УДК 519.2

А.Н. ГОЛОДНИКОВ

О НЕПАРАМЕТРИЧЕСКОМ КРИТЕРИИ НЕЙМАНА–ПИРСОНА ДЛЯ ПРОВЕРКИ СЛОЖНОЙ ГИПОТЕЗЫ ОБ ОДНОМ СРЕДНЕМ

Рассмотрена задача построения критерия Неймана–Пирсона для проверки сложной непараметриче-ской гипотезы об одном сред-нем. Показано, что эта задача сводится к задачам линейного программирования. Получаемый при этом оптимальный критерий является довольно общим, поскольку при его построении не требуется незави-симости наблюдений.

Введение. Еще в 1928–1933 гг Нейман и Пирсон предложили в работах [1,2] формальную процедуру построения оптимальных кри-териев для проверки статистических гипотез. Была доказана лемма Неймана–Пирсона о наилучших критических областях, из которой следует, что наилучшим критерием для слу-чая, когда нулевая и конкурирующая гипоте-зы – простые, является отношение правдоподобия. Однако в более общем случае, когда обе конкурирующие гипотезы – сложные, задача построения гипотез в рамках этой формальной процедуры не решена до сих пор.

В данной статье рассматривается задача построения критерия Неймана–Пирсона для проверки сложной непараметрической гипотезы об одном среднем. Показано, что эта задача сводится к задачам линейного программирования.

Постановка задачи. В соответствии с про-цедурой Неймана–Пирсона [3] рассмат-рива-ются: вектор наблюдаемых случайных пере-менных $\mathbf{Y} = (Y_1, \ldots, Y_n)$; выборочное пространство W, состоящее из всех возможных выборочных точек $y = (y_1, \ldots, y_n)$; два конкурирующих множества допустимых простых гипотез K и Ω; плотности распределения (функции частот) $p_{\mathbf{Y}}(y \mid H)$, $p_{\mathbf{Y}}(y \mid \overline{H})$, свя-занные с простыми гипотезами $H \in K$ и $\overline{H} \in \Omega$, и отношение

$$\lambda(y) = \frac{\sup\limits_{H \in K} p_{\mathbf{Y}}(y \mid H)}{\sup\limits_{\overline{H} \in \Omega} p_{\mathbf{Y}}(y \mid \overline{H})}. \tag{1}$$

При заданном уровне значимости α выражения (1) используется для описания критической области, содержащей все точки y', для которых $\lambda(y') < t_{\alpha}$, и не содержащей ни одной точки y'', для которых $\lambda(y'') > t_{\alpha}$, где число t_{α} определяется из условия

$$P\{\lambda(y) \le t_{\alpha} \mid H\} = \alpha. \tag{2}$$

1. Непараметрический критерий при проверке гипотез о значении среднего одной генеральной совокупности. Пусть $x = (x_1, x_2, \ldots, x_n)$ – выборка объема n из некоторой одномерной генеральной совокупности. Требуется проверить гипотезу о том, что истинное значение среднего μ этой генеральной совокупности совпадает с некоторым стандартом μ_0, меньше или больше этого стандарта. Предполагается, что все значения исследуемой генеральной совокупности принадлежат конечному отрезку $[A, B]$, и ее среднее равно μ.

Рассмотрим следующие возможные гипотезы относительно среднего μ:
1) $K : \mu = \mu_0$; $\Omega : \mu < \mu_0$; 2) $K : \mu = \mu_0$; $\Omega : \mu > \mu_0$; 3) $K : \mu < \mu_0$; $\Omega : \mu > \mu_0$; 4) $K : \mu > \mu_0$; $\Omega : \mu < \mu_0$.

Пусть $F_{\mathbf{X}}(x_1, x_2, \ldots x_n)$ – функция совместного распределения вектора наблюдаемых случайных переменных $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ при условии, что верна простая гипотеза $H \in K$ и $G_{\mathbf{X}}(x_1, x_2, \ldots x_n)$ – функция совместного распределения этого вектора при условии, что верна конкурирующая простая гипотеза $\overline{H} \in \Omega$. Поскольку выборка осуществляется из одной генеральной совокупности, то для любой пары случайных переменных X_i и X_j, их маргинальные распределения равны. Это условие эквивалентно тому, что для любого значения $a \in [A, B]$ выполняются следующие равенства:

$$F_{X_1}(a) = F_{X_2}(a), F_{X_1}(a) = F_{X_3}(a), \ldots, F_{X_1}(a) = F_{X_n}(a), \tag{3}$$

$$G_{X_1}(a) = G_{X_2}(a), G_{X_1}(a) = G_{X_3}(a), \ldots, G_{X_1}(a) = G_{X_n}(a), \tag{4}$$

