ISSN 1314-3379

BOOTSTRAP СТАТИСТИЧЕСКИ ТЕСТОВЕ ЗА УСТАНОВЯВАНЕ НА РАЗЛИКИ В РАЗСЕЙВАНЕТО НА ИЗВАДКИ

Н. Михайлова, Н. Д. Николова, К. Тенекеджиев

BOOTSTRAP STATISTICALTESTS TO DETECT DIFFERENCES IN SAMPLE DEVIATION

N. Mihaylova, N.D. Nikolova, K. Tenekedjiev

***Резюме**: Изследва се еднаквостта на разпръскването на две разпределения на непрекъснат параметър, за който са налице две извадки от измервания. Част от сравненията се правят с тестови параметри отношенията на числовите характеристики дисперсии в двете извадки, а в останалите сравнения тестови параметри са отношенията на интерквартилните обхвати в двете извадки. Разпределенията на тестовите параметри при вярна нулева хипотеза се построяват с Bootstrap компютърна симулация. Разработени са два алгоритъма (един за дисперсии и един за интерквартилните обхвати) за изчисляване на p_{value}на Bootstrapтестовете за еднаквост на разсейването в две генерални съвкупности, които отчитат и типа на теста (едностранен или двустранен).*

***Ключови думи**: Bootstrapсимулация, числови характеристики, тестване на хипотези*

I. ВЪВЕДЕНИЕ

Нека са налице две едномерни извадки от непрекъснат параметър, условно наречени Извадка 1 и Извадка 2:

$$Z^1 = \left\{ z_1^1, z_2^1, ..., z_{n_1}^1 \right\} \quad (1)$$

$$Z^2 = \left\{ z_1^2, z_2^2, ..., z_{n_2}^2 \right\} \quad (2)$$

Нека измерените наблюдения в Извадка 1 и в Извадка 2 са представители на две генерални съвкупности наречени Генерална Съвкупност 1 и Генерална Съвкупност 2. Целта е да се определи доколко разликата в условията, при които извадките са измерени, са повлияли върху разсейването на признака. Еквивалентна постановка на въпроса е да се счита, че двете извадки са формирани от два различни непрекъснати признака, като е необходимо да се тества възможността двата признака да са еднакви разсейвания.

Отлика между двете генерални съвкупности може да се измерва като разлика между законите на разпределение на разглеждания параметър, описващи двете Генерални Съвкупности. Например, в [1] е представен BootstrapKuipertест за еднаквост на две едномерни непрекъснати разпределения.Друг пример е даден в [6], където се представя аналитичен рангов тест на Wilcoxon, който решава същия проблем при ординални непрекъснати променливи. В[10] е използван аналитичен Kolmogorov-Smirnofftест за решаване на същата задача.

Друг подход за измерване на различиятамежду двете Генерални Съвкупности е да се сравнят стойностите c_1 и c_2 на избрана числова характеристика *C*на двата закона на разпределение, описващи двете Генерални Съвкупности. Едни от най-използваните за тази цел числови характеристики са математическото очакване, дисперсията, медианата и интерквартилния обхват (размах). Първите две са моменти, а последните две – квантили, като първото и третото характеризират положението, а второто и четвъртото – разсейването. Стойностите c_1 и c_2 са детерминирани, но почти винаги неизвестни. Те могат да се оценят на $\hat{c}_1$ и $\hat{c}_2$ след обработка съответно на Извадка 1 и на Извадка 2.Тъй като извадките от генералните съвкупности са случайни, то $\hat{c}_1$ и $\hat{c}_2$ са реализации на

случайните величини $\hat{C}_1$ и $\hat{C}_2$, където първата представлява оценката на c_1 получени от произволна случайна извадка с размер n_1 от Генерална Съвкупност 1, а втората представлява оценката на c_2 получена от произволна случайна извадка с размер n_2 от Генерална Съвкупност 2.

Нека $s = f(c_1, c_2)$ е някаква функция на c_1 и c_2, измерваща различиятамеждустойностите на числовата характеристика *C*на вероятностни разпределения в Генералните Съвкупности 1 и 2. Стойността *s* на функцията може да се нарече измерител на различията между Генералните Съвкупности. Например, когато *C*е числова характеристика на разсейването, то е удобно да избира $s = c_1 / c_2$. Тук s е детерминирана но неизвестна стойност.Същата функция може да се използва заизмерване на различиятамеждуоценките на числовата характеристика *C*в Извадки 1 и 2: $\hat{s} = f(\hat{c}_1, \hat{c}_2)$. Стойността $\hat{s}$ на функцията може да се нарече измерител на различията между Извадките. Тъй като извадките от генералните съвкупности са случайни, то $\hat{s}$ е реализации на случайната величини $\hat{S}$, представляваща измерителя на различията между Извадките за произволна случайна извадка с размер n_1 от Генерална Съвкупност 1, и произволна случайна извадка с размер n_2 от Генерална Съвкупност 2. Символично може да се запише, че $\hat{S} = f(\hat{C}_1, \hat{C}_2)$.

Статистическите тестове, които решават проблема за еднаквостта на числовата характеристика на разсейването *C*на законите на разпределение на случайната величина в Генерална Съвкупност 1 и 2, оценяват типичността на $\hat{s}$ при предпоставка, че Генералните Съвкупности 1 и 2 са с еднакви стойности на числовата характеристика *C*, т.е. $c_1=c_2$. Оценителят $\hat{S}$ се нарича тестова статистика и представлява случайна величина, която трябва да генерира толкова по-различни от единица представители $\hat{s}$, колкото по-голямо е различието между истинските стойности c_1 и c_2. Очевидно е, че формулата за изчисляване на $\hat{s}$ при конкретните извадки неявно ще зависи и от приетата процедура за изчисляване на извадковите числови характеристики $\hat{c}_1$ и $\hat{c}_2$.

Например в един от най-разпространените статистически тестове за различиемежду дисперсиите на двете Генерални Съвкупности, наречен аналитичен F-тест,за оценител $\hat{S}$ се използва частното от отношението на двете дисперсии, като по-голямата винаги е в числител[8]. Търсеното p_{value} се получава при предпоставка за нормално разпределение на данните от аналитично интегриране на *F*-разпределение (Johnson, Kotz et al. 1995). Използваният алгоритъм е реализиран във функцията *vartest2* на MATLAB (Mathworks 2013).

