УДК 519.24 + 621.391

НОВЫЙ НЕПАРАМЕТРИЧЕСКИЙ СТАТИСТИЧЕСКИЙ КРИТЕРИЙ ДЛЯ ЗАДАЧ С ТРЕМЯ ВЫБОРКАМИ, БОЛЕЕ ЭФФЕКТИВНЫЙ, ЧЕМ КРИТЕРИЙ УИТНИ*

Г. И. Салов

Институт вычислительной математики и математической геофизики СО РАН,
630090, г. Новосибирск, просп. Академика Лаврентьева, 6
E-mail: sgi@ooi.sscc.ru

Представлен новый непараметрический статистический критерий для проверки гипотезы однородности трёх выборок против альтернативной гипотезы, состоящей в том, что случайные величины одной из этих выборок имеют тенденцию быть стохастически больше случайных величин каждой из двух других выборок в отдельности. Известный критерий Уитни эквивалентен частному случаю нового критерия. Рассмотрено использование этих критериев в реальной трудной задаче обнаружения заданного протяжённого объекта на зашумлённом изображении в ситуации с возможным появлением некоторого другого более крупного «мешающего» объекта.

Ключевые слова: три выборки, критерий однородности, непараметрический критерий, обнаружение объектов, зашумлённое изображение.

Введение. Предлагаемая работа является дополнением к [1–3], но может читаться и независимо от них. Усложним (обобщим) пример задачи обнаружения объекта, приведённый в работе [1].

Пример реальной трудной задачи. Пусть, как и прежде, наблюдается зашумлённое изображение. Предположим теперь, что в отличие от примера в [1] в поле зрения наблюдателя может попасть: 1) либо весь заданный важный протяжённый объект, 2) либо часть окрестности контура некоторого более крупного, но не представляющего никакого интереса («мешающего») объекта, 3) либо может не попасть ни тот ни другой. Для наблюдателя важен именно заданный объект. Появление же мешающего объекта может привести к наиболее нежелательной ложной (ННЛ) «тревоге». Это необходимо учесть.

Пусть $X_1, \ldots, X_m$ — совокупность независимых результатов наблюдений (измерений), полученных в m «точках» области возможного положения важного объекта, и пусть с целью обнаружения такого объекта в случае его присутствия по обе стороны от этой области (симметрично относительно наибольшей средней линии области) берутся ещё две совокупности независимых наблюдений $Y_1, \ldots, Y_n$ и $Z_1, \ldots, Z_n$. Пусть, кроме того, наблюдателю из «физического» смысла и/или прошлого опыта известно, что если в течение наблюдений ни важного для него, ни мешающего объекта в поле зрения не было (гипотеза H_0), то $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$, $Z_1, \ldots, Z_n$ можно рассматривать как стохастически независимые случайные величины с одной и той же непрерывной функцией распределения вероятностей, предположим $F(x)$ («обычная» однородность). Если в течение наблюдений в поле зрения находился только весь важный для наблюдателя объект (гипотеза H_1), то каждая из величин X_i будет иметь уже другую непрерывную функцию распределения вероятностей, скажем $G \le F$, $G \not\equiv F$, т. е. величины $X_1, \ldots, X_m$ будут стохастически больше как величин $Y_1, \ldots, Y_n$, так и $Z_1, \ldots, Z_n$ по отдельности. Наконец, если в поле зре-

*Работа выполнена при поддержке Президиума РАН (программа № 32) и Российского фонда фундаментальных исследований (грант № 13-07-00068).

ния находилась часть окрестности контура мешающего более крупного объекта, то наряду с величинами X_i ещё и каждая из величин Y_j (или Z_j) будет иметь другую функцию распределения, пусть для простоты она совпадает с функцией $G(x)$. Как и в [1], наблюдателю относительно функций $F(x)$ и $G(x)$ известно только то, что они непрерывные (типичный на практике случай).

Проблема состоит в том, чтобы по трём независимым совокупностям (выборкам) $X_1, \ldots, X_m, Y_1, \ldots, Y_n, Z_1, \ldots, Z_n$ распознать (обнаружить) тот случай, когда присутствует именно важный объект. Требуется указать тест (критерий), который приводил бы к желаемому результату с максимальной (или близкой к ней) вероятностью и в то же время был бы по возможности слабо чувствительным к случаю попадания в поле зрения мешающего объекта. Наличие последнего требования противоположного характера делает эту задачу весьма трудной. Среди известных критериев подходящим в рассматриваемой задаче является широко применимый непараметрический статистический критерий, предложенный Уитни в [4].

Цель данной работы — указать критерий, имеющий часто бо́льшую вероятность обнаружения при не большей вероятности ННЛ-тревоги, т. е. более эффективный, чем критерий Уитни.

Новый критерий. Критерий Уитни (для краткости обозначим его через Wh) основан на статистиках U_1 и U_2 двух критериев Манна — Уитни (MW) [5] и предписывает отклонять гипотезу H_0 об однородности в пользу H_1, когда одновременно

$$U_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{I}\{X_i > Y_j\} > C, \qquad U_2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{I}\{X_i > Z_j\} > C. \tag{1}$$

Здесь и далее $\mathbf{I}\{A\}$ обозначает функцию-индикатор события A равную 1, если событие A произошло, и 0 в противном случае, число C подбирается по заранее заданному уровню значимости критерия.