где $F_{X_i}(x), G_{X_i}(x), i = 1,2,\ldots,n$ – маргинальные функции распределения случайной переменной $X_i, i = 1,2,\ldots,n$, которые вычисляются по формулам:

$$F_{X_i}(x) = \int\limits_{X_1 \in [A,B]} \cdots \int\limits_{X_{i-1} \in [A,B]} \int\limits_{X_i \in [A,x]} \int\limits_{X_{i+1} \in [A,B]} \cdots \int\limits_{X_n \in [A,B]} dF_{\mathbf{X}}(x_1, x_2, \ldots, x_n), \; i = 1, \ldots, n,$$

$$G_{X_i}(x) = \int\limits_{X_1 \in [A,B]} \cdots \int\limits_{X_{i-1} \in [A,B]} \int\limits_{X_i \in [A,x]} \int\limits_{X_{i+1} \in [A,B]} \cdots \int\limits_{X_n \in [A,B]} dG_{\mathbf{X}}(x_1, x_2, \ldots, x_n), \; i = 1, \ldots, n,$$

В случае 1) и 2) любая простая гипотеза $H \in K$ предполагает, что среднее значение каждой из маргинальных функций распределения $F_{X_1}(x)$, $F_{X_2}(x)$,..., $\ldots, F_{X_n}(x)$ равно μ_0. В силу их равенства это условие достаточно записать для одной из них, например для $F_{X_1}(x)$:

$$\int_A^B x dF_{X_1}(x) = \int\limits_{X_1 \in [A,B]} x_1 \int\limits_{X_2 \in [A,B]} \cdots \int\limits_{X_n \in [A,B]} dF_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = \mu_0 . \tag{5}$$

Соответственно, в случае 3) любая простая гипотеза $H \in K$ налагает на $F_{X_1}(x)$ ограничение

$$\int_A^B x dF_{X_1}(x) = \int\limits_{X_1 \in [A,B]} x_1 \int\limits_{X_2 \in [A,B]} \cdots \int\limits_{X_n \in [A,B]} dF_{\mathbf{X}}(x_1, x_2, \ldots, x_n) < \mu_0 , \tag{6}$$

а в случае 4) – ограничение

$$\int_A^B x dF_{X_1}(x) = \int\limits_{X_1 \in [A,B]} x_1 \int\limits_{X_2 \in [A,B]} \cdots \int\limits_{X_n \in [A,B]} dF_{\mathbf{X}}(x_1, x_2, \ldots, x_n) > \mu_0 . \tag{7}$$

Аналогично, если предположить, что верна простая конкурирующая гипотеза $\overline{H} \in \Omega$, то в случаях 1) и 4) маргинальное распределение $G_{X_1}(x)$ должно удовлетворять условию

$$\int_A^B x dG_{X_1}(x) = \int\limits_{X_1 \in [A,B]} x_1 \int\limits_{X_2 \in [A,B]} \cdots \int\limits_{X_n \in [A,B]} dG_{\mathbf{X}}(x_1, x_2, \ldots, x_n) < \mu_0 , \tag{8}$$

а в случаях 2) и 3) – условию

$$\int_A^B x dG_{X_1}(x) = \int\limits_{X_1 \in [A,B]} x_1 \int\limits_{X_2 \in [A,B]} \cdots \int\limits_{X_n \in [A,B]} dG_{\mathbf{X}}(x_1, x_2, \ldots, x_n) > \mu_0 . \tag{9}$$

Все маргинальные функции распределения должны удовлетворять нормировочным ограничениям. В силу соотношений (3) и (4) эти ограничения достаточно выписать для маргинальных распределений $F_{X_1}(x)$ и $G_{X_1}(x)$. Таким образом, должны выполняться следующие ограничения:

$$\int_A^B dF_{X_1}(x) = 1 , \tag{10}$$

$$\int_A^B dG_{X_1}(x) = 1 . \tag{11}$$

Обозначим: $\Gamma_1 = \{ F_{\mathbf{X}}(x_1, x_2, \ldots x_n)$: удовлетворяются ограничения (3), (5) и (10)}; $\Gamma_2 = \{ F_{\mathbf{X}}(x_1, x_2, \ldots x_n)$: удовлетворяются ограничения (3), (6) и (10)};
$\Gamma_3 = \{ F_{\mathbf{X}}(x_1, x_2, \ldots x_n)$: удовлетворяются ограничения (3), (7) и (10)};

$L_1 = \{ G_{\mathbf{X}}(x_1, x_2, \dots x_n)$: удовлетворяются ограничения (4), (8) и (11)$\}$;
$L_2 = \{ G_{\mathbf{X}}(x_1, x_2, \dots x_n)$: удовлетворяются ограничения (4), (9) и (11)$\}$.

