ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

АКТУАЛНОСТ НА ПРОБЛЕМА

Едно от често срещаните сърдечни заболявания в световен мащаб е исхемичната болест на сърцето (ИБС), усложнена с исхемична митрална регургитация (ИМР), при която изходящата клапа на лявата камера не функционира добре и разрешава връщане в сърцето на част от изтласканата кръв. Изследванията от последните години показват тенденция това заболяване да се среща не само при хора в напреднала възраст, но и сред пациенти между 40 и 50 години. Счита се, че прогнозата за пациентите с ИМР е по-лоша в сравнение с тази на пациенти с МР от друг произход, поради обвързаността на ИМР с увреждане на лявата камера на сърцето. Изследванията върху този проблем за българската популация са доста ограничени.

В дисертация на доц. д-р Пламен Панайотов от 2013 г. се идентифицират критерии за оценка на предоперативното състояние на пациенти с ИБС, усложнена със значима хронична ИМР и се анализират резултатите от оперативното им лечение. Формулирани са алгоритми за класификация на оперативното лечение и е оценен ефекта на оперативните интервенции върху обратното ремоделиране на сърдечни кухини при различните подгрупи пациенти. Данните са събрани между януари 2007 и юни 2011 г. в Отделението по кардиохирургия към УМБАЛ "Св. Марина" – ЕАД, град Варна, където за периода са се лекували 1196 пациенти от ИБС с някаква степен на МР. От тях подходящи за участие в изследването са били 140 пациенти.

Настоящата дисертационна работа представя осъвременени версии на използваните в коментирания по-горе труд количествени процедури, математическите алгоритми и симулационни подходи за статистически анализ, които са разработени от докторантката под ръководството на доц. Наталия Николова и консултирани с проф. Кирил Тенекеджиев. Обект на дисертацията се явява информационният аспект на споменатото изследване, като новопредложените статистически техники в неговите рамки може да се използват за решаването на някои типични проблеми, срещащи се при изграждането на медицински информационни системи.

Настоящата дисертация е насочена към разработване и адаптиране на техники, процедури, алгоритми и подходи, които да намират приложение в разнообразни медицински информационни системи, които при липса на типични контролни групи и при произволни едномодални и платообразни функции на предпочитание върху отделните параметри на състоянието, от една страна да отчитат взаимовръзките между статистическите тестове за произволна двойки извадки, а от друга страна да са в състояние да доказват ефект на провеждани медицински процедури. Формализирани са серии от симулационни статистически тестове, които оценяват ефекта на различните курсове на лечение върху серия параметри на състоянието на пациентите. Предложеният информационен инструментариум е илюстриран върху конкретен информационен масив от медицински данни за пациенти, като успешно са адресирани някои от изброените по-горе проблеми при изграждането на медицински информационни системи.

Научната актуалността на проблема е неоспорима. Симулационната техника Буутстрап са модерно и мощно средство за преодоляване трудностите, свързани често невъзможните за удовлетворяване предпоставки при статистическите доказателства така характерни за класическата статистика. Съвременната тенденция в много сфери на науката класическата статистика да се заменя със симулационни методи е следвана в настоящият научен труд и е в полза на доклазването на нейната ефективност.

НАУЧНА ЗНАЧИМОСТ И НОВОСТ

Разработката има съществени научно-приложни новости. Показва се важността и ролята на информационните медицински системи в съвременното здравеопазване. Подходите тук напълно замества необходимостта да се провеждат стандартно познатите и изисквани широко-мащабни медицински изследвания, при които се обхващат милиони пациенти в продължение на около 20 години, преди да се публикуват достоверни резултати въз основа на огромния обем информация. Широкото въвеждане на информационни медицински системи е също и начин за възприемане на медицината, основана на доказателства, което се доказва с настоящата разработка.

Поставените задачи изискват използване на богата серия от техники, което е постигнато в разработката. В първичната обработка на данните се използват многообещаващи Буутстрап алгоритми и кластерни подходи към класическия проблем за компютърно симулационно моделиране по метода Монте Карло. Прилагат се емпирични специфични за областта знания, екстрахирани от медицински специалисти, относно типа на предпочитанията върху наблюдавани медицински параметри. Изпползват се принципите на модерния софт-компютинг за интегрирането на получените статистически резултати от кластерните Буутстрап процедури с емпирично установените динамични промени в състоянието на пациентите. Съществен елемент на новост е установяването на влиянието на допълнително въздействие без типични да

се използват класическите контролни групи. Намерени са задачи от сферата на медицината, върху които е демонстрирано приложението на новопредложените алгоритми, процедури и подходи.

ПРАКТИЧЕСКА ПОЛЕЗНОСТ НА РЕЗУЛТАТИТЕ

Към документите на докторантката са приложени две писма за използване на представените алгоритми, техники и процедури в медицински научни изследвания. Едното писмо удостоверява използване на техниките в Семелвайския медицински университет, Будапеща, Унгария, в рамките на научен проект, финансиран от унгарски научни фондове. Второто писмо декларира използване на медицинска информационна система за пациенти с исхемична болест на сърцето, усложнена с исхемична митрална регургитация в Клиниката по кардиохирургия към УМБАЛ „Св. Марина“, като статистическите алгоритми и процедури, вложени в системата, са разработени от маг. Нели Костадинова Михайлова под ръководството на доц. Наталия Николова и проф. Кирил Тенекеджиев.

АПРОБАЦИЯ НА РЕЗУЛТАТИТЕ

Идеите и подходите, които са изложени в дисертационния труд, са представени, обсъдени и изцяло отпечатани (или под печат) на английски език в следните международни конференции:

IFAC TECIS, Созопол, България, 24-27 Септември 2015

EUREKA 2015: Fifth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support, April 22-24 April, Mexico City, Mexico

Международна конференция "Автоматика и информатика'2013", стр. I-135 – I-138, 3-7 октомври 2013, София, България

Окончателните резултати от дисертационния труд са докладвани на разширен катедрен съвет на катедра “Информационни технологии ” при ВВМУ „Н. Вапцаров“, гр. Варна, състоял се на 30.09.2015.

ПУБЛИКУВАНЕ НА РЕЗУЛТАТИТЕ

Представените нови концепции в дисертационния труд са обект на публикации в Доклади на Съюза на учените – Варна, както и в Наука, труд и капитал.

СТРУКТУРА И ОБЕМ НА РАБОТАТА

Дисертационният труд е в обем от 1162 стр., от които 6 стр. литература и 12 стр. съдържание, въведение, таблица на съкращенията и благодарности. Дисертацията съдържа Въведение, 3 глави, 2 приложения и заключение. Приложенията са с обем от 150 стр. Използваните литературни източници са 139, от които 16 са на кирилица, а 123 са на латиница.

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. ОБЕКТ НА ИЗСЛЕДВАНЕТО И ОБЗОР НА СТАТИСТИЧЕСКИ МЕТОДИ, ИЗПОЛЗВАНИ В МЕДИЦИНСКИТЕ ИНФОРМАЦИОННИ СИСТЕМИ

Едно от често срещаните сърдечни заболявания в световен мащаб е исхемичната болест на сърцето (ИБС), усложнена с исхемична митрална регургитация (ИМР), при която изходящата клапа на лявата камера не функционира добре и разрешава връщане в сърцето на част от изтласканата кръв. В изследване, представено в дисертация на доц. д-р Пламен Панайотов, д.м. през 2013 г. се идентифицират критерии за оценка на предоперативното състояние на пациенти с ИБС, усложнена със значима хронична ИМР и се анализират резултатите от оперативното им лечение. Формулирани са алгоритми за класификация на оперативното лечение и е оценен ефекта на оперативните интервенции върху обратното ремоделиране на сърдечни кухини при различните подгрупи пациенти. Данните са събрани между януари 2007 и юни 2011 г. в Отделението по кардиохирургия към МБАЛ "Св. Марина" – ЕАД, град Варна. Подходящи за участие в изследването са били 140 пациенти

Настоящата дисертационна работа представя осъвременени версии на използваните в дисертацията на доц. Панайотов количествени процедури, математическите алгоритми и симулационни подходи за статистически анализ, които са разработени от докторантката под ръководството на доц. Наталия Николова и консултирани с проф. Кирил Тенекеджиев. Обект на дисертацията се явява информационният аспект на споменатото изследване, като новопредложените статистистически техники в неговите рамки може да се използват за решаването на някои типични проблеми, срещащи се при изграждането на медицински информационни системи. Настоящият научен труд няма никакви научни претенции относно медицинския аспект в дисертацията на доц. Панайотов, който, без да се ограничава до, включва: събраните данни за пациентите, избора на конкретни методи за оперативно лечение, избора на диагностични признаци за оценка на ефективността на пластиката на митрална клапа върху анализираните пациенти, предлагането на нов диагностичен признак ’модифицирана фракция на изтласкване’ за оценка на степента на митралната регургитация преди и след операцията, създадените три алгоритъма за избор на оперативно лечение и за класификация на пациентите в подгрупи, медицинската интерпретация на статистическите резултати.

Литературният обзор обхваща 139 източника. От тях, 48 са след 2005 г., като 10 са източници на български, а останалите 38 са на английски език. Останалите 91 са публикациите до 2005 г., от които 6 на български език и 85 на английски език.

От обзора на статистическата литература следват няколко извода:

- Огромен брой статистически тестове с високи изисквания и нереалистични допускания, поради което използването им е затруднено
- Буутстрап симулационните процедури са универсален заместител на класически статистически тестове, но се избягват, тъй като рандомизирания им характер дава различни резултати при всяко изчисление
- Липсват специфични числени алгоритми на Буутстрап статистическите тестове
- Почти липсват изследвания за група тестове, приложени върху една извадка от данни и за взаимодействието между тестовете
- Контролна група в медицинските проучвания, поради етични и хуманни ограничения, често отсъства и не се предлага алтернатива

Целта на дисертацията е да се разработят и апробират върху реален медицински проблем алгоритмичните основи на Буутстрап-базирана симулационна система за обработка на медицинска информация, която при липса на типични контролни групи и при произволни едномодални и платообразни функции на предпочитание върху отделните параметри на състоянието, от една страна да отчита взаимовръзките между всички статистически тестове за произволна двойки извадки, а от друга страна да е в състояние да доказва ефекти на въздействие на медицински процедури.

Поставената цел може да се постигне чрез решаването на следните *задачи*:

1. Формиране на кластери от стандартизирани Буутстрап статистически тестове, оценяващи както различията в непрекъснат или дискретен параметър между извадки от две популации, така и ненулевостта на промяната в непрекъснат параметър по двойка сдвоени извадки от една популация

2. Разработване на алгоритми, които при произволни едномодални и платообразни функции на предпочитание да определят благоприятността както на различията в непрекъснат или дискретен параметър между извадки от две популации, така и на промяната в непрекъснат параметър по двойка сдвоени извадки от една популация.

3. Разработване на алгоритми, които при произволно избрано ниво на значимост да класифицират статистическата значимост както на различията в непрекъснат или дискретен параметър между извадки от две популации, така и на промяната в непрекъснат параметър по двойка сдвоени извадки от една популация.

4. Разработване на алгоритми, които да класифицират в отделни типове на съвместна благоприятност и значимост както различията в непрекъснат или дискретен параметър между извадки от две популации, така и промяната в непрекъснат параметър по двойка сдвоени извадки от една популация.

5. Създаване на алгоритъм за оценка на ефект от допълнително въздействие върху дискретен или непрекъснат параметър на състоянието, работещ без типична контролна група, чрез сравняване на извадки както преди, така и след допълнителното въздействие.

6. Създаване на алгоритъм за оценка на ефект от допълнително въздействие върху непрекъснат параметър на състоянието, работещ без типична контролна група, чрез сравняване на промяната в двойка сдвоени извадки вследствие на допълнителното въздействие с промяната в двойка сдвоени извадки, неподложени на допълнителното въздействие.

7. Апробиране на предложените тестове и алгоритми при реален медицински проблем за доказване на ефекта от допълнително въздействие.

8. Апробиране на предложените тестове и алгоритми при реален медицински проблем за сравнение между диагностичните способности на параметри за установяване на ефект от допълнително въздействие с независимо доказано влияние.

ГЛАВА 2. АЛГОРИТМИЧНИ ОСНОВИ НА СИСТЕМА ЗА ОБРАБОТКА НА МЕДИЦИНСКА ИНФОРМАЦИЯ

2.1. Буутстрап тестове

Сравняват се двойки извадки.

Първи кластер – двете извадки са едномерни и съдържат стойностите на избран непрекъснат параметър. Пациентите във всяка от двете извадки отговарят на различни ограничения. Пациентите в първата извадка се отличават от тези във втората точно по един фактор. Така влиянието на този фактор върху избрания непрекъснат параметър ще може да се оцени чрез сравнение на извадките.

Втори кластер – двете извадки са едномерни и съдържат стойностите на избран дискретен признак. Пациентите във всяка от двете извадки отговарят на различни ограничения. Пациентите в първата извадка се отличават от тези във втората точно по един фактор. Така влиянието на този фактор върху избрания дискретен признак ще може да се оцени чрез сравнение на извадките.

Трети кластер – извадката е двумерна и съдържа стойностите на избран параметър, измерен в две времеви точки. Пациентите в извадката отговарят на дадени ограничения. Втората времева точка трябва да бъде преди първата, като двойката е от тип "първа времева точка – втора времева точка". Така влиянието на оперативната намеса върху избрания обобщен непрекъснат тримерен параметър ще може да се оцени чрез оценка на промяната между двете.

2.1.1. Тестове от първи кластер

Нека са налице две едномерни извадки Извадка 1 и Извадка 2 от непрекъснат параметър Z:

$$Z^1 = \left\{ z_1^1, z_2^1, ..., z_{n_1}^1 \right\} \tag{2.1.1.1}$$

$$Z^2 = \left\{ z_1^2, z_2^2, ..., z_{n_2}^2 \right\} \tag{2.1.1.2}$$

Наблюдения в Извадка 1 и в Извадка 2 са представители на Популация 1 и Популация 2. Целта е да се определи дали Популация 1 има същите характеристики като Популация 2. Закона на разпределението на непрекъсната случайна величина Z в дадена популация е кумулативна функция на разпределение $F(.)$:

$$F(z) = P(Z \le z), \qquad \text{за } z \in (-\infty; +\infty) \tag{2.1.1.3}$$

2.1.1.1. Буутстрап Kuiper тест за еднаквост на непрекъснати популационни разпределения

Апроксимация по извадка на $F(.)$ на непрекъсната променлива е емпиричната кумулативна функция на разпределение (ECDF):

$$CDF_1(z) = \frac{1}{n_1} \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_1} 1, \qquad \text{за } z \in (-\infty; +\infty) \tag{2.1.1.1.1}$$

$$CDF_2(z) = \frac{1}{n_2} \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_2} 1, \qquad \text{за } z \in (-\infty; +\infty) \tag{2.1.1.1.2}$$

Наблюденията в извадките са i.i.d. (independent identically distributed). Задачата е да се провери дали кумулативните функции на разпределение $F_1(.)$ и $F_2(.)$ на Z се отличават в двете популации. H_0 е, че непрекъснатите разпределения на двете популации са еднакви. H_1 е, че тези разпределения са различни.

За измерител на различията между две произволни кумулативни функции на разпределение *CDF'*(.) и *CDF"*(.) е предложен функционалът Kuiper статистика (*KU*):

$$KU = KU\left(CDF',CDF''\right) = \sup_z\left(CDF'(z) - CDF''(z)\right) + \sup_z\left(CDF''(z) - CDF'(z)\right) \quad (2.1.1.1.3)$$

Ако $KU(.,.)$ се пресметне за $CDF_1(.)$ и $CDF_2(.)$ апроксимирани от двете извадки, то (2.1.1.1.3) ще даде реалната оценка ku_{re} на Kuiper статистиката:

$$ku_{re} = KU\left(CDF_1,CDF_2\right) = \sup_z\left(CDF_1(z) - CDF_2(z)\right) + \sup_z\left(CDF_2(z) - CDF_1(z)\right) \quad (2.1.1.1.4)$$

Ако популационните разпределения са апроксимирани с емпирични кумулативни функции на разпределение, то пресмятането на реалната оценка ku_{re} на KU може да се опрости до:

$$ku_{re} = \max_{k=1,2,...,n_1}\left(CDF_1\left(z_k^1\right) - CDF_2\left(z_k^1\right)\right) + \max_{k=1,2,...,n_2}\left(CDF_2\left(z_k^2\right) - CDF_1\left(z_k^2\right)\right) \quad (2.1.1.1.5)$$

Тестовата статистика може да се получи директно от наблюденията в извадките:

$$ku_{re} = \max_{k=1,2,...,n_1}\left(\frac{k}{n_1} - \frac{1}{n_2}\sum_{\substack{r=1 \\ z_r^2 \le z_k^1}}^{n_2} 1\right) + \max_{k=1,2,...,n_2}\left(\frac{k}{n_2} - \frac{1}{n_1}\sum_{\substack{r=1 \\ z_r^1 \le z_k^2}}^{n_1} 1\right) \quad (2.1.1.1.6)$$

Нека разпределенията на двете популации са еднакви и равни на $F_{tr}(.)$. Тези две еднакви разпределения не са известни на експериментатора, но се реализират статистически чрез $(Z^1 - Z^2)$. По извадките може да се изчисли реалната оценка ku_{re} на Kuiper статистиката. От $F_{tr}(.)$ биха могли да се генерират безброй много *хипотетични* двойки извадки: $\left(Z_1^1 - Z_1^2\right)$, $\left(Z_2^1 - Z_2^2\right)$, $\left(Z_3^1 - Z_3^2\right)$, ... Експериментаторът би могъл да изчисли хипотетичните реализации ku_1, ku_2, ku_3, ... на Kuiper статистиката по всяка от двойките хипотетични извадки.

Разликата $KU - ku_{tr} = KU - 0 = KU$ е случайна величина, наречена *root* и неговото разпределение описва изцяло количествената несигурност в реалната тестова статистика ku_{re}. Задача тук е да се намери разпределението на *root* без да е известно $F_{tr}(.)$ и без да са известни $\left(Z_1^1 - Z_1^2\right)$, $\left(Z_2^1 - Z_2^2\right)$, $\left(Z_3^1 - Z_3^2\right)$, ... Тъй като истинското $F_{tr}(.)$ е неизвестно, то би могло да се замени с негова оценка. Ако H_0 е вярно, то $F_{tr}(.)$ е реализирано статистически (n_1+n_2) пъти в извадките (2.1.1.1) и (2.1.1.2). Нека двете извадки Z^1 и Z^2 се обединят в една обща извадка, съдържаща $(n_1 + n_2)$ наблюдения:

$$Z^{1\cup 2} = \left\{z_1^1, z_2^1, ..., z_{n_1}^1, z_1^2, z_2^2, ..., z_{n_2}^2\right\} = \left\{z_1^{1\cup 2}, z_2^{1\cup 2}, ..., z_{n_1+n_2}^{1\cup 2}\right\} \quad (2.1.1.1.7)$$

Тогава, $F_{tr}(.)$ може да се апроксимира с ECDF, построена върху общата извадка (2.1.1.1.6):

$$CDF_{1\cup 2}(z) = \frac{1}{n_1+n_2}\sum_{\substack{k=1 \\ z_k^{1\cup 2} \le z}}^{n_1+n_2} 1, \quad \text{за } z \in (-\infty;+\infty) \quad (2.1.1.1.8)$$

В компютърния свят от $CDF_{1\cup 2}(.)$ може да се генерират N_b синтетични $\left(Z_q^{1,s} - Z_q^{2,s}\right)$ за $q=1,2,\ldots,N_b$:

$$Z_q^{1,s} = \left\{z_{1,q}^{1,s}, z_{2,q}^{1,s}, ..., z_{n_1,q}^{1,s}\right\} \quad (2.1.1.1.9)$$

$$Z_q^{2,s} = \left\{z_{1,q}^{2,s}, z_{2,q}^{2,s}, ..., z_{n_2,q}^{2,s}\right\} \quad (2.1.1.1.10)$$

За q-тата двойка $\left(Z_q^{1,s} - Z_q^{2,s}\right)$ се изчислява синтетичната Kuiper статистика ku_q^s:

$$ku_q^s = \max_{k=1,2,...,n_1}\left(\frac{k}{n_1} - \frac{1}{n_2}\sum_{\substack{r=1 \\ z_{r,q}^{2,s} \le z_{k,q}^{1,s}}}^{n_2} 1\right) + \max_{k=1,2,...,n_2}\left(\frac{k}{n_2} - \frac{1}{n_1}\sum_{\substack{r=1 \\ z_{r,q}^{1,s} \le z_{k,q}^{2,s}}}^{n_1} 1\right) \quad (2.1.1.1.11)$$

Получава се синтетична извадка $S_{KU^s} = \left\{ku_1^s, ku_2^s, ..., ku_{N_b}^s\right\}$ от N_b реализации на KU^s. Ако N_b е голямо число (например 100000), то кумулативното разпределение на KU^s може да се апроксимира с много голяма точност, като ECDF, построена върху синтетичната извадка S_{KU^s}:

$$CDF_{KU^s}(ku) = \frac{1}{N_b}\sum_{\substack{q=1 \\ ku_q^s \le ku}}^{N_b} 1 \quad \text{, за } ku \in (-\infty;+\infty) \quad (2.1.1.1.12)$$

$KU^s - ku_{tr,cw} = KU^s - 0 = KU^s$ е случайна величина, наречена $root_{cv}$. Основното допускане при Буутстрап метода е, че разпределението на $root_{cv}$ в компютърния свят е същото като разпределението на *root* в истинския.

