ИЗВЕСТИЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ

№ 9 ФИЗИКА 1995

Б. Ю. ЛЕМЕШКО, С. Н. ПОСТОВАЛОВ

СТАТИСТИЧЕСКИЙ АНАЛИЗ ОДНОМЕРНЫХ НАБЛЮДЕНИЙ ПО ЧАСТИЧНО ГРУППИРОВАННЫМ ДАННЫМ

Программное обеспечение, реализующее решение задачи выбора закона распределения, наиболее хорошо описывающего выборочные данные, является дальнейшим развитием программной системы «Статистический анализ одномерных наблюдений случайных величин» [1]. Реализованные возможности позволяют оценивать параметры и проверять согласие по 26 наиболее часто используемым на практике распределениям и их *смесям*.

Характер выборочных данных, по которым осуществляется анализ распределений, *может быть различным.* Наиболее общим случаем является частично группированная выборка [2]. Выборка является *негруппированной,* если выборочные значения представляют собой индивидуальные значения наблюдений из области определения случайной величины. Выборка является *группированной,* если область определения случайной величины разбита на κ непересекающихся интервалов граничными точками

$$x_0 < x_1 < \ldots < x_{\kappa-1} < x_\kappa,$$

где x_0 — нижняя грань области определения случайной величины ξ, x_κ — верхняя грань области определения случайной величины ξ, и зафиксированы количества наблюдений n_i, попавших в i-й интервал значений. Выборка является *частично группированной,* если имеющаяся в нашем распоряжении информация связана с множеством непересекающихся интервалов, которые делят область определения случайной величины так, что каждый интервал принадлежит к одному из двух типов:

а) i-й интервал принадлежит к первому типу, если число n_i известно, но индивидуальные значения x_{ij}, $j=\overline{1, n_i}$, неизвестны;

б) i-й интервал принадлежит ко второму типу, если известно не только число n_i, но и все индивидуальные значения x_{ij}, $j=\overline{1, n_i}$.

Область определения случайной величины в этом случае можно представить в виде $X=X_{(1)} \cup X_{(2)}$, где $X_{(1)}$ — множество интервалов первого типа, а $X_{(2)}$ — множество интервалов второго типа.

Когда перед нами возникает задача выбора распределения, с которым наиболее хорошо согласуются данные экспериментов, то последовательность наших действий описывается следующим алгоритмом.

Ограничиваем класс распределений, из которого мы будем выбирать подходящий закон распределения вероятностей.

Далее для выбранных распределений оцениваем параметры и проверяем гипотезы о согласии.

Выбираем то распределение, которое наиболее хорошо согласуется с выборкой.

При использовании общепринятой методики проверки гипотез по критериям согласия, когда гипотеза о согласии с данным распределением не отвергается, если вычисленное значение статистики не превышает критического, соответствующего заданному уровню значимости α, обычно оказывается, что нет причин отказаться от целого ряда распределений. В этом случае сохраняется несколько возможных альтернатив. Мы же должны остановиться на том распределении, согласие с которым наиболее хорошее.

В описываемом программном обеспечении при проверке гипотез о согласии для каждой используемой статистики S_i, $i=\overline{1, m}$, вычисляются вероятности вида $P\{S_i > S_i^*\} = \int\limits_{S_i^*}^{\infty} g_i(s)\,ds$, где S_i^* — найденное по выборке значение соответствующей статистики, $g_i(s)$ — функция плотности распределения статистики S_i при условии, что гипотеза H_0 является истинной. S_i^* является функционалом, зависящим от конкретных выборки и закона распределения, т. е. $S_i^* = S_i^*(\bar{X}, f(x, \hat{\theta}))$, где через $\bar{X}$ обозначена выборка случайной величины. Допустим, что на основании первичных предположений мы выделили множество законов распределений, к которым может принадлежать рассматриваемая выборка, пронумеровали эти законы, обозначив через R множество индексов функций плотности $f_j(x, \hat{\theta})$, $j \in R$, оценили по данной выборке параметры законов распределений, вычислили значения статистик $S_{ij}^* = S_i^*(\bar{X}, f_j(x, \hat{\theta}))$ и вероятности $P\{S_i > S_{ij}^*\} = \alpha_{ij}$. Тогда при проверке гипотезы о согласии с j-м распределением по i-у критерию, если $\alpha_{ij} > \alpha$, где α — задаваемый исследователем уровень значимости, нет повода отвергать гипотезу о согласии с j-м распределением в соответствии с i-м критерием. Пусть в соответствии с используемыми критериями нет оснований отвергать гипотезу о согласии с множеством законов, помеченных индексами из $R_1 \subset R$. Тогда мы должны выбрать тот закон распределения случайной величины $f_l(x, \hat{\theta})$, для которого $\forall i \;\; \alpha_{il} = \max\limits_{j \in R_1} \alpha_{ij}$.

