МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
МАРІУПОЛЬСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ЕКОНОМІКО-ПРАВОВИЙ ФАКУЛЬТЕТ
КАФЕДРА СИСТЕМНОГО АНАЛІЗУ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

До захисту допустити:
В.о. зав. кафедри

Ганна МАРТИНЮК

«26» грудня 2023 р.

«БАГАТОФАКТОРНИЙ АНАЛІЗ ДАНИХ З МЕТОЮ ПОШУКУ ТА ВИКОРИСТАННЯ В АЛГОРИТМАХ КЕРУВАННЯ НОВИХ ЗНАНЬ»

Кваліфікаційна робота
здобувача вищої освіти другого (магістерського) рівня
освітньо-професійної програми
«Системний аналіз»
Сергеєва Данила Андрійовича
Науковий керівник:
Мартинюк Ганна Вадимівна,
кандидат технічних наук, доцент,
в.о. завідувача кафедри системного аналізу та інформаційних технологій
Рецензент:
Гнатюк Сергій Олександрович,
д.т.н., професор, декан факультету комп'ютерних наук та технологій Національного авіаційного університету

Кваліфікаційна робота захищена
з оцінкою задовільно 64 (D)
Секретар ЕК

«16» січня 2024 р.

КИЇВ
2024

ЗМІСТ

Вступ

В умовах сучасної економіки для підтримання конкурентоспроможності на високому рівні, а також для грамотного ведення ринкової діяльності фірмі необхідно проводити маркетингові дослідження для моніторингу її поточної діяльності.

Нестача і недостовірність інформації найчастіше є причиною неправильних прогнозів, а отже, і ухвалення неефективних управлінських рішень, тоді як грамотний аналіз даних у більшості випадків дає змогу уникнути цієї проблеми.

Одним із найважливіших завдань маркетингу і маркетингового дослідження є встановлення причинно-наслідкових зв'язків, виявлення закономірностей бізнес-процесу. Багатофакторний дисперсійний аналіз слугує інструментом дослідження впливу набору чинників, які є якісними змінними, на залежну кількісну змінну (обсяг і частота покупок, розмір доходу, споживча оцінка, рейтинг фірми тощо). При цьому в ролі якісних змінних можуть виступати характеристики як споживачів (наприклад стать, вік, рівень доходу), так і самої фірми (інтенсивність і концепція рекламної кампанії, варіанти упаковки, географічне розташування). Як чинники можуть бути розглянуті і зовнішні впливи, такі як економічна ситуація в країні, її кліматичні та культурні особливості, пора року.

У цій роботі розглядається застосування різних моделей багатофакторного аналізу із метою пошуку та використання в алгоритмах керування нових знань.

Об'єктом застосування було обрано конкретну фірму - аптечну мережу (на прохання власника фірми аналізовані дані наводять у знеособленому форматі).

Необхідно також зазначити, що сфера застосування багатофакторного аналізу не обмежується маркетинговими дослідженнями. Розглянутий метод

широко використовують у найрізноманітніших галузях науки, зокрема в психології, соціології, медицині, біології та агрономії

Метою роботи було дослідження особливостей застосування різних моделей багатофакторного аналізу.

Для досягнення поставленої мети було сформульовано такі завдання:

1. Вибрати оптимальні моделі для проведення дослідження
2. Провести багатофакторний аналіз за обраними моделями
3. Провести інтерпретацію отриманих результатів
4. Інформацію, отриману під час аналізу, узагальнити у висновку за підсумками дослідження планувалося:

1. Виявити переваги та недоліки досліджуваного методу
2. Оцінити потенційну цінність інформації, отриманої в результаті застосування методу

Предметом роботи є багатофакторний аналіз даних з метою пошуку та використання в алгоритмах керування нових знань.

Наукова новизна отриманих результатів: в дипломній роботі "Багатофакторний аналіз даних з метою пошуку та використання в алгоритмах керування нових знань" було реалізовано наступні нові підходи та відкриття:

Розробка інноваційної методології багатофакторного аналізу: описати, як була вдосконалена традиційна методологія багатофакторного аналізу, включаючи нові алгоритми чи методи обробки даних.

Практичне значення отриманих результатів: багатофакторний аналіз даних є потужним інструментом для отримання цінних інсайтів з великої кількості інформації і використовується в різних галузях для покращення прийняття рішень та досягнення більшої ефективності.

Структура дипломної роботи складається зі вступу, основної частини, висновків, списку використаних джерел та літератури.

РОЗДІЛ 1. БАГАТОФАКТОРНИЙ ДИСПЕРСІЙНИЙ АНАЛІЗ

1.1. Принцип дисперсійного аналізу

Багатофакторний дисперсійний аналіз є статистичним методом аналізу результатів спостережень, які залежать від різних чинників, що діють одночасно, з метою вибору найбільш значущих чинників і оцінки їхнього впливу на досліджуваний процес. За допомогою дисперсійного аналізу встановлюються зміни дисперсії результатів експерименту в разі зміни рівнів досліджуваного чинника. Якщо дисперсії відрізнятимуться значною мірою, то випливає висновок про значущий вплив чинника на середнє значення спостережуваної випадкової величини.

Нульовою гіпотезою в дисперсійному аналізі є твердження про рівність середніх значень:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_j$$

Альтернативною гіпотезою буде припущення про порушення хоча б однієї з цих рівностей. Нехай на випадкову величину *X* впливає деякий якісний чинник *F* , що має *p* рівнів, а кількість спостережень на кожному рівні чинника однакова і дорівнює q.

x - генеральне середнє значення всіх спостережень.

Уведемо позначення:

$S_{total} = \sum_{j=1}^{p} \sum_{i=1}^{q} (x_{ij} - \overline{x})^2$ - загальна сума квадратів відхилень спостережуваних значень від загального середнього;

$S_{BG} = q \sum_{i=1}^{q} (\overline{x}_i - \overline{x})^2$ - факторна (або міжгрупова, between-group) сума квадратів відхилень групових середніх від загального середнього, характерно що зумовлює розсіювання між групами;

Σ Σ

$$S_{WG} = \sum_{j=1}^{p} \sum_{i=1}^{q} (x_{ij} - \overline{x}_j)^2$$ остаточна внутрішньогрупова, within-group) сума квадратів відхилень спостережуваних значень групи від свого групового середнього, що характеризує розсіювання всередині груп, причому,

$$S_{total} = S_{BG} + S_{WG}.$$

Розділивши суми квадратів на відповідне їм число ступенів свободи, отримаємо загальну, факторну і залишкову дисперсії:

$$MS_{total} = \frac{S_{total}}{N-1}, \qquad MS_{BG} = \frac{S_{BG}}{p-1}, \qquad MS_{WG} = \frac{S_{WG}}{N-p}.$$

Якщо справедливою є гіпотеза H_0 , то експериментальні групи є випадковими вибірками з однієї й тієї самої генеральної сукупності, то тоді факторна та залишкова дисперсії є незміщеними оцінками дисперсії цієї сукупності, а отже, різняться незначно. Формально залишкова і факторна дисперсії порівнюються за допомогою
F-критерію, або критерію Фішера.

Для цього за формулою $\frac{MS_{BG}}{MS_{WG}}$

рахується значення статистики. Критичне ж значення F-критерію визначається бажаним рівнем значущості та властивостями F-розподілу, форму якого повністю задають ступені свободи, відповідні залишковій і факторній дисперсіям.

За наявності кількох чинників, аналогічні обчислення і перевірку за критерієм Фішера проводять для кожного з них. Формули для проведення дво- і трифакторного дисперсійного аналізу можна знайти в додатку. Також у цьому випадку, крім основної нульової гіпотези, проводять перевірку ще однієї гіпотези, згідно з якою комбінація чинників не справляє ефекту взаємодії на значення залежної змінної. Формальний запис цієї гіпотези буде розглянуто нижче.

Принцип дисперсійного аналізу є одним із ключових методів в статистиці та аналізі даних, який використовується для визначення різниці між середніми значеннями двох або більше груп. Цей метод був розроблений Рональдом Фішером в 1920-х роках і став фундаментальним інструментом у багатьох галузях науки, включаючи медицину, психологію, економіку та соціологію.

Основна ідея дисперсійного аналізу полягає в порівнянні дисперсії (внутрішньої і міжгрупової) в даних, щоб визначити, чи існує статистично значуща різниця між групами. Для цього розглядаються два типи дисперсії:

1. Внутрішньогрупова дисперсія: Ця дисперсія вимірює різницю між спостереженнями всередині кожної групи. Вона вказує на розкид даних всередині кожної групи та її динаміку.
2. Міжгрупова дисперсія: Ця дисперсія вимірює різницю між середніми значеннями кожної групи. Вона показує, наскільки великою є різниця між середніми значеннями різних груп.

Для проведення дисперсійного аналізу використовуються різні статистичні тести, зокрема аналіз варіації (ANOVA). Основна ідея цього тесту полягає в порівнянні відсотка зміни між групами з загальної варіації в даних. Якщо велика частина варіації пояснюється міжгруповою дисперсією, то це свідчить про те, що існує статистично значуща різниця між групами.

Один з основних підходів до дисперсійного аналізу - це однофакторний ANOVA, який використовується для порівняння трьох або більше груп.

Даний тест розраховує F-статистику, яка визначає ступінь різниці між групами та ступінь зміни всередині груп.

Ще однією важливою частиною дисперсійного аналізу є оцінка статистичної значущості різниці між групами. Для цього використовується рівень значущості (p-значення), який вказує на ймовірність того, що спостережувана різниця між групами може бути випадковою.

Якщо p-значення менше заданого рівня значущості (зазвичай 0,05), то можна вважати, що існує статистично значуща різниця між групами. У цьому випадку відхиляють нульову гіпотезу про відсутність різниці.

Принцип дисперсійного аналізу може бути застосований в різних сферах досліджень і аналізу даних. Наприклад, у медицині він може бути використаний для порівняння ефективності різних методів лікування, у психології - для дослідження впливу різних факторів на психічне здоров'я, у економіці - для аналізу впливу економічних чинників на доходи різних груп населення.

Також важливо враховувати обмеження та припущення, які супроводжують дисперсійний аналіз. Один із головних припущень - це незалежність спостережень та нормальний розподіл даних. Якщо ці припущення не виконуються, то результати дисперсійного аналізу можуть бути неточними або недостовірними.

У підсумку, принцип дисперсійного аналізу є важливим інструментом в аналізі даних, який дозволяє визначити статистично значущі різниці між групами та робити висновки на основі цих різниць. Він знаходить застосування в різних галузях та дослідженнях і допомагає науковцям та аналітикам отримувати більш глибоке розуміння даних та явищ, які вони вивчають.

1.2. Модель багатофакторного аналізу

Модель багатофакторного аналізу є важливим інструментом в науці та дослідженнях, який дозволяє аналізувати взаємодію різних факторів на результати досліджень або спостережень. Ця модель допомагає виявити, як різні змінні впливають на результат та як вони можуть взаємодіяти між собою. Вона широко використовується в різних галузях, включаючи психологію, медицину, соціологію, економіку та інші.

