Serdica
Bulgariacae mathematicae publicationes

Сердика
Българско математическо списание

ОДНО РЕШЕНИЕ ЗАДАЧИ ОБ ОДНОРОДНОСТИ

Л. Н. БОЛЬШЕВ, М. С. НИКУЛИН

Рассматривается задача о проверке однородности нескольких независимых выборок элементы каждой из которых независимы и одинаково непрерывно распределены. Предполагается, что соответствующие функции распределения принадлежат одному и тому же заданному параметрическому семейству. Гипотеза однородности заключается в утверждении, что всем выборкам соответствует одна и та же функция распределения из заданного параметрического семейства. Соответствующие этой функции значения параметров предполагаются неизвестными, поэтому речь идет о проверке сложной гипотезы. В этой ситуации конструируется критерий, статистика которого при неограниченном увеличении объемов выборок имеет распределение, сходящееся к распределению хи-квадрат (центральному — в случае справедливости гипотезы однородности, и нецентральному — при близких альтернативах). Обсуждаются свойства этого критерия для решения проблемы двух выборок в предположении, что параметрическое семейство распределений зависит только от параметров сдвига и масштаба.

Аналогичная задача для случая двух выборок рассматривалась в статьях В. Мюрти и А. Гафаряна (1960) и Г. Чейза (1972). Однако решения, предложенные упомянутыми авторами, приводят к потере в мощности, поскольку статистики критериев несимметричны относительно элементов обеих выборок. Кроме того, распределения этих статистик с ростом числа наблюдений сходятся к нестандартному распределению, отличному от распределения хи-квадрат.

1. Рассмотрим r выборок $\xi_{i,1}, \ldots, \xi_{i,n_i}$ $(i=1,\ldots,r; r\geq 2)$. Предполагается, что все ξ_{ij} взаимно независимы, причем $\xi_{i1}, \ldots, \xi_{in_i}$ распределены одинаково, и при всех значениях $i=1,\ldots,r$ соответствующие функции распределения принадлежат параметрическому семейству

$$F(x \mid \theta) = \int_{-\infty}^{x} f(x \mid \theta)dx$$

($x \in R_1$, $\theta = (\theta_1, \ldots, \theta_s)^{\mathrm{T}} \in \Theta \subset R_s$, Θ — открытое множество). Таким образом распределения вероятностей, соответствующие каждой из выборок, известны с точностью до значения параметра θ. Пусть истинное значение неизвестного параметра θ для i-й выборки $(i=1,\ldots,r)$ есть $\theta^i = (\theta_1^i, \ldots, \theta_s^i)^{\mathrm{T}} \in \Theta$. Данное сообщение посвящено применению критерия хи-квадрат для статистической проверки гипотезы H_0, согласно которой $\theta^1 = \ldots = \theta^r = \theta^0$ (θ^0 — истинное значение параметра θ в предположении, что гипотеза H_0 верна). Зафиксируем вектор $p = (p_1, \ldots, p_k)^{\mathrm{T}}$ такой, что $p_1 > 0, \ldots, p_k > 0$, $p_1 + \cdots + p_k = 1$ $(k > s+1)$. Положим $x_j(\theta) = F^{-1}(p_1 + \cdots + p_j, \theta)$, $j = 1, \ldots,$

Сердика. Българско математическо списание. Том. 1, 1975, с. 104—109.

$k-1$; $x_0(\theta)=-\infty$, $x_k(\theta)=+\infty$. Пусть $\bar{\theta}_N$ — асимптотически эффективная по Рао оценка неизвестного параметра θ^0, вычисленная по всем наблюдениям ξ_{ij} в количестве $N=n_1+\cdots+n_r$. В таком случае (см. Д. Мур [1])

$$\sqrt{N}(\bar{\theta}_N-\theta^0)=\frac{1}{\sqrt{N}}\sum_{i=1}^{r}\sum_{j=1}^{n_i}BA(\xi_{ij})+o_P(1),$$

где $B=B(\theta^0)$ — некоторая невырожденная матрица порядка $s\times s$, элементы которой, вообще говоря, зависят от θ^0, $A(\xi_{ij})$ — градиент функции $\log f(\xi_{ij}|\theta)$ в точке $\theta=\theta^0$:

$$A(\xi_{ij})=\frac{\partial}{\partial\theta}\log f(\xi_{ij}\,|\,\theta)\,|_{\theta=\theta^0},$$

$o_P(1)$ — случайный вектор размерности s, сходящийся по вероятности к нулевому вектору при $N\to\infty$. Будем предполагать, что при $N\to\infty$ объемы выборок $n_1,\ldots,n_r$ неограниченно увеличиваются так, что $c<n_i/N<C$, $i=1,\ldots,r$, где c и C — некоторые положительные константы. Потребуем также, чтобы матрица $L=B+B^{\mathrm{T}}-BIB^{\mathrm{T}}$ была невырождена ($I=\mathrm{E}AA^{\mathrm{T}}$ — информационная матрица Фишера).