В друг пример за установяване различие в интерквартилните обхвати на двете Генерални Съвкупности, наречен аналитиченAnsari-Bradley тест, за отхвърляне на нулевата хипотеза се използва W-статистика [6], която е непараметрична. Стойността на p_{value} се намира "точно" на базата на пълно изброяване на възможностите [3]. Когато сбора на наблюденията в двете извадки е над 69 или когато по-малката от двете извадки има повече от 15 измервания, тогава времето за изчисление на този тест става неприемливо дълго и затова се използва нормално апроксимиран аналитичен Ansari-Bradley тест и *W*-статистиката се заменя с апроксимирана нормална *W**-статистика, при предпоставка за нормалност на данните [5]. Използваният алгоритъм е реализиран във функцията *ansaribradley* на MATLAB [9], като подаваните към софтуера извадки са центрирани така, че медианите им да са нули. Ограниченията на аналитичнияAnsari-Bradley тест са свързани с предпоставките, че наблюденията в двете Генерални

Съвкупности се подчиняват на еднакъв тип разпределения и, че медианите на Генералните Съвкупности съвпадат с медианите на Извадките. Допълнително, теста е асимптотичен, което изисква голям брой наблюдения и в двете извадки за получаване на надеждни резултати.

В настоящата разработкасе предлагат две групи статистически тестове, които ще установяват различията в характеристиките на двете генерални съвкупности чрез използване на числовите характеристики дисперсияи интерквартилен обхват (iqr). Ще се представят алгоритми за изчисляване на p_{value}на базата на Bootstrapкомпютърна симулация оценяваща разпределението на тестовата статистика при предпоставка за еднаквост на избраните числовите характеристики на разпределенията в двете Генерални Съвкупности.

II.АНАЛИЗ НА РАЗЛИЧИЯТА В ДИСПЕРСИИТЕ

В тази точкаразличиятамежду двата закона на разпределение определящи двете Генерални Съвкупности ще да се сравнят по стойноститена числова характеристика дисперсия (общо наречена по-гореC) . Нека v^1 и v^2 (общо наречени по-горе c_1 и c_2) са дисперсиите на наблюденията в двете Генерални Съвкупности. Тогаваоценките (общо наречени по-горе $\hat{c}_1$ и $\hat{c}_2$) на v^1 и v^2 от двете Извадки са съответно:

$$\hat{v}^1 = \sum_{k=1}^{n_1}\left(z_k^1 - \sum_{r=1}^{n_1} z_r^1 \Big/ n_1 \right)^2 \Big/ \left(n_1 - 1\right) \tag{3}$$

$$\hat{v}^2 = \sum_{k=1}^{n_2}\left(z_k^2 - \sum_{r=1}^{n_2} z_r^2 \Big/ n_2 \right)^2 \Big/ \left(n_2 - 1\right) \tag{4}$$

Извадковите дисперсии $\hat{v}^1$ и $\hat{v}^2$ са реализации на случайните величини $\hat{V}^1$ и $\hat{V}^2$ (общо наречени по-горе $\hat{C}_1$ и $\hat{C}_2$), където първата представлява оценката на дисперсията на произволна случайна извадка с размер n_1 от Генерална Съвкупност 1, а втората представлява оценката на дисперсиятана произволна случайна извадка с размер n_2 от Генерална Съвкупност 2.Група от три статистически теста (един двустранен и два едностранни) се опитва да отговори на въпроса: 'Има ли разлика в дисперсиите на двете Генерални Съвкупности?'.При всички тестове в тази група нулевата хипотеза H_0 е, че двете дисперсии са равни. Следователно, ако H_0 е вярно, то случайнатавеличина $\hat{\Delta}_{1-2}^{var} = \hat{V}^1 / \hat{V}^2$ (общо наречена по-горе $\hat{S}$), би имала стойности близки до единица. Тогава единствената наблюдаваната реализация (общо наречена по-горе $\hat{s}$), на $\hat{\Delta}_{1-2}^{var}$ е:

$$\hat{\delta}_{1-2}^{var} = \hat{v}^1 / \hat{v}^2 \tag{5}$$

Последната би трябвало да е около единица. Всяко „голямо“ отклонение на $\hat{\delta}_{1-2}^{var}$ от единица хвърля съмнение върху валидността на H_0. Затова $\hat{\Delta}_{1-2}^{var}$ може да бъде използвана като тестова статистика за еднаквост на дисперсиите на двете Генерални Съвкупности.

Алтернативните хипотези H_1 са различни при трите тестав групата.

При двустранния тест, H_1 е, че дисперсиите на двете ГенералниСъвкупности са различни. При едностранните тестове, H_1 зависи от изчислените дисперсии на двете извадки. Ако Извадка 1 има по-голяма дисперсия от Извадка 2, то H_1 е, че дисперсията на Генерална Съвкупност 1 е по-голяма от тази на Генерална Съвкупност 2. Ако Извадка 1 има по-малка дисперсия от Извадка 2, то H_1 е, че дисперсията на Генерална Съвкупност 1 е по-малка от тази на Генерална Съвкупност 2. В много редкия случай,

когато Извадка 1 има равна дисперсия с тази наИзвадка 2, то едностранен тест не се извършва.

И така ще се разгледат три статистически теста за еднаквост на дисперсии с общaH_0, а именно „дисперсиите на двете Генерални Съвкупностиса еднакви: $v^1 = v^2$ “. Според алтернативните си хипотези, тестовете биват:

- Двустранен, при който H_1 е „дисперсиите на двете Генерални Съвкупности са различни: $v^1 \neq v^2$ “ , като тестът ще се изпълнява винаги;

- Едностранен от първи тип, при който H_1 е „дисперсията на Генерална Съвкупност1 е по-голяма от тази на Генерална Съвкупност 2: $v^1 > v^2$ “ , като тестът ще се изпълнява само ако $\hat{v}^1 > \hat{v}^2$ (т.е. ако Извадка 1 има по-голяма дисперсия от Извадка 2);

- Едностранен от втори тип, при който H_1 е при който H_1 е „дисперсиятана Генерална Съвкупност1 е по-малка от тази на Генерална Съвкупност 2: $v^1 < v^2$ “, като тестът ще се изпълнява само ако $\hat{v}^1 < \hat{v}^2$ (т.е. ако Извадка 1 има по-малкадисперсия от Извадка 2).