Основними поняттями, пов'язаними з багатофакторним аналізом, є фактори та залежні змінні. Фактори - це змінні, які досліджуються в контексті дослідження і можуть впливати на результат. Наприклад, у дослідженні ефективності лікування факторами можуть бути тип лікування, вік пацієнтів і стать. Залежні змінні - це ті змінні, результати яких ми спостерігаємо і аналізуємо. У вищезгаданому прикладі це може бути здоров'я пацієнтів після лікування.

Багатофакторний аналіз дозволяє враховувати багато факторів одночасно і визначити, як вони взаємодіють між собою. Це дозволяє отримати більш точні та повні дані про вплив різних факторів на результати дослідження. Однак це також вимагає великої кількості даних та складного статистичного аналізу.

Існують різні методи багатофакторного аналізу, одним із найпоширеніших є аналіз дисперсії (ANOVA) і лінійна регресія. ANOVA використовується для визначення, чи існує статистично значуща різниця між групами, і чи впливають фактори на результати дослідження. Лінійна регресія дозволяє встановити ступінь впливу кожного фактора на залежну змінну.

Багатофакторний аналіз також дозволяє враховувати взаємодію між факторами. Наприклад, може виявитися, що ефект лікування залежить від віку пацієнтів, і що у різних вікових груп ефективність лікування різна.

Додатково до вже згаданих методів, багатофакторний аналіз може включати в себе розрахунок коефіцієнтів взаємодії між факторами. Це допомагає визначити, чи є взаємодія між факторами статистично значущою і як ця взаємодія впливає на результат.

Важливо пам'ятати, що багатофакторний аналіз має свої припущення і обмеження. Один з них - незалежність факторів. Це означає, що фактори не повинні бути взаємозалежніми. Якщо це припущення порушується, результати аналізу можуть бути неточними. Також важливо враховувати розмір вибірки, оскільки маленька вибірка може призвести до недостовірних результатів.

Трифакторний дисперсійний аналіз є ефективним інструментом у статистичному аналізі, який використовується для вивчення взаємодії трьох факторів на залежну змінну. Цей метод дозволяє досліджувати, як окремі фактори впливають на результат, а також як ці фактори можуть взаємодіяти між собою.

Перш за все, важливо розуміти, що таке фактор у контексті дисперсійного аналізу. Фактор - це категоріальна змінна, яка має декілька рівнів або категорій. Наприклад, у дослідженні впливу дієти, віксельної та статі на вагу людини, "дієта", "вік" та "стать" будуть розглядатися як фактори.

Для проведення трифакторного дисперсійного аналізу потрібно мати вибірку, в якій кожен спостережуваний об'єкт належить до певного комбінації рівнів усіх трьох факторів. Таким чином, кожному спостереженню можна присвоїти трійкове число, що відповідає комбінації рівнів трьох факторів. Наприклад, 1-1-1, 1-1-2, 1-2-1 та ін.

Одним з основних припущень дисперсійного аналізу є припущення про однорідність дисперсій. Це означає, що дисперсії залежної змінної є однаковими для всіх комбінацій рівнів факторів. Це припущення може перевіряти за допомогою статистичних тестів, таких як тест Левена або Бартлетта.

Коли припущення про однорідність дисперсій виконується, можна провести аналіз залежності між факторами та залежною змінною. Одним з головних показників, який оцінюється в дисперсійному аналізі, є F-статистика. Ця статистика відображає відношення між факторованою та внутрішньою дисперсією. Велике значення F-статистики вказує на те, що є статистично значуща взаємодія між факторами.

Якщо аналіз показує статистично значущу взаємодію між факторами, то слід додатково вивчати цю взаємодію. Для цього можна проводити пост-хок тестування, таке як тест Тьюкі, щоб визначити, між якими конкретними групами факторів існує статистично значуща відмінність.

Трифакторний дисперсійний аналіз може бути особливо корисним у ситуаціях, коли дослідники хочуть вивчити взаємодію між трьома факторами, а не просто вплив кожного фактора окремо. Наприклад, у медичних дослідженнях може бути цікаво дослідити, як доза ліку, стать пацієнта та тривалість лікування взаємодіють між собою на ефективність лікування.

Окрім того, трифакторний дисперсійний аналіз може бути використаний для вивчення взаємодії між різними факторами в різних контекстах. Наприклад, у соціологічних дослідженнях може бути цікаво вивчити, як економічний статус, освіта та регіон проживання впливають на ставлення людей до питань громадської політики.

Побудуємо модель трифакторного дисперсійного аналізу. Нехай на випадкову величину *X* впливають фактори *A, B, C*, що мають *a*, *b* і *c* рівнів відповідно. Позначимо через y_{ijkt} результат t-го вимірювання, проведеного на рівні *i* фактора *A*, рівні *j* фактора *B* і рівні *k* фактора *C*. Модель матиме такий вигляд[4] :

$$y_{ijkt} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkt}$$

де:

μ - глобальне середнє значення ознаки *y*;

μ_{ijk} - середнє значення спостережень на перетині рівнів *i*, *j* і *k* факторів *A*, *B* і *C* відповідно;

$\alpha_i = \mu_{i..} - \mu$ - ефект рівня *i* фактора *A*, де $\mu_{i..}$ - середнє значення при-

знака *y* на і-му рівні фактора А;

β_j і γ_k - ефекти рівня *j* фактора *B* і рівня *k* фактора *C* відповідно;

$(\alpha\beta)_{ij} = \mu_{ij.} - (\mu + \alpha_i + \beta_j)$ - ефект взаємодії для комбінації рівня

i фактора *A* і рівня *j* фактора *B*, де $\mu_{ij.}$ - середнє значення ознаки *y* на перетині і-го рівня фактора *A* і j-го рівня фактора *B*;

Аналогічно визначаються $(\beta\gamma)_{jk}$ і $(\alpha\gamma)_{ik}$;

$(\alpha\beta\gamma)_{ijk} = \mu_{ijk} - (\mu + (\alpha\beta\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + \alpha_i + \beta_j + \gamma_k)$ - ефект взаємодії для комбінації рівня *i* фактора *A*, рівня *j* фактора *B* і рівня *k* фактора *C*;

ε_{ijkt} - випадкова помилка t-го вимірювання на перетині рівнів *i*, *j* і *k* факторів *A*, *B* і *C* відповідно;

Нульові гіпотези можна записати таким чином:

H_0 A : $\alpha_i = 0, \forall i$

H_0 B : $\beta_j = 0, \forall$j H_0

C : $\gamma_k = 0, \forall k$

Кожна зі сформульованих гіпотез еквівалентна гіпотезі про рівність середніх рівнів фактора, сформульованій у попередньому параграфі.

Оскільки на залежну змінну діє більше, ніж один фактор, додаються нульові гіпотези про наявність ефекту взаємодії факторів:

$$H_{0_{AB}} : (\alpha\beta)_{ij} = 0, \forall i, j$$

$$H_{0_{BC}} : (\beta\gamma)_{jk} = 0, \forall j, k$$

$$H_{0_{AC}} : (\alpha\gamma)_{ik} = 0, \forall i, k$$

$$H_{0_{ABC}} : (\alpha\beta\gamma)_{ijk} = 0, \forall i, j, k$$

1.3. Основні припущення дисперсійного аналізу

Класичні методи дисперсійного аналізу ґрунтуються на таких передумовах:

- Усі вибірки мають випадковий і незалежний характер
- Розподіл вихідних випадкових величин нормальний
- Дисперсії експериментальних даних однакові на різних рівнях досліджуваного фактора (умова гомоскедастичності)

1.3.1. Перевірка умови нормальності розподілу генеральної сукупності

У цій роботі було ухвалено рішення про використання критерію Шапіро-Вілка для перевірки гіпотези про нормальність розподілу генеральної сукупності. Цей критерій було обрано, оскільки вивчення його потужності показало, що він є одним із найефективніших критеріїв перевірки нормальності розподілу випадкових величин. Серед недоліків цього методу можна згадати його зміщеність за малих обсягів вибірок стосовно альтернатив, більш порівняно з нормальним законом. Але оскільки в цьому дослідженні обсяг вибірок доволі великий (n = 100), критерій Шапіро-Вілка можна вважати оптимальним і прийняти як основний інструмент перевірки гіпотези нормальності розподілу генеральної сукупності.

Тест заснований на відношенні оптимальної лінійної незміщеної оцінки дисперсії до її звичайної оцінки методом максимальної правдоподібності. Статистика критерію має вигляд

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$W = \frac{1}{s^2}[\sum_{i=1}^{k} a_{n-i+1}(x_{n-i+1} - x_i)]^2,$$

$$s^2 = \sum_{i=1}^{n}(x_i - \bar{x});\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Де

Коефіцієнти a_{n-i+1} можна знайти в таблиці коефіцієнтів критерію Шапіро-Вілка. Якщо $W < W(\alpha)$, то нульова гіпотеза нормальності розподілу відхиляється на рівні значущості α, критичні значення $W(\alpha)$ можна знайти в таблиці процентних точок критерію $W(\alpha)$.[1]

1.3.2. Перевірка умови гомоскедастичності

Умова гомоскедастичності може бути перевірена кількома способами, що включають критерії Гартлі, Кохрана, Левене, Фліґнера-Кілліна та Бартлетта. Деякі з цих критеріїв є надто чутливими до порушення умови нормальності (критерій Бартлетта), критерій Флігнера-Кілліна хоч і є непараметричним, але передбачає рівність медіан тестованих вибірок. Крім того, непараметричні критерії значно поступаються в потужності параметричним. Дослідження показали, що критерій Кохрана є найпотужнішим із перелічених критеріїв, зберігаючи цю властивість і при порушенні умови нормальності. Тому вибір було зроблено на користь критерію Кохрана.

Нульова гіпотеза для m вибірок може бути записана так:

$$H_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_m.$$

Альтернативною є гіпотеза про порушення хоча б однієї з цих рівностей.

Статистика цього критерію виражається формулою

$$Q = \frac{S_{max}}{S_1^2 + S_2^2 + \ldots + S_m^2},$$

де $S_{max} = max(S^2, S^2, \ldots, S^2)$, m - число вибірок, S^2 - оцінки вибр.
1 i

рочних дисперсій. Критичне значення критерію можна обчислити таким чином:

$$C_{UL}(\alpha, n, m) = [1 + \frac{m -}{F_c(\alpha/m, (n-1), (}$$

де n - кількість спостережень у кожній вибірці, F_c критичне значення розподілу Фішера. Якщо $C_j > C_{UL}$ хоча б для одного j, то нульова гіпотеза відхиляється.

РОЗДІЛ 2. ДВОФАКТОРНИЙ ДИСПЕРСІЙНИЙ АНАЛІЗ

2.1. Двофакторний дисперсійний аналіз із повторними вимірами

Визначення 2.1: *Дисперсійний аналіз із повторними вимірюваннями*

- це такий вид дисперсійного аналізу, для якого на кожному рівні досліджуваного чинника вимірювання проводять на одних і тих самих суб'єктах. Розглянемо його для випадку двох чинників.