Пусть $\mu_i=(\mu_{i1},\ldots,\mu_{ik})^{\mathrm{T}}$ — результат группировки случайных величин $\xi_{i1},\ldots,\xi_{in_i}$ $(i=1,\ldots,r)$ по интервалам $(-\infty;x_1(\bar{\theta}_N)]$, $(x_1(\bar{\theta}_N);x_2(\bar{\theta}_N)],\ldots$ $(x_{k-1}(\bar{\theta}_N);+\infty)$, и пусть $\tilde{\mu}_i=(\mu_{i1},\ldots,\mu_{i,k-1})^{\mathrm{T}}$, $i=1,\ldots,r$, $\tilde{p}=(p_1,\ldots,p_{k-1})^{\mathrm{T}}$. Как показал Мур [1], если гипотеза H_0 имеет место, то вектор $\nu_i=n_i^{-1/2}(\tilde{\mu}_i-n_i\tilde{p})$, $i=1,\ldots,r$, при $N\to\infty$ распределен асимптотически нормально с нулевым вектором средних и невырожденной ковариационной матрицей

$$\Xi_i=\Xi_i(\theta^0)=D-\tilde{p}\,\tilde{p}^{\mathrm{T}}-\frac{n_i}{N}W^{\mathrm{T}}LW,$$

где D — диагональная матрица с элементами $p_1,\ldots,p_{k-1}$ на главной диагонали, $W=W(\theta^0)=\|w_{ij}\|$ — матрица порядка $s\times(k-1)$ с элементами

$$w_{ij}=w_{ij}(\theta^0)=\int_{x_{j-1}(\theta^0)}^{x_j(\theta^0)}\frac{\partial}{\partial\theta_i}f(x\,|\,\theta)\,|_{\theta=\theta^0}dx.$$

Очевидно, что вектор $\nu=(\nu_1^{\mathrm{T}},\nu_2^{\mathrm{T}},\ldots,\nu_r^{\mathrm{T}})^{\mathrm{T}}$ при $N\to\infty$ распределен асимптотически нормально с нулевым вектором средних размерности $r(k-1)$ и невырожденной ковариационной матрицей U, которая состоит из блоков U_{ij}, $i,j=1,\ldots,r$, где

$$U_{ij}=\begin{cases}\Xi_i, & \text{если } i=j,\\ -\dfrac{\sqrt{n_in_j}}{N}W^{\mathrm{T}}LW, & \text{если } i\neq j.\end{cases}$$

Если матрица $\Omega=[L^{-1}-WB^{-1}W^{\mathrm{T}}]^{-1}$ существует, то существует матрица U^{-1}, и квадратичная форма

$$Y^2(\theta^0)=\nu^{\mathrm{T}}U^{-1}(\theta^0)\nu$$

имеет в пределе при $N \to \infty$ хи-квадрат распределение с $r(k-1)$ степенями свободы. Используя для обращения матрицы U обобщение на блочные матрицы результата упражнения 2.9 учебника Рао ([2], с. 45), легко убедиться, что квадратичную форму $Y^2(\theta^0)$ можно представить в виде

$$Y^2(\theta^0) = \sum_{i=1}^{r} X_i^2 + N^{-1} \sum_{i=1}^{s} \sum_{j=1}^{s} \omega_{ij} a_i a_j,$$

где

$$X_i^2 = \sum_{j=1}^{k} (\mu_{ij} - n_i p_j)^2 / n_i p_j, \quad a_i = \sum_{l=1}^{k} \mu_l w_{il}(\theta^0)/p_l,$$

$\mu_l = \mu_{1l} + \cdots + \mu_{rl}$, ω_{ij} — элемент матрицы $\Omega = \Omega(\theta^0)$.