Търсенитep_{value}на трите теста се изчисляват, като разпределението на $\hat{\Delta}_{1-2}^{var}$ при вярна нулева хипотеза се намира с компютърна Bootstrap симулация, като първо наблюдаваните (оригинални) извадки се мащабират така, че дисперсиите им да съвпаднат, и второ синтетичните извадки се получават с теглене с връщане от мащабираните извадки [10; 4].Единствената предпоставка при тези тестове са, че извадките са i.i.d. (independentidenticallydistributed). Разбира се, че изчисляването на p_{value} на тези тестове става със значително повече изчислителни усилия, отколкото при аналитичните тестове.

Ако H_0 е вярно, то двете извадки могат да се мащабират, така че дисперсиите им да са еднакви:

$$Z_{var,b}^1 = \left\{ z_{var,1}^1, z_{var,2}^1, ..., z_{var,n_1}^1 \right\}, \text{ където } z_{var,k}^1 = \frac{z_k^1}{\hat{v}^1} \frac{\left(n_1 \hat{v}^1 + n_2 \hat{v}^2 \right)}{\left(n_1 + n_2 - 2 \right)}, \text{ за } k=1, 2, \ldots, n_1 \quad (6)$$

$$Z_{var,b}^2 = \left\{ z_{var,1}^2, z_{var,2}^2, ..., z_{var,n_2}^2 \right\}, \text{ където } z_{var,k}^2 = \frac{z_k^2}{\hat{v}^2} \frac{\left(n_1 \hat{v}^1 + n_2 \hat{v}^2 \right)}{\left(n_1 + n_2 - 2 \right)}, \text{ за } k=1, 2, \ldots, n_2 \quad (7)$$

От мащабираните извадки последователно ще се теглят чрез връщане N_b на брой синтетични двойки извадки с размери n_1 и n_2 , като q-тата двойка (q=1,2,...,N_b) има вида:

$$Z_{var,q}^{1,s} = \left\{ z_{var,1,q}^{1,s}, z_{var,2,q}^{1,s}, ..., z_{var,n_1,q}^{1,s} \right\} \quad (8)$$

$$Z_{var,q}^{2,s} = \left\{ z_{var,1,q}^{2,s}, z_{var,2,q}^{2,s}, ..., z_{var,n_2,q}^{2,s} \right\} \quad (9)$$

За q-тата двойка извадки може да бъде изчислена отношението на дисперсиите на синтетичните извадки $\hat{\delta}_{1-2,q}^{var,s}$ съгласно (3), (4) и (5).

Ако N_b е голямо число, примерно 100 000, то разпределението на статистиката „частно на дисперсии“ при вярно H_0, може да се построи по честотни методи [11]. Всъщност необходимо е да се пресметнат p_{value} на трите теста, а не самото разпределение. При двустранен тест използвайки честотната интерпретация на вероятността, p_{value}е относителният брой синтетични $\hat{\delta}_{1-2,q}^{var,s}$ статистики, чийто стойносте

не в интервала между по-малката ипо-голямата стойност измеждунаблюдаваната статистика $\hat{\delta}_{1-2}^{var}$ и нейната реципрочна $1/\hat{\delta}_{1-2}^{var}$.

$$p_{value} = \frac{\sum_{q=1}^{N_b} 1}{\hat{\delta}_{1-2,q}^{var,s} \geq max\left(\hat{\delta}_{1-2}^{var}, 1/\hat{\delta}_{1-2}^{var}\right)} + \frac{\sum_{q=1}^{N_b} 1}{\hat{\delta}_{1-2,q}^{var,s} \leq min\left(\hat{\delta}_{1-2}^{var}, 1/\hat{\delta}_{1-2}^{var}\right)} \qquad (10)$$

$$N_b \qquad\qquad N_b$$

При едностранен тест от първи тип,използвайкичестотната интерпретация на вероятността,p_{value} ще е относителния брой на синтетични $\hat{\delta}_{1-2,q}^{var,s}$ статистики, чийто стойност е не по-малка от по-голямата стойност измежду наблюдаваната статистика $\hat{\delta}_{1-2}^{var}$ и нейната реципрочна $1/\hat{\delta}_{1-2}^{var}$:

$$p_{value} = \frac{\sum_{q=1}^{N_b} 1}{\hat{\delta}_{1-2,q}^{var,s} \geq max\left(\hat{\delta}_{1-2}^{var}, 1/\hat{\delta}_{1-2}^{var}\right)} \qquad (11)$$

$$N_b$$

На практика едностранният тест от първи тип се прилага, когато $\hat{v}^1 > \hat{v}^2$ и тогава условието в сумата на (11) става $\hat{\delta}_{1-2,q}^{var,s} \geq \hat{\delta}_{1-2}^{var}$, защото $\hat{\delta}_{1-2}^{var} > 1$.

При едностранен тест от втори тип,използвайкичестотната интерпретация на вероятността,p_{value} ще е относителния брой на синтетични $\hat{\delta}_{1-2,q}^{var,s}$ статистики, чийто стойност е не по-голяма от по-малката стойност измеждунаблюдаваната статистика $\hat{\delta}_{1-2}^{var}$ и нейната реципрочна $1/\hat{\delta}_{1-2}^{var}$:

$$p_{value} = \frac{\sum_{q=1}^{N_b} 1}{\hat{\delta}_{1-2,q}^{var,s} \leq min\left(\hat{\delta}_{1-2}^{var}, 1/\hat{\delta}_{1-2}^{var}\right)} \qquad (12)$$

$$N_b$$

Аналогично, едностранен тест от втори тип се проверява само когато $\hat{v}^1 < \hat{v}^2$ и тогава условието на сумата от формулата (12) става $\hat{\delta}_{1-2,q}^{var,s} \leq \hat{\delta}_{1-2}^{var}$, защото $\hat{\delta}_{1-2}^{var} < 1$.

Алгоритъм за изчисляване на трите p_{value} при Bootstrap тестове за равенство на дисперсиите

1. Изчисли дисперсията на първа извадка $\hat{v}^1$ по формулата (3).
2.Изчисли дисперсията навтората извадка $\hat{v}^2$ по формулата (4).
3. Изчисли наблюдаваната статистика 'частно на дисперсиите' $\hat{\delta}_{1-2}^{var}$ по формула (5).
4. Формирай първата мащабирана извадка $Z_{var,b}^1$ по формула (6).
5. Формирай второтамащабирана извадка $Z_{var,b}^2$ по формула (7).
6. Избери броя на псевдо-реалностите N_b.
7. Инициирай цикъл по номер на псевдо-реалността: q=1.
8. Формирай синтетична извадка $Z_{var,q}^{1,s}$:

а) инициирай цикъл по номер на генерирано наблюдение на първа извадка: k=1;

б) изтегли с връщане от $Z^1_{var,b}$, *k*-тото наблюдение на $Z^{1,s}_{var,q}$ както следва:

*Генерирай случайно равномерно разпределено в интервала $(0-1]$ число: *rand*.

