Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
факультет соціології
кафедра методології та методів соціологічного дослідження

КВАЛІФІКАЦІЙНА РОБОТА

на тему:

«ПОРІВНЯЛЬНИЙ АНАЛІЗ АЛГОРИТМІВ КЛАСИФІКАЦІЇ ОБ'ЄКТІВ ЗА НЕПЕРЕРВНОЮ МЕТРИЧНОЮ ОЗНАКОЮ»

Галузь знань: 054 «Соціологія»
Освітня програма: «Соціальні технології»
Освітній ступінь: бакалавр
Кваліфікація: бакалавр соціології

Виконавець:
Каракай Даніїл Юрійович,
студент 4 курсу

Науковий керівник:

Середа Олексій Сергійович, асистент
кафедри методології та методів
соціологічних досліджень

Бакалаврська робота допущена до захисту
рішенням кафедри методології та методів соціологічного дослідження

Протокол №_____ від «___» _____ 20__ р.

Зав. кафедри _____ доцент Сидоров М.В-С.
підпис

Київ 2020

Реєстрація _____ _____

номер дата підпис лаборанта кафедри

Рекомендовано до захисту _____ _____

підпис наукового керівника ініціали, прізвище наукового керівника

Результат захисту _____ _____

оцінка дата захисту

Голова ЕК _____ _____

підпис ініціали, прізвище

Члени ЕК _____ _____

підпис ініціали, прізвище

_____ _____

підпис ініціали, прізвище

_____ _____

підпис ініціали, прізвище

_____ _____

підпис ініціали, прізвище

Секретар ЕК _____ _____

підпис ініціали, прізвище

ЗМІСТ

ВСТУП **4**

РОЗДІЛ І. Задача класифікації об'єктів в емпіричному соціологічному дослідженні **6**

1.1 Сутність понять систематизації, класифікації, типології 6

1.1.1 Систематизація 6

1.1.2 Класифікація 7

1.1.3 Типологія 10

1.2 Критерії класифікації даних 12

1.3 Використання класифікації за однією змінною в соціальних науках 17

1.3.1. Абсолютний підхід до дохідної стратифікації 18

1.3.1. Відносний підхід до дохідної стратифікації 20

Висновки до Розділу І 21

РОЗДІЛ ІІ. Методи одновимірної класифікації даних **23**

2.1 Рівні інтервали 24

2.1 Квантилі 26

2.3 Стандартне відхилення 27

2.4 Природні межі Дженкса 29

2.5 Head/tail breaks 33

Висновки до Розділу ІІ 35

РОЗДІЛ ІІІ. Порівняльний аналіз алгоритмів одновимірної класифікації за рівнем доходів респондентів **36**

3.1 Масив даних 37

3.2 Критерії оцінки 40

3.3 Одиниці аналізу 41

3.4 Результати 42

Висновки до Розділу ІІІ 49

ВИСНОВКИ **52**

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ **56**

ДОДАТКИ **60**

ВСТУП

Класифікація є одним з основних видів людської інтелектуальної діяльності. Для того щоб зрозуміти факти і явища, розробити принципи, які б пояснювали закономірності і зв'язки, що існують між ними, їх необхідно згрупувати, адже робота з якісно згрупованими даними дозволяє значно спростити аналіз соціальної, або будь-якої іншої реальності. З іншого боку невдала класифікація даних може позбавити зібраний емпіричний матеріал, як би якісно він не був зібраний, всякої наукової та практичної цінності.

Протягом усього минулого століття питанням розробки методів класифікації займалися вчені різних галузей, що пов'язано з розвитком математики. Однак більша частина літератури з класифікації, зокрема з кластерного аналізу, була написана протягом останніх десятиліть і пишеться по цей день.

Більшість літератури з цього питання, яка з'явилася протягом останнього часу, присвячена методам багатовимірної класифікації даних, тобто класифікації об'єктів за великою кількістю ознак. Однак питанню методів класифікації об'єктів за однією метричною ознакою приділено значно менше уваги. Уваги цьому питанню приділено мало як і в англомовній та російськомовній, так і в україномовній літературі. Мотивацією написання цієї роботи в тому числі є відсутність соціологічної літератури, яка б досконало висвітлювала суть, множину підходів та особливості застосування подібної класифікації.

Варто також відзначити і різноманіття способів застосування класифікації за однією ознакою: часто вона є інструментом для попередньої обробки даних або виступає у якості проміжного етапу на шляху до «сутнісної» класифікації (що сама по собі має наукову цінність) або прогнозування. Її використовують

для стиснення та узагальнення даних, а також для знаходження об'єктів, що «вибиваються» з-поміж решти масиву.

Таким чином, все вище сказане дає чітко зрозуміти необхідність детального різностороннього вивчення даної проблеми з метою розробки практичних рекомендацій для подальшого використання алгоритмів одновимірної класифікації об’єктів на практиці.

Ключові слова: класифікація за однією ознакою, групування, дохідна стратифікація, природні межі Дженкса.

Об'єкт: процедура класифікації об’єктів в емпіричному соціологічному дослідженні

Предмет: пізнавальні можливості алгоритмів одновимірної класифікації об’єктів за метричною ознакою у соціологічному дослідженні

Мета: здійснити порівняльний аналіз алгоритмів класифікації об’єктів за неперервною метричною ознакою

Завдання:

1. Дати визначення поняттю класифікація та співвіднести його з іншими суміжними поняттями.
2. Описати підходи та критерії класифікації об’єктів за неперервною метричною ознакою.
3. Описати роль процедури класифікації об’єктів за неперервною метричною ознакою у емпіричному соціологічному дослідженні та підходи до її реалізації.
4. Визначити переваги та недоліки алгоритмів одновимірної класифікації об’єктів за неперервною метричною ознакою.

РОЗДІЛ І. Задача класифікації об’єктів в емпіричному соціологічному дослідженні

1.1 Сутність понять систематизації, класифікації, типології

Огляд літератури з цього питання дозволяє засвідчити, що єдиного підходу до визначення змісту цих суміжних понять не існує. До прикладу, поняття типологізація і класифікація можуть як ототожнюватись, так і відрізнятися. В той же час типологізація згадується як в якості окремого поняття, що позначає більш досконалий процес групування об’єктів порівняно з класифікацією, так і в якості різновиду останньої. [Настасяк, 2015: с. 38-42]

1.1.1 Систематизація

Найбільш широким поняттям, що включає в себе всі інші терміни, які будуть описані нижче є систематизація. Філософський енциклопедичний словник визначає систематизацію наступним чином: «Систематизація - в науці специфічна форма дослідження, пізнавальний процес упорядкування деякої множини розрізнених об’єктів і знання про них. Упорядкування здійснюється шляхом встановлення єдності і відмінності елементів, що підлягають систематизації, визначення місця кожного елемента відносно один одного [Шинкарук, 2002: с. 584] Дослідження такого роду супроводжується логічними операціями порівняння, абстрагування, класифікації, аналізу і синтезу, опису та пояснення.

Втім такий підхід до визначення систематизації не є єдиним. Так, радянський соціолог М.С. Каган в своїй роботі «Системное рассмотрение типов группировки», поміж іншого, яскраво ілюструє відсутність єдиної термінології в питанні сутності поняття систематизація та співвідношення його з іншими суміжними поняттями. В роботі зазначається, що вона може розглядатися як більш широке поняття по відношенню до класифікації, як синонім до

класифікації і навіть як окремий, принципово відмінний від класифікації, спосіб пізнання. [Каган, 2018: с. 51-52]

1.1.2 Класифікація

Класифікацію (від латин. classis – розряд, група і facio – роблю) варто розглядати в якості основної форми систематизації знання, «…поняття якої представляють собою впорядковані групи, за якими розподілені об'єкти деякої предметної області на підставі їх подібності в певних властивостях..» , або результатом такої систематизації [Субботин, 2001: с. 9].

Один із співзасновників нумерологічного напряму в класифікації, мова про який піде нижче, Роберт Сокал давав класифікації наступне визначення «Класифікація це впорядкування об’єктів по їх схожості». [Райзин, 1980: с. 8].

Здійснення класифікації передбачає насамперед існування деяких множин об’єктів. Об'єктом в свою чергу можна назвати все що завгодно, включаючи процеси та дії. Об’єкти, з яких складаються ці множини, повинні представляти собою відокремлені, відносно автономні одиниці, що однозначно відрізняються одне від одного, і в той же час є подібними за певними постійними властивостями. Одним із важливих структурних елементів класифікації також є принцип, за яким сформовані групи поєднуються в єдину систему.

Для подальшого пояснення виникнення плутанини вживання означених в темі підрозділу понять, вважаю важливим зупинитися на поділі класифікації на штучну та природню.

Штучною називають класифікацію, яка представляє її об’єкт в зручному для огляду, запам'ятовування і розпізнавання вигляді. Розподіл об'єктів по групах тут здійснюється на підставі деякого мінімального числа їх постійних, проте не обов'язково істотних для цих об'єктів властивостей. При цьому вибираються такі властивості, які найбільш помітні і які чіткіше, надійніше, ніж інші властивості, відрізняють один від одного об'єкти різних груп.

Встановлений порядок самих груп в такій класифікації носить також зовнішній, формальний характер.

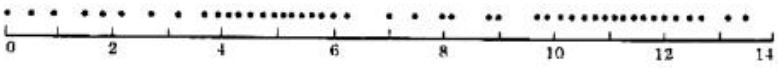
Ідеальна **природна класифікація** передбачає урахування всіх властивостей об'єктів, що класифікуються. Природна класифікація - класифікація, в основі якої знаходиться суттєва ознака, що визначається природою досліджуваних предметів і явищ, їх єством, на відміну від штучної класифікації, в основі якої лежить ознака, що має значення з практичної точки зору для цілей дослідження.

Незважаючи на популярність такого підходу до класифікації класифікації, існує думка щодо неприпустимості подібного розподілу. Основним аргументом проти є той факт, що природна класифікація систематизує реальні об’єкти дійсності, що містять в собі як відомі людині властивості, так і такі, що ще не відкритті для нашого усвідомлення, а отже така класифікація є принципово неможливою. Іншим аргументом може слугувати складність визначення критеріїв, за якими варто розподіляти властивості на суттєві та несуттєві. І дійсно, в питанні класифікації об’єктів реальності не останню роль, якщо не вирішальну, відіграє суб’єктивні вподобання дослідника в виборі тих чи інших наборів ознак та класифікаційних методів.

В підручниках зі статистики часто можна зустріти визначення групування, як розподіл множини об'єктів за суттєвою ознакою, здійсненого з метою приведення даних до більш зручного та впорядкованого виду. [Васнев, 2001: с. 16] Також в учбовій літературі зі статистики можна зіштовхнутися з думкою про те, що класифікація відрізняться від групування тим, що перша створюється на основі кількісних, а не якісних ознак, а також постійністю та незмінністю [Туктарова, 2010: с. 27] В підручнику радянського соціолога Ядова В.А. «Социологическое исследование: методология программа методы» схоже, до першого наведеного в абзаці, визначення дається терміну «просте групування», однак тут, на відміну від попереднього прикладу, поняття

класифікації та групування не протиставляються одне одному – просте групування тут розглядається, як одновимірний різновид класифікації. При цьому ці слова використовуються як синоніми [Ядов, 1972: с. 184].

а)



б)

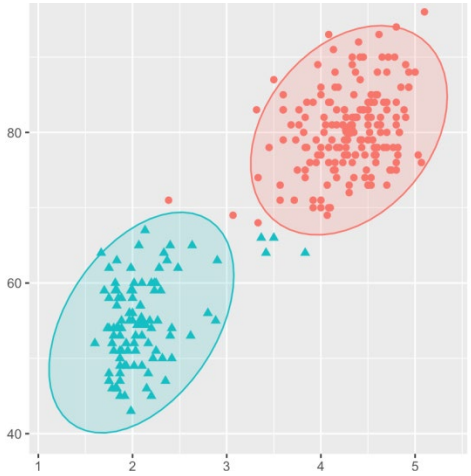


Рис 1.1.1 Приклад протиставлення а) групування та б) класифікації

На мою думку, використання терміну групування для опису подібної процедури (без уточнення того, що це групування є простим) невдале вже хоча б тому, що групування може розглядатися як в якості синоніма до слова класифікація, так і в більш широкому контексті.

Цікавим для розгляду також є співставлення понять класифікації та опису. Коли людина дає визначення якомунебудь об’єкту реальності, вона здійснює операцію доволі подібну до класифікації. Вище було дано визначення класифікації. Для того аби дати визначення поняттю спершу було сказано чим цей предмет є,(основною формою систематизації знання) а потім були перелічені ознаки, сукупність яких відрізняє цей предмет від інших таких предметів. Мені видається, що різниця між цими поняття полягає кількісній наповненості груп.

1.1.3 Типологія

Філософський академічний словник визначає типологію як «1) метод наукового пізнання, в основі якого лежить розчленування систем об'єктів та їх групування за допомогою узагальненої ідеалізованої моделі або типу. 2) Результат типологічного опису або співставлення» [Ильичев, 1983: с. 685]

В результаті такого роду класифікації об'єкти співставляються з деяким типом – зразковим об'єктом групи, що розглядається як такий, що найбільш виражено представляє властивості, які є ключовими для віднесення об'єкту до цієї групи. Всі об'єкти, які мають з ним більше схожого ніж з будь-яким іншим зразковим об'єктом, групуються навколо цього зразкового елемента. Межі такої класифікації не є точно проведеними, адже існують об'єкти, які важко з впевненістю віднести до одного чи іншого типу.

Такий зразковий об'єкт не обов'язково має існувати в емпіричній реальності. Так, для дослідження соціальної реальності відомим німецьким соціологом М. Вебером була запропонована концепція ідеальних типів. Ідеальний тип є певним артефактом, абстракцією, що агрегує в собі реальні риси об'єктів, створюючи тим самим концептуальний образ, що підкреслює особливості цієї реальності.

Дослідник, використовуючи цей концепт, свідомо відсторонюється від існуючої реальності, від всіх її конкретних особливостей і підкреслює, часто до крайності, її окремі властивості для вирішення поставленого перед собою завдання. Така конструкція, при правильному використані, може відігравати значну роль в аналізі емпіричної реальності та стати основою для розуміння останньої.

В соціологічній літературі присутній поділ типологізації на теоретичну та емпіричну. [Ядов, 1972: с. 186] Сутність теоретичної було розглянуто вище, коли ми описували концепцію ідеального типу за Вебером. Основну ідею

емпіричної типологізації, що може бути застосовано до соціології, сформулював американський соціолог Пауль Лазарсфельд.

В статті Бабич Н.С, присвяченій концепції типологізації Пауля Лазарсфельд, зазначається, що з точки зору останнього, основою будь-якої типології є набір всіх її можливих критеріїв. Якщо дослідник має справу з невеликою кількістю ознак, то множина їх комбінацій може бути представлена, або в вигляді системи координат, (в випадку, коли ознака кількісна) або у вигляді перехресної таблиці (в випадку, коли ознака якісна). Це представлення є аналітичним інструментом, що має назву «простір властивостей» [Бабич, 2012: с. 1-2]

В соціологічній літературі, присвяченій методам досліджень, також можна зустріти наступне розрізнення понять типології та класифікації : «Відмітимо, що якщо використаний (используемый) нами термін «типологія» тісно пов’язаний з змістовним характером відповідного розбиття сукупності на групи, з певним видом пізнання, то термін «класифікація» такими властивостями не володіє». [Андреенков, 1982: с. 11]

З огляду на те, що класифікація здійснюється за окремою ознакою, вона не продукує знань про весь набір суттєвих ознак, а тому є нейтральною до сутності явищ. Саме це відрізняє її від типологізації, метою якої є формування уявлень про єдиний набір суттєвих ознак, що розкривають сутність відповідного явища. Натомість систематизація – це впорядкування об’єктів у межах класифікаційних груп, об’єднання їх у ціле, у єдиний цілісний комплекс, систему.