Квадратичную форму $Y^2(\theta^0)$ нельзя непосредственно использовать для проверки гипотезы H_0, так как она зависит от неизвестного параметра θ^0. Но из свойств регулярности $F(x \mid \theta)$ на $R_1 \times \Theta$ и самостоятельности $\bar{\theta}_N$ следует, что при $N \to \infty$ статистика $Y^2(\theta_N)$ и квадратичная форма $Y^2(\theta^0)$ асимптотически одинаково распределены. Таким образом, для проверки гипотезы H_0 можно использовать статистику $Y^2(\bar{\theta}_N)$, которая имеет в пределе при $N \to \infty$ хи-квадрат распределение с $r(k-1)$ степенями свободы.

2. Предположим, что проверяемая гипотеза H_0 неверна, и на самом деле имеет место гипотеза H_1, согласно которой соотношение $\theta^1 = \theta^2 = \cdots = \theta^r = \theta^0$ не выполняется. Обозначим $\theta^* = (n_1\theta^1 + \cdots + n_r\theta^r)/N$. Так как $\bar{\theta}_N - \theta^* = o_P(1)$ $(N \to \infty)$, то при гипотезе H_1

$$\mathbf{P}\{x_{l-1}(\bar{\theta}_N) < \xi_{ij} \leq x_l(\bar{\theta}_N)\} = p_l + c_{il} + o(1),$$

где $c_{il} = (\theta^i - \theta^*)^{\mathrm{T}} W_l(\theta^*)$, $W_l(\theta^*) = (w_{1l}, \ldots, w_{sl})^{\mathrm{T}}$ — l-й столбец матрицы $W(\theta^*)$. Поступая так же, как в работе Д. М. Чибисова [3], убеждаемся, что если верна гипотеза H_1, то статистика $Y^2(\bar{\theta}_N)$ будет иметь в пределе при $N \to \infty$ нецентральное распределение хи-квадрат с $r(k-1)$ степенями свободы и параметром нецентральности

$$\varkappa(\theta^*) = \sum_{i=1}^{r} \sum_{l=1}^{k} n_i c_{il}^2 / p_l.$$

3. Рассмотрим теперь задачу об однородности в случае двух выборок в предположении, что параметрическое семейство распределений зависит только от параметров сдвига и масштаба. Итак, пусть имеем две выборки $\xi_{1,1}, \ldots, \xi_{1,n_1}$ и $\xi_{2,1}, \ldots, \xi_{2,n_2}$, причем все ξ_{ij} взаимно независимы, $\xi_{i1}, \ldots, \xi_{in_i}$ $(i = 1, 2)$ распределены одинаково и соответствующие функции распределения принадлежат параметрическому семейству $\{G[(x-m)/b]\}$, причем $|x| < \infty$, $|m| < \infty$, $b > 0$ и $g(y) = G'(y) > 0$ при всех действительных значениях y.

Обозначим истинные значения параметров m и b для первой и второй выборок соответственно m_1, b_1 и m_2, b_2. Проверяется гипотеза H_0,

согласно которой $m_1=m_2$, $b_1=b_2$. Пусть $\hat{m}$ и $\hat{b}$ — оценки максимального правдоподобия для m и b, вычисленные по всем $N=n_1+n_2$ наблюдениям в случае справедливости гипотезы H_0. Зафиксируем, как и раньше, вектор $p=(p_1,\ldots,p_k)^{\mathrm{T}}$ и определим на действительной прямой точки $x_0, x_1,\ldots,x_k$ по правилу: $x_0=-\infty$, $x_k=+\infty$,

$$x_i=G^{-1}(p_1+\cdots+p_i),\quad i=1,\ldots,k-1.$$

Пусть $\mu_1=(\mu_{1,1},\ldots,\mu_{1,k})^{\mathrm{T}}$ и $\mu_2=(\mu_{2,1},\ldots,\mu_{2,k})^{\mathrm{T}}$ — векторы, получающиеся в результате группировки элементов первой и второй выборок соответственно по полуинтервалам $(\hat{m}+\hat{b}x_{l-1},\ \hat{m}+\hat{b}x_l]$, $l=1,\ldots,k$. В этом случае векторы $\nu_1=n_1^{-1/2}(\widetilde{\mu_1}-n_1\widetilde{p})$ и $\nu_2=n_2^{-1/2}(\widetilde{\mu_2}-n_2\widetilde{p})$ при $n_1\to\infty$ и $n_2\to\infty$ асимптотически нормально распределены с нулевыми векторами средних и матрицами ковариаций Ξ_1 и Ξ_2 соответственно, где

$$\Xi_i=D-\widetilde{p}\widetilde{p}^{\mathrm{T}}-\frac{n_i}{N}W^{\mathrm{T}}I^{-1}W\quad(i=1,2),$$