*Намери номера на *k*-тото изтеглено наблюдение: *numt*=$\lceil rand \times n_1 \rceil$

* Определи *k*-тото изтеглено наблюдение: $z^{1,s}_{var,k,q} = z^1_{var,numt}$

в) премини към следващото изтеглено наблюдение: *k*=*k*+1

г) провери дали $Z^1_{var,b}$ е попълнено: ако $k \le n_1$, иди към 8б

д) запомни синтетичната извадка: $Z^{1,s}_{var,q} = \left\{ z^{1,s}_{var,1,q}, z^{1,s}_{var,2,q}, ..., z^{1,s}_{var,n_1,q} \right\}$

9. Формирай синтетична извадка $Z^{2,s}_{var,q}$:

а) инициирай цикъл по номер на генерираното наблюдение на втора извадка: *k*=1;

б) изтегли с връщане от Z^2_{var} , *k*-тото наблюдение на $Z^{2,s}_{var,q}$

* Генерирай случайно равномерно разпределено в интервала $(0-1]$ число *rand*.

* Намери номера на *k*-тото изтеглено наблюдение:*numt*=$\lceil rand \times n_2 \rceil$

* Определи *k*-тото изтеглено наблюдение: $z^{2,s}_{var,k,q} = z^2_{var,numt}$

в) Премини към следващото изтеглено наблюдение: *k*=*k*+1

г) Провери дали $Z^{2,s}_{var,q}$ е попълнено: ако $k \le n_2$, иди към 9б

д) запомни синтетичната извадка: $Z^{2,s}_{var,q} = \left\{ z^{2,s}_{var,1,q}, z^{2,s}_{var,2,q}, ..., z^{2,s}_{var,n_2,q} \right\}$

10. Изчисли статистиката 'частно на дисперсии' в *q*-тата псевдо-реалност като $\hat{\delta}^{var,s}_{1-2,q} = \dfrac{\sum_{k=1}^{n_1} \left(z^{1,s}_{var,k,q} - \sum_{r=1}^{n_1} z^{1,s}_{var,k,q} \Big/ n_1 \right)^2 (n_2 - 1)}{\sum_{k=1}^{n_2} \left(z^{2,s}_{var,k,q} - \sum_{r=1}^{n_2} z^{2,s}_{var,k,q} \Big/ n_2 \right)^2 (n_1 - 1)}$. Ако знаменателя е нула, а числителя не е, то положи $\hat{\delta}^{var,s}_{1-2,q} = \infty$, а ако и числителя и знаменателя са нули, то положи $\hat{\delta}^{var,s}_{1-2,q} = 1$:

11. Мини към следващата псевдо-реалност: *q*=*q*+1.

12. Провери за край на симулацията: ако $q \le N_b$, иди към т. 8.

13. Запомни генерираните стойности на статистиките: $F^{var,s} = \left\{ \hat{\delta}^{var,s}_{1-2,1}, \hat{\delta}^{var,s}_{1-2,2}, ..., , \hat{\delta}^{var,s}_{1-2,N_b} \right\}$

14. Изчисли p_{value} за двустранен статистически тест по (10). Дефинирай H_1'дисперсиите на двете Генерални Съвкупности са различни: $v^1 \ne v^2$ '.

15. Ако $\hat{v}^1 > \hat{v}^2$ изчисли p_{value} за едностранен статистически тест по формула (11).Дефинирай H_1"дисперсията на Генерална Съвкупност1 е по-голяма от тази на Генерална Съвкупност 2: $v^1 > v^2$ " .

16. Ако $\hat{v}^1 < \hat{v}^2$, изчисли p_{value} за едностранен статистически тест съгласно (12). Дефинирай H_1"дисперсията на Генерална Съвкупност1 е по-малка от тази на Генерална Съвкупност 2: $v^1 < v^2$ ".

III. АНАЛИЗ НА РАЗЛИЧИЯТА В ИНТЕРКВАРТИЛНИТЕ ОБХВАТИ

В тази точкаразличиятамежду двата закона на разпределение определящи двете Генерални Съвкупности ще се сравнят по стойноститена числова характеристика интерквартилен обхват (общо наречена по-горе*C*). Нека r^1 и r^2 (общо наречени по-горе c_1 и c_2) са интерквартилните обхвати на наблюденията в двете Генерални Съвкупности.

Нека елементите от извадките Z^1 и Z^2от (1) и (2) са сортирани и са получени извадки $Z^{1,sort}$ и $Z^{2,sort}$:

$$Z^{1,sort} = \left\{ z_1^{1,sort}, z_2^{1,sort}, ..., z_{n_1}^{1,sort} \right\} \tag{13}$$

където $z_1^{1,sort} \le z_2^{1,sort} \le ... \le z_{n_1}^{1,sort}$

$$Z^{2,sort} = \left\{ z_1^{2,sort}, z_2^{2,sort}, z_3^{2,sort}, ..., z_{n_2}^{2,sort} \right\} \tag{14}$$

където $z_1^{2,sort} \le z_2^{2,sort} \le z_3^{2,sort} \le ... \le z_{n_2}^{2,sort}$.