В [1] для задач с двумя выборками были введены новые статистики и основанный на них новый непараметрический критерий, обозначенный через S_E. Результаты исследований в [1, 2] дают основания считать, что S_E-критерий может быть более мощным, чем MW-критерий, в широком ряде случаев. Поэтому интересно выяснить, что получится, если в критерии Уитни два MW-критерия заменить двумя S_E-критериями. Не желательно, чтобы с увеличением мощности критерия (вероятности обнаружения) возросла и вероятность ННЛ-тревоги.

Для этого возьмём $n = 2\nu$ чётным и введём следующие события ($i = 1, \ldots, m$, $j = 1, \ldots, \nu$):

$$E^+_{1ij} = \{X_i > \max(Y_j, Y_{\nu+j})\}, \quad E^-_{1ij} = \{X_i < \min(Y_j, Y_{\nu+j})\}, \quad E^0_{1ij} = \bar{E}^+_{1ij} \cap \bar{E}^-_{1ij},$$

$$E^+_{2ij} = \{X_i > \max(Z_j, Z_{\nu+j})\}, \quad E^-_{2ij} = \{X_i < \min(Z_j, Z_{\nu+j})\}, \quad E^0_{2ij} = \bar{E}^+_{2ij} \cap \bar{E}^-_{2ij},$$

а также считающие их количества статистики ($q = 1, 2$)

$$S^+_{Eq} = \sum_{i=1}^{m} \sum_{j=1}^{\nu} \mathbf{I}\{E^+_{qij}\}; \qquad S^-_{Eq} = \sum_{i=1}^{m} \sum_{j=1}^{\nu} \mathbf{I}\{E^-_{qij}\}; \qquad S^0_{Eq} = \sum_{i=1}^{m} \sum_{j=1}^{\nu} \mathbf{I}\{E^0_{qij}\}, \tag{2}$$

принимающие значения от 0 до $m\nu$ с суммой $S^+_{Eq} + S^-_{Eq} + S^0_{Eq} = m\nu$.

Используя статистики (2), получим новый критерий

$$S^+_{Eq} > h(S^0_{Eq}), \quad q = 1, 2. \tag{3}$$

Для краткости обозначим его через $\mathbf{S}_E$.

Лемма. Критерий Уитни эквивалентен частному случаю $\mathbf{S}_E$-критерия (3), когда $h(z)$ — линейная убывающая функция вида $2h(z) \equiv C - z$, $z = 0, 1, \ldots, m\nu$, и критерию, отклоняющему гипотезу H_0, когда $S^+_{Eq} - S^-_{Eq} > C - m\nu$, $q = 1, 2$, где C — число, входящее в критерий Уитни (1).

Доказательство непосредственно следует из равенств (эквивалентности) событий

$$\{S^+_{Eq} > (C - S^0_{Eq})/2\} = \{2S^+_{Eq} + S^0_{Eq} > C\} = \{U_q > C\}.$$

Ясно, что при редукции пары статистик (S^+_{Eq}, S^-_{Eq}) к простой разности $S^+_{Eq} - S^-_{Eq}$ возможна потеря некоторой информации о выборках.

Характеристики критериев Уитни и $\mathbf{S}_E$. Для вычисления уровня значимости критерия Уитни можно воспользоваться, например, следующим предложением 1.

Обозначим через $\mathfrak{P}(u)$ множество всех упорядоченных $(m+1)$-разбиений $\mathfrak{p} = \mathfrak{p}(n) = (n_0, n_1, \ldots, n_m)$ числа n (т. е. $n = n_0 + n_1 + \ldots + n_m$, где каждое целое число $n_i \geq 0$ и порядок следования чисел n_i существен), для которых

$$\sum_{i=0}^{m} (m - i) n_i = u. \tag{4}$$

Пусть $\mathfrak{p}' = (n'_0, n'_1, \ldots, n'_m) \in \mathfrak{P}(u_1)$ и $\mathfrak{p}'' = (n''_0, n''_1, \ldots, n''_m) \in \mathfrak{P}(u_2)$ — два таких $(m+1)$-разбиения числа n.

Удобно положить $0! = 1$.

Предложение 1 [6]. Пусть U_1 и U_2 — статистики из (1). Тогда при гипотезе H_0 вероятность

$$\mathbf{P}\{U_1 = u_1, U_2 = u_2 \,|\, H_0\} = \frac{m!(n!)^2}{(m+2n)!} \sum_{\mathfrak{p}' \in \mathfrak{P}(u_1)} \sum_{\mathfrak{p}'' \in \mathfrak{P}(u_2)} \prod_{i=0}^{m} \binom{n'_i + n''_i}{n'_i}.$$

В связи с поставленной задачей обнаружения и с тем, что и критерий Уитни, и предлагаемый новый критерий наиболее чувствительны к сдвигам распределений, интересно сравнить их для семейств распределений, отличающихся в основном лишь сдвигами. К сожалению, найти в явном виде точные формулы для вычисления мощности и вероятности ННЛ-тревоги критериев, как правило, не представляется возможным. Существует лишь несколько отдельных случаев, когда их удается получить. Наиболее просто это сделать в случаях с экспоненциальными (этот случай особенно важен для практических приложений) и прямоугольными распределениями вида

$$H_0\colon F(x) = 1 - \mathrm{e}^{-x} \quad \text{при } x \geq 0, \tag{5}$$

$$H_1\colon G(x) = 1 - \mathrm{e}^{-(x-\theta)} \quad \text{при } x \geq \theta,\ \theta > 0 \tag{6}$$

и соответственно

$$H_0\colon F(x) = x \quad \text{при } 0 \leq x \leq 1, \tag{7}$$

$$H_1\colon G(x) = \frac{x - \theta}{1 - \theta} \quad \text{при } \theta \leq x \leq 1,\ 0 \leq \theta < 1. \tag{8}$$

Для этих распределений имеем $F(x_h) = b_0 + b_1 G(x_h)$,

$$1 - F(x_h) = b_1[1 - G(x_h)];$$
$$F(x_h) - F(x_k) = b_1[G(x_h) - G(x_k)], \quad x_h > x_k \geq \theta, \tag{9}$$

где для (5), (6) $b_0 = 1 - \exp(-\theta)$, а для (7), (8) $b_0 = \theta$, при этом в обоих случаях $b_1 = 1 - b_0$. Данный факт позволяет получить явные точные формулы и для вероятностей ННЛ-тревоги, и для мощности рассматриваемых здесь критериев.