Търсената апроксимация $CDF_{KU}(.)$ на условното разпределение на KU, ако H_0 е вярна, съвпада с вече построената апроксимация на разпределението на KU^s. Тогава:

$$p_{value} = 1 - CDF_{KU}(ku_{re}) + P(KU = ku_{re}) = \frac{1}{N_b} \sum_{\substack{q=1 \\ ku_q^s \geq ku_{re}}}^{N_b} 1 \qquad (2.1.1.1.13)$$

Разработен е алгоритъм за намиране на p_{value} на Буутстрап Kuiper теста.

Алгоритъм 2.1.1.1 за изчисляване на p_{value} на Буутстрап Kuiper тест за еднаквост на популационни разпределения на непрекъснат параметър

1. Изчисли реалната оценка на Kuiper статистика ku_{re} по (2.1.1.1.6)
2. Ако ku_{re}=0, то положи p_{value}=1 и край
3. Формирай обща извадка $Z^{1\cup 2}$ по (2.1.1.1.7)
4. Избери броя N_b на псевдо-реалностите
5. Инициирай цикъл по псевдо-реалности: q=1
6. Формирай първа синтетична извадка $Z_q^{1,s}$ в q-тата псевдореалност:
   6.1. Инициирай цикъл по номер на наблюдение на първа синтетична извадка: k=1
   6.2. Изтегли с връщане от $Z^{1\cup 2}$ k-тото наблюдение на $Z_q^{1,s}$:
      6.2.1. Генерирай случайно равномерно разпределено число $rand \in (0;1]$
      6.2.2. Намери номера на k-тото изтеглено наблюдение $numt = \lceil rand \times (n_1 + n_2) \rceil$, където $\lceil x \rceil$ е най-малкото цяло число, не по-малко от x (т.е. $\lceil \ \rceil$ е оператор за закръгляне нагоре)
      6.2.3. Определи k-тото изтеглено наблюдение $z_{k,q}^{1,s} = z_{numt}^{1\cup 2}$
   6.3. Премини към следващото изтеглено наблюдение: k=k+1
   6.4. Провери дали $Z_q^{1,s}$ е попълнено: ако $k \leq n_1$, то иди към точка 6.2
   6.5. Запомни първата синтетична извадка $Z_q^{1,s} = \{z_{1,q}^{1,s}, z_{2,q}^{1,s}, ..., z_{n_1,q}^{1,s}\}$
7. Формирай втора синтетичната извадка $Z_q^{2,s}$ в q-тата псевдореалност
   7.1. Инициирай цикъл по номер на наблюдение на втора синтетична извадка: k=1
   7.2. Изтегли с връщане от $Z^{1\cup 2}$ k-тото наблюдение на $Z_q^{2,s}$:
      7.2.1. Генерирай случайно равномерно разпределено число $rand \in (0;1]$
      7.2.2. Намери номера на k-тото изтеглено наблюдение $numt = \lceil rand \times (n_1 + n_2) \rceil$
      7.2.3. Определи k-тото изтеглено наблюдение $z_{k,q}^{2,s} = z_{numt}^{1\cup 2}$
   7.3. Премини към следващо изтеглено наблюдение: k=k+1
   7.4. Провери дали $Z_q^{2,s}$ е попълнено: ако $k \leq n_2$, то иди към точка 7.2
   7.5. Запомни втората синтетична извадка $Z_q^{2,s} = \{z_{1,q}^{2,s}, z_{2,q}^{2,s}, ..., z_{n_2,q}^{2,s}\}$
8. Изчисли Kuiper статистиката ku_q^s в q-тата псевдо-реалност по (2.1.1.1.11)
9. Мини към следваща псевдо-реалност: q=q+1
10. Провери за край на симулацията: ако $q \leq N_b$, то иди към точка 6
11. Запомни синтетичните стойности на тестовите Kuiper статистики $S_{KU^s} = \{ku_1^s, ku_2^s, ..., ku_{N_b}^s\}$
12. Изчисли p_{value} по (2.1.1.1.13) и край

2.1.1.4. Буутстрап тестове за равенство на популационни дисперсии

Една от най-разпространените оценки по извадка на дисперсия $\pm^2$ на популационен закон на разпределение на непрекъсната случайна величина е извадкова дисперсия VAR:

$$VAR_1 = \sum_{k=1}^{n_1} \left(z_k^1 - \sum_{r=1}^{n_1} z_r^1 \Big/ n_1 \right)^2 \Big/ (n_1 - 1) \qquad (2.1.1.4.1)$$

$$VAR_2 = \sum_{k=1}^{n_2}\left(z_k^2 - \sum_{r=1}^{n_2} z_r^2 \Big/ n_2 \right)^2 \Big/ \left(n_2 - 1\right) \qquad (2.1.1.4.2)$$

Оценките (2.1.1.4.1) и (2.1.1.4.2) са състоятелни и неизместени, ако наблюденията в извадките са i.i.d.

Задачата е да се провери дали дисперсиите $\pm_1^2$ и $\pm_2^2$ на непрекъснатата случайна величина Z се отличават в двете популации. Нулевата хипотеза H_0 е, че дисперсиите на непрекъснатите разпределения на двете популации са еднакви ($\pm_1^2 = \pm_2^2$). Алтернативните хипотези H_1 са различни за двустранен Буутстрап тест за дисперсии и за едностранните Буутстрап тестове за дисперсии. При двустранния тест, H_1 е че $\pm_1^2 \neq \pm_2^2$. При Първия едностранен тест H_1 е, че $\pm_1^2 > \pm_2^2$. При Втория едностранен тест H_1 е, че $\pm_1^2 < \pm_2^2$. В настоящата разработка, ако $VAR_1 > VAR_2$, то за едностранен тест се приема Първият едностранен тест, а ако $VAR_1 < VAR_2$, то за едностранен тест се приема Вторият едностранен тест. В много редкия случай, когато $VAR_1 = VAR_2$, то нито Първият, нито Вторият едностранен тест се извършва, техните p_{value} се фиксират на 0.5, а който и да е от двата теста се нарича едностранен тест. Двустранният тест също не се извършва, а p_{value} се полага за 1. По-нататъшното разглеждане е само ситуацията, в която $VAR_1 > 0$ и $VAR_2 > 0$.

За измерител на различията между v' и v'' е предложен критерия на Фишер:

$$\mathrm{i}_{1:2}^{VAR} = \mathrm{i}_{1:2}^{VAR}\left(v', v''\right) = v' / v'' \qquad (2.1.1.4.3)$$

Ако $\mathrm{i}_{1:2}^{VAR}(.,.)$ се пресметне за VAR_1 и VAR_2, то

$$\mathrm{a}_{1:2,re}^{VAR} = VAR_1 / VAR_2 \qquad (2.1.1.4.4)$$

Ако в (2.1.1.4.4) за оценки VAR_1 и VAR_2 на популационните дисперсии на непрекъсната променлива се използват извадковите дисперсии от (2.1.1.4.1) и (2.1.1.4.2), то:

$$\mathrm{a}_{1:2,re}^{VAR} = \frac{\left(n_2 - 1\right)\sum_{k=1}^{n_1}\left(z_k^1 - \sum_{r=1}^{n_1} z_r^1 \Big/ n_1 \right)^2}{\left(n_1 - 1\right)\sum_{k=1}^{n_2}\left(z_k^2 - \sum_{r=1}^{n_2} z_r^2 \Big/ n_2 \right)^2} \qquad (2.1.1.4.5)$$

Нека кумулативните функции на разпределение на двете популации $F_1(.)$ и $F_2(.)$ са с еднакви дисперсии, равни на $\pm_{tr}^2$. Тези две разпределения не са известни на експериментатора, но се реализират статистически чрез двойката извадки (Z^1 – Z^2). По извадките може да се изчисли реалната оценка $\mathrm{a}_{1:2,re}^{VAR}$ на статистиката отношение между извадкови дисперсии. От $F_1(.)$ и $F_2(.)$ биха могли да се генерират безброй много *хипотетични* двойки извадки: $\left(Z_1^1 - Z_1^2\right)$, $\left(Z_2^1 - Z_2^2\right)$, $\left(Z_3^1 - Z_3^2\right)$, ... Експериментаторът би могъл да изчисли хипотетичните реализации $\mathrm{a}_{1:2,1}^{VAR}$, $\mathrm{a}_{1:2,2}^{VAR}$, $\mathrm{a}_{1:2,3}^{VAR}$, … .

Разликата $\mathrm{i}_{1:2}^{VAR} - \mathrm{a}_{1:2,tr}^{VAR} = \mathrm{i}_{1:2}^{VAR} - 1$ е случайна величина, наречена *root* и неговото разпределение описва изцяло количествената несигурност в реалната тестова статистика $\mathrm{a}_{1:2,re}^{VAR}$. Задачата тук е да се намери разпределението на *root* без да са известни $F_1(.)$, $F_2(.)$ и $\pm_{tr}^2$ и без да са известни хипотетичните двойки извадки $\left(Z_1^1 - Z_1^2\right)$, $\left(Z_2^1 - Z_2^2\right)$, $\left(Z_3^1 - Z_3^2\right)$, ... Тъй като истинските $F_1(.)$, $F_2(.)$ и $\pm_{tr}^2$ са неизвестни, то те биха могли да се заменят с техни оценки.

Ако H_0 е вярно, то неизвестната дисперсия в двете популации $\pm_{tr}^2$ може да се замени с оценка на общата дисперсия VAR по наблюденията в двете извадки:

$$VAR = \frac{\left(n_1 - 1\right)VAR_1 + \left(n_2 - 1\right)VAR_2}{n_1 + n_2 - 2} \qquad (2.1.1.4.6)$$

Ако H_0 е вярно, то $F_1(.)$ и $F_2(.)$ са реализирани статистически n_1 и n_2 пъти в извадките Z^1 и Z^2, но мащабирана така, че извадковата дисперсия вместо VAR_1 и VAR_2 да стане VAR, което е най-добрата оценка на неизвестната $\pm_{tr}^2$. Тогава, $F_1(.)$ и $F_2(.)$се апроксимират с ECDF по мащабираните извадки

$$Z_m^1 = \left\{ z_{m,1}^1, z_{m,2}^1, \ldots, z_{m,n_1}^1 \right\}, \text{ където } z_{n,k}^1 = z_k^1 \sqrt{VAR} / \sqrt{VAR_1}, \text{ за } k=1, 2, \ldots, n_1 \qquad (2.1.1.4.7)$$

$$CDF_{1,m}\left(z\right) = \frac{1}{n_1} \sum_{\substack{k=1 \\ z_{m,k}^1 \le z}}^{n_1} 1, \quad \text{за } z \in \left(-\infty; +\infty\right) \qquad (2.1.1.4.8)$$

$$Z_m^2 = \left\{ z_{m,1}^2, z_{m,2}^2, \ldots, z_{m,n_2}^2 \right\}, \text{ където } z_{n,k}^2 = z_k^2 \sqrt{VAR} / \sqrt{VAR_2} \text{ , за } k=1, 2, \ldots, n_2 \qquad (2.1.1.4.9)$$

$$CDF_{2,m}(z) = \frac{1}{n_2} \sum_{\substack{k=1 \\ z_{m,k}^2 \le z}}^{n_2} 1, \qquad \text{за } z \in (-\infty; +\infty) \qquad (2.1.1.4.10)$$

В компютърния свят от известните $CDF_{1,m}(.)$ и $CDF_{2,m}(.)$ може да се генерират N_b синтетични двойки извадки $\left(Z_{m,q}^{1,s} - Z_{m,q}^{2,s} \right)$ за q=1,2,…,N_b чрез теглене чрез връщане от мащабираните извадки:

$$Z_q^{1,s} = \left\{ z_{1,q}^{1,s}, z_{2,q}^{1,s}, \ldots, z_{n_1,q}^{1,s} \right\} \qquad (2.1.1.4.11)$$

$$Z_q^{2,s} = \left\{ z_{1,q}^{2,s}, z_{2,q}^{2,s}, \ldots, z_{n_2,q}^{2,s} \right\} \qquad (2.1.1.4.12)$$

$$VAR_1^s = \sum_{k=1}^{n_1} \left(z_{1,q}^{1,s} - \sum_{r=1}^{n_1} z_{1,q}^{1,s} \Big/ n_1 \right)^2 \Big/ (n_1 - 1) \qquad (2.1.1.4.13)$$

$$VAR_2^s = \sum_{k=1}^{n_2} \left(z_{1,q}^{2,s} - \sum_{r=1}^{n_2} z_{1,q}^{2,s} \Big/ n_2 \right)^2 \Big/ (n_2 - 1) \qquad (2.1.1.4.14)$$

Сега може да се изчисли синтетичната Фишер тестова статистика $\mathsf{a}_{1:2,q}^{VAR,s}$:

$$\mathsf{a}_{1:2,q}^{VAR,s} = \begin{cases} VAR_1^s / VAR_2^s & \text{ако } VAR_2^s > 0 \\ +\infty & \text{ако } VAR_1^s > 0 \text{ и } VAR_2^s = 0 \\ 1 & \text{ако } VAR_1^s = VAR_2^s = 0 \end{cases} \qquad (2.1.1.4.15)$$

Получава се синтетична извадка $S_{\mathsf{i}_{1:2}^{VAR,s}} = \left\{ \mathsf{a}_{1:2,1}^{VAR,s}, \mathsf{a}_{1:2,2}^{VAR,s}, \ldots, \mathsf{a}_{1:2,N_b}^{VAR,s} \right\}$ от N_b реализации на $\mathsf{i}_{1:2}^{VAR,s}$. Ако N_b е голямо число (например 100000), то:

$$CDF_{\mathsf{i}_{1:2}^{VAR,s}}(\mathsf{a}) = \frac{1}{N_b} \sum_{\substack{q=1 \\ \mathsf{a}_{1:2,q}^{VAR,s} \le \mathsf{a}}}^{N_b} 1 \qquad \text{, за } \mathsf{a} \in (-\infty; +\infty) \qquad (2.1.1.4.16)$$

Разликата $\mathsf{i}_{1:2}^{VAR,s} - \mathsf{a}_{1:2,tr,cw}^{VAR,s} = \mathsf{i}_{1:2}^{VAR,s} - 1$ е случайна величина, наречена $root_{cv}$. В случая *root* съвпада с $\mathsf{i}_{1:2}^{VAR} - 1$, а $root_{cv}$ съвпада с $\mathsf{i}_{1:2}^{VAR,s} - 1$. Оттук следва, че разпределенията на $\mathsf{i}_{1:2}^{VAR}$ и $\mathsf{i}_{1:2}^{VAR,s}$ съвпадат. Нека се въведат горна и долна помощни граници за реалната извадкова Фишерова статистика:

$$\mathsf{a}_{1:2,re}^{VAR,u} = max\left\{ \mathsf{a}_{1:2,re}^{VAR}, 1 / \mathsf{a}_{1:2,re}^{VAR} \right\} \qquad (2.1.1.4.17)$$

$$\mathsf{a}_{1:2,re}^{VAR,d} = min\left\{ \mathsf{a}_{1:2,re}^{VAR}, 1 / \mathsf{a}_{1:2,re}^{VAR} \right\} \qquad (2.1.1.4.18)$$

Могат да се изведат формули за намиране на Pvalue. При първия едностранен тест:

$$p_{value,1}^{1T} = 1 - CDF_{\mathsf{i}_{1:2}^{VAR}}\left(\mathsf{a}_{1:2,re}^{VAR,u} \right) + P\left(\mathsf{i}_{1:2}^{VAR} = \mathsf{a}_{1:2,re}^{VAR,u} \right) = \frac{1}{N_b} \sum_{\substack{q=1 \\ \mathsf{a}_{1:2,q}^{VAR,s} \ge \mathsf{a}_{1:2,re}^{VAR,u}}}^{N_b} 1 \qquad (2.1.1.4.19)$$

При втория едностранен тест:

$$p_{value,2}^{1T} = CDF_{\mathsf{i}_{1:2}^{VAR}}\left(\mathsf{a}_{1:2,re}^{VAR,d} \right) = \frac{1}{N_b} \sum_{\substack{q=1 \\ \mathsf{a}_{1:2,q}^{VAR,s} \le \mathsf{a}_{1:2,re}^{VAR,d}}}^{N_b} 1 \qquad (2.1.1.4.20)$$

При двустранния тест:

$$p_{value}^{2T} = CDF_{\mathsf{i}_{1:2}^{VAR}}\left(\mathsf{a}_{1:2,re}^{VAR,u} \right) - CDF_{\mathsf{i}_{1:2}^{VAR}}\left(\mathsf{a}_{1:2,re}^{VAR,d} \right) - P\left(\mathsf{i}_{1:2}^{VAR} = \mathsf{a}_{1:2,re}^{VAR,u} \right) =$$

$$= 1 - \frac{1}{N_b} \sum_{\substack{q=1 \\ \mathsf{a}_{1:2,re}^{VAR,d} < \mathsf{a}_{1:2,q}^{VAR,s} < \mathsf{a}_{1:2,re}^{VAR,u}}}^{N_b} 1 \qquad (2.1.1.4.21)$$

Алгоритъм 2.1.1.4 за изчисляване на трите p_{value} при Буутстрап тестове за равенство на дисперсии

1. Изчисли дисперсията VAR_1 на Извадка 1 по (2.1.1.4.1)
2. Изчисли дисперсията VAR_2 на Извадка 2 по (2.1.1.4.2)

3. Ако $VAR_1>0$ и $VAR_2=0$, то положи $p_{value}^{2T}=0$ за двустранен статистически тест, положи $p_{value}^{1T}=0$ за едностранен статистически тест, дефинирай H_1 за едностранен тест 'дисперсията на Популация 1 е по-голяма от тази на Популация 2' и край

4. Ако $VAR_1=0$ и $VAR_2>0$, то положи $p_{value}^{2T}=0$ за двустранен статистически тест, положи $p_{value}^{1T}=0$ за едностранен статистически тест, дефинирай H_1 за едностранен тест 'дисперсията на Популация 1 е по-малка от тази на Популация 2' и край

5. Ако $VAR_1=0$ и $VAR_2=0$, то положи $p_{value}^{2T}=1$ за двустранен статистически тест, положи $p_{value}^{1T}=1$ за едностранен статистически тест и край

6. Изчисли реалната тестова статистика $a_{1:2,re}^{VAR}$ по (2.1.1.4.5)

7. Ако $a_{1:2,re}^{VAR}=1$, то положи $p_{value}^{2T}=1$ за двустранен статистически тест, положи $p_{value}^{1T}=0.5$ за едностранен статистически тест и край

8. Изчисли общата дисперсия *VAR* на наблюденията в двете извадки по (2.1.1.4.6)

