АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я

№ 6 1984

КРАТКИЕ СООБЩЕНИЯ

УДК 519.283

С. Е. ЖАРИНОВ
(Петропавловск-Камчатский)

ГРАФИЧЕСКИЙ МЕТОД ПРОВЕРКИ УНИМОДАЛЬНОСТИ

Введение. При обработке экспериментальных данных некоторые параметры считаются случайными величинами, имеющими непрерывные функции распределения (ФР) вероятностей. Плотности соответствующих распределений в таком случае оценивают по выборке (например, с помощью гистограмм), а по форме и особенностям полученных оценок делают выводы о природе изучаемого явления или процесса.

Важная характеристика интерпретируемого распределения — мультимодальность. На рис. 1 показана типичная гистограмма, анализ которой позволяет предположить наличие по крайней мере двух «горбов» в исходном распределении, что связано с «провалом» в средней части. Однако последний, вообще говоря, может быть обусловлен и случайностью выборки. Таким образом, возникает вопрос о статистической значимости выдвинутого предположения. Предлагаемый графический метод позволяет проверять гипотезу унимодальности.

Описание метода. Идею поясним с помощью графиков, приведенных на рис. 2, которые иллюстрируют качественное отличие ФР в унимодальном (*а*) и бимодальном (*б*) случаях. Очевидно, что максимумы и минимумы плотностей соответствуют точкам перегиба ФР, и в унимодальном случае такая точка лишь одна. Предположим теперь, что $F_2(x)$ — оценка истинной функции распределения $F(x)$, полученная по выборке. Построив достаточно узкую полосу

$$W_\varepsilon = \{(x, y) \mid \max [F_2(x) - \varepsilon, 0] \leqslant y \leqslant \min [F_2(x) + \varepsilon, 1];\ x \in (-\infty, \infty)\},$$

можно добиться того, что в нее невозможно будет «вписать» никакую непрерывную неубывающую функцию с одной точкой перегиба (типа $F_1(x)$), а определив вероятность события $F(x) \subset W_\varepsilon$, можно судить и об уровне статистической значимости данного утверждения. В работе для этой цели используется критерий Колмогорова — Смирнова.

Пусть $X = (x_1, \ldots, x_N)$ — выборка из распределения с непрерывной функцией распределения $F(x)$. Рассмотрим эмпирическую ФР:

$$S_N(x) = \begin{cases} 0, & x < x_{(1)}; \\ i/N, & x_{(i)} \leqslant x < x_{(i+1)}; \\ 1, & x_{(N)} \leqslant x, \end{cases} \tag{1}$$

где $x_{(i)}$ — i-я порядковая статистика X. Используя статистику Колмогорова — Смирнова

$$D_N = \sup_x |S_N(x) - F(x)|$$

и задав доверительную вероятность $P_д$, можно найти доверительный интервал для ФР $F(x)$ на уровне значимости $\alpha = 1 - P_д$:

$$\max [S_N(x) - \varepsilon_\alpha, 0] \leqslant F(x) \leqslant \min [S_N(x) + \varepsilon_\alpha, 1],\quad x \in (-\infty, \infty), \tag{2}$$

где ε_α определяется из соотношения $P\{D_N > \varepsilon_\alpha\} = \alpha$. Соответствующие критические значения ε_α табулированы (см., например, [1]), но при больших N ($N \geqslant 80$) в принципе достаточно приближения [2]

$$\varepsilon_\alpha(N) \approx K_\alpha/\sqrt{N}. \tag{3}$$

Здесь K_α — постоянная, зависящая лишь от уровня значимости:

$$K_{0,10} = 1{,}23;\quad K_{0,05} = 1{,}36;\quad K_{0,01} = 1{,}63.$$

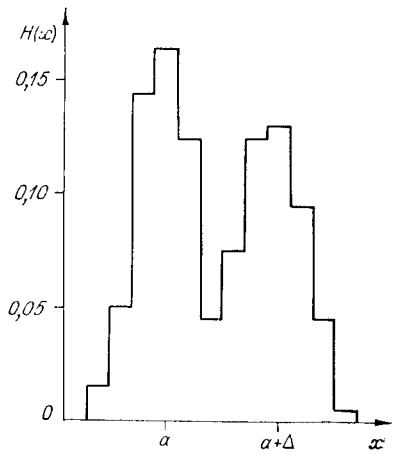

Рис. 1. Гистограмма, построенная по выборке объемом $N_1 = 200$ из распределения (4) ($\Delta/\sigma=4$).