Вероятность ННЛ-тревоги критерия Уитни, т. е. вероятность принятия гипотезы H_1 в том случае (обозначим его через H_1^*), когда на самом деле в поле зрения присутствует лишь часть окрестности контура мешающего объекта, может быть подсчитана с помощью следующего предложения.

Предложение 2. Пусть в случае H_1^* каждая из величин Y_j имеет распределение (5) (или (7)), а каждая из случайных величин X_i и Z_j — распределение (6) (или (8)). Тогда вероятность $\mathbf{P}\{U_1 = u_1, U_2 = u_2 \,|\, H_1^*\}$ с учётом (9) можно представить в виде

$$m!(n!)^2 \sum_{\mathfrak{p}' \in \mathfrak{P}(u_1)} \sum_{\mathfrak{p}'' \in \mathfrak{P}(u_2)} \frac{1}{n_0'! n_0''!} \prod_{i=1}^{m} \binom{n_i' + n_i''}{n_i'} \sum_{r=0}^{n_0'} \binom{n_0'}{r} \frac{(n_0' + n_0'' - r)!}{(m + 2n - r)!} b_0^r b_1^{n-r}. \tag{10}$$

Д о к а з а т е л ь с т в о. Поскольку функции F и G (по предположению) непрерывны, без потери общности можно считать, что величины $X_1, \ldots, X_m$ имеют различные значения. Обозначим через $X_{(1)} < X_{(2)} < \ldots < X_{(m)}$ эти величины, расположенные в порядке возрастания.

Пусть $x_1 < x_2 < \ldots < x_m$. Введём в рассмотрение $m + 1$ открытых промежутков $I_0 = (-\infty, x_1)$, $I_h = (x_h, x_{h+1})$, $h = 1, 2, \ldots, m - 1$, $I_m = (x_m, \infty)$. При наличии в поле зрения мешающего объекта (при H_1^*) в соответствии с предположениями для вероятности p_h попадания случайной величины Y_j в промежуток I_h, $h = 0, 1, \ldots, m$, имеем следующие очевидные равенства:

$$p_0 = F(x_1), \qquad p_h = F(x_{h+1}) - F(x_h), \quad h = 1, 2, \ldots, m-1, \; p_m = 1 - F(x_m),$$

тогда как для вероятности q_h попадания случайной величины Z_j в промежуток I_h, $h = 0, 1, \ldots, m$, будем иметь уже

$$q_0 = G(x_1), \qquad q_h = G(x_{h+1}) - G(x_h), \quad h = 1, 2, \ldots, m-1, \; q_m = 1 - G(x_m).$$

Отсюда в силу предположения о независимости наблюдений вероятность того, что среди всех n наблюдений Y_j ровно n_0' наблюдений, а среди всех n наблюдений Z_j ровно n_0'' наблюдений будут меньше x_1, ровно n_1' и соответственно n_1'' наблюдений будут заключены между x_1 и x_2 и т. д., наконец, среди всех n наблюдений Y_j ровно n_m' наблюдений, а среди всех n наблюдений Z_j ровно n_m'' наблюдений будут больше x_m, равна произведению двух полиномиальных распределений

$$\frac{(n!)^2}{n_0'! n_1'! \ldots n_m'! n_0''! n_1''! \ldots n_m''!} p_0^{n_0'} p_1^{n_1'} \ldots p_m^{n_m'} q_0^{n_0''} q_1^{n_1''} \ldots q_m^{n_m''}.$$

Если при этом $X_{(1)} = x_1, X_{(2)} = x_2, \ldots, X_{(m)} = x_m$, то согласно (1) будем иметь

$$U_1 = \sum_{i=0}^{m} (m - i) n_i', \quad U_2 = \sum_{i=0}^{m} (m - i) n_i''.$$

Следовательно,

$$\mathbf{P}\{U_1 = u_1, U_2 = u_2 \,|\, X_{(1)} = x_1, X_{(2)} = x_2, \ldots, X_{(m)} = x_m; H_1^*\} =$$

$$= (n!)^2 \sum_{\mathfrak{p}' \in \mathfrak{P}(u_1)} \sum_{\mathfrak{p}'' \in \mathfrak{P}(u_2)} \Big(\prod_{k=0}^{m} n_h'! n_h''! \Big)^{-1} [F(x_1)]^{n_0'} [F(x_2) - F(x_1)]^{n_1'} \times$$

$$\times [F(x_3) - F(x_2)]^{n_2'} \ldots [F(x_m) - F(x_{m-1})]^{n_{m-1}'} [1 - F(x_m)]^{n_m'} \times$$

$$\times [G(x_1)]^{n_0''} [G(x_2) - G(x_1)]^{n_1''} \ldots [G(x_m) - G(x_{m-1})]^{n_{m-1}''} [1 - G(x_m)]^{n_m''}.$$

