УДК [311.214:004.4]:303.446 **Doi: 10.31767/su.4(91)2020.04.02**
JEL Classification: C13, C18, C88

Н. В. Ковтун,
доктор економічних наук, професор,
завідувач кафедри,
Київський національний університет імені Тараса Шевченка,
E-mail: kovtun_natali@ukr.net
ResearcherID: M-6596-2017,
ORCID: https://orcid.org/0000-0002-2935-8597;
А.-Н. Я. Фаталієва,
Біостатистик I, ЕДЕТЕК,
E-mail: nastya1nana@gmail.com
ORCID: https://orcid.org/0000-0001-5541-8509

Програмна реалізація відновлення пропущених даних: порівняльний аналіз

Проведено порівняльний аналіз можливостей застосування різних програмних продуктів для вирішення проблеми відновлення даних на прикладі вибірки, для якої симульовані різні варіанти пропусків даних. Дослідження дало змогу виявити слабкі та сильні сторони розглянутих програмних продуктів, а також визначити ефективність застосування того чи іншого методу за різних обсягів пропущеної інформації.

Найпростішим інструментом відновлення пропусків визначено пакет прикладних програм Statistica, який пропонує користувачу лише прості методи обробки пропущених даних. Ця програма допоможе впоратися з пропущеними даними при незначному обсязі пропусків (до 10%). SPSS пропонує ширший спектр методів відновлення даних порівняно зі Statistica, водночас має зрозуміліший інтерфейс для користувача проти мов програмування R чи SAS. В останніх зазначених програмних середовищах можна використовувати різні методи відновлення даних від найпростіших до найскладніших, таких як, наприклад, множинна імпутація. Отже, R та SAS є найпотужнішими програмами з відновлення даних, проте і складнішими для користувачів, оскільки потребують знання мови програмування.

Встановлено, що жодне з розглянутих програмно-аналітичних середовищ не має вбудованих процедур обробки категоріальних даних. У програмних середовищах R та SAS є певні підходи, які можна реалізувати за аналогією для упорядкованих категорій, проте це не покриває всі потреби аналізу досліджень, реалізованих у вигляді опитувань і результати яких здебільшого представлені як відповіді на запитання. Методи, які застосовуються для відновлення кількісних даних, не можуть бути поширені на категоріальні, навіть якщо для кодування відповідей використані цифри.

Дослідження безперечно довело той факт, що до відновлення даних у різних програмних середовищах, так само, як і до вибору можливих способів застосування тих чи інших способів імпутації у різних середовищах, слід підходити дуже обережно. У кожному конкретному випадку проблема імпутації має вирішуватися на основі ретельного аналізу існуючої бази даних з урахуванням не тільки особливостей самих даних і обсягу пропусків, а й специфіки конкретного дослідження.

Робота з пропущеними даними охоплює широкий спектр проблем, серед яких вивчення природи пропусків, вибір методології обробки й відновлення даних залежно від їхньої природи та від типу, а також використання різних програмних засобів відновлення даних.

У подальшому планується оцінити ефективність відновлювальної здатності методів, реалізованих у різних пакетах прикладних програм, а також розробити методологічні засади відновлення пропусків для категоріальних даних та реалізувати їх на практиці.

Ключові слова: *пропущені дані, типи пропусків, засоби імпутації даних, SPSS Statistics, Statistica, програмне середовище R, SAS.*

Вступ. Проблема пропусків даних притаманна будь-якому статистичному спостереженню, особливо якщо воно є тривалим у часі й охоплює велику кількість об'єктів спостереження. Пропуски знижують статистичну потужність критеріїв, а також можуть бути причиною систематичних помилок, що, своєю чергою, знижує якість результатів статистичного аналізу.

Існує багато причин виникнення таких пропусків [1–4]: це відсутність можливості збирання даних за окремими одиницями сукупності, вибуття суб'єктів (респондентів), соціальна (психоло-

гічна) чутливість до запитань, складності ідентифікації відповіді, а також несистематичні помилки при збиранні або введенні даних або втрата частини інформації. Процент пропусків може бути різним. Як правило, визначають такі діапазони пропусків даних [2]:

1) менше 5% пропусків вважається несуттєвим і не впливає на результати дослідження;

2) 5–10% пропусків можна відновити, використовуючи достатньо прості методи імпутації;

3) 10–20% пропусків вимагають ретельного підходу до вибору методу імпутації залежно від типу й особливостей виникнення пропусків, а також мети статистичного аналізу;

4) втрата 20% і більше даних ставлять під сумнів результати дослідження.