9. Формирай първата мащабирана извадка Z_m^1 по (2.1.1.4.7)

10. Формирай втората мащабирана извадка Z_m^2 по (2.1.1.4.9)

11. Избери броя на псевдо-реалностите N_b

12. Инициирай цикъл по псевдо-реалности: q=1

13. Формирай първа синтетична извадка $Z_q^{1,s}$ в q-тата псевдореалност:

13.1. Инициирай цикъл по номер на наблюдение на първа синтетична извадка: k=1

13.2. Изтегли с връщане от Z_m^1, k-тото наблюдение на $Z_q^{1,s}$:

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

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

13.2.3. Определи k-тото изтеглено наблюдение $z_{k,q}^{1,s}=z_{m,numt}^1$

13.3. Премини към следващото изтеглено наблюдение: $k=k+1$

13.4. Провери дали $Z_q^{1,s}$ е попълнено: ако $k\le n_1$, то иди към точка 13.2

13.5. Запомни първата синтетична извадка $Z_q^{1,s}=\left\{z_{q,1}^{1,s},z_{2,q}^{1,s},\ldots,z_{n_1,q}^{1,s}\right\}$

14. Изчисли дисперсията VAR_1^s на първа синтетична извадка от q-тата псевдореалност (2.1.1.4.13)

15. Формирай втора синтетична извадка $Z_q^{2,s}$ в q-тата псевдореалност :

15.1. Инициирай цикъл по номер на наблюдение на втора синтетична извадка: k=1

15.2. Изтегли с връщане от Z_m^2, k-тото наблюдение на $Z_q^{2,s}$:

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

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

15.2.3. Определи k-тото изтеглено наблюдение $z_{k,q}^{2,s}=z_{m,numt}^2$

15.3. Премини към следващото изтеглено наблюдение: $k=k+1$

15.4. Провери дали $Z_q^{2,s}$ е попълнено: ако $k\le n_2$, то иди към точка 15.2

15.5. Запомни втората синтетична извадка $Z_q^{2,s}=\left\{z_{q,1}^{2,s},z_{2,q}^{2,s},\ldots,z_{n_2,q}^{2,s}\right\}$

16. Изчисли дисперсията VAR_2^s на втора синтетична извадка от q-тата псевдо реалност (2.1.1.4.14)

17. Изчисли $a_{1:2,q}^{VAR,s}$ в q-тата псевдо-реалност по (2.1.1.4.15)

18. Мини към следващата псевдо-реалност: $q=q+1$

19. Провери за край на симулацията: ако $q\le N_b$, то иди към точка 13

20. Запомни синтетичните стойности на тестовите статистики $S_{i_{1:2}^{VAR,s}}=\left\{a_{1:2,1}^{VAR,s},a_{1:2,2}^{VAR,s},\ldots,a_{1:2,N_b}^{VAR,s}\right\}$

21. Изчисли горната помощна граница $a_{1:2,re}^{VAR,u}$ за реалната извадкова Фишерова статистика по (2.1.1.4.17)

22. Изчисли долната помощна граница $a_{1:2,re}^{VAR,d}$ за реалната извадкова Фишерова статистика по (2.1.1.4.18)

23. Изчисли p_{value}^{2T} за двустранен статистически тест по (2.1.1.4.21)

24. Ако $VAR_1>VAR_2$, то изчисли p_{value}^{1T} за едностранен статистически тест по формулата за $p_{value,1}^{1T}$ (2.1.1.4.19) и дефинирай H_1 за едностранен тест 'дисперсията на Популация 1 е по-голяма от тази на Популация 2'

25. Ако $VAR_1<VAR_2$, то изчисли p_{value}^{1T} за едностранен статистически тест по формулата за $p_{value,2}^{1T}$ (2.1.1.4.20), дефинирай H_1 'дисперсията на Популация 1 е по-малка от тази на Популация 2' и край

По сходни процедури се провеждат и Буутстрап тестовете за равенство на популационни средни, на популационни медиани и на популационни интерквартилни обхвати.

2.1.2. Тестове от втори кластер

Нека са налице две едномерни извадки от дискретен параметър Z - Извадка 1 и Извадка 2:

$$Z^1=\left\{z_1^1,z_2^1,...,z_{n_1}^1\right\} \tag{2.1.2.1}$$

$$Z^2=\left\{z_1^2,z_2^2,...,z_{n_2}^2\right\} \tag{2.1.2.2}$$

Нека измерените наблюдения в Извадка 1 и в Извадка 2 са представители на Популация 1 и Популация 2. Целта е да се определи дали Популация 1 има същите характеристики като Популация 2. Нека дискретния параметър Z има T различни дискрети, номерирани с 1, 2, ..., T. Нека множеството $G=\{g_1, g_2, \ldots, g_t\}$, да съдържа номерата на всички дискрети, срещащи се в Извадка 1 и/или в Извадка 2. Ако t е 1, то всички обекти и в двете извадки съдържат един и същ дискрет. Тогава дискретният параметър Z е тривиално статистически неразличим в двете популации. Затова може да се разгледа само случаят, когато t е поне 2.

Закона на разпределението на дискретната случайна величина Z с T дискрета в дадена популация може да се изрази чрез вероятностна масова функция на разпределение p(.):

$$p\left(z\right)=P\left(Z=z\right)=\begin{cases}p_z & \text{, за z}\in\{1,2,\ldots,\text{T}\}\\ 0 & \text{, за z}\notin\{1,2,\ldots,\text{T}\}\end{cases}, \tag{2.1.2.3}$$

$$p\left(g_r\right)=P\left(Z=g_r\right)=p_{g_r}, \qquad \text{за } r=1, 2, \ldots, t \tag{2.1.2.4}$$

2.1.2.1. Буутстрап Pearson тест за еднаквост на дискретни популационни разпределения

Една от най-разпространените апроксимации по извадка на популационна вероятностна масова функция на разпределение p(.) на дискретна променлива е емпиричната масова функция на разпределение (EPMF), където вероятностите p_{g_r} в p(.) за всяко от събитията $Z=g_r$ (r=1,2,…,t) са оценени и заменени с относителните честоти $\neg_{g_r}$ за същите събития в извадката:

$$PMF_1\left(g_r\right)=\neg_{g_r,1}=\frac{1}{n_1}\sum_{\substack{k=1\\ z_k^1=g_r}}^{n_1}1, \qquad \text{за } r=1, 2, \ldots, t \tag{2.1.2.1.1}$$

$$PMF_2\left(g_r\right)=\neg_{g_r,2}=\frac{1}{n_2}\sum_{\substack{k=1\\ z_k^2=g_r}}^{n_2}1, \qquad \text{за } r=1, 2, \ldots, t \tag{2.1.2.1.2}$$

Наблюденията в извадките са i.i.d. Задачата е да се провери дали вероятностните масови функции на разпределение p_1(.) и p_2(.) на случайната дискретна величина Z се отличават в двете популации. Нулевата хипотеза H_0 е, че дискретните разпределения на двете популации са еднакви, а алтернативната хипотеза H_1 е, че тези разпределения са различни.

Нека две масови функции на разпределение на дискретна случайна величина Z с t дискрета са означени с PMF'(.) и PMF''(.).За измерител на различията между две произволни масови функции на разпределение PMF'(.) и PMF''(.) е предложен функционалът Pearson статистика (PN):

$$PN=PN\left(PMF',PMF''\right)=$$

$$=\sum_{r=1}^{t}\frac{n'\left(PMF'\left(g_r\right)-PMF_{,\cup,}\left(g_r\right)\right)^2}{PMF_{,\cup,}\left(g_r\right)}+\sum_{r=1}^{t}\frac{n''\left(PMF''\left(g_r\right)-PMF_{,\cup,}\left(g_r\right)\right)^2}{PMF_{,\cup,}\left(g_r\right)} \tag{2.1.2.1.3}$$

$$PMF_{,\cup,}\left(g_r\right)=\frac{n'PMF'\left(g_r\right)+n''PMF''\left(g_r\right)}{n'+n''}, \quad \text{за } r=1, 2, \ldots, t \tag{2.1.2.1.4}$$

Ако функционалът PN(.,.) се пресметне за емпиричните масови функции на разпределение PMF_1(.) и PMF_2(.) апроксимирани от двете дискретни извадки (2.1.2.1) и (2.1.2.2), то

$$pn_{re} = PN\left(PMF_1, PMF_2\right) =$$

$$= \sum_{r=1}^{t} \frac{n_1\left(PMF_1\left(g_r\right) - PMF_{1\cup 2}\left(g_r\right)\right)^2}{PMF_{1\cup 2}\left(g_r\right)} + \sum_{r=1}^{t} \frac{n_2\left(PMF_2\left(g_r\right) - PMF_{1\cup 2}\left(g_r\right)\right)^2}{PMF_{1\cup 2}\left(g_r\right)} \qquad (2.1.2.1.5)$$

$$PMF_{1\cup 2}\left(g_r\right) = \frac{n_1 PMF_1\left(g_r\right) + n_2 PMF_2\left(g_r\right)}{n_1 + n_2}, \qquad \text{за } r=1, 2, \ldots, t \qquad (2.1.2.1.6)$$

В (2.1.2.1.6), $PMF_{1\cup 2}\left(.\right)$ е емпиричната масова функция на разпределение получена при предпоставка, че двете дискретни извадки Z^1 и Z^2 са събрани в една обща извадка $Z^{1\cup 2}$:

$$Z^{1\cup 2} = \left\{z_1^1, z_2^1, \ldots, z_{n_1}^1, z_1^2, z_2^2, \ldots, z_{n_2}^2\right\} = \left\{z_1^{1\cup 2}, z_2^{1\cup 2}, \ldots, z_{n_1+n_2}^{1\cup 2}\right\} \qquad (2.1.2.1.7)$$

В случай, че популационните разпределения са апроксимирани с емпирични масови функции на разпределение (2.1.2.1.1) и (2.1.2.1.2), то пресмятането на реалната оценка pn_{re} на PN може да се опрости до:

$$pn_{re} = PN\left(PMF_1, PMF_2\right) = \sum_{r=1}^{t} \frac{n_1\left(\neg_{g_r,1} - \neg_{g_r,1\cup 2}\right)^2}{\neg_{g_r,1\cup 2}} + \sum_{r=1}^{t} \frac{n_2\left(\neg_{g_r,2} - \neg_{g_r,1\cup 2}\right)^2}{\neg_{g_r,1\cup 2}}, \qquad (2.1.2.1.8)$$

$$\neg_{g_r,1\cup 2} = \frac{n_1 \neg_{g_r,1} + n_2 \neg_{g_r,2}}{n_1 + n_2}, \qquad \text{за } r=1, 2, \ldots, t \qquad (2.1.2.1.9)$$

След заместване на (2.1.2.1.9) в (2.1.2.1.8) се получава:

$$pn_{re} = n_1 n_2 \sum_{r=1}^{t} \frac{\left(\neg_{g_r,1} - \neg_{g_r,2}\right)^2}{n_1 \neg_{g_r,1} + n_2 \neg_{g_r,2}} \qquad (2.1.2.1.10)$$

Ако се заместят (2.1.2.1.1), (2.1.2.1.2) в (2.1.2.1.10) то

$$pn_{re} = \frac{1}{n_1 n_2} \sum_{r=1}^{t} \frac{\left(n_2 \sum\limits_{\substack{k=1 \\ z_k^1 = g_r}}^{n_1} 1 - n_1 \sum\limits_{\substack{k=1 \\ z_k^2 = g_r}}^{n_2} 1\right)^2}{\sum\limits_{\substack{k=1 \\ z_k^1 = g_r}}^{n_1} 1 + \sum\limits_{\substack{k=1 \\ z_k^2 = g_r}}^{n_2} 1} \qquad (2.1.2.1.11)$$

Ако H_0 е вярна, то Популации 1 и 2 са с еднакви вероятностни масови функции на разпределение, които са поотделно еднакви с $p_{tr}(.)=p_1(.)=p_2(.)$, където:

$$p_{tr}\left(g_r\right) = P\left(Z = g_r\right) = p_{g_r,tr}, \quad \text{за } r=1, 2, \ldots, t \qquad (2.1.2.1.12)$$

Нека вероятностните масови функции на разпределение в двете популации са еднакви и равни на $p_{tr}(.)$. Тези две еднакви разпределения не са известни на експериментатора, но се реализират статистически чрез двойката дискретни извадки ($Z^1 - Z^2$). От $p_{tr}(.)$ биха могли да се генерират безброй много *хипотетични* двойки извадки, които съдържат съответно n_1 и n_2 наблюдения: $\left(Z_1^1 - Z_1^2\right)$, $\left(Z_2^1 - Z_2^2\right)$, $\left(Z_3^1 - Z_3^2\right)$, ... Експериментаторът би могъл да изчисли хипотетичните реализации pn_1, pn_2, pn_3, ... на Pearson статистиката

Разликата $PN–pn_{tr}=PN–0=PN$ е случайна величина, наречена *root* и неговото разпределение описва изцяло количествената несигурност в реалната тестова статистика pn_{re}. Задача тук е да се намери разпределението на *root* без да е известно $p_{tr}(.)$ и без да са известни хипотетичните двойки извадки $\left(Z_1^1 - Z_1^2\right)$, $\left(Z_2^1 - Z_2^2\right)$, $\left(Z_3^1 - Z_3^2\right)$, ... Ако H_0 е вярно, то $p_{tr}(.)$ е реализирано статистически (n_1+n_2) пъти в извадките (2.1.2.1) и (2.1.2.2). Двете извадки Z^1 и Z^2 са обединени в обща дискретна извадка $Z^{1\cup 2}$ (2.1.2.1.7), съдържаща $(n_1 + n_2)$ наблюдения. Тогава, $p_{tr}(.)$ може да се апроксимира с емпиричната масова функция на разпределение $PMF_{1\cup 2}\left(.\right)$ от (2.1.2.1.6), построена върху общата дискретна извадка $Z^{1\cup 2}$

В компютърния свят, истинските $p_{tr}(.)$ на двете популации са заменени с $CDF_{1\cup 2}\left(.\right)$ от (2.1.2.1.16) и

$$pn_{re,cw} = PN\left(PMF_{1\cup 2}, PMF_{1\cup 2}\right) =$$

$$= \sum_{r=1}^{t} \frac{n_1\left(PMF_{1\cup 2}(g_r) - PMF_{1\cup 2}(g_r)\right)^2}{PMF_{1\cup 2}(g_r)} + \sum_{r=1}^{t} \frac{n_2\left(PMF_{1\cup 2}(g_r) - PMF_{1\cup 2}(g_r)\right)^2}{PMF_{1\cup 2}(g_r)} = \quad (2.1.2.1.13)$$

$$= \sum_{r=1}^{t} \frac{n_1(0)^2}{PMF_{1\cup 2}(g_r)} + \sum_{r=1}^{t} \frac{n_2(0)^2}{PMF_{1\cup 2}(g_r)} = 0 + 0 = 0$$

В компютърния свят от известното $PMF_{1\cup 2}(.)$ може да се генерират N_b синтетични двойки извадки $\left(Z_q^{1,s} - Z_q^{2,s}\right)$ за q=1,2,…,N_b чрез теглене чрез връщане от $Z^{1\cup 2}$ на двойки синтетични извадки:

$$Z_q^{1,s} = \left\{z_{1,q}^{1,s}, z_{2,q}^{1,s}, ..., z_{n_1,q}^{1,s}\right\} \quad (2.1.2.1.14)$$

$$Z_q^{2,s} = \left\{z_{1,q}^{2,s}, z_{2,q}^{2,s}, ..., z_{n_2,q}^{2,s}\right\} \quad (2.1.2.1.15)$$

За q-тата двойка извадки $\left(Z_q^{1,s} - Z_q^{2,s}\right)$ по (2.1.2.1.11) се изчислява синтетичната Pearson статистика pn_q^s:

$$pn_q^s = \frac{1}{n_1 n_2} \sum_{r=1}^{t} \frac{\left(n_2 \sum_{\substack{k=1 \\ z_{k,q}^{1,s}=g_r}}^{n_1} 1 - n_1 \sum_{\substack{k=1 \\ z_{k,q}^{2,s}=g_r}}^{n_2} 1\right)^2}{\sum_{\substack{k=1 \\ z_k^1=g_r}}^{n_1} 1 + \sum_{\substack{k=1 \\ z_k^2=g_r}}^{n_2} 1} \quad (2.1.2.1.16)$$

Получава се синтетична извадка $S_{PN^s} = \left\{pn_1^s, pn_2^s, ..., pn_{N_b}^s\right\}$ от N_b реализации на PN^s. Ако N_b е голямо число (например 100000), то кумулативното разпределение на непрекъснатата случайна величина PN^s може да се апроксимира с много голяма точност, като ECDF, построена върху синтетичната извадка S_{PN^s}:

$$CDF_{PN^s}(pn) = \frac{1}{N_b} \sum_{\substack{q=1 \\ pn_q^s \le pn}}^{N_b} 1 \quad \text{, за } pn \in (-\infty; +\infty) \quad (2.1.2.1.17)$$

Разликата $PN^s - pn_{tr,cw} = PN^s - 0 = PN^s$ е случайна величина, наречена $root_{cv}$. Търсената апроксимация $CDF_{PN}(.)$ на условното разпределение на PN, ако H_0 е вярна, съвпада с вече построената в (2.1.2.1.13) апроксимация на разпределението на PN^s. При разглеждания тест, търсеното p_{value} е:

$$p_{value} = 1 - CDF_{PN}(pn_{re}) + P(PN = pn_{re}) = \frac{1}{N_b} \sum_{\substack{q=1 \\ pn_q^s \ge pn_{re}}}^{N_b} 1 \quad (2.1.2.1.18)$$

2.1.2.2. Буутстрап тестове за равенство на популационни вероятности за попадане в дискрет g_r

Една от най-разпространените оценки по извадка на популационната вероятност p_{g_r} за събитието 'дискретната случайна величина Z да заеме дискрет номер g_r' е извадковата относителна честота $\neg_{g_r}$ на същото събитие $Z=g_r$. Съответните стойности на $\neg_{g_r}$ за извадките (2.1.2.1) и (2.1.2.2) са:

$$\neg_{g_{r,1}} = \frac{1}{n_1} \sum_{\substack{k=1 \\ z_k^1=g_r}}^{n_1} 1 \quad (2.1.2.2.1)$$

$$\neg_{g_{r,2}} = \frac{1}{n_2} \sum_{\substack{k=1 \\ z_k^2=g_r}}^{n_2} 1 \quad (2.1.2.2.2)$$

Оценките (2.1.2.2.1) и (2.1.2.2.2) са състоятелни, неизместени и ефективни, ако наблюденията в извадките са i.i.d. Задачата е да се провери дали в двете популации се отличават вероятности $p_{g_r,1}$ и $p_{g_r,2}$ за събитието 'дискретната случайна величина Z да заеме дискрет g_r'. Нулевата хипотеза $H_0^{g_r}$ е, че популационните вероятности за $Z=g_r$ са равни в двете популации ($p_{g_r,1} = p_{g_r,2}$). Алтернативните хипотези $H_1^{g_r}$ са различни за двустранен Буутстрап тест за дискретни вероятности и за едностранните Буутстрап

тестове за дискретни вероятности. При двустранния тест, $H_1^{g_r}$ е, че $p_{g_r,1} \neq p_{g_r,2}$. При Първия едностранен тест $H_1^{g_r}$ е, че $p_{g_r,1} > p_{g_r,2}$. При Втория едностранен тест $H_1^{g_r}$ е, че $p_{g_r,1} < p_{g_r,2}$. В настоящата разработка, ако $\neg_{g_r,1} > \neg_{g_r,2}$, то за едностранен тест се приема Първият едностранен тест, а ако $\neg_{g_r,1} < \neg_{g_r,2}$, то за едностранен тест се приема Вторият едностранен тест. В много редкия случай, когато $\neg_{g_r,1} = \neg_{g_r,2}$, то нито Първият, нито Вторият едностранен тест се извършва, техните p_{value} се фиксират на 0.5, а който и да е от двата теста се нарича едностранен тест. Двустранният тест също не се извършва, а неговото p_{value} се полага за 1.

