УДК 519.24

О применении критериев проверки однородности средних

Б. Ю. Лемешко, С. Б. Лемешко, И. В. Веретельникова, А. Ю. Новикова[1]

Рассмотрены предпосылки применения параметрических и непараметрических критериев проверки гипотез о равенстве математических ожиданий. Методами статистического моделирования исследована сходимость распределений статистик непараметрических критериев к асимптотическим. Проведен сравнительный анализ мощности параметрических и непараметрических критериев. Даны рекомендации по применению критериев.

Ключевые слова: проверка гипотез, однородность средних, параметрические критерии, непараметрические критерии, распределения статистик, статистическое моделирование, мощность критерия.

1. Введение

К критериям проверки гипотез об однородности математических ожиданий (об однородности средних) прибегают при контроле средств измерений, при статистическом анализе результатов экспериментов, при статистическом управлении качеством для проверки наличия возмущения в ходе процесса. Эти критерии часто используются при анализе результатов биомедицинских исследований.

В общем случае проверяемая гипотеза о равенстве математических ожиданий, соответствующих k выборкам, имеет вид

$$H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k$$

при конкурирующей гипотезе

$$H_1 \colon \mu_{i_1} \neq \mu_{i_2},$$

где неравенство выполняется хотя бы для одной пары индексов i_1, i_2.

Для проверки гипотезы H_0 может использоваться ряд параметрических и непараметрических критериев.

Основным предположением, обуславливающим возможность применения параметрических критериев, является принадлежность анализируемых выборок нормальному закону.

Непараметрические критерии свободны от этого требования. Однако мало кто обращает внимание на то, что возможность использования асимптотических результатов, касающихся непараметрических критериев однородности средних, обусловлена выполнением предположения об однородности законов сравниваемых выборок.

С одной стороны, давно известно, что параметрические критерии однородности средних достаточно устойчивы к нарушению предположения о нормальности. Это мнение в существенной степени опирается на опыт использования критериев. С другой стороны, иногда

[1] Исследования выполнены при поддержке Министерства образования и науки РФ в рамках государственной работы «Обеспечение проведения научных исследований» (№ 1.4574.2017/6.7) и проектной части государственного задания (№ 1.1009.2017/4.6).

звучат достаточно резкие критические замечания, указывающие на возможную некорректность выводов, в связи с тем, что исследователи при использовании параметрических критериев (вследствие объективных и субъективных причин) пренебрегают проверкой нормальности наблюдений. Нередко такая критика в силу «отсутствия в природе» нормального закона сопровождается требованием использовать только непараметрические критерии.

Вместе с тем применение непараметрических критериев имеет свои нюансы. Например, при ограниченных объёмах выборок дискретные распределения их статистик могут существенно отличаться от непрерывных асимптотических. Как правило, непараметрические критерии уступают в мощности параметрическим, аналогами которых они являются.

В настоящей статье, которую можно рассматривать как развитие [1, 2], делается попытка ответить на вопрос, насколько важно убедиться в принадлежности анализируемых выборок нормальному закону при использовании параметрических критериев проверки однородности средних. Приводятся результаты сходимости распределений статистик непараметрических критериев к асимптотическим, приводятся результаты сравнительного анализа мощности наиболее популярных критериев проверки однородности средних.

2. Параметрические критерии

Применение критерия сравнения двух выборочных средних при **известных дисперсиях** предусматривает вычисление статистики

$$z = \left(\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2\right) \Big/ \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\,, \tag{1}$$

где $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, n_i – объем i-й выборки, $i = 1, 2$. В случае принадлежности анализируемых выборок нормальным законам статистика z подчиняется стандартному нормальному закону. Если в статистике (1) опустить μ_1 и μ_2, то будет проверяться гипотеза о равенстве математических ожиданий. Критерий двусторонний, проверяемая гипотеза отклоняется при больших по модулю значениях статистики (1).

Применение **критерия Стьюдента** обусловлено предположением о равенстве неизвестных дисперсий. Статистика критерия имеет вид [3]:

$$t = \left(\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2\right) \Big/ \sqrt{\left[\frac{n_1 + n_2}{n_1 n_2}\right]\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right]}\,, \tag{2}$$

где

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2 .$$

При проверке гипотезы о равенстве математических ожиданий в статистике (2) опускаются μ_1 и μ_2. При справедливости гипотезы H_0 и принадлежности выборок нормальному закону статистика (2) подчиняется t_ν-распределению Стьюдента с числом степеней свободы $\nu = n_1 + n_2 - 2$.

При неравных объемах выборок $n_1 \neq n_2$ статистика **критерия сравнения двух выборочных средних при неизвестных и неравных дисперсиях**, предложенная в [4], имеет вид [3]

$$t = \left(\bar{x}_1 - \bar{x}_2 - \mu_1 + \mu_2\right) \Big/ \sqrt{s_1^2/n_1 + s_2^2/n_2}\,. \tag{3}$$

Критерий также двусторонний. В случае нормального закона и справедливости гипотезы H_0 статистика (3) подчиняется t_ν- распределению Стьюдента с числом степеней свободы

$$\nu = \left(s_1^2/n_1 + s_2^2/n_2\right)^2 \Bigg/ \left[\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}\right]. \quad (4)$$

Задача, связанная с поиском распределения статистики (3), долгое время носила название проблемы Беренса–Фишера [5–7]. Её настоящее решение получено в работах [8–10].

В случае равенства неизвестных дисперсий статистика (3) эквивалентна статистике (2). При неравенстве дисперсий всегда $\nu < n_1 + n_2 - 2$. Чем больше разница в дисперсиях, соответствующих анализируемым выборкам, тем сильнее распределение статистики (3) отличается от распределения статистики (2).

Распределения статистик (1) и (2) при справедливости H_0 известны точно, а решение проблемы Беренса–Фишера носит приближенный характер. В то же время это решение обладает хорошей точностью.

Отметим, что при $n_1 + n_2 > 200$ различие между критериями со статистиками (1) – (3) практически исчезает, а соответствующие распределения Стьюдента практически не отличаются от стандартного нормального закона.