Тогава оценките (общо наречени по-горе $\hat{c}_1$ и $\hat{c}_2$) на r^1 и r^2 от двете Извадки в двете извадки ще са съответно:

$$\hat{r}^1 = \begin{cases} \dfrac{z_{3k+1}^{1,sort} + z_{3k}^{1,sort}}{2} - \dfrac{z_{k+1}^{1,sort} + z_k^{1,sort}}{2} & \text{при } n_1 = 4k \quad \text{,за} \quad k \in \\ \dfrac{z_{3k+2}^{1,sort} + 3 * z_{3k+1}^{1,sort}}{4} - \dfrac{3 * z_{k+1}^{1,sort} + z_k^{1,sort}}{4} & \text{при } n_1 = 4k+1 \text{ ,за} \quad k \in \\ z_{3k+2}^{1,sort} - z_{k+1}^{1,sort} & \text{при } n_1 = 4k+2 \text{ ,за} \quad k \in \\ \dfrac{3 * z_{3k+3}^{1,sort} + z_{3k+2}^{1,sort}}{4} - \dfrac{z_{k+2}^{1,sort} + 3 * z_{k+1}^{1,sort}}{4} & \text{при } n_1 = 4k+3 \text{ ,за} \quad k \in \end{cases} \tag{15}$$

$$\hat{r}^2 = \begin{cases} \dfrac{z_{3k+1}^{2,sort} + z_{3k}^{2,sort}}{2} - \dfrac{z_{k+1}^{2,sort} + z_k^{2,sort}}{2} & \text{при } n_2 = 4k \quad \text{,за} \quad k \in \\ \dfrac{z_{3k+2}^{2,sort} + 3 * z_{3k+1}^{2,sort}}{4} - \dfrac{3 * z_{k+1}^{2,sort} + z_k^{2,sort}}{4} & \text{при } n_2 = 4k+1 \text{ ,за} \quad k \in \\ z_{3k+2}^{2,sort} - z_{k+1}^{2,sort} & \text{при } n_2 = 4k+2 \text{ ,за} \quad k \in \\ \dfrac{3 * z_{3k+3}^{2,sort} + z_{3k+2}^{2,sort}}{4} - \dfrac{z_{k+2}^{2,sort} + 3 * z_{k+1}^{2,sort}}{4} & \text{при } n_2 = 4k+3 \text{ ,за} \quad k \in \end{cases} \tag{16}$$

Извадковите дисперсии $\hat{r}^1$ и $\hat{r}^2$ са реализации на случайните величини $\hat{R}^1$ и $\hat{R}^2$ (общо наречени по-горе $\hat{C}_1$ и $\hat{C}_2$), където първата представлява оценката на интерквартилния обхват на произволна случайна извадка с размер n_1 от Генерална Съвкупност 1, а втората представлява оценката на интерквартилния обхват на произволна случайна извадка с размер n_2 от Генерална Съвкупност 2. Група от три статистически теста (един двустранен и два едностранни) се опитва да отговори на въпроса: 'Има ли разлика в интерквартилните обхватина двете Генерални Съвкупности?'. При всички тестове в тази група нулевата хипотеза H_0 е, че дватаинтерквартилни обхвата са равни. Следователно, ако H_0 е вярно, то случайната величина $\hat{\Delta}_{1-2}^{iqr} = \hat{R}^1 / \hat{R}^2$ (общо наречена по-горе $\hat{S}$), би имала стойности близки до единица. Тогава единствената наблюдаваната реализация (общо наречена по-горе $\hat{s}$), на $\hat{\Delta}_{1-2}^{iqr}$ е:

$$\hat{\delta}_{1-2}^{iqr} = \hat{r}^1 / \hat{r}^2 \tag{17}$$

Последната би трябвало да е около единица. Всяко „голямо“ отклонение на $\hat{\delta}_{1-2}^{iqr}$ от единица хвърля съмнение върху валидността на H_0. Затова $\hat{\Delta}_{1-2}^{iqr}$ може да бъде използвана като тестова статистика за еднаквост на интерквартилния обхват на двете Генерални Съвкупности.

Алтернативните хипотези H_1 са различни при трите тестав групата.

При двустранния тест, H_1 е, че интерквартилните обхвати на двете ГенералниСъвкупности са различни. При едностранните тестове, H_1 зависи от изчислените интерквартилни обхвати на двете извадки. Ако Извадка 1 има по-голям интерквартилен обхват от Извадка 2, то H_1 е, че интерквартилният обхват на Генерална Съвкупност 1 е по-голям от тази на Генерална Съвкупност 2. Ако Извадка 1 има по-малък интерквартилен обхват от Извадка 2, то H_1 е, че интерквартилният обхват на Генерална Съвкупност 1 е по-малък от този на Генерална Съвкупност 2. В много редкия случай, когато Извадка 1 има равен интерквартилен обхват с този на Извадка 2, то едностранен тест не се извършва.

И така ще се разгледат три статистически теста за еднаквост на интерквартилни обхвати с общатаH_0, а именно „интерквартилните обхвати на двете Генерални Съвкупностиса еднакви: $r^1 = r^2$ “. Според алтернативните си хипотези, тестовете биват:

- Двустранен, при който H_1 е „интерквартилните обхвати на двете Генерални Съвкупности са различни: $r^1 \neq r^2$ “ , като тестът ще се изпълнява винаги;

- Едностранен от първи тип, при който H_1 е „интерквартилният обхват на Генерална Съвкупност1 е по-голям от тази на Генерална Съвкупност 2: $r^1 > r^2$ “ , като тестът ще се изпълнява само ако $\hat{r}^1 > \hat{r}^2$ (т.е. ако Извадка 1 има по-голям интерквартилен обхват от Извадка 2);

- Едностранен от втори тип, при който H_1 е при който H_1 е „интерквартилният обхват на Генерална Съвкупност1 е по-малък от този на Генерална Съвкупност 2: $r^1 < r^2$ “, като тестът ще се изпълнява само ако $\hat{r}^1 < \hat{r}^2$ (т.е. ако Извадка 1 има по-малък интерквартилен обхват от Извадка 2).

Търсенитеp_{value}на трите теста се изчисляват, като разпределението на $\hat{\Delta}_{1-2}^{iqr}$ при вярна нулева хипотеза се намира с компютърна Bootstrap симулация, като първо наблюдаваните (оригинални) извадки се мащабират така, че интерквартилните им обхвати да съвпаднат, и второ синтетичните извадки се получават с теглене с връщане от мащабираните извадки [10; 4].Единствената предпоставка при тези тестове са, че извадките са i.i.d. (independentidenticallydistributed). Разбира се, че изчисляването на p_{value} на тези тестове става със значително повече изчислителни усилия, отколкото при аналитичните тестове.