Як ми бачимо, поділ за подібним принципом описувався нами вище при розгляді різновидів класифікації. Типологію та природню класифікацію можна просто розглядати як синоніми. Навіть більше, такий підхід знаходить в філософському словнику під редакцією радянсько філософа Івана Фролова. Типологія в ньому визначається, як «класифікація за суттєвими ознаками».

[Фролов, 2001: с. 247] Терміну класифікація в соціологічній літературі, цілком може відповідати термін штучна класифікація. Окремим підвидом штучної класифікації за однією змінною є «просте групування».

Підсумовуючи, можна сказати, що у кожній предметній області виникає своя термінологія, в зв'язку з чим може виникнути плутанина. Тим не менш, виділивши певний час вивченню питання визначення окреслених понять, дослідник спрощує сприйняття цих термінів в різних дисциплінах і може знайти їм відповідники.

1.2 Критерії класифікації даних

Загалом на проблему класифікації даних існує декілька філософських поглядів – реалістичний та номіналістичний.

Коріння реалістичної концепції класифікації знаходиться в філософії Платона. Ця концепція передбачає, що в основі всього різноманіття речей лежать постійні «сутності», а задача науки полягає в тому, щоби відкрити та описати ці сутності, що складають істину природу речей. Розподіл речей по групам відбувається на основі тих сутностей, до яких вони відносяться і від яких отримують назву.

Відповідно до цієї логіки, якщо ряд об'єктів мають загальну назву, вони мають також загальну «сутність». Наприклад, хоча олівців може бути багато, існує лише одна «сутність» олівця. Адже коли ми вживаємо слово «олівець» ми розуміємо щось відмінне від кожного конкретного олівця. Так само як відображення олівця в дзеркалі є лише відображенням олівця, але ніяк не «реальним» олівцем, так і різні окремі олівці нереальні, а є лише копіями «сутності», яка являє собою один єдиний олівець. Концепція має метафізичний характер. Таким чином, реалісти ставлять перед собою задачу відкрити вже існуючу систему, яка може бути тільки одна.

Інша, номіналістична концепція класифікації, також має давню історію та сягає корінням в середньовічну філософію. Номіналісти вважають, що ніяких загальних сутностей не існує, а існують тільки одиничні об'єкти. Групи ж, в які вони об'єднуються, формуються по тим ознакам, які створює людський розум, що надає тим чи іншим проявам реальності імена. З цієї точки зору будь який із варіантів класифікації одних і тих самих об'єктів є рівноправним, оскільки є породженням діяльності людського розуму. Таким чином, в нашій роботі можна виділити перший підхід до встановлення критерію класифікації – **довільний.**

Представлені концепції є двома крайностями в підходах до класифікації. Зрозуміло, що існують підходи і методи класифікації однієї множини речей, які краще відображають наше уявлення про існуючу реальність ніж інші (або принаймні краще справляються с задачами, які ставить перед собою дослідник) та спрощуючи реальність, одночасно краще відкривають дорогу до отримання знання про неї. При цьому потрібно усвідомлювати, що цілком можливо ми ніколи не зможемо дізнатися про абсолютно всі властивості об'єктів для того аби здійснити ідеальну класифікацію

Описаний в попередньому підрозділі поділ на природні та штучні класифікації певною мірою є способом узгодження цих двох концепцій. Штучна класифікація являє собою в першу чергу інструмент для аналізу, систему опису множини об'єктів, зручною для огляду, розпізнавання та запам'ятовування. Виходячи з цього, критерієм якості побудови такої класифікації повинна бути зручність аналітичного інструменту .

При створенні природньої класифікації виходять з врахування всієї сукупності властивостей об'єктів, що піддаються класифікації. Об'єднання відбувається на основі їх найбільшої подібності між собою, тобто на основі постійних спільних властивостей , що визначають множину інших властивостей цих об'єктів, які по цій причині є джерелом максимальної інформації про ці об'єкти [Субботин, 2001: с. 34]

В той же час фактом є те, що існуючі класифікації, якими б правильними вони не здавалися на сьогоднішній день, необхідно переглядати по мірі отримання нового наукового знання, оскільки ті ознаки, підстави для класифікації, які «відображають суть речей» зараз, можуть виглядати безглуздо в майбутньому.

Усвідомлення цього факту призвело до кризи інтуїтивного підходу до створення природніх класифікацій. Це вимагало від науки формулювання об'єктивних критеріїв розподілу елементів за групами. Спробою відповіді на це питання став розвиток кількісного підходу з розробкою математичних алгоритмів здійснення такої класифікації (нумерична таксономія). Вважалось, що використання методів стандартної обробки даних та оцінки результатів дасть можливість створювати такі класифікації, так як це дасть можливість зменшити суб'єктивний вплив дослідника та підвищити об'єктивність класифікації.

Поступово ілюзії щодо математичних методів розсіялися. Методів і способів автоматичної кластеризації стало так багато, що останні вже й самі могли бути об'єктами класифікації. Втім, причиною розчарування стала не їх кількість, а те, що вони могли давати неузгоджені результати для одних і тих самих даних. Зрозуміло, що ці відмінності в результатах обумовлені технічними причинами (тобто відмінностями використовуваних методів), але часто виникають ситуації, коли у нас немає можливості їх зняти, оскільки неясно, яку нумеричну класифікацію з великого числа розроблених слід вважати найкращою, і на яких підставах.

Втім, розвиток нумеролічного напрямку класифікації приніс багато корисних доробок, в тому числі в плані критерії якості класифікації. Так американський статистик Джозеф Крусскал, брат іншого відомого американського статистика, визначав критерій природності груп, утворених в результаті кластеризації, наступним чином: «Ми називаємо кластер природнім,

якщо належність до нього природнім чином і доволі добре визначається даними» [Райзин, 1980: с. 24]

В тій же роботі визначаються два критерії формування кластерів. Ними можуть слугувати компактність кластерів та/або чіткість меж, що їх роз'єднують.

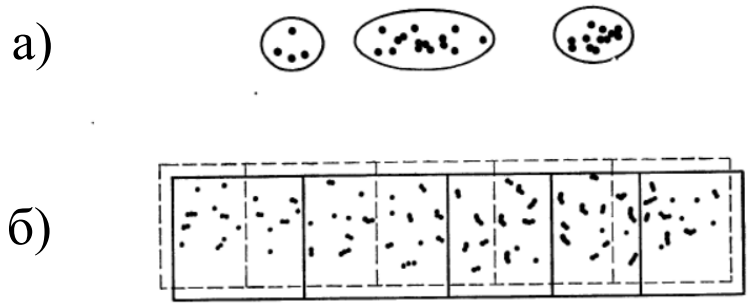


Рис 2.1.1 Приклад а) природньої та б) не природньої класифікації

Схожу позицію висловлював і співзасновник нумеричної таксономії Роберт Сокал: «Природна політетична класифікація дозволяє зробити два типи прогнозів щодо значень ознак. Ці значення повинні бути однорідними всередині таксонів, але між таксонами розподіл значень повинен якісно відрізнятися». [Райзин, 1980: с. 10]. Оскільки це не є очевидним, варто пояснити, що таке політетична група. Політетична група – це група, яка характеризується великим числом ознак, кожна з яких взята окремо не може бути підставою для віднесення об'єкта до групи.

Питанню критеріїв класифікації конкретно за однією метричною ознакою приділяється мало уваги. Так, наприклад, представлення статистичних даних в класах при групуванні за однією ознакою рідко розглядається з точки зору статистики. Більшість правил групування, що можна зустріти в учбовій літературі, має скоріше рекомендаційний характер, що направлений на зручність сприйняття. Так само і в емпіричних роботах, де дослідники застосовують ті чи інші підходи та алгоритми класифікації об'єктів за однією метричною ознакою для подальшого аналізу короткий огляд яких буде

представлено нижче, рідко можна зустріти аргументацію того, чому було здійснено саме таке групування, а не інше.

В випадку статистичних даних можна встановити наступне правило для виділення класів: якщо статистично відмінні дані об’єднуються в один клас, то змістовність відображення структури даних буде зменшено в порівнянні з вихідними даними. В той же час якщо класи, які не є суттєво відмінними з точки зору статистики, будуть розділені на два або більше класів, то таке групування буде надавати спотворену інформацію про структуру даних. [Lajos, 1987: c. 143]

Поняття «якісна відмінність» і «внутрішня однорідність» взаємно доповнюють один одного. На перший погляд може здатися, що одне з них повністю визначає інше. Це не так, хоча на емпіричному рівні досягнення того чи іншого може відбуватися за допомогою одних і тих же засобів.

Якісна відмінність розуміється як різнотипність об'єктів, що належать різним групам. Якісно відмінними є не всі внутрішньо однорідні групи. Відмінність груп не може визначатися тільки процесом розбиття. Після процедури розбиття на класи, яка в певному сенсі носить формальний характер, з'являється необхідність в об'єднанні деяких класів в однотипні групи або в їх поділі на однотипні групи.

Внутрішня однорідність може розумітися і як кількісна однорідність, і як якісна. Внутрішня однорідність - близькість, схожість, однотипність об'єктів, що належать до певної групи. Вона досягається на декількох стадіях типологічного аналізу. Зрозуміло, будь-яка математична модель розбиття емпіричних об'єктів дозволяє виділити однорідні, якісно відмінні один від одного класи. Але це тільки один із способів досягнення однотипності.

1.3 Використання класифікації за однією змінною в соціальних науках

Класифікація об’єктів за однією ознакою має широке застосування в якості одного з інструментів аналізу соціальних явищ та часто використовується при проведенні емпіричних соціологічних досліджень. Наприклад, така класифікація часто використовується в якості інструменту, що слугує потребам здійснення геопросторового аналізу даних.

Так, класифікацію об’єктів для візуалізації просторових даних можна застосувати для вивчення просторових закономірностей електоральних, політичних орієнтацій, тенденцій споживання певного комерційного продукту або послуги. Також існує можливість дослідити структуру географічного розповсюдження ціннісних установок, певної моделі соціальної поведінки, культурних патернів, стереотипів тощо. Фактично одновимірна класифікація просторових даних при здійсненні геопросторового аналізу дозволяє спростити візуалізацію перетину географічного та соціального простору, дослідження їх динамічної взаємодії, що в свою чергу дозволяє змоделювати або екстраполювати розвиток соціальних тенденцій.

Іншим яскравим прикладом застосування одновимірної класифікації за неперервною метричною ознакою є стратифікація за рівнем доходів, яка є однією з поширених підходів до аналізу структури суспільства. Не зважаючи на відносну простоту, такий підхід, при усвідомлені обмежень, дозволяє вирішувати широкий спектр задач, пов’язаних з аналізом структури суспільства.

Зокрема такий інструмент дозволяє отримати кількісні оцінки найменш та найбільш благополучних груп у фінансовому відношенні, визначити ступінь дохідної нерівності в суспільстві, відслідковувати процеси її динаміки з подальшим аналізом соціальних та економічних чинників, під впливом яких відбуваються зміни в структурі дохідної нерівності.

В силу того, що основним індикатором для її побудови є універсальний показник рівня доходів, саме така модель стратифікації зручна для міжнародних порівнянь співвідношення соціальних груп, за умови вироблення єдиних принципів встановлення меж останніх. Велика її роль і в сфері соціальної політики, де на підставі показника доходів визначаються групи бідних і нужденних, які стають об'єктом соціальної підтримки.

Втім, попри, або можливо навпаки – через, очевидну важливість вивчення питання дохідної стратифікації, на сьогоднішній день світова наукова спільнота не дійшла до єдиного розуміння того, як повинна виглядати шкала доходів при побудові моделі дохідної стратифікації. Найбільші питання викликає те, де саме повинні проходити межі різних дохідних груп, щоб забезпечувати їх відносну гомогенність та диференціювати їх між собою. Так, традиційно виділяють два підходи до вирішення цього питання – абсолютний та відносний. [Хахулина, 1995: с. 18]

1.3.1. Абсолютний підхід до дохідної стратифікації

Абсолютні методи стратифікації за доходами припускають, що кордони між прибутковими групами задаються апріорі, як якась певна сума грошових одиниць. Ці методи зазвичай використовуються для аналізу соціальної структури країн, що розвиваються. [Тихонова, 2017: с. 26]

В рамках абсолютного підходу виділяють такі напрямки розрахунку меж дохідних груп: [Аникин, 2018: с. 240 -243]

1) Межа бідності та її мультиплікація
2) Методика Світового Банку

Наказ «Про затвердження Методики комплексної оцінки бідності» від 2017 року визначає межу монетарної бідності як «рівень доходу (витрат), нижче від якого є неможливим задоволення особою основних потреб.» В тому ж

місці зазначається, що вартісне значення межі монетарної бідності є основою для віднесення особи до категорії бідних.[1]

Розрахунок меж монетарної бідності, що проводять державні статистичні служби, доволі часто здійснюються на основі співвідношення доходів або розходів громадян з прожитковим мінімумом, тобто з величиною, що достатньою для забезпечення нормального функціонування організму людини, збереження його здоров'я, набору продуктів харчування, а також мінімального набору непродовольчих товарів та послуг, необхідних для задоволення основних соціальних і культурних потреб особистості. [2]

Класифікація доходів, здійснена Світовим Банком, розроблялася з метою аналізу рівня бідності на основі єдиних стандартних показників для всього світу. Так в 1990 році група дослідників Світового Банку запропонували вимірювати чисельність бідного населення за допомогою стандартів, які прийняті в найбідніших країнах. Після цього ці дані були переведені в єдину валюту на основі паритету купівельної спроможності. Було виявлено що значення межі бідності в досліджуваних країнах коливалася навколо 1 долара США. Цей показник ліг в основу розробки універсального показника рівня бідності.[3]

Іншими словами, класифікація за метричною ознакою доходу утворює дві групи населення – «бідні» і «не бідні», межі груп встановлювались аналітично на основі експертної оцінки дослідників. Згодом були спроби здійснення більш диференційованого групування. Наводжу приклад подібної класифікації, здійсненої Світовим Банком, де межі дохідних груп встановлені наступним чином [Vakis, 2015: c. 8]:

---

[1] Наказ про затвердження Методики комплексної оцінки бідності [Електронний ресурс]. – 2017. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/z0728-17.
[2] Наказ про прожитковий мінімум [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/966-14.
[3] FAQs: Global Poverty Line Update [Електронний ресурс] – Режим доступу до ресурсу: https://www.worldbank.org/en/topic/poverty/brief/global-poverty-line-faq.

- Надзвичайно бідні (2,5 долари на день)
- Помірно бідні (до 4 доларів на день)
- Незахищені (від 4 до 10 доларів на день)
- Середній клас (від 10 до 50 доларів на день)
- Заможні (більше 50 доларів на день)

1.3.1. Відносний підхід до дохідної стратифікації

Відповідно до відносного підходу, межі дохідних груп при класифікації об'єктів за однією змінною визначаються не через встановлений абсолютний стандарт, а враховуючи ступінь соціальної нерівності в конкретній державі або регіоні, кожен з яких має власні локальні стандарти доходів. В цьому випадку бідність визначається як недостатність ресурсів, які необхідні для підтримки того рівня життя, який вважається прийнятим в межах даної культури.

Єдиного способу визначення порогу бідності, як і інших меж дохідних груп, не існує. Втім, більшість способів групування базуються навколо наступних понять з області математичної статистики: [Foster, 2013: с. 2]

1. Мода (або заданий дослідником відсоток від неї)
2. Середнє значення (або заданий дослідником відсоток від неї)
3. Квантиль

Так, Міністерство економіки України в 2010 році до бідних відносила тих людей, рівень доходу яких був менший за 75 відсотків від значення медіани доходу. [22, стр 34] Євростат для подібного показника для конкретної країни визначає межу бідності на рівні 60 відсотків від медіани доходу. [4]

Поширеним підходом до групування об'єктів по доходу є групування через квінтелі. Класифікація за квінтелями передбачає, що «середній» середній клас у вузькому сенсі - це середні 20 відсотків розподілу доходів (Q3). Нижче

[4] Посилання на офіційний сайт Євростату: https://ec.europa.eu/eurostat/web/products-datasets/product?code=sdg_10_30

знаходиться «нижній» середній клас (Q2), вище - «верхній» середній клас (Q4). Три квінтиля, взяті разом (наприклад, Q2 + Q3 + Q4, що становить 60 відсотків від розподілу доходів), утворюють весь середній клас. Q1 - найнижчий клас, Q5 - найвищий клас доходу. Перевагою такого підходу є його простота, очевидним недоліком – принципова незмінність кількісної наповненості виділених груп.