F-критерий может применяться для проверки гипотезы об однородности математических ожиданий по k выборкам объемом n в случае справедливости предположения о постоянстве (о равенстве) дисперсий [11]. Статистика критерия задаётся выражением

$$F = \frac{Q_1/(k-1)}{Q_2/[k(n-1)]}, \quad (5)$$

где

$$Q_1 = n\sum_{i=1}^{k}\left(\bar{x}_i - \bar{\bar{x}}\right)^2, \quad Q_2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2, \quad \bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad \bar{\bar{x}} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}.$$

Например, в задачах контроля качества Q_1 является мерой различия в уровнях настройки между k выборками, в то время как Q_2 определяет различие в уровнях настройки внутри этих k выборок.

Если все выборки извлекаются из нормальной генеральной совокупности, то при справедливости гипотезы H_0 статистка (5) подчиняется F_{ν_1,ν_2}-распределению Фишера со степенями свободы $\nu_1 = k - 1$ и $\nu_2 = k(n-1)$ [11]. Критерий правосторонний. Проверяемая гипотеза отклоняется при больших значениях статистики (5).

3. Об устойчивости параметрических критериев

При исследовании степени устойчивости параметрических критериев в данном случае использовалась та же методика компьютерного моделирования и анализа статистических закономерностей, что и в других наших работах, например, при подготовке рекомендаций [12, 13]. Распределения статистик (1) – (3) и (5) при справедливой проверяемой гипотезе H_0 исследовались для различных законов, в частности, в случае принадлежности наблюдений обобщённому нормальному закону с плотностью

$$f(x) = \frac{\theta_2}{2\theta_1\Gamma(1/\theta_2)}\exp\left\{-\left(\frac{|x-\theta_0|}{\theta_1}\right)^{\theta_2}\right\} \quad (6)$$

с различными значениями параметра формы θ_2. При $\theta_2 = 2$ выражение (6) дает плотность нормального закона распределения. При больших значениях θ_2 распределение (6) стремится к равномерному, при малых θ_2 получаем симметричные законы с «тяжелыми хвостами».

Например, на рис. 1 показаны полученные в результате моделирования распределения статистики (1) в случае принадлежности наблюдаемых величин законам распределения семейства (6) при различных значениях параметра формы, а также при показательном законе с плотностью

$$f(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right).$$

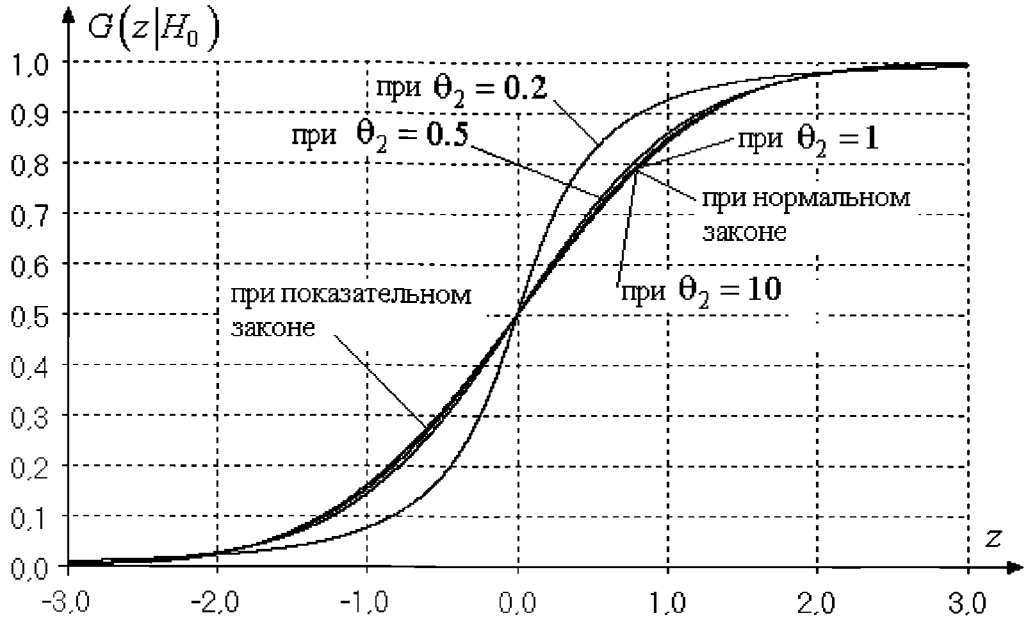

Рис. 1. Эмпирические распределения статистики (1) при различных законах распределения наблюдаемых величин и справедливости гипотезы H_0 при объемах выборок $n_1 = n_2 = 10$

Конечно, распределение статистики (1) зависит от законов, которым принадлежат анализируемые выборки. Асимметрия наблюдаемых законов приводит к отличию распределения статистики от стандартного нормального, однако это отличие не настолько велико, чтобы приводить к серьезным ошибкам при использовании критерия. В случае симметричных законов наблюдается устойчивость распределения статистики к значительным отклонениям наблюдаемых законов от нормального (вплоть до равномерного): распределения статистики существенно отличаются от стандартного нормального только в случае законов с «тяжелыми хвостами» (например, при распределении Коши или в случае закона (6) при малых значениях θ_2 (см. рис. 1, при $\theta_2 = 0.5$ и $\theta_2 = 0.2$).

С ростом объемов выборок устойчивость критерия к отклонениям наблюдаемых законов от нормального повышается.

Результаты исследования критериев со статистиками (2) и (3) позволяют сделать аналогичные выводы о характере зависимости распределений этих статистик от вида наблюдаемых законов.

Распределения статистики (4) F-критерия также устойчивы к отклонениям законов, соответствующих анализируемым выборкам, от нормального. Однако следует подчеркнуть, что применение данного критерия для проверки однородности средних предполагает примерное равенство дисперсий анализируемых выборок. При невыполнении данного условия распределения статистики $G(F|H_0)$ становятся отличными от соответствующего F_{ν_1,ν_2}-распределения. Если отношение максимальной дисперсии к минимальной соответствующих анализируемых выборок не превышает 4, то отклонение распределения статистики $G(F|H_0)$ от F_{ν_1,ν_2}-распределения Фишера не превосходит величины 0.01.

