УДК 519.2

Ю.В. Потапов

МЕТОД ВЕРОЯТНОСТНОЙ БУМАГИ ДЛЯ ЗАДАЧИ ОДНОРОДНОСТИ

Предлагается глазомерный метод анализа правдоподобных видов альтернатив к гипотезе однородности двух выборок.

В математической статистике известен глазомерный метод вероятностной бумаги (Probability Plot) для анализа задачи согласия [1. §9.11; 2. §5.1].

Для имеющейся выборки повторных и независимых наблюдений $\{x_i\}_{i=1}^{n}$ проверяется сложная гипотеза согласия $H_0^C: F_{ГС}(x) \equiv F_0\left[(x-\mu)/\sigma\right]$ о том, что наблюдения извлечены из генеральной совокупности с функцией распределения определенного непрерывного и строго монотонного типа F_0 при неизвестных значениях параметров сдвига μ и масштаба σ (К примеру, проверяется, что выборочные данные являются нормальными с неизвестным математическим ожиданием и с неизвестным среднеквадратическим отклонением).

В методе вероятностной бумаги просматривается вариационный ряд выборки $\{x_{(r)}\}_{r=1}^{n}$ и для каждой порядковой статистики $x_{®}$ как для квантиля гипотетического распределения F_0 вычисляется значение $y_r = F_0^{-1}(\text{¢}_r)$, где ¢_r – какая-либо состоятельная оценка нижнего квантильного уровня у порядковой статистики ранга r. На практике часто [2. С. 163] берут $\text{¢}_r = (2r-1)/2r$, , т.е. как середину «скачка» в точке $x_{(r)}$ у эмпирической функции распределения на выборке $F_Э(x) = \frac{1}{n}\sum_{r=1}^{n} E(x - x_{(r)})$, где $E(z) = \begin{cases} 1, \text{если } z > 0; \\ 0, \text{если } z \le 0. \end{cases}$ – функция Хевисайда.

Затем на графике в координатах XOY откладываются точки

$$\left\{\left(x_{(r)}, F_0^{-1}\left[\frac{2r-1}{2n}\right]\right)\right\}_{r=1}^{n} \qquad (1)$$

и «на глаз» анализируется, укладываются ли они стохастически на прямую. Если это так, то выдвинутую гипотезу согласия H_0^C можно признать правдоподобной.

Теоретическим обоснованием метода служит то, что в силу известного факта стохастической сходимости эмпирической функции распределения $F_Э(x)$ к истинной функции распределения генеральной совокупности $F_{ГС}(x)$ оценка квантильного уровня ¢_r стохастически сходится к величине $F_{ГС}(x_{(r)})$ [1. С. 257]. Но при верности гипотезы согласия H_0^C величина y_r стохастически сходится к значению $F_0^{-1}\left[F_0\left(\frac{x_{(r)}-\mu}{\sigma}\right)\right] \equiv \frac{x_{(r)}-\mu}{\sigma}$. Это и означает, что случайные точки (1) стохастически должны укладываться на прямую

$$y = \frac{x-\mu}{\sigma}. \qquad (2)$$

Рассмотрим задачу однородности. В приложениях математической статистики под ней подразумевают проблему отсутствия эффекта обработки [2. §3.5]. Формально же задача заключается в следующем. Имеется исходная выборка наблюдений $\{x_i\}_{i=1}^{n_x}$ – значения некоторого показателя (например, урожайности) до обработки, а затем получена выборка наблюдений $\{y_j\}_{j=1}^{n_y}$ – значения этого же показателя после обработки (после внесения удобрений). Выборки предполагаются повторными и независимыми, извлеченными соответственно из генеральных совокупностей с функциями распределения $F(x)$ и $G(y)$. Эти функции в общей ситуации не известны, но предполагаются непрерывными и строго монотонными. Проверяется гипотеза однородности [1. § 9.13] $H_0: F(z) \equiv G(z)$ о том, что обе выборки фактически извлечены из одной и той же генеральной совокупности, т.е. что эффекта обработки нет (урожайность не изменилась после внесения удобрений).