За измерител на различията между две произволни популационни вероятности за $Z=g_r$ се използва функцията разлика между две относителни честоти:

$$\mathfrak{a}_{1-2}^{p,g_r} = \mathfrak{a}_{1-2}^{p,g_r}(p', p'') = p' - p'' \qquad (2.1.2.2.3)$$

Реалната стойност на тестовата статистика е:

$$¢_{1-2,re}^{p,g_r} = \frac{1}{n_1} \sum_{\substack{k=1 \\ z_k^1 = g_r}}^{n_1} 1 - \frac{1}{n_2} \sum_{\substack{k=1 \\ z_k^2 = g_r}}^{n_2} 1 \qquad (2.1.2.2.5)$$

Нека вероятностните масови функции на разпределение на двете популации $p_1(.)$ и $p_2(.)$ са с еднакви вероятности за попадане в дискрет g_r, равни на $p_{g_r,tr}$. Тези две разпределения не са известни на експериментатора, но се реализират статистически чрез двойката извадки (Z^1 – Z^2). По извадките може чрез (2.1.2.2.5) да се изчисли реалната оценка $¢_{1-2,re}^{p,g_r}$ на статистиката разлика между извадкови средни стойности. От дискретните разпределения $p_1(.)$ и $p_2(.)$ биха могли да се генерират безброй много *хипотетични* двойки извадки: $\left(Z_1^1 - Z_1^2\right)$, $\left(Z_2^1 - Z_2^2\right)$, $\left(Z_3^1 - Z_3^2\right)$, ... Експериментаторът би могъл да изчисли хипотетичните реализации $¢_{1-2,1}^{p,g_r}, ¢_{1-2,2}^{p,g_r}, ¢_{1-2,3}^{p,g_r}, \ldots$ на тестовата статистика по всяка от двойките хипотетични извадки.

Разликата $\mathfrak{a}_{1-2}^{p,g_r} - ¢_{1-2,tr}^{p,g_r} = \mathfrak{a}_{1-2}^{p,g_r} - 0 = \mathfrak{a}_{1-2}^{p,g_r}$ е *root* и неговото разпределение описва изцяло количествената несигурност в реалната тестова статистика $¢_{1-2,re}^{p,g_r}$. Задачата тук е да се намери разпределението на *root* без да са известни $p_1(.)$, $p_2(.)$ и $p_{g_r,tr}$ и без да са известни хипотетичните двойки извадки $\left(Z_1^1 - Z_1^2\right)$, $\left(Z_2^1 - Z_2^2\right)$, $\left(Z_3^1 - Z_3^2\right)$, ...

Ако $H_0^{g_r}$ е вярно, то неизвестните вероятности за Z=g_r в двете популации $p_{g_r,tr}$ могат да се заменят с относителната честота $\neg_{g_r}$ на събитието Z=g_r, пресметната по наблюденията в двете извадки:

$$\neg_{g_r} = \frac{n_1 \neg_{g_r,1} + n_2 \neg_{g_r,2}}{n_1 + n_2} = \frac{\sum_{\substack{k=1 \\ z_k^1 = g_r}}^{n_1} 1 + \sum_{\substack{k=1 \\ z_k^2 = g_r}}^{n_2} 1}{n_1 + n_2} \qquad (2.1.2.2.6)$$

Ако $H_0^{g_r}$ е вярно, то $p_1(.)$ е реализирано статистически n_1 пъти в извадката Z^1 от (2.1.2.1), но променена така, че относителната честота за попадане в дискрет g_r вместо $\neg_{g_r,1}$ да стане $\neg_{g_r}$, което е най-добрата оценка на неизвестната $p_{g_r,tr}$. Ако броят на дискретите в Z^1 и Z^2 е t>2, то ще съществуват безкрайно много двойки от емпирични масови функции $PMF_1(.)$ и $PMF_2(.)$, които да апроксимират популационните масови функции $p_1(.)$, $p_2(.)$, така че да са изпълнени гореуказаните условия:

$$PMF_1(g_r) = PMF_2(g_r) = \neg_{g_r} \qquad (2.1.2.2.7)$$

Една такава двойка е две еднакви емпирични масови функции $PMF_1(.) = PMF_2(.) = PMF_{1\cup 2}(.)$, апроксимиращи обединените в една обща извадка $Z^{1\cup 2}$ наблюдения от двете извадки Z^1 и Z^2:

$$Z^{1\cup 2} = \left\{z_1^1, z_2^1, \ldots, z_{n_1}^1, z_1^2, z_2^2, \ldots, z_{n_2}^2\right\} = \left\{z_1^{1\cup 2}, z_2^{1\cup 2}, \ldots, z_{n_1+n_2}^{1\cup 2}\right\} \qquad (2.1.2.2.8)$$

Тогава, емпиричната масова функция, построена върху общата извадка (2.1.2.2.8) ще е:

$$PMF_{1\cup 2}(g_r) = \frac{1}{n_1 + n_2} \sum_{\substack{k=1 \\ z_k^{1\cup 2} = g_r}}^{n_1+n_2} 1, \text{ за } r=1, 2, \ldots, t \qquad (2.1.2.2.9)$$

В компютърния свят от известната $PMF_{1\cup 2}(.)=PMF_1(.)=PMF_1(.)$ може да се генерират N_b синтетични двойки извадки $\left(Z_q^{1,s}-Z_q^{2,s}\right)$ за q=1,2,…,N_b. Генерацията от $PMF_{1\cup 2}(.)$ се свежда до теглене чрез връщане от общата извадка $Z^{1\cup 2}$ на двойки синтетични извадки, съдържащи съответно n_1 и n_2 наблюдения:

$$Z_q^{1,s}=\left\{z_{1,q}^{1,s},z_{2,q}^{1,s},...,z_{n_1,q}^{1,s}\right\} \quad (2.1.2.2.10)$$

$$Z_q^{2,s}=\left\{z_{1,q}^{2,s},z_{2,q}^{2,s},...,z_{n_2,q}^{2,s}\right\} \quad (2.1.2.2.11)$$

За q-тата двойка извадки $\left(Z_q^{1,s}-Z_q^{2,s}\right)$ по (2.1.2.2.5) се изчислява синтетичната тестова статистика, която е една реализация на случайната величина $\mathsf{a}_{1-2}^{p,g_r,s}$ в компютърния свят:

$$¢_{1-2,q}^{p,g_r,s}=\frac{1}{n_1}\sum_{\substack{k=1\\ z_{k,q}^{1,s}=g_r}}^{n_1}1-\frac{1}{n_2}\sum_{\substack{k=1\\ z_{k,q}^{2,s}=g_r}}^{n_2}1 \quad (2.1.2.2.12)$$

Получава се синтетична извадка $S_{\mathsf{a}_{1-2}^{p,g_r,s}}=\left\{¢_{1-2,1}^{p,g_r,s},¢_{1-2,2}^{p,g_r,s},...,¢_{1-2,N_b}^{p,g_r,s}\right\}$ от N_b реализации на $\mathsf{a}_{1-2}^{p,g_r,s}$. Ако N_b е голямо число (например 100000), то кумулативното разпределение на случайната величина $\mathsf{a}_{1-2}^{p,g_r,s}$ може да се апроксимира с много голяма точност, като ECDF, построена върху синтетичната извадка $S_{\mathsf{a}_{1-2}^{p,g_r,s}}$:

$$CDF_{\mathsf{a}_{1-2}^{p,g_r,s}}(¢)=\frac{1}{N_b}\sum_{\substack{q=1\\ ¢_{1-2,q}^{p,g_r,s}\le ¢}}^{N_b}1 \quad \text{, за } ¢\in(-\infty;+\infty) \quad (2.1.2.2.13)$$

Разликата $\mathsf{a}_{1-2}^{p,g_r,s}-¢_{1-2,tr,cw}^{p,g_r,s}=\mathsf{a}_{1-2}^{p,g_r,s}-0=\mathsf{a}_{1-2}^{p,g_r,s}$ е $root_{cv}$. Нека се въведат горна и долна помощни граници за реалната статистика разлика между относителни честоти за $Z=g_r$:

$$¢_{1-2,re}^{p,g_r,u}=max\left\{¢_{1-2,re}^{p,g_r},-¢_{1-2,re}^{p,g_r}\right\} \quad (2.1.2.2.14)$$

$$¢_{1-2,re}^{p,g_r,d}=min\left\{¢_{1-2,re}^{p,g_r},-¢_{1-2,re}^{p,g_r}\right\} \quad (2.1.2.2.15)$$

Извеждат се формули за Pvalue. При първия едностранен тест:

$$p_{value,1}^{1T,g_r}=1-CDF_{\mathsf{a}_{1-2}^{M}}\left(¢_{1-2,re}^{p,g_r,u}\right)+P\left(\mathsf{a}_{1-2}^{p,g_r}=¢_{1-2,re}^{p,g_r,u}\right)=\frac{1}{N_b}\sum_{\substack{q=1\\ ¢_{1-2,q}^{p,g_r,s}\ge ¢_{1-2,re}^{p,g_r,u}}}^{N_b}1 \quad (2.1.2.2.16)$$

При втория едностранен тест,

$$p_{value,2}^{1T,g_r}=CDF_{\mathsf{a}_{1-2}^{p,g_r}}\left(¢_{1-2,re}^{p,g_r,d}\right)=\frac{1}{N_b}\sum_{\substack{q=1\\ ¢_{1-2,q}^{p,g_r,s}\le ¢_{1-2,re}^{p,g_r,d}}}^{N_b}1 \quad (2.1.1.2.17)$$

При двустранния тест,

$$p_{value}^{2T,g_r}=CDF_{\mathsf{a}_{1-2}^{p,g_r}}\left(¢_{1-2,re}^{p,g_r,u}\right)-CDF_{\mathsf{a}_{1-2}^{p,g_r}}\left(¢_{1-2,re}^{p,g_r,d}\right)-P\left(\mathsf{a}_{1-2}^{p,g_r}=¢_{1-2,re}^{p,g_r,u}\right)=$$

$$=1-\frac{1}{N_b}\sum_{\substack{q=1\\ ¢_{1-2,re}^{p,g_r,d}<¢_{1-2,q}^{p,g_r,s}<¢_{1-2,re}^{p,g_r,u}}}^{N_b}1 \quad (2.1.2.2.18)$$

2.1.2.3. Изчисляване на всички (3t+1) pvalue при Буутстрап тестове от втори кластер

Всички тестови статистики на тестовете от втори кластер се изчислят в една и съща псевдо реалност.

Алгоритъм 2.1.2.3. за изчисляване на p_{value} за всички (3t+1) Буутстрап тестове от втори кластер

1. Преброй броят t на наличните дискрети в извадките (2.1.2.1) и (2.1.2.2)
2. Формирай множеството на наличните дискрети $G=\{g_1, g_2, \ldots, g_t\}$ така, че да съдържа номерата на всички дискрети, срещащи се в Извадка 1 и/или в Извадка 2 дадени в (2.1.2.1) и (2.1.2.2)
3. Ако t=1, то положи за търсените p_{value} както следва:
   3.1. За Буутстрап Pearson тест: p_{value}=1
   3.2. За двустранен Буутстрап тест за равенство на вероятности за Z=g_1: p_{value}^{2T,g_1} =1

3.3. За едностранен Буутстрап тест за равенство на вероятности за Z=g_1: p_{value}^{1T,g_1}=0.5

3.4. Край

4. Инициирай цикъл по дискрети: r=1

5. Пресметни извадковата относителна честота $\neg_{g_r,1}$ на събитие $Z=g_r$ от дискретна Извадка 1 (2.1.2.1) по (2.1.2.2.1)

6. Пресметни извадковата относителна честота $\neg_{g_r,2}$ на събитие $Z=g_r$ от дискретна Извадка 2 (2.1.2.2) по (2.1.2.2.2)

7. Изчисли реалната тестова статистика $¢_{1-2,re}^{p,g_r}$ разлика между относителни честоти за $Z=g_r$ по (2.1.2.2.4)

8. Изчисли горната помощна граница $¢_{1-2,re}^{p,g_r,u}$ за реалната извадкова статистика разлика между относителни честоти за $Z=g_r$ по (2.1.2.2.14)

9. Изчисли долната помощна граница $¢_{1-2,re}^{p,g_r,d}$ за реалната извадкова статистика разлика между относителни честоти за $Z=g_r$ по (2.1.2.2.15)

10. Премини към следващ дискрет: r=r+1

11. Провери дали всички дискрети са обработени: ако $r \le t$, то иди към точка 5

12. Изчисли реалната оценка на Pearson статистика pn_{re} по (2.1.2.1.10)

13. Ако pn_{re}=1, то положи за търсените p_{value} както следва:

13.1. За Буутстрап Pearson тест: p_{value}=1

13.2. Инициирай цикъл по дискрети: r=1

13.3. За двустранен Буутстрап тест за равенство на вероятности за Z=g_r: p_{value}^{2T,g_r}=1

13.4. За едностранен Буутстрап тест за равенство на вероятности за Z=g_r: p_{value}^{1T,g_r}=0.5

13.5. Премини към следващ дискрет: r=r+1

13.6. Провери дали всички дискрети са обработени: ако $r \le t$, то иди към точка 13.3

13.7. Край

14. Формирай обща извадка $Z^{1\cup 2}$ по (2.1.2.2.8)

15. Избери броя N_b на псевдо-реалностите

16. Инициирай цикъл по псевдо-реалности: q=1

17. Формирай първа синтетична извадка $Z_q^{1,s}$ в q-тата псевдо реалност:

17.1. Инициирай цикъл по номер на наблюдение на първа синтетична извадка: k=1

17.2. Изтегли с връщане от $Z^{1\cup 2}$ k-тото наблюдение на $Z_q^{1,s}$:

17.2.1. Генерирай случайно равномерно разпределено число $rand \in (0;1]$

17.2.2. Намери номера на k-тото изтеглено наблюдение $numt = \lceil rand \times (n_1 + n_2) \rceil$, където $\lceil x \rceil$ е най-малкото цяло число, не по-малко от x (т.е. $\lceil\ \rceil$ е оператор за закръгляне нагоре)

17.2.3. Определи k-тото изтеглено наблюдение $z_{k,q}^{1,s} = z_{numt}^{1\cup 2}$

17.3. Премини към следващото изтеглено наблюдение: k=k+1

17.4. Провери дали $Z_q^{1,s}$ е попълнено: ако $k \le n_1$, то иди към точка 17.2

17.5. Запомни първата синтетична извадка $Z_q^{1,s} = \left\{ z_{1,q}^{1,s}, z_{2,q}^{1,s}, ..., z_{n_1,q}^{1,s} \right\}$

18. Формирай втора синтетичната извадка $Z_q^{2,s}$ в q-тата псевдореалност:

18.1. Инициирай цикъл по номер на наблюдение на втора синтетична извадка: k=1

18.2. Изтегли с връщане от $Z^{1\cup 2}$ k-тото наблюдение на $Z_q^{2,s}$:

18.2.1. Генерирай случайно равномерно разпределено число $rand \in (0;1]$

18.2.2. Намери номера на k-тото изтеглено наблюдение $numt = \lceil rand \times (n_1 + n_2) \rceil$

18.2.3. Определи k-тото изтеглено наблюдение: $z_{k,q}^{2,s} = z_{numt}^{1\cup 2}$

18.3. Премини към следващо изтеглено наблюдение: k=k+1

18.4. Провери дали $Z_q^{2,s}$ е попълнено: ако $k \le n_2$, то иди към точка 18.2

18.5. Запомни втората синтетична извадка $Z_q^{2,s} = \left\{ z_{1,q}^{2,s}, z_{2,q}^{2,s}, ..., z_{n_2,q}^{2,s} \right\}$

19. Изчисли Pearson статистиката pn_q^s в q-тата псевдо-реалност по (2.1.2.1.16)

20. Инициирай цикъл по дискрети: r=1

21. Ако $\neg_{g_r,1} \neq \neg_{g_r,2}$, то изчисли тестовата статистика $¢_{1-2,q}^{p,g_r,s}$ в q-тата псевдо-реалност по (2.1.2.2.12)

22. Премини към следващ дискрет: r=r+1

23. Провери дали всички дискрети са обработени: ако $r \leq t$, то иди към точка 21

24. Мини към следваща псевдо-реалност: q=q+1

25. Провери за край на симулацията: ако $q \leq N_b$, то иди към точка 17

26. Запомни синтетичните стойности на тестовите Pearson статистики $S_{PN^s} = \left\{ pn_1^s, pn_2^s, ..., pn_{N_b}^s \right\}$

27. Инициирай цикъл по дискрети: r=1.

28. Ако $\neg_{g_r,1} \neq \neg_{g_r,2}$, то запомни $S_{a_{1-2}^{p,g_r,s}} = \left\{ ¢_{1-2,1}^{p,g_r,s}, ¢_{1-2,2}^{p,g_r,s}, ..., ¢_{1-2,N_b}^{p,g_r,s} \right\}$

29. Премини към следващ дискрет: r=r+1

30. Провери дали всички дискрети са обработени: ако $r \leq t$, то иди към точка 28

31. Изчисли p_{value} на Буутстрап Pearson теста по (2.1.2.1.18)

32. Инициирай цикъл по дискрети: r=1

33. Ако $\neg_{g_r,1} \neq \neg_{g_r,2}$, то изчисли p_{value}^{2T,g_r} за двустранен Буутстрап тест за еднаквост на популационни вероятности за попадане в дискрет номер g_r по (2.1.2.2.18)

34. Ако $\neg_{g_r,1} > \neg_{g_r,2}$, то изчисли p_{value}^{1T,g_r} за едностранен Буутстрап тест за еднаквост на популационни вероятности за попадане в дискрет номер g_r по формулата за $p_{value,1}^{1T,g_r}$ (2.1.2.2.16) и дефинирай $H_1^{g_r}$ за едностранен тест 'вероятността за Z=g_r при Популация 1 е по-голяма от тази на Популация 2'

35. Ако $\neg_{g_r,1} < \neg_{g_r,2}$, то изчисли p_{value}^{1T,g_r} за едностранен Буутстрап тест за еднаквост на популационни вероятности за попадане в дискрет номер g_r по формулата за $p_{value,2}^{1T,g_r}$ (2.1.2.2.17) и дефинирай $H_1^{g_r}$ за едностранен тест 'вероятността за Z=g_r при Популация 1 е по-малка от тази на Популация 2'

36. Ако $\neg_{g_r,1} = \neg_{g_r,2}$, то положи $p_{value}^{2T,g_r} = 1$ за двустранен Буутстрап тест за еднаквост на популационни вероятности за попадане в дискрет номер g_r и p_{value}^{1T,g_r} =0.5 за едностранен Буутстрап тест за еднаквост на популационни вероятности за попадане в дискрет номер g_r'

37. Премини към следващ дискрет: r=r+1

38. Провери дали всички дискрети са обработени: ако $r \leq t$, то иди към точка 33, иначе край

2.1.3. Тестове от трети кластер

Нека Извадка 1 е едномерна извадка Z^1, която съдържа n наблюдения на непрекъснат параметър Z за дадени обекти, обикновено измерени след настъпването на някакво събитие за всеки обект. Нека Извадка 2 е едномерна извадка Z^2, която съдържат стойностите на същия непрекъснат параметър Z за същите обекти, но измерени в друг момент от време, обикновено измерени преди настъпването на събитието за конкретния обект. Затова Извадка 1 и Извадка 2 се наричат сдвоени или зависими извадки:

$$Z^1 = \left\{ z_1^1, z_2^1, ..., z_n^1 \right\} \quad (2.1.3.1)$$

$$Z^2 = \left\{ z_1^2, z_2^2, ..., z_n^2 \right\} \quad (2.1.3.2)$$

Нека Извадка 3 е едномерна извадка Z^3 , която съдържа промяната aZ на стойностите на непрекъснатия параметър Z от Извадка 2 до Извадка 1. Затова Извадка 3 се нарича извадка на промяната:

$$Z^3 = \left\{ z_1^3, z_2^3, ..., z_n^3 \right\}, \text{ където } z_k^3 = z_k^1 - z_k^2, \text{ за } k=1, 2, \ldots, n \quad (2.1.3.3)$$

Нека обектите в Извадка 1, Извадка 2 и Извадка 3 са представители на една Популация. Целта е да се определи дали вследствие на настъпилото събитие между Извадка 2 и Извадка 1, има съществена промяна aZ в стойностите на непрекъснатия параметър Z за Популацията. За този кластер се провеждат тестове за промяна в средната стойност и медианата, които са сходни с вече представените.