Данные результаты подтверждают общую закономерность: параметрические критерии, связанные с проверкой гипотез о математических ожиданиях, весьма устойчивы к отклонениям наблюдаемых законов от нормального. Это справедливо даже в случае многомерных случайных величин [14].

4. Непараметрические критерии однородности средних

Строго говоря, применение непараметрических критериев проверки гипотез о равенстве математических ожиданий предполагает, что анализируемые выборки принадлежат законам, которые могут отличаться разве что параметрами сдвига. И проверка гипотезы об однородности направлена именно на обнаружение возможного сдвига.

Нарушение данного предположения отражается на распределениях статистик критериев, имеющих место при справедливости проверяемой гипотезы H_0 о равенстве математических ожиданий. Однако на этом факте, как правило, не акцентируют внимания, полагая, что заметные изменения распределений статистик (при справедливости H_0) проявляются лишь при существенно (на порядок) отличающихся параметрах масштаба закона, соответствующего анализируемым выборкам. В действительности ситуация не так оптимистична: существенные изменения распределений статистик происходят при значительно меньших отличиях в параметрах масштаба.

Ранговый **критерий Манна и Уитни** [15–18] является непараметрическим аналогом t-критерия для сравнения двух средних значений непрерывных распределений. Критерий Манна и Уитни представляет собой развитие критерия Уилкоксона [19].

Для вычисления статистики Уилкоксона две независимые выборки объединяют в одну объёмом n_1+n_2 и упорядочивают. По объединенной выборке определяют сумму рангов R_1, соответствующую элементам первой выборки, и сумму рангов второй R_2. Статистика критерия Уилкоксона [19] имеет вид

$$U=\min\{U_1,U_2\}, \tag{7}$$

где

$$U_1=n_1n_2+n_1(n_1+1)/2-R_1,$$
$$U_2=n_1n_2+n_2(n_2+1)/2-R_2.$$

Дискретные распределения $G(U|H_0)$ статистики сильно зависят от размера выборок, что затрудняет применение критерия. В критерии Манна–Уитни (Манна–Уитни–Уилкоксона) вместо U-статистики используется нормализованная статистика

$$\tilde{z}=\frac{U-n_1n_2/2}{\sqrt{n_1n_2(n_1+n_2+1)/12}}. \tag{8}$$

Дискретное распределение статистики (8) в случае справедливости проверяемой гипотезы H_0 хорошо приближается стандартным нормальным законом при $n_1+n_2>60$, если объем каждой из выборок не слишком мал: $n_1\ge 8$, $n_2\ge 8$. При меньших объемах выборок следует учитывать, что достигаемый уровень значимости p_{value}, вычисляемый по значению статистики $\tilde{z}$ в соответствии с функцией распределения стандартного нормального закона, может заметно отличаться от истинного.

H-**критерий Краскела–Уаллиса** [20, 21] является развитием U-критерия для проверки гипотезы о равенстве средних по k выборкам. Объединенную выборку объёмом $n=\sum_{i=1}^{k} n_i$

упорядочивают и вычисляют суммы рангов R_i для i-й выборки, $i = \overline{1,k}$. Статистика для проверки гипотезы H_0 имеет вид

$$H = \left[\frac{12}{n(n+1)}\right]\left[\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n+1). \tag{9}$$

Статистика H представляет собой дисперсию ранговых сумм. При больших n_i и k в случае справедливости проверяемой гипотезы H_0 распределение статистики хорошо аппроксимируется χ^2_{k-1}-распределением [21]. В описаниях критерия говорится, что χ^2_{k-1}-распределением практически можно пользоваться при $n_i \geq 5$, $k \geq 4$.

Исследования показали, что на практике в случае $k = 2$ можно пренебречь дискретностью при $n_i \geq 30$ (см. рис. 2).

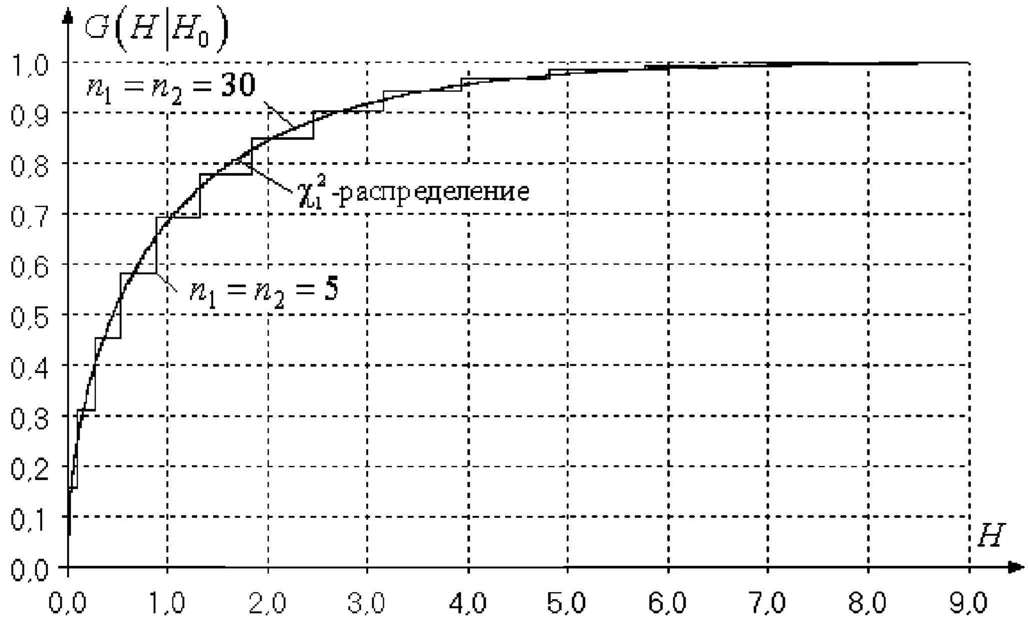

Рис. 2. Сходимость распределения статистики (9) Краскела–Уаллиса к χ^2_{k-1}-распределению при $k = 2$

С ростом числа выборок влияние дискретности быстро убывает. При $k = 3$ распределение статистики достаточно хорошо приближается χ^2_{k-1}-распределением начиная с $n_i \geq 20$, а при $n_i \geq 30$ согласие распределения статистики с χ^2_{k-1}-распределением не отклоняется по всем применяемым критериям согласия [12, 22–24]. При $k \geq 5$ согласие распределения статистики с χ^2_{k-1}-распределением не отклоняется при $n_i \geq 20$.