В емпіричних роботах, де дослідники застосовують ті чи інші підходи та алгоритми класифікації об'єктів за однією метричною ознакою для подальшого аналізу, рідко можна зустріти аргументацію того, чому було здійснено саме таке групування, а не інше. В якості прикладу спроби продемонструвати таку аргументацію можна навести роботу російського соціолога Тетяни Богомолової. [Богомолова, 2006] В роботі для представлення економічної стратифікації населення Росії в якості ознаки для класифікації був вибраний дохід на одну споживчу одиницю, який в свою чергу був нормований на регіональний прожитковий мінімум. Задля змістовного відображення структури дохідної нерівності в роботі були використані нерівні інтервали. Адекватність моделі перевірялась за допомогою критерія Колмогорова-Смірнова, що порівнює теоретичну функцію розподілу з емпіричною.

Висновки до Розділу I

Підсумовуючи все вище сказане, можна стверджувати, що незважаючи на відсутність єдиної термінології, факт чого зумовлює необхідність постійного уточнення понять під час проведення наукових досліджень, класифікація є широким поняттям, яке може розглядатися як метод наукового пізнання, результат якого є цінним сам по собі. В такому випадку вона може розглядатися як синонім типології. Також вона може розглядатися як утилітарний інструмент, використання якого має на меті спростити аналіз соціальних та природніх явищ, однак не продукує наукового знання.

Об’єкти, з яких складаються множини, що піддаються класифікації, повинні представляти собою відокремлені, відносно автономні одиниці, що однозначно відрізняються одне від одного, і в той же час є подібними за певними постійними властивостями. При цьому оцінка якості виконання цих критеріїв групування може бути суб’єктивною і варіюватися в залежності від поставлених перед дослідником завдань. В випадку класифікації на основі кількісних показників до певною міри стає можливо математично обґрунтувати однорідність та диференційованість груп.

Класифікація широко використовується в соціальних науках в тому числі класифікація за однією метричною ознакою. При цьому навіть для здійснення класифікації за однією ж самою ознакою використовуються різні методи групування даних. Втім питання якості такої класифікації розглядається рідко.

РОЗДІЛ ІІ. Методи одновимірної класифікації даних

Як же було зазначено раніше в значній частині учбової літератури операція об'єднання значень у відносно однорідні групи за однією ознакою називається групуванням. В цій же літературі можна зустріти поділ групування на наступні види: аналітичне, типологічне та структурне групування. [Васнев, 2001: с. 16]

Останнє представляє собою розділення однорідної сукупності на групи по тій чи іншій класифікаційній ознаці із не нульовою дисперсією.

В якості прикладу групування такого виду можна навести вікові та статеві групи, в попередньому розділі був розглянутий досвід групування за доходом. Основною задачею групування такого плану є вивчення структурного складу тої чи іншої сукупності змін на основі тої чи іншої класифікаційної ознаки. Вивчення структурних змін в свою чергу допомагає вивчати закономірності соціальних явищ. При групуванні даних за метричною ознакою закономірно виникає два питання: «Скільки потрібно виділити груп?» та «Якого розміру повинні бути ці групи?» Від в залежності від того як даються відповіді на них, в залежності від того як групуються вихідні дані, можуть бути виявлені різні властивості досліджуваного явища. Нижче будуть розглянуті способи дати відповідь на ці питання.

При розгляді методів простого групування в цій роботі я буду орієнтуватися на класифікацію методів одномірної класифікації, яка є прийнятою в картографії [Evans, 1977: с. 100-102], доцільність якої знаходить своє відображення в статистичних бібліотеках, присвячених цьому питанню – рівні інтервали, квантилі, стандартне відхилення та природні межі Дженкса. Побіжно будуть наведені приклади модифікації цих алгоритмів та описано алгоритм класифікації,призначений для розподілу з довгим хвостом.

2.1 Рівні інтервали

В тому випадку, якщо ми вважаємо, шо розмір груп, які ми виділяємо, має бути однаковий, відповіді на вищеокреслені питання зазвичай можна знайти в літературі присвяченій побудові гістограм. Вважається, що при об'єднанні, класів повинно бути стільки, щоб можна було виділити особливості розподілу випадкової величини. Таким чином, групування даних за невеликим числом класів за рідким виключенням є небажаним, так як скорочення кількості класів забезпечує спрощення за рахунок втрати корисних деталей. В той же час велике число класів може призвести до втрати сенсу у самому групуванні, як інструменті аналізу. [Evans, 1977: с.98]

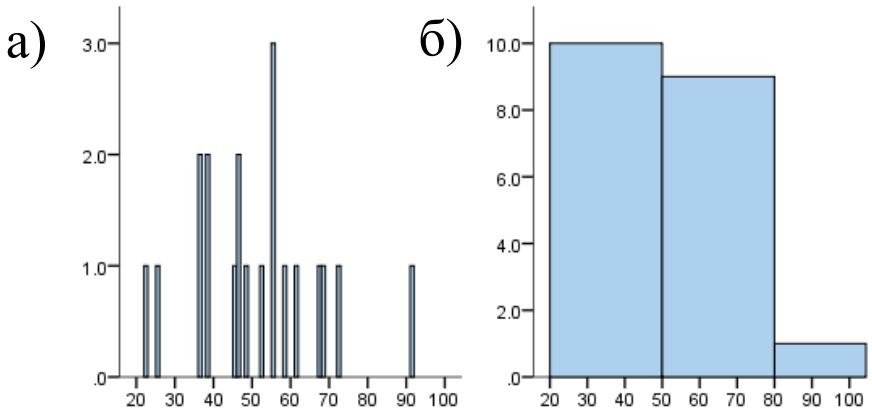


Рис 2.1.1 Приклади невдалого вибору кількості класів: а) виділення занадто великої кількості груп; б) малої кількості груп.

Для визначення числа класів існують різні формули. Наведемо приклади деяких з них:

1. Формула Стерджеса

$$k = [1 + 3{,}22 * \lg N]$$

k – кількість інтервалів;

N – кількість об’єктів;

[x] – ціла частина числа.

Даний підхід часто наводиться в учбовій літературі зі статистики, однак піддається критиці через непридатність роботи з великим числом об’єктів (більше 200) та вимоги нормальності розподілу даних [Hyndman, 1995: с. 1]. В цієї формули існує модифікація, яка робить її більш придатною для розбиття на інтервали ненормально розподілених даних.

2. Формула Скотта

$$h = \frac{3{,}5 * \sigma}{\sqrt[3]{n}}$$

h – ширина інтервалу;

σ – стандартне відхилення.

Формула вважається більш адекватною, ніж описана вище. Так як при розрахунку використовується стандарте відхилення, вимога нормальності розподілу даних зберігається. Кількість груп в цьому випадку розраховується шляхом поділу довжини діапазону значень на встановлену довжину інтервалу.

3. Формула Фрідмана-Дьяконісона

$$h = 2\frac{IQR(x)}{\sqrt[3]{n}}$$

h – ширина інтервалу;

IQR – міжквартильний розмах.

Варто зазначити, що ні одна із представлених формул, як і тих, що тут не представлені, не є універсальною і всі вони носять скоріше характер рекомендації. Цілком можливо, що в радянській та пострадянській учбовій літературі присвяченій основам статистики формула Стерджеса є більш поширеною через простоту обрахунку. На сьогоднішній день в більшості статистичних бібліотек на різних мовах програмування більшість з цих формул закладені в якості атрибуту функції побудови гістограм, що значно спрощує користувачу процедуру вибору і порівняння різних варіантів групування.

Отже, метод рівних інтервалів розбиває діапазон значень ознаки на під діапазони рівного розміру. Цей спосіб групування даних є простим та інтуїтивно зрозумілим незважаючи на те, що визначення оптимальної кількості груп потребує певних зусиль, якщо вона не була визначена дослідником апріорно. Основним недоліком методу є його сильна залежність від нормальності розподілу. Окрім цього, за умови наявності значних розривів, деякі групи можуть бути взагалі не наповненими.

2.1 Квантилі

Як було продемонстровано **в розділі 1.3,** в соціальних емпіричних дослідженнях в якості одного із способів групування даних за однією метричною ознакою часто використовують квантилі. Розглянемо його «найдрібніший» різновид – процентиль.

Найбільш простим способом зрозуміти, що таке процентиль це уявити, що рівно 100 людей мають якийсь числовий атрибут, наприклад зріст. Після чого відсортувати їх по зростанню. В такому випадку перша людина в цьому відсортованому ряді метафорично буде відповідати першому процентилю, десята – десятому (дециль), двадцята - двадцятому (квінтель), двадцять п’ята - двадцять п’ятому (квартиль), п’ятдесята - п’ятдесятому (медіана). Зі сказаного стає очевидно, що розрахунок процентилів заснований на впорядкуванні рангів. Через порядкову природу числове вираження процентилів в цілих числах є умовним.

Порядкова природа цього підходу є його головним недоліком. Очевидно, що розміри виділених таким чином груп на протязі всього діапазону значень не будуть відрізнятися. При нормальному розподілі ознаки, навколо середнього значення будуть знаходитися групи з меншим розміром діапазону значень, а ті, що віддалені від середнього відповідно будуть мати більший діапазон значень. Відповідно, різниця між середнім значенням ознаки, наприклад в десятому та

двадцятому процентилях, це зовсім не те саме, що різниця в п’ятдесятому та шістдесятому процентилях. Іншим великим недоліком з точки зору класифікації є те, що об’єкти з одним і тим самим значенням єдиного притаманного їм атрибуту можуть попасти в різні класи. Основною перевагою підходу є його простота в сприйнятті та обчислені.

2.3 Стандартне відхилення

В емпіричних дослідженнях психологів часто можна зустріти групування результатів оцінок психологічних тестів за допомогою переведення значень «сирих балів» в стандартизовані оцінки, що показують місце індивідуального результату тесту відносно показників обстежуваної вибірки. Основними різновидами таких оцінок є z-бали, t-бали, стени та станайни.

Z-бали, або z-значення представляють собою самі базові з всіх стандартизованих оцінок, та використовується при розрахунку всіх інших. Вони засновані на гіпотезі про нормальний розподіл. Кожний нормальний розподіл є варіацією стандартного нормального розподілу, що за визначенням має середнє значення, яке дорівнює нулю та стандартне відхилення, що дорівнює одиниці.

Переведення сирих значень в еквіваленті z-значення передбачає просту операцію лінійного перетворення нормально розподілених даних до вигляду стандартного нормального розподілу, в результаті чого середнє значення та стандартне відхилення z-оцінок відрізняються від оригіналу, хоча форма розподілу оригіналу зберігається. Це означає, що якщо розподіл оригіналу значно відрізняється від нормального, процедура стандартизації ніяк не може змінити форму розподілу ознаки. Одним із способів вирішення проблеми ненормальності розподілу полягає в тому, щоб нормалізувати результати тестів шляхом представлення їх у вигляді процентилів. Сама операція вирахування z-значень виглядає наступним чином: від кожного елемента вектору значень

віднімається середнє арифметичне. Отримана різниця ділиться на стандартне відхилення всього вектору значень.

Таке перетворення прекрасно підходить для того, аби порівнювати результати отримані на різних вибірках, втім z-оцінки не є дуже зручними в плані встановлення меж класів, оскільки більшість значень, отриманих в результаті такого перетворення, матимуть велику дробову частину. В якості недоліку можна розглядати і те, що такий розподіл має від'ємні значення, що не завжди зручно для сприйняття. Для подолання окреслених незручностей в психометричній практиці застосовують і інші системи скорингу. Відрізняються вони між собою тільки встановленим значенням середнього та стандартними відхиленням:

1. Т-оцінка

Розподіл значень такої шкали має середнє значення 50 та стандартне відхилення, що дорівнює 10, звідки і походить назва оцінки (десять - ten). Її можна розрахувати за формулою:

$$\mathrm{h} = 50 + 10 * \frac{x - m}{\sigma}$$

Оскільки Т-бали часто округляються до найбільшого цілого числа, вони утворюють доволі просту шкалу, придатну до групування за цілими значеннями.

2. Стени

Стени, або «стандартні десятки», широко використовуються в тестах оцінки особистості. Вони приймають значення від одного до десяти. Середнє значення по шкалі в такому випадку знаходиться в точці 5,5. Стандартне відхилення дорівнює двом. Зручність використання таких оцінок в основному полягає в простоті сприйняття .

3. Станайни

Стенайни, або «стандартні дев’ятки», як можна здогадатися з назви, утворюють шкалу, значення якої варіюються від одного до дев’яти. Вона представляє собою різновид стена, однак на відміну від першого, має ціле середнє значення. Як і стени, станайни розраховуються за формулую розрахунку z-значень, однак до чисельника додається число, яке і буде середнім значенням розподілу, а чисельник множиться на цікаве нам значення стандартного відхилення.

2.4 Природні межі Дженкса

Метод класифікації Дженкса, або, як його ще називають, методом оптимізації Дженкса був представлений світові в 1977 році і розроблявся спеціально для аналізу географічних даних. Алгоритм був розроблений на базі методу «точної оптимізації» Фішера та представляє собою один із методів кластеризації даних та призначений для виявлення природних меж груп, утворених при класифікації за метричною ознакою. [North, 2009: с.1]

Сутність терміну «природні межі» сягає корінням до описаного в першому розділі підходу до поділу класифікації на природню, тобто найкращу можливу, яка випливає з урахування всіх можливих ознак, які властиві об’єктові, що піддається класифікації, та на штучну, призначення якої є інструментальним і яка сама по собі не є цінною. Однак класифікації можуть підлягати і абстрактні об’єкти, в тому числі такі, які мають тільки одну властивість.

Під природними межами груп, в випадку розбиття метричної ознаки на інтервали, є такі межі інтервалів, при яких варіація всередині кожної з груп є максимальною , а отже значення в середині виділених діапазонів є максимально близькими один до одного. Втім, не зважаючи на чіткість меж груп, що виділяються, алгоритм не може гарантувати їх статистичну відмінність, що є важливим при роботі з вибірковими сукупностями.

На відміну від вищеописаних стандартних підходів до розбиття груп на інтервали, межі яких вираховуються достатньо просто, вирахування природніх меж за допомогою алгоритму оптимізації Дженкса вручну є доволі проблематичним. Ручний перебір комбінацій може зайняти багато часу навіть у машини, а виконання таких дій людиною займе значно більше часу. Окрім цього, на відміну від машини, людина може здійснити помилку при розрахунках та виборі найкращої моделі.

В значній частині наукової літератури кластеризація визначається як багатовимірний вимір класифікації даних, в зв'язку з цим виникає питання чи можна розглядати алгоритм класифікації Дженкса в якості різновиду кластеризації даних. Втім, серед наукової літератури можна знайти згадки про цей метод просто в якості алгоритму кластеризації, або в якості одновимірного різновиду відомого алгоритму кластеризації k-means[5]. Тому відповідь на поставлене вище питання залишається предметом дискусій. Варто зазначити, що станом на сьогоднішній день, окрім класифікації Дженкса, існує ще один алгоритм, що представляє собою одновимірну реалізацію методу k-means.

Взагалі, суть кластеризації полягає в тому, щоб поділяти певну множину об'єктів на кілька порівняно однорідних груп, які називають кластерами. Кластери будуються таким чином, аби відстань між об'єктами з одного кластера була значно меншою, ніж відстань між об'єктами, що належать до різних кластерів. Інакше кажучи, об'єкти всередині одного кластера більш подібні між собою, ніж об'єкти з різних кластерів.