2.2. Статистически изводи, базирани върху сравнения с псевдо контролни групи

2.2.1. Експериментално отчитане на статистически значими ефекти от въздействие чрез контролни групи

Нека се търси ефектът от дадено въздействие R върху избран параметър X, характеризиращ състоянието на обектите от някаква популация P. За решаването на такава задача най-често експериментално

се измерват стойностите на X преди и след въздействието R за дадена група от принадлежащи на P обекти, наречена *експериментална група*. Ако се разглежда само промяната на параметъра в експерименталната група, вследствие на въздействието, то съществува опасност да се получат неверни изводи. Поради това, в експерименталните изследвания се използва *контролна група*, съставена също от обекти, принадлежащи на P, които обаче не се подлагат на изследваното въздействие R. Има няколко методологии, които разрешават да се получават верни изводи от експериментални изследвания при използване на контролни и на експериментални групи. В разпространения случай, когато изследваното въздействие действа едновременно с друго базово въздействие B, чийто ефект не може да се преустанови по време на експеримента (например времето), то изследваният параметър на състоянието се измерва за двете групи преди и след въздействието. Получават се 4 извадки от данни, както следва:

1) E_0 съдържа стойностите на X за експерименталната група в началото на експеримента.
2) K_0 съдържа стойностите на X за контролната група в началото на експеримента.
3) E_e съдържаща стойностите на X за експерименталната група в края на експеримента, след като групата е подложена на базовото въздействие B и на изследваното въздействие R.
4) K_e съдържа стойностите на X за контролната група в края на експеримента, след като групата е подложена на базовото въздействие B.

2.2.2. Необходимост от псевдо контролни групи в медицинските изследвания

В някои експериментални изследвания, контролна група може просто да не съществува по най-разнообразни причини. Това е особено характерно за медицинските изследвания, при които се тества ефектът на лекарство или на медицинска процедура върху реални пациенти. Описани са три типични ситуации, където използването на класическа контролна група е проблематично като теория и спорно като практика.

2.2.3. Експериментално отчитане на статистически значими ефекти от въздействие чрез контролни групи

Нека се търси ефектът от дадено въздействие R, действащо едновременно с друго базово въздействие B, върху избран параметър X, характеризиращ състоянието на обектите от някаква популация P, наречена *целева*. Нека преди и след въздействията B и R, експериментално да са измерени стойностите на X за дадена група от принадлежащи на P обекти, наречена *експериментална група*. Търсеният ефект може да се установи в сравнение с ефекта от базовото въздействие B върху параметъра X, характеризиращ състоянието на обектите от някаква сходна на P популация Q, наречена *псевдо контролна*. Нека преди и след въздействието B, експериментално да са измерени стойностите на X за дадена група от принадлежащи на Q обекти, наречена *псевдо контролна* група.

Получават се 4 извадки от данни, както следва:

1) E_0 съдържа стойностите на X за експерименталната група в началото на експеримента.
2) PC_0 съдържа стойностите на X за псевдо контролната група в началото на експеримента.
3) E_e съдържа стойностите на X за експерименталната група в края на експеримента, след като групата е подложена на базовото въздействие B и на изследваното въздействие R.
4) PC_e съдържа стойностите на X за псевдо контролната група в края на експеримента, след като групата е подложена на базовото въздействие B.

Ефектът на изследваното въздействие R върху избрания параметър X ще се търси по два начина. При първия начин се тестват за еднаквост популациите P и Q както преди въздействията чрез извадките E_0 и PC_0, така и след въздействията чрез извадките E_e и PC_e. Сравнението между установените статистически разлики преди и след въздействията ще покаже един аспект от ефектът на R върху X. При вторият начин се тестват за нулевост както промяната на X вследствие на въздействията R и B при популацията P чрез сдвоените извадки E_0 и E_e, така и промяната на X вследствие на въздействието B при популацията Q чрез сдвоените извадки PC_0 и PC_e. Сравнението между статистически установените ненулевости на промените в двете популации ще покаже друг аспект от ефекта на R върху X.

2.2.4. Тестване за еднаквост на непрекъснат параметър в две популации по две извадки

Нека Извадка 1 да съдържа стойностите на непрекъснатия параметър X за обектите от експерименталната група, а Извадка 2 да съдържа стойностите на X за обектите от псевдо контролната група. Нека оптималните стойности на параметъра X са от $X_{d,opt}$ до $X_{u,opt}$ като състоянието става по-неблагоприятно както с намаляването на X от $X_{d,opt}$ надолу, така и с увеличаването на X от $X_{u,opt}$ нагоре.

За експерименталната група, извадковите числови характеристики на разпределението на X могат да се изчислят от Извадка 1. Значението на означенията е: M_E – извадкова средна стойност в експерименталната група; MED_E – извадкова медиана в експерименталната група; VAR_E – извадкова дисперсия в

експерименталната група; IQR_E – извадков интерквартилен обхват в експерименталната група. Аналогично, за псевдо контролната група, извадковите числови характеристики на разпределението на X могат да се изчислят от Извадка 2. Значението на означенията е: M_{PC} – извадкова средна стойност в псевдо контролната група; MED_{PC} – извадкова медиана в псевдо контролната група; VAR_{PC} – извадкова дисперсия в псевдо контролната група; IQR_{PC} – извадков интерквартилен обхват в псевдо контролната група.

Разработен е Алгоритъм 2.2.4.1 за определяне благоприятността на различията между популациите от непрекъснат параметър, подобен на този, който е показан по-долу за случая на благоприятността на различията между популациите от дискретен параметър (Алгоритъм 2.2.5.1).

Предлага се статистическата значимост на различията между популациите от непрекъснат параметър да се установява чрез 9 статистически теста, като от първи кластер. Нека $Pvalue_i$ е вероятността да се отхвърли погрешно нулевата хипотеза в Тест i, а нивото на значимост за всички тестове е ~ . Разработен е Алгоритъм 2.2.4.2 за определяне статистическата значимост на различията между популациите от непрекъснат параметър, подобен на този, който е показан по-долу за случая на статистическата значимост на нулевостта на промяната в непрекъснат параметър при една популация (Алгоритъм 2.2.6.2).

След определянето на статистическата значимост и благоприятността на различията между целевата и псевдо контролната популация в стойностите на непрекъснатия параметър X, то е възможно да се класифицира в пет категории еднаквостта на X в двете популации, както следва:

Категория '+1' – непрекъснатият параметър X показва статистически значимо по-благоприятно състояние в целевата популация, отколкото в псевдо контролната популация.

Категория '+1/2' – непрекъснатият параметър X показва гранично статистически значимо по-благоприятно състояние в целевата популация, отколкото в псевдо контролната популация.

Категория '0' – непрекъснатият параметър X не показва статистически значима разлика в състоянието при целевата и при псевдо контролната популация.

Категория '–1/2' – непрекъснатият параметър X показва гранично статистически значимо по-неблагоприятно състояние в целевата популация, отколкото в псевдо контролната популация.

Категория '–1' – непрекъснатият параметър X показва статистически значимо по-неблагоприятно състояние в целевата популация, отколкото в псевдо контролната популация.

Различията между двете популации от непрекъснатия параметър на състоянието могат да се класифицират като последователно се проверят 14 условия. Тази класификация може да се извърши много удобно автоматично по следния алгоритъм:

Алгоритъм 2.2.4.3 за класификация на различията между две популации от непрекъснат параметър

1. Ако средната стойност на X в целевата популация P е статистически значимо по-благоприятна от тази в псевдо контролната популация Q, то продължи, иначе иди към точка 2.
   1.1. Ако медианата на X в целевата популация P е или статистически значимо по-благоприятна, или гранично статистически значимо по-благоприятна от, или статистически неразличима от, или неутрална спрямо тази в псевдо контролната популация Q, то класифицирай в категория '+1' и край.
   1.2. Ако медианата на X в целевата популация P е гранично статистически значимо по-неблагоприятна от тази в псевдо контролната популация Q, то продължи, иначе иди към точка 1.3.
      1.2.1. Ако разпределенията на X в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '+1' и край.
      1.2.2. Класифицирай в категория '+1/2' и край.
   1.3. Тук медианата на X в целевата популация P е статистически значимо по-неблагоприятна от тази в псевдо контролната популация Q. Класифицирай в категория '0' и край.
2. Ако средната стойност на X в целевата популация P е гранично статистически значимо по-благоприятна от тази в псевдо контролната популация Q, то продължи, иначе иди към точка 3.
   2.1. Ако медианата на X в целевата популация P е статистически значимо по-благоприятна от тази в псевдо контролната популация Q, то класифицирай в категория '+1' и край.
   2.2. Ако медианата на X в целевата популация P е или гранично статистически значимо по-благоприятна от, или статистически неразличима от, или неутрална спрямо тази в псевдо контролната популация Q, то продължи, иначе иди към точка 2.3.
      2.2.1. Ако разпределенията на X в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '+1' и край.
      2.2.2. Класифицирай в категория '+1/2' и край.

2.3. Ако медианата на *X* в целевата популация *P* е гранично статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*, то класифицирай в категория ’0’ и край.
2.4. Тук медианата на *X* в целевата популация *P* е статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*.
2.4.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория ’–1’ и край.
2.4.2. Класифицирай в категория ’–1/2’ и край.

3. Ако средната стойност на *X* в целевата популация *P* е статистически незначимо по-благоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 4.
3.1. Ако медианата на *X* в целевата популация *P* е статистически значимо по-благоприятна от тази в псевдо контролната популация *Q*, то класифицирай в категория ’+1’ и край.
3.2. Ако медианата на *X* в целевата популация *P* е гранично статистически значимо по-благоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 3.3.
3.2.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория ’+1’ и край.
3.2.2. Класифицирай в категория ’+1/2’ и край.
3.3. Ако медианата на *X* в целевата популация *P* е или статистически незначимо по-благоприятна от, или неутрална спрямо тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 3.4.
3.3.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са статистически неразличими, то класифицирай в категория ’+1’ и край.
3.3.2. Ако разпределенията на *X* в двете популации са гранично статистически значимо различни и характеристиките на разсейването в двете популации са статистически неразличими, то класифицирай в категория ’+1/2’ и край.
3.3.3. Класифицирай в категория ’0’ и край.
3.4. Ако медианата на *X* в целевата популация *P* е статистически незначимо по-неблагоприятна от тази в псевдо контролната популация *Q*, то класифицирай в категория ’0’ и край.
3.5. Ако медианата на *X* в целевата популация *P* е гранично статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 3.6.
3.5.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория ’–1’ и край.
3.5.2. Класифицирай в категория ’–1/2’ и край.
3.6. Тук медианата на *X* в целевата популация *P* е статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*. Класифицирай в категория ’–1’ и край.

4. Ако средната стойност на *X* в целевата популация *P* е неутрална спрямо тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 5.
4.1. Ако медианата на *X* в целевата популация *P* е статистически значимо по-благоприятна от тази в псевдо контролната популация *Q*, то класифицирай в категория ’+1’ и край.
4.2. Ако медианата на *X* в целевата популация *P* е гранично статистически значимо по-благоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 4.3.
4.2.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория ’+1’ и край.
4.2.2. Класифицирай в категория ’+1/2’ и край.
4.3. Ако медианата на *X* в целевата популация *P* е статистически незначимо по-благоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 4.3.
4.3.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са статистически неразличими, то класифицирай в категория ’+1’ и край.
4.3.2. Ако разпределенията на *X* в двете популации са гранично статистически значимо различни и характеристиките на разсейването в двете популации са статистически неразличими, то класифицирай в категория ’+1/2’ и край.
4.3.3. Класифицирай в категория ’0’ и край.
4.4. Ако медианата на *X* в целевата популация *P* е неутрална спрямо тази в псевдо контролната популация *Q*, то класифицирай в категория ’0’ и край.

4.5. Ако медианата на *X* в целевата популация *P* е статистически незначимо по-неблагоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 4.6.
4.5.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '–1' и край.
4.5.2. Класифицирай в категория '–1/2' и край.
4.6. Ако медианата на *X* в целевата популация *P* е гранично статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 4.7.
4.6.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '–1' и край.
4.6.2. Класифицирай в категория '–1/2' и край.
4.7. Тук медианата на *X* в целевата популация *P* е статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*. Класифицирай в категория '–1' и край.

5. Ако средната стойност на *X* в целевата популация *P* е статистически незначимо по-неблагоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 6.
5.1. Ако медианата на *X* в целевата популация *P* е статистически значимо по-благоприятна от тази в псевдо контролната популация *Q*, то класифицирай в категория '+1' и край.
5.2. Ако медианата на *X* в целевата популация *P* е гранично статистически значимо по-благоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 3.3.
5.2.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '+1' и край.
5.2.2. Класифицирай в категория '+1/2' и край.
5.3. Ако медианата на *X* в целевата популация *P* е статистически незначимо по-благоприятна от тази в псевдо контролната популация *Q*, то класифицирай в категория '0' и край.
5.4. Ако медианата на *X* в целевата популация *P* е или статистически незначимо по-неблагоприятна от, или неутрална спрямо тази в псевдо контролната популация *Q*, то продължи, иначе иди към т. 3.4.
5.4.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са статистически неразличими, то класифицирай в категория '–1' и край.
5.4.2. Ако разпределенията на *X* в двете популации са гранично статистически значимо различни и характеристиките на разсейването в двете популации са статистически неразличими, то класифицирай в категория '–1/2' и край.
5.4.3. Класифицирай в категория '0' и край.
5.5. Ако медианата на *X* в целевата популация *P* е гранично статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 3.6.
5.5.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '–1' и край.
5.5.2. Класифицирай в категория '–1/2' и край.
5.6. Тук медианата на *X* в целевата популация *P* е статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*. Класифицирай в категория '–1' и край.

6. Ако средната стойност на *X* в целевата популация *P* е гранично статистически значимо по-неблагоприятна от тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 7.
6.1. Ако медианата на *X* в целевата популация *P* е статистически значимо по-благоприятна от тази в псевдо контролната популация *Q*, то класифицирай в категория '0' и край.
6.1.1. Ако разпределенията на *X* в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '+1' и край.
6.1.2. Класифицирай в категория '+1/2' и край.
6.2. Ако медианата на *X* в целевата популация *P* е гранично статистически значимо по-благоприятна от тази в псевдо контролната популация *Q*, то класифицирай в категория '0' и край.
6.3. Ако медианата на *X* в целевата популация *P* е или гранично статистически значимо по-неблагоприятна от, или статистически неразличима от, или неутрална спрямо тази в псевдо контролната популация *Q*, то продължи, иначе иди към точка 6.4.

6.3.1. Ако разпределенията на X в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '–1' и край.

6.3.2. Класифицирай в категория '–1/2' и край.

6.4. Тук медианата на X в целевата популация P е статистически значимо по-неблагоприятна от тази в псевдо контролната популация Q. Класифицирай в категория '–1' и край.

7. Тук средната стойност на X в целевата популация P е статистически значимо по-неблагоприятна от тази в псевдо контролната популация Q.

7.1. Ако медианата на X в целевата популация P е статистически значимо по-благоприятна от тази в псевдо контролната популация Q, то класифицирай в категория '0' и край.

7.2. Ако медианата на X в целевата популация P е гранично статистически значимо по-благоприятна от тази в псевдо контролната популация Q, то продължи, иначе иди към точка 7.3.

7.2.1. Ако разпределенията на X в двете популации са статистически значимо различни и характеристиките на разсейването в двете популации са или статистически неразличими, или гранично статистически различни, то класифицирай в категория '–1' и край.

7.2.2. Класифицирай в категория '–1/2' и край.

7.3. Тук медианата на X в целевата популация P е или статистически значимо по-неблагоприятна от, или гранично статистически значимо по-неблагоприятна от, или статистически неразличима от, или неутрална спрямо тази в псевдо контролната популация Q. Класифицирай в категория '–1' и край

2.2.5. Тестване за еднаквост на дискретен параметър в две популации по две извадки

Нека дискретният параметър X има T различни дискрети, номерирани с 1, 2, ..., T. Нека Извадка 1 да съдържа стойностите на дискретния параметър X за обектите от експерименталната група, а Извадка 2 да съдържа стойностите на X за обектите от псевдо контролната група. Нека множеството $G=\{g_1, g_2, \ldots, g_t\}$, да съдържа номерата на всички t дискрета, срещащи се в Извадка 1 и в Извадка 2. Ако t е 1, то всички обекти и в двете извадки съдържат един и същ дискрет. Тогава дискретният параметър на състоянието X е тривиално статистически неразличим в двете популации. Затова може да се разгледа само случаят, когато t е поне 2. Тогава G може да се раздели на три непресичащи се подмножества: непразно множество G^+, съдържащо номерата на всички сравнително благоприятни дискрети; непразно множество G^-, съдържащо номерата на всички сравнително неблагоприятни дискрети; евентуално празно множество G^0, съдържаща номерата на всички сравнително неутрални дискрета. Множествата G^+, G^- и G^0 се наричат съответно *плюс множество*, *минус множество* и *нула множество*.

За експерименталната група относителните честоти за попадане на X във всеки от g_r (за r=1,2,..,t) могат да се изчислят от Извадка 1. Оттам могат да се изчислят и относителните честоти за попадане на X в плюс и минус множествата G^+ и G^-. Значението на означенията е: $\neg^E_{g_r}$ – извадкова относителна честота за попадане на X в дискрет номер g_r при експерименталната група (за r=1,2,..,t); $«^E_+ = \sum_{r\in G^+} \neg^E_r$ – обща извадкова относителна честота за попадане на X в плюс множеството G^+ при експерименталната група; $«^E_- = \sum_{r\in G^-} \neg^E_r$ – обща извадкова относителна честота за попадане на X в минус множеството G^- при експерименталната група.

Аналогично, за псевдо контролната група, относителните честоти за попадане на X във всеки от g_r (за r=1,2,..,t) могат да се изчислят от Извадка 2. Оттам могат да се изчислят и относителните честоти за попадане на X в плюс и минус множествата G^+ и G^-. Значението на означенията е: $\neg^{PC}_{g_r}$ – извадкова относителна честота за попадане на X в дискрет номер g_r при псевдо контролната група (за r=1,2,..,t); $«^{PC}_+ = \sum_{r\in G^+} \neg^{PC}_r$ – обща извадкова относителна честота за попадане на X в плюс множеството G^+ при псевдо контролната група; $«^{PC}_- = \sum_{r\in G^-} \neg^{PC}_r$ – обща извадкова относителна честота за попадане на X в минус множеството G^- при псевдо контролната група.

Алгоритъм 2.2.5.1 за определяне благоприятността на различията между популациите от дискретен параметър

За относителни честоти за благоприятен дискрет ($g_r \in G^+$):

Ако $\neg^{PC}_{g_r} < \neg^E_{g_r}$, то вероятността за благоприятния дискрет номер g_r при целевата популация P се приема за по-благоприятна от тази при псевдо контролната популация Q.

Ако $\neg_{g_r}^{PC} > \neg_{g_r}^{E}$, то вероятността за благоприятния дискрет номер g_r при целевата популация P се приема за по-неблагоприятна от тази при псевдо контролната популация Q.

Ако $\neg_{g_r}^{PC} = \neg_{g_r}^{E}$, то вероятността за благоприятния дискрет номер g_r при целевата популация P се приема за неутрална спрямо тази при псевдо контролната популация Q.

За относителни честоти за неблагоприятен дискрет ($g_r \in G^-$):

Ако $\neg_{g_r}^{PC} < \neg_{g_r}^{E}$, то вероятността за неблагоприятния дискрет номер g_r при целевата популация P се приема за по-неблагоприятна от тази при псевдо контролната популация Q.