Зрозуміло, що визначення меж коливання доволі умовне. Але безперечним є те, що чим більша частка даних відсутня, тим менша надійність висновків і тим складніше обґрунтувати достовірність результатів статистичного аналізу.

Один із способів вирішення цієї проблеми полягає в тому, щоб повністю виключити записи, які мають пропуски; це призводить до зменшення вибірки, а отже, впливає на достовірність результатів статистичного аналізу і потужність статистичного висновку. Проте некоректна зроблена імпутація також може вплинути як на статистичну похибку, так і на довірчі інтервали та врешті решт призвести до зниження рівня довіри до результатів статистичного аналізу, а іноді – навіть до викривлення результатів статистичного спостереження. Наявність пропусків даних, так само, як і аналіз тільки повних спостережень (після виключення спостережень з пропусками), може призвести до отримання зміщених результатів і як наслідок – до викривлення висновків за результатами дослідження і прийняття неправильних рішень.

Аналіз останніх досліджень і публікацій. В Україні дослідженню проблем, пов'язаних з відновленням даних у контексті програмної реалізації, приділялась і приділяється незначна увага, тоді як за кордоном упродовж останніх 30 років ця проблематика вивчається активно і з наростаючим інтересом, про що свідчать публікації останніх років [5–8]. Статистичним методам імпутації даних присвячено чимало робіт таких дослідників, як О. Злоба, А. Куталієв, О. Міщук, В. Саріогло, Р. Ткаченко, І. Яцків, Р. Дж. А. Літтл, М. О'Келлі, Б. Ратіч, Д. Б. Рубін тощо. О. Злоба та І. Яцків досліджують ефективність методів відновлення пропущених даних у цілому. В. Саріогло досліджує процедури імпутації при обробці даних вибіркових обстежень домогосподарств, О. Міщук та Р. Ткаченко розглядають методи заповнення пропущених даних екологічного моніторингу. В роботах М. О'Келлі, Б. Ратіч розглядається метод множинної імпутації пропущених даних у клінічних дослідженнях та особливості його застосування в програмі SAS. Отже, проблема відновлення пропущених даних потребує всебічного дослідження та опрацювання, а саме, розробки методології та програмної реалізації зважаючи на особливості застосування цього інструменту в тій чи іншій сфері.

Більшість статистичних методів розраховані на те, що дослідник працює з повним набором даних (матриці, вектори, пошукові таблиці). Задля усунення негативних наслідків, які можуть мати місце у випадку наявності пропусків даних, використовують різноманітні підходи та методи, спрямовані на відновлення пропущених значень, які надають можливість замінити пропуски відповідними умовними (розрахунковими) значеннями.

Проте дуже часто має місце ситуація, коли дослідники взагалі не приділяють уваги наявним пропускам і обробляють неповні дані. Досить низький рівень культури обробки даних з пропусками знаходить своє відображення й у сучасному стані статистичного програмного забезпечення. Відновлення даних вручну не є ефективним засобом оброблення пропусків. У роботі [9] ми вже довели вплив частки пропусків на результати статистичних розрахунків і якість статистичного висновку, виявили слабкі та сильні сторони кожного методу імпутації даних (для повністю випадкових (MCAR) і випадкових (MAR) типів пропусків), а також визначили ефективність застосування того чи іншого методу при різних частках пропущеної інформації. Результати дослідження також показали, що до процесу імпутації слід підходити дуже обережно. Проблема імпутації має вирішуватись у кожному конкретному випадку на основі ретельного аналізу існуючої бази даних з урахуванням не тільки особливостей самих даних і обсягу пропусків, а й специфіки конкретного дослідження.

Мета представленої роботи – дослідити різні підходи до реалізації автоматизованих методів відновлення пропущених даних і здійснити їх порівняльний аналіз у різних програмно-аналітичних середовищах: Statistica, SAS, R та SPSS.

Методологія. Насамперед розглянемо типи пропусків даних і методи їх відновлення. Відповідно до класифікації Літтла і Рубіна [4], пропуски можна розділити на три групи:

1. Повністю випадкові пропуски (missing completely at random (MCAR)).