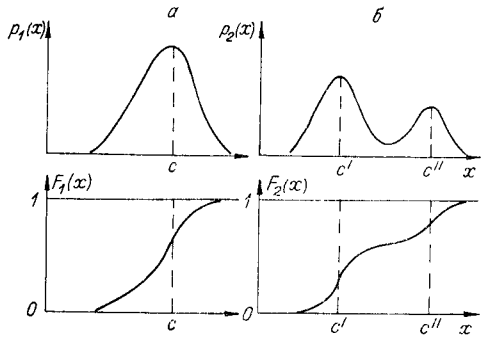

Рис. 2. Плотности и функции распределения.

Таким образом, для решения поставленной задачи требуется по выборке построить эмпирическую ФР (1), вычислить критическое значение ε_α по формуле (3) и проанализировать доверительную полосу (2). Если в нее нельзя вписать функцию с одним перегибом, то гипотезу об унимодальности следует отвергнуть на уровне значимости α.

Заметим, что последнее условие не поддается формализации, однако вопрос фактически сводится к тому, можно ли вписать в среднюю часть доверительной полосы отрезок прямой линии (т. е. критическое распределение, равномерное в области «провала»), что в каждом конкретном случае определяется визуально. Проиллюстрируем это на примере.

Пример. Пусть случайная величина имеет плотность распределения

$$p(x) = (1/2\sqrt{2\pi}\,\sigma)\exp\{-(x-a)^2/2\sigma^2\} + (1/2\sqrt{2\pi}\,\sigma)\exp\{-(x-a-\Delta)^2/2\sigma^2\}, \quad (4)$$

представляющую собой смесь двух нормальных распределений с одинаковыми дисперсиями и весами, но разными математическими ожиданиями, отличающимися на Δ.

При $\Delta = 4\sigma$ было получено две выборки объемом $N_1 = 200$ и $N_2 = 800$. Построенные доверительные полосы показаны на рис. 3. Видно, что в первом случае на уровне значимости $\alpha=0{,}10$ ($\varepsilon_{0,10}(200) = 0{,}087$) гипотезу об унимодальности отклонить не удается (допустимая функция с одним перегибом показана на рис. 3, *а* тонкой линией), тогда как во втором случае нулевая гипотеза, без сомнения, отклоняется даже на уровне значимости $\alpha = 0{,}05$ ($\varepsilon_{0,05}(800) = 0{,}048$).

Обсуждение. В рассмотренном примере потребовалось большое количество измерений для принятия правильного решения, хотя «горбы» на гистограмме достаточно ярко выражены и при меньшем объеме выборки (см. рис. 1). Это свидетельствует о невысокой мощности критерия для гипотез рассматриваемого типа.

Для оценки чувствительности предложенного метода к величине «провала» были проведены модельные эксперименты с семейством распределений (4) при различных значениях параметра Δ. Известно, что бимодальность в этом случае имеет место при $\Delta > 2\sigma$, а при $\Delta > 6\sigma$ минимум становится практически нулевым. Усредненная зависимость Δ/σ от критического объема выборки, необходимого для отклонения нулевой гипотезы на разных уровнях значимости, показана на рис. 4. Полученная зависимость позволяет (хотя бы качественно) оценить, достаточно ли име-

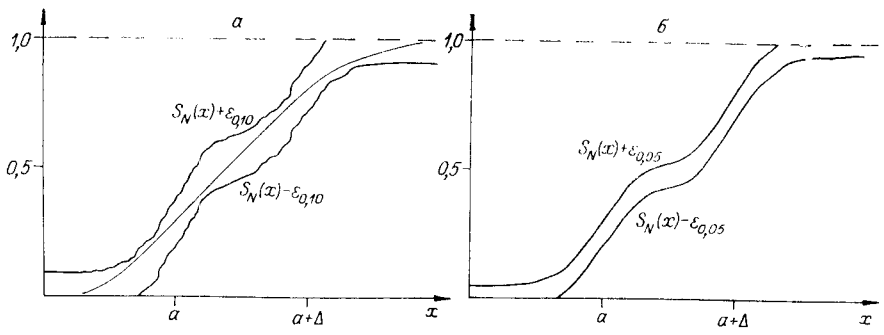

Рис. 3. Доверительные полосы, построенные по выборке из распределения (4) ($\Delta/\sigma = 4$):

а — $N = 200$, $\alpha=0{,}10$; *б* — $N=800$, $\alpha=0{,}05$.