Повна сума квадратів у цьому випадку розбивається таким чином:

$$S_{total} = S_A + S_B + S_{AB} + S_{WG},$$

де S_{WG} , своєю чергою, розбивається на S_{Err} і S_S .

S_S - внутрішньогрупова сума квадратів для фактора суб'єкти, S_{Err} - сума квадратів помилки. Річ у тім, що внаслідок того, що на кожному рівні фактора вимірювання проводять на одних і тих самих суб'єктах, ці суб'єкти можуть розглядатися як окремий фактор. Усі подальші обчислення здійснюють аналогічно до простого випадку, з тією лише різницею, що для обчислення внутрішньогрупової дисперсії замість S_{WG} використовують S_{Err} (у разі класичного дисперсійного аналізу S_{WG} збігається з S_{Err}). Така модель дисперсійного аналізу має перевагу порівняно з класичною моделлю: за рахунок зменшення суми квадратів помилки збільшується значення F-статистики, а це своєю чергою призводить до підвищення потужності критерію для виявлення значущих відмінностей середніх.

Для проведення дисперсійного аналізу з повтореннями необхідно, щоб, крім трьох основних припущень, виконувалося ще й припущення *сферичності*.

Визначення 2.2: *Сферичність* - властивість, згідно з якою, дисперсії

різниць між різними рівнями фактора з повторними вимірюваннями рівні. Нульова гіпотеза про те, що набір вибірок задовольняє умову сферичності, перевіряється за допомогою тесту Моучлі. У разі порушення умови сферичності ймовірність помилкового відхилення нульової гіпотези стає більшою, ніж рівень значущості α, для усунення цього ефекту виконують виправлення статистики F-критерію за методом Грінгауса-Гейсера.

Двофакторний дисперсійний аналіз із повторними вимірами є одним із потужних статистичних методів, який використовується для вивчення впливу двох або більше факторів на змінну величину. Цей метод дозволяє досліджувати взаємодію між факторами та визначати їхні впливи на результати експерименту. У даній статті ми розглянемо основні принципи і процедури двофакторного дисперсійного аналізу із повторними вимірами та його практичне застосування.

Двофакторний дисперсійний аналіз із повторними вимірами передбачає використання двох факторів (незалежних змінних), які можуть впливати на залежну змінну, і багаторазове вимірювання цієї залежної змінної у різних умовах. Цей метод особливо корисний у випадках, коли ми хочемо вивчити, як різні фактори впливають на змінну величину в різні моменти часу або в різних умовах досліду.

Для проведення двофакторного дисперсійного аналізу із повторними вимірами, спершу необхідно зібрати дані, які включають у себе значення залежної змінної для кожного комбінованого значення обох факторів. Наприклад, якщо ми вивчаємо вплив двох факторів - типу лікування і часу, то для кожного пацієнта ми будемо мати значення залежної змінної для кожної комбінації типу лікування і часу. Це може бути зроблено за допомогою спеціально розробленого експерименту або досліду.

Далі, для проведення аналізу, необхідно обчислити середні значення залежної змінної для кожної комбінації факторів. Це допомагає отримати загальне уявлення про вплив кожного фактора і їх взаємодію на результати

експерименту.

Основними кроками в аналізі є розрахунок сум квадратів відхилень (SST), які представляють загальну дисперсію результатів, і поділ цієї дисперсії на два компоненти: міжгрупову дисперсію (SSB) і внутрішньогрупову дисперсію (SSW). Міжгрупова дисперсія відображає вплив факторів на результати, тоді як внутрішньогрупова дисперсія відображає варіабельність в середніх значеннях в межах кожної комбінації факторів.

Для оцінки впливу факторів і їх взаємодії на результати експерименту проводиться аналіз дисперсії (ANOVA). Основна ідея ANOVA полягає в порівнянні внутрішньогрупової дисперсії з міжгруповою дисперсією. Якщо міжгрупова дисперсія виявляється статистично значущою, то це свідчить про те, що принаймні один з факторів має вплив на залежну змінну.

У двофакторному дисперсійному аналізі із повторними вимірами також вивчається взаємодія між факторами. Для цього проводиться аналіз взаємодії, який допомагає визначити, чи є статистично значущою вплив факторів один на одного. Якщо аналіз взаємодії показує статистично значущий результат, то це свідчить про те, що ефект одного фактора залежить від значень іншого фактора.

У додаток до основного аналізу, важливо також провести статистичні тести для перевірки статистичної значущості різниць між групами або комбінаціями факторів. Зазвичай для цього використовуються t-тести, або інші аналогічні методи, залежно від специфіки досліду.

Застосування двофакторного дисперсійного аналізу із повторними вимірами може бути знайдено в багатьох галузях науки та практики. Наприклад, в медицинському дослідженні він може бути використаний для вивчення впливу двох методів лікування на здоров'я пацієнтів протягом певного періоду часу. У психологічних дослідженнях він може бути застосований для вивчення впливу двох факторів на психологічні показники у різні моменти часу.

Двофакторний дисперсійний аналіз із повторними вимірами також

використовується в експериментальних дослідженнях та виробничому контролі для визначення ефективності двох або більше факторів на якість продукції або процес виробництва. Він дозволяє визначити оптимальні умови для досягнення бажаних результатів і покращення якості продукції.У підсумку, двофакторний дисперсійний аналіз із повторними вимірами є потужним інструментом для вивчення впливу двох або більше факторів на залежну змінну у різних умовах або в різні моменти часу. Він дозволяє досліджувати міжгрупові та внутрішньогрупові варіації, а також взаємодію між факторами. Цей метод має широкий спектр застосування і важливий для багатьох галузей науки та практики, де необхідно вивчити вплив багатьох факторів на результати досліджень чи виробництва.

2.2. Опис даних

Об'єктом дослідження в цій роботі була база даних мережі аптек (назву не розголошують у зв'язку з комерційною таємницею), яка містила інформацію про дохід за кожною одиницею товару в трьох різних філіях, деталізовану за місяцями (було подано дані за 2015 рік).

Аптечна мережа містить три філії, одна з яких розташована на першому поверсі торговельного центру (далі - *філія 3*), інші дві торгові точки розташовані в незалежних торговельних приміщеннях (*філія 1*, *філія 2*).

Метою дослідження було з'ясувати, чи впливають пора року і місце розташування конкретної філії на дохід мережі за основною категорією товарів - лікарськими препаратами. Багатофакторний дисперсійний аналіз було обрано основним інструментом аналізу. Математичні розрахунки та побудову графіків проводили за допомогою статистичного пакета R. Оскільки вимірювання за кожним товаром здійснюють на всіх рівнях обох чинників (для вимірювань на різних рівнях використовують одні й ті самі найменування товарів), аналіз проводили за моделлю двофакторного дисперсійного аналізу з

повтореннями.

2.3. Збір і підготовка даних

Із генеральної сукупності товарів категорії *лікарські засоби*, представлених в асортименті кожної з філій, випадковим чином було відібрано 100 найменувань. Для кожної пори року було обчислено сумарний дохід за кожним товаром. Аналогічні обчислення було проведено для решти двох філій мережі. Таким чином, було сформовано дані, що складаються з дванадцяти вибірок (для всіх поєднань рівнів чинників *філія* і *сезон*) або груп. Для подальшого опрацювання дані було сформовано в таблицю, що містить тринадцять стовпців (стовпчик "найменування" і дванадцять стовпців, кожен з яких представляє собою одну з вищезазначених груп). За допомогою функції read.delim було здійснено імпорт даних в R.

2.4. Перевірка даних

Перед проведенням дослідження необхідно переконатися, що підготовлені дані задовольняють основним положенням дисперсійного аналізу, описаним у попередньому розділі.

1. Дані було відібрано з генеральної сукупності випадковим чином, а отже, кожна вибірка має випадковий і незалежний характер.
2. Для перевірки припущення про нормальний розподіл залежної змінної, якою є змінна "дохід, було використано функцію shapiro.test, що здійснює перевірку нульової гіпотези про нормальний розподіл вибірки за критерієм Шапіро-Вілка. Оскільки для кожної з тестованих груп p-значення (імовірність помилки першого роду) перевищило значення заданого рівня значущості.

Перевірку гомоскедастичності груп було здійснено за допомогою

функції cochran.test, що є програмною реалізацією критерію Кохрана. Імовірність помилки першого роду, як і в попередньому пункті, перевищила задане значення $\alpha = 0.05$, що дало змогу зробити вибір на користь прийняття нульової гіпотези про рівність дисбалансів генеральних сукупностей, з яких тестовані вибірки було вилучено. Переконавшись, що дані задовольняють вихідним положенням ANOVA, можна переходити до роботи з ними.

2.5. Дисперсійний аналіз

Через особливості роботи в R з даними, що містять повторні вимірювання, таблицю потрібно трансформувати у формат long (на один рядок припадає одне спостереження). Для цього використовується функція melt. Дані в новому форматі збережено у змінній longData. Нова таблиця складається з трьох стовпчиків: "найменування", "групи", що містить імена вихідних стовпчиків, з якої взято інформацію про дохід, і "дохід", що являє собою стовпчик із даними про дохід за кожним найменуванням. Можна помітити, що стовпчик "групи" містить інформацію як про зону, так і про філію. Розділимо ці ознаки. Знаючи, що перші 400 рядків містять інформацію про першу філію, кожні 100 з яких - про один із чотирьох сезонів, і що та сама логіка справедлива для наступних 800 рядків, створимо стовпчики "сезон" і "філія". Для цього використовуємо функцію gl. Тепер можна переходити безпосередньо до аналізу.

Загалом, побудуємо графік boxplot, щоб оцінити дані графічно.
Результат можна побачити нижче.

Ліва колонка: дохід, стовбці
1- зима, 2- весна, 3-літо, 4-осінь
Філіали: червоний – 1, зелений -2, блакитний -3.