2. Випадкові пропуски (missing at random (MAR)).

3. Неігноровані пропуски (non-ignorable missingness).

Для обробки перших двох видів пропусків застосовують вісім основних методів:

1) аналіз повних спостережень (видалення по рядках, або порядкове видалення, listwise deletion);

2) попарне видалення (pairwise deletion);

3) підстановка середнього по вибірці (mean substitution);

4) метод хот-дек (hot deck);

5) регресійний аналіз (regression);

6) метод максимальної правдоподібності (maximum likelihood estimation);

7) підстановка за допомогою факторного аналізу (factor analysis substitution);

8) модель множинного відновлення даних (multiple imputations method).

Статистичні дані, як відомо, можуть бути виражені у кількісній та некількісній формах, тобто представлені специфічними категоріями. Переважно вважається, що імпутація пропущених значень можлива тільки для кількісних даних, проте це не так: для відновлення категоріальних даних існують певні процедури, які базуються на інших методологічних засадах і потребують окремої уваги (аналіз таких процедур виходить за межі дослідження). У цій роботі ми робимо акцент саме на методах, спрямованих на відновлення пропущених значень, що є кількісними ознаками, в основному – неперервними.

Результати. Розглянемо процедури з відновлення даних, які пропонуються у зазначених вище пакетах прикладних програм (ППП). Так, у пакеті IBM® SPSS Statistics допускається відновлення двох видів пропущених значень:

1. Пропуски, які визначаються системою (System-defined missing values): якщо в матриці даних є незаповнені чисельні комірки, система SPSS самостійно ідентифікує їх як пропущені значення. Цей факт відображається в матриці даних за допомогою коми (“,”).

2. Пропущені значення, що задаються користувачем (User-defined missing values): якщо в певних випадках у змінних відсутні значення, користувач може за допомогою кнопки Missing оголосити ці значення як пропущені [10].

Процедура “Аналіз пропущених значень” у SPSS виконує три основні функції:

1) описує структуру пропущених даних: Де розташовані пропущені значення? Наскільки широку область вони охоплюють? Чи є тенденція до пропуску значень у декількох спостереженнях у пар змінних? Чи приймають дані крайні значення? Чи мають пропуски випадковий характер?

2) оцінює середні, середньоквадратичні відхилення, коваріації і кореляції для різних методів обробки пропущених значень: порядкове та попарне видалення, регресія, оцінка максимальної правдоподібності. Попарний метод виводить також частоти повних пар спостережень.

3) імпутує на місце пропущених значень оціночні значення, використовуючи метод регресії, оцінку максимальної правдоподібності або більш точний метод множинної імпутації [11].

У R середовищі пропущені дані позначаються символом NA (not available, або немає в наявності). Неприпустимі значення (наприклад, ділення на 0) позначаються як NaN (not a number – не є числом). На відміну від SAS, у R використовуються однакові позначення для пропущених значень у текстових і числових даних. Крім того, в пакеті R є кілька функцій, призначених для виявлення пропущених значень. Функція is.na () дозволяє перевірити дані на наявність пропущених значень. Можна виділити таку класифікацію методів обробки пропущених значень:

1) видалення пропущених значень (по рядках та попарно);

2) оцінка максимальної правдоподібності;

3) заміщення пропущених значень (одиночна та множина імпутація) [12].

У програмі Statistica порожнім коміркам надається певний спеціальний код – код пропущених даних (Missing Data Code), значення якого за замовчуванням дорівнює 99999. Спосіб використання пропущених даних можна підібрати індивідуально для кожної процедури аналізу. Там, де це можливо, користувачеві надано вибір способу обробки пропущених даних: видалення їх з обчислень по рядках або попарно, заміна на середні значення, а також їх перетворення або інтерполяція (наприклад, у модулі “Часові ряди”). Щоб дізнатися про конкретні способи використання пропущених даних у певних процедурах, потрібно натиснути кнопку довідки або клавішу f1 у відповідному діалоговому вікні завдання аналізу [13].