Статистика непараметрического **критерия Ван дер Вардена** для анализа 2-х выборок вычисляется в соответствии с выражением [25]:

$$V = \sum_{i=1}^{n_2} u_{R_i/(n_1+n_2+1)}, \tag{10}$$

где u_γ – γ-квантиль стандартного нормального закона, R_i, $i = \overline{1,n_2}$ – ранг i-го наблюдения, например, второй выборки (как в выражении (10) для статистики) в общем вариационном ряду объединенной выборки из $n_1 + n_2$ наблюдений.

Считается [26], что при $n_1, n_2 \geq 20$ распределение статистики (10) удовлетворительно описывается нормальным законом с нулевым математическим ожиданием и дисперсией

$$D[V] = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} \sum_{i=1}^{n_1+n_2} u^2_{i/(n_1+n_2+1)}.$$

Нормализованная статистика

$$V^* = \frac{V}{\sqrt{D[V]}} \tag{11}$$

должна подчиняться стандартному нормальному закону.

Критерий двусторонний, проверяемая гипотеза отклоняется при больших по модулю значениях статистики (11).

Исследование распределений статистики (11) показало, что при $n_1 + n_2 \geq 40$ отличием распределения $G(V^* | H_0)$ статистики (11) от стандартного нормального закона можно пренебречь. Величина отклонения не имеет практического значения уже при $n_1 + n_2 \geq 20$. Дискретность распределения статистики хорошо заметна при $n_1 = n_2 = 5$, а при $n_1 = n_2 = 10$ распределение статистики визуально совпадает со стандартным нормальным законом.

Из непараметрических критериев однородности средних критерий Ван дер Вардена, по-видимому, является наиболее предпочтительным.

Критерий Фишера–Йэйтса–Терри–Гёфдинга, рассмотренный в работах [27–29], очень близок к критерию Ван дер Вардена. В качестве меток в критерии выбраны математические ожидания соответствующих порядковых статистик в выборке объёмом $n = n_1 + n_2$ из стандартного нормального закона. Статистика критерия имеет вид:

$$S = \sum_{i=1}^{n_2} u_{(R_i - 3/8)/(n+1/4)}, \tag{12}$$

где u_γ – γ-квантиль стандартного нормального закона, R_i, $i = \overline{1, n_2}$ – ранг i-го наблюдения, например, второй выборки в общем вариационном ряду объединенной выборки из $n_1 + n_2$ наблюдений.

Так же как и в случае критерия Ван дер Вардена, статистика (12) достаточно хорошо описывается нормальным законом с нулевым математическим ожиданием и дисперсией

$$D[S] = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} \sum_{i=1}^{n_1+n_2} u^2_{(i-3/8)/(n_1+n_2+1/4)},$$

а нормализованная статистика

$$S^* = \frac{S}{\sqrt{D[S]}} \tag{13}$$

– стандартным нормальным законом.

По своим свойствам и мощности критерий со статистикой (13) эквивалентен критерию Ван дер Вардена.

Статистика k**-выборочного критерия Ван дер Вардена** для проверки гипотезы о равенстве средних имеет вид

$$T = (n-1) \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{j=1}^{n_i} u_{R_{ij}/(n+1)} \right)^2 \Bigg/ \sum_{i=1}^{n} u^2_{i/(n+1)}, \tag{14}$$

где $n = \sum_{i=1}^{k} n_i$, u_γ – γ-квантиль стандартного нормального закона, R_{ij}, $j = \overline{1, n_i}$ – ранг j-го элемента i-й выборки в вариационном ряду объединённой выборки объёма n.

При справедливости проверяемой гипотезы H_0 статистика (14) хорошо описывается χ^2_{k-1}-распределением. Отклонением распределения статистики от χ^2_{k-1}-распределения можно пренебречь при $n_i > 30$.

При $n_i = 10$ в интервале значений функции распределения статистики от 0.9 до 1 функция χ^2_{k-1}-распределения максимально отклоняется от реального распределения статистики в сторону меньших значений на величину порядка 0.005. То есть на такую величину может быть завышен достигнутый уровень значимости p_{value}, вычисляемый по χ^2_{k-1}-распределению.

Напомним, что распределения статистик непараметрических критериев однородности средних при справедливости проверяемой гипотезы H_0 не зависят от закона, которому принадлежат выборки, если все они принадлежат одному виду закона. Но если при справедливости H_0 это разные законы, то распределение статистик всё-таки меняется (даже в случае принадлежности выборок законам, отличающимся только параметрами масштаба).

В качестве примера на рис. 3 приведены распределения статистики многовыборочного критерия Ван дер Вардена, соответствующие двум ситуациям: в первом случае все анализируемые выборки принадлежали одному и тому же закону, во втором – обобщённому нормальному закону (6) с различными значениями параметра формы θ_2 (0.5, 2 и 6) и одинаковыми параметрами масштаба θ_1.