причем в данном случае $W=(u,v)$, где $u=(u_1,\ldots,u_{k-1})^{\mathrm{T}}$, $v=(v_1,\ldots,v_{k-1})^{\mathrm{T}}$, $u_i=g(x_i)-g(x_{i-1})$, $v_i=x_ig(x_i)-x_{i-1}g(x_{i-1})$, $i=1,\ldots,k$. Элементы информационной матрицы $I=\|I_{ij}\|$, $i,j=1,2$, выражаются формулами

$$I_{11}=\int_{-\infty}^{\infty}(g'/g)^2g\,dx,\quad I_{22}=\int_{-\infty}^{\infty}x^2(g'/g)^2g\,dx-1,$$

$$I_{12}=I_{21}=\int_{-\infty}^{\infty}x(g'/g)^2g\,dx.$$

Напомним, что $g(x)=dG(x)/dx$.

Таким образом, матрицы Ξ_1 и Ξ_2 не зависят от неизвестных параметров и невырождены (см. [1] и [4]).

Вектор $\nu=(\nu_1^{\mathrm{T}},\nu_2^{\mathrm{T}})^{\mathrm{T}}$ при $N=n_1+n_2\to\infty$ распределен асимптотически нормально с невырожденной ковариационной матрицей

$$U=\begin{Vmatrix}\Xi_1 & -\frac{\sqrt{n_1n_2}}{N}W^{\mathrm{T}}I^{-1}W\\ -\frac{\sqrt{n_1n_2}}{N}W^{\mathrm{T}}I^{-1}W & \Xi_2\end{Vmatrix},$$

причем

$$U^{-1}=\begin{Vmatrix}C & 0\\ 0 & C\end{Vmatrix}+\frac{1}{N}\begin{Vmatrix}n_1M & \sqrt{n_1n_2}M\\ \sqrt{n_1n_2}M & n_2M\end{Vmatrix},$$

где $C=D^{-1}+(1/p_k)\mathbf{1}\cdot\mathbf{1}^{\mathrm{T}}$, $\mathbf{1}=(1,\ldots,1)^{\mathrm{T}}$ — вектор размерности $(k-1)$, все компоненты которого равны единице,

$$M=\frac{1}{\lambda_1\lambda_2-\lambda_3^2}[\lambda_2yy^{\mathrm{T}}-\lambda_3(zy^{\mathrm{T}}+yz^{\mathrm{T}})+\lambda_1zz^{\mathrm{T}}],$$

$$y = D^{-1}u - \frac{u_k}{p_k}\,\mathbf{1}, \quad z = D^{-1}v - \frac{v_k}{p_k}\,\mathbf{1},$$

$$\lambda_1 = I_{11} - \sum_{i=1}^{k}\frac{u_i^2}{p_i}, \quad \lambda_2 = I_{22} - \sum_{i=1}^{k}\frac{v_i^2}{p_i}, \quad \lambda_3 = I_{12} - \sum_{i=1}^{k}\frac{u_i v_i}{p_i}.$$

Согласно результатам из п. 1 статистика $Y^2 = \nu^{\mathrm{T}} U^{-1}\nu$ имеет в пределе при $N = n_1 + n_2 \to \infty$ распределение хи-квадрат с $2(k-1)$ степенями свободы. Используя представление матрицы U^{-1}, получаем, что

$$Y^2 = \sum_{i=1}^{k}(\mu_{1i} - n_1 p_i)^2/n_1 p_i + \sum_{i=1}^{k}(\mu_{2i} - n_2 p_i)^2/n_2 p_i$$
$$+ [N(\lambda_1\lambda_2 - \lambda_3^2)]^{-1}\left[\lambda_1\left(\sum_{i=1}^{k} v_i\mu_i/p_i\right)^2 + \lambda_2\left(\sum_{i=1}^{k} u_i\mu_i/p_i\right)^2\right.$$
$$\left. - 2\lambda_3\left(\sum_{i=1}^{k} u_i\mu_i/p_i\right)\left(\sum_{j=1}^{k} v_j\mu_j/p_j\right)\right],$$

где $\mu_i = \mu_{1i} + \mu_{2i}$, $i = 1, \ldots, k$.