Важливо відзначити, що як і у випадку неієрархічних методів кластеризації, таких як k-means, різновидом якого можна вважати алгоритм

[5] Приклади згадок в якості різновиду k-means в рецензованих журналах: An unsupervised data-driven method to discover equivalent relations in large Linked Datasets / Z.Zhan, A. Gentile, E. Blomqvist, F. Ciravegna. // semantic web journal. – 2016. – №8. та Peters M. A scalable preference model for autonomous decision-making / M. Peters, M. Saar-Tsechansky, W. Ketter. // Machine Learning. – 2018.; Приклад згадки в якості алгоритму кластеризації - Mac Carrona P. Calling Dunbar's numbers / P. Mac Carrona, K. Kaskib, R. Dunbara. // Social Networks. – 2016. – №47. – С. 151–155.

пошуку природних меж Дженкса, кількість кластерів (k) для розбиття задається користувачем. Звичайно, не всі значення k призведуть до побудови однорідних кластерів, що можна відрізнити один від одного, тому доцільно запускати алгоритм кілька разів з різними значеннями k. Після чого обирати для якого з k певні характеристики або графіки виглядають краще, або приймати кластерне рішення, що, на думку дослідника, призводить до найбільш змістовної інтерпретації. Також можна прийняти це рішення автоматичним способом, тобто дозволити комп'ютеру спробувати всі (або багато) можливих значень k і вибрати той, який найкращим чином відповідає деякому числовому критерію. [Kaufman, 2005: c. 38]

Послідовність виконання алгоритму класифікації природніх меж Дженкса виглядає наступним чином:

Крок 1 Здійснюється вибір метричної ознаки (x) для класифікації та обирається число класів для розбиття (k).

Крок 2 Набір k-1 випадкових або рівномірних значень формується в діапазоні від найбільшого до найменшого значення ознаки, які використовуються в якості меж початкових груп.

Крок 3 Для всього діапазону в цілому та для кожної групи окремо розраховується середнє значення та сума квадратів відхилень від вирахуваного середнього.

Крок 4 Після перевірки суми квадратів відхилень від середнього в середині кожної групи, приймається рішення перемістити один об’єкт групи з найбільшим значенням вищезгаданої суми до групи з її меншим значенням.

Операція повторюється до тих пір, допоки сума квадратів відхилень від середнього в середині класів не стане мінімальною. Згаданий числовий критерій для оцінки якості розбиття, що може бути застосований дослідником для прийняття рішення вибору кількості груп, розраховується як різниця між сумою

квадратів відхилень для всього діапазону та сумою квадратів відхилень для кожного класу. [Smith, 2005: с. 64]

Огляд літератури присвяченої пошуку оптимальної кількості груп, дозволив знайти ще один підхід. Суть його полягає в тому, щоб перевіряти утворені групи на статистичну відмінність за допомогою t-критерію Стьюдента та збільшувати кількість класів до тих пір, поки не будуть отримані (або не отримані) статистично відмінні класи [North, 2009: с. 36]

За схожою логікою в 2010 році була запропонована «модифікація» алгоритму, заснована на пошуку статистично відмінних груп. [Sun, 2010: с. 294] В цьому підході значення сортуються в порядку зростання. Для кожного значення розраховується довірчий інтервал. В випадку, якщо межі інтервалів в двох значень не пересікаються – в цій точці встановлюється розрив між класами. Мета полягає в тому, щоб сформувати класи так, щоб два значення, які мають найтісніший контакт розривів класів, були статистично різними і, отже, спостереження присвоювалися до класів, які статистично розрізняються за значеннями.

Як вже було зазначено вище, ручний перебір всіх комбінацій може зайняти багато часу навіть з допомогою обчислювальної техніки. Це накладає певні обмеження на використання алгоритму Дженкса на великих масивах даних, тим більше при пошуку оптимальних меж інтервалів великої кількості груп. По мірі розвитку можливостей обчислювальної техніки ця проблема скоріш за все буде усунена. Хороша новина полягає в тому, що оскільки соціологи зазвичай мають справу з вибірковими сукупностями, проблем з класифікацією об’єктів за метричною ознакою на невелику кількість груп виникати не повинно.

2.5 Head/tail breaks

Вище були наведені методи, які використовуються за умови нормальності розподілу ознаки, на основі якої здійснюється класифікації. Однак не всі ознаки об’єктів, що зустрічає людина є нормально розподіленими. Одним з перших цю закономірність помітив італійський соціолог так економіст Ф. Парето. Він зробив два висновки, перший з яких полягав у тому, що 80% італійської землі належало 20% населення, а 80% суспільних благ належить 20% населення. Згодом подібні спостереження були здійснені по відношенню до інших соціальних та природніх явищ. Так, наприклад, було встановлено, що в містах коротких вулиць значно більше ніж довгих ».[Jiang, 2007] Подібний вигляд має і вікова структура населення в багатьох країнах. Загальний висновок справедливий по відношенню до багатьох речей, що нас оточують, можна сформулювати наступним чином – малих речей є значно більше ніж великих.

При класифікації об’єктів для проведення аналізу в таких випадках існував би сенс приділяти достатню увагу низькочастотним явищам, так як вони мають тенденцію здійснювати більш суттєвий вплив у порівнянні з явищами високочастотними. Низькочастотні події інколи містять в собі набагато більше інформації, ніж події високочастотні, через неймовірність їх природи. Наприклад заголовки новин - низькочастотні події. Вони відбуваються рідко, а тому достойні висвітлення в засобах масової інформації: вони в якомусь сенсі передають більш важливу інформацію у порівняні з низькочастотними подіями.

Розподіли, які відображають подібний стан речей, називають розподілами з довгими хвостами. На відміну від нормального розподілу, який має форму купола з двома тонкими хвостами, що швидко досягають осі X, розподіли з довгими хвостами, до яких відносяться наприклад логнормальний розподіл, або розподіл Парето, теоретично має довгий хвіст схилений в право, що ніколи не досягає вищезгаданої Осі. Середнє значення такого розподілу розбиває його на

«голову», де знаходиться велика кількість великих значень і «хвіст» , що складається з малої кількості великих значень.

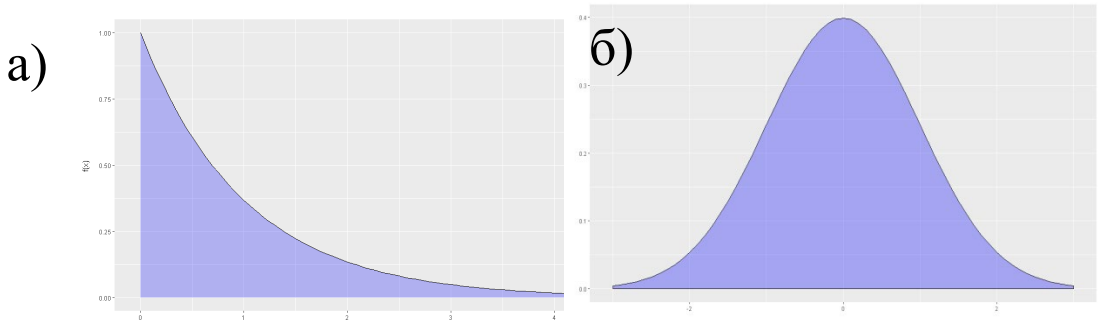


Рис 2.1.1 Ілюстрація розподілв: а) розподіл з довгим хвостом; б)нормальний розподіл.

Для класифікації об’єктів розподілених таким чином шведський професор геоінформатики Бін Чьянг в 2010 році запропонував схему, в якій ширина інтервалів і кількість класів визначається «самою природою даних».[Jiang, 2010: с. 5] В основі цієї схеми лежить описаний вище поділ розподілу на голову і хвіст. Очевидно, що після одного такого поділу через середнє значення, структура розподілу значень в середині хвоста буде приблизно такою самою, як в початковому варіанті. Це означає, що хвіст також може бути поділений на дві частини навколо середнього арифметичного і так далі.

Отже схема представляє собою рекурсивну функцію, яка розподіляє діапазон значень метричної ознаки на «голову» і хвіст» навколо середнього арифметичного після чого «хвіст» оголошується головою і операція повторюється доти, поки «хвіст» має вигляд розподілу з довгим хвостом.

Перевагою такого методу є його простота і виконання структурного групування ненормально розподілених значень. Очевидним недоліком, що витікає з природи схеми класифікації – непридатність для роботи з нормальним розподілом.

Висновки до Розділу II

Підсумовуючи написане в цьому розділі можна сказати, що кожен з представлених методів групування даних має свої переваги та недоліки. Більшість і з них розраховані на данні, що мають нормальний розподіл по зрозумілій причині, так як більшість соціальних та природніх явищ розподілені нормально. До ненормально розподілених даних можна застосовувати групування за квантилями. В випадку, якщо дані мають логнормальний розподіл, або схожий на нього, доцільно використовувати спеціально розроблений для цього алгоритм.

Метод класифікації Дженкса є одним із найскладніших представлених тут алгоритмів групування за однією метричною ознакою. Втім, класифікація Дженкса за умов правильного використання дає найкраще уявлення про структуру досліджуваних даних, адже його застосування дає максимально гомогенні групи.

На сьогоднішній день не існує жодного алгоритму класифікації даних за метричною ознакою, застосування якого гарантовано би створювало статистично відмінні групи незважаючи на те, що ідея його створення не є нова. Враховуючи сьогоднішні можливості обчислювальної техніки, вважаю, що такий алгоритм може бути реалізований шляхом модифікації алгоритму Дженкса, за допомогою введення додатковою умови перевірки на однорідність при переборі значень.

РОЗДІЛ ІІІ. Порівняльний аналіз алгоритмів одновимірної класифікації за рівнем доходів респондентів

Спроба співставлення об’єктів з ідеальними типами (що, власне і представлють з себе такі поняття як бідність, або середній клас і т.п) на основі всього лише однієї ознаки накладає значні обмеження в пізнанні сутності явищ. Незважаючи на те, що в більшості теоретичних підходах до осмислення явища бідності або нерівності загалом і фігурує дохід, сам по собі він не може повністю визначати здатність людини задовольняти свої базові потреби. Так, наприклад, цілком справедливо можна сказати, що якщо у індивіда є земельна ділянка, достатньо велика для того, щоб прокормлювати себе і свою сім’ю, він знаходиться в кращих матеріальних умовах ніж точно така сама людина з таким самим доходом без такої ділянки. Розробка критеріїв для створення стійкої типології є складною теоретичною та емпіричною проблемою, вирішення якої не є предметом даної роботи.

В цій роботі буде порівнюватися різний досвід виділення груп за метричною ознакою, а саме монетарних меж класів, що були виділені під час групування за доходом. Слово «клас» тут використовується в широкому сенсі для позначення впорядкованих гомогенних підмножин множини відносно автономних об’єктів, що виділені за певним принципом. Тобто, коли в тексті буде вживатися термін «середній клас», це означатиме лише те, що виділена за допомогою механічної процедури група знаходиться посередині, а не те, що люди, які були виділені під час класифікації, проведеної в рамках даної роботи, відносяться до «середнього класу», або навіть те, що останній існує в Україні.

Враховуючи, що в значній кількості робіт, де фігурує механічна класифікація за доходом, не наводиться ніякого обґрунтування причин, чому дані були згруповані саме таким чином, а не іншим, в цій роботі буде здійснена

спроба такого обґрунтування та продемонстровано, які з методів групування будуть краще справлятися з задачею виділення дохідних груп.

Об'єкт: процедура за метричною ознакою класифікації об'єктів в емпіричному соціологічному дослідженні

Предмет: процедура класифікації за метричною ознакою об'єктів при визначенні монетарних меж дохідних груп.

Мета: порівняння методів одновимірної класифікації на прикладі визначення монетарних меж дохідних груп.

Гіпотези:

1. Групи, отримані шляхом кластеризації даних, не будуть статистично відрізнятись між собою.
2. Класифікація, розроблена для побудови дохідної шкали російського суспільства буде найбільш наближеною до класифікації, отриманою шляхом застосування алгоритму Дженкса.
3. Результатом кластеризації за допомогою алгоритму k-means будуть менш однорідні групи ніж створені за допомого алгортма природних меж Дженкса.

3.1 Масив даних

Для здійснення порівняння був відібраний масив даних загальнонаціонального опитування «Українське суспільство – 2010». Опитування проведене методом роздаткового анкетування Інститутом соціології НАН України у квітні 2010 року. Вибірка дослідження складала 1800 респондентів. Вона пропорційно репрезентує доросле (віком понад 18 років) населення 24 областей України, АР Крим та м. Севастополя. Статистична похибка вибірки дорівнює 2,3%. Основною вимогою до масиву була наявність в ньому інформації про рівень доходу на одну людину та репрезентативність всього дорослого населення України.

Для вивчення дохідної стратифікації дослідник має визначитися, що саме він розуміє під доходом. Так, в якості одиниці спостереження можна розглядати як індивіда, так і господарство. Дохід домогосподарства з одного боку відображає потенційні можливості реалізації його членів, з іншого – відображає об'єктивні обмеження, що виникають внаслідок необхідності врахування спільних інтересів сім'ї. З цієї причини в якості ознаки, за якою була здійсненна класифікація, в масиві даних була відібрана зміна v218 «L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашої сім'ї».

В **таблиці 3.1** приведена описова статистика відібраної змінної. Як можна помітити, не всі респонденти (8,9% респондентів не відповіли на запитання) дали відповідь на питання про середньодушовий дохід домогосподарства. Це означає, що поза нашою увагою залишаються відповіді не тільки відповіді важко досяжних груп з понад низьким та понад високим доходом, а також значна частка тих, хто з якихось причин не надає інформації про своє матеріальне становище.

Високе значення коефіцієнту варіації, свідчить що більшість значень відповідей респондентів згруповані навколо середнього значення розподілу. Середнє значення та медіана, значно відрізняється один від одного, що може свідчити про наявність викидів в розподілі. Останнє твердження було підтверджено в результаті побудови гістограми частот розподілу представленого на **рисунку 3.1**.

Таблиця 3.1

Описова статистика змінної v218

Статистики		
L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашої сім'ї»		
N	Валидные	1655
	Пропущенные	147
Среднее		1084,638
Медиана		810
Коэф. вариации		85%
Минимум		0
Максимум		17000
Процентили	25	600
	50	810
	75	1300

Для подальшої коректності вибору критеріїв порівняння методів класифікації, була здійснена перевірка нормальності розподілу відібраної змінної за допомогою критерію Колмогорова-Смірнова. Значимість критерію не перевищила значення 0.05, що свідчить про ненормальність розподілу даних. (див. дод. Г)

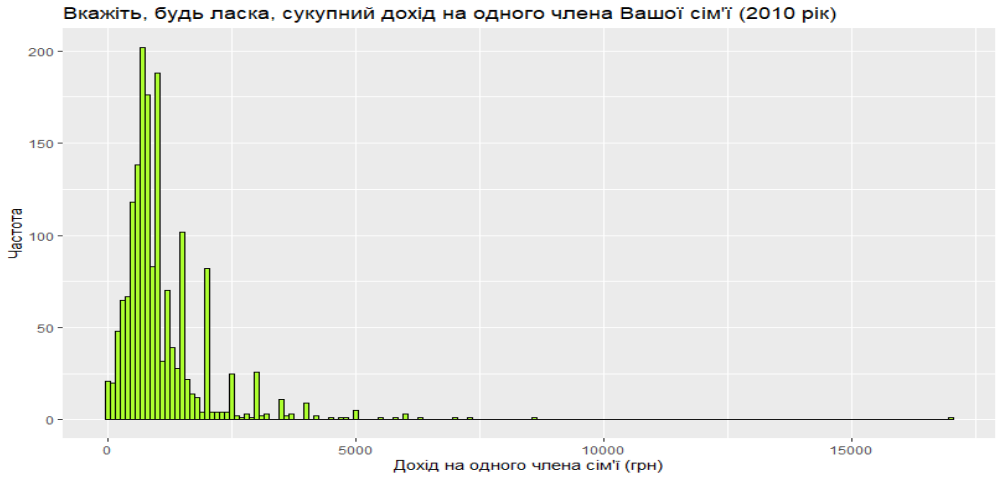


Рис 3.1.1 Гістограма частот доходу на одного члена сім'ї. «Українське суспільство 2010»

3.2 Критерії оцінки

У **розділі 1.2** в результаті огляду літератури за темою були визначені два основних критерії якості будь-якої класифікації - це однорідність груп та чіткі межі, що дозволяють з впевненістю відносити об’єкти до кожної з них. Придатної для будь-якої, а отже придатних і для кожної, здійсненої в рамках даної емпіричної частини роботи, класифікації.