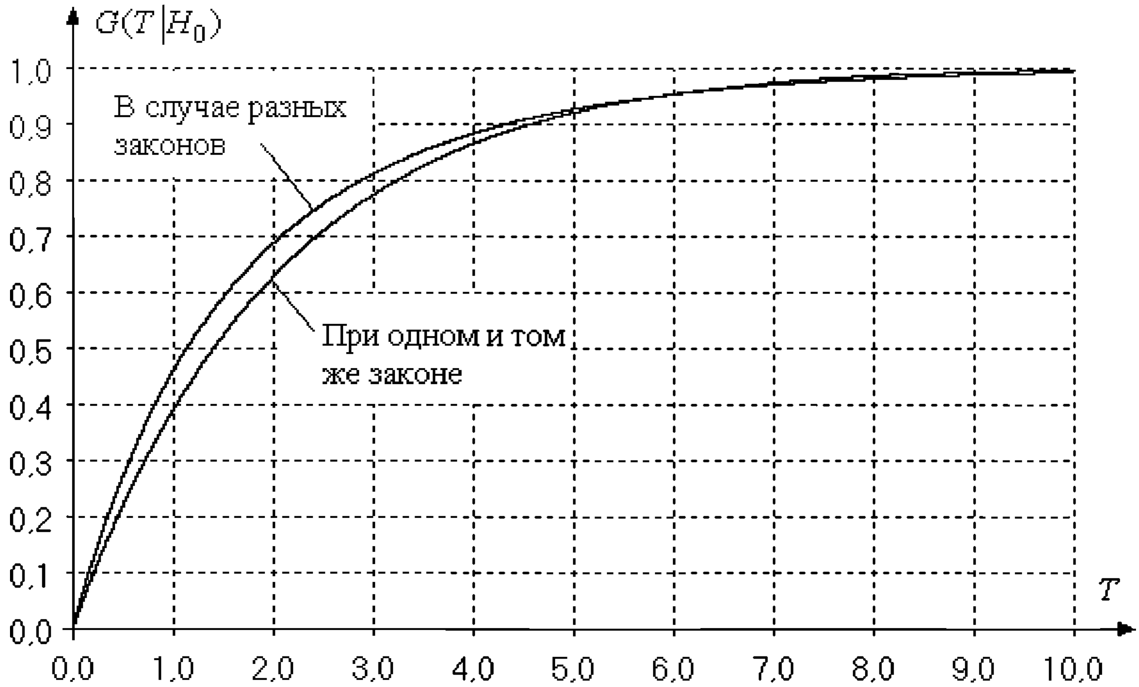

Рис. 3. Эмпирические распределения статистики многовыборочного критерия Ван дер Вардена при $k = 3$ и объемах выборок $n_i = 20$

В подобной ситуации аналогичную зависимость демонстрируют и распределения статистики H-критерия Краскела–Уаллиса и всех двухвыборочных непараметрических критериев однородности средних.

5. Сравнительный анализ мощности критериев

С результатами проверки гипотез связывают ошибки двух видов. Ошибка первого рода заключается в отклонении справедливой гипотезы H_0. Вероятность ошибки первого рода, как правило, обозначают α (уровень значимости). Если справедлива некоторая конкурирующая гипотеза H_1, а в процессе проверки гипотеза H_0 не отклоняется, то происходит ошибка второго рода, вероятность которой обозначают β. Задание вероятности одной из ошибок для пары гипотез H_0 и H_1 определяет вероятность другой. Мощность критерия при заданном уровне значимости α определяется величиной $1-\beta$: чем выше мощность, тем лучше он различает гипотезы H_0 и H_1.

Для вычисления мощности критерия со статистикой S необходимо знание условных распределений статистики $G(S|H_0)$ и $G(S|H_1)$ при справедливости H_0 и H_1 соответственно. Распределения $G(S|H_0)$ статистик непараметрических критериев однородности средних при ограниченных объёмах выборок отличаются от асимптотических, а вид распределений $G(S|H_1)$ параметрических и непараметрических критериев, как правило, неизвестен. Поэтому реально оценить мощность критериев можно только на основании статистического моделирования.

Мощность рассматриваемых критериев исследовалась относительно различных конкурирующих гипотез. В данном случае мощность критериев сравнивалась при одинаковых дисперсиях анализируемых выборок, принадлежащих нормальным законам, относительно трёх конкурирующих гипотез H_1^1: $\mu_2=\mu_1+0.1\sigma$; H_1^2: $\mu_2=\mu_1+0.2\sigma$; H_1^3: $\mu_2=\mu_1+0.5\sigma$.

Естественно, что на результатах сравнительного анализа в определённой степени отразилась существенная дискретность распределений статистик некоторых непараметрических критериев.

Вычисленные на основании результатов моделирования распределений статистик оценки мощности критериев для различных значений уровня значимости α и при одинаковых объемах 2-х сравниваемых выборок ($n_i=n$) представлены в табл. 1 – 3, где критерии упорядочены по убыванию мощности. Выборки распределений статистик моделировались при числе имитационных экспериментов $N=10^6$, что позволило оценивать значения мощности с погрешностью в пределах $\pm10^{-3}$.

Какие выводы можно сделать на основе результатов, представленных в таблицах? Во-первых, очевидно, что параметрические критерии обладают несколько большей мощностью по сравнению с непараметрическими, но это преимущество весьма незначительно.

t-критерий Стьюдента со статистикой (2) слегка уступает в мощности z-критерию со статистикой (1) при известных дисперсиях. В рассматриваемой ситуации равенства дисперсий критерий со статистикой (3) эквивалентен критерию со статистикой (2) и имеет ту же мощность.

F-критерий в случае анализа 2-х выборок эквивалентен критерию со статистикой (3), применяемому при неравенстве неизвестных дисперсий.

Непараметрические критерии Ван дер Вардена практически не уступают в мощности параметрическому F-критерию.

В случае анализа 2-х выборок ранговые $\tilde{z}$-критерий Манна–Уитни и H-критерий Краскела–Уаллиса асимптотически эквивалентны и слегка уступают в мощности критериям Ван дер Вардена и чуть больше – параметрическим критериям.

Некоторый разнобой в оценках мощности $\tilde{z}$-критерия Манна–Уитни и H-критерия Краскела–Уаллиса, отраженный в таблицах, является следствием дискретности распределений этих статистик. Из-за дискретности действительные уровни значимости для этих критериев отличаются от заданных в таблице значений α и несколько превышают их. Этим же объясняется «преимущество» в некоторых случаях H-критерия Краскела–Уаллиса перед параметрическими критериями.