Чтобы получить вероятность $\mathbf{P}\{U_1 = u_1, U_2 = u_2 \,|\, H_1^*\}$, надо проинтегрировать последнее выражение по распределению порядковых статистик $X_{(1)}, \ldots, X_{(m)}$ при H_1^*. Элемент этого распределения равен $m! dG(x_1) \ldots dG(x_m)$ внутри области, где $\theta < x_1 < x_2 < \ldots < x_m < < \infty$, и 0 вне её (см., например, [7]). Отсюда с учётом (9) имеем ($n_i = n_i' + n_i''$, $i = 0, 1, \ldots, m$)

$$\mathbf{P}\{U_1 = u_1, U_2 = u_2 \,|\, H_1^*\} = m!(n!)^2 \sum_{\mathfrak{p}' \in \mathfrak{P}(u_1)} \sum_{\mathfrak{p}'' \in \mathfrak{P}(u_2)} \Big(\prod_{k=0}^{m} n_k'! n_k''! \Big)^{-1} b_1^{n - n_0'} \times$$

$$\times \int_{\theta}^{\infty} [1 - G(x_m)]^{n_m} dG(x_m) \int_{\theta}^{x_m} [G(x_m) - G(x_{m-1})]^{n_{m-1}} dG(x_{m-1}) \times$$

$$\times \int_{\theta}^{x_{m-1}} [G(x_{m-1}) - G(x_{m-2})]^{n_{m-2}} dG(x_{m-2}) \ldots \int_{\theta}^{x_3} [G(x_3) - G(x_{m-2})]^{n_2} dG(x_2) \times$$

$$\times \int_{\theta}^{x_2} [G(x_2) - G(x_1)]^{n_1} [b_0 + b_1 G(x_1)]^{n_0'} [G(x_1)]^{n_0''} dG(x_1).$$

Используя теперь формулу бинома Ньютона и переходя последовательно к новым переменным $y_1, \ldots, y_m$, полагая $y_i = G(x_i)$, $i = 1, \ldots, m$, получаем

$$\mathbf{P}\{U_1 = u_1, U_2 = u_2 \,|\, H_1^*\} = m!(n!)^2 \sum_{\mathfrak{p}' \in \mathfrak{P}(u_1)} \sum_{\mathfrak{p}'' \in \mathfrak{P}(u_2)} \Big(\prod_{k=0}^{m} n_k'! n_k''! \Big)^{-1} \sum_{r=0}^{n_0'} \binom{n_0'}{r} b_0^r b_1^{n-r} \times$$

$$\times \int_{0}^{1} [1 - y_m]^{n_m} dy_m \int_{0}^{y_m} [y_m - y_{m-1}]^{n_{m-1}} dy_{m-1} \int_{0}^{y_{m-1}} [y_{m-1} - y_{m-2}]^{n_{m-2}} dy_{m-2} \times \ldots$$

$$\ldots \times \int\limits_0^{y_3} [y_3 - y_2]^{n_2} dy_2 \int\limits_0^{y_2} [y_2 - y_1]^{n_1} y_1^{n_0 - r} dy_1. \qquad (11)$$

Согласно (3.7) и (3.9) из [1] многократный интеграл в (11) равен

$$\frac{(n_0 - r)! n_1! \ldots n_m!}{(n_0 - r + n_1 + n_2 + \ldots + n_m + m)!} = \frac{(n_0 - r)! n_1! \ldots n_m!}{(m + 2n - r)!}. \qquad (12)$$

Подставляя (12) в (11), приходим к представлению (10). Предложение 2 доказано.

Вероятность ННЛ-тревоги критерия Уитни равна

$$\sum_{u_1 > C} \sum_{u_2 > C} \mathbf{P}\{U_1 = u_1, U_2 = u_2 \,|\, H_1^*\}.$$

Для вычисления мощности критерия Уитни, т. е. вероятности отклонения основной гипотезы H_0, когда верна H_1 (в поле зрения присутствует именно важный объект), может служить следующее предложение.

Предложение 3 [3]. Пусть в случае H_1 каждая из величин Y_j и Z_j имеет распределение (5) (или (7)), а каждая из случайных величин X_i — распределение (6) (или (8)). Тогда вероятность $\mathbf{P}\{U_1 = u, U_2 = u_2 \,|\, H_1\}$ с учётом (9) может быть записана в виде

$$m!(n!)^2 \sum_{\mathfrak{p}' \in \mathfrak{P}(u_1)} \sum_{\mathfrak{p}'' \in \mathfrak{P}(u_2)} \prod_{i=0}^{m} \binom{n_i' + n_i''}{n_i'} \sum_{r=0}^{n_0' + n_0''} \frac{b_0^r b_1^{2n-r}}{(m + 2n - r)! r!}.$$

Перейдём теперь к характеристикам нового $\mathbf{S}_E$-критерия (3). Введём сначала необходимые обозначения.

Пусть величина $\mathfrak{P}$ — множество тех упорядоченных $(m+1)^2$-разбиений $\mathfrak{p} = \mathfrak{p}(\nu)$ числа ν вида

$$\mathfrak{p}(\nu) = (\nu_{00}, \nu_{01}, \ldots, \nu_{0m}, \nu_{10}, \nu_{11}, \ldots, \nu_{1m}, \ldots, \nu_{m(m-1)}, \nu_{mm}),$$

для которых имеет место неравенство $u > h(m\nu - u - v)$, где

$$u = \sum_{h=0}^{m-1} \sum_{k=0}^{m-1} (m - \max(h,k)) \nu_{hk}; \qquad v = \sum_{h=1}^{m} \sum_{k=1}^{m} \min(h,k) \nu_{hk}. \qquad (13)$$

Пусть

$$\mathfrak{p}'(\nu) = (\nu_{00}', \nu_{01}', \ldots, \nu_{0m}', \nu_{10}', \nu_{11}', \ldots, \nu_{1m}', \ldots, \nu_{m(m-1)}, \nu_{mm}') \in \mathfrak{P},$$

$$\mathfrak{p}''(\nu) = (\nu_{00}'', \nu_{01}'', \ldots, \nu_{0m}'', \nu_{10}'', \nu_{11}'', \ldots, \nu_{1m}'', \ldots, \nu_{m(m-1)}'', \nu_{mm}'') \in \mathfrak{P}$$

— два подобных $(m+1)^2$-разбиения числа ν.