Оскільки в цьому дослідженні порівнюється робота методів класифікації за однією метричною ознакою доходу на масиві даних, зібраних в рамках вибіркового дослідження, видається не позбавленою сенсу перевірка груп на статистичну відмінність між собою. Враховуючи ненормальний характер розподілу даних та чутливість t-критерію Стьюдента до викидів, для здійснення подібної перевірки в дослідженні був використаний U-критерій Манна — Уітні.

В випадку попарних порівнянь комбінації навіть трьох груп, значно зростає вірогідність зробити помилку другого роду, прийняти в якості правильної нульові гіпотезу про рівність групових середніх. Це означає, що для того аби перевірити відмінність всіх груп між собою, варто зменшити рівень значущості настільки, щоб не допустити цієї помилки. З цією метою при порівнянні груп в дослідженні використовувалась поправка Бонферроні, основна ідея якої полягає в тому, що якщо вірогідність припуститися помилки другого роду зростає пропорційно кількості проведених парних порівнянь, то рівень значущості, який цікавить дослідника варто поділити на кількість таких порівнянь. Враховуючи, що кількість груп для порівняння не є великою, з поміж множини доступних поправок на множинне порівняння, вважаю використання вищезгаданої поправки прийнятним.

Отже, в якості критерію, для відповіді на питання «чи відрізняються групи між собою?» був вибраний U-критерій Манна — Уітні. В якості показника однорідності груп ми будемо використовувати сумарну дисперсію в середині

виділених класів – чим менша сумарна внутрішньогрупова дисперсія, тим кращою є класифікація з точки зору критерію однорідності утворюваних класів.

3.3 Одиниці аналізу

Вибір кількості груп для класифікації був зумовлений традицією виділення 5 груп при побудові моделей дохідної стратифікації, що дало можливість порівнювати не тільки стандартні методи класифікації даних, але й специфічні для цього напрямку класифікації, теоретично встановленні межі дохідних груп.

В ході дослідження класифікація об'єктів здійснювалася наступним чином:

- Рівні інтервали
- Квінтелі
- Стандартне відхилення
- K-means
- Природні межі Дженкса
- Класифікація Світового банку
- Класифікація, що наведена в статті В.Бобкова[6]

Розрахунок меж більшості груп, виділених за ознакою доходу, здійснювався за допомогою бібліотеки «classInt», що реалізована в R - мові програмування для статистичної обробки даних. В бібліотеці наведено два варіанта реалізації алгоритму Дженкса, в цій роботі межі груп вираховувались за допомогою алгоритму «fisher», відповідно до рекомендацій, наданих в документації бібліотеки[7]. Для розрахунку меж дохідних груп за класифікацією наведеній в праці В.Бобкова, був уточнений прожитковий мінімум станом на

[6] Бобков В.Н., Колмаков И.Б. Выявление социальной структуры и неравенства распределения денежных доходов населения Российской Федерации // Экономика региона. 2017. Т. 13. № 4.
[7] Посилання на документацію бібліотеки: https://cran.r-project.org/web/packages/classInt/classInt.pdf

момент проведення дослідження. Так само для класифікації за межами, встановленими Світовим банком, був уточнений курс долара станом на цей же період – квітень 2010 року[8].

3.4 Результати

Рівні інтервали

Таблиця 3.2

Описова статистика груп виділених методом рівних інтервалів

Рівні інтервали		Статистика
0-3400	Частота	1 610,0
	Среднее	980,9
	Медиана	800,0
	Коэф. Вариации	61,0
	Дисперсия	354 218,5
3401-6800	Частота	41,0
	Среднее	4 288,4
	Медиана	4 000,0
	Коэф. Вариации	19,9
	Дисперсия	728 940,5
6801-10200	Частота	3,0
	Среднее	7 650,0
	Медиана	7 350,0
	Коэф. Вариации	11,0
	Дисперсия	707 500,0
13601-17000	Частота	1,0
	Среднее	NA
	Медиана	NA
	Коэф. Вариации	NA
	Дисперсия	NA

Метод рівних інтервалів передбачає розбиття діапазону значень метричної ознаки на групи, розмір меж яких є однаковим. Кількість утворених груп в результаті використання цього виявилась меншою ніж кількість заданих результатів в результаті чого було утворено 4 групи. Існування «пустої» групи стало можливим через наявність значення, яке значно відрізняється від

[8] Інформація про курс долара та прожитковий мінімум станом на квітень 2010 року бралась з архіву сайта Мінфін: https://index.minfin.com.ua/exchange/archive/nbu/2010-04-01/

середнього по розподілу. В даному випадку метод виявися мало придатним з практичної точки зору, оскільки майже всі об’єкти потрапили до першої групи. Всі виділені групи статистично значимо відрізняються одне від одного, окрім випадку порівняння групи з найбільш дохідною групою. Останнє пов’язане з малою наповненістю цієї групи.

Квінтелі

Таблиця 3.3

Описова статистика груп виділених методом квінтелей

Квінтелі		Статистика
0-550	Частота	339,0
	Среднее	353,2
	Медиана	400,0
	Дисперсия	22 865,4
	Коэф. Вариации	42,8
551-750	Частота	340,0
	Среднее	670,7
	Медиана	700,0
	Дисперсия	2 787,0
	Коэф. Вариации	7,9
751-1000	Частота	442,0
	Среднее	903,8
	Медиана	900,0
	Дисперсия	7 856,3
	Коэф. Вариации	9,8
1001-1500	Частота	272,0
	Среднее	1 330,3
	Медиана	1 300,0
	Дисперсия	23 160,4
	Коэф. Вариации	11,4
1501-17000	Частота	262,0
	Среднее	2 618,3
	Медиана	2 000,0
	Дисперсия	1 977 606,5
	Коэф. Вариации	53,7

Використання методу групування метричної ознаки за допомогою квінтельного підходу виділяє 5 груп, межі яких встановлюються таким чином,

щоб наповненість кожної з груп була однаковою. Для подолання проблеми, належності об’єктів з однаковими числовими характеристиками до різних груп, мною була здійснена незначна модифікація методу, яка полягала в закриті лівої межі інтервалу при перекодуванні змінної в категоріальну ознаку (клас). Всі 5 виділених груп статистично значимо відрізнялися одне від одної.

Стандартне відхилення

Таблиця 3.4

Описова статистика груп виділених стандартним відхиленням

Стандартне відхилення		Статистика
0-1084	Частота	1 127,0
	Среднее	668,6
	Медиана	700,0
	Дисперсия	63 180,0
	Коэф. Вариации	37,6
1085-5719	Частота	519,0
	Среднее	1 871,9
	Медиана	1 500,0
	Дисперсия	632 112,8
	Коэф. Вариации	42,5
5720-10354	Частота	8,0
	Среднее	6 631,3
	Медиана	6 150,0
	Дисперсия	932 098,2
	Коэф. Вариации	14,6
14990-17000	Частота	1,0
	Среднее	NA
	Медиана	NA
	Коэф. Вариации	NA
	Дисперсия	NA

Межі груп при використанні цього методу встановлюються на рівні одного стандартного відхилення від середнього значення розподілу. Внаслідок ненормальності розподілу відповідей респондентів, ліва межа першого інтервалу та права межа останнього знаходилась значно далі від мінімального та максимального значення діапазону розподілу в зв’язку з чим перші були

урізані до розміру останніх. Як і у випадку з рівними інтервалами, одна з груп виявилася «пустою», отже в результаті використання цього методу було утворено всього 4 наповнені групи. Всі групи, які виявилися достатньо наповненими для перевірки гіпотези про їх відмінність, статистично значимо відрізнялися одне від одного.

k-means

Таблиця 3.5

Описова статистика груп виділених методом k-means

k-means		Статистика
0-475	Частота	225,0
	Среднее	276,3
	Медиана	300,0
	Дисперсия	16 701,2
	Коэф. Вариации	46,8
476-877	Частота	635,0
	Среднее	679,8
	Медиана	700,0
	Дисперсия	11 784,2
	Коэф. Вариации	16,0
878-1460	Частота	435,0
	Среднее	1 076,5
	Медиана	1 000,0
	Дисперсия	22 280,2
	Коэф. Вариации	13,9
1461-2950	Частота	284,0
	Среднее	1 842,9
	Медиана	1 800,0
	Дисперсия	123 780,7
	Коэф. Вариации	19,1
2951-17000	Частота	76,0
	Среднее	4 073,4
	Медиана	3 500,0
	Дисперсия	3 627 755,6
	Коэф. Вариации	46,8

Метод кластеризації k-means полягає в мінімізації дисперсії в середині кластерів (груп) та її максимізації між ними. В ході роботи алгоритму випадковим чином відбираються центри кластерів, після чого до групи

приписується найближчий до центра кластера об’єкт. Алгоритм виконується до тих пір, поки групи не стануть стійкими, тобто в результаті повторення операцій алгоритму не будуть виділятися одні й ті ж групи. Наслідком застосування цього методу стали 5 наповнених груп, кожна з яких статистично значимо відрізняється одне від одної.

Природні межі Дженкса

Таблиця 3.6

Описова статистика груп виділених методом Дженкса

Природні межі Дженкса		Статистика
0-1128	Частота	1 157,0
	Среднее	679,8
	Медиана	700,0
	Дисперсия	66 259,9
	Коэф. Вариации	37,9
1129-2341	Частота	386,0
	Среднее	1 574,2
	Медиана	1 500,0
	Дисперсия	91 265,5
	Коэф. Вариации	19,2
2342-4350	Частота	94,0
	Среднее	3 046,4
	Медиана	3 000,0
	Дисперсия	269 038,7
	Коэф. Вариации	17,0
4351-12800	Частота	17,0
	Среднее	5 738,2
	Медиана	5 500,0
	Дисперсия	1 198 602,9
	Коэф. Вариации	19,1
12801-17000	Частота	1,0
	Среднее	NA
	Медиана	NA
	Коэф. Вариации	NA
	Дисперсия	NA

Принцип роботи алгоритму класифікації Дженкса також заснований на мінімізації дисперсії всередині груп, однак на відмінну від попереднього

методу, розпочинає свою роботу не від випадково встановлених центрів груп, а впорядковуючи об’єкти в порядку зростання. Результатом використання алгоритму природних меж Дженкса стали 5 груп. Всі групи, які виявилися достатньо наповненими для перевірки гіпотези про їх відмінність, статистично значимо відрізнялися одне від одного. Виділені групи мали найменшу середню арифметичну між групову дисперсію.

Класифікація Світового Банку

Таблиця 3.7

Описова статистика груп виділених за класифікацією Світового Банку

Класифікація світового банку		Статистика
2.5$	Частота	342,0
	Среднее	355,1
	Медиана	400,0
	Дисперсия	23 072,1
	Коэф. Вариации	42,8
2.5-4$	Частота	591,0
	Среднее	741,4
	Медиана	740,0
	Дисперсия	8 929,1
	Коэф. Вариации	12,7
4-10$	Частота	610,0
	Среднее	1 368,1
	Медиана	1 300,0
	Дисперсия	131 472,1
	Коэф. Вариации	26,5
10-50$	Частота	111,0
	Среднее	3 458,6
	Медиана	3 000,0
	Дисперсия	1 350 163,6
	Коэф. Вариации	33,6
50$	Частота	1,0
	Среднее	NA
	Медиана	NA
	Коэф. Вариации	NA
	Дисперсия	NA

Межі груп в даному випадку були задані директивно відповідно до 5-ти групового варіанту класифікації, що використовує Світовий Банк. Застосування критерію Манна-Утні дало можливість підтвердити відмінність всіх груп, що були достатньо наповненими для його використання.

Класифікація, наведена в статті В.Бобкова

Таблиця 3.8

Описова статистика груп виділених за класифікацією, що наведена в статті В. Бобкова

bobkov_test5		Статистика
1 ПМ	Частота	834,0
	Среднее	565,6
	Медиана	600,0
	Дисперсия	43 457,1
	Стд. отклонение	36,9
1 - 3 ПМ	Частота	738,0
	Среднее	1 347,6
	Медиана	1 200,0
	Дисперсия	188 443,1
	Стд. отклонение	32,2
3 - 7 ПМ	Частота	75,0
	Среднее	3 527,7
	Медиана	3 200,0
	Дисперсия	528 354,3
	Стд. отклонение	20,6
7 - 11 ПМ	Частота	7,0
	Среднее	6 750,0
	Медиана	6 300,0
	Дисперсия	955 833,3
	Стд. отклонение	14,5
11 ПМ	Частота	1,0
	Среднее	NA
	Медиана	NA
	Коэф. Вариации	NA
	Дисперсия	NA

Як і в попередньому випадку, межі груп, виділених на основі метричної ознаки доходу було встановлено директивно на основі значення прожиткового

мінімуму. Всі виділені групи статистично значимо відрізняються одне від одного, окрім випадку порівняння групи з найбільш дохідною групою, в яку потрапила відповідь тільки одного респондента.

Висновки до Розділу ІІІ

В результаті дослідження було виявлено, що абсолютно всі, використанні в дослідженні методи класифікації об'єктів за ознакою доходу, утворюють статистично відмінні групи, навіть за умови використання поправки Бонферроні, що значно знижує шанси на прийняття нульової гіпотези. (див. дод. Б) Це означає, що у випадку масиву 2010 року, абсолютно всі використані нами методи поділу змінної доходу на 5 класів дали необхідну диференціацію для їх подальшого порівняння. Цей факт також спростував першу гіпотезу дослідження, що ґрунтувалася на побоюваннях щодо виділення занадто щільних груп.

Таблиця 3.9

Однорідність в середині кожного з класів та середнє значення дисперсії. Метод Дженкса та теоретична класифікація.

Дисперсія всередині груп				
№	Дженкс	k-means	Світовий Банк	Стратифікація Бобкова
1	66259,9	16701,2	23072,1	43457,1
2	91265,5	11784,2	1200,0	632112,8
3	269038,7	22280,2	528354,3	932098,2
4	1198602,9	123780,7	1350163,6	955833,3
5	0,0	3627755,6	0,0	0,0
Середнє	**325033,4**	760460,4	380558,0	512700,3

Також можемо констатувати, що алгоритм класифікації Дженкса значно краще справляється з здачею виділення однорідних груп ніж алгоритм кластеризації k-means. Це пояснюються в першу чергу тим, що перший передбачає відсортовування об'єктів перед здійсненням групування. Втім, з точки зору наповненості груп, що також може бути можливо з інструментальної точки зору, k-means видається більш оптимальним варіантом групування.

Серед представлених груп, найменшим, після Дженкса, середнім значенням внутрішньогрупових дисперсій володіють групи, виділені за допомогою класифікації Світового банку. З одного боку це спостереження може слугувати ще одним підтвердженням слушності думки В. Паніотто, щодо того, що на даному етапі економічного розвитку країни, (станом на 2009 рік) абсолютний підхід до виділення межі бідності має використовуватися в якості основного. [Paniotto, 2009]. Це справедливо, адже в випадку застосування абсолютного підходу, принаймні реалізованого в версії Світового Банку, виділяються не тільки об'єктивні можливості задовольнити базові потреби, але й однорідні з точки зору доходу групи.