Нека се търси интерквартилният обхват на обединената извадка от двете извадки Z^1 и Z^2от (1) и (2), която съдържа $n=n_1 + n_2$наблюдения:

$$Z_b^{1-2} = \left\{ z_1^1, z_2^1, ..., z_{n_1}^1, z_1^2, z_2^2, ..., z_{n_2}^2 \right\} \tag{18}$$

Нека Z_b^{1-2} се сортира:

$$Z^{1-2,sort} = \left\{ z_1^{1-2}, z_2^{1-2}, ..., z_n^{1-2} \right\} \tag{19}$$

където $z_1^{1-2} \leq z_2^{1-2} \leq ... \leq z_n^{1-2}$

Интерквартилния обхват на обединената извадка, може да се пресметне, като:

$$\hat{r}^{1-2}=\begin{cases}\dfrac{z_{3k+1}^{1-2,sort}+z_{3k}^{1-2,sort}}{2}-\dfrac{z_{k+1}^{1-2,sort}+z_{k}^{1-2,sort}}{2} & \text{при } n=4k \quad ,\text{за} \quad k\in \\ \dfrac{z_{3k+2}^{1-2,sort}+3*z_{3k+1}^{1-2,sort}}{4}-\dfrac{3*z_{k+1}^{1-2,sort}+z_{k}^{1-2,sort}}{4} & \text{при } n=4k+1 \ ,\text{за} \quad k\in \\ z_{3k+2}^{1-2,sort}-z_{k+1}^{1-2,sort} & \text{при } n=4k+2 \ ,\text{за} \quad k\in \\ \dfrac{3*z_{3k+3}^{1-2,sort}+z_{3k+2}^{1-2,sort}}{4}-\dfrac{z_{k+2}^{1-2,sort}+3*z_{k+1}^{1-2,sort}}{4} & \text{при } n=4k+3 \ ,\text{за} \quad k\in \end{cases} \tag{20}$$

Ако H_0 е вярно, то двете извадки могат да се мащабират, така че интерквартилните им обхвати да са еднакви:

$$Z_{iqr,b}^{1}=\left\{z_{iqr,1}^{1},z_{iqr,2}^{1},...,z_{iqr,n_1}^{1}\right\}, \text{ където } z_{var,k}^{1}=\frac{z_k^1}{\hat{r}^1}\hat{r}^{1-2}, \text{ за } k=1, 2, \ldots, n_1 \tag{21}$$

$$Z_{iqr,b}^{2}=\left\{z_{iqr,1}^{2},z_{iqr,2}^{2},...,z_{iqr,n_2}^{2}\right\}, \text{ където } z_{var,k}^{2}=\frac{z_k^2}{\hat{r}^2}\hat{r}^{1-2}, \text{ за } k=1, 2, \ldots, n_2 \tag{22}$$

От мащабираните извадки последователно ще се теглят чрез връщане N_b на брой синтетични двойки извадки с размери n_1 и n_2, като q-тата двойка (q=1,2,...,N_b) има вида:

$$Z_{iqr,q}^{1,s}=\left\{z_{iqr1,q}^{1,s},z_{iqr,2,q}^{1,s},.....,z_{iqrn_1,q}^{1,s}\right\} \tag{23}$$

$$Z_{iqr,q}^{2,s}=\left\{z_{iqr,1,q}^{2,s},z_{iqr,2,q}^{2,s},.....,z_{iqr,n_2,q}^{2,s}\right\} \tag{24}$$

За q-тата двойка извадки може да бъде изчислена отношението на дисперсиите на синтетичните извадки $\hat{\delta}_{1-2,q}^{iqr,s}$ съгласно (15), (16) и (17).

Ако N_b е голямо число, примерно 100 000, то разпределението на статистиката „частно на интерквартилни обхвати“ при вярно H_0, може да се построи по честотни методи [11]. Всъщност необходимо е да се пресметнат p_{value} на трите теста, а не самото разпределение. При двустранен тест използвайки честотната интерпретация на вероятността, p_{value}е относителният брой синтетични $\hat{\delta}_{1-2,q}^{iqr,s}$ статистики, чийто стойност е не в интервала между по-малката ипо-голямата стойност измежду наблюдаваната статистика $\hat{\delta}_{1-2}^{iqr}$ и нейната реципрочна $1/\hat{\delta}_{1-2}^{iqr}$.

$$p_{value}=\frac{\sum_{q=1}^{N_b}1}{N_b}_{\hat{\delta}_{1-2,q}^{iqr,s}\geq max\left(\hat{\delta}_{1-2}^{iqr},1/\hat{\delta}_{1-2}^{iqr}\right)}+\frac{\sum_{q=1}^{N_b}1}{N_b}_{\hat{\delta}_{1-2,q}^{iqr,s}\leq min\left(\hat{\delta}_{1-2}^{iqr},1/\hat{\delta}_{1-2}^{iqr}\right)} \tag{25}$$

При едностранен тест от първи тип,използвайкичестотната интерпретация на вероятността,p_{value} ще е относителния брой на синтетични $\hat{\delta}_{1-2,q}^{iqr,s}$ статистики, чийто стойност е не по-малка от по-голямата стойност измежду наблюдаваната статистика $\hat{\delta}_{1-2}^{iqr}$ и нейната реципрочна $1/\hat{\delta}_{1-2}^{iqr}$:

$$p_{value} = \frac{\sum_{q=1}^{N_b} 1}{N_b} \quad \hat{\delta}_{1-2,q}^{iqr,s} \geq max\left(\hat{\delta}_{1-2}^{iqr}, 1/\hat{\delta}_{1-2}^{iqr}\right) \tag{26}$$

На практика едностранният тест от първи тип се прилага, когато $\hat{r}^1 > \hat{r}^2$ и тогава условието в сумата на (26) става $\hat{\delta}_{1-2,q}^{iqr,s} \geq \hat{\delta}_{1-2}^{iqr}$, защото $\hat{\delta}_{1-2}^{iqr} > 1$.

При едностранен тест от втори тип,използвайкичестотната интерпретация на вероятността,p_{value} ще е относителния брой на синтетични $\hat{\delta}_{1-2,q}^{iqr,s}$ статистики, чийто стойност е не по-голяма от по-малката стойност измежду наблюдаваната статистика $\hat{\delta}_{1-2}^{iqr}$ и нейната реципрочна $1/\hat{\delta}_{1-2}^{iqr}$:

$$p_{value} = \frac{\sum_{q=1}^{N_b} 1}{N_b} \quad \hat{\delta}_{1-2,q}^{iqr,s} \leq min\left(\hat{\delta}_{1-2}^{iqr}, 1/\hat{\delta}_{1-2}^{iqr}\right) \tag{27}$$

Аналогично, едностранен тест от втори тип се проверява само когато $\hat{r}^1 < \hat{r}^2$ и тогава условието на сумата от формулата (27) става $\hat{\delta}_{1-2,q}^{iqr,s} \leq \hat{\delta}_{1-2}^{iqr}$, защото $\hat{\delta}_{1-2}^{iqr} < 1$.