Рассмотрим статистику $Y(X_1, X_2, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$. Эта статистика генерирует случайную величину Y, значения которой совпадают с выборочным средним: $y = \frac{1}{n}\sum_{i=1}^{n} x_i$.

В терминах процедуры Неймана–Пирсона множество выборочных средних является выборочным пространством W. В случае, когда верна простая гипотеза $H \in K$, функция распределения $\Phi(y \mid H)$ случайной величины Y определяется по формуле

$$\Phi(y \mid H) = \Phi_{F_{\mathbf{X}}}(y) = \int\limits_{\sum_{i=1}^{n} x_i \le ny} \cdots \int dF_{\mathbf{X}}(x_1, x_2, \dots, x_n),$$

а в случае, когда верна конкурирующая гипотеза $\overline{H} \in \Omega$, ее функция распределения

$$\Phi(y \mid \overline{H}) = \Phi_{G_{\mathbf{X}}}(y) = \int\limits_{\sum_{i=1}^{n} x_i \le ny} \cdots \int dG_{\mathbf{X}}(x_1, x_2, \dots, x_n).$$

Пусть $\varphi(y \mid H) = \varphi_{F_{\mathbf{X}}}(y)$ обозначает плотность распределения случайной величины Y (или ее функцию частот) при условии, что верна простая гипотеза $H \in K$, а $\varphi(y \mid \overline{H}) = \varphi_{G_{\mathbf{X}}}(y)$ обозначает плотность распределения случайной величины Y (или ее функцию частот) при условии, что верна конкурирующая гипотеза $\overline{H} \in \Omega$.

При рассмотрении гипотез для случаев 1) – 4) необходимо вычислить выражения:

$$\lambda(y) = \frac{\sup\limits_{H \in K} p_{\mathbf{Y}}(y \mid H)}{\sup\limits_{\overline{H} \in \Omega} p_{\mathbf{Y}}(y \mid \overline{H})} = \frac{\sup\limits_{H \in K} \varphi(y \mid H)}{\sup\limits_{\overline{H} \in \Omega} \varphi(y \mid \overline{H})} = \frac{\sup\limits_{F_{\mathbf{X}} \in \Gamma_1} \varphi_{F_{\mathbf{X}}}(y)}{\sup\limits_{G_{\mathbf{X}} \in L_1} \varphi_{G_{\mathbf{X}}}(y)}; \quad \lambda(y) = \frac{\sup\limits_{F_{\mathbf{X}} \in \Gamma_1} \varphi_{F_{\mathbf{X}}}(y)}{\sup\limits_{G_{\mathbf{X}} \in L_2} \varphi_{G_{\mathbf{X}}}(y)};$$

$$\lambda(y) = \frac{\sup\limits_{F_{\mathbf{X}} \in \Gamma_2} \varphi_{F_{\mathbf{X}}}(y)}{\sup\limits_{G_{\mathbf{X}} \in L_2} \varphi_{G_{\mathbf{X}}}(y)}; \quad \lambda(y) = \frac{\sup\limits_{F_{\mathbf{X}} \in \Gamma_3} \varphi_{F_{\mathbf{X}}}(y)}{\sup\limits_{G_{\mathbf{X}} \in L_1} \varphi_{G_{\mathbf{X}}}(y)}.$$