В математической статистике известны критерии для проверки сформулированной гипотезы H_0. Выбор подходящего теста однородности зависит от характера возможных альтернатив H_1 к гипотезе H_0. Для анализа правдоподобных на полученных выборочных данных альтернатив к гипотезе однородности H_0 нами и предлагается методика, развивающая идею метода вероятностной бумаги.

При решении задачи однородности все равно какую из двух введенных выше выборок считать первой, а какую второй. Условимся за первую рассматривать выборку меньшего объема и обозначим ее через x. Вторую выборку обозначим y, причем $n_x \le n_y$.

Построим вариационные ряды выборок $\{x_{(r_x)}\}_{r_x=1}^{n_x}$ и $\{y_{(r_y)}\}_{r_y=1}^{n_y}$ и по аналогии с (1) рассмотрим на графике в координатах XOY точки:

$$\left\{\left(x_{(r_x)}, G_Э^{-1}\left[\frac{2r_x-1}{2n_x}\right]\right)\right\}_{r_x=1}^{n_x} \equiv \{(x_{(r_x)}, y_{(\tilde{r}_y)})\}_{r_x=1}^{n_x}, \qquad (3)$$

где за оценку неизвестной функции распределения $G(y)$ генеральной совокупности для второй выборки взята ее эмпирическая функция распределения

$$G_Э(y) = \frac{1}{n_y}\sum_{r_y=1}^{n_y} E(y - y_{(r_y)}).$$

Обращение функции $G_Э(y)$ в (3) фактически сводится (рис. 1) к поиску в вариационном ряду второй выборки такого ранга $\tilde{r}_y$, который удовлетворяет условиям:

$$\frac{\tilde{r}_y - 1}{n_y} \le \frac{2r_x - 1}{2n_x} < \frac{\tilde{r}_y}{n_y}. \qquad (4)$$

Разрешение условий (4) дает значение ранга $\tilde{r}_y$ как функции от ранга r_x в вариационном ряду первой выборки в виде

$$\tilde{r}_y = \tilde{r}_y(r_x) = \text{целая часть}\left[\frac{n_y}{n_x}\left(r_x - \frac{1}{2}\right) + 1\right], \qquad (5)$$

т.е. номера рангов порядковых статистик второй выборки в (3) оказываются «просеянными» по правилу (5) для

номеров рангов $r_x=\overline{1,n_x}$ на первой выборке. Например, если $n_x=10$, а $n_y=15$, то «просеивание» дает результаты – r_x:1,2,3,4,5,6,7,8,9,10; $\tilde{r}_y$:1,3,4,6,7,9,10,12,13,15. В частном случае при $n_x=n_y$ «просеивания» нет.

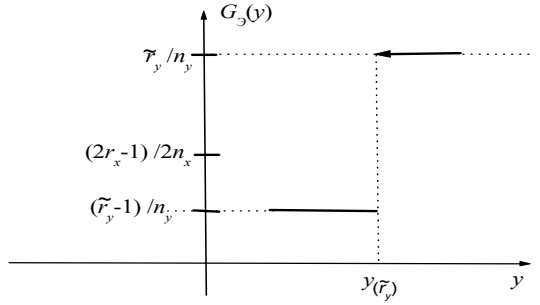

Рис. 1

В силу качеств сходимости эмпирической функции распределения величина $y_{(\tilde{r}_y)}$ в (3) при (5) стохастически сходится к значению $G^{-1}[F(x_{(r_x)})]$, т.е. точки (3) стохастически должны укладываться на линию с уравнением

$$y = G^{-1}[F(x)]. \qquad (6)$$

В частности, если функции распределения F и G принадлежат одному типу, т.е. в задаче однородности имеет место сдвиго-масштабная альтернатива $H_1^{CM}: F(z) \equiv G\left[(z-\Delta)/\Theta\right]$, то линия (6) подобно (2) оказывается прямой (рис. 2)

$$y = (x-\Delta)/\Theta. \qquad (7)$$

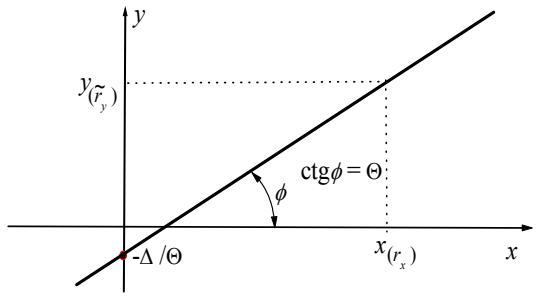

Рис. 2

Решить, укладываются ли случайные точки (3) на прямую, можно не только «на глаз», но и более формально по значимости выборочного коэффициента корреляции между рядами x и y в (3). А восстановить теоретическую прямую (7), т.е. оценить параметры Δ и Θ по точкам (3), можно методом наименьших квадратов.