Алгоритъм за изчисляване на трите p_{value} при Bootstrap тестове за равенство на интерквартилни обхвати

1. Сортирай първата извадка Z^1 от (1) и получи $Z^{1,sort}$от (13).
2. Изчисли интерквартилния обхват на първата извадка $\hat{r}^1$ по формула (15).
3. Сортирай втората извадка Z^2 от (2) и получи $Z^{2,sort}$от (14).
4. Изчисли интерквартилния обхват на втората извадка $\hat{r}^2$ по формула (16).
5. Изчисли наблюдаваната статистика 'частно на интерквартилните обхвати' $\hat{\delta}_{1-2}^{iqr}$ по формула (17).
6. Сортирай общата извадка Z_b^{1-2} от (18) и получи $Z^{1-2,sort}$ от формула (19).
7. Изчисли интерквартилния обхватна обединената извадка $\hat{r}^{1-2}$ по формула (20).
8. Формирай първата мащабирана извадка $Z_{iqr,b}^1$ по формула (21).
9. Формирай втората мащабирана извадка $Z_{iqr,b}^2$ по формула (22).
10. Избери броя на псевдо-реалностите N_b.
11. Инициирай цикъл по номер на псевдо-реалността: q=1.
12. Формирай синтетична извадка $Z_{iqr,q}^{1,s}$:

а) инициирай цикъл по номер на генерираното наблюдение на първа извадка: k=1;

б) изтегли с връщане от $Z_{iqr,b}^1$, k-тото наблюдение на $Z_{iqr,q}^{1,s}$

*Генерирай случайно равномерно разпределено в интервала $(0-1]$ число: *rand*.

*Намери номера на k-тото изтеглено наблюдение: $numt=\lceil rand \times n_1 \rceil$

* Определи k-тото изтеглено наблюдение: $z_{iqr,k,q}^{1,s} = z_{iqr,numt}^1$

в) премини към следващото изтеглено наблюдение: k=k+1

г) провери дали $Z_{iqr,q}^{1,s}$ е попълнено: ако $k \leq n_1$, иди към 12б

д) запомни синтетичната извадка: $Z_{iqr,q}^{1,s} = \left\{z_{iqr,1,q}^{1,s}, z_{iqr,2,q}^{1,s}, ..., z_{iqr,n_1,q}^{1,s}\right\}$

13. Сортирай $Z_{iqr,q}^{1,s}$ до $Z_{iqr,q}^{1,sort,s} = \left\{z_{iqr,1,q}^{1,sort,s}, z_{iqr,2,q}^{1,sort,s}, ..., z_{iqr,n_1,q}^{1,sort,s}\right\}$, където $z_{iqr,1,q}^{1,sort,s} \leq z_{iqr,2,q}^{1,sort,s} \leq ... \leq z_{iqr,n_1,q}^{1,sort,s}$.

14. Изчисли интерквартилния обхват на $Z_{iqr,q}^{1,s}$ съгласно

$$\hat{r}_q^1 = \begin{cases} \dfrac{z_{iqr.3k+1,q}^{1,sort,s} + z_{iqr,3k,q}^{1,sort,s}}{2} - \dfrac{z_{iqr,k+1,q}^{1,sort,s} + z_{iqr,k,q}^{1,sort,s}}{2} & \text{при } n_1 = 4k \quad ,\text{за} \quad k \in \\ \dfrac{z_{iqr,3k+2,q}^{1,sort,s} + 3* z_{iqr,3k+1,q}^{1,sort,s}}{4} - \dfrac{3* z_{iqr,k+1,q}^{1,sort,s} + z_{iqr,k,q}^{1,sort,s}}{4} & \text{при } n_1 = 4k+1 \ ,\text{за} \quad k \in \\ z_{iqr,3k+2,q}^{1,sort,s} - z_{iqr,k+1,q}^{1,sort,s} & \text{при } n_1 = 4k+2 \ ,\text{за} \quad k \in \\ \dfrac{3* z_{iqr,3k+3,q}^{1,sort,s} + z_{iqr,3k+2,q}^{1,sort,s}}{4} - \dfrac{z_{iqr,k+2,q}^{1,sort,s} + 3* z_{iqr,k+1,q}^{1,sort,s}}{4} & \text{при } n_1 = 4k+3 \ ,\text{за} \quad k \in \end{cases}$$

15. Формирай синтетична извадка $Z_{iqr,q}^{2,s}$:

а) инициирай цикъл по номер на генерираното наблюдение на втора извадка: *k*=1;

б) изтегли с връщане от $Z_{iqr,b}^{2}$, *k*-тото наблюдение на $Z_{iqr,q}^{2,s}$

*Генерирай случайно равномерно разпределено в интервала $(0-1]$ число: *rand*.

*Намери номера на *k*-тото изтеглено наблюдение: *numt*=$\lceil rand \times n_2 \rceil$

* Определи *k*-тото изтеглено наблюдение: $z_{iqr,k,q}^{2,s} = z_{iqr,numt}^{2}$

в) премини към следващото изтеглено наблюдение: *k*=*k*+1

г) провери дали $Z_{iqr,q}^{2,s}$ е попълнено: ако $k \le n_2$, иди към 15б

д) запомни синтетичната извадка: $Z_{iqr,q}^{2,s} = \left\{ z_{iqr,1,q}^{2,s}, z_{iqr,2,q}^{2,s}, ..., z_{iqr,n_2,q}^{2,s} \right\}$

16. Сортирай $Z_{iqr,q}^{2,s}$ до $Z_{iqr,q}^{2,sort,s} = \left\{ z_{iqr,1,q}^{2,sort,s}, z_{iqr,2,q}^{2,sort,s}, ..., z_{iqr,n_2,q}^{2,sort,s} \right\}$, където $z_{iqr,1,q}^{2,sort,s} \le z_{iqr,2,q}^{2,sort,s} \le ... \le z_{iqr,n_2,q}^{2,sort,s}$.