Обычно такой вывод можно сделать однозначно. Однако вполне возможно (и это бывает довольно часто для различных, но близких законов распределения), что выводы по разным критериям указывают на предпочтительность того или иного закона. Это означает, что решения задачи выбора распределения по разным критериям не совпадают. Такая «несогласованность» объясняется различием мер, используемых в критериях. Следовательно, мы имеем естественную многокритериальную задачу принятия решения. Так как все критерии измеряются в единой шкале, то решить её можно, сформировав простой компромиссный критерий вида $\max\limits_{j \in R_1} \sum\limits_{i=1}^{m} \omega_i \alpha_{ij}$, где ω_i — весовой коэффициент i-го критерия, $\sum\limits_{i=1}^{m} \omega_i = 1$.

При решении задач статистического анализа и, в частности, вычислении оценок параметров распределений чрезвычайно важное значение приобретает проблема наличия в выборке *аномальных измерений*. В практике решения таких задач широко известно, что наличие даже одного аномального наблюдения приводит к оценкам, которые совершенно не вяжутся с выборочными данными. Вообще говоря, наличие выбросов отражается на качестве всех выводов.

Понятно желание каждого исследователя, чтобы найденные оценки были как можно менее чувствительны к аномальным наблюдениям. Так как в противном случае прежде чем переходить к оцениванию, приходится использовать процедуры исключения грубых ошибок измерений, что выливается в не совсем простую задачу. В данном случае следует подчеркнуть достоинство оценок, использующих *группирование* исходных выборочных данных, так как очевидно, что они менее чувствительны к случайным выбросам. Группирование выборки позволяет резко снизить влияние аномальных наблюдений, а иногда и совсем исключить влияние случайных выбросов.

Продемонстрируем сказанное на следующем примере. Была смоделирована выборка по нормальному закону с нулевым математическим ожиданием и единичной дисперсией, состоящая из 500 наблюдений. По ней были найдены оценки максимального правдоподобия: $\mu = 1{,}027$ и $\sigma = 1{,}017$. Затем в данной выборке увеличили первое наблюдение на 20 и снова провели соответствующий анализ. Его результаты приведены на рис. 1.

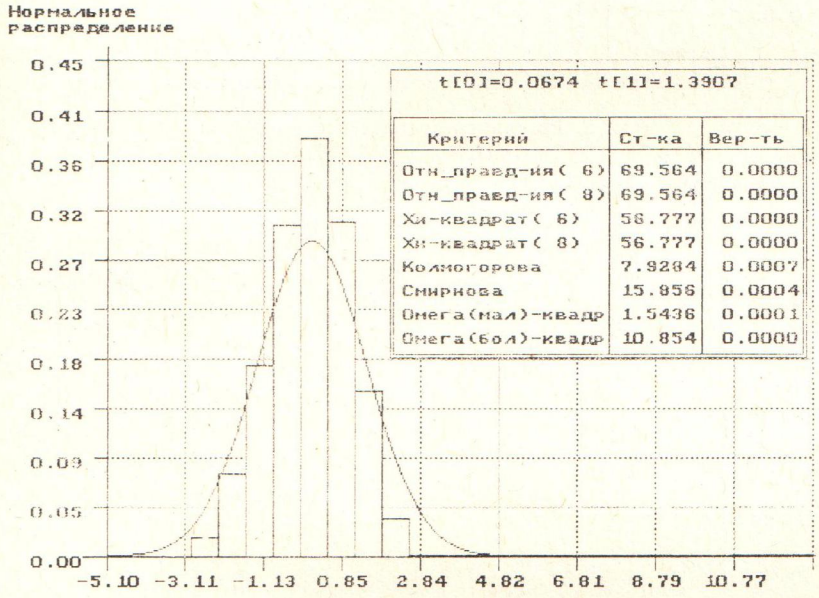

Рис. 1. Результаты анализа при наличии «аномального» наблюдения

Как и следовало ожидать, наиболее существенно изменилась оценка среднеквадратичного отклонения. Согласие по всем критериям отвергается.

А далее осуществили группирование выборки с «аномальным» наблюдением, провели оценивание по группированной выборке и проверили согласие. Результаты представлены на рис. 2. Как видим, «случайная» ошибка в данных практически не повлияла на оценки параметров.