Вычисление этих выражений сводится к решению следующих оптимизационных задач в пространстве функций распределения:

$$\sup_{F_{\mathbf{X}} \in \Gamma_1} \varphi_{F_{\mathbf{X}}}(y), \tag{12}$$

$$\sup_{F_{\mathbf{X}} \in \Gamma_2} \varphi_{F_{\mathbf{X}}}(y), \tag{13}$$

$$\sup_{F_{\mathbf{X}} \in \Gamma_3} \varphi_{F_{\mathbf{X}}}(y), \tag{14}$$

$$\sup_{G_{\mathbf{X}} \in L_1} \varphi_{G_{\mathbf{X}}}(y), \tag{15}$$

$$\sup_{G_{\mathbf{X}} \in L_2} \varphi_{G_{\mathbf{X}}}(y). \tag{16}$$

Эти задачи можно с любой заданной точностью аппроксимировать задачами линейного программирования путем аппроксимации функций совместного распределения $F_{\mathbf{X}}(x_1, x_2, \dots x_n)$ и $G_{\mathbf{X}}(x_1, x_2, \dots x_n)$ соответствующими ступенчатыми функциями распределения $\tilde{F}_{\mathbf{X}}(x_1, x_2, \dots x_n)$ и $\tilde{G}_{\mathbf{X}}(x_1, x_2, \dots x_n)$. Для этого разобъем отрезок $[A, B]$ на M равных интервала равной длины $\delta = (B - A) / M$. Ступенчатые функции распределения $\tilde{F}_{\mathbf{X}}(x_1, x_2, \dots x_n)$ и $\tilde{G}_{\mathbf{X}}(x_1, x_2, \dots x_n)$ могут иметь ненулевые скачки только в точках $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$, координаты которых представимы в виде $\tilde{x}_j = A + i_j \delta$, $i_j = 0,1, \dots, M$, $j = 1,2, \dots, n$. Таким образом, в n-мерном гиперкубе построена равномерная сетка, узлами которой являются точки $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$ с координатами $\tilde{x}_1 = A + i_1 \delta$, $\tilde{x}_2 = A + i_2 \delta$, ..., $\tilde{x}_n = A + i_n \delta$. Будем обозначать такие узлы $x_{i_1, i_2, \dots i_n}$, а скачки ступенчатых функций распределения $\tilde{F}_{\mathbf{X}}(x_1, x_2, \dots x_n)$ и $\tilde{G}_{\mathbf{X}}(x_1, x_2, \dots x_n)$ в этих точках – $p_{i_1, i_2, \dots, i_n}$ и $q_{i_1, i_2, \dots, i_n}$ соответственно. Тогда ограничения (10)–(16) можно записать в следующем виде:

$$\sum_{i_2, i_3, \dots, i_n} p_{k, i_2, \dots, i_n} = \sum_{i_1, i_3, \dots, i_n} p_{i_1, k, i_3, \dots, i_n}, \quad k = 0,1, \dots, M,$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$\sum_{i_2, i_3, \dots, i_n} p_{k, i_2, \dots, i_n} = \sum_{i_1, i_2, \dots, i_{n-1}} p_{i_1, i_2, \dots, i_{n-1}, k}, \quad k = 0,1, \dots, M, \tag{17}$$

$$\sum_{i_2, i_3, \dots, i_n} q_{k, i_2, \dots, i_n} = \sum_{i_1, i_3, \dots, i_n} q_{i_1, k, i_3, \dots, i_n}, \quad k = 0,1, \dots, M,$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$\sum_{i_2, i_3, \dots, i_n} q_{k, i_2, \dots, i_n} = \sum_{i_1, i_2, \dots, i_{n-1}} q_{i_1, i_2, \dots, i_{n-1}, k}, \quad k = 0,1, \dots, M, \tag{18}$$

$$\sum_{i=0}^{M} i\delta \sum_{i_2, i_3, \dots, i_n} p_{i, i_2, \dots, i_n} = \mu_0, \tag{19}$$

$$\sum_{i=0}^{M} i\delta \sum_{i_2,i_3,\dots,i_n} p_{i,i_2,\dots,i_n} < \mu_0, \tag{20}$$

$$\sum_{i=0}^{M} i\delta \sum_{i_2,i_3,\dots,i_n} p_{i,i_2,\dots,i_n} > \mu_0, \tag{21}$$

$$\sum_{i=0}^{M} i\delta \sum_{i_2,i_3,\dots,i_n} q_{i,i_2,\dots,i_n} < \mu_0, \tag{22}$$

$$\sum_{i=0}^{M} i\delta \sum_{i_2,i_3,\dots,i_n} q_{i,i_2,\dots,i_n} > \mu_0. \tag{23}$$