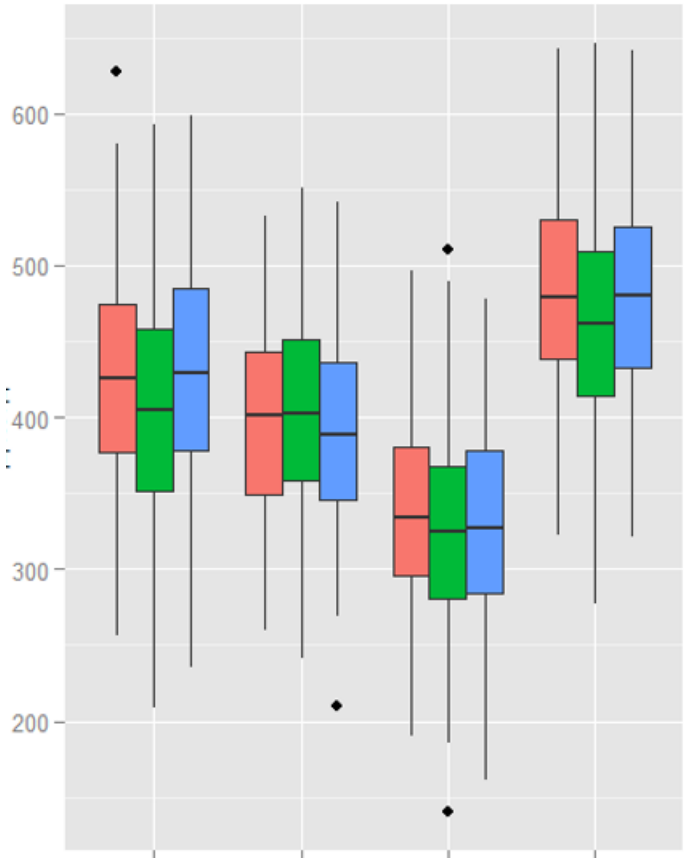


Рис. 1: графік boxplot

Як видно з графіка, деяка тенденція до зниження розміру доходу спостерігається в літній період, а в осінній період він, навпаки, перевищує аналогічне значення для інших сезонів. Подальший аналіз дасть змогу виявити значущість цих відмінностей, а також зробити висновки про наявність (відсутність) значущих відмінностей між розміром доходу серед трьох філій. Проведемо дисперсійний аналіз за допомогою функції aov_ez пакета afex. Застосувавши до побудованої моделі функцію summary, виведемо на екран отриманий результат.

Univariate type iii RKepeated-Measures ANOVA Assuming sphericity

SS num Df Error SS den Df F Pr(>F)

(Intercept) 196501016 1 = $42914 99 35831.8168 < 2e-16 ***

сезон 3238953 3 1301470 297 =246.3801 < 2e-16 ***

24575 2 964142 198 2.5234 0.08276 .

сезон: Філіал $2797 6 2700122 594 1.9358 0.07302 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.02 ‘#7 0.05 *.7 O.2 S' 1

Mauchly Tests for Sphericity

Test statistic p-value

сезон 0.95689 0.50620

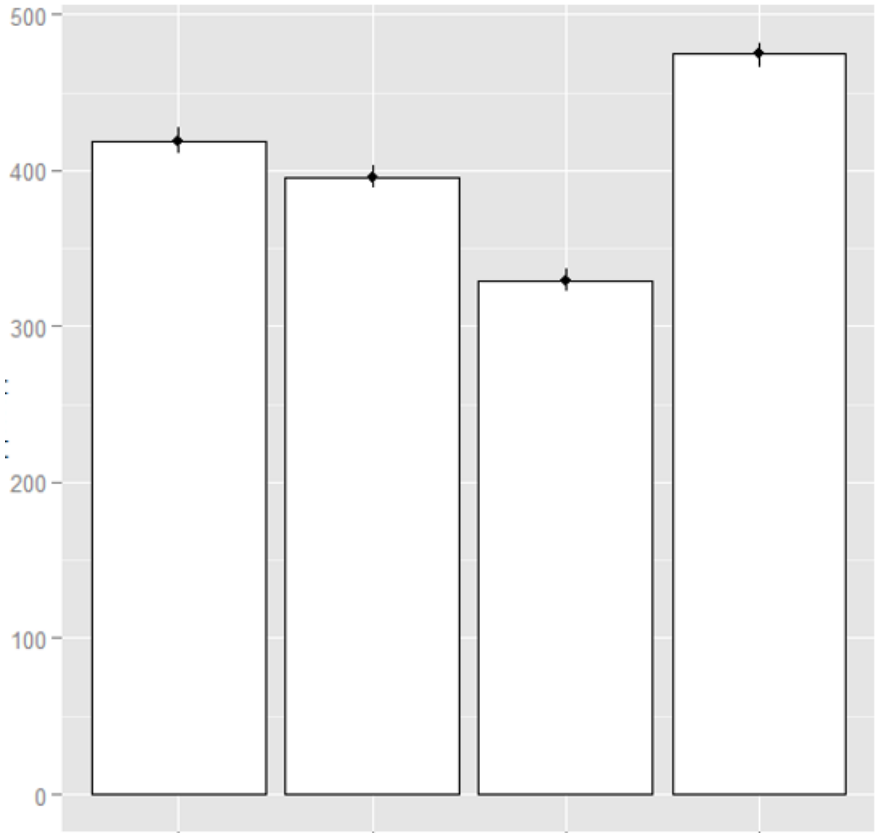
Філіал 0.97199 0.24856

Сезон: Філіал 0.84535 0.70061

Аналіз на значущість відмінностей рівнів чинників слід розпочати з оцінки результатів перевірки на сферичність за допомогою тесту Моучлі. За результатами цього тесту можна зробити висновок, що властивість сферичності для кожного чинника і для їхньої взаємодії (*р-значення*> $\alpha = 0.05$), тому немає необхідності використовувати виправлені р-значення. Подивившись на результати й оцінивши р-значення (імовірності помилки за відхилення нульової гіпотези), можна побачити, для яких чинників і комбінацій нульова гіпотеза може бути прийнята. Такий висновок ми можемо зробити для фактора *філія* - згідно з цим висновком, фактор *філія* не мав значного впливу на розмір доходу. Гіпотеза приймається і для взаємодії *сезон-філія* - це означає, що фактор *сезон* мав однаковий вплив на розмір доходу в кожній із трьох філій. Аналіз дав змогу виявити наявність значущих відмінностей між рівнями фактора *сезон*.

Щоб візуально оцінити отримані результати, побудуємо графік.

1. Сезон



Де ліві рядки – дохід, а стовбці зима-весна-літо-осінь

Рис. 3: сезон

Згідно з результатами дисперсійного аналізу, фактор *сезон* є значущим. Це означає, що середнє значення доходу за кожним сезоном, що припадає на кожен товар, різниться хоча б для двох пір року. Дисперсійний аналіз перевіряє нульову гіпотезу про наявність відмінності між середніми значеннями рівнів фактора. Щоб з'ясувати, де саме лежить ця відмінність, необхідно скористатися методом контрастів.

Загальновідомим є той факт, що в літній період рівень продажів в аптеках як правило знижується, тому доцільно вважати, що

значення рівня *літо* контрольною вибіркою. За допомогою першого контрасту порівняємо значення доходу в літній період і значення доходу в інші періоди року. Далі, можна помітити, що згідно з побудованим графіком, середнє значення доходу в осінній період було дещо вищим за решту, тож за допомогою другого контрасту порівняємо рівні зима і весна з рівнем осінь. Третій контраст дасть змогу порівняти між собою дохід у зимовий і весняний періоди. Створимо змінні для цих контрастів:

```
ZVOvsL<-c(1, 1,-3,1)
ZVvsO<-c(1, 1, 0,-2)
ZvsV<-c(1, -1, 0,0)
```

За допомогою функції contrast проведемо додатковий аналіз за методом контрастів.

```
Contrast
Всі_літо
Зимавесна_осінь
Зима_весна
```

estimate	SE	df	t.ratio	p.value
300.18641	13.239417	297	22.674	<.0001
-134.60478	9.361682	297	-14.378	<.0001
23.12772	5.404969	297	4.279	<.0001

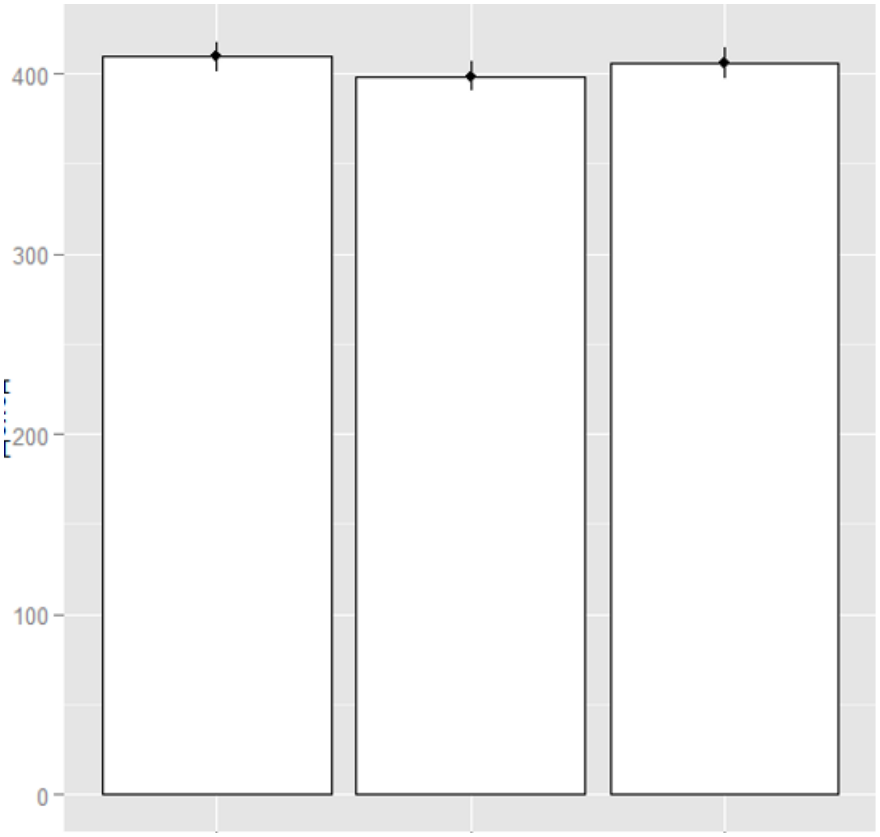
Рис. 4: метод контрастів

Як можна бачити на рисунку 4, *p-значення* для кожного зі встановлених контрастів виявилося досить малим, щоб вважати кожен із них значимим. Проводячи інтерпретацію результатів, отриманих під час дослідження чинника *сезон*, можна зробити висновок, що дохід у літній період істотно нижчий, ніж у решту пір року, дохід у першому півріччі значно нижчий, ніж восени, а в разі порівняння зимового та весняного періоду виявляється

перевага значення середнього доходу взимку. Іншими словами, рівень доходу досліджуваного підприємства має сезонний характер.

2. Філія

За результатами дисперсійного аналізу значущих відмінностей між середніми значеннями доходу за одиницю товару за трьома філіями виявлено не було, що абсолютно узгоджується з візуальним уявленням.