У програмно-аналітичному середовищі SAS пропущені числові дані позначаються як “.”, а текстові залишаються порожніми. Як правило, процедури SAS, обробляють та аналізують дані, оминаючи відсутні значення. Тобто за замовчуванням використовуються методи видалення за рядками або попарного видалення залежно від процедури [14]. У цьому ППП можна імпутувати пропуски за допомогою всіх існуючих методів: підстановка середнього по вибірці, метод хот-дек, регресійний аналіз, оцінка максимальної правдоподібності, підстановка за допомогою факторного аналізу та метод множинної імпутації. Процедури proc mi, proc mianalyze – потужний інструмент для обробки та відновлення даних у SAS [15].

Задля ілюстрації роботи розглянутих вище процедур було використано дані про пацієнтів, хворих на анорексію, вага яких вимірювалася до та після лікування. Упродовж періоду дослідження частина пацієнтів отримала лікування (46 осіб), а інша належала до контрольної групи і лікування не отримувала (26 осіб) [16].

Описова статистика даних для дослідження наведена в табл. 1 (власні розрахунки авторів за даними [16]), що є стандартним способом узагальнення даних у клінічних випробуваннях.

Таблиця 1

Описова статистика для даних

(кг)

Статистики	Отримали лікування		Контрольна група	
	Вага до лікування	Вага після лікування	Вага до лікування	Вага після лікування
Середня вага, $\bar{x}$	37,60	39,68	36,99	36,79
Стандартне відхилення, σ	2,205	3,913	2,589	2,152
Мінімум	31,8	32,3	32,0	33,1
Медіана	37,56	39,08	36,58	36,60
Максимум	43,0	47,0	41,6	40,6

Для проведення порівняльного аналізу програмної реалізації відновлення пропусків за допомогою програмного коду випадковим чином було симульовано набори даних, де для змінної "вага після лікування" пропущено 5%, 10%, 25% і 50% значень спостережень.

Розглядаючи повні дані як генеральну сукупність, а симульовані з пропусками як вибірку, було пораховано середнє значення на пропущених даних та довірчі межі (табл. 2, власні розрахунки авторів у середовищі SAS).

Таблиця 2

Середня вага після лікування

(кг)

Частка пропусків, %	Отримали лікування		Контрольна група	
	Середня вага, $\bar{x}$	Довірчі межі	Середня вага, $\bar{x}$	Довірчі межі
0	39,68	x	36,79	x
5	39,64	(38,45; 40,83)	36,65	(35,75; 37,55)
10	39,79	(38,57; 41,01)	36,80	(35,87; 37,73)
25	39,51	(38,18; 40,84)	37,19	(36,25; 38,12)
50	41,27	(39,57; 42,98)	37,11	(35,91; 38,31)

Розраховані довірчі межі включають середнє значення, пораховане на повних даних, – 39,68 кг та 36,79 кг для пацієнтів, які отримали та не отримали лікування відповідно. Проте чим більша частка пропущених даних, тим більше відхиляється середня, обчислена на повних даних, від середньої, отриманої на неповних даних (особливо при 50% пропусків). Зважаючи на те, що 5% пропусків майже не призвели до суттєвих відхилень середньої ваги в обох групах, процедуру відновлення даних було застосовано для 10% пропусків і вище.

При застосуванні методу порядкового видалення було отримано такі самі результати, оскільки цей метод включає тільки ті спостереження, що не містять пропущених даних. Оскільки пропуски має лише одна змінна, то в цьому випадку застосовувати метод попарного видалення не має сенсу. Якщо необхідно провести аналіз, ураховуючи інші наявні зміні, то тоді доречно застосувати метод попарного видалення.

Метод безумовної імпутації (підстановка середнього по вибірці) можна використати в програмі Statistica, SAS або R. Проте з'ясувалося, що в програмі Statistica цей метод реалізований не в усіх модулях. Так, при використанні модуля множинної регресії зазначений метод можна застосувати, тоді як у модулі описової статистики наявні лише два найпростіші методи. При незначному обсязі даних дослідник, працюючи у ППП Statistica, може лише вручну підставити середнє замість пропусків.

Метод хот-дек (заповнення пропусків з упередженим підбором) можна реалізувати лише в програмі SAS. Результати двох вищезазначених методів наведено в табл. 3 (власні розрахунки авторів у середовищі SAS).

Для групи пацієнтів, які отримали лікування, довірчі межі включають середнє значення, обраховане на повних даних. У другій контрольній групі бачимо, що метод безумовної імпутації погано спрацював для 25% та 50% пропусків, а метод хот-дек – для 50% пропущених даних. Можемо зробити висновок, що вищезазначені методи краще застосовувати при обсягах пропущених даних до 25%.