Предложение 4 [3]. Уровень значимости $\mathbf{S}_E$-критерия даётся формулой

$$\mathbf{P}\{S_{Eq}^+ > h(S_{Eq}^0),\ q = 1, 2 \,|\, H_0\} = \frac{m!(\nu!)^2}{(m + 2n)!} \sum_{\mathfrak{p}' \in \mathfrak{P}} \sum_{\mathfrak{p}'' \in \mathfrak{P}} \Big(\prod_{k=0}^{m} s_k! \Big) \Big(\prod_{h,k=0}^{m} \nu_{hk}'! \nu_{hk}''! \Big)^{-1}, \qquad (14)$$

где

$$s_k = \sum_{h=0}^{m} (\nu'_{kh} + \nu'_{hk} + \nu''_{kh} + \nu''_{hk}). \tag{15}$$

Обратимся к вероятности ННЛ-тревоги $\mathbf{S}_E$-критерия.

Предложение 5. Пусть выполнены условия предложения 2. Тогда с учётом (9) справедливо равенство

$$\mathbf{P}\{S_q^+ > h(S_q^0),\ q = 1,2 \,|\, H_1^*\} =$$

$$= m!(\nu!)^2 \sum_{\mathfrak{p}' \in \mathfrak{P}} \sum_{\mathfrak{p}'' \in \mathfrak{P}} \Big(\prod_{k=1}^{m} s_k!\Big)\Big(\prod_{h,k=0}^{m} \nu'_{hk}!\nu''_{hk}!\Big)^{-1} \sum_{r=0}^{s'_0} \binom{s'_0}{r} \frac{(s_0 - r)! b_0^r b_1^{n-r}}{(m + 2n - r)!},$$

где s_k, $k = 0, 1, \ldots, m$, как и прежде, определяются формулой (15)

$$s'_0 = \sum_{h=0}^{m} (\nu'_{0h} + \nu'_{h0}).$$

Доказательство. Пусть, как и выше, $x_1 < x_2 < \ldots < x_m$ и вместе с $m+1$ открытыми промежутками $I_0 = (-\infty, x_1)$, $I_k = (x_k, x_{k+1})$, $k = 1, 2, \ldots, m-1$, $I_m = (x_m, \infty)$. Введём в рассмотрение $(m+1)^2$ открытых прямоугольников вида $I_{hk} = I_h \times I_k$, $h, k = 0, 1, \ldots, m$.

В силу предположений (ср. также начало доказательства предложения 2) при гипотезе H_1^* вероятность попадания пары $(Y_j, Y_{\nu+j})$ в прямоугольник I_{hk} равна $p_{hk} = p_h p_k$, тогда как вероятность попадания пары $(Z_j, Z_{\nu+j})$ в прямоугольник I_{hk} равна $q_{hk} = q_h q_k$. Отсюда вероятность того, что из всех ν пар $(Y_j, Y_{\nu+j})$ ровно ν'_{00}, а среди всех ν пар $(Z_j, Z_{\nu+j})$ ровно ν''_{00} пар попадут в прямоугольник I_{00}, а также соответственно ровно ν'_{hk} и ν''_{hk} пар попадут в прямоугольник I_{hk} $(h, k = 0, 1, \ldots, m)$, определяется произведением двух полиномиальных распределений

$$\frac{\nu!}{\nu'_{00}!\nu'_{01}! \ldots \nu'_{0m}!\nu'_{10}!\nu'_{11}! \ldots \nu'_{m(m-1)}!\nu'_{mm}!} p_{00}^{\nu'_{00}} p_{01}^{\nu'_{01}} \ldots p_{0m}^{\nu'_{0m}} p_{10}^{\nu'_{10}} p_{11}^{\nu'_{11}} \ldots p_{m(m-1)}^{\nu'_{m(m-1)}} p_{mm}^{\nu'_{mm}} \times$$

$$\times \frac{\nu!}{\nu''_{00}!\nu''_{01}! \ldots \nu''_{0m}!\nu''_{10}!\nu''_{11}! \ldots \nu''_{m(m-1)}!\nu''_{mm}!} q_{00}^{\nu''_{00}} q_{01}^{\nu''_{01}} \ldots q_{0m}^{\nu''_{0m}} q_{10}^{\nu''_{10}} q_{11}^{\nu''_{11}} \ldots q_{m(m-1)}^{\nu''_{m(m-1)}} q_{mm}^{\nu''_{mm}} =$$

$$= (\nu!)^2 \Big(\prod_{h,k=0}^{m} \nu'_{hk}!\nu''_{hk}!\Big)^{-1} p_0^{s'_0} p_1^{s'_1} \ldots p_m^{s'_m} q_0^{s''_0} q_1^{s''_1} \ldots q_m^{s''_m},$$

где

$$s'_k = \sum_{h=0}^{m} (\nu'_{hk} + \nu'_{kh}), \qquad s''_k = \sum_{h=0}^{m} (\nu''_{hk} + \nu''_{kh}), \quad k = 0, 1, \ldots, m.$$