4. Если существует достаточная статистика, то целесообразно использовать другие способы для проверки гипотезы H_0, поскольку эта гипотеза носит параметрический характер. Рассмотрим в качестве примера проблему проверки однородности двух выборок (см. п. 3) при дополнительном предположении, что $G(x) = \Phi(x)$ — функция стандартного нормального распределения. Пусть $\Phi[(x-m_1)/\sigma_1]$ и $\Phi[(x-m_2)\sigma_2]$ — функции распределения элементов первой и второй выборок соответственно. В этом случае достаточной статистикой является вектор $(\bar{\xi}_1, \bar{\xi}_2, s_1^2, s_2^2)$, где

$$\bar{\xi}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\xi_{ij}, \quad s_i^2 = \frac{1}{f_i}\sum_{j=1}^{n_i}(\xi_{ij} - \bar{\xi}_i)^2, \quad f_i = n_i - 1, \quad i = 1, 2.$$

Компоненты достаточной статистики — взаимно независимые случайные величины, причем $\bar{\xi}_i = m_i + \zeta_i\sigma_i/n_i$, $s_i^2 = \chi_i^2\sigma_i^2/f_i$ $(i = 1, 2)$, ζ_1 и ζ_2 подчиняются стандартному нормальному распределению, а χ_1^2 и χ_2^2 подчиняются распределениям хи-квадрат с f_1 и f_2 степенями свободы соответственно.

Положим $M_i = \sqrt{2/f_i}\,\Gamma[(f_i+1)/2]/\Gamma(f_i/2)$, $i = 1, 2$, и рассмотрим статистику

$$Z^2 = \frac{(f_1+f_2)(f_1+f_2-2)}{(f_1+f_2-1)(f_1 s_1^2 + f_2 s_2^2)}\left[\frac{(\bar{\xi}_1 - \bar{\xi}_2)^2}{n_1^{-1} + n_2^{-1}} + \frac{(s_1/M_1 - s_2/M_2)^2}{1/M_1^2 + 1/M_2^2 - 2}\right].$$

Если гипотеза H_0 верна (т. е. $m_1 = m_2$ и $\sigma_1 = \sigma_2$), то с помощью элементарных преобразований можно убедиться, что

$$Z^2 = \frac{f_1+f_2-2}{f_1+f_2-1}\left[t^2 + \frac{f_1+f_2}{1/M_1^2 + 1/M_2^2 - 2}\left(\frac{1}{M_1\sqrt{f_1}}\sqrt{\beta} - \frac{1}{M_2\sqrt{f_2}}\sqrt{1-\beta}\right)^2\right],$$

где t и β — взаимно независимые случайные величины, причем t подчиняется распределению Стьюдента с количеством степеней свободы f_1+f_2, а β подчиняется бета-распределению с параметрами $f_1/2$ и $f_2/2$. Если $N=n_1+n_2\to\infty$, то распределение Z^2 стремится к распределению хи-квадрат с двумя степенями свободы.

При альтернативной гипотезе, близкой к H_0, статистика Z^2 имеет распределение, аппроксимирующееся нецентральным распределением хи-квадрат с двумя степенями свободы и параметром нецентральности

$$\frac{(f_1+f_2)(f_1+f_2-2)}{(f_1+f_2-1)(f_1\sigma_1^2+f_2\sigma_2^2)}\left[\frac{(m_1-m_2)^2}{n_1^{-1}+n_2^{-1}}+\frac{(\sigma_1-\sigma_2)^2}{1/M_1^2+1/M_2^2-2}\right].$$

ЛИТЕРАТУРА

1. D. S. Mooore. A chi-square statistic with random cell boundaris. *Ann. Math. Statist* **42**, 1971, No. 1, 147—156.
2. С. Р. Рао. Линейные статистические методы и их применения. Москва, 1968.
3. Д. М. Чибисов. Некоторые критерии типа хи-квадрат для непрерывных распределений. *Теория вероят. и ее примен.*, **16**, 1971, № 1, 3—20.
4. М. С. Никулин. Критерии хи-квадрат для непрерывных распределений с параметрами сдвига и масштаба. *Теория вероят. и ее примен.*, **17**, 1973, № 3, 583—592.
5. V. K. Murty, A. V. Gafarian. Limiting distribution of some variations of the chi-square statistic. *Ann. Math. Statist.*, **41**, 1970, No. 1, 188—194.
6. G. R. Chase. Chi-square test when parameters are estimated independently of the sample. *J. Amer. Statist. Assoc.*, **67**, 1972, 609—611.

Математический институт им. В. А. Стеклова АН СССР

Москва 117333 СССР

Поступила 12. 5. 1974