Существенно увеличить количество моделей, используемых для описания реальных случайных величин, можно за счет применения смесей распределений. Функция распределения смеси из s распределений имеет вид

$$F(x) = \sum_{i=1}^{s} w_i F_i(x, \theta_i), \quad \sum_{i=1}^{s} w_i = 1,$$

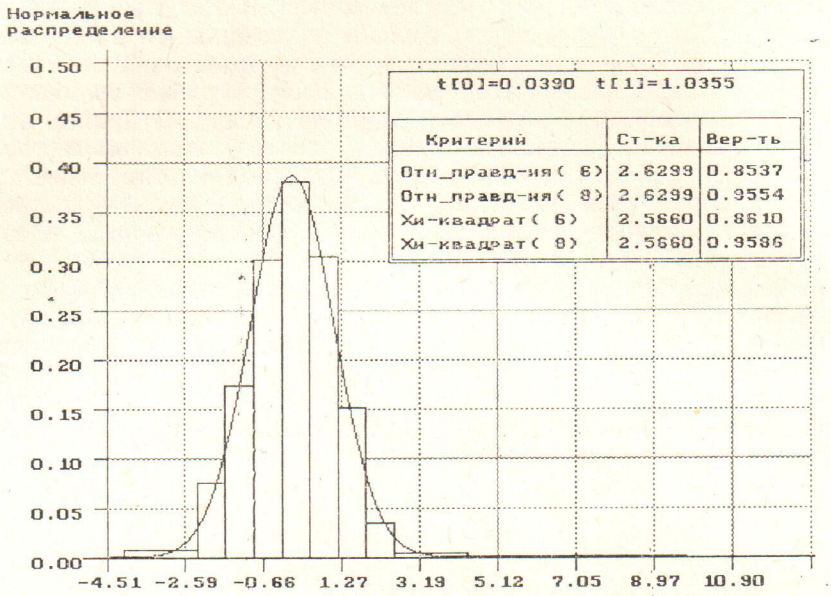

Рис. 2. Результаты робастного оценивания

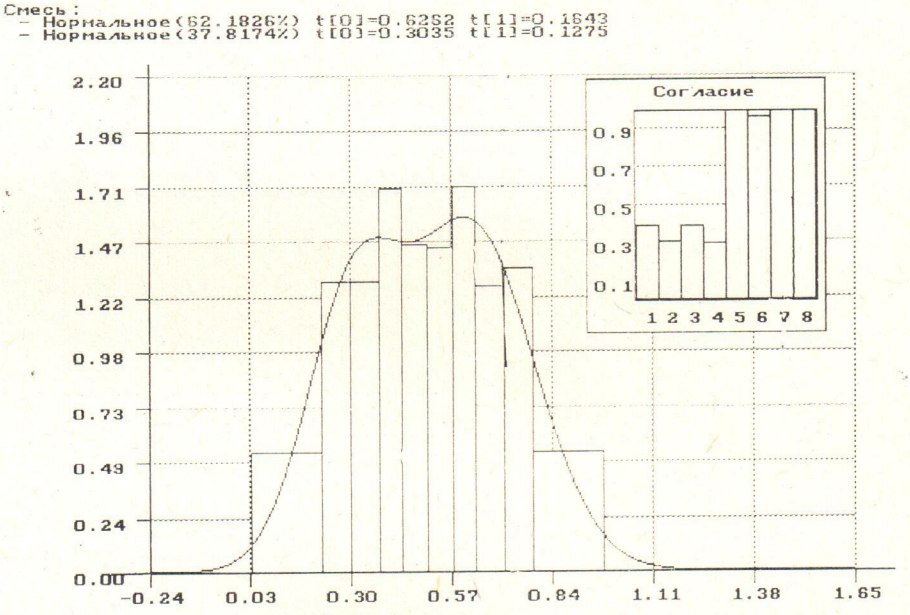

Рис. 3. Оценивание параметров и проверка гипотез согласия смеси двух нормальных распределений по группированной выборке

где s — число распределений в смеси, w_i — параметры смеси, F_i—i-я функция распределения, θ_i — вектор её параметров. Когда параметры смеси принадлежат интервалу $[0, 1]$, мы имеем классическую смесь (рис. 3), получаемую, например, объединением выборок. Если же какой-то параметр $w_i \notin [0, 1]$, то одно из распределений входит в смесь со знаком минус и, таким образом, вычитается из других распределений.