Отсюда, учитывая (9), имеем ($s_i = s'_i + s''_i$, $i = 0, 1, \ldots, m$)

$$\mathbf{P}\{S_q^+ > h(S_q^0),\ q = 1, 2 \,|\, X_{(1)} = x_1, X_{(2)} = x_2, \ldots, X_{(m)} = x_m;\ H_1^*\} =$$

$$= (\nu!)^2 \sum_{\mathfrak{p}' \in \mathfrak{P}} \sum_{\mathfrak{p}'' \in \mathfrak{P}} \Big(\prod_{h,\,k=0}^{m} \nu'_{hk}!\nu''_{hk}! \Big)^{-1} [F(x_1)]^{s'_0} [F(x_2) - F(x_1)]^{s'_1} \times \ldots$$

$$\ldots \times [F(x_m) - F(x_{m-1})]^{s'_{m-1}} [1 - F(x_m)]^{s'_m} [G(x_1)]^{s''_0} [G(x_2) - G(x_1)]^{s''_1} \times \ldots$$

$$\ldots \times [G(x_m) - G(x_{m-1})]^{s''_{m-1}} [1 - G(x_m)]^{s''_m} =$$

$$= (\nu!)^2 \sum_{\mathfrak{p}' \in \mathfrak{P}} \sum_{\mathfrak{p}'' \in \mathfrak{P}} \Big(\prod_{h,\,k=0}^{m} \nu'_{hk}!\nu''_{hk}! \Big)^{-1} b_1^{n - s'_0} [b_0 + b_1 G(x_1)]^{s'_0} [G(x_1)]^{s''_0} \times$$

$$\times [G(x_2) - G(x_1)]^{s_1} \ldots [G(x_m) - G(x_{m-1})]^{s_{m-1}} [(1 - G(x_m))]^{s_m}. \qquad (16)$$

Чтобы получить вероятность $\mathbf{P}\{S_q^+ > h(S_q^0),\ q = 1, 2 \,|\, H_1^*\}$, надо проинтегрировать выражение (16) по распределению величин $X_{(1)}, \ldots, X_{(m)}$, элемент которого при H_1^*, как и прежде, равен $m!dG(x_1) \ldots dG(x_m)$ внутри области, где $\theta < x_1 < x_2 < \ldots < x_m < \infty$, и 0 вне её. Преобразуем сначала выражение под знаком суммы в (16), опуская произведение и используя формулу бинома Ньютона, к виду

$$\sum_{r=0}^{s'_0} \binom{s'_0}{r} b_0^r b_1^{n-r} [G(x_1)]^{s_0 - r} [G(x_2) - G(x_1)]^{s_1} \times$$

$$\times [G(x_3) - G(x_2)]^{s_2} \ldots [G(x_m) - G(x_{m-1})]^{s_{m-1}} [(1 - G(x_m))]^{s_m}. \qquad (17)$$

Интегрируя теперь (17) по распределению величин $X_{(1)}, \ldots, X_{(m)}$ при H_1^*, последовательно переходя к новым переменным $y_1, \ldots, y_m$ по формуле $y_i = G(x_i)$, $i = 1, \ldots, m$, а также используя (11), левую часть (12) и равенства

$$\sum_{k=0}^{m} s_k = \sum_{k=0}^{m} \sum_{h=0}^{m} (\nu'_{hk} + \nu'_{kh} + \nu''_{hk} + \nu''_{kh}) = 4\nu = 2n,$$

получаем

$$m! \sum_{r=0}^{s'_0} \binom{s'_0}{r} b_0^r b_1^{n-r} \int_0^1 [1 - y_m]^{s_m} dy_m \int_0^{y_m} [y_m - y_{m-1}]^{s_{m-1}} y_{m-1} \times \ldots$$

$$\ldots \times \int_0^{y_3} [y_3 - y_2]^{s_2} dy_2 \int_0^{y_2} [y_2 - y_1]^{s_1} y_1^{s_0 - r} dy_1 = m! \Big(\prod_{k=1}^{m} s_k! \Big) \sum_{r=0}^{s'_0} \binom{s'_0}{r} \frac{(s_0 - r)! b_0^r b_1^{n-r}}{(m + 2n - r)!}.$$

Подставляя этот результат в (16), приходим к требуемому. Предложение 5 доказано.

Приведём, наконец, утверждение о мощности $\mathbf{S}_E$-критерия.

Предложение 6 [3]. Пусть выполнены условия предложения 3. Тогда с учётом (9)

$$\mathbf{P}\{S_q^+ > h(S_q^0),\ q = 1, 2 \,|\, H_1\} =$$

$$= m!(\nu!)^2 \sum_{\mathfrak{p}'(\nu)\in\mathfrak{P}} \sum_{\mathfrak{p}''(\nu)\in\mathfrak{P}} \Big(\prod_{k=1}^{m} s_k!\Big)\Big(\prod_{h,k=0}^{m} \nu'_{hk}!\nu''_{hk}!\Big)^{-1} \sum_{r=0}^{s_0} \binom{s_0}{r} \frac{(s_0-r)!b_0^r b_1^{n-r}}{(m+2n-r)!},$$

где s_k, $k = 0, 1, \ldots, m$, как и выше, определяются формулой (15).