Ако $\neg_{g_r}^{PC} > \neg_{g_r}^{E}$, то вероятността за неблагоприятния дискрет номер g_r при целевата популация P се приема за по-благоприятна от тази при псевдо контролната популация Q.

Ако $\neg_{g_r}^{PC} = \neg_{g_r}^{E}$, то вероятността за неблагоприятния дискрет номер g_r при целевата популация P се приема за неутрална спрямо тази при псевдо контролната популация Q.

За относителни честоти за неутрален дискрет ($g_r \in G^0$):

Ако съществува, то вероятността за неутралния дискрет номер g_r при целевата популация P се приема за неутрална спрямо тази при псевдо контролната популация Q.

За общи относителни честоти за попадане в плюс множеството:

Ако $«_+^{PC} < «_+^{E}$, то общата вероятност за попадане в плюс множеството при целевата популация P се приема за по-благоприятна от тази при псевдо контролната популация Q.

Ако $«_+^{PC} > «_+^{E}$, то общата вероятност за попадане в плюс множеството при целевата популация P се приема за по-неблагоприятна от тази при псевдо контролната популация Q.

Ако $«_+^{PC} = «_+^{E}$, то общата вероятност за попадане в плюс множеството при целевата популация P се приема за неутрална спрямо тази при псевдо контролната популация Q.

За общи относителни честоти за попадане в минус множеството:

Ако $«_-^{PC} < «_-^{E}$, то общата вероятност за попадане в минус множеството при целевата популация P се приема за по-неблагоприятна от тази при псевдо контролната популация Q.

Ако $«_-^{PC} > «_-^{E}$, то общата вероятност за попадане в минус множеството при целевата популация P се приема за по-благоприятна от тази при псевдо контролната популация Q.

Ако $«_-^{PC} = «_-^{E}$, то общата вероятност за попадане в минус множеството при целевата популация P се приема за неутрална спрямо тази при псевдо контролната популация Q.

Предлага се статистическата значимост на различията между популациите от дискретен параметър да се установява чрез $(2t+1)$ статистически теста като от втори кластер. Нека $Pvalue_i$ е вероятността да се отхвърли погрешно нулевата хипотеза в Тест i, а нивото на значимост за всички тестове да е ~ . Разработен е Алгоритъм 2.2.5.2 за определяне статистическата значимост на различията между популациите от дискретен параметър, подобен на този, който е показан по-долу за случая на статистическата значимост на нулевостта на промяната в непрекъснат параметър при една популация (Алгоритъм 2.2.6.2).

Възможно е да се класифицира в пет категории еднаквостта на X в двете популации, както следва:

Категория ’+1’ – дискретният параметър X показва статистически значимо по-благоприятно състояние в целевата популация, отколкото в псевдо контролната популация.

Категория ’+1/2’ – дискретният параметър X показва гранично статистически значимо по-благоприятно състояние в целевата популация, отколкото в псевдо контролната популация.

Категория ’0’ – дискретният параметър X не показва статистически значима разлика в състоянието при целевата и при псевдо контролната популация.

Категория ’–1/2' – дискретният параметър X показва гранично статистически значимо по-неблагоприятно състояние в целевата популация, отколкото в псевдо контролната популация.

Категория ’–1' – дискретният параметър X показва статистически значимо по-неблагоприятно състояние в целевата популация, отколкото в псевдо контролната популация.

Различията между двете популации от дискретния параметър на състоянието могат да се класифицират като **последователно** се проверят 18 условия. Класификацията може да се извърши по-удобно автоматично по разработен Алгоритъм 2.2.5.3 за класификация на различията между две популации от дискретен

параметър, подобен на този, който е показан по-горе за случая на класификация на различията между две популации от непрекъснат параметър (Алгоритъм 2.2.4.3).

2.2.6. Тестване за нулевост на промяната на непрекъснат параметър при една популация по две сдвоени извадки

Нека Извадка 1 да съдържа стойностите на непрекъснатия параметър X за обектите от експерименталната или от псевдо контролната група в края на експеримента след базово въздействие B (и евентуално след изследваното въздействие R), а Извадка 2 да съдържа стойностите на X за обектите от същата група в началото на експеримента преди базово въздействие B (и евентуално преди изследваното въздействие R). Нека оптималните стойности на параметъра X са от $X_{d,opt}$ до $X_{u,opt}$ като състоянието става по-неблагоприятно както с намаляването на X от $X_{d,opt}$ надолу, така и с увеличаването на X от $X_{u,opt}$ нагоре. Когато предпочитанията върху X растат безкрайно с нарастването на стойността му, тогава може да се положи $X_{d,opt} = X_{u,opt} = +\infty$. Когато предпочитанията върху X намаляват безкрайно с нарастването на стойността му, тогава може да се положи $X_{d,opt} = X_{u,opt} = -\infty$.

За крайната група, извадковите числови характеристики на разпределението на X могат да се изчислят от Извадка 1. Значението на означенията е: M_e – извадкова средна стойност в крайната група; MED_e – извадкова медиана в крайната група. Аналогично, за началната група извадковите числови характеристики на разпределението на X могат да се изчислят от Извадка 2. Значението на означенията е: M_0 – извадкова средна стойност в началната група; MED_0 – извадкова медиана в началната група.

Извадка 3 съдържа разликите на стойностите на X от крайната Извадка 1 и стойностите на X от същите обекти в началната Извадка 2. Извадковите числови характеристики на разпределението на промяната $\mathrm{a}X$ могат да се изчислят от Извадка 3. Значението на означенията е: M_{a} – извадкова средна стойност на промяната; MED_{a} – извадкова медиана на промяната. Следва да се отбележи, че докато винаги $M_{\mathrm{a}} = M_e - M_0$, то в общия случай $MED_{\mathrm{a}} \neq MED_e - MED_0$.

Разработен е Алгоритъм 2.2.6.1 за определяне благоприятността на промяната в непрекъснат параметър при една популация, подобен на този, който е показан по-горе за случая на благоприятността на различията между популациите от дискретен параметър (Алгоритъм 2.2.5.1).

Предлага се статистическата значимост на нулевостта на промяната в непрекъснат параметър при една популация да се установява чрез 4 статистически теста. Нека $Pvalue_i$ е вероятността да се отхвърли погрешно нулевата хипотеза в Тест i, а нивото на значимост за всички тестове е $\sim$.

Алгоритъм 2.2.6.2 за определяне статистическата значимост на нулевостта на промяната в непрекъснат параметър при една популация

За средни стойности:

Ако $Pvalue_2 \le Pvalue_1 \le \sim$ и $M_{\mathrm{a}} > 0$, то популационната средна стойност на промяната се приема за статистически значимо положителна.

Ако $Pvalue_2 \le Pvalue_1 \le \sim$ и $M_{\mathrm{a}} < 0$, то популационната средна стойност на промяната се приема за статистически значимо отрицателна.

Ако $Pvalue_2 \le \sim < Pvalue_1$ и $M_{\mathrm{a}} > 0$, то популационната средна стойност на промяната се приема за гранично статистически значимо положителна.

Ако $Pvalue_2 \le \sim < Pvalue_1$ и $M_{\mathrm{a}} < 0$, то популационната средна стойност на промяната се приема за гранично статистически значимо отрицателна.

Ако $\sim < Pvalue_2 \le Pvalue_1$, то популационната средна стойност на промяната се приема за статистически неразличима от нула.

От една страна, $Pvalue_2$ на едностранния тест е не по-голям от $Pvalue_1$ на двустранния тест и следователно винаги $Pvalue_2 \le Pvalue_1$. От друга страна, ако средната стойност на промяната е нула ($M_{\mathrm{a}} = 0$), то $Pvalue_1$ се приема за 1, а $Pvalue_2$ се приема за 0.5, което е много по-голямо от всяка разумна стойност на $\sim$ и следователно тогава винаги $\sim < Pvalue_2 \le Pvalue_1$. Оттук следва, че други случаи не съществуват.

За медиани:

Ако $Pvalue_4 \le Pvalue_3 \le \sim$ и $MED_{\mathrm{a}} > 0$, то популационната медиана на промяната се приема за статистически значимо положителна.

Ако $Pvalue_4 \le Pvalue_3 \le \sim$ и $MED_{\mathrm{a}} < 0$, то популационната медиана на промяната се приема за статистически значимо отрицателна.

Ако $Pvalue_4 \le \sim < Pvalue_3$ и $MED_a > 0$, то популационната медиана на промяната се приема за гранично статистически значимо положителна.

Ако $Pvalue_4 \le \sim < Pvalue_3$ и $MED_a < 0$, то популационната медиана на промяната се приема за гранично статистически значимо отрицателна.

Ако $\sim < Pvalue_4 \le Pvalue_3$, то популационната медиана на промяната се приема за статистически неразличима от нула.

От една страна, $Pvalue_4$ на едностранния тест е не по-голям от $Pvalue_3$ на двустранния тест и следователно винаги $Pvalue_4 \le Pvalue_3$. От друга страна, ако медианата на промяната е нула ($MED_a = 0$), то $Pvalue_3$ се приема за 1, а $Pvalue_4$ се приема за 0.5, което е много по-голямо от всяка разумна стойност на $\sim$ и следователно тогава винаги $\sim < Pvalue_4 \le Pvalue_3$. Оттук следва, че други случаи не съществуват.

След определянето на статистическата значимост на знака и благоприятността на промяната a_X в стойностите на X в една популация, то е възможно a_X в популацията да се класифицира в пет категории:

Категория '+1' – непрекъснатият параметър X показва статистически значима благоприятна промяна на състоянието в популацията.

Категория '+1/2' – непрекъснатият параметър X показва гранично статистически значима благоприятна промяна на състоянието в популацията.

Категория '0' – непрекъснатият параметър X не показва статистически значима промяна на състоянието в популацията.

Категория '–1/2' – непрекъснатият параметър X показва гранично статистически значима неблагоприятна промяна на състоянието в популацията.

Категория '–1' – непрекъснатият параметър X показва статистически значима неблагоприятна промяна на състоянието в популацията.

Промяната a_X на непрекъснатия параметър на състоянието в популацията може да се класифицира като последователно се проверят 6 условия. Класификацията може да се извърши по-удобно автоматично по разработен Алгоритъм 2.2.6.3 за класификация на промяната в непрекъснат параметър при една популация, подобен на този, който е показан по-горе за случая на класификация на различията между две популации от непрекъснат параметър (Алгоритъм 2.2.4.3).

2.2.7. Метод на псевдо контролната група

Нека намирането на ефекта на изследваното въздействие R върху избрания параметър X чрез сравняването на различията в популациите P и Q преди въздействията чрез извадките E_0 и PC_0, със същите след въздействията чрез извадките E_e и PC_e, да се нарече *метод на псевдо контролната група* (МПКГ).

Основната идея на МПКГ е, че ако според стойностите на X в извадките, първоначално се наблюдава или статистически значимо, или гранично статистически значимо по-неблагоприятно, а след въздействията се наблюдава или статистически значимо, или гранично статистически значимо по-благоприятно състояние в целевата популация P, отколкото в псевдо контролната популация Q, то влиянието на изследваното въздействие R върху избрания параметър на състоянието се счита за статистически значимо благоприятно и обратно.

Нека b е число от множеството {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията, в която първоначално са класифицирани различията между стойностите на параметъра на състоянието X в целевата популация P и в псевдо контролната популация Q. Нека e е число от множеството {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията, в която след въздействията са класифицирани различията между стойностите на параметъра на състоянието X в целевата популация P и в псевдо контролната популация Q. Наредената двойка (b, e) се нарича *тип на МПКГ*, който зависи от данните в четирите извадки E_0, PC_0, E_e и PC_e.

Използвайки типа МПКГ, благоприятността и значимостта на влиянието на изследваното въздействие R върху избрания параметър X може да се класифицира в 5 категории, както следва:

Категория 'ДА+' – R има статистически значимо благоприятно влияние върху параметър X.

Категория 'ГР+' – R има гранично статистически значимо благоприятно влияние върху параметър X.

Категория 'НЕ' – липса на статистически значимо влияние на въздействие R върху параметър X.

Категория 'ГР–' – R има гранично статистически значимо неблагоприятно влияние върху параметър X.

Категория 'ДА–' – R има статистически значимо неблагоприятно влияние върху параметър X.

В резултат на сравненията на данните от четирите извадките E_0, PC_0, E_e и PC_e могат да възникнат 25 различни типа МПКГ, които се класифицират в съответните категории. Може да се изведе просто емпирично

правило за класификация на благоприятността и значимостта на влиянието на въздействие R върху параметър X, използвайки типа на МПКГ:

$$\text{Класифицирай } (b,e) \text{ на МПКГ в } \begin{cases} \text{'ДА+'} & \text{, ако } b-e<-1/2 \\ \text{'ГР+'} & \text{, ако } b-e=-1/2 \\ \text{'НЕ'} & \text{, ако } b-e=0 \\ \text{'ГР}-\text{'} & \text{, ако } b-e=+1/2 \\ \text{'ДА}-\text{'} & \text{, ако } b-e>+1/2 \end{cases} \qquad (2.2.7.2.1)$$

Може окончателно да се формализира МПКГ в следния, удобен за компютърна реализация, алгоритъм, който ползва на входа си данните в четирите извадки E_0, PC_0, E_e и PC_e:

Алгоритъм 2.2.7.3 за оценка на ефекта от въздействие R върху параметър на състоянието X по МПКГ

1. Ако параметърът на състоянието X е непрекъснат, то продължи, иначе иди в точка 2.
   - 1.1. За началната експериментална извадка E_0, изчисли извадковите числови характеристики на разпределението M_E, MED_E, VAR_E и IQR_E от параграф 2.2.4.1.
   - 1.2. За началната псевдо контролна извадка PC_0, изчисли извадковите числови характеристики на разпределението M_{PC}, MED_{PC}, VAR_{PC} и IQR_{PC} от параграф 2.2.4.1.
   - 1.3. За началото на експеримента, по Алгоритъм 2.2.4.1 определи благоприятността на различията между целевата популация P и псевдо контролната популация Q.
   - 1.4. Изчисли $Pvalue_i$ (за i=1,2,...,9) на 9-те статистически теста от параграф 2.2.4.2, които сравняват началните извадки E_0 и PC_0.
   - 1.5. За началото на експеримента, по Алгоритъм 2.2.4.2 определи статистическата значимост на различията между P и Q.
   - 1.6. За началото на експеримента, по Алгоритъм 2.2.4.3 класифицирай различията между P и Q в една от петте категории '+1', '+1/2', '0', '–1/2' или '–1'.
   - 1.7. Определи b като число от множеството {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията от т. 1.6.
   - 1.8. За крайната експериментална извадка E_e, изчисли извадковите числови характеристики на разпределението M_E, MED_E, VAR_E и IQR_E от параграф 2.2.4.1.
   - 1.9. За крайната псевдо контролна извадка PC_e, изчисли извадковите числови характеристики на разпределението M_{PC}, MED_{PC}, VAR_{PC} и IQR_{PC} от параграф 2.2.4.1.
   - 1.10. За края на експеримента, по Алгоритъм 2.2.4.1 определи благоприятността на различията между целевата популация P и псевдо контролната популация Q.
   - 1.11. Изчисли $Pvalue_i$ (за i=1,2,...,9) на деветте статистически теста от параграф 2.2.4.2, които сравняват крайните извадки E_e и PC_e.
   - 1.12. За края на експеримента, по Алгоритъм 2.2.4.2 определи статистическата значимост на различията между целевата популация P и псевдо контролната популация Q.
   - 1.13. За края на експеримента, по Алгоритъм 2.2.4.3 класифицирай различията между P и Q в една от петте категории '+1', '+1/2', '0', '–1/2' или '–1'.
   - 1.14. Определи e като число от множеството {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията от т.1.13.
2. Тук параметърът на състоянието X е дискретен.
   - 2.1. Намери номерата g_r и броят t на всички дискрети, срещащи се в началните извадки E_0 и PC_0. Формирай множеството $G=\{g_1, g_2, \ldots, g_t\}$.
   - 2.2. Раздели G от точка 2.1 на непресичащите се плюс множество G^+, минус множество G^- и нула множество G^0 от параграф 2.2.5.
   - 2.3. За началната експериментална извадка E_0, изчисли извадковите числови характеристики на разпределението $\neg^{E}_{g_k}$ (за r=1,2,..,t), «$^{E}_{+}$ и «$^{E}_{-}$ от параграф 2.2.5.1.
   - 2.4. За началната псевдо контролна извадка PC_0, изчисли извадковите числови характеристики на разпределението $\neg^{PC}_{g_k}$ (за r=1,2,..,t), «$^{PC}_{+}$ и «$^{PC}_{-}$ от параграф 2.2.5.1.
   - 2.5. За началото на експеримента, по Алгоритъм 2.2.5.1 определи благоприятността на различията между целевата популация P и псевдо контролната популация Q.
   - 2.6. Изчисли $Pvalue_i$ (за i=1,2,...,2t+1) на (2t+1)-те статистически теста от параграф 2.2.5.2, които сравняват началните извадки E_0 и PC_0.
   - 2.7. За началото на експеримента, по Алгоритъм 2.2.5.2 определи статистическата значимост на различията между целевата популация P и псевдо контролната популация Q.
   - 2.8. За началото на експеримента, по Алгоритъм 2.2.5.3 класифицирай различията между P и Q в една от петте категории '+1', '+1/2', '0', '–1/2' или '–1'.

2.9. Определи b като число от {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията, определена в точка 2.8.
2.10. Намери номерата g_r и броят t на всички дискрети, срещащи се в крайните извадки E_e и PC_e. Формирай множеството $G=\{g_1, g_2, \ldots, g_t\}$.
2.11. Раздели G от точка 2.10 на непресичащите се плюс множество G^+, минус множество G^- и нула множество G^0 от параграф 2.2.5.
2.12. За крайната експериментална извадка E_e, изчисли извадковите числови характеристики на разпределението $\neg_{g_k}^{E}$ (за r=1,2,..,t), «$_{+}^{E}$ и «$_{-}^{E}$ от параграф 2.2.5.1.
2.13. За крайната псевдо контролна извадка PC_e, изчисли извадковите числови характеристики на разпределението $\neg_{g_k}^{PC}$ (за r=1,2,..,t), «$_{+}^{PC}$ и «$_{-}^{PC}$ от параграф 2.2.5.1.
2.14. За края на експеримента, по Алгоритъм 2.2.5.1 определи благоприятността на различията между целевата популация P и псевдо контролната популация Q.
2.15. Изчисли $Pvalue_i$ (за i=1,2,...,2t+1) на (2t+1)-те статистически теста от параграф 2.2.5.2, които сравняват крайните извадки E_e и PC_e.
2.16. За края на експеримента, по Алгоритъм 2.2.5.2 определи статистическата значимост на различията между целевата популация P и псевдо контролната популация Q.
2.17. За края на експеримента, по Алгоритъм 2.2.5.3 класифицирай различията между P и Q в една от петте категории '+1', '+1/2', '0', '–1/2' или '–1'.
2.18. Определи e като число от множеството {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията от т. 1.17.

3. Формирай типа на МПКГ, като наредената двойка (b, e).
4. По формула (2.2.7.2.1), използвайки типа на МПКГ, класифицирай благоприятността и значимостта на влиянието на R върху параметър X, в една от петте категории 'ДА+', 'ГР+', 'НЕ', 'ГР–' или 'ДА–'.

2.2.8. Модифициран метод на псевдо контролната група

Нека намирането на ефекта на въздействие R върху непрекъснат параметър X чрез сравняването на нулевостта на промяната a_X на параметъра на състоянието в целевата популация P, вследствие на въздействията R и B, чрез сдвоените извадки E_0 и E_e с нулевостта на същата в псевдо контролната популация Q, вследствие на въздействието B, чрез сдвоените извадки PC_0 и PC_e да се нарече *модифициран метод на псевдо контролната група* (ММПКГ).

Основната идея на ММПКГ е, че ако според стойностите на X в извадките, в целевата популация P се наблюдава или статистически значима, или гранично статистически значима благоприятна промяна на състоянието, а в псевдо контролната популация Q се наблюдава или статистически значима, или гранично статистически значима неблагоприятна промяна на същото, то влиянието на изследваното въздействие R върху избрания параметър на състоянието се счита за статистически значимо благоприятно и обратно.