Рядки- дохід

Стовпці – Філіал-1, Філіал-2, Філіал-3 відповідно

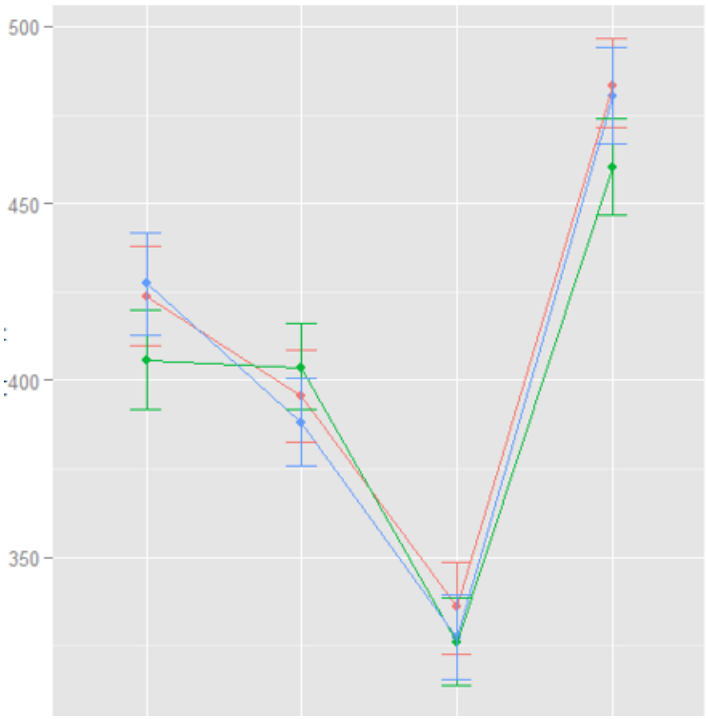
Рис. 5: філія

Такий результат не потребує проведення додаткових досліджень.

Інтерпретувати отриманий результат можна так: торгові точки приносять приблизно однаковий дохід, незалежно від їхнього місця розташування. Примітним є і той факт, що відмінностей між рівнем доходу від продажів у філії, розташованій у торговому центрі та продажів у двох інших торгових

точках виявлено не було.

3. Сезон і філія



Рядки- дохід

Стовпці – зима_весна_літо_осінь відповідно

Червоний – філіал 1, зелений філіал 2, блакитний – філіал 3

Рис. 6: ефект взаємодії

Взаємодія чинників *сезон* і *філія*, згідно з проведеним аналізом, також не є значущою. Це означає, що вплив фактора сезон на всіх рівнях фактора філія проявляється однаково. Інакше кажучи, явище сезонності проявляється однаково в кожній із трьох філій.

За результатами дослідження можна дійти висновку, що з двох досліджуваних чинників, істотний вплив на рівень доходу справляє лише

один із них - сезон. Ефекту взаємодії чинників виявлено не було, а отже, явище сезонності проявляється однаково в кожній із трьох філій. Підбиваючи підсумок, можна сказати, що середній дохід, одержуваний від продажу низки товарів категорії "лікарські препарати", варіюється залежно від сезону, досягаючи свого максимуму в осінній період і суттєво знижуючись у літній. Рівні доходу в різних точках продажу не мають значущих відмінностей.

РОЗДІЛ 3. ТРИФАКТОРНИЙ ДИСПЕРСІЙНИЙ АНАЛІЗ: ЗМІШАНА МОДЕЛЬ

Визначення 3.1: *Багатофакторний дисперсійний аналіз за змішаною моделлю* - це такий різновид дисперсійного аналізу, що охоплює і міжгрупові (змінні, на різних рівнях яких вимірювання проводять на одних і тих самих суб'єктах), і внутрішньогрупові змінні. Розглянемо цю модель на прикладі двофакторного аналізу. Нехай *A* - міжгрупова змінна, а *B* - внутрішньогрупова. Повна сума квадратів у цьому випадку розбивається таким чином:

$$S_{total} = S_A + S_B + S_{AB} + S_{E_a} + S_{E_b} + S_{WG},$$

где:

$$S_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}\sum_{k=1}^{t} y_{ijk}^2 - Nt\bar{y}_{...}^2$$

$$S_A = t\sum_{i=1}^{a} n\bar{y}_{i..}^2 - Nt\bar{y}_{...}^2$$

$$S_B = N\sum_{k=1}^{t} \bar{y}_{..k}^2 - Nt\bar{y}_{...}^2$$

$$S_{AB} = \sum_{i=1}^{a}\sum_{k=1}^{t} n\bar{y}_{i.k}^2 - Nt\bar{y}_{...}^2 - S_A - S_B$$

$$S_{E_a} = t\sum_{i=1}^{a}\sum_{j=1}^{n} \bar{y}_{ij.}^2 - t\sum_{i=1}^{a} n_i\bar{y}_{i..}^2$$

$$S_{E_b} = S_{total} - S_A - S_B - S_{AB} - S_{E_a}$$

N - загальна кількість спостережень, *a* - кількість рівнів фактора *A*, *t* - кількість рівнів фактора *B*, *n* - кількість спостережень у кожній комірці.[17] Подальші обчислення проводяться аналогічно класичній моделі.

3.1 Планування

Дослідження, проведене в рамках розділу 1, можна розширити, додавши до нього ще один фактор. Таким фактором було обрано категорію товарів.

Новий варіант дослідження передбачає розширення списку товарів, що аналізуються, за рахунок додавання до них найменувань із двох категорій, що становлять більшу частину асортименту мережі аптек - *БАДи* (біологічно активні добавки) і *предмети особистої гігієни*. З появою третього фактора, додалося ще три взаємодії, вплив яких також необхідно перевірити. Нова незалежна змінна - категорія, не є внутрішньогруповою (вимірювання для кожного рівня цієї змінної проводять на різних товарах), отже, дослідження набуває вигляду трифакторного дисперсійного аналізу, який проводять за змішаною моделлю (дві внутрішньогрупові змінні та одна міжгрупова).

3.2Збір даних

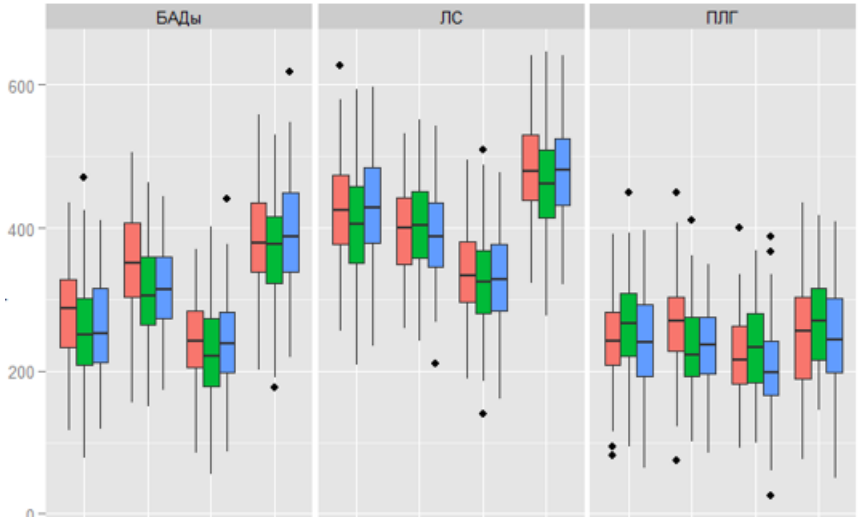
Як і в попередньому розділі, з генеральної сукупності кожної з трьох вищезазначених категорій товарів, представлених в асортименті кожної з філій, випадковим чином було відібрано найменувань -100 Для кожної пори року було обчислено сумарний дохід за кожним товаром. Такі маніпуляції було проведено для кожної з трьох філій. Таким чином, дані, підготовлені до опрацювання, являли собою таблицю, що містила чотирнадцять стовпчиків (стовпчики "найменування" і "категорія", і дванадцять стовпців з інформацією про дохід - по чотири стовпці, відповідно до кількості пір року, для кожної з трьох філій).

3.3 Дисперсійний аналіз

Після проведення перевірки отриманих вибірок на нормальність розподілу генеральних сукупностей і на рівність дисперсій, можна переходити до безпосередньої роботи з даними. Як і у випадку з дво-факторним аналізом, наявність повторних вимірювань зумовлює необхідність приведення таблиці у формат long за допомогою функції melt. Дані в новому форматі збережено у змінній Data1. Нова таблиця складається з шести колонок: "найменування", "категорія", ім'я колонки, з якої взято інформацію про дохід ("групи"),

колонка, що містить дані про дохід ("дохід"), а також дві колонки, які містять імена рівнів чинників сезон і філія. Перед початком аналізу побудуємо і проаналізуємо графік boxplot (рис. 7).

Дивлячись на графік, можна помітити, що сезонність найбільше виражена для категорій *БАДи* та *лікарські засоби*, тоді як рівень доходу за товарами категорії *предмети особистої гігієни* слабко коливається протягом року. Дисперсійний аналіз дасть змогу вирішити, чи варто прийняти або відкинути це припущення, і зробити подальші висновки.



Рядки- дохід

Стовпці – зима_весна_літо_осінь відповідно

Червоний – філіал 1, зелений філіал 2, блакитний – філіал 3

Рис. 7: графік boxplot

Аналіз на значущість відмінностей рівнів чинників слід розпочати з оцінки результатів перевірки на сферичність за допомогою тесту Моучлі. Звернувшись до результатів тесту Моучлі, що містяться в таблиці проведеного дисперсійного аналізу (рис. 8), можна дійти висновку, що властивість сферичності виконується для кожного чинника і для їхніх взаємодій (*p-значення*> 0.05).

Подивившись на результати й оцінивши *p-значення*, можна побачити для яких чинників і комбінацій нульова гіпотеза може бути відкинута. Такий висновок ми можемо зробити для всіх чинників і для всіх чинників. Оскільки взаємодія трьох чинників є значущою, інтерпретація головних ефектів чинників і взаємодії пар чинників може виявитися недостовірною. Це означає, що основні висновки мають ґрунтуватися саме на інтерпретації ефекту взаємодії трьох чинників. Тому варто відразу перейти до розгляду цього головного ефекту.

Значимість взаємодії трьох чинників означає, що одна або кілька подвійних взаємодій значно різняться вздовж рівнів третьої змінної.

Для початку побудуємо графік, що ілюструє взаємодію факторів *сезон* і *філія* для кожного з рівнів фактора *категорія*. Можна припустити, що ця взаємодія чинить більший вплив на третій рівень фактора *категорія*, ніж на два інші. Про це свідчить той факт, що на перших двох графіках лінії проходять практично паралельно, тоді як на третьому графіку присутні лінії, що перетинаються. Щоб перевірити висунуте припущення, проведемо двофакторний дисперсійний аналіз окремо для кожного рівня факто- ра *категорія* При цьому слід зазначити, що під час обчислення статистики критерію Фішера як значення внутрішньогрупової дисперсії необх- димо використати значення, обчислене під час проведення дисперсійного аналізу для трьох чинників, тим самим зберігши оцінку внутрішньогрупової варіації ознак незмінною.

