенно сдвинуто вправо" относительно $F(x)$, например, $G(x)=F(x - A)$, где A достаточно велико. Если альтернативная гипотеза справедлива, то при $A \to \infty$ вероятность равенства

$$X_n = \max(X_1, X_2, ..., X_n)$$

стремится к 1, поэтому естественно применять решающее правило следующего вида:

если $x_{max.} > d$, то принять H_1,

если $x_{max.} \le d$, то принять H_0, (1)

где d - параметр решающего правила, значение которого следует определять из вероятностно-статистических соображений.

При справедливости нулевой гипотезы

$$P\{\max_{1\le i\le n} X_i \le d\} = \{F(d)\}^n.$$

Статистический критерий проверки гипотезы H_0, основанный на решающем правиле вида (1), имеет уровень значимости α, если

$$P\{\max_{1\le i\le n} X_i > d\} = 1-\{F(d)\}^n = \alpha,$$

т.е.

$$F(d) = \sqrt[n]{1-\alpha}. \qquad (2)$$

Из соотношения (2) определяют граничное значение $d = d(\alpha, n)$ в решающем правиле (1).

При больших n и малых α

$$F(d) = \sqrt[n]{1-\alpha} = 1-\frac{\alpha}{n}+O\left(\frac{\alpha^2}{n^2}\right), \qquad (3)$$

поэтому в качестве хорошего приближения к $d(\alpha, n)$ рассматривают $(1-\alpha/n)$ - квантиль распределения $F(x)$.

Пусть правило отбраковки задано в соответствии с выражениями (1) и (2) с некоторой функцией распределения F, однако выборка берется из функции распределения G, мало отличающейся от F в смысле расстояния Колмогорова

$$\rho(F,G) = \sup_x |F(x)-G(x)| \le \delta. \qquad (4)$$

С помощью соотношения (3) получаем, что величина $\gamma = G(d)$ для d из уравнения (2) находится между $\gamma_1 = \max(0,\ 1-\frac{\alpha}{n}-\delta)$ и $\gamma_2 = \min(1-\frac{\alpha}{n}+\delta,\ 1)$.

Уровень значимости критерия, построенного для F, при применении к наблюдениям из G есть $1-\gamma^n$ и может принимать любые значения в отрезке $[1-\gamma_2^n;\ 1-\gamma_1^n]$. В частности, при $\delta = 0{,}01$, $\alpha = 0{,}05$, $n = 5$ возможные значения уровня значимости заполняют отрезок [0; 0,1], т.е. уровень значимости может быть в 2 раза выше номинального, а если n возрастает до 30, то

максимальный уровень значимости есть 0,297, т.е. почти в 6 раз выше номинального. При дальнейшем росте n верхняя граница для уровня значимости, как нетрудно видеть, приближается к 1.

Рассмотрим и другой вопрос - насколько правило отбраковки с уровнем значимости α для G может отличаться от такового для F при справедливости неравенства (4). С использованием соотношения (3) заключаем, что из

$$G(d) = 1 - \frac{\alpha}{n} \qquad (5)$$

следует, что $\gamma_1 \le F(d) \le \gamma_2$, где γ_1 и γ_2 выписаны выше. Решение уравнения (5) может принимать любое значение в отрезке $[F^{-1}(\gamma_1);\ F^{-1}(\gamma_2)]$. В частности, при α = 0,05 и n = 5 для стандартного нормального распределения F имеем $d(\alpha, n)$ = 2,319, при δ = 0,01 решение уравнения (5) может принимать любое значение в отрезке [2,054; + ∞], при δ = 0,005 - любое значение в отрезке [2,170; 2,576].

При использовании любого другого расстояния между функциями распределения выводы о неустойчивости правил отбраковки также справедливы. Отметим, что проведенные рассмотрения выполнены в рамках "общей схемы устойчивости" (см. об устойчивости статистических процедур [19 - 22] и др.).

Рассмотренные примеры показывают, что при конкретном значении δ = 0,01 в неравенстве (4) весьма неустойчивы как уровни значимости при фиксированном правиле отбраковки, так и параметр d правила отбраковки при фиксированном уровне значимости. Обсудим, насколько реалистично определение функции распределения с точностью $\delta \le 0{,}01$.