Когда реальная выборка действительно является смесью наблюдений, применение в качестве искомого закона смеси распределений даёт хорошие результаты.

Как уже говорилось, при анализе осуществляется проверка гипотез о согласии по критериям: χ^2 Пирсона, отношения правдоподобия, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса.

Параметрические критерии χ^2 Пирсона и отношения правдоподобия предусматривают группирование данных. Использование в разработанном программном обеспечении наряду с равномерным и равновероятным группированием *асимптотически оптимального* группирования обеспечивает *максимальную мощность критериев* согласия отношения правдоподобия и χ^2 Пирсона при близких конкурирующих гипотезах [1] и *минимум асимптотической дисперсии* при оценивании параметров по группированным данным.

При построении асимптотически оптимальных граничных точек интервалов решается задача

$$\max_{x<x_1<\ldots x_{\kappa-1}<x_\kappa} \det M_\Gamma(\theta),$$

где $M_\Gamma(\hat{\theta}) = \sum_{i=1}^{\kappa} \frac{\nabla P_i(\hat{\theta})\,\nabla^T P_i(\hat{\theta})}{P_i(\hat{\theta})}$ — информационная матрица Фишера, $P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x,\theta)\,dx$ — вероятность попадания в соответствующий интервал.

В тех случаях, когда исходные данные представляют собой группированную или частично группированную выборку, применение непараметрических критериев Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса [3] затруднительно, так как оказываются неизвестными значения соответствующих статистик. В этом случае предложен следующий подход. Для каждой статистики находятся оценки сверху и снизу, и на основании верхней и нижней границы вероятности согласия делаются статистические выводы.

Соответствующая критерию Колмогорова статистика имеет вид

$$D_n = \sup_x |F_n(x) - F(x)|,$$

где $F_n(x)$ — эмпирическая функция распределения, $F(x)$ — теоретическая, согласие с которой проверяется, n — объем выборки.

Пусть задана частично группированная выборка. Введем обозначения:

$$N_i = \sum_{j=1}^{i} n_j,\ N_{-1}=0,\ N_0=n_0,\ N_{\kappa-1}=n,\ N_{ij}=N_i+j.$$

Эмпирическая функция распредления $F_n(x)$ полностью определена для интервалов второго типа:

$$F_n(x) = N_{i-1,j}/n,\ \forall x \in [x_{ij},\ x_{i,j+1}) \subseteq [x_i,\ x_{i+1}) \subseteq X_{(2)},\ j=1,\ \ldots,\ n_i,$$
$$(x_{i,n_i+1} \equiv x_{i+1}),$$

а также во всех граничных точках x_i, $i=0,\ \ldots,\ \kappa$:

$$F_n(x_i) = N_{i-1}/n.$$

На интервалах первого типа нам известно только, что

$$G_\kappa^-(x) = N_{i-1}/n \le F_n(x) \le N_i/n = G_\kappa^+(x),\ x \in [x_i,\ x_{i+1}) \subseteq X_{(1)}.$$

Мы можем ограничить D_n снизу следующим образом:

$$D_n=\sup_x|F_n(x)-F(x)|=\max\{\sup_{X_{(1)}}|F_n(x)-F(x)|,\ \sup_{X_{(2)}}|F_n(x)-F(x)|\},$$

$$D_n\geq\max\{\max_{i=0,\ldots,\kappa}|N_{i-1}/n-F(x_i)|,\ \sup_{X_{(2)}}|F_n(x)-F(x)|\}=\underline{D_{n\kappa}}.$$

Найдем теперь оценку сверху. Функции $G_\kappa^+(x)$ и $G_\kappa^-(x)$ построены так, что $\forall x\in X_{(1)}\ \ G_\kappa^-(x)\leq F_n(x)\leq G_\kappa^+(x)$. Тогда

$$G_\kappa^-(x)-F(x)\leq F_n(x)-F(x)\leq G_\kappa^+(x)-F(x),$$

$$F(x)-G_\kappa^+(x)\leq F(x)-F_n(x)\leq F(x)-G_\kappa^-(x).$$

Обозначим через $A=\{x\in X_{(1)}: F_n(x)\geq F(x)\}$ и $B=\{x\in X_{(1)}: F_n(x)<F(x)\}$, найдем, что

$$D_n=\max\{\sup_{A\subseteq X_{(1)}}(F_n(x)-F(x)),\ \sup_{B\subseteq X_{(1)}}(F(x)-F_n(x)),\ \sup_{X_{(2)}}|F_n(x)-F(x)|\},$$