Нека b е число от множеството {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията, в която е класифицирана промяната a_X на стойностите на параметъра на състоянието X в целевата популация P. Нека e е число от множеството {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията, в която е класифицирана промяната a_X на стойностите на параметъра на състоянието X в псевдо контролната популация Q. Тъй като ММПКГ няма смисъл при дискретни параметри, то X е винаги непрекъснат параметър и тогава b и e се определят съгласно точка 2.2.6.3. Наредената двойка (b, e) се нарича *тип на ММПКГ*, който зависи от данните в четирите извадки E_0, PC_0, E_e и PC_e.

Използвайки типа ММПКГ, благоприятността и значимостта на влиянието на изследваното въздействие R върху избрания параметър X може да се класифицира в 5 категории, както следва:

Категория 'ДА+' – R има статистически значимо благоприятно влияние върху параметър X.
Категория 'ГР+' – R има гранично статистически значимо благоприятно влияние върху параметър X.
Категория 'НЕ' – липса на статистически значимо влияние на въздействие R върху параметър X.
Категория 'ГР–' – R има гранично статистически значимо неблагоприятно влияние върху параметър X.
Категория 'ДА–' – R има статистически значимо неблагоприятно влияние върху параметър X.

В резултат на сравненията на данните от четирите извадките E_0, PC_0, E_e и PC_e могат да възникнат 25 различни типа ММПКГ, които се класифицират в съответните категории. От класификацията може да се изведе просто емпирично правило за класификация на благоприятността и значимостта на влиянието на въздействие R върху параметър X, използвайки типа на ММПКГ:

$$\text{Класифицирай } (b,e) \text{ на ММПКГ в } \begin{cases} \text{'ДА+'} & \text{, ако} \quad b-e > +1/2 \\ \text{'ГР+'} & \text{, ако} \quad b-e = +1/2 \\ \text{'НЕ'} & \text{, ако} \quad b-e = 0 \\ \text{'ГР–'} & \text{, ако} \quad b-e = -1/2 \\ \text{'ДА–'} & \text{, ако} \quad b-e < -1/2 \end{cases} \qquad (2.2.8.2.1)$$

Представените алгоритми могат да се обединят за окончателно формализиране на ММПКГ в следния, удобен за компютърна реализация, алгоритъм, който ползва на входа си данните в четирите извадки E_0, PC_0, E_e и PC_e:

Алгоритъм 2.2.8.3 за оценка на ефекта от въздействие *R* върху параметър на състоянието *X* по ММПКГ

1. За крайната експериментална извадка E_e, изчисли извадковите числови характеристики на разпределението M_e и MED_e от параграф 2.2.6.1.
2. За началната експериментална извадка E_0, изчисли извадковите числови характеристики на разпределението M_0 и MED_0 от параграф 2.2.6.1.
3. Формирай експерименталната извадка на промяната, съдържаща разликите между съответните стойностите на *X* от крайната експериментална извадка E_e и от началната експериментална извадка E_0.
4. За експерименталната извадка на промяната, изчисли извадковите числови характеристики на разпределението на промяната M_a и MED_a от параграф 2.2.6.1.
5. За целевата популация *P*, по Алгоритъм 2.2.6.1 определи благоприятността на промяната a_X в непрекъснатия параметър на състоянието *X*.
6. Изчисли $Pvalue_i$ (за i=1,2,...,4) на 4-те статистически теста от параграф 2.2.6.2, които оценяват ненулевостта на промяната между експерименталните извадки E_e и E_0.
7. За целевата популация *P*, по Алгоритъм 2.2.6.2 определи статистическата значимост на нулевостта на промяната a_X в непрекъснатия параметър на състоянието *X*.
8. За целевата популация *P*, по Алгоритъм 2.2.6.3 класифицирай промяната a_X в непрекъснатия параметър на състоянието *X* в една от петте категории '+1', '+1/2', '0', '–1/2' или '–1'.
9. Определи *b* като число от {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията, определена в точка 8.
10. За крайната псевдо контролна извадка PC_e, изчисли извадковите числови характеристики на разпределението M_e и MED_e от параграф 2.2.6.1.
11. За началната псевдо контролна извадка PC_0, изчисли извадковите числови характеристики на разпределението M_0 и MED_0 от параграф 2.2.6.1.
12. Формирай псевдо контролната извадка на промяната, съдържаща разликите между съответните стойностите на *X* от PC_e и от PC_0.
13. За псевдо контролната извадка на промяната, изчисли извадковите числови характеристики на разпределението на промяната M_a и MED_a от параграф 2.2.6.1.
14. За псевдо контролната популация *Q*, по Алгоритъм 2.2.6.1 определи благоприятността на промяната a_X в непрекъснатия параметър на състоянието *X*.
15. Изчисли $Pvalue_i$ (за i=1,2,...,4) на 4-те статистически теста от параграф 2.2.6.2, които оценяват ненулевостта на промяната между псевдо контролните извадки PC_e и PC_0.
16. За псевдо контролната популация *Q*, по Алгоритъм 2.2.6.2 определи статистическата значимост на нулевостта на промяната a_X в непрекъснатия параметър на състоянието *X*.
17. За псевдо контролната популация *Q*, по Алгоритъм 2.2.6.3 класифицирай промяната a_X в непрекъснатия параметър на състоянието *X* в една от петте категории '+1', '+1/2', '0', '–1/2' или '–1'.
18. Определи *e* като число от {+1, +1/2, 0, –1/2, –1}, съвпадащо с категорията, определена в точка 17.
19. Формирай типа на ММПКГ, като наредената двойка (*b*, *e*).
20. По формула (2.2.8.2.1), използвайки типа на ММПКГ, класифицирай благоприятността и значимостта на влиянието на *R* върху *X*, в една от петте категории 'ДА+', 'ГР+', 'НЕ', 'ГР–' или 'ДА–'.

ГЛАВА 3. ЕКСПЕРИМЕНТАЛНО АПРОБИРАНЕ НА АЛГОРИМИТЕ

3.1. Информационна база за апробиране на системата

В Отделението (сега Клиника) по кардиохирургия към УМБАЛ "Св. Марина" – ЕАД, град Варна, за периода януари 2007 – юни 2011 г. е събрана информация за 140 пациенти, страдащи от ИБС, усложнена със значима хронична ИМР. В електонна Excel таблица се съхраняват имената, структурата и стойностите на 75 непрекъснати параметъра, 85 дискретни параметъра и 13 идентификатори. Някои параметри са измерени преди операцията, веднага след операцията и късно след операцията - *строени*, като всяка тройка се нарича *обобщен тримерен параметър*. Налични са 12 непрекъснати, 21 обобщени тримерни непрекъснати, 43

дискретни и 14 обобщени тримерни дискретни параметри. Вероятностите за принадлежност към групите са субективно оценени от медицински специалисти, чрез решаване на преференциални уравнения от маг. Снежана Иванова, направени при разработка на нейния дисертационен труд под ръководството на доц. д-р Наталия Николова. Програмният интерфейс за достъп до информационния масив, прехвърлянето му в MATLAB формат и генерацията на произволни извадки за провеждане на тестове от Първи, Втори и Трети кластер също са разработени от маг. Снежана Иванова. Процедурата за оценяване на вероятностите, програмния интерфейс, както и получените условни и абсолютни вероятности излизат извън обхвата на настоящата дисертация. Предложените алгоритмични основи на обща мединска информационна система са апробирани върху 3 идентификатора, 11 обобщени двумерни непрекъснати параметъра и един обобщен двумерен дискретен параметър, като всички параметри са измерени един път преди операцията и един път късно след операцията (от 6 до 54 месеца след нея). За всеки пациент се формира 27-мерен запис от:

1. Идентификатори:
1.1. Идентификационен номер на пациента.
1.2. Подгрупа на пациента, дискретен признак с 4 дискрета: 0 – за A1, 1 – за A2, 2 – за B1 и 3 – за B2.
1.3. Вероятност за принадлежност на пациента към съответната му подгрупа.
2. Двумерни непрекъснати параметри:
2.1. Индекс на теледиастолния обем на лява камера (LVEDV_Index) в *ml/m²*
2.2. Индекс на телесистолния обем на лява камера (LVESV_Index) в *ml/m²*
2.3. Индекс на обема на ляво предсърсие (LA_Volume_Index) в *ml/m²*
2.4. Вена контракта (Vena_Contracta) в *mm*
2.5. Регургитационен обем през митралната клапа (RegVol) в *ml*
2.6. Коаптационна дължина (Coaptation_Length) в *mm*
2.7. Тентинг площ (Tenting_Area) в cm^2
2.8. Тентинг височина (Tenting_Height) в *mm*
2.9. PISA радиус (PISAr) в *cm*
2.10. Модифицирана фракция на изтласкване (Calculated_Real_EF) в %
2.11. Фракция на изтласкване (EF_Number) в %
3. Двумерни дискретни параметри:
3.1. Реална степен на митрална регургитация (Real_MR) с дискрети: 0, 0-I, I, I-II, II, II-III и III

Формиран е масив от 11 обобщени двумерни непрекъснати параметри, 1 обобщен двумерен дискретен параметър; 24 записа за A1, 32 записа за A2, 32 записа за B1 и 22 записа за B2. В резултат за всяка от подгрупите може да се формира масив от 11 обобщени двумерни непрекъснати параметри и 1 обобщен двумерен дискретен параметър, съдържащи съответно 24 записа за подгрупа A1, 32 записа за подгрупа A2, 32 записа за подгрупа B1 и 22 записа за подгрупа B2.

3.2. Влияние на анулопластиката чрез Буутстрап с размерно-еднаква генерация от ECDF

Нека се търси ефектът от въздействието оперативна анулопластика (*R*), действащо едновременно с базовото въздействие оперативна реваскулоризация (*B*), върху някой от параметрите 3.1 или от 2.1 до 2.10.

Този параметър характеризира състоянието на целевата популация (*P*) от пациенти със значима ИМР подложени на комбинирана процедура. Преди и след комбинирана процедура, експериментално са измерени стойностите на параметъра за пациентите от експерименталната група A. Ефектът на анулопластиката се установява в сравнение с ефекта на изолираната процедура върху същия параметър (*X*), който сега характеризира състоянието на псевдо контролната популация (*Q*) от пациенти със значима ИМР, подложени на изолирана реваскулоризация. Преди и след изолираната процедура, експериментално са измерени стойностите на параметъра за пациентите от псевдо контолната група B.

Класифицирането на пациентите в групи A или B става по 5-стъпков алгоритъм. Двете групи са информационно разделени на по две относително по-хомогенни подгрупи. Това разрешава да се получат много по-ясни статистически изводи при сравненията от една страна на подгрупите A1 или B1 на пациентите със сравнително по-запазено сърдечно и общо състояние и от друга страна на подгрупите A2 или B2 на пациентите със сравнително по-увредено сърдечно и общо състояние. И в двата случая МПКГ може и се прилага както върху дискретния параметър, така и върху 10 непрекъснати параметри, докато ММПКГ може и се прилага само върху 10 непрекъснати параметъра.

Статистическите тестове са извършени чрез Буутстрап метод с по 2000 псевдо реалности. Синтетичните извадки се формират с размерно-еднаква генерация от емпирични кумулативни функции на разпределение. Нивото на значимост при всички тестове е прието за 0.05.

Четирите извадки от данни, са както следва:

E_0 съдържа стойностите на избрания параметър за A1 преди комбинираната операция.

PC_0 съдържа стойностите на избрания параметър за B1 преди изолираната процедура.

E_e съдържа стойностите на избрания параметър за A1 късно след комбинираната операция.

PC_e съдържа стойностите на избрания параметър за B1 късно след изолираната процедура.

Получените стойности на Pvalue за статистическите тестове, сравняващи две извадки от пациенти, принадлежащи на групи A1 и B1 за непрекъснатите параметри на състояние, показани в колона 1, са дадени в Таблица 3.2.1.1.1. Деветте теста, указани в таблицата, са като тестовете за първи кластер. В колона 2, 'преди' означава, че теста е извършен преди операцията, а 'след' – че е извършен късно след операцията. Получените стойности на Pvalue за статистическия тест, сравняващ две извадки от пациенти, принадлежащи на групи A1 и B1 за дискретния параметър на състояние Real_MR, са показани в Таблица 3.2.1.1.2. Осемте теста, указани в таблицата, са като от втори кластер. Тестове от 2 до 8 са изпълнени в два варианта – двустранен (в 1 и 3 ред) и едностранен (във 2 и 4 ред). В колона 2, 'преди' означава, че теста е извършен преди операцията, а 'след' – че е извършен късно след операцията.

Получените стойности на Pvalue за статистическите тестове, оценяващи нулевостта на промяната вследствие на операцията на непрекъснати параметри на състояние, показани в колона 1, или от група A1 или от група B1, са показани в Таблица 3.2.1.2.1. Четирите теста, указани в таблицата, са както от трети кластер. В колона 2 е означена групата, за която се оценява нулевостта на промяната.

В Таблица 3.2.1.3.1 се обобщава ефекта на анулопластиката върху единадесетте параметъра при сравнение на групите A1 и B1. В колони 3 и 5 значението на символите е: 'НЕ' – липса на статистически значимо влияние; 'ГР+' – гранично статистически значимо благоприятно влияние; 'ГР–' – гранично статистически значимо неблагоприятно влияние; 'ДА+' – статистически значимо благоприятно влияние; 'ДА–' – статистически значимо неблагоприятно влияние. В колона 7 значението на символите е: 'НЕ' – липса на медицински съществено влияние; 'ОТН+' – медицински относително съществено благоприятно влияние; 'ОТН–' – медицински относително съществено неблагоприятно влияние; 'ДА+' – медицински съществено благоприятно влияние; 'ДА–' – медицински съществено неблагоприятно влияние. В колони 4 и 6 са дадени типовете резултати от съответния тест. Резултатът от всеки метод, сочещ неблагоприятно влияние на анулопластиката върху даден параметър е даден на жълт фон.

Таблица 3.2.1.1.1. *Pvalue* на статистически тестове за еднаквост на групи A1 и B1 по непрекъснати параметри, проведени преди и късно след операцията, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	**Време**	**Тестове**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
LVEDV_Index	преди	0.97	0.37	0.19	0.46	0.21	0.165	0.090	0.22	0.21
	след	0.92	0.80	0.40	0.75	0.32	0.73	0.37	0.63	0.43
LVESV_Index	преди	0.99	0.30	0.15	0.21	0.0040	0.91	0.48	0.46	0.28
	след	0.23	1.0	0.51	0.57	0.17	0.92	0.44	0.11	0.090
LA_Volume_Index	преди	0.685	0.030	0.020	0.10	0.020	0.030	0.0005	0.66	0.48
	след	0.13	0.22	0.115	0.92	0.30	0.020	0	0.77	0.55
Vena_Contracta	преди	0	0	0	0	0	0.0035	0	1	1
	след	0.0005	0	0	0.070	0.070	0.40	0.22	0.51	0.47
RegVol	преди	0	0	0	0	0	0.0050	0.0040	0.030	0.030
	след	0.0005	0	0	0.19	0.17	0.050	0.040	0.33	0.31
Coaptation_Length	преди	0	0	0	0.26	0.24	0.0035	0.0010	1	1
	след	0	0	0	0	0	0.15	0.060	0.67	0.38
Tenting_Area	преди	0	0	0	0.0040	0.0005	0	0	0.10	0.10
	след	0.0005	0	0	0	0	0.42	0.20	0.78	0.43
Tenting_Hight	преди	0	0	0	0	0	0.0085	0.010	0.27	0.27
	след	0.30	0.85	0.43	1	1	0.29	0.16	0.21	0.17
PISAr	преди	0.0060	0	0	0.0015	0	0.0065	0	0.52	0.39
	след	0	0	0	0.0005	0	0.87	0.45	0.72	0.46
Calculated_Real_EF	преди	0	0	0	0	0	0.58	0.30	0.73	0.415
	след	0.060	0.010	0.0035	0.020	0	0.17	0.080	0.62	0.36

От Таблица 3.2.1.3.1 се вижда, че:

1) Според МПКГ от 11 параметъра при групи A1 и B1, анулопластиката има статистически значимо благоприятно влияние върху 9 параметъра, анулопластиката има гранично статистически значимо благоприятно влияние върху 1 параметър и липсва статистически значимо влияние на анулопластиката върху 1 параметър.

2) Според ММПКГ от 10 параметъра при групи A1 и B1, анулопластиката има гранично статистически значимо благоприятно влияние върху 1 параметър, липсва статистически значимо влияние на анулопластиката върху 8 параметъра и анулопластиката има гранично статистически значимо неблагоприятно влияние върху 1 параметър.

3) Според ММПКГ от 10 параметъра при групи A1 и B1, анулопластиката има медицински съществено благоприятно влияние върху 6 параметъра, анулопластиката има медицински относително съществено благоприятно влияние върху 3 параметъра и липсва медицински съществено влияние на анулопластиката върху 1 параметър.

4) Общата картина от всички параметри показва статистически значимо и медицински съществено по-благоприятно влияние на анулопластиката при пациенти от група A1 спрямо влиянието на изолираната процедура в псевдо-контролната група B1.

Таблица 3.2.1.1.2. *Pvalue* на статистически тестове за еднаквост на групи A1 и B1 по дискретни параметри, проведени преди и късно след операцията, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	**Време**	**Тестове**							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Real_MR	Преди (двустранен)	0				0	0.085	0	0
	Преди (едностранен)					0	0.050	0	0
	След (двустранен)	0	0	0.20	0.10		0.11	0.50	0.51
	След (едностранен)		0	0.11	0.050		0.050	0.25	0.27

Аналогични процедури за приложение за анализ влиянието на анулопластиката върху пациенти от групи A2-B2.