Більш потужні методи, такі як метод регресійного аналізу, множинної імпутації, оцінка максимальної правдоподібності, було реалізовано в SAS та SPSS.

Таблиця 3

Середня вага після лікування після проведення процедури відновлення пропусків у ППП Statistica та SAS

(кг)

Частка пропусків, %	Отримали лікування		Контрольна група	
	Середня вага, $\bar{x}$	Довірчі межі	Середня вага, $\bar{x}$	Довірчі межі
Метод безумовної імпутації				
0	39,68	x	36,79	x
10	39,72	(38,58; 40,86)	37,11	(36,28; 37,94)
25	39,37	(38,31; 40,43)	37,74	(37,07; 38,41)
50	40,33	(39,54; 41,12)	38,14	(37,32; 38,96)
Метод хот-дек				
10	39,90	(38,74; 41,06)	36,46	(35,57; 37,35)
25	39,40	(38,19; 40,62)	37,67	(36,69; 38,65)
50	40,00	(38,90; 41,09)	38,53	(37,34; 39,72)

Результати обчислень наведені в табл. 4, 5 (власні розрахунки авторів у середовищах SAS і SPSS відповідно).

Таблиця 4

Середня вага після лікування після проведення процедури відновлення пропусків у SAS

(кг)

Частка пропусків, %	Отримали лікування		Контрольна група	
	Середня вага, $\bar{x}$	Довірчі межі	Середня вага, $\bar{x}$	Довірчі межі
Метод регресійного аналізу				
0	39,68	x	36,79	x
10	39,67	(38,51; 40,83)	36,78	(35,92; 37,65)
25	39,66	(38,50; 40,82)	36,79	(35,93; 37,66)
50	39,66	(38,49; 40,82)	36,78	(35,91; 37,65)
Оцінка максимальної правдоподібності				
10	39,67	(38,51; 40,83)	36,78	(35,92; 37,65)
25	39,67	(38,50; 40,83)	36,79	(35,92; 37,66)
50	39,66	(38,50; 40,83)	36,78	(35,91; 37,65)
Метод множинної імпутації				
10	39,67	(39,32; 40,03)	36,78	(36,53; 37,04)
25	39,67	(39,31; 40,02)	36,78	(36,53; 37,04)
50	39,66	(39,31; 40,02)	36,78	(36,53; 37,04)

Таблиця 5

Середня вага після лікування після проведення процедури відновлення пропусків в SPSS

(кг)

Частка пропусків, %	Отримали лікування		Контрольна група	
	Середня вага, $\bar{x}$	Довірчі межі	Середня вага, $\bar{x}$	Довірчі межі
Метод регресійного аналізу				
1	**2**	**3**	**4**	**5**
0	39,68	x	36,79	x
10	39,85	(38,70; 41,00)	37,15	(36,15; 38,15)

Продовження табл. 5

1	2	3	4	5
25	39,69	(38,53; 40,85)	37,06	(36,19; 37,93)
50	39,89	(38,79; 40,99)	37,88	(36,50; 39,27)
Оцінка максимальної правдоподібності				
Частка пропусків, %	**Отримали лікування, кг**		**Контрольна група, кг**	
	Середня вага, $\bar{x}$	**Довірчі межі**	**Середня вага, $\bar{x}$**	**Довірчі межі**
10	39,67	(38,51; 40,83)	36,78	(35,92; 37,65)
25	39,67	(38,51; 40,83)	36,78	(35,92; 37,65)
50	39,67	(38,51; 40,83)	36,78	(35,92; 37,65)
Метод множинної імпутації				
Частка пропусків, %	**Отримали лікування**		**Контрольна група**	
	Середня вага, $\bar{x}$	**Довірчі межі**	**Середня вага, $\bar{x}$**	**Довірчі межі**
10	39,67	(39,17; 40,17)	36,78	(36,42; 37,15)
25	39,67	(39,17; 40,17)	36,78	(36,42; 37,15)
50	39,48	(38,98; 39,98)	37,12	(36,68; 37,57)

У результаті дослідження викладених методів, реалізованих у статистичних прикладних програмах, можна говорити про таке:

1) метод регресійного аналізу краще реалізований в SAS, оскільки середнє на відновлених даних практично не відрізняється не тільки від середнього, отриманого на повних даних, а й від результатів, обчислених з використанням інших, більш потужних методів;

2) на даних із 50% пропусків SAS дає точніші оцінки, ніж SPSS; більше того, результати, отримані в SAS різними методами, практично не відрізняються один від одного, що свідчить про їхню однакову потужність;

3) у SPSS при 50% пропусків доцільніше користуватися оцінками максимальної правдоподібності, ніж методом множинної імпутації або методом регресійного аналізу; крім того, регресійний аналіз у SPSS не дає стійких результатів, а тому цей ППП слід застосовувати дуже обережно незалежно від кількості пропусків.