Есть два подхода к определению функции распределения результатов наблюдений: эвристический подбор с последующей проверкой с помощью критериев согласия и вывод из некоторой вероятностной модели.

Пусть с помощью критерия согласия Колмогорова проверяется гипотеза о том, что выборка взята из распределения *F*. Пусть функции распределения *F* и *G* удовлетворяют соотношению (4). Пусть на самом деле выборка взята из распределения *G*, а не *F*. При каких δ не удастся различить *F* и *G*? Для определенности, при каких δ гипотеза согласия с *F* будет приниматься не менее чем в 50% случаев?

Критерий согласия Колмогорова основан на статистике

$$\lambda_n = \sqrt{n}\rho(F_n, H), \quad (6)$$

где расстояние ρ между функциями распределения определено выше в формуле (4); *H* - та функция распределения, согласие с которой проверяется, а F_n - эмпирическая функция распределения (т.е. $F_n(x)$ равно доле наблюдений, меньших *x*, в выборке объема *n*). Как показал А.Н. Колмогоров в 1933 г., функция распределения случайной величины λ_n при росте объема выборки *n* сходится к некоторой функции распределения $K(x)$, которую ныне называют функцией Колмогорова [3, 23]. При этом $K(1{,}36) = 0{,}95$ и $K(0{,}83) = 0{,}50$.

Поскольку выборка взята из распределения *G*, то с вероятностью 0,50

$$\rho(F_n, G) < 0{,}83/\sqrt{n} \quad (7)$$

(при больших *n*). Тогда для рассматриваемой выборки с учетом неравенства (4) и неравенства треугольника для расстояния Колмогорова и симметричности этого расстояния имеем

$$\rho(F_n, F) \le \rho(F_n, G) + \rho(G\ , F) = \rho(F_n, G) + \rho(F\ , G) < 0{,}83/\sqrt{n} + \delta.$$

Если

$$0{,}83/\sqrt{n} + \delta \le 1{,}36/\sqrt{n},$$

т.е.

$$\delta\sqrt{n} \le 0{,}53, \quad (8)$$

то, согласно формуле (6), гипотеза согласия принимается по крайней мере с той же вероятностью, с которой выполнено неравенств (7), т.е. с вероятностью не менее 0,50. Для δ = 0,01 это условие выполняется при $n \leq$ 2809. Таким образом, для определения функции распределения с точностью $\delta \leq 0{,}01$ с помощью критерия согласия Колмогорова необходимо несколько тысяч наблюдений, что для большинства прикладных задач нереально.

При втором из названных выше подходов к определению функции распределения ее конкретный вид выводится из некоторой системы аксиом, в частности, из некоторой модели порождения соответствующей случайной величины. Например, из модели суммирования вытекает нормальное распределение, а из мультипликативной модели перемножения - логарифмически нормальное распределение. Как правило, при выводе используется предельный переход. Так, из Центральной Предельной Теоремы теории вероятностей вытекает, что сумма независимых случайных величин может быть приближена нормальным распределением. Однако более детальный анализ, в частности, с помощью неравенства Берри - Эссеена (см. выше) показывает, что для гарантированного достижения точности $\delta \leq 0{,}01$ необходимо более полутора тысяч слагаемых. Такого количества слагаемых реально, конечно, указать почти никогда нельзя. Это означает, что при решении практических задач теория дает возможность лишь сформулировать гипотезу о виде функции распределения, а проверять ее надо с помощью анализа реальной выборки объема, как показано выше, не менее нескольких тысяч. Таким образом, в большинстве реальных ситуаций определить функцию распределения с точностью $\delta \leq 0{,}01$ невозможно.

Итак, показано, что правила отбраковки, основанные на использовании конкретной функции распределения, являются крайне неустойчивыми к отклонениям от нее распределения элементов выборки, а гарантировать

отсутствие подобных отклонений невозможно. Поэтому *отбраковка по классическим правилам математической статистики не является научно обоснованной*, особенно при больших объемах выборок. Указанные правила целесообразно применять лишь для выявления "подозрительных" наблюдений, вопрос об отбраковке которых должен решаться из соображений соответствующей предметной области, а не из формально-математических соображений [24].