Таблица 1. Мощность критериев относительно $H_1^1 : \mu_2 = \mu_1 + 0.1\sigma$

z-критерий при известных дисперсиях					
α	$n = 10$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
0.1	0.108	0.117	0.126	0.142	0.184
0.05	0.056	0.61	0.068	0.079	0.109
0.01	0.012	0.014	0.016	0.020	0.032
t-критерий Стьюдента при неизвестных и равных дисперсиях					
0.1	0.108	0.117	0.126	0.141	0.183
0.05	0.055	0.061	0.067	0.078	0.108
0.01	0.012	0.013	0.016	0.020	0.031
t-критерий Стьюдента при неизвестных и неравных дисперсиях					
0.1	0.108	0.117	0.126	0.141	0.183
0.05	0.055	0.061	0.067	0.078	0.108
0.01	0.012	0.013	0.016	0.020	0.031
F-критерий					
0.1	0.108	0.116	0.125	0.141	0.183
0.05	0.055	0.060	0.067	0.078	0.108
0.01	0.012	0.013	0.015	0.020	0.031
Критерий Ван дер Вардена					
0.1	0.108	0.116	0.125	0.141	0.183
0.05	0.055	0.061	0.067	0.078	0.108
0.01	0.012	0.013	0.015	0.020	0.031
Многовыборочный критерий Ван дер Вардена					
0.1	0.108	0.116	0.125	0.141	0.182
0.05	0.055	0.060	0.067	0.078	0.107
0.01	0.012	0.013	0.015	0.020	0.031
H-критерий Краскела–Уаллиса					
0.1	0.113	0.118	0.126	0.140	0.179
0.05	0.057	0.063	0.066	0.077	0.105
0.01	0.013	0.014	0.015	0.019	0.030
$\tilde{z}$-критерий Манна–Уитни					
0.1	0.102	0.113	0.124	0.138	0.179
0.05	0.051	0.060	0.066	0.076	0.105
0.01	0.011	0.013	0.015	0.019	0.030

Таблица 2. Мощность критериев относительно $H_1^2 : \mu_2 = \mu_1 + 0.2\sigma$

z-критерий при известных дисперсиях					
α	$n = 10$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
0.1	0.133	0.167	0.200	0.263	0.411
0.05	0.073	0.097	0.122	0.170	0.293
0.01	0.018	0.027	0.036	0.058	0.125
t-критерий Стьюдента при неизвестных и равных дисперсиях					
0.1	0.131	0.166	0.198	0.261	0.408
0.05	0.071	0.095	0.119	0.167	0.291
0.01	0.017	0.025	0.035	0.056	0.122
t-критерий Стьюдента при неизвестных и неравных дисперсиях					
0.1	0.131	0.166	0.198	0.261	0.408
0.05	0.071	0.095	0.119	0.167	0.291
0.01	0.017	0.025	0.035	0.056	0.122
F-критерий					
0.1	0.131	0.165	0.198	0.261	0.408

0.05	0.071	0.094	0.119	0.168	0.290
0.01	0.016	0.025	0.035	0.056	0.121
Критерий Ван дер Вардена					
0.1	0.130	0.163	0.197	0.259	0.407
0.05	0.070	0.093	0.118	0.166	0.289
0.01	0.016	0.025	0.034	0.056	0.120
Многовыборочный критерий Ван дер Вардена					
0.1	0.129	0.163	0.196	0.259	0.406
0.05	0.070	0.093	0.118	0.166	0.288
0.01	0.016	0.025	0.034	0.056	0.120
H-критерий Краскела–Уаллиса					
0.1	0.136	0.164	0.196	0.254	0.395
0.05	0.073	0.096	0.115	0.162	0.279
0.01	0.019	0.025	0.033	0.054	0.114
$\tilde{z}$ -критерий Манна–Уитни					
0.1	0.121	0.158	0.192	0.252	0.395
0.05	0.063	0.091	0.115	0.161	0.279
0.01	0.015	0.023	0.033	0.053	0.114

Таблица 3. Мощность критериев относительно $H_1^3 : \mu_2 = \mu_1 + 0.5\sigma$

z -критерий при известных дисперсиях					
α	$n = 10$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
0.1	0.301	0.477	0.615	0.803	0.971
0.05	0.201	0.354	0.492	0.706	0.942
0.01	0.073	0.162	0.262	0.470	0.833
t -критерий Стьюдента при неизвестных и равных дисперсиях					
0.1	0.287	0.466	0.607	0.799	0.970
0.05	0.185	0.339	0.478	0.696	0.940
0.01	0.060	0.145	0.246	0.455	0.825
t -критерий Стьюдента при неизвестных и неравных дисперсиях					
0.1	0.287	0.466	0.607	0.799	0.970
0.05	0.185	0.339	0.478	0.696	0.940
0.01	0.060	0.145	0.246	0.455	0.825
F-критерий					
0.1	0.287	0.464	0.606	0.799	0.970
0.05	0.184	0.338	0.477	0.697	0.940
0.01	0.060	0.144	0.245	0.453	0.824
Критерий Ван дер Вардена					
0.1	0.280	0.457	0.600	0.795	0.969
0.05	0.179	0.333	0.471	0.691	0.938
0.01	0.058	0.141	0.239	0.449	0.821
Многовыборочный критерий Ван дер Вардена					
0.1	0.280	0.456	0.599	0.794	0.969
0.05	0.179	0.330	0.471	0.691	0.938
0.01	0.058	0.140	0.239	0.448	0.821
H-критерий Краскела–Уаллиса					
0.1	0.286	0.452	0.592	0.781	0.963
0.05	0.184	0.331	0.457	0.679	0.930
0.01	0.064	0.139	0.231	0.434	0.802
$\tilde{z}$ -критерий Манна–Уитни					
0.1	0.259	0.441	0.586	0.779	0.963
0.05	0.162	0.321	0.457	0.674	0.930
0.01	0.053	0.132	0.231	0.431	0.802

В табл. 4 – 5 представлены оценки мощности k-выборочных критериев при $k=3$, 5 относительно конкурирующих гипотез, при которых $k-1$ выборке соответствовало математическое ожидание μ, а k-й выборке – $\mu+0.2\sigma$ и $\mu+0.5\sigma$ (при одинаковых объемах сравниваемых выборок $n_i=n$). Здесь особенно заметно преимущество критерия Ван дер Вардена перед критерием Краскела–Уаллиса.