17. Изчисли интерквартилния обхват на $Z_{iqr,q}^{2,s}$ съгласно

$$\hat{r}_q^2 = \begin{cases} \dfrac{z_{iqr.3k+1,q}^{2,sort,s} + z_{iqr,3k,q}^{2,sort,s}}{2} - \dfrac{z_{iqr,k+1,q}^{2,sort,s} + z_{iqr,k,q}^{2,sort,s}}{2} & \text{при } n_2 = 4k \quad ,\text{за} \quad k \in \\ \dfrac{z_{iqr,3k+2,q}^{2,sort,s} + 3* z_{iqr,3k+1,q}^{2,sort,s}}{4} - \dfrac{3* z_{iqr,k+1,q}^{2,sort,s} + z_{iqr,k,q}^{2,sort,s}}{4} & \text{при } n_2 = 4k+1 \ ,\text{за} \quad k \in \\ z_{iqr,3k+2,q}^{2,sort,s} - z_{iqr,k+1,q}^{2,sort,s} & \text{при } n_2 = 4k+2 \ ,\text{за} \quad k \in \\ \dfrac{3* z_{iqr,3k+3,q}^{2,sort,s} + z_{iqr,3k+2,q}^{2,sort,s}}{4} - \dfrac{z_{iqr,k+2,q}^{2,sort,s} + 3* z_{iqr,k+1,q}^{2,sort,s}}{4} & \text{при } n_2 = 4k+3 \ ,\text{за} \quad k \in \end{cases}$$

18. Изчисли статистиката 'частно на интерквартилни обхвати' в *q*-тата псевдо-реалност като $\hat{\delta}_{1-2,q}^{iqr,s} = \hat{r}_q^{1,s} / \hat{r}_q^{1,s}$. Ако знаменателя е нула, а числителя не е, то положи $\hat{\delta}_{1-2,q}^{iqr,s} = \infty$, а ако и числителя и знаменателя са нули, то положи $\hat{\delta}_{1-2,q}^{iqr,s} = 1$.

19. Мини към следващата псевдо-реалност: *q*=*q*+1.

20. Провери за край на симулацията: ако $q \le N_b$, иди към т. 12.

21. Запомни генерираните стойности на статистиките: $F^{iqr,s} = \left\{ \hat{\delta}_{1-2,1}^{iqr,s}, \hat{\delta}_{1-2,2}^{iqr,s}, ..., , \hat{\delta}_{1-2,N_b}^{iqr,s} \right\}$

22. Изчисли p_{value} за двустранен статистически тест по (25). Дефинирай H_1'интерквартилните обхвати на двете Генерални Съвкупности са различни: $r^1 \ne r^2$ '.

23. Ако $\hat{r}^1 > \hat{r}^2$ изчисли p_{value} за едностранен статистически тест по формула (26).Дефинирай H_1'интерквартилния обхват на Генерална Съвкупност1 е по-голям от този на Генерална Съвкупност 2: $r^1 > r^2$ ' .

24. Ако $\hat{r}^1 < \hat{r}^2$ изчисли p_{value} за едностранен статистически тест по формула (27).Дефинирай H_1'интерквартилния обхват на Генерална Съвкупност1 е по-малък от този на Генерална Съвкупност 2: $r^1 < r^2$ ' .

IV. ЗАКЛЮЧЕНИЕ

В разработкатаса представени две процедури, чрез които се проверява еднаквостта в характеристиките на разсейване на две Генерални Съвкупности, представени с по една Извадка от наблюдения на едномерен непрекъснат параметър. При първата процедура за тестова статистикае използвано частнотона дисперсиите между извадките, а при втората – частното на интерквартилните обхвати. В двата разработени алгоритъма се използват Bootstrapсимулационни процедури, които при за еднакви разсейвания, оценяват на разпределенията на тестовите статистикии изчисляват на p_{value}при едностранните или при двустранни тестове. Апробацията на предложените тестове ще се търси в приложението им върху числови данни от реални експерименти.

БИБЛИОГРАФИЯ

1. Николова, Н. Д., Михайлова, Н., Тенекеджиев, К., BootstrapKuiper тест за еднаквост на едномерни непрекъснати разпределения, Международна конференция "Автоматика и информатика", София, 4-6 Октомври, стр., I-135-I-138, 2013а

2. Николова, Н.Д., Михайлова, Н., Тенекеджиев, К., Статистическо доказване на разлики между извадки чрез Bootstrap тест за проверка на хипотези, Наука, Труд и Капитал, бр. 2, стр. 36-44, 2013б

3. Ansari, A., Bradley, R., Rank-Sum Tests for Dispersions, Annals of Mathematical Statistics Volume 31(No. 4): 1174-1189, 1960

4. Efron, B., Tibshirani, R. J., An Introduction to the Bootstrap. Chapman & Hall, 1993

5. Hajek, J., Sidak, Z. et al, Theory of Rank Tests. London: Academic Press. San Diego, 1999

6. Hollander, M., D. A. Wolfe. NonparametricStatistical Methods. Hoboken, NJ: John Wiley & Sons,Inc., 1999

7. Johnson, N., Kotz, S. et al., Continuous Univariate Distributions, Wiley, 1995

8. Markowski, C., Markowski, E., Condition for the Effectiveness of a Preliminary Test of Variance, The American Statistician, Vol. 44, No. 4 322-326, 1990

9. Mathworks, T., Statistics Toolbox: User's Guide, The MathWorks, Inc., USA, 2013

10. Press, W. H., S. A. Teukolski, W. T. Vetterling, B. P. Flannery, Numerical Recipes – The Art of Scientific Computing, Cambridge University Press, 1992

11.Tenekedjiev, K., Dimitrakiev, D., Nikolova, N.D., Building Frequentist Distributions of Continuous Random Variables, Machine Mechanics, Volume 47, pp.164-168, 2002

За контакти:
докт. Нели Михайлова, Висше Военноморско училище "Никола Йонков Вапцаров", Варна 9026, ул "Васил Друмев" №73,
тел. +3592552245,mihaylova_neli@abv.bg
доц. д-р Наталия Николова, Висше Военноморско училище "Никола Йонков Вапцаров", Варна 9026, ул "Васил Друмев" №73,
*тел. +359885401514,natalianik@gmail.com
проф. д.т.н. инж. Кирил Тенекеджиев, Висше Военноморско училище "Никола Йонков Вапцаров", Варна 9026, ул "Васил Друмев" №73,
*тел. +359885401503,Kiril.Tenekedjiev@fulbrightmail.org