$$D_n\leq\max\{\sup_{A\subseteq X_{(1)}}(G_\kappa^+(x)-F(x)),\ \sup_{B\subseteq X_{(1)}}(F(x)-G_\kappa^-(x)),\ \sup_{X_{(2)}}|F_n(x)-F(x)|\},$$

$$D_n\leq\max\{\sup_{X_{(1)}}|G_\kappa^+(x)-F(x)|,\ \sup_{X_{(1)}}|F(x)-G_\kappa^-(x)|,$$

$$\sup_{X_{(2)}}|F_n(x)-F(x)|\}=\overline{D_{n\kappa}}.$$

Таким образом, $\underline{D_{n\kappa}}\leq D_n\leq\overline{D_{n\kappa}}$ и, так как функция распределения является монотонно возрастающей, то $p_{\min}\leq p\leq p_{\max}$, где $p=1-K(g(D_n))$, $p_{\min}=1-K(g(\overline{D_{n\kappa}}))$, $p_{\max}=1-K(g(\underline{D_{n\kappa}}))$, $K(\lambda)$ — функция распределения Колмогорова, а $g(y)=\sqrt{(6ny+1)^2/36n}$ [3].

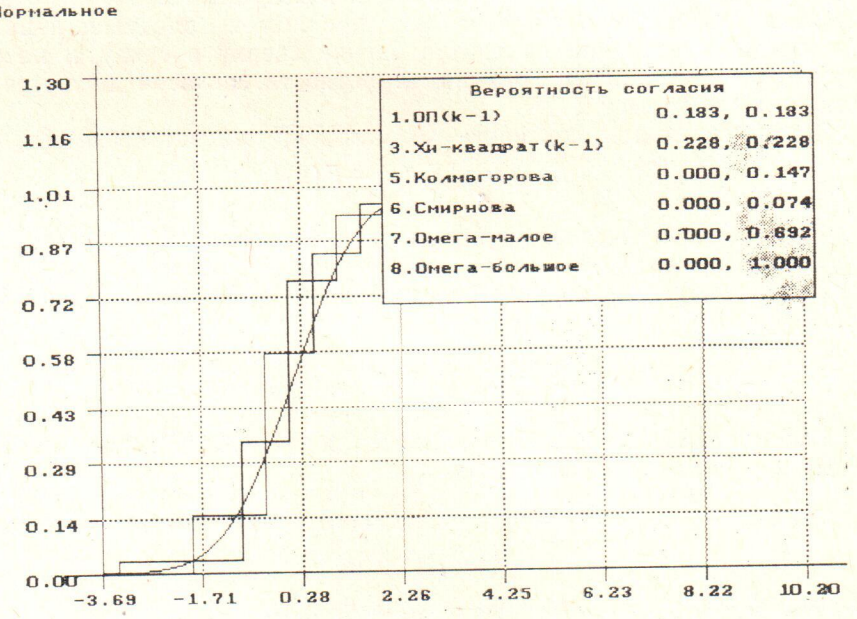

Рис. 4. Проверка согласия нормального распределения с параметрами $\mu=0{,}10$ и $\sigma=1{,}10$ по группированным данным

Аналогичные оценки получены для остальных статистик.

Следовательно, для рассмотренных критериев, при заданном уровне значимости α, возможны следующие выводы: гипотезу о согласии

следует отклонить, если $p_{max} \leq \alpha$; гипотезу о согласии не следует отвергать, если $p_{min} > \alpha$.

На рис. 4 приведен пример с нормальным распределением. Ступенчатые функции обозначают верхнюю и нижнюю предельную границу для неизвестной эмпирической функции распределения. При заданном уровне значимости $\alpha = 0{,}15$ гипотеза о согласии с нормальным распределением с параметрами $\mu = 0{,}1$ и $\sigma = 1{,}1$ проходит по критериям χ^2 Пирсона, отношения правдоподобия, ω^2 и Ω^2 Мизеса и отвергается по критериям Колмогорова и Смирнова.

СПИСОК ЛИТЕРАТУРЫ

1. Денисов В. И., Лемешко Б. Ю., Цой Е. Б. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов. — Новосибирск, 1993. — 347 с.
2. Куллдорф Г. Введение в теорию оценивания по группированным и частично группированным выборкам. — М.: Наука, 1966. — 176 с.
3. Большев Н. Л., Смирнов Н. В. Таблицы математической статистики. — М.: Наука, 1983. — 416 с.

Новосибирский государственный
технический университет