Таблица 3.2.1.2.1. *Pvalue* на статистически тестове за нулевост на промяната на непрекъснати параметри, проведени в група A1 и в група B1, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	**Група**	**Тестове**			
		1	**2**	**3**	**4**
LVEDV_Index	A1	0.010	0.0030	0.12	0.11
	B1	0.28	0.14	0.71	0.42
LVESV_Index	A1	0.0025	0	0.010	0.010
	B1	0.080	0.046	0.17	0.090
LA_Volume_Index	A1	0.050	0.040	0.12	0.060
	B1	0.040	0.025	0.85	0.42
Vena_Contracta	A1	0	0	0	0
	B1	0.0090	0.0050	0.16	0.010
RegVol	A1	0	0	0	0
	B1	0.010	0.0060	0.050	0.040
Coaptation_Length	A1	0	0	0	0
	B1	0	0	0.16	0
Tenting_Area	A1	0	0	0	0
	B1	0	0	0.080	0
Tenting_Hight	A1	0	0	0	0
	B1	0.0080	0.0025	0.17	0.10
PISAr	A1	0	0	0	0
	B1	0.0025	0.0020	0.29	0.0085
Calculated_Real_EF	A1	0	0	0	0
	B1	0.010	0.0070	0.235	0.23

Таблица 3.2.1.3.1. Ефект на анулопластиката върху единадесет параметъра при сравнение на групите A1 и B1 чрез Буутстрап с размерно-еднаква генерация от ECDF

Номер	Параметър	МПКГ		ММПКГ		
		Статистическа значимост	Тип	Статистическа значимост	Тип	Медицинска същественост
1	LVEDV_Index	НЕ	(0, 0)	НЕ	(0, 0)	ОТН+
2	LVESV_Index	ГР+	(–1/2, 0)	ГР+	(+1, +1/2)	ОТН+
3	LA_Volume_Index	ДА+	(–1, 0)	ГР–	(+1/2, +1)	НЕ
4	Vena_Contracta	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+
5	RegVol	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+
6	Coaptation_Length	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+
7	Tenting_Area	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+
8	Tenting_Hight	ДА+	(–1, 0)	НЕ	(+1, +1)	ОТН+
9	PISAr	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+
10	Calculated_Real_EF	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+
11	Real_MR	ДА+	(–1, +1)			

Таблица 3.2.2.1.1. *Pvalue* на статистически тестове за еднаквост на групи A2 и B2 по непрекъснати параметри, проведени преди и късно след операцията, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	**Време**	**Тестове**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
LVEDV_Index	преди	0.27	0	0	0.0010	0	0.020	0.0010	0.97	0.57
	след	0.79	0.0020	0.0015	0.34	0.25	0.0010	0	0.37	0.35
LVESV_Index	преди	0.89	0.0040	0.0035	0.030	0	0.020	0.0010	0.66	0.43
	след	0.615	0.0040	0.0015	0.50	0.32	0	0	0.38	0.38
LA_Volume_Index	преди	0.0010	0	0	0.0015	0	0.84	0.43	0.55	0.21
	след	0.87	0.12	0.060	0.78	0.41	0.10	0.035	0.825	0.46
Vena_Contracta	преди	0	0	0	0.0020	0.0020	0.0005	0.0005	1	1
	след	0.0010	0.0030	0.0010	0.040	0	0.075	0.050	0.33	0.29
RegVol	преди	0	0	0	0	0	0.040	0.040	0.29	0.29
	след	0.010	0.040	0.020	0.11	0.020	0.88	0.43	0.24	0.23
Coaptation_Length	преди	0	0	0	0.14	0.14	0.27	0.16	0.40	0.18
	след	0.040	0	0	0.0010	0	0.75	0.39	1	1
Tenting_Area	преди	0.010	0	0	0	0	0.26	0.18	0.30	0.29
	след	0.12	0.040	0.020	0.25	0.030	0.10	0.020	0.41	0.17
Tenting_Hight	преди	0.030	0	0	0	0	0.40	0.22	1	1
	след	0.55	0.55	0.27	1	1	0.13	0.030	1	1
PISAr	преди	0	0	0	0	0	0.010	0.010	0	0
	след	0.0020	0.0025	0	0.010	0	0.24	0.080	0.98	0.42
Calculated_Real_EF	преди	0.0005	0.0005	0	0.0035	0	0.32	0.16	0.40	0.30
	след	0.090	0.37	0.17	0.55	0.19	0.040	0.010	0.62	0.48

Таблица 3.2.2.1.2. *Pvalue* на статистически тестове за еднаквост на групи A2 и B2 по дискретни параметри, проведени преди и късно след операцията, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	Време	Тестове							
		1	2	3	4	5	6	7	8
Real_MR	Преди (двустранен)	0				0	0.085	0	0
	Преди (едностранен)					0	0.050	0	0
	След (двустранен)	0	0	0.20	0.10		0.11	0.50	0.51
	След (едностранен)		0	0.11	0.485		0.050	0.25	0.27

От Таблица 3.2.2.3.1 се вижда, че:

1) Според МПКГ от 11 параметъра при групи A2 и B2, анулопластиката има статистически значимо благоприятно влияние върху 9 параметъра и липсва статистически значимо влияние на анулопластиката върху 2 параметъра.

2) Според ММПКГ от 10 параметъра при групи A2 и B2, анулопластиката има статистически значимо благоприятно влияние върху 4 параметъра, анулопластиката има гранично статистически значимо благоприятно влияние върху 1 параметър, липсва статистически значимо влияние на анулопластиката върху 4 параметъра и анулопластиката има гранично статистически значимо неблагоприятно влияние върху 1 параметър.

3) Според ММПКГ от 10 параметъра при групи A2 и B2, анулопластиката има медицински съществено благоприятно влияние върху 8 параметъра, липсва медицински съществено влияние на анулопластиката върху 1 параметър и анулопластиката има медицински относително съществено неблагоприятно влияние върху 1 параметър.

4) Общата картина от всички параметри показва статистически значимо и медицински съществено по-благоприятно влияние на анулопластиката при пациенти от група A2 спрямо влиянието на изолираната процедура в псевдо-контролната група B2.

Таблица 3.2.2.2.1. *Pvalue* на статистически тестове за нулевост на промяната на непрекъснати параметри, проведени в група A2 и в група B2, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	Група	Тестове			
		1	2	3	4
LVEDV_Index	A2	0	0	0.030	0.030
	B2	0.0020	0.0020	0.12	0.11
LVESV_Index	A2	0.085	0.040	0.17	0.030
	B2	0.0010	0.0010	0.11	0.11
LA_Volume_Index	A2	0	0	0	0
	B2	0.12	0.060	0.040	0.010
Vena_Contracta	A2	0	0	0	0
	B2	0.0040	0.0035	0.63	0.23
RegVol	A2	0	0	0	0
	B2	0.19	0.11	1	1
Coaptation_Length	A2	0	0	0	0
	B2	0.0095	0.0050	0.33	0.10
Tenting_Area	A2	0	0	0	0
	B2	0.94	0.465	0.52	0.10
Tenting_Hight	A2	0	0	0	0
	B2	0.42	0.21	1	1
PISAr	A2	0	0	0	0
	B2	0.050	0.030	1	1
Calculated_Real_EF	A2	0	0	0.0050	0.0050
	B2	0.030	0.020	0.50	0.10

Таблица 3.2.2.3.1. Ефект на анулопластиката върху единадесет параметъра при сравнение на групите A2 и B2 чрез Буутстрап с размерно-еднаква генерация от ECDF

Номер	Параметър	МПКГ		ММПКГ		
		Статистическа значимост	Тип	Статистическа значимост	Тип	Медицинска същественост
1	LVEDV_Index	НЕ	(–1, –1)	НЕ	(+1, +1)	НЕ
2	LVESV_Index	НЕ	(–1, –1)	ГР–	(+1/2, +1)	ОТН–
3	LA_Volume_Index	ДА+	(–1, 0)	ДА+	(+1, –1)	ДА+
4	Vena_Contracta	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+
5	RegVol	ДА+	(–1, +1)	ДА+	(+1, 0)	ДА+
6	Coaptation_Length	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+
7	Tenting_Area	ДА+	(–1, +1)	ДА+	(+1, 0)	ДА+
8	Tenting_Hight	ДА+	(–1, 0)	ДА+	(+1, 0)	ДА+
9	PISAr	ДА+	(–1, +1)	ГР+	(+1, +1/2)	ДА+
10	Calculated_Real_EF	ДА+	(–1, 0)	НЕ	(+1, +1)	ДА+
11	Real_MR	ДА+	(–1, +1)			

3.3. Сравнение на EF_Number и Calculated_Real_EF чрез Буутстрап с размерно-еднаква генерация от ECDF

Методите МПКГ и ММПКГ могат не само да доказват ефект на някакво медицинско въздействие. Ако ефектът на допълнителното въздействие е доказан предварително, то МПКГ и ММПКГ биха могли да сравняват по два или повече параметъра по техните диагностични способности да измерват ефекта на допълнителното въздействие. Тъй като в точка 3.2 положителният ефект на анулопластиката върху пациенти страдащи от ИБС със значима ИМР е безспорно доказан от два метода с по 11 и 10 параметъра, то широко използваният EF_Number може да се сравни с предложения в дисертацията на доц. Панайотов непрекъснат параметър Calculated_Real_EF.

Статистическите тестове са извършени чрез Буутстрап метод с по 2000 псевдо реалности. Синтетичните извадки се формират с размерно-еднаква генерация от емпирични кумулативни функции на разпределение. Нивото на значимост при всички тестове е прието за 0.05. Получените стойности на Pvalue за статистическите тестове, сравняващи две извадки от пациенти, принадлежащи на групи A1 и B1 за непрекъснатите параметри на състояние EF_Number и Calculated_Real_EF, са показани в Таблица 3.3.1.1.1. Деветте теста, указани в таблицата, са като от първи кластер.

Таблица 3.3.1.1.1. *Pvalue* на статистически тестове за еднаквост на групи A1 и B1 по непрекъснати параметри, проведени преди и късно след операцията, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	**Време**	**Тестове**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
EF_Number	преди	0.24	0.50	0.25	0.96	0.55	0.060	0.020	0.20	0.15
	след	0.40	0.62	0.29	0.71	0.51	0.53	0.28	0.77	0.39
Calculated_Real_EF	преди	0	0	0	0	0	0.58	0.30	0.73	0.415
	след	0.060	0.0090	0.0035	0.020	0	0.17	0.080	0.62	0.36

Получените стойности на Pvalue за статистическите тестове, оценяващи нулевостта на промяната вследствие на операцията на непрекъснати параметри на състояние EF_Number и Calculated_Real_EF или от група A1 или от група B1, са показани в Таблица 3.3.1.2.1. Четирите теста, указани в таблицата, са като в трети кластер.

Таблица 3.3.1.2.1. *Pvalue* на статистически тестове за нулевост на промяната на непрекъснати параметри, проведени в група A1 и в група B1, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	**Група**	**Тестове**			
		1	**2**	**3**	**4**
EF_Number	A1	0.050	0.030	0.080	0
	B1	0.040	0.030	0.24	0.16
Calculated_Real_EF	A1	0	0	0	0
	B1	0.010	0.0070	0.235	0.23

В Таблица 3.3.1.3.1 се обобщава информационната стойност на двата параметъра относно благоприятния ефект на анулопластиката при сравнение на A1 и B1. От Таблица 3.3.1.3.1 се вижда, че:

1) Според МПКГ при групи A1 и B1, анулопластиката има статистически значимо благоприятно влияние върху Calculated_Real_EF, но няма статистически значимо влияние върху EF_Number.

2) Според ММПКГ при групи A1 и B1, анулопластиката няма статистически значимо влияние върху Calculated_Real_EF, но има гранично статистически значимо неблагоприятно влияние върху EF_Number.

3) Според ММПКГ при групи A1 и B1, анулопластиката има медицински съществено благоприятно влияние върху Calculated_Real_EF, но няма медицински съществено влияние върху EF_Number.

4) Общата картина показва, че Calculated_Real_EF има статистически значима и медицински съществена по-висока информационна стойност за благоприятен ефект на анулопластиката отколкото EF_Number.

Таблица 3.3.1.3.1. Сравнение на информационната стойност на два параметъра за групите A1 и B1 чрез Буутстрап с размерно-еднаква генерация от ECDF

Номер	Параметър	МПКГ		ММПКГ		
		Статистическа значимост	Тип	Статистическа значимост	Тип	Медицинска същественост
1	EF_Number	НЕ	(0, 0)	ГР-	(+1/2, +1)	НЕ
2	Calculated_Real_EF	ДА+	(–1, +1)	НЕ	(+1, +1)	ДА+

Аналогичен анализ е проведен и за A2 и B2.

Таблица 3.3.2.1.1. *Pvalue* на статистически тестове за еднаквост на A2 и B2 по непрекъснати параметри, проведени преди и късно след операцията, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	**Време**	**Тестове**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
EF_Number	преди	0.53	0.24	0.12	1	1	0.85	0.43	0.67	0.385
	след	0.105	0.11	0.060	0.70	0.41	1	0.52	0.80	0.43
Calculated_Real_EF	преди	0	0	0	0	0	0.58	0.30	0.73	0.41
	след	0.060	0.0090	0.0035	0.020	0	0.17	0.080	0.62	0.36

В Таблица 3.3.2.3.1 се обобщава информационната стойност на двата параметъра относно благоприятния ефект на анулопластиката при сравнение на A2 и B2. От Таблица 3.3.2.3.1 се вижда, че:

1) Според МПКГ при групи A2 и B2, анулопластиката има статистически значимо благоприятно влияние върху Calculated_Real_EF, но няма статистически значимо влияние върху EF_Number.

2) Според ММПКГ при групи A2 и B2, анулопластиката няма статистически значимо влияние върху Calculated_Real_EF, но има статистически значимо неблагоприятно влияние върху EF_Number.

3) Според ММПКГ при групи A2 и B2, анулопластиката има медицински съществено благоприятно влияние върху Calculated_Real_EF, но има медицински съществено неблагоприятно влияние върху EF_Number.

4) Общата картина от всички сравнения показва, че Calculated_Real_EF има статистически значима и медицински съществена по-висока информационна стойност за благоприятния ефект на анулопластиката, отколкото EF_Number.

Таблица 3.3.2.2.1. *Pvalue* на статистически тестове за нулевост на промяната на непрекъснати параметри, проведени в група A2 и в група B2, чрез Буутстрап с размерно-еднаква генерация от ECDF

Параметър	**Група**	**Тестове**			
		1	**2**	**3**	**4**
EF_Number	A1	0.51	0.27	0.61	0.43
	B1	0	0	0.030	0.030
Calculated_Real_EF	A1	0	0	0	0
	B1	0.010	0.0070	0.235	0.23

Таблица 3.3.2.3.1. Сравнение на информационната стойност на два параметъра за групите A2 и B2 чрез Буутстрап с размерно-еднаква генерация от ECDF

Номер	Параметър	МПКГ		ММПКГ		
		Статистическа значимост	Тип	Статистическа значимост	Тип	Медицинска същественост
1	EF_Number	НЕ	(0, 0)	ДА--	(0, +1)	ДА-
2	Calculated_Real_EF	ДА+	(–1, 0)	НЕ	(+1, +1)	ДА+

ЗАКЛЮЧЕНИЕ

Дисертацията има следните **претенции за приноси**:

1. Създ0адени са три кластера от Буутстрап статистически тестове, както следва:

А) Първи кластер от 9 теста за оценка на различията в непрекъснат параметър между извадки от две популации, реализиран в 5 алгоритъма, пресмятащи Pvalue съответно на: а) Kuiper тест за непрекъснати разпределения; б) двустранен и едностранен тест за средни стойности (без претенции за оригинален принос); в) двустранен и едностранен тест за медиани; г) двустранен и едностранен тест за дисперсии; д) двустранен и едностранен тест за интерквартилни обхвати;

Б) Втори кластер от (2t+1) теста за оценка на различията в дискретен параметър между извадки от две популации, реализиран в алгоритъм, едновременно пресмятащ Pvalue на двустранен тест за дискретни вероятности за всички t дискрета, на едностранен тест за същите и на Pearson тест за дискретни разпределения;

В) Трети кластер от 4 теста за оценка на ненулевостта на промяната в непрекъснат параметър по двойка сдвоени извадки от една популация, реализиран в 2 алгоритъма, пресмятащи Pvalue съответно за: а) двустранен и едностранен тест за средна стойност на промяната; б) двустранен и едностранен тест за медиана на промяната.

2. Разработени са 3 алгоритъма, които при произволни едномодални и платообразни функции на предпочитание определят благоприятността съответно на: а) различията в непрекъснат параметър между извадки от две популации; б) различията в дискретен параметър между извадки от две популации; в) промяната в непрекъснат параметър по двойка сдвоени извадки от една популация.

3. Разработени са 3 алгоритъма, които при дадено ниво на значимост класифицират в отделни типове на съвместна благоприятност и значимост съответно: а) различията в непрекъснат параметър между извадки от две популации; б) различията в дискретен параметър между извадки от две популации; в) промяната в непрекъснат параметър по двойка сдвоени извадки от една популация.

4. Алгоритмизиран е новопредложен метод МПКГ за оценка на ефект от допълнително въздействие върху дискретен или непрекъснат параметър на състоянието, използващ псевдо контролната група, чрез сравняване на двойка извадки от експерименталната и псевдо контолната групи преди допълнителното въздействие и на двойка извадки от двете групи след допълнителното въздействие.

5. Алгоритмизиран е новопредложен метод ММПКГ за оценка на ефект от допълнително въздействие върху непрекъснат параметър на състоянието, използващ псевдо контролната група, чрез сравняване на промяната от допълнителното въздействие в експерименталната група в двойка сдвоени извадки с промяната в псевдо контролната група в друга двойка сдвоени извадки.

6. Предложените тестове и алгоритми са апробирани върху пациенти с ИБС, усложнена със значима ИМР, като от една страна чрез обработка на група от дискретни и непрекъснати признаци е доказано положителното влияние на анулопластиката, а от друга страна са сравнени по диагностична способност за определяне степента на митрална регургитация параметрите ’модифицирана фракция на изтласкване’ и ’фракция на изтласкване’.

Някои **насоки за бъдеща работа** могат да се намерят в следните аспекти:

- Фокусиране на анализа върху реални размити извадки, при които пациентите имат различна степен на принадлежност към съответната генерална съвкупност;
- Приложение на Буутстрап тестовете върху размити извадки;
- Формиране на синтетични извадки по начини, различни от размерно-еднаква генерация от емпирични кумулативни функции на разпределение.

Основни елементи на представения дисертационен труд са **публикувани в шест труда:**

1. Nikolova, N.D., Mihaylova, N., Tenekedjiev, K., Bootstrap Tests for Mean Value Differences over Fuzzy Samples, Proc. IFAC TECIS, Sozopol, Bulgaria, 24-27 September, 2015b (in print)

2. Mihaylova, N., Nikolova, N., Tenekedjiev, K., Outlier Detection in Regression Analysis Using Multiple Testing Procedures, EUREKA 2015: Fifth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support, April 22-24 April, Mexico City, Mexico, 2015 (in print)

3. Nikolova, N., Mihaylova, N., Tenekedjiev, K., Testing Equality of 1D Continuous Distributions Using Bootstrap Kuiper Test Over Fuzzy Samples, EUREKA 2015: Fifth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support, April 22-24 April, Mexico City, Mexico, 2015a (in print)

4. Михайлова, Н., Николова, Н.Д., Тенекеджиев, К., Bootstrap статистически тестове за установяване

на разлики в разсейването на извадки, Доклади на Съюза на учените – Варна, България, Серия Морски науки, стр. 161-172, 2014

5. 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. 3-4, стр. 79-90, 2013b

6. Николова, Н.Д, Михайлова, Н., Тенекеджиев, К., Bootsrap Куипер тест за еднаквост на едномерни непрекъснати разпределения, Международна конференция "Автоматика и информатика'2013", стр. I-135 – I-138, 3-7 октомври 2013, София, България, 2013a

BOOTSTRAP-BASED SIMULATION SYSTEM FOR ANALYSIS OF MEDICAN INFORMATION

PhD Student: Neli Kostadinova Mihaylova
Scientific Supervisor: Assoc. Prof. Natalia Nikolova, PhD
Scientific Consultant: Prof. Kiril Tenekedjiev, DSc. Eng.

Ischemic heart condition with ischemic mitral regurgitation is a global heart condition that relates to malfunction of the valve of the left chamber. Current research shows a tendency for this medical condition to occur not only with middle-aged people, but also with patients between 40-50 years of age. The survey in Bulgaria on this issue is rather limited.

The dissertation of Assoc. Prof. PhD Plamen Panayotov, MD from 2013 identifies criteria to asses the pre-surgical status of patients with ischemic heart condition with ischemic mitral regurgitation. It also analyses the results fof the surgery. Algorithms are formulated to classify the surgical treatment and the effect of surgical interventions over the recovery of the heart for different groups of patients is assessed. The data was collected betweem January 2007 and June 2011 in the Cardiosurgery Ward of St. Marina Hospital – Varna. For this aforementioned period, this ward has 1196 patients with some level of mitral regurgitation. Only 140 were found suitable for the study.

The current thesis offers updated version of the quantitative techniques, mathematical algorithms and simulation approaches, previoly discussed in the thesis of Assoc. Prof. Panayotov. These opdated methodology was developed by the PhD student, under the scientific supervision of Assoc. prof. Natalia Nikolvoa and the scientific consultation of Prof. Kiril Tenekedjiev.

The scope of this work is the information aspect of the aforementioned study. The newly proposed statistical techniques may be used to solve some typical problems arising in the elaboration of medical information systems. The thesis develops and adapts techniques and procedures, algorithms and approaches, which may be appied in a variety of medical information systems. Under lack of control group and under random single-modal and valley preferences over the parameters, it is possible to account for the interrelation between the statistical tests over a random pair of samples, and on the other side it is possible to prove the effect of a particular medical treatment. A series of simulation statistical tests are formulated. These assess the effect of different courses of treatment based on a series of parameters of patients' condition. The proposed information tools is illustrated over medical data for patiens, and some of the above mentioned problems in creating medical information systems were addressed.

The thesis has defined 10 propositions for contributions based on the results. Main parts of the thesis are disseminated in 6 papers, two of which in journals, and of the other 4 conference papers – two in conferences abroad, and two in international conferences in Bulgaria