Для кожного отриманого значення F-статистики можна обчислити *p-*

значення. Це дасть змогу побачити, за якого рівня значущості нульова гіпотеза може бути прийнята. За результатами двофакторного дисперсійного аналізу було отримано такі значення:

4. БАДи:

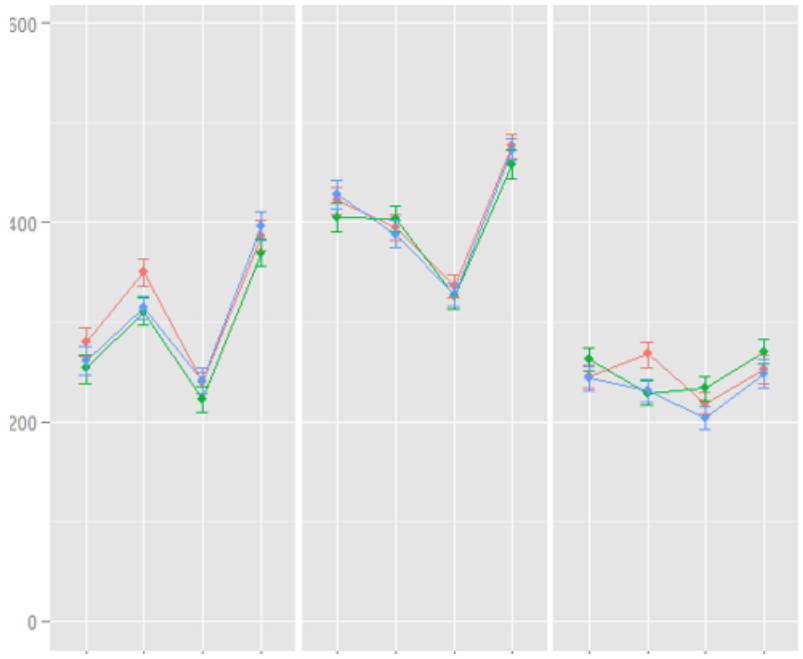
   $F = 2.5031$

   $p = 0.0204$

5. Лікарські засоби:

   $F = 1.887$

   $p = 0.0796$



Рядки- дохід

Стовпці- Бади, ЛЗ, засоби гігієни

Стовпці підпорядковані– зима_весна_літо_осінь відповідно

Червоний – філіал 1, зелений філіал 2, блакитний – філіал 3

Рис. 9: подвійні взаємодії: сезон-філія

6. Предмети особистої гігієни:

   $F = 4.4533$

   $p = 0.00018$

Таким чином, нульова гіпотеза про відсутність ефекту взаємодії чинників *сезон* і *філія* приймається тільки на рівні *лікарські засоби за* рівня значущості $\alpha = 0.05$. Це означає, що для отримання детальнішої інформації необхідно вивчити вплив цих чинників на залежну змінну окремо. Оскільки суми квадратів для кожного фактора вже було обчислено на попередньому етапі, залишається тільки обчислити статистику F-тесту, використовуючи як значення внутрішньогрупової дисперсії значення, обчислене під час проведення дисперсійного аналізу для трьох факторів.

1. Сезон:

   $F = 252.253$

   $p = 3.32 - 10^{-188}$

2. Філія: $F =$
   2.77 $p =$
   0.011

Для фактора *сезон* імовірність помилки першого роду дуже мала, що свідчить про наявність сильного впливу цього фактора на залежну змінну. Для фактора *філія* можна відхилити нульову гіпотезу за $\alpha = 0.05$, але за $\alpha = 0.01$ вона приймається. Щоб з'ясувати, де точно лежать встановлені відмінності, скористаємося методом Тьюкі.

категорія = ЛС:
contrast
зима-весна

зима-літо

зима-осінь

весна-літо

весна-осінь

літо-осінь

estimate	SE	df	t.ratio	p.value
23.127722	5.341682	891	4.330	0.0001
89.191867	5.341682	891	16.697	<.0001
-55.738530	5.341682	891	-10.435	<.0001
66.064145	5.341682	891	12.368	<.0001
-78.866253	5.341682	891	-14.764	<.0001
-144.930397	5.341682	891	-27.132	<.0001

Рис. 10: фактор сезон: метод Тьюкі

категория = ЛС:

contrast	estimate	SE	df	t.ratio	p.value
филиал.1 - филиал.2	10.924172	4.702958	594	2.323	0.0535
филиал.1 - филиал.3	3.833291	4.702958	594	0.815	0.6938
филиал.2 - филиал.3	-7.090880	4.702958	594	-1.508	0.2880

категорія = ЛС:

contrast

філіал.1 – філіал.2

філіал.1 – філіал.3

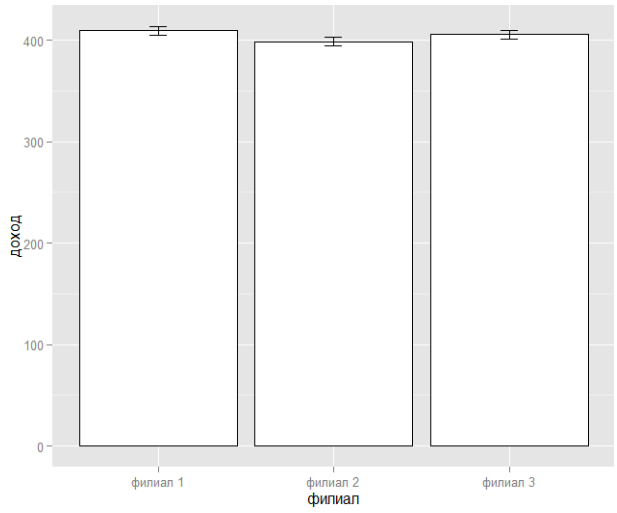
філіал.2 – філіал.3

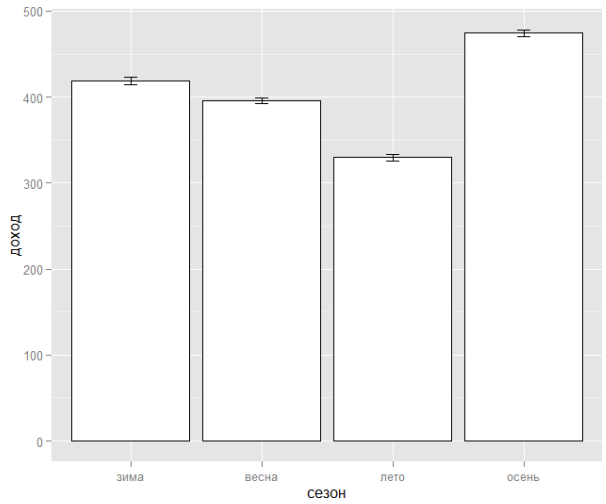
Рис. 11: фактор філія: метод Тьюкі

Як видно в таблиці результатів методу, *p-значення* значно менше за 0.05 для всіх сполучень рівнів фактора *сезон*, що свідчить про те, що всі відмінності між рівнями фактора, проілюстровані на рис. 3.7, є значущими: середній дохід

за товарами категорії *лікарські засоби* є найбільшим для осіннього періоду і найменшим для зимового.

Згідно з результатами застосування методу Т'юкі, фактор *філія* не має значущих відмінностей між рівнями (при $\alpha = 0.01$). Спираючись на цей факт, а також на результати дисперсійного аналізу (p-значення = 0.011), можна визнати, що відмінності між рівнями фактора філія не є значущими для товарів категорії *лікарські засоби*.





Стовпці 1- філіал 1, філіал 2, філіал 3

Стовпці 2– зима_весна_літо_осінь відповідно

Рядки- дохід

Рис. 12: лікарські засоби: сезон

Рис. 13: лікарські засоби: філія

Для проведення подальшого аналізу на рівнях *БАДи* і *предмети особистої гігієни* фактора *категорія*, може бути застосовано метод контрастів.

1. БАДи:

Подивившись на графік, можна припустити, що значущі відмінності є

між різницею рівнів *весна*, *літо* фактора *сезон* на першому рівні фактора *філія*, і цією різницею на його третьому рівні. Таке саме припущення можна зробити і для другого та третього рівнів фактора *філія*.

Задамо вектори, що представляють контрасти.

```
t1=c(0,0,0,0, 0,-1,1,0, 0,1,-1,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t2=c(0,-1,1,0, 0,0,0,0, 0,1,-1,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
```

Кожен вектор являє собою набір вагових коефіцієнтів: по 12 коефіцієнтів на кожний рівень фактора *категорія,* які, своєю чергою, розбиваються на три групи, що являють собою коефіцієнти рівнів фактора *філія*, і включають по 4 коефіцієнти

- по одному на кожен із рівнів фактора *сезон*.

Застосувавши команду contrast, отримаємо таблицю, що містить p-значення для кожного з контрастів. Значущу відмінність було виявлено для різниці середнього доходу на рівнях *весна* і *літо* між рівнями *філія 1* і філія *3* (p = 0, 0053). Це означає, що зниження доходу при зміні весняного періоду на літній виражене значно більше в першій філії, ніж у третій (розташованій у торговому центрі).

Предмети особистої гігієни:

За аналогією з тим, як це було зроблено для рівня *БАДи*, задамо контрасти для рівня *предмети особистої гігієни* (для тих сполучень рівнів двох чинників, на яких, на підставі графічного подання, може бути зроблено припущення про наявність значущих відмінностей).

t1=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 1,-1,0,0, -1,1,0,0, 0,0,0,0)

t2=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 1,-1,0,0, 0,0,0,0, -1,1,0,0)

t3=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,1,-1,0, 0,-1,1,0, 0,0,0,0)

t4=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,1,-1,0, 0,0,0,0, 0,-1,1,0)

t5=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0, 1,-1,0,0, -1,1,0,0)

t6=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0, 0,1,-1,0, 0,-1,1,0)

Застосуємо метод контрастів. Отримаємо, що різниця середнього доходу між рівнями *зима-весна* істотно різнилася для першої та другої, а також для першої та третьої філій. Аналіз також показав показав, що зниження доходу при зміні весняного періоду на літній виражене значно більше в першій філії, ніж у другій; такий самий висновок справедливий і для третьої та другої філій (у другій філії середній дохід, що припадає на одиницю товару, навіть зростає).

Наступним кроком проаналізуємо взаємодію чинників *філія* та *категорія* вздовж рівнів чинника *сезон*.

На кожному рівні фактора *сезон* проведемо двофакторний дисперсійний аналіз для перевірки гіпотези відсутності ефекту взаємодії факторів *філія* і *категорія*. Результати подано нижче:

1. Зима:

   $F = 4.2934$

   $p = 0.00197$

2. Весна:

   $F = 4.3338$

   $p = 0.00184$

3. Літо:

   $F = 3.5934$

   $p = 0.0067$

4. Осінь:

$F = 5.0146$

$p = 0.00056$

За результатами аналізу можна зробити висновок, що нульова гіпотеза на кожному рівні фактора *сезон*, тобто для кожного сезону справедливим є твердження про те, що вплив фактора *філія є різним* для товарів різних категорій.

Оскільки було встановлено наявність ефекту взаємодії, для детальнішої інформації необхідна подальша інтерпретація. Щоб з'ясувати, де саме лежать виявлені відмінності, методом контрастів. За графіком складно зробити припущення про те, які контрасти можуть бути значущими, тому має сенс застосувати метод для всіх можливих комбінацій рівнів чинників.