Таблица 4. Мощность критериев относительно $H_1^2 : \mu_k = \mu_1 + 0.2\sigma$

Критерий	α	$k=3$		$k=5$	
		$n=30$	$n=50$	$n=30$	$n=50$
F-критерий	0.1	0.191	0.254	0.175	0.232
	0.05	0.112	0.160	0.101	0.142
	0.01	0.031	0.052	0.026	0.043
Многовыборочный критерий Ван дер Вардена	0.1	0.190	0.252	0.174	0.231
	0.05	0.111	0.159	0.100	0.141
	0.01	0.031	0.051	0.026	0.043
H-критерий Краскела–Уаллиса	0.1	0.187	0.247	0.172	0.225
	0.05	0.109	0.155	0.098	0.137
	0.01	0.030	0.049	0.025	0.041

Таблица 5. Мощность критериев относительно $H_1^3 : \mu_k = \mu_1 + 0.5\sigma$

Критерий	α	$k=3$		$k=5$	
		$n=30$	$n=50$	$n=30$	$n=50$
F-критерий	0.1	0.618	0.823	0.585	0.807
	0.05	0.488	0.727	0.454	0.705
	0.01	0.252	0.491	0.226	0.468
Многовыборочный критерий Ван дер Вардена	0.1	0.612	0.819	0.581	0.804
	0.05	0.482	0.723	0.449	0.701
	0.01	0.248	0.487	0.223	0.463
H-критерий Краскела–Уаллиса	0.1	0.598	0.806	0.564	0.786
	0.05	0.468	0.706	0.431	0.680
	0.01	0.238	0.467	0.208	0.437

На практике, применяя рассматриваемые критерии для проверки гипотезы об однородности математических ожиданий, обычно задаются только вероятностью α ошибки первого рода. Процедуры контроля предусматривают чаще всего небольшие объемы выборок. Как правило, не заходит речи о вероятности β ошибки второго рода: не отклонить проверяемую гипотезу при справедливости конкурирующей. В то же время в процедуре контроля при задании α желательно гарантировать величину $\beta \le \alpha$.

В данном случае по результатам анализа мощности мы видим, что при конкурирующей гипотезе H_1^1 для $\alpha=0.1$ и объёмах выборок $n=100$ вероятность ошибки второго рода составит величину $\beta=1-0.184=0.816$ для z-критерия со статистикой (1). При $\alpha=0.1$ и объемах выборок $n=100$ данный критерий обеспечит величину $\beta=0.119$ для более далекой альтернативы $\mu_1+0.4\sigma$ и величину $\beta=0.029<0.1$ относительно H_1^3. А чтобы с заданным качеством различать гипотезы H_0 и H_1^1, необходимо иметь выборки объемом порядка 1625 наблюдений!

Какие конкурирующие гипотезы можно с таким же качеством (при $\alpha \le 0.1$ и $\beta \le 0.1$) различать по выборкам объемом $n=10$? Альтернативы, когда μ_2 отличается от μ_1 на величину не менее чем 1.28σ! При $n=20$, когда μ_2 отличается от μ_1 на величину примерно 0.9σ, при $n=30$ – примерно 0.74σ, при $n=50$ – приблизительно 0.555σ, при $n=100$ – на 0.41σ.

6. Заключение

Показанная устойчивость параметрических критериев проверки однородности математических ожиданий подчеркивает возможность использования классических результатов в условиях нарушения стандартного предположения о нормальности. Если закон (законы) распределения анализируемых выборок отличается от нормального, но нет оснований полагать, что наблюдаемые величины принадлежат законам с «тяжелыми хвостами», применение параметрических критериев со статистиками (1) – (3) будет приводить к корректным выводам. По крайней мере, погрешности при оценке достигнутого уровня значимости p_{value} будут невелики.

Если дисперсии анализируемых выборок неизвестны и, возможно, различны, лучше воспользоваться критерием со статистикой (3), так как при малых объемах выборок распределение статистики (2) будет существенно отличаться от $t_{n_1+n_2-2}$-распределения Стьюдента.

При $n_1+n_2>200$ для всех критериев со статистиками (1) – (3) в качестве распределений статистик можно использовать стандартный нормальный закон.

Двухвыборочный критерий Ван дер Вардена является лучшей непараметрической заменой критериям со статистиками (2) и (3), так как практически не уступает им по мощности.

Недостатком $\tilde{z}$-критерия Манна–Уитни является существенная дискретность распределения статистики, что отражается на точности оценок p_{value}. Это же замечание относится и к H-критерию Краскела–Уаллиса. В то же время оба эти критерия не столь уж сильно уступают по мощности критериям со статистиками (1) – (3).

Применение F-критерия проверки однородности математических ожиданий нескольких выборок целесообразно, если есть основание полагать, что дисперсии, соответствующие выборкам, примерно одинаковы. В противном случае от него следует отказаться и воспользоваться многовыборочным критерием Ван дер Вардена, обладающим практически той же мощностью, или критерием Краскела–Уаллиса, который по мощности немного уступает F-критерию.

Используя непараметрические критерии однородности средних, не следует забывать о том, что область их корректного применения, также как и непараметрических критериев однородности дисперсий, сильно ограничена предположением об однородности законов [30].