Висновки. У дослідженні обґрунтовано, що робота з пропущеними даними охоплює широкий спектр проблем. Серед них – вивчення природи пропусків, вибір методології обробки й відновлення даних залежно як від їхньої природи, так і від типу, а також використання різних програмних засобів відновлення даних.

Отже, найзручнішим та найпростішим засобом відновлення пропусків є ППП Statistica. Однак способи відновлення даних тут обмежені наявними програмними функціями, тобто Statistica допоможе впоратися з проблемою за незначного обсягу пропусків (до 10%). Наступним за простотою програмним середовищем оброблення пропущених даних є SPSS, що порівняно зі Statistica пропонує ширший спектр методів відновлення даних, а також зрозуміліший інтерфейс для користувача проти мов програмування R чи SAS. Проте останні залишаються найпотужнішими програмами відновлення даних, надаючи можливість не тільки обробляти великі масиви даних зі значною часткою пропусків, а й застосовувати різні методи відновлення даних від найпростіших до найскладніших процедур. Водночас робота у середовищах R та SAS має свою специфіку, потребує знання відповідних мов програмування і спеціальної фахової підготовки.

Слід зазначити, що жодне з розглянутих програмно-аналітичних середовищ не має вбудованих процедур обробки категоріальних даних. Є певні підходи, які можна реалізувати за аналогією для упорядкованих категорій у програмних середовищах R та SAS, проте це не охоплює всі потреби аналізу досліджень, які реалізовані у вигляді опитувань і результати яких здебільшого представлені як відповіді на запитання. Незважаючи на те, що при використанні цифр для кодування відповідей код умовно відображає кількісну цифрову послідовність, він по суті не є числом і йому не притаманні властивості числа, а отже, методи, застосовувані для кількісних даних, не можуть бути поширені на категоріальні. Подальше дослідження буде орієнтовано на оцінку ефективності та відновлюваної здатності методів і процедур аналізованих ППП, а також на розроблення методологічних засад відновлення пропусків для категоріальних даних та вивчення можливостей практичної реалізації цих процесів у різних програмно-аналітичних середовищах.

Список використаних джерел

1. The Prevention and Treatment of Missing Data in Clinical Trials / R. J. Little et al. *The New England Journal of Medicine*. 2012. Vol. 367, № 14. P. 1355–1360. URL: http://www.nejm.org/doi/pdf/10.1056/nejmsr1203730
2. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных. *Computer Modelling & New Technologies*. 2002. Vol. 6, № 1. P. 51–61.
3. Кутлалиев А. Х. Метод множественного восстановления данных. *Социологические методы в современной исследовательской практике:* сб. статей. URL: https://publications.hse.ru/mirror/pubs/share/folder/21tn35z9vl/direct/92272011
4. Литтл Р. Дж. А., Рубин Д. Б. Статистический анализ данных с пропусками. Москва: Финансы и статистика, 1990. 336 с.
5. Ratitch B., & O'Kelly M. Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures. *Proceedings of PharmaSUG*. 2011. URL: https://www.pharmasug.org/proceedings/2011/SP/PharmaSUG-2011-SP04.pdf
6. Ratitch B., O'Kelly M., Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics*. 2013. Vol. 12, Is. 6. P. 337–347.
7. Yuan Y. Sensitivity Analysis in Multiple Imputation for Missing Data. *Proceedings of PharmaSUG*. 2014. URL: http://support.sas.com/resources/papers/proceedings14/SAS270-2014.pdf
8. Smuk M. Missing Data Methodology: Sensitivity analysis after multiple imputation: PhD thesis. London School of Hygiene & Tropical Medicine. 2015. URL: https://researchonline.lshtm.ac.uk/id/eprint/2212896/1/2015_EPH_PhD_SMUK_M.pdf
9. Ковтун Н. В., Фаталієва А.-Н. Я. Нові тенденції у доказовій статистиці: проблеми імпутації даних // Статистика України. 2019. № 4. С. 4–13. Doi: 10.31767/su.4(87)2019.04.01.
10. IBM SPSS Statistics 25 Documentation. URL: https://www.ibm.com/support/pages/ibm-spss-statistics-25-documentation#en
11. Анализ пропущенных значений. Документация к IBM SPSS Statistics Subscription. IBM Knowledge Center. URL: https://www.ibm.com/support/knowledgecenter/ru/SSLVMB_sub/statistics_mainhelp_ddita/spss/mva/idh_miss.html
12. Наглядная статистика. Используем R! / Шипунов А. Б. и др. 2014. URL: https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf
13. StatSoft, Inc. Электронный учебник по статистике. Москва: StatSoft, 2012. URL: http://www.statsoft.ru/home/textbook/default.htm
14. Missing data in SAS. Introduction to SAS. UCLA: Statistical Consulting Group. URL: https://stats.idre.ucla.edu/sas/modules/missing-data-in-sas/
15. SAS 9.4 Product Documentation. SAS. Resources / Documentation. URL: https://support.sas.com/documentation/94/
16. Rdatasets. Vincent Arel-Bundock's Github projects. URL: https://vincentarelbundock.github.io/Rdatasets/datasets.html