Сравнение характеристик критериев. Числовые примеры. Приведём теперь некоторые числовые результаты, которые в какой-то мере характеризуют рассматриваемые критерии с разных сторон. В [2] в качестве подходящих критических значений $h(z)$, $z = 0, \ldots, m\nu$, критерия $S_{Eq}^+ > h(S_{Eq}^0)$, на основе которого построен $\mathbf{S}_E$-критерий (3), предлагается использовать критические значения, получаемые с помощью известной леммы Неймана — Пирсона для случая с альтернативой, близкой к нулевой гипотезе, вида

$$G = (1-a)F + aF^2. \tag{18}$$

В выборе (18) учитывался тот факт, что при этой альтернативе для достаточно малых $a > 0$ MW-критерий $U_q > C$, как эквивалентный критерию Вилкоксона, является наиболее мощным (имеющим наибольшую вероятность принять гипотезу H_1, когда она верна) среди ранговых критериев (см., например, [8, гл. 6, п. 12, задача 28]). При критических значениях $h(z)$, $z = 0, \ldots, m\nu$, полученных для ряда фиксированных значений a^* параметра a, критерий $S_{Eq}^+ > h(S_{Eq}^0)$ оказался более мощным, чем MW-критерий, в довольно широкой области значений параметра a в (18), а именно для $a > \gamma(m, n)$, где граничное значение γ убывает с ростом m и n (более подробно см. в [2]).

Случай с $m = 5$, $n = 2\nu = 4$. Пусть в (1) $C = 15$. При таком C уровень значимости MW-критерия равен 0,0(952380) — бесконечная периодическая десятичная дробь, а уровень значимости Wh-критерия (1) оказался равным 0,0(257904). В [2] подходящими критическими значениями $h(z)$, $z = 0, 1, \ldots, 10$, S_E-критерия оказались критерии, полученные, в частности, при $a^* = 0{,}53$. Они представлены в табл. 1. При них уровень значимости S_E-критерия равен 0,0(925), а уровень значимости $\mathbf{S}_E$-критерия — 0,0(264698), т. е. несколько больше уровня значимости Wh-критерия. Поэтому лишь для более корректного сравнения мощности критериев была сделана замена $h(0) = 9$, после которой уровень значимости $\mathbf{S}_E$-критерия (3) стал равен 0,0(250268), что уже меньше уровня значимости Wh-критерия.

В табл. 2, 3 приводятся результаты вычислений мощности и вероятности ННЛ-тревоги Wh- и $\mathbf{S}_E$-критериев для различных значений θ в распределениях (6) и (8). Числа в таблицах приведены с точностью, достаточной для сравнения; первая и вторая строки принадлежат Wh-критерию, третья и четвёртая — $\mathbf{S}_E$-критерию. При этом в первой и третьей строках приведены значения вероятности ННЛ-тревоги критериев, а во второй и четвёртой — мощности. Столбец с $\theta = 0{,}0$ содержит уровни значимости критериев.

Таблица 1

Значения $h(z)$, $z = 0, 1, \ldots, 10$, S_E-критерия при $m = 5$, $n = 4$

z	0	1	2	3	4	5	6	7	8	9	10
h	8	7	7	6	5	4	4	3	2	1	0

Таблица 2

Мощности и вероятности ННЛ-тревоги Wh- и $\mathbf{S}_E$-критериев в случае с $m = 5$, $n = 4$ и альтернативой (6)

θ										
0,0	0,01	0,1	0,2	0,5	1,0	1,5	2,0	2,5	3,0	4,0
0,0257	0,02639	0,0317	0,0377	0,0539	0,072	0,0832	0,088	0,091	0,093	0,0945
0,0257	0,02701	0,0404	0,0614	0,1660	0,419	0,6440	0,791	0,878	0,928	0,9749
0,0250	0,02565	0,0313	0,0377	0,0550	0,074	0,0834	0,087	0,088	0,089	0,0898
0,0250	0,02630	0,0410	0,0657	0,1995	0,523	0,7737	0,904	0,962	0,985	0,9979

Таблица 3

Мощности и вероятности ННЛ-тревоги Wh- и $\mathbf{S}_E$-критериев в случае с $m = 5$, $n = 4$ и альтернативой (8)

θ									
0,0	0,01	0,1	0,2	0,3	0,5	0,7	0,8	0,9	0,95
0,0257	0,02639	0,0321	0,0391	0,046	0,062	0,077	0,0847	0,090	0,093
0,0257	0,02701	0,0413	0,0672	0,108	0,259	0,519	0,6827	0,849	0,928
0,0250	0,02565	0,0317	0,0392	0,047	0,063	0,078	0,0846	0,088	0,089
0,0250	0,02631	0,0421	0,0720	0,124	0,321	0,640	0,8113	0,945	0,985

Случай с $m = 7$, $n = 2\nu = 6$. При $C = 30$ уровень значимости MW-критерия равен 0,0903 (с точностью до четырёх цифр после запятой), а Wh-критерия — 0,0249. Подобно предыдущему случаю сначала при $a^* = 0{,}55$ были получены критические значения S_E-критерия, приведённые в табл. 4, с которыми его уровень значимости оказался равным 0,0898.

Чтобы при сравнении критериев уровень значимости $\mathbf{S}_E$-критерия был меньше уровня значимости Wh-критерия, проведена замена $h(14) = 7$.

В табл. 5 и 6 сравниваются мощности и вероятности ННЛ-тревоги критериев для различных значений θ в (6) и (8) в этом случае. Как и выше, первая и вторая строки принадлежат Wh-критерию, а третья и четвёртая — $\mathbf{S}_E$-критерию. В первой и третьей строках указаны значения вероятности ННЛ-тревоги критериев, во второй и четвёртой — мощности.