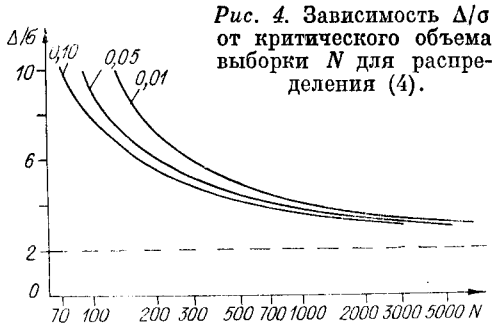

Рис. 4. Зависимость Δ/σ от критического объема выборки N для распределения (4).

ющихся данных для значимого утверждения о мультимодальности в каждом конкретном случае. Так, при получении гистограммы типа рис. 1 ($\Delta/\sigma = 4$) для этого требуется при $\alpha = 0{,}10$ выборка объемом не менее 600 измерений.

Заключение. К достоинствам предложенного графического метода проверки гипотезы унимодальности распределений следует отнести его непараметричность, простоту и наглядность. Малая мощность критерия снижает эффективность его практического применения. Возможно, использование других критериев согласия, например критерия χ^2, позволит улучшить эту характеристику.

ЛИТЕРАТУРА

1. **Большев Л. Н., Смирнов Н. В.** Таблицы математической статистики.— М.: Наука, 1965.
2. **Кендалл М. Дж., Стьюарт А.** Статистические выводы и связи.— М.: Наука, 1973.

Поступило в редакцию 23 февраля 1981 г.

УДК 55.51

В. А. ИВАНОВ, Г. А. ИВАНЧЕНКО, Н. Н. КАРЛСОН, Н. С. ЯКОВЕНКО
(Новосибирск)

ПРОГРАММНОЕ И ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ГЕОЛОГО-ГЕОФИЗИЧЕСКОЙ БАЗЫ ДАННЫХ

Совместный анализ фотоизображений территории и геолого-геофизических данных (ГГД) этой местности имеет определяющее значение при выявлении районов, перспективных для разведки полезных ископаемых, а также для оценки их запасов [1]. Для анализа фотоснимков на комплексе цифровой обработки изображений [2], управляемом мини-ЭВМ ЕС-1010, используется эксплуатируемая на протяжении последних пяти лет база данных цифровых изображений (БДЦИ) [3, 4].

Основная трудность обработки ГГД на ЭВМ заключена в представлении их в виде карт. В настоящем сообщении описывается первая очередь геолого-геофизической базы данных (ГГБД), программное обеспечение которой позволяет осуществлять ввод/вывод, хранение, редактирование и анализ ГГД совместно со снимками.

Ввод/вывод графической информации. Ввод закодированной графической информации (реперы, точки, изолинии и т. п.) с карты на магнитную ленту и вывод данных на бумагу производится графопостроителем-кодировщиком (г/к) «Планшет» [5]. В составе г/к «Планшет» имеются управляющая микроЭВМ «Электроника-60М», связанная с мини-ЭВМ «Nord-100», и «Электроника 100-25». ЭВМ «Nord-100» служит для накопления данных и записи на НМЛ, а также чтения с НМЛ и вывода на «Планшет». «Электроника 100-25» используется как инструментальная ЭВМ для поддержки управляющего программного обеспечения г/к «Планшет». Применение «Планшета» позволяет осуществлять ввод/вывод графической информации в формате 840×1200 мм^2 с разрешением 0,01 мм. Практическая точность установки визира для ввода координаты не хуже 0,5 мм. Файл временного хранения данных дает возможность накопить до 10^5 закодированных точек.

Системы баз данных. Следуя терминологии [6], объектами ГГБД являются различные точки на кодируемой карте. Объект имеет три атрибута: признак (реперная точка, изолиния, профиль, линеамента и прочие точки), значение геолого-геофизической величины (например, значение температуры в скважине) и координату кодируемой точки (1 дискрет = 0,1 мм). Домен атрибутов — множество целых чисел. Атрибут «признак» — ключ для всего набора объектов (т. е. значение ключа однозначно идентифицирует каждый объект во всем наборе).

Физическая база данных размещается на магнитных дисках ЕС-5052 емкостью 7,5 Мб и поддерживается системами управления файлами FMS-10 и FMS-D операционной системы DOS-10 мини-ЭВМ ЕС-1010. При помощи существующего в комплексе функционально полного набора цифровых средств ввода/вывода изображений система управления БДЦИ позволяет вводить снимки в файлы малого (256×256 байт) и большого (1024×1024 байт) форматов, визуализировать их на черно-белых и цветных полутоновых дисплеях, а также осуществлять их хранение и обработку.

Система управления ГГБД позволяет данные, записанные на МЛ, а также на перфолентах и перфокартах, вводить с возможным преобразованием формата в файлы, а в случае необходимости результат анализа выводить в аналогичном виде на