$$\sum_{i=0}^{M} \sum_{i_2,i_3,\dots,i_n} p_{i,i_2,\dots,i_n} = 1, \tag{24}$$

$$\sum_{i=0}^{M} \sum_{i_2,i_3,\dots,i_n} q_{i,i_2,\dots,i_n} = 1. \tag{25}$$

Функции частот $\varphi_{\tilde{F}_{\mathbf{X}}}(y)$ и $\varphi_{\tilde{G}_{\mathbf{X}}}(y)$ можно записать в виде

$$\varphi_{\tilde{F}_{\mathbf{X}}}(y) = \sum_{i_1,i_2,\dots i_n:\,(i_1+i_2+\dots+i_n)\delta = ny} p_{i_1,i_2,\dots,i_n},$$

$$\varphi_{\tilde{G}_{\mathbf{X}}}(y) = \sum_{i_1,i_2,\dots i_n:\,(i_1+i_2+\dots+i_n)\delta = ny} q_{i_1,i_2,\dots,i_n}.$$

И, наконец, все переменные должны быть не отрицательными:

$$p_{i_1,i_2,\dots i_n} \ge 0,\ i_1 = 0,1,\dots,M;\ i_2 = 0,1,\dots,M;\dots;\ i_n = 0,1,\dots,M, \tag{26}$$

$$q_{i_1,i_2,\dots i_n} \ge 0,\ i_1 = 0,1,\dots,M;\ i_2 = 0,1,\dots,M;\dots;\ i_n = 0,1,\dots,M. \tag{27}$$

Таким образом, при фиксированном значении y задача (12) сводится к следующей задаче линейного программирования:

$$\max_{\tilde{F}_{\mathbf{X}}} \varphi_{\tilde{F}_{\mathbf{X}}}(y) = \max_{p_{i_1,i_2,\dots,i_n}} \left[\sum_{i_1,i_2,\dots i_n:\,(i_1+i_2+\dots+i_n)\delta = ny} p_{i_1,i_2,\dots,i_n} \right], \tag{28}$$

при линейных ограничениях (17), (19), (24), (26); задача (13) сводится к задаче максимизации (28) при линейных ограничениях (17), (20), (24), (26); задача (14) сводится к задаче максимизации (28) при линейных ограничениях (17), (21), (24), (26).

Аналогично, при фиксированном значении y задача (15) сводится к следующей задаче линейного программирования:

$$\max_{\tilde{G}_{\mathbf{X}}} \varphi_{\tilde{F}_{\mathbf{X}}}(y) = \max_{q_{i_1,i_2,\dots,i_n}} \left[\sum_{i_1,i_2,\dots i_n:\,(i_1+i_2+\dots+i_n)\delta = ny} q_{i_1,i_2,\dots,i_n} \right], \tag{29}$$

при линейных ограничениях (18), (22), (25), (27); задача (16) сводится к задаче максимизации (29) при линейных ограничениях (18), (23), (25), (27).

Заключение. В результате проведенных исследований задача построения критерия Неймана–Пирсона для проверки сложной непараметрической гипотезы об одном среднем была сведена к задачам линейного программирования с сильно разряженными матрицами.

О.М. Голоднiков

ПРО НЕПАРАМЕТРИЧНИЙ КРИТЕРІЙ НЕЙМАНА–ПІРСОНА ДЛЯ ПЕРЕВІРКИ СКЛАДНОЇ ГІПОТЕЗИ ПРО ОДНЕ СЕРЕДНЄ

Розглядається задача побудови критерію Неймана–Пірсона для перевірки складної непараметричної гіпотези стосовно одного середнього. Показано, що ця задача зводиться до задачі лінійного програмування. Оптимальний критерій, який при цьому отримується, є загальним тому, що при його побудові не вимагається незалежність спостережень.

A.N. Golodnikov

ABOUT NON–PARAMETRIC NEYMAN–PEARSON'S ONE SAMPLE MEAN TEST

The Neyman–Pearson's problem of developing composit one sample mean test is considered. It is shown that this problem is reduced to the linear programming. The resulting test is general enough, since it does not require knowledge of underlying distribution, nor independence of observations.

1. *Neyman J., Pearson E.* On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part I // Biometrika. –1928. – 20A. – P. 175 – 240.
2. *Neyman J., Pearson E.* On the Problem of the Most Efficient Tests of Statistical Hypotheses // Philosophical Transactions of the Royal Society. –1933. – Ser. A. –231. – P. 289 – 337.
3. *Нейман Ю.* Вводный курс теории вероятностей и математической статистики. – М.: Наука, 1968. – 448 с.

Получено 15.05.2006