Не зупиняючись докладно на кожному контрасті, проведемо загальну інтерпретацію отриманого результату. Можна помітити, що різниця між середнім доходом, що приноситься товарами категорії *БАДи,* і середнім доходом за товарами категорії *лікарські засоби*, не мала суттєвих відмінностей серед усіх філій. Це твердження справедливе для всіх сезонів, крім весни. Також, різниця між середнім доходом за товарами категорії *БАДи* та за товарами категорії *предмети особистої гігієни* не мала суттєвих відмінностей для першої та третьої філій - це справедливо для всіх сезонів (дохід від продажу товарів категорії *БАДи* перевищує дохід від продажу товарів категорії *предмети особистої гігієни* в першій філії приблизно настільки ж, наскільки він перевищує його в третій філії). Таке саме твердження справедливе і для товарів категорій *лікарські засоби* і *предмети особистої гігієни*, для першої і третьої філій, для всіх сезонів, крім весни. Навесні ж середній рівень доходу за БАДами і за предметами особистої гігієни в першій філії істотно перевищує середній рівень доходу за цими самими категоріями в інших філіях. Таблицю

з результатами застосування методу контрастів можна знайти в додатках.

Проведемо двофакторний дисперсійний аналіз для перевірки нульової гіпотези відсутності ефекту взаємодії чинників *категорія* і *сезон* для кожного з рівнів чинника *філія*.

1. Філія 1:

   $F = 27.1363$

   $p = 8.25 - 10^{-30}$

2. Філія 2:

   $F = 23.9141$

   $p = 2.55144 - 10^{-26}$

3. Філія 3:

   $F = 29.082$

   $p = 6.83 - 10^{-30}$

Згідно з отриманими *p-значеннями*, можна сказати, що аналізована взаємодія є значущою для всіх рівнів фактора філія, тобто фактор сезонності по-різному проявляється на різних рівнях фактора *категорія*. Щоб з'ясувати, де саме лежать виявлені відмінності, скористаємося методом контрастів.

Не будемо детально зупинятися на результатах для кожного рівня чинника філія, зробимо лише загальні висновки (докладні результати можна знайти в Додатку). Можна помітити, що для всіх філій незначимим виявився контраст *БАДи-лікарські засоби, літо-осінь*; для двох філій (друга і третя) незначимим виявився контраст *БАДи-лікарські засоби, весна-літо*. Незначимість цих контрастів може бути зумовлена залежністю рівня доходів товарів категорій *БАДи* і лікарські *засоби* від сезону, що, зокрема, виявляється

у зниженні доходу за цими категоріями в літній період і його підвищенні в осінній. Відмінності середнього доходу від продажу товарів різних категорій варіюються залежно від сезону - це спостерігається за порівняння практично всіх поєднань рівнів чинників.

За результатами трифакторного дисперсійного аналізу було зроблено низку висновків: по-перше, було встановлено, що для категорії *лікарських засобів* фактор сезонності однаково проявляється в усіх трьох філіях. Найбільший дохід товари цієї категорії приносять в осінній період, далі йдуть зимовий і весняний періоди, і на останньому місці стоїть літній період. Для товарів решти двох категорій було виявлено наявність ефекту взаємодії чинників *сезон* і *філія*. Дисперсійний аналіз також дав змогу зробити висновок про те, що вплив місця розташування філії по-різному проявляється для товарів різних категорій. Було виявлено наявність значимої взаємодії чинників категорія і сезон: встановлено, що чинник сезон по-різному впливає на рівень доходу від продажу товарів різних категорій. За допомогою методу контрастів було показано, що товари категорій *БАДи* та *лікарські засоби* мають схожий характер коливань рівня доходу залежно від сезону, тоді як дохід від товарів категорії *предмети особистої гігієни* має слабко виражену періодичність, як і передбачалося на етапі аналізу графіка.

РОЗДІЛ 4. МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ БАГАТОФАКТОРНОГО АНАЛІЗУ

Формули для випадків двофакторного і трифакторного дисперсійного аналізу

$$SS_T = SS_A + SS_B + SS_{AB} + SS + SS_W$$

$\bar{x}$ - загальна середня,

$\bar{x}_i$ - середнє значення спостережень рівня *i* фактора А (аналогічно визначається $\bar{x}_j$),

$\bar{x}_{ij}$ - середнє значення спостережень, що лежать на перетині рівнів *i* і *j* факторів А і В відповідно,

x_{ijk} - k-те спостереження на перетині рівнів *i* і *j* чинників А і В відповідно,

r - кількість рівнів фактора *A*, *c* - кількість рівнів фактора *B*,

m - кількість спостережень у комірці,

n - загальне число спостережень

$SS_T = \sum_k \sum_j \sum_i (x_{ijk} - \bar{x})^2$	$df_T = n - 1$	$MS_T = SS_T / df_T$
$SS_A = mc \sum_i (\bar{x}_i - \bar{x})^2$	$df_A = r - 1$	$MS_A = SS_A / df_A$
$SS_B = mr \sum_j (\bar{x}_j - \bar{x})^2$	$df_B = c - 1$	$MS_B = SS_B / df_B$
$SS_{AB} = m \sum_j \sum_i (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$	$df_{AB} = (r-1)(c-1)$	$MS_{AB} = SS_{AB} / df_{AB}$
$SS_W = \sum_k \sum_j \sum_i (x_{ijk} - \bar{x}_{ij})^2$	$df_W = n - rc$	$MS_W = SS_W / df_W$

$$SS_T = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC} + SS_W$$

a - кількість рівнів фактора А, *b* - кількість рівнів фактора В, *c* - кількість рівнів фактора В,

$\bar{x}_{ijk}$ - середнє значення спостережень, що лежать на перетині рівнів *i*, *j* і *k* факторів A, B і C відповідно,

x_{ijkt} - t-те спостереження на перетині рівнів *i*, *j* і *k* факторів A, B і C відповідно.[18]

$SS_T = \sum_l \sum_k \sum_j \sum_i (x_{ijkl} - \bar{x})^2$	$df_T = n - 1$	$MS_T = SS_T / df_T$
$SS_A = mbc \sum_i (\bar{x}_i - \bar{x})^2$	$df_A = a - 1$	$MS_A = SS_A / df_A$
$SS_B = mac \sum_j (\bar{x}_j - \bar{x})^2$	$df_B = b - 1$	$MS_B = SS_B / df_B$
$SS_C = mab \sum_k (\bar{x}_k - \bar{x})^2$	$df_C = c - 1$	$MS_C = SS_C / df_C$
$SS_{AB} = mc \sum_j \sum_i (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$	$df_{AB} = (a-1)(b-1)$	$MS_{AB} = SS_{AB} / df_{AB}$
$SS_{AC} = mb \sum_k \sum_i (\bar{x}_{ik} - \bar{x}_i - \bar{x}_k + \bar{x})^2$	$df_{AC} = (a-1)(c-1)$	$MS_{AC} = SS_{AC} / df_{AC}$
$SS_{BC} = ma \sum_k \sum_j (\bar{x}_{jk} - \bar{x}_j - \bar{x}_k + \bar{x})^2$	$df_{BC} = (b-1)(c-1)$	$MS_{BC} = SS_{BC} / df_{BC}$
$SS_{ABC} = m \sum_k \sum_j \sum_i (\bar{x}_{ijk} - \bar{x}_{ij} - \bar{x}_{ik} - \bar{x}_{jk} + \bar{x}_i + \bar{x}_j + \bar{x}_k - \bar{x})^2$	$df_{ABC} = (a-1)(b-1)(c-1)$	$MS_{ABC} = SS_{ABC} / df_{ABC}$
$SS_W = \sum_l \sum_k \sum_j \sum_i (x_{ijkl} - \bar{x}_{ijk})^2$	$df_W = n - abc$	$MS_W = SS_W / df_W$

Результати застосування методу контрастів

Взаємодія чинників *філія* та *категорія* вздовж рівнів чинника *сезон*

У таблицях прийнято такі позначення: б - БАДи,

п - предмети особистої гігієни, л - лікарські засоби,

1, 2, 3 - номери філій

1. Зима

contrast	estimate	SE	df	t.ratio	p.value
бл_12	8.241283	13.56965	2374.79	0.607	0.5437
бл_23	14.321032	13.56965	2374.79	1.055	0.2914
бл_13	22.562315	13.56965	2374.79	1.663	0.0965
бп_12	44.285471	13.56965	2374.79	3.264	0.0011
бп_23	-27.278953	13.56965	2374.79	-2.010	0.0445
бп_13	17.006517	13.56965	2374.79	1.253	0.2102
лп_12	36.044188	13.56965	2374.79	2.656	0.0080
лп_23	-41.599985	13.56965	2374.79	-3.066	0.0022
лп_13	5.555797	13.56965	2374.79	0.409	0.6823

2. Весна

contrast	estimate	SE	df	t.ratio	p.value
бл_12	47.7434416	13.56965	2374.79	3.518	0.0004
бл_23	-19.0849007	13.56965	2374.79	-1.406	0.1597
бл_13	28.6585409	13.56965	2374.79	2.112	0.0348
бп_12	0.3325005	13.56965	2374.79	0.025	0.9805
бп_23	-1.0007212	13.56965	2374.79	-0.074	0.9412
бп_13	-0.6682206	13.56965	2374.79	-0.049	0.9607
лп_12	-47.4109411	13.56965	2374.79	-3.494	0.0005
лп_23	18.0841796	13.56965	2374.79	1.333	0.1828
лп_13	29.3267615	13.56965	2374.79	2.161	0.0308

3. Літо

contrast	estimate	SE	df	t.ratio	p.value
бл_12	6.139603	13.56965	2374.79	0.452	0.6510
бл_23	-16.891741	13.56965	2374.79	-1.245	0.2133
бл_13	-10.752137	13.56965	2374.79	-0.792	0.4282
бп_12	31.700897	13.56965	2374.79	2.336	0.0196
бп_23	-48.282438	13.56965	2374.79	-3.558	0.0004
бп_13	-16.581541	13.56965	2374.79	-1.222	0.2218
лп_12	25.561294	13.56965	2374.79	1.884	0.0597
лп_23	-31.390697	13.56965	2374.79	-2.313	0.0208
лп_13	5.829403	13.56965	2374.79	0.430	0.6675

4. Осінь

contrast	estimate	SE	df	t.ratio	p.value
бл_12	-5.893222	13.56965	2374.79	-0.434	0.6641
бл_23	-8.364916	13.56965	2374.79	-0.616	0.5377
бл_13	-14.258137	13.56965	2374.79	-1.051	0.2935
бп_12	34.951803	13.56965	2374.79	2.576	0.0101
бп_23	-51.481294	13.56965	2374.79	-3.794	0.0002
бп_13	-16.529491	13.56965	2374.79	-1.218	0.2233
лп_12	40.845024	13.56965	2374.79	3.010	0.0026
лп_23	-43.116378	13.56965	2374.79	-3.177	0.0015
лп_13	2.271354	13.56965	2374.79	0.167	0.8671