З іншого боку, за умови перевірки адекватності моделі для російського суспільства, можна стверджувати, що структура дохідної диференціації останнього відрізняється від структури суспільства українського, а отже модель запропонована В.Бобковим з точки зору якості класифікації не може використовуватися в якості еталону.

На відміну від алгоритму Дженкса, всі інші, приведені вище, (**таблиця 3.1**) способи класифікації об'єктів за доходом краще виділяють однорідні групи на рівні нижчих класів. При дослідженні соціальної стратифікації населення загалом, часто виникає зміщення акценту в сторону найбідніших груп, що видається цілком логічним у випадку, коли державні органи соцзахисту намагаються зрозуміти. які групи населення потребують найбільшої допомоги, або в випадку соціологічних та економічних досліджень, об'єктом яких

виступає бідність. Однак у випадку побудови 5-групової моделі класифікації доходів домогосподарств України у цілому з метою найкращого відображення його структури вибір методу природних меж Дженкса є найбільш вдалим. В подальшому в виділених групах, умовно можна буде виділяти підгрупи бідного населення.

«Класичні» методи групування даних за однією метричною ознакою показали менш задовільний результат. Можна було б стверджувати, що значення внутрішньогрупових дисперсій при класифікації за допомогою рівних інтервалів, квінтелів та стандартного відхилення знаходяться приблизно на одному і тому самому рівні – рівні, в цілому кращому з точки зору класифікації за однією змінною ніж використання k-means чи моделі дохідної стратифікації російського суспільства, здійсненої В. Бобковим. Однак у випадку використання рівних інтервалів та стандартного відхилення утворилися «пусті» класи, що не дає нам права порівнювати результати отримані для 5 груп. (див. дод. В) Причина останнього полягає в чутливості алгоритмів до «викидів» в розподілі даних.

Метод групування через квінтелі має серйозні недоліки з точки зору адекватності представлення уявлень про бідність, так як частка «бідних» за такого підходу є принципово незмінною. Тим не менш, принаймні на цьому масиві даних, цей метод залишається цілком адекватним, а головне простим інструментом класифікації доходів, з точки зору виділення відносно однорідних та відмінних між собою груп.

Підсумовуючи можна стверджувати, що використання більшості з представлених методів групування з метою виділення меж однорідних дохідних класів є доцільним. В майбутньому ці методи можна буде застосовувати при побудові моделі дохідної стратифікації українського суспільства, шляхом перевірки відтворюваності меж дохідних груп на різних

масивах даних всеукраїнських опитувань та роботи над теоретичним обґрунтуванням, або спростування, доцільності виділення саме таких меж.

ВИСНОВКИ

Отже, в даній роботі піднімалося питання про методи класифікації, а саме про методи класифікації даних за однією метричною ознакою та їх використання у емпіричному соціологічному дослідженні**.** У першому розділі мною було розглянуто поняття класифікації та його співвідношення з іншими суміжними поняттями. Було встановлено, що єдиного підходу до визначення цих понять не існує. Часто вони використовуватися як синоніми.

Так, **класифікацію** в широкому значенні можна розглядати як процес або результат виділення груп об’єктів за певним принципом; 2) як процес або результат виділення груп об’єктів за певною (часто однією) ознакою, що не продукує знань про весь набір суттєвих ознак об’єктів, а тому є нейтральним до їх «сутності»; 3) як процес або результат виділення груп об’єктів за сукупністю множини деяких ознак.; 4) як елементарну процедуру впорядкування даних за категоріальною ознакою, що передує їхньому аналізу.

При вживанні поняття класифікації в другому, виділеному вище, значенні воно часто протиставляється поняттю типологізації. Останнє може розглядатися і як різновид класифікації, що здійснюється шляхом співвідношення об’єктів з деяким вибраним реальним або теоретичним типом. В третьому та четвертому значенні поняття класифікація вживається в статистичній літературі і протиставляється поняттю групування. По відношенню до систематизації, класифікація, в свою чергу, може розглядатися, як один із її методів, як одного із методів специфічної форми наукового дослідження, пізнавального процесу упорядкування деякої множини розрізнених об’єктів і знання про них.

В роботі були описані критерії класифікації об’єктів. Класи, або групи виділені в результаті класифікації повинні містити в собі об’єкти, що є **схожими між собою**. В той же час, ці групи повинні **відрізнятися** один від одного. Питання щодо того, на якій підставі вважати об’єкти подібними залишається відкритим. З точки зору наукового агностицизму, виділення однорідних груп на основі всіх ознак об’єкта є принципово неможливим, так як ми принципово не можемо знати про всі ознаки, якими володіє об’єкт. З іншої сторони, навіть якщо вважати, що можемо, то виникає питання, чи мають всі ці ознаки однакову вагу. В зв’язку з цим оцінка якості виконання критеріїв класифікації є суб’єктивною і може варіюватися в залежності від поставлених перед дослідником завдань.

В випадку класифікації на основі **метричних показників** до певною міри стає можливим математично обґрунтування однорідності груп за допомогою алгоритмів кластеризації. Суть кластеризації полягає в тому, щоб поділити певну множину об'єктів на кілька відносно однорідних груп, які називають кластерами. Кластери будуються таким чином, аби відстань між об'єктами всередині одного кластера була більшою, ніж об'єктів з різних кластерів. Мірою такої відстані між об’єктами в випадку класифікації за однією метричною ознакою може виступати дисперсія. В випадку, коли класифікуються дані, отримані під час проведення вибіркового дослідження важливо, аби виділені групи статистично значимо відрізнялись одне від одної.

Також в роботі були описані підходи до класифікації за неперервною метричною ознакою доходу у емпіричному соціологічному дослідженні. **Абсолютний підхід** до встановлення меж класів передбачає, що останні встановлюються на основі деякого експертного консенсусу та є незмінними. **Відносний підхід** визначає ці межі, ґрунтуючись на природі розподілу даних. Найбільш популярними способами виділення меж таких груп є виділення через певну частку медіани або середнього значення. Часто межами груп, при

побудові шкали дохідної стратифікації, можуть виступати межі рівнонаповнених інтервалів.

В другому розділі була розкрита сутність часто застосовуваних **методів** класифікації об’єктів за метричною ознакою, а саме: метод рівних інтервалів, метод рівнонаповнених інтервалів, виділення меж груп на основі стандартного відхилення, алгоритм Дженкса-Фішера. Також був описаний алгоритм, призначений для класифікац ії об’єктів, розподілених логнормально.

Основною перевагою методу рівнонаповнених та рівних інтервалів є їх інтуїтивна зрозумілість виконання алгоритмів, що зменшує ризики невірної інтерпретації отриманих в ході дослідження результатів. Окрім цього, в ході емпіричного дослідження, що метод відносно добре виконує задачу виділення однорідних груп при класифікації доходів. Головним недоліком першого є можливість потрапляння об’єктів з однаковими властивостями до різних класів. Головним недоліком другого є сильна залежність від нормальності розподілу, за умови не виконання якої можуть з’явитися пусті, ненаповнені класи, приклад чого був продемонстрований **в розділі 3**.

Від нормальності розподілу. особливо від відсутності викидів, залежить і метод класифікації за стандартним відхиленням. В роботі також був здійснений огляд методів шкалювання, в основі яких лежить стандартне відхилення , а саме: t-бали¸ стени, станайни. Таке шкалювання здійснюється за допомого лінійного перетворення початкових даних з їх округленням до найближчого цілого числа. Основною перевагою такого способу є зручність сприйняття. Окрім того, використання такого методу, принаймні на першому етапі групування знімає питання кількості груп.

Метод класифікації Дженкса, за умов правильного використання, дає найкраще уявлення про структуру досліджуваних даних, адже його застосування дає максимально гомогенні групи. Однак цей метод не дає відповіді на питання оптимальної кількості груп, що виділяються. В випадку

коли дані розподілені логнормально, доцільно використовувати також алгоритм head/tail.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Foster M. The OECD approach to measure and monitor income poverty across countries / M. Foster, H. Levy. – 2013. – С. 2.
2. Birg´e L. How many bins should be put in a regular histogram / L. Birg´e, Y. Rozenholc. // *ESAIM: Probability and Statistics*. – 2006. – С. 24–26.
3. Dallinger U. The endangered middle class? A comparative analysis of the role played by income redistribution / Ursula Dallinger. // *Journal of European Social Policy*. – 2013. – №23.
4. Evans I. The Selection of Class Intervals / Ian S. Evans. // *Transactions of the Institute of British Geographers, New Series*. – 1977. – С. 98–102.
5. Hyndman R. The problem with Sturges' rule for constructing histograms [Електронний ресурс] / Rob J Hyndman. – 1995. – Режим доступу до ресурсу: https://robjhyndman.com/papers/sturges.pdf.
6. Jiang B. Head/tail Breaks: A New Classification Scheme for Data with a Heavy-tailed Distribution / Bin Jiang. – 2010.
7. Jiang B. A topological pattern of urban street networks: universality and peculiarity, *Physica A: Statistical Mechanics and its Applications,* Bin Jiang. – 2007.
8. Statistical Mechanics and its Applications, 384, 647-655.
9. Kaufman Leonard, Peter J. Rousseeuw. 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, New Jersey : *John Wiley & Sons, Inc*., 2005.
10. Lajos Stegena. Statistical Determination of Class Intervals for Maps / Lajos Stegena, Ferenc Csillag. // *The Cartographic Journal*. – 1987. – №24. – С. 142–146.
11. Medeiros M. The rich and the poor: the construction of an affluence line from the poverty line // Social Indicators Research. 2006. Vol. 78. No. 1.
12. Millennium development goals. National report. – 2010. – С. 34.

13. North M. A Method for Implementing a Statistically Significant Number of Data Classes in the Jenks Algorithm / Matthew A. North. – 2009. – С. 35–38.
14. Paniotto V. What Poverty Criteria Are Best for Ukraine? / V. Paniotto, N. Kharchenko. // *Problems of Economic Transition*. – 2008. – С. pp. 5–12.
15. Scott D. Normal Reference Rule / D. Scott, R. Kosar. // *Wiley StatsRef: Statistics Reference Online*. – 2015.
16. Smith R. Comparing traditional methods for selecting class intervals on choropleth maps / Richard M. Smith. // *The Professional Geographer*. – 2005. – №38. – С. 62 – 67.
17. Sun M. Incorporating Data Quality Information in Mapping American Community Survey Data / M. Sun, D. Wong. *// Cartography and Geographic Information Science*. – 2010. – №37. – С. 285–299.
18. Vakis R. Left Behind: Poverty in Latin America and the Caribbean (overview) / R. Vakis, J. Rigolini, L. Lucchetti. – *Washington: International Bank for Reconstruction and Development* / The World Bank, 2015. – 44 с.
19. Wang H. Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming / H. Wang, M. Song. // *The R Journal*. – 2011. – С. 29–33.
20. Андреенков В. Г. Типология и классификация в социологических исследованиях / В. Г. Андреенков, Ю. Н. Толстова. – Москва: Наука, 1982. – 296 с.
21. Аникин В. Экономическая стратификация: об определении границ доходных групп / В. Аникин, Ю. Лежнина. // *Социологичесское обозрение.* – 2018. – №1. – С. 237 –273.
22. Бабич Н. С. Концепция типологических операций Пауля Лазарсфельда: сущность и роль в современной западной социологии / Н. С. Бабич. // *Актуальные вопросы современной науки*. – 2012.

23. Бардасов С. А. Гистограммы. Критерии оптимальности / С. А. Бардасов. – Тюмень: Тюменский государственный университет, 2014. – 96 с.
24. Богомолова Т. Ю. Бедность в Современной Росии: измерение и анализ / Т. Ю. Богомолова. // *Социология*. – 2006. – №4. – С. 90–94.
25. Васнев С.А. Статистика: Учебное пособие / Васнев С.А.. – Москва: МГУП, 2001. – 170 с.
26. Гржибовский А. М. Анализ трех и более независимых групп количественных данных / А. М. Гржибовский. // *Экология человека*. – 2008. – №3. – С. 50–58.
27. Ефремова О. А. Методы интервальных оценок и классификация числовых характеристик данных агентства по печати и средствам массовой информации Республики Башкортостан / О. А. Ефремова, С. А. Мишустин // *Геоинформационные технологии в проектировании и создании корпоративных информационных систем* / О. А. Ефремова, С. А. Мишустин. – Уфа, 2014. – С. 138–142.
28. Ильичев Л. Ф. Ильичев Философский энциклопедический словарь / Л. Ф.Ильичев, П. Н. Федосеев, С. М. Ковалев, В. Г. Панов. – Москва: Советская энциклопедия, 1983. – 840 с.
29. Каган М. С. Проблемы методологии гуманитарного познания. Избранные труды для вузов / М. С. Каган. – Москва: Юрайт, 2018. – 321 с. – (Антология мысли).
30. Настасяк, І. Ю. Співвідношення поняття «типологізація» з суміжними поняттями / Настасяк, І. Ю. // *Вісник Львівського університету*. – 2015. – №61. – С. 37–44.
31. Носс И. Н. Введение в практику психологического эксперимента / И. Н. Носс. – Москва: ПЭР СЭ, 2006. – 304 с.
32. Райзин Д. Вэн (ред.) Классификация и кластер / Дж. Вэн Райзин. – Москва: Мир, 1980. – 394 с.

33. Субботин А. Л. Классификация / А. Л. Субботин. – Москва, 2001. – 94 с.
34. Тихонова Н. В. Стратификация по доходу в России: специфика модели и вектор изменений. Статья 1 / Н. В. Тихонова. // *Общественные науки и современность.* – 2017. – №2. – С. 23–35.
35. Тихонова Н. Е. Доходная стратификации в России: кросс-страновой и динамический анализ // *Социологический журнал*, 2017. Том. 23. № 4. С. 31-50.
36. Туктарова Ф. К. Общая теория статистики Конспект лекций / Ф. К. Туктарова. – Пенза: ПГУ, 2010. – 94 с.
37. Фролов И. Т. Философский словарь / И. Т. Фролов. – Москва: Республика, 2001. – 720 с.
38. Хахулина Л. Распределение доходов: бедные и богатые в постсоциалистических обществах (некоторые результаты сравнительного анализа) / Л. Хахулина, М. Тучек. // *Мониторинг общественного мнения*. – 1995. – №1. – С. 18–22.
39. Шинкарук В.І. Філософський енциклопедичний словник / Шинкарук В.І. – Київ: НАН України, Ін-т філософії ім. Г. С. Сковороди, 2002. – 742 с. – (Абрис).
40. Ядов В. А. Социологическое исследование: методология, программа, методы / В. А. Ядов. – Москва: Наука, 1972. – 266 с.