Выше для простоты изложения рассмотрен лишь случай полностью известного распределения *F*, для которого изучено правило отбраковки, заданное формулами (1) и (2). Аналогичные выводы о крайней неустойчивости правил отбраковки справедливы, если "истинное распределение" принадлежит какому-либо параметрическому семейству, например, нормальному, Вейбулла - Гнеденко, гамма.

Параметрическим методам отбраковки, основанным на моделях тех или иных параметрических семейств распределений, посвящены тысячи книг и статей. Приходится признать, что они имеют в основном внутриматематический интерес. При обработке реальных данных следует применять устойчивые методы (см. [19 - 22] и др.), в частности, непараметрические [25, 26].

Литература

1. Орлов А.И. Основные этапы становления статистических методов // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 97. С. 73–85.
2. Орлов А.И. Современная прикладная статистика // Заводская лаборатория. Диагностика материалов. 1998. Т.64. №3. С. 52-60.
3. Орлов А.И. Прикладная статистика. — М.: Экзамен, 2006. — 671 с.
4. Орлов А.И. Состояние и перспективы развития прикладной и теоретической статистики // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2016. № 115. С. 202–226.

5. Орлов А.И. Эконометрика. Изд. 4-е, доп. и перераб. Учебник для вузов. – Ростов-на-Дону: Феникс, 2009. - 572 с.

6. Орлов А.И. Современное состояние непараметрической статистики // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2015. № 106. С. 239 – 269.

7. Орлов А.И. Структура непараметрической статистики (обобщающая статья) // Заводская лаборатория. Диагностика материалов. 2015. Т.81. №7. С. 62-72.

8. Налимов В.В. Применение математической статистики при анализе вещества. - М.: ГИФМЛ, 1960. - 430 с.

9. Clancey V.J. Statistical methods in chemical analyses // Nature. 1947. V.159. № 4036. P.339-340.

10. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. - Л.: Энергоатомиздат, 1985. - 248 с.

11. Новицкий П.В. Основы информационной теории измерительных устройств. - Л.: энергия, 1968. - 248 с.

12. Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. Диагностика материалов. 1991. Т.57. №7. С.64-66.

13. Боровков А.А. Теория вероятностей. - М.: Наука, 1976. - 352 с.

14. Петров В.В. Суммы независимых случайных величин. - М.: Наука, 1972. - 416 с.

15. Золотарев В.М. Современная теория суммирования независимых случайных величин. - М.: Наука, 1986. - 416 с.

16. Егорова Л.А., Харитонов Ю.С., Соколовская Л.В. О применении непараметрического Х-критерия Ван-дер-Вардена при статистической обработке результатов наблюдений // Заводская лаборатория. Диагностика материалов. 1976. Т.42. №10. С. 1237-1239.

17. Орлов А.И. Проверка статистической гипотезы однородности математических ожиданий двух независимых выборок: критерий Крамера-Уэлча вместо критерия Стьюдента // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2015. № 110. С. 197–218.

18. Артемьев Б.Г., Голубов С.М. Справочное пособие для работников метрологических служб.- М.: Изд-во стандартов, 1982. - 280 с.

19. Орлов А.И. Устойчивость в социально-экономических моделях. — М.: Наука, 1979. — 296 с.

20. Орлов А.И. Устойчивые математические методы и модели // Заводская лаборатория. Диагностика материалов. 2010. Т.76. №3. С.59-67.

21. Орлов А.И. Устойчивые экономико-математические методы и модели. Saarbrücken (Germany), Lambert Academic Publishing, 2011. 436 с.

22. Орлов А.И. Новый подход к изучению устойчивости выводов в математических моделях // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 100. С. 146-176.

23. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. - 416 с.

24. Орлов А.И. Неустойчивость параметрических методов отбраковки резко выделяющихся наблюдений. // Заводская лаборатория. Диагностика материалов. 1992. Т.58. №7. С.40-42.

25. Орлов А.И. Структура непараметрической статистики (обобщающая статья) // Заводская лаборатория. Диагностика материалов. 2015. Т.81. №7. С. 62-72.

26. Орлов А.И. Современное состояние непараметрической статистики // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2015. № 106. С. 239 – 269.

References

1. Orlov A.I. Osnovnye jetapy stanovlenija statisticheskih metodov // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2014. № 97. S. 73–85.

2. Orlov A.I. Sovremennaja prikladnaja statistika // Zavodskaja laboratorija. Diagnostika materialov. 1998. T.64. №3. S. 52-60.