References

1. Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., & Farrar, J. T., et al. (2012). The Prevention and Treatment of Missing Data in Clinical Trials. *The New England Journal of Medicine, Vol. 367,* 14. Retrieved from http://www.nejm.org/doi/pdf/10.1056/nejmsr1203730
2. Zloba, E., & Yatskiv, I. (2002). Statisticheskie metody vosstanovleniia propushchennykh dannykh [Statistical methods for missing data recovering]. *Computer Modelling & New Technologies, Vol. 6 (1),* 51–61 [in Russian].
3. Kutlaliev, A. (2011). Metod mnozhestvennoho vosstanovleniia dannykh [Multiple Data Imputation Method]. *Sotsiolohicheskie metody v sovremennoi issledovatelskoi praktike – Sociological methods in modern research practice,* 201–208. Retrieved from https://publications.hse.ru/mirror/pubs/share/folder/21tn35z9vl/direct/92272011 [in Russian].
4. Little, R. J. A., & Rubin, D. B. (1990). *Statisticheskii analiz dannykh s propuskami [Statistical analysis with missing data].* Moscow: Finance and Statistics [in Russian].
5. Ratitch, B., & O'Kelly, M. (2011). Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures. *Proceedings of PharmaSUG 2011.* Retrieved from https://www.pharmasug.org/proceedings/2011/SP/PharmaSUG-2011-SP04.pdf
6. Ratitch B., O'Kelly, M., & Tosiello, R. (2013). Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics, Vol. 12, Is. 6,* 337–347.

7. Yuan, Y. (2014). Sensitivity Analysis in Multiple Imputation for Missing Data. *Proceedings of PharmaSUG 2014*. Retrieved from https://support.sas.com/resources/papers/proceedings14/SAS270-2014.pdf

8. Smuk, M. (2015) Missing Data Methodology: Sensitivity analysis after multiple imputation. *PhD thesis*. London School of Hygiene & Tropical Medicine. Retrieved from https://researchonline.lshtm.ac.uk/id/eprint/2212896/1/2015_EPH_PhD_SMUK_M.pdf

9. Kovtun, N. V., & Fataliieva, A.-N. Y. (2019). New Trends in Evidence-based Statistics: Data Imputation Problems. *Statystyka Ukrainy – Statistics of Ukraine, 87 (4)*, 4–13. Retrieved from https://doi.org/10.31767/su.4(87)2019.04.01

10. IBM SPSS Statistics 25 Documentation. (2018). *www.ibm.com*. Retrieved from https://www.ibm.com/support/pages/ibm-spss-statistics-25-documentation#en

11. Missing Value Analysis. IBM SPSS Statistics Subscription documentation. IBM Knowledge Center. *www.ibm.com*. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_sub/statistics_kc_ddita_cloud/spss/product_landing_cloud.html

12. Shipunov, A. B., Baldin, E. M., Volkova, P. A., Korobeinikov, A. I., Nazarova, S. A., & Petrov, S. V. (2014). *Nahliadnaia statystyka. Ispolzuem R! [Visual statistics. Let us use R!]*. Retrieved from https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf [in Russian].