ДОДАТКИ

Додаток А

Значення дисперсії всередині виділених груп

Рівні інтервали	
№	Дисперсія
1	354218,5
2	728940,5
3	707500,0
4	NA
5	0,0
Середнє	447664,8

Квінтелі	
№	Дисперсія
1	22865,4
2	2787,0
3	7856,3
4	23160,4
5	1977606,5
Середнє	406855,1

k-means	
№	Дисперсія
1	16701,2
2	11784,2
3	22280,2
4	123780,7
5	3627755,6
Середнє	760460,4

Дженкс	
№	Дисперсія
1	66259,9
2	91265,5
3	269038,7
4	1198602,9
5	0,0
Середнє	325033,4

Стандартне відхилення	
№	Дисперсія
1	63180,0
2	632112,8
3	932098,2
4	NA
5	0,0
Середнє	406847,8

Світовий Банк	
№	Дисперсія
1	23072,1
2	1200,0
3	528354,3
4	1350163,6
5	0,0
Середнє	380558,0

Стратифікація Бобкова	
№	Дисперсія
1	43457,1
2	632112,8
3	932098,2
4	955833,3
5	0,0
Середнє	512700,3

Додаток Б

Порівняння груп виділених за допомогою рівних інтервалів

-W= v218 BY equal_test5(1 2)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	1296855,000
Z	-10,965
Асимпт. знч. (двухсторонняя)	,000

/M-W= v218 BY equal_test5(2 3)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	861,000
Z	-2,901
Асимпт. знч. (двухсторонняя)	,004
Точная знч. [2*(1-сторонняя Знач.)]	,000

/M-W= v218 BY equal_test5(1 3)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	1296855,000
Z	-3,001
Асимпт. знч. (двухсторонняя)	,003

/M-W= v218 BY equal_test5(2 5)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	861,000
Z	-1,717
Асимпт. знч. (двухсторонняя)	,086
Точная знч. [2*(1-сторонняя Знач.)]	,048

/M-W= v218 BY equal_test5(1 5)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	1296855,000
Z	-1,734
Асимпт. знч. (двухсторонняя)	,083

/M-W= v218 BY equal_test5(3 5)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	6,000
Z	-1,342
Асимпт. знч. (двухсторонняя)	,180
Точная знч. [2*(1-сторонняя Знач.)]	,500

Порівняння груп виділених за допомогою квінтелів

/M-W= v218 BY quantile_test5(1 2)

/M-W= v218 BY quantile_test5(2 4)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	57630,000
Z	-22,670
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY quantile_test5(1 3)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	57970,000
Z	-21,433
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY quantile_test5(2 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	57630,000
Z	-24,224
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY quantile_test5(1 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	57970,000
Z	-21,186
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY quantile_test5(3 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	57630,000
Z	-21,368
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY quantile_test5(1 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	97903,000
Z	-22,768
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY quantile_test5(3 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	57630,000
Z	-21,120
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY quantile_test5(2 3)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	97903,000
Z	-22,496
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY quantile_test5(4 5)
```

Статистики критерия

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	57970,000
Z	-24,271
Асимпт. знч. (двухсторонняя)	,000

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	37128,000
Z	-20,106
Асимпт. знч. (двухсторонняя)	,000

Порівняння груп виділених за допомогою k-means

```
/M-W= v218 BY kmeans_test5(1 2)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	25425,000
Z	-22,417
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY kmeans_test5(2 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	201930,000
Z	-24,363
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY kmeans_test5(1 3)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	25425,000
Z	-21,316
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY kmeans_test5(2 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	201930,000
Z	-14,375
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY kmeans_test5(1 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	25425,000
Z	-19,510
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY kmeans_test5(3 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	94830,000
Z	-22,926
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY kmeans_test5(1 5)
```

```
/M-W= v218 BY kmeans_test5(3 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	25425,000
Z	-13,086
Асимпт. знч. (двухсторонняя)	,000

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	94830,000
Z	-14,256
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY kmeans_test5(2 3)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	201930,000
Z	-27,953
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY kmeans_test5(4 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	40470,000
Z	-13,609
Асимпт. знч. (двухсторонняя)	,000

Порівняння груп виділених за допомогою алгоритму Дженкса

```
/M-W= v218 BY fisher_test5(1 2)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	669903,000
Z	-29,512
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY fisher_test5(2 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	74691,000
Z	-7,072
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY fisher_test5(1 3)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	669903,000
Z	-16,193
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY fisher_test5(2 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	74691,000
Z	-1,753
Асимпт. знч. (двухсторонняя)	,080
Точная знч. [2*(1-сторонняя Знач.)]	,005

/M-W= v218 BY fisher_test5(1 4)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	669903,000
Z	-7,113
Асимпт. знч. (двухсторонняя)	,000

/M-W= v218 BY fisher_test5(3 4)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	4465,000
Z	-6,629
Асимпт. знч. (двухсторонняя)	,000

/M-W= v218 BY fisher_test5(1 5)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	669903,000
Z	-1,737
Асимпт. знч. (двухсторонняя)	,082

/M-W= v218 BY fisher_test5(3 5)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	4465,000
Z	-1,750
Асимпт. знч. (двухсторонняя)	,080
Точная знч. [2*(1-сторонняя Знач.)]	,021

/M-W= v218 BY fisher_test5(2 3)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	74691,000
Z	-15,162
Асимпт. знч. (двухсторонняя)	,000

/M-W= v218 BY fisher_test5(4 5)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	153,000
Z	-1,659
Асимпт. знч. (двухсторонняя)	,097
Точная знч. [2*(1-сторонняя Знач.)]	,111

Порівняння груп виділених за допомогою стандартного відхилення

/M-W= v218 BY sd_test5(1 2)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	635628,000
Z	-32,691
Асимпт. знч. (двухсторонняя)	,000

/M-W= v218 BY sd_test5(25)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	134940,000
Z	-1,740
Асимпт. знч. (двухсторонняя)	,082

/M-W= v218 BY sd_test5(1 3)

/M-W= v218 BY sd_test5(35)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	635628,000
Z	-4,900
Асимпт. знч. (двухсторонняя)	,000

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	36,000
Z	-1,576
Асимпт. знч. (двухсторонняя)	,115
Точная знч. [2*(1-сторонняя Знач.)]	,222

```
/M-W= v218 BY sd_test5(1 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	635628,000
Z	-1,738
Асимпт. знч. (двухсторонняя)	,082

```
/M-W= v218 BY sd_test5(23)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	134940,000
Z	-4,886
Асимпт. знч. (двухсторонняя)	,000

Порівняння груп виділених за класифікацією Світового Банку

```
/M-W= v218 BY WB5(1 2)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	58653,000
Z	-25,577
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY WB5(2 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	174936,000
Z	-16,856
Асимпт. знч. (двухсторонняя)	,000

/M-W= v218 BY WB5(1 3)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	58653,000
Z	-25,755
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY WB5(2 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	174936,000
Z	-1,751
Асимпт. знч. (двухсторонняя)	,080

/M-W= v218 BY WB5(1 4)

```
/M-W= v218 BY WB5(3 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	58653,000
Z	-15,937
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY WB5(1 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	186355,000
Z	-16,943
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY WB5(3 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	58653,000
Z	-1,751
Асимпт. знч. (двухсторонняя)	,080
Точная знч. [2*(1-сторонняя Знач.)]	,006

```
/M-W= v218 BY WB5(2 3)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	186355,000
Z	-1,758
Асимпт. знч. (двухсторонняя)	,079

```
/M-W= v218 BY WB5(4 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	174936,000
Z	-30,105
Асимпт. знч. (двухсторонняя)	,000

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	6216,000
Z	-1,739
Асимпт. знч. (двухсторонняя)	,082
Точная знч. [2*(1-сторонняя Знач.)]	,018

Порівняння груп виділених за класифікацією Бобкова

/M-W= v218 BY bobkov_test5(1 2)

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	348195,000
Z	-34,322
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY bobkov_test5(1 3)
```

```
/M-W= v218 BY bobkov_test5(2 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	272691,000
Z	-4,601
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY bobkov_test5(2 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	348195,000
Z	-14,417
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY bobkov_test5(1 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	348195,000
Z	-4,584
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY bobkov_test5(1 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	348195,000
Z	-1,739
Асимпт. знч. (двухсторонняя)	,082

```
/M-W= v218 BY bobkov_test5(2 3)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	272691,000
Z	-14,386
Асимпт. знч. (двухсторонняя)	,000

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	272691,000
Z	-1,746
Асимпт. знч. (двухсторонняя)	,081

```
/M-W= v218 BY bobkov_test5(3 4)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	2850,000
Z	-4,437
Асимпт. знч. (двухсторонняя)	,000