Из сопоставления значений, приведённых в таблицах, достаточно хорошо видно, что при удалении θ от $\theta_0 = 0$ в (6) и (8) мощность и вероятность ННЛ-тревоги $\mathbf{S}_E$-критерия достаточно быстро растут и вскоре становятся больше, чем у Wh-критерия, точнее, что касается вероятности ННЛ-тревоги, то лишь незначительно больше. Но это происходит только при не очень больших значениях θ, где значения мощности и вероятности ННЛ-

Таблица 4

Значения $h(z)$, $z = 0, 1, \ldots, 21$, S_E-критерия при $m = 7$, $n = 6$

z	0	1	2	3	4	5	6	7	8	9	10
h	15	15	14	14	13	13	12	11	11	10	9
z	11	12	13	14	15	16	17	18	19	20	21
h	9	8	7	6	6	5	4	3	2	1	0

Таблица 5

Мощности и вероятности ННЛ-тревоги Wh- и $\mathbf{S}_E$-критериев в случае с $m = 7$, $n = 6$ и альтернативой (6)

θ									
0,0	0,01	0,1	0,2	0,5	1,0	1,5	2,0	2,5	3,0
0,0249	0,0257	0,0327	0,0403	0,0600	0,0789	0,086	0,088	0,0898	0,0901
0,0249	0,0265	0,0447	0,0750	0,2355	0,5918	0,828	0,935	0,9763	0,9914
0,0244	0,0252	0,0325	0,0404	0,0604	0,0783	0,084	0,086	0,0874	0,0876
0,0244	0,0260	0,0455	0,0793	0,2618	0,6433	0,867	0,955	0,9852	0,9950

Таблица 6

Мощности и вероятности ННЛ-тревоги Wh- и $\mathbf{S}_E$-критериев в случае с $m = 7$, $n = 6$ и альтернативой (8)

θ									
0,0	0,01	0,1	0,2	0,3	0,5	0,6	0,8	0,9	0,95
0,0249	0,02575	0,0331	0,0420	0,0514	0,0693	0,0768	0,087	0,0896	0,0901
0,0249	0,02656	0,0460	0,0837	0,1467	0,3752	0,5366	0,860	0,9647	0,9913
0,0244	0,02522	0,0329	0,0392	0,0517	0,0694	0,0764	0,085	0,0872	0,0876
0,0244	0,02607	0,0470	0,0892	0,1610	0,4164	0,5872	0,895	0,9771	0,9949

тревоги критериев являются настолько малыми, что не представляют практического интереса; наблюдатели (пользователи) «обычно не интересуются мощностью критерия при альтернативах с θ, близких к θ_0, но желают контролировать вероятность обнаружения альтернатив, достаточно удалённых от θ_0» [8, с. 145].

При дальнейшем же удалении θ от θ_0 скорость роста вероятности ННЛ-тревоги $\mathbf{S}_E$-критерия, как это видно из табл. 2, 3, 5, 6, заметно снижается. В результате эта вероятность при $\theta > \theta^*(m, n)$ становится уже меньше вероятности ННЛ-тревоги Wh-критерия. Следует отметить также, что граничное значение $\theta^* = \theta^*(m, n)$ убывает с ростом m и n.

Заключение. В данной работе рассматривается новый непараметрический статистический критерий для задач с тремя выборками. Сравнение нового критерия с критерием Уитни в случаях с экспоненциальными и равномерными альтернативными распределениями показывает, что мощность (в задаче обнаружения — это вероятность обнаружения) нового критерия заметно больше, чем мощность критерия Уитни. При этом вероятность наиболее нежелательной «ложной тревоги», когда в поле зрения наблюдателя вместо заданного важного объекта попадает часть окрестности контура более крупного мешающего объекта, не только не оказалась заметно больше аналогичной вероятности критерия Уитни, а наоборот, установилась даже несколько меньше её (неожиданный замечательный факт). Так как непараметрические критерии сравнительно не очень чувствительны к отклонениям от использованных при их построении конкретных альтернативных распределений, то можно надеяться, что эти два преимущества нового критерия будут иметь место в более широком ряде случаев, а не только в проверенных с экспоненциальными и равномерными распределениями. Это чрезвычайно важно для многих приложений при самых различных условиях.

СПИСОК ЛИТЕРАТУРЫ

1. **Салов Г. И.** Новый статистический критерий для задач с двумя и тремя выборками, более мощный, чем критерии Вилкоксона и Уитни // Автометрия. 2011. **47**, № 4. С. 58–70.
2. **Салов Г. И.** О мощности одного нового статистического критерия и двухвыборочного критерия Вилкоксона // Автометрия. 2014. **50**, № 1. С. 44–59.
3. **Салов Г. И.** Новый непараметрический статистический критерий для задач с тремя выборками, частный случай которого эквивалентен критерию Уитни // Сибирский журнал вычислительной математики. 2014. **17**, № 4. С. 389–397.
4. **Whitney D. R.** A bivariate extension of the U statistic // Ann. Math. Stat. 1951. **22**, N 2. P. 274–282.
5. **Mann H. B., Whitney D. R.** On a test of whether one of two random variables is stochastically larger than the other // Ann. Math. Stat. 1947. **18**, N 1. P. 50–60.
6. **Салов Г. И.** О мощности непараметрических критериев для обнаружения протяженных объектов на случайом фоне // Автометрия. 1997. № 3. С. 60–75.
7. **Wilks S. S.** Order statistics // Bull. Amer. Math. Soc. 1948. **54**, N 1. P. 5–50.
8. **Леман Э.** Проверка статистических гипотез. М.: Наука, 1964. 498 с.

Поступила в редакцию 17 апреля 2014 г.