13. StatSoft, Inc. (2012). *Elektronnyy uchebnik po statistike [Electronic textbook on statistics]*. Moscow: StatSoft. Retrieved from http://www.statsoft.ru/home/textbook/default.htm [in Russian].

14. Missing data in SAS. Introduction to SAS. UCLA: Statistical Consulting Group. *stats.idre.ucla.edu*. Retrieved from https://stats.idre.ucla.edu/sas/modules/missing-data-in-sas/

15. SAS 9.4 Product Documentation. SAS. Resources / Documentation. *support.sas.com*. Retrieved from https://support.sas.com/documentation/94/

16. Rdatasets. Vincent Arel-Bundock's Github projects. *vincentarelbundock.github.io*. Retrieved from https://vincentarelbundock.github.io/Rdatasets/datasets.html

N. V. Kovtun,
DSc in Economics, Professor,
Department of Statistics and Demography,
Taras Shevchenko National University of Kyiv,
E-mail: kovtun_natali@ukr.net
Researcher ID: M-6596-2017,
ORCID: https://orcid.org/0000-0002-2935-8597;
A.-N. Ya. Fataliieva,
EDETEK Inc, Biostatistician I,
E-mail: nastya1nana@gmail.com
ORCID: https://orcid.org/0000-0001-5541-8509

Software Implementation of Missing Data Recovery: Comparative Analysis

The paper contains a comparative analysis of the possibilities of using different software products to solve the problem of missing data on the example of the sample for which different variants of data skips are simulated. The study provided an opportunity to identify the strengths and weaknesses of these software products, as well as to determine the effectiveness of a particular method for different amounts of missed information.

Thus, the easiest way to handle the situation with missing data is Statistica, but there are offered only simple methods of processing data with missing values in Statistica. So, this program will help to cope with the missed data when there is a small number of omissions (up to 10%). SPSS offers a wider range of data imputation methods than Statistica, and at the same time it offers a more user-friendly interface compared to the R or SAS programming language. In the R and SAS software environments, you can use different methods of missing data imputation from the simplest to the most complex, such as, for example, multiple imputation. Thus, R and SAS are the most powerful missing data recovery programs, but they are more complex for users because they require knowledge of the programming language.

It is found out that none of the mentioned software-analytical environments has built-in procedures for processing categorical data with missing values. There are approaches that can be implemented by analogy for ordered categories in R and SAS software environments, but it does not cover all the needs of the analysis of research, which are implemented in the form of surveys with the results that are mostly presented as answers.

The methods used to impute quantitative data cannot be applied to categorical data, even if numbers are used to encode responses.

The study undoubtedly proved that handling the missing data, as well as the choosing of possible ways to use certain methods of data imputation in different software environments should be approached very carefully and the problem of imputation should be solved in each case based on careful analysis of the existing database, considering not only the characteristics of the data and the number of gaps, but also the specific of a particular study.

Dealing with missing data involves a wide range of the issues, which includes both the exploration of the nature of gaps, the methodology for data processing and imputation, depending not only on their nature but also on the type and the use of various software environments on missing data imputation.

It is planned in future research to assess the effectiveness of the recoverability of imputation methods in different software environments, as well as to develop methodological principles for restoring gaps for categorical data and implement them into practice.

Key words: *missing data, interval types, data imputation techniques, SPSS Statistics, Statistica, software environment R, SAS.*

Бібліографічний опис для цитування:

Ковтун Н. В., Фаталієва А.-Н. Я. Програмна реалізація відновлення пропущених даних: порівняльний аналіз. *Статистика України*. 2020. № 4. С. 12–20. Doi: 10.31767/su.4(91)2020.04.02.

Bibliographic description for quoting:

Kovtun, N. V., & Fataliieva, A.-N. Ya. (2020). Prohramna realizatsiia vidnovlennia propushchenykh danykh: porivnialnyi analiz [Software Implementation of Missing Data Recovery: Comparative Analysis]. *Statystyka Ukrainy – Statistics of Ukraine, 4,* 12–20. Doi: 10.31767/su.4(91)2020.04.02.