```
/M-W= v218 BY bobkov_test5(3 5)
```

Статистики критерия

	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	2850,000
Z	-1,749
Асимпт. знч. (двухсторонняя)	,080
Точная знч. [2*(1-сторонняя Знач.)]	,026

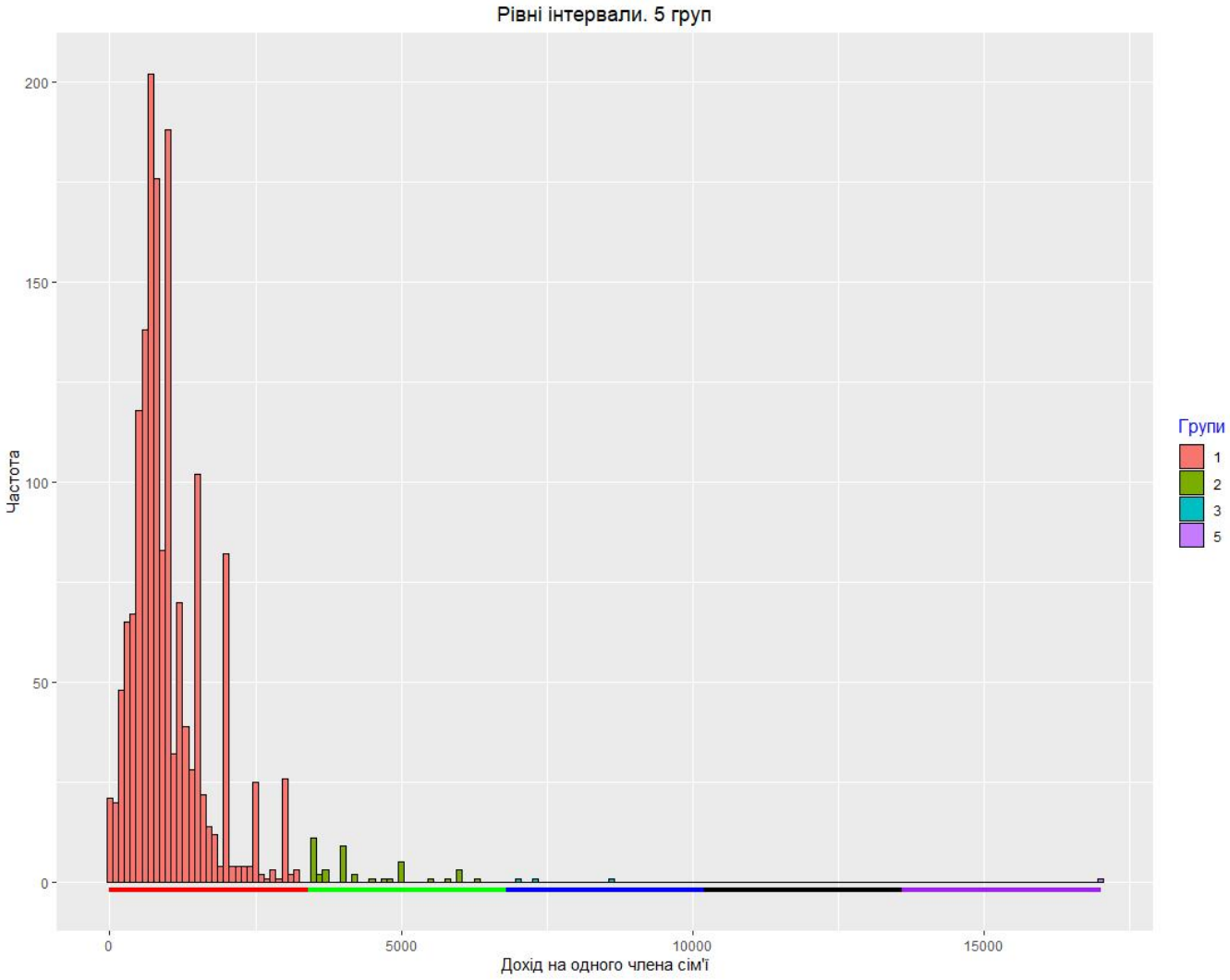
```
/M-W= v218 BY bobkov_test5(4 5)
```

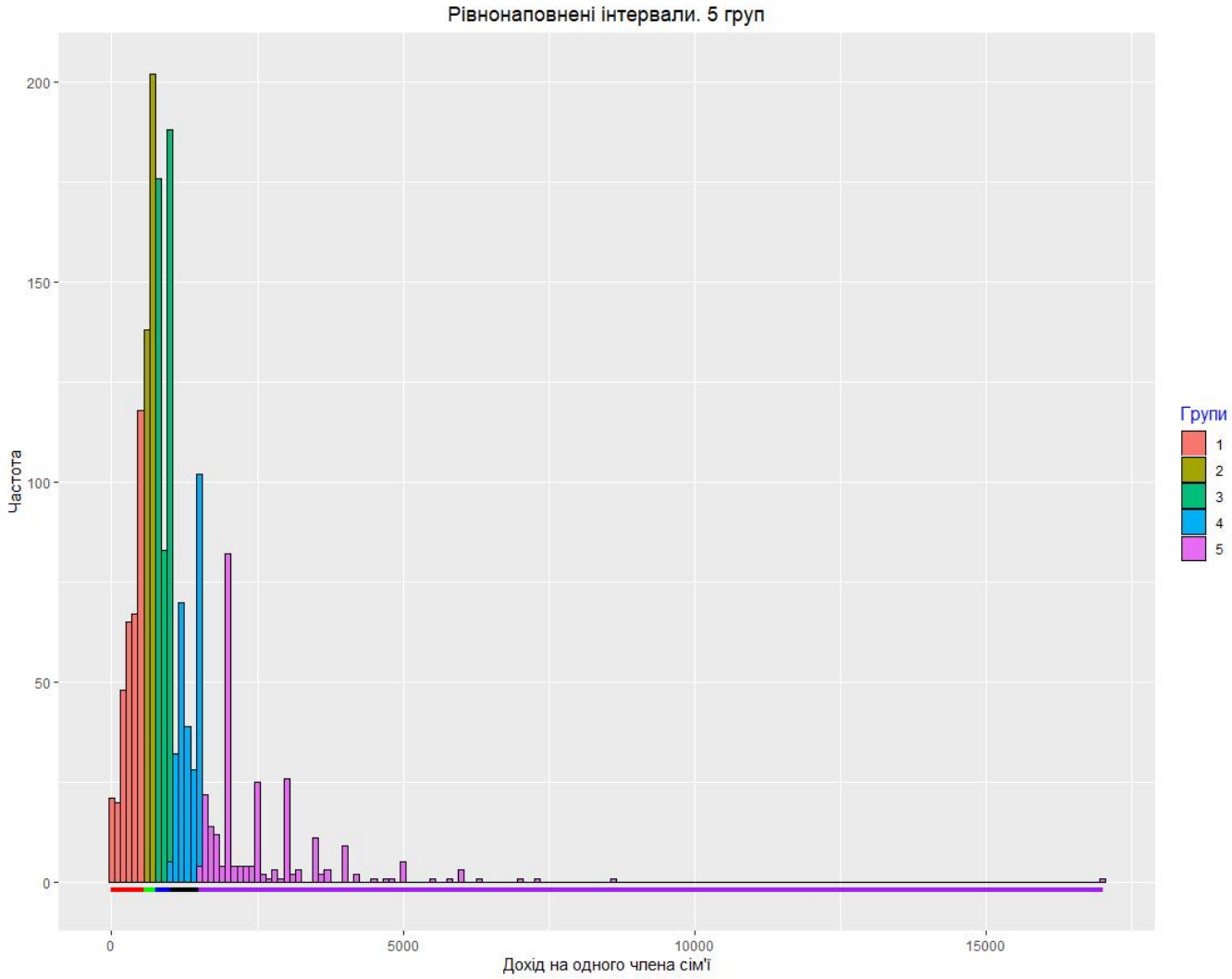
Статистики критерия

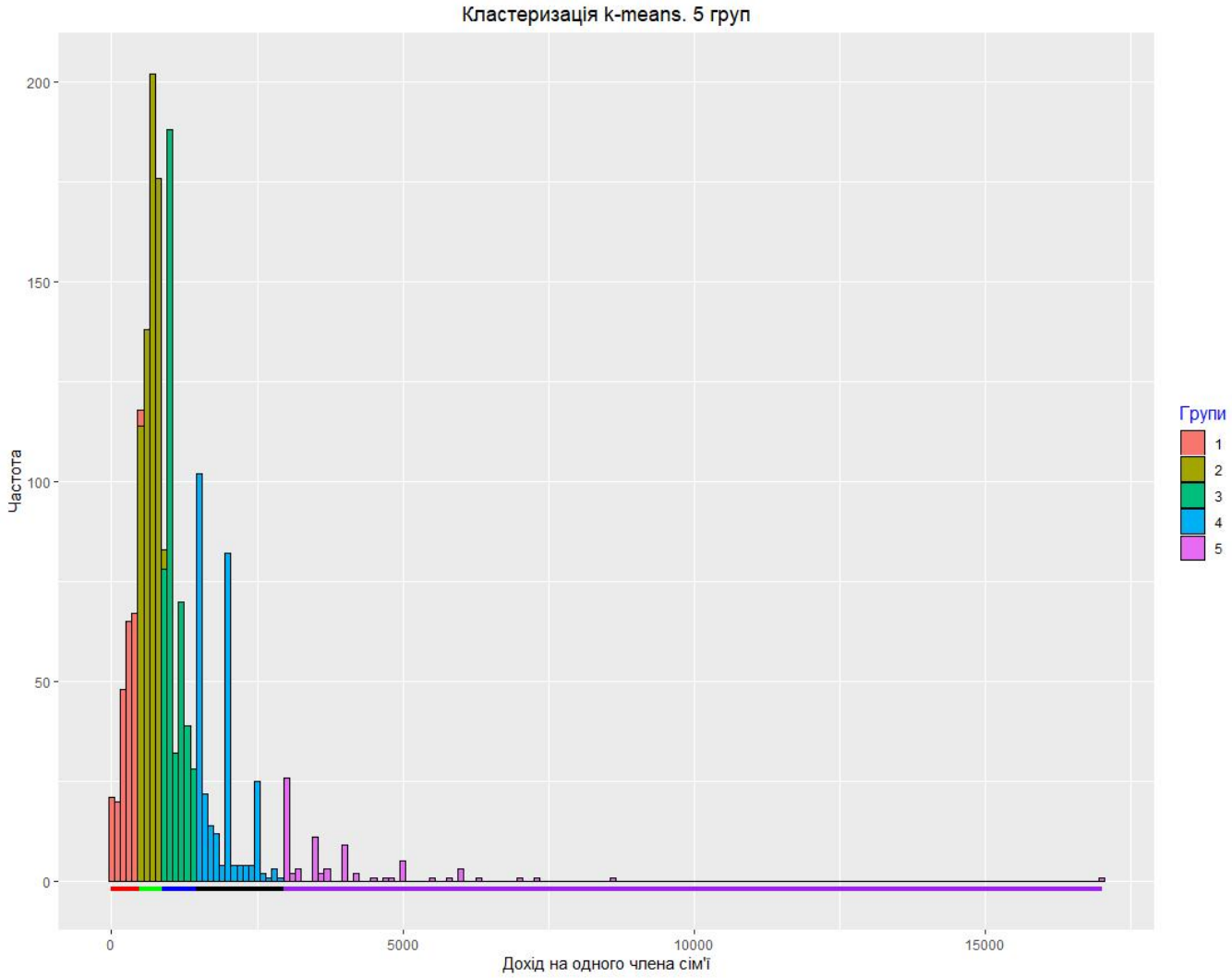
	L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
Статистика U Манна-Уитни	0,000
Статистика W Уилкоксона	28,000
Z	-1,565
Асимпт. знч. (двухсторонняя)	,118
Точная знч. [2*(1-сторонняя Знач.)]	,250

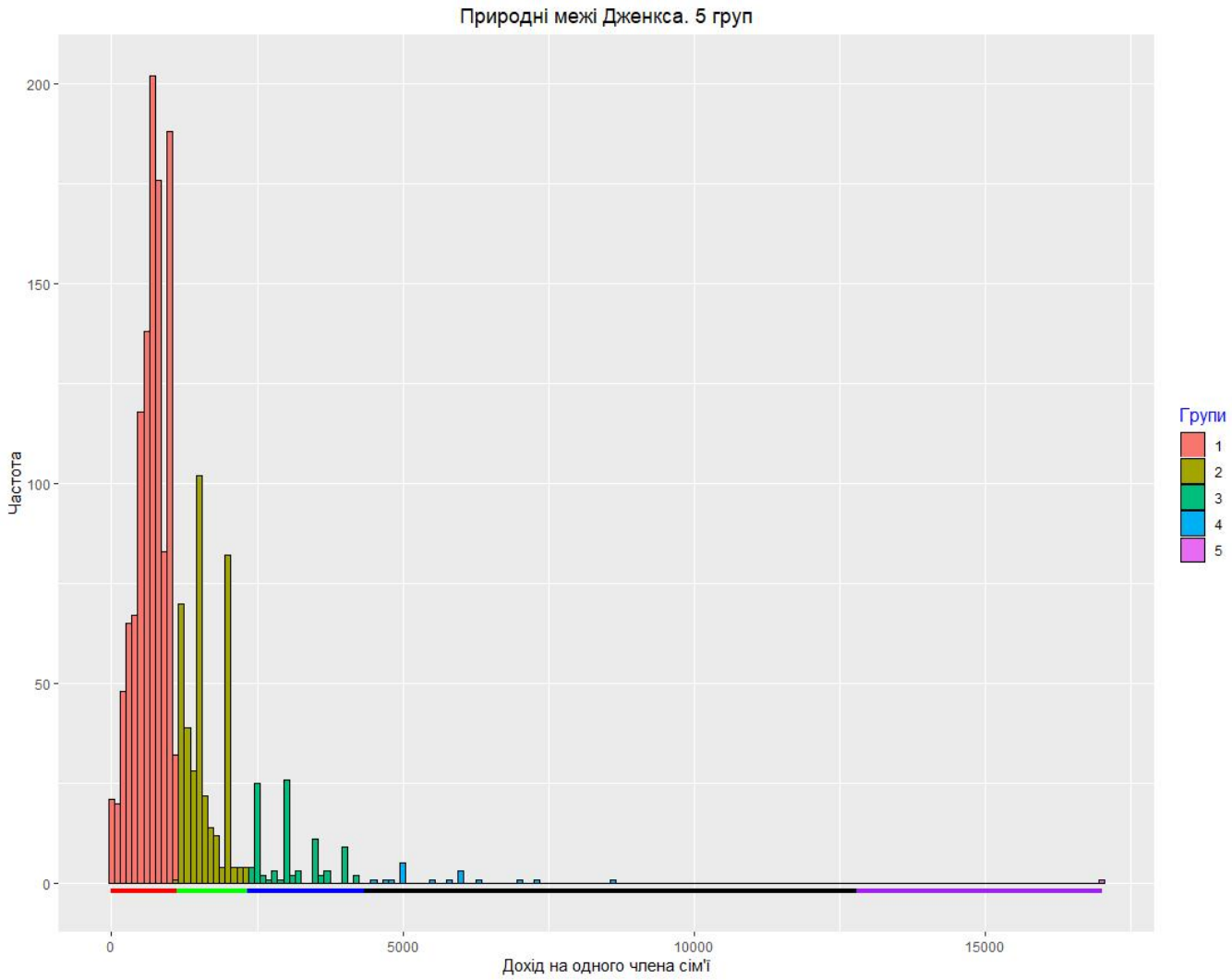
Додаток В

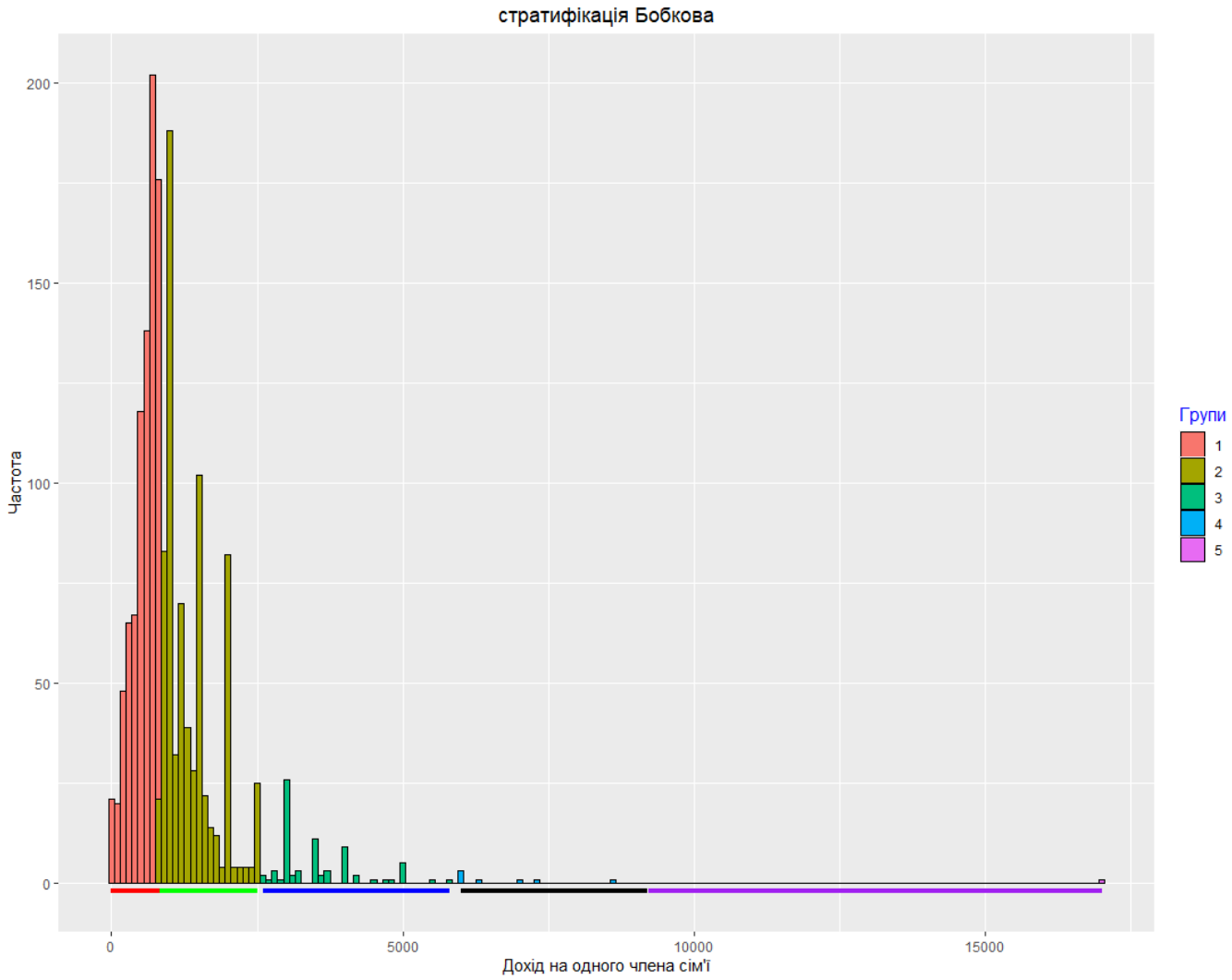
Діаграми частот відповідей респондентів на запитання про середньодушовий дохід домогосподарства

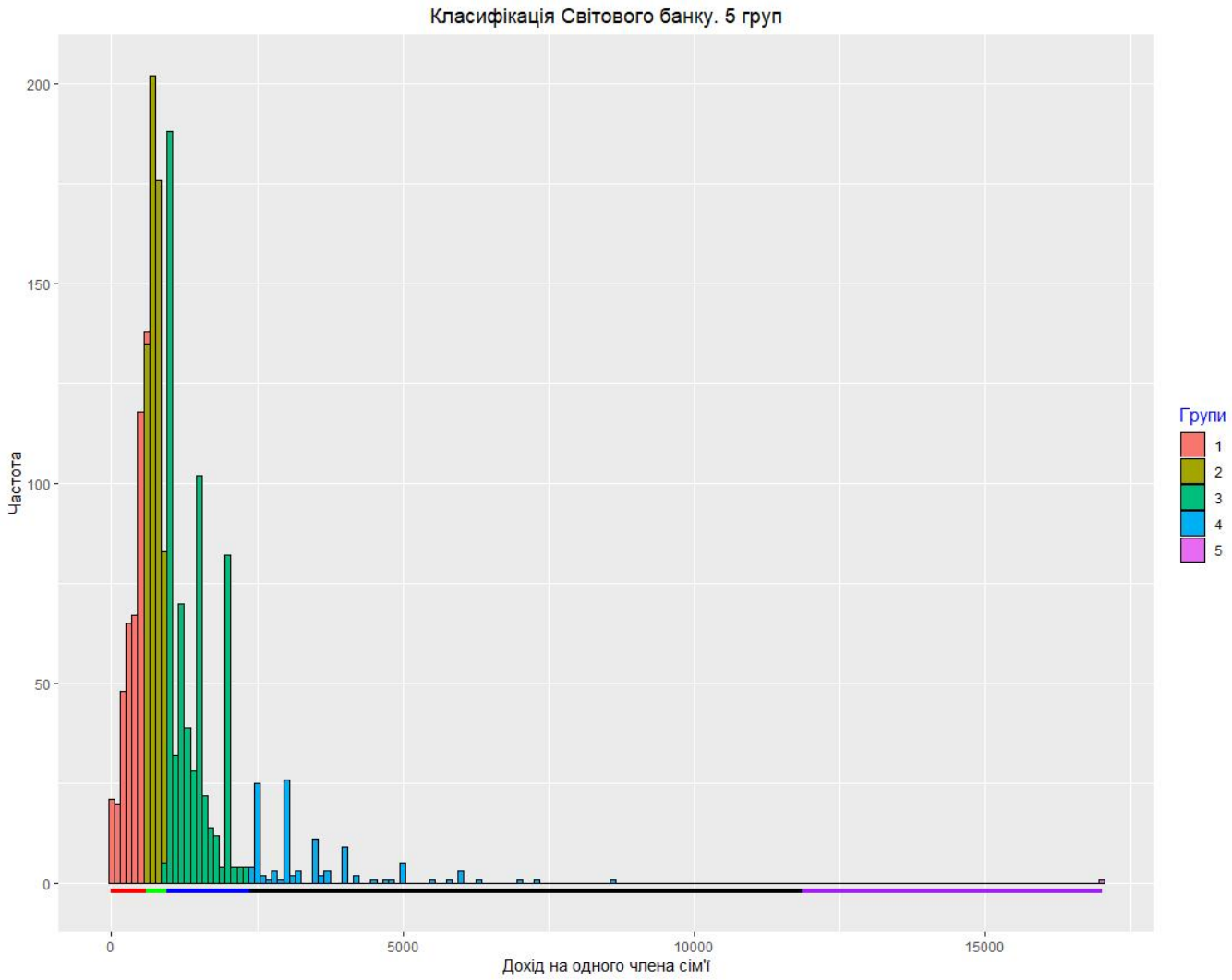



Рівнонаповнені інтервали. 5 груп
Частота
200
150
100
50
0
0
5000
10000
15000
Дохід на одного члена сім'ї
Групи
1
2
3
4
5


Кластеризація k-means. 5 груп
Частота
200
150
100
50
0
0
5000
10000
15000
Дохід на одного члена сім'ї
Групи
1
2
3
4
5


Природні межі Дженкса. 5 груп
Частота
200
150
100
50
0
0
5000
10000
15000
Дохід на одного члена сім'ї
Групи
1
2
3
4
5


стратифікація Бобкова
Частота
200
150
100
50
0
0
5000
10000
15000
Дохід на одного члена сім'ї
Групи
1
2
3
4
5


Класифікація Світового банку. 5 груп
Частота
200
150
100
50
0
0
5000
10000
15000
Дохід на одного члена сім'ї
Групи
1
2
3
4
5

Додаток Г

Перевірка нормальності розподілу змінної v218

Одновыборочный критерий Колмогорова-Смирнова

		L5. Вкажіть, будь ласка, сукупний дохід на одного члена Вашо
N		1655
Нормальные параметры[a,b]	Среднее	1084.6381
	Стд. отклонение	926.97967
Разности экстремумов	Модуль	.214
	Положительные	.214
	Отрицательные	-.143
Статистика Z Колмогорова-Смирнова		8.694
Асимпт. знч. (двухсторонняя)		.000

a. Сравнение с нормальным распределением.

b. Оценивается по данным.