3. Orlov A.I. Prikladnaja statistika. — M.: Jekzamen, 2006. — 671 s.

4. Orlov A.I. Sostojanie i perspektivy razvitija prikladnoj i teoreticheskoj statistiki // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2016. № 115. S. 202–226.

5. Orlov A.I. Jekonometrika. Izd. 4-e, dop. i pererab. Uchebnik dlja vuzov. – Rostov-na-Donu: Feniks, 2009. - 572 s.

6. Orlov A.I. Sovremennoe sostojanie neparametricheskoj statistiki // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2015. № 106. S. 239 – 269.

7. Orlov A.I. Struktura neparametricheskoj statistiki (obobshhajushhaja stat'ja) // Zavodskaja laboratorija. Diagnostika materialov. 2015. T.81. №7. S. 62-72.

8. Nalimov V.V. Primenenie matematicheskoj statistiki pri analize veshhestva. - M.: GIFML, 1960. - 430 s.

9. Clancey V.J. Statistical methods in chemical analyses // Nature. 1947. V.159. № 4036. P.339-340.

10. Novickij P.V., Zograf I.A. Ocenka pogreshnostej rezul'tatov izmerenij. - L.: Jenergoatomizdat, 1985. - 248 s.

11. Novickij P.V. Osnovy informacionnoj teorii izmeritel'nyh ustrojstv. -L.: jenergija, 1968. - 248 s.

12. Orlov A.I. Chasto li raspredelenie rezul'tatov nabljudenij javljaetsja normal'nym? // Zavodskaja laboratorija. Diagnostika materialov. 1991. T.57. №7. S.64-66.

13. Borovkov A.A. Teorija verojatnostej. - M.: Nauka, 1976. - 352 s.

14. Petrov V.V. Summy nezavisimyh sluchajnyh velichin. - M.: Nauka, 1972. - 416 s.

15. Zolotarev V.M. Sovremennaja teorija summirovanija nezavisimyh sluchajnyh velichin. - M.: Nauka, 1986. - 416 s.

16. Egorova L.A., Haritonov Ju.S., Sokolovskaja L.V. O primenenii neparametricheskogo H-kriterija Van-der-Vardena pri statisticheskoj obrabotke rezul'tatov nabljudenij // Zavodskaja laboratorija. Diagnostika materialov. 1976. T.42. №10. S. 1237-1239.

17. Orlov A.I. Proverka statisticheskoj gipotezy odnorodnosti matematicheskih ozhidanij dvuh nezavisimyh vyborok: kriterij Kramera-Ujelcha vmesto kriterija St'judenta // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2015. № 110. S. 197–218.

18. Artem'ev B.G., Golubov S.M. Spravochnoe posobie dlja rabotnikov metrologicheskih sluzhb.- M.: Izd-vo standartov, 1982. - 280 s.

19. Orlov A.I. Ustojchivost' v social'no-jekonomicheskih modeljah. — M.: Nauka, 1979. — 296 s.

20. Orlov A.I. Ustojchivye matematicheskie metody i modeli // Zavodskaja laboratorija. Diagnostika materialov. 2010. T.76. №3. S.59-67.

21. Orlov A.I. Ustojchivye jekonomiko-matematicheskie metody i modeli. Saarbrücken (Germany), Lambert Academic Publishing, 2011. 436 s.

22. Orlov A.I. Novyj podhod k izucheniju ustojchivosti vyvodov v matematicheskih modeljah // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2014. № 100. S. 146-176.

23. Bol'shev L.N., Smirnov N.V. Tablicy matematicheskoj statistiki. – M.: Nauka, 1983. - 416 s.

24. Orlov A.I. Neustojchivost' parametricheskih metodov otbrakovki rezko vydeljajushhihsja nabljudenij. // Zavodskaja laboratorija. Diagnostika materialov. 1992. T.58. №7. S.40-42.

25. Orlov A.I. Struktura neparametricheskoj statistiki (obobshhajushhaja stat'ja) // Zavodskaja laboratorija. Diagnostika materialov. 2015. T.81. №7. S. 62-72.

26. Orlov A.I. Sovremennoe sostojanie neparametricheskoj statistiki // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2015. № 106. S. 239 – 269.