Литература

1. *Лемешко Б. Ю., Лемешко С. Б.* Об устойчивости и мощности критериев проверки однородности средних // Измерительная техника. 2008. № 9. С. 23–28.
2. *Lemeshko B. Yu. Lemeshko S. B.* Power and robustness of criteria used to verify the homogeneity of means // Measurement Techniques. 2008. V. 51, № 9. P. 950–959.
3. *Закс Л.* Статистическое оценивание. М.: Статистика, 1976. 598 с.
4. *Welch B. L.* The Significance of the Difference Between Two Means when the Population Variances are Unequal // Biometrika. 1938. V. 29, №. 3/4. P. 350–362.
5. *Behrens W. U.* Ein Beitrag Zur Fehlerberechnung bei wenigen Beobachtungen // Landw. Jb., 1929. B. 68, S. 807–837.
6. *Fisher R. A.* The fiducial argument in statistical inference // Annals of Eugenics, Lond., 1935. V. 6, № 4. P. 391–398.

7. *Fisher R. A.* The asymptotic approach to Behrens's integral, with further tables for the *d* test of significance // Annals of Eugenics, Lond., 1941. V. 11. P. 141–172.
8. *Welch B. L.* The generalization of "Student's" problem when several different population variances are involved // Biometrika. 1947. V. 34, № 1/2. P. 28–35.
9. *Scheffe H.* On Solutions of the Behrens-Fisher Problem, Based on the t-Distribution // The Annals of Mathematical Statistics. 1943. V. 14, № 1. P. 35–44.
10. *Scheffe H.* Practical Solutions of the Behrens-Fisher Problem // Journal of the American Statistical Association. 1970. V. 65, № 332. P. 1501–1508.
11. *Миттаг X. -Й., Ринне X.* Статистические методы обеспечения качества. М.: Машиностроение, 1995. 602 с.
12. Р 50.1.033–2001. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Ч. I. Критерии типа хи-квадрат. М.: Изд-во стандартов, 2002. 87 с.
13. Р 50.1.037–2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Ч. II. Непараметрические критерии. М.: Изд-во стандартов, 2002. 61 с.
14. *Лемешко Б. Ю., Помадин С. С.* Корреляционный анализ наблюдений многомерных случайных величин при нарушении предположений о нормальности // Сибирский журнал индустриальной математики. 2002. Т. 5, № 3. С. 115–130.
15. *Mann H. B., Whitney D. R.* On a test of whether one of two random variables is stochastically larger than the other // Ann. Math. Statist. 1947. V. 18. P. 50–60.
16. *Milton R. C.* An extended table of critical values for the Mann–Whitney (Wilcoxon) two-sample statistic // J. Amer. Statist. Ass. 1964. V. 59. P. 925–934.
17. *Hollander M., Wolfe D. A.* Non-parametric Statistical Methods. 2nd ed. New York: Wiley, 1999.
18. *Conover W. J., Johnson M. E., Johnson M. M.* A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data // Technometrics. 1981. V. 23, № 4. P. 351–361.
19. *Wilcoxon F.* Individual comparisons by ranking methods // Biometrics Bulletin. 1945. № 1. P. 80–83.
20. *Kruskal W. H., Wallis W. A.* Use of ranks in one-criterion variance analysis // J. Amer. Statist. Assoc. 1952. V. 47. P. 583–621.
21. *Kruskal W. H., Wallis W. A.* Use of ranks in one-criterion variance analysis // J. Amer. Statist. Assoc. 1953. V. 48. P. 907–911.
22. *Лемешко Б. Ю.* Непараметрические критерии согласия. Руководство по применению. М.: ИНФРА-М, 2014. 163 с.
23. *Лемешко Б. Ю., Горбунова А. А., Лемешко С. Б., Рогожников А. П.* О решении проблем применения некоторых непараметрических критериев согласия // Автометрия. 2014. Т. 50, № 1. С. 26–43.
24. *Lemeshko B. Yu., Gorbunova A. A., Lemeshko S. B., Rogozhnikov A. P.* Solving problems of using some nonparametric goodness-of-fit tests // Optoelectronics, Instrumentation and Data Processing. 2014. V. 50, № 1. P. 21–35.
25. *Ван дер Варден Б. Л.* Математическая статистика. М.: Иностранная литература, 1960. 435 с.
26. *Кобзарь А. И.* Прикладная математическая статистика. Для инженеров и научных работников. М.: Физматлит, 2006. 816 с.
27. *Fisher R. A., Yates F.* Statistical tables for biological, agricultural and medical research. 6th Ed. Oliver & Boyd, Edinburgh and London, 1963. 146 p.
28. *Terry M. E.* Some rank order test which are most powerful against specific parametric alternatives // The Annals of Mathematical Statistics. 1952. V. 23. P. 346–366.
29. *Hoeffding W.* Optimum non-parametric test // Proc. 11-th Berkeley Symp., 1950. P. 83–92.

30. *Лемешко Б. Ю.* Критерии проверки гипотез об однородности. Руководство по применению. М: ИНФРА–М, 2017. 208 с.

Статья поступила в редакцию 9.06.2017.

Лемешко Борис Юрьевич
д.т.н., профессор, г.н.с. кафедры теоретической и прикладной информатики НГТУ (630073, Новосибирск, просп. Карла Маркса, 20), e-mail: lemeshko@ami.nstu.ru.

Лемешко Станислав Борисович
к.т.н., с.н.с. кафедры теоретической и прикладной информатики НГТУ, e-mail: skyer@mail.ru.

Веретельникова Ирина Викторовна
аспирант кафедры теоретической и прикладной информатики НГТУ, e-mail: ira-veterok@mail.ru.

Новикова Алена Юрьевна
магистрант кафедры теоретической и прикладной информатики НГТУ, e-mail: alena.shestakova.92@inbox.ru.

On the application of checking criteria for homogeneity of means

B. Lemeshko, S. Lemeshko, I. Veretelnikova, A. Novikova

Prerequisites for application of parametric and nonparametric criteria for hypotheses testing on the equality of means are examined. Convergence of statistical distributions of nonparametric criteria to asymptotic ones is investigated by methods of statistical modeling. Comparative analysis of parametric and non-parametric criteria powers has been carried out. Recommendations on criteria application are given.

Keywords: hypotheses testing, homogeneity of means, parametric tests, nonparametric tests, statistics distributions, statistic simulation, test power.