Анализируя расположение точек (3) на плоскости, можно конкретизировать проблему однородности. Если точки явно не укладываются на прямую, то это означает, что эффект обработки между первой и второй выборками есть, и он сложно проявляется в искажении формы распределения генеральной совокупности от F к G. Здесь к общей гипотезе однородности H_0 напрашивается общая альтернатива $H_1: F(z) \neq G(z)$, а для проверки гипотез должен быть использован универсальный тест однородности типа ω^2 Розенблатта – Смирнова [3. С. 86].

Но и в такой ситуации можно уточнить общую альтернативу H_1 по характеру выпуклости кривой (6), на которую стохастически укладываются точки (3). К примеру, если на самом деле к гипотезе H_0 имеет место односторонняя альтернатива вида $H_1^{ОД}: F(z) < G(z)$, то, как ясно из рис. 3, ординаты y точек кривой (6) всегда оказываются меньше абсцисс x, а значит, кривая (6) должна иметь прогнутый вид – как на рис. 4. Для проверки таких гипотез может быть использован универсальный односторонний тест однородности типа теста Колмогорова – Смирнова [3. С. 83].

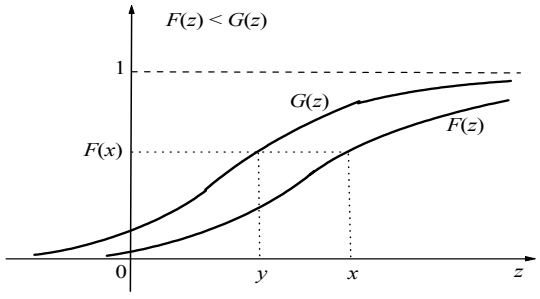

Рис. 3

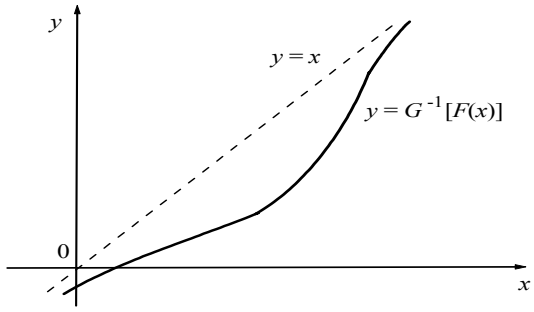

Рис. 4

Если же точки (3) стохастически укладываются на прямую, то это означает, что эффект обработки между первой и второй выборками проявляется лишь в сдвиго-масштабных искажениях, но не в изменении формы распределения генеральной совокупности. При этом от общей гипотезы однородности H_0 можно перейти к частным параметрическим гипотезам о сдвиге $H_0^{\Delta}: \Delta=0$ или о масштабе $H_0^{\Theta}: \Theta=1$ с подходящими односторонними альтернативами (в зависимости от значений оценок параметров Δ и Θ прямой (7)). А для проверки таких гипотез может быть использован ранговый тест однородности типа Вилкоксона [3. С. 93].

ЛИТЕРАТУРА

1. *Тарасенко Ф.П.* Непараметрическая статистика. Томск: Изд-во Том. ун-та, 1976. 294 с.
2. *Тюрин Ю.Н., Макаров А.А.* Анализ данных на компьютере /Под ред. В.Э. Фигурнова. М.: ИНФРА-М; Финансы и статистика, 1995. 384 с.
3. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики, 3-е изд. М.: Наука, 1983. 416 с.

Статья представлена кафедрой прикладной информатики факультета информатики Томского государственного университета, поступила в научную редакцию номера 3 декабря 2001 г.