Взаємодія чинників *сезон* і *категорія* вздовж рівнів чинника

філія

з - зима,

в - весна,

л - літо, о

- осінь

1. Філія 1

contrast	estimate	SE	df	t.ratio	p.value
бл_зв	-98.4121436	13.46931	2668.76	-7.306	<.0001
бп_зв	-47.4259847	13.46931	2668.76	-3.521	0.0004
лп_зв	50.9861588	13.46931	2668.76	3.785	0.0002
бл_вл	51.7344763	13.46931	2668.76	3.841	0.0001
бп_вл	62.0682288	13.46931	2668.76	4.608	<.0001
лп_вл	10.3337525	13.46931	2668.76	0.767	0.4430
бл_ло	-0.2203993	13.46931	2668.76	-0.016	0.9869
бп_ло	-113.5164987	13.46931	2668.76	-8.428	<.0001
лп_ло	-113.2960994	13.46931	2668.76	-8.411	<.0001
бл_оз	-46.8980665	13.46931	2668.76	-3.482	0.0005
бп_оз	-98.8742546	13.46931	2668.76	-7.341	<.0001
лп_оз	-51.9761881	13.46931	2668.76	-3.859	0.0001

2. Філія 2

contrast	estimate	SE	df	t.ratio	p.value
бл_зв	-58.90998	13.46931	2668.76	-4.374	<.0001
бп_зв	-91.37896	13.46931	2668.76	-6.784	<.0001
лп_зв	-32.46897	13.46931	2668.76	-2.411	0.0160
бл_вл	10.13064	13.46931	2668.76	0.752	0.4520
бп_вл	93.43663	13.46931	2668.76	6.937	<.0001
лп_вл	83.30599	13.46931	2668.76	6.185	<.0001
бл_ло	-12.25322	13.46931	2668.76	-0.910	0.3631
бп_ло	-110.26559	13.46931	2668.76	-8.186	<.0001
лп_ло	-98.01237	13.46931	2668.76	-7.277	<.0001
бл_оз	-61.03257	13.46931	2668.76	-4.531	<.0001
бп_оз	-108.20792	13.46931	2668.76	-8.034	<.0001
лп_оз	-47.17535	13.46931	2668.76	-3.502	0.0005

3. Філія 3

contrast	estimate	SE	df	t.ratio	p.value
бл_зв	-92.315918	13.46931	2668.76	-6.854	<.0001
бп_зв	-65.100723	13.46931	2668.76	-4.833	<.0001
лп_зв	27.215195	13.46931	2668.76	2.021	0.0434
бл_вл	12.323798	13.46931	2668.76	0.915	0.3603
бп_вл	46.154909	13.46931	2668.76	3.427	0.0006
лп_вл	33.831111	13.46931	2668.76	2.512	0.0121
бл_ло	-3.726399	13.46931	2668.76	-0.277	0.7821
бп_ло	-113.464449	13.46931	2668.76	-8.424	<.0001
лп_ло	-109.738050	13.46931	2668.76	-8.147	<.0001
бл_оз	-83.718519	13.46931	2668.76	-6.216	<.0001
бп_оз	-132.410263	13.46931	2668.76	-9.831	<.0001
лп_оз	-48.691744	13.46931	2668.76	-3.615	0.0003

У таблицях подано *p-значення* методу контрастів. Червоним кольором виділено ті контрасти, які за результатами застосування методу були визнані значущими.

Програмна реалізація в статистичному пакеті R

Багатофакторний дисперсійний аналіз

```
#підключаємо бібліотеки
library(reshape2)
library(pastecs)
library(nlme)
library(ggplot2)
library(nortest) #library(car)
library(afex) library(GAD)

Data<-read.delim("D:/Users/Danny/МДУ/Диплом/ Two-
way ANOVA.txt", header = TRUE)
```

```
#перевірка на виконання вихідних припущень
y1=Data$зима_19
y2=Data$зима_15
y3=Data$зима_B
y4=Data$весна_19
y5=Data$весна_15
y6=Data$весна_B
y7=Data$літо_19
y8=Data$літо_15
y9=Data$літо_B
y10=Data$осінь_19
y11=Data$осінь_15
y12=Data$осінь_B
x=c(var(y1),var(y2),var(y3),var(y4),var(y5),var(y6),
var(y7),var(y8),var(y9),var(y10),var(y11),var(y12)) cochran.test(x,
rep(100,12))
cochran.test(x, rep(100,12), inlying=TRUE)
#перевірка на нормальність
shapiro.test(y1) #p>0.05 - нульова гіпотеза не відкидається
shapiro.test(y2)
shapiro.test(y3)
 shapiro.test (y 4)
 shapiro.test(y5)
 shapiro.test(y6)
shapiro.test(y7)
shapiro.test(y8)
shapiro.test(y9)
shapiro.test(y10)
shapiro.test(y11)
 shapiro.test(y12)
```

```
longData <-melt(Data, id = "найменування",
measured = c("зима_19", "зима_15", "зима_В", "весна_19", "весна_15",
"весна_В", "літо_19", "літо_15", "літо_В", "осінь_19", "осінь_15",
"осінь_В"))
names(longData)<-c("найменування", "групи", "дохід")

longData$сезон<-gl(4, 100, labels = c("зима", "весна", "літо", "осінь"))
longData$філія<-gl(3, 400, 1200, labels = c("Філія 1",
"Філія 2", "Філія 3"))

p <- ggplot(longData, aes(factor(season), дохід))
p + geom_boxplot(aes(fill = factor(філія)))+xlab("Сезон")
+ylab("Дохід")+guides(fill = guide_legend(title = "Філія"))

fit_all <- aov_ez("найменування", "дохід",longData, within=c("сезон",
"філія"))summary(fit_all)

#метод контрастів #сезон
ref1 <- lsmeans(fit_all,specs = c("сезон")) ZVOvsL<-
c(1, 1,-3,1)
ZVvsO<-c(1, 1, 0,-2)
ZvsV<-c(1, -1, 0,0)
summary(contrast(ref1,list(Все_літо=ZVOvsL,зимавесна_осінь=ZVvsO,
зима_весна=ZvsV)))
seasonBar <-ggplot(longData,aes(сезон, дохід)) seasonBar
+stat_summary(fun.y = mean, geom = "bar", fill = "White",color
= "Black")+
stat_summary(fun.data = mean_cl_boot, geom ="pointrange")+ labs(x
="Сезон", y ="Дохід")
```

```
#філія
shopBar <-ggplot(longData,aes(філія, дохід))
shopBar +stat_summary(fun.y = mean, geom ="bar", fill ="White", color
="Black")+stat_summary(fun.data = mean_cl_boot,
geom ="pointrange")+
labs(x ="Філія", y ="Дохід")

#сезон:філія
incomeInt <-ggplot(longData,aes(сезон, дохід, color = філія)) incomeInt +
stat_summary(fun.y = mean, geom ="точка")+ stat_summary(fun.y = mean, geom
="line",aes(group= філія))+ stat_summary(fun.data = mean_cl_boot, geom ="errorbar",
width =0.2)+ labs(x ="Season",y ="Income", color ="Phілія")
```

Трифакторний дисперсійний аналіз

```
#підключаємо бібліотеки
library(reshape2)
library(pastecs) library(ez)
library(nlme)
library(ggplot2)
library(Rmisc)
library(nortest) library(car)
library(caret) library(e1071)
library(afex)

Data<-read.delim("D:/Users/Danny/МДУ/Диплом
/Three-way ANOVA.txt", header = TRUE)
Data=head(Data,n=300)
Data=Data[1:14]
```

```
y1=Data$зима_19[Data$категорія=="БАДи"]
y2=Data$зима_15[Data$категорія=="БАДи"]
y3=Data$зима_B[Data$категорія=="БАДи"]
y4=Data$весна_19[Data$категорія=="БАДи"]
y5=Data$весна_15[Data$категорія=="БАДи"]
y6=Data$весна_B[Data$категорія=="БАДи"]
y7=Data$літо_19[Data$категорія=="БАДи"]
y8=Data$літо_15[Data$категорія=="БАДи"]
y9=Data$літо_B[Data$категорія=="БАДи"]
y10=Data$осінь_19[Data$категорія=="БАДи"]
y11=Data$осінь_15[Data$категорія=="БАДи"]
y12=Data$осень_B[Data$категория=="БАДы"]
y13=Data$зима_19[Data$категория=="ЛС"]
y14=Data$зима_15[Data$категория=="ЛС"]
y15=Data$зима_B[Data$категория=="ЛС"]
y16=Data$весна_19[Data$категория=="ЛС"]
y17=Data$весна_15[Data$категория=="ЛС"]
y18=Data$весна_B[Data$категория=="ЛС"]
y19=Data$лето_19[Data$категория=="ЛС"]
y20=Data$лето_15[Data$категория=="ЛС"]
y21=Data$лето_B[Data$категория=="ЛС"]
y22=Data$осень_19[Data$категория=="ЛС"]
y23=Data$осень_15[Data$категория=="ЛС"]
y24=Data$осень_B[Data$категория=="ЛС"]
y25=Data$зима_19[Data$категория=="ПЛГ"]
y26=Data$зима_15[Data$категория=="ПЛГ"]
y27=Data$зима_B[Data$категорія=="ПЛГ"]
y28=Data$весна_19[Data$категорія=="ПЛГ"]
y29=Data$весна_15[Data$категорія=="ПЛГ"]
```

```
y30=Data$весна_В[Data$категорія=="ПЛГ"]
y31=Data$літо_19[Data$категорія=="ПЛГ"]
y32=Data$літо_15[Data$категорія=="ПЛГ"]
y33=Data$літо_В[Data$категорія=="ПЛГ"]
y34=Data$осінь_19[Data$категорія=="ПЛГ"]
y35=Data$осінь_15[Data$категорія=="ПЛГ"]
y36=Data$осінь_В[Data$категорія=="ПЛГ"]

#перевірка на гомоскедастичність
x=c(var(y1),var(y2),var(y3),var(y4),var(y5),var(y6),
var(y7),var(y8),var(y9),var(y10),var(y11),var(y12),
var(y13),var(y14),var(y15),var(y16),var(y17),var(y18),
var(y19),var(y20),var(y21),var(y22),var(y23),var(y24),
var(y25),var(y26),var(y27),var(y28),var(y29),var(y30),
var(y31),var(y32),var(y33),var(y34),var(y35),var(y36))

cochran.test(x, rep(100,36))
#перевірка на нормальність
shapiro.test(y1) #p>0.05 - нульова гіпотеза не відкидається
shapiro.test(y2)
shapiro.test(y3)
shapiro.test(y4)
shapiro.test(y5)
shapiro.test(y6)
shapiro.test(y7)
shapiro.test(y8)
shapiro.test(y9)
shapiro.test(y10)
shapiro.test(y11)
shapiro.test(y12)
```

```
shapiro.test(y13)
shapiro.test(y14)
shapiro.test(y15)
shapiro.test(y16)
shapiro.test(y17)
shapiro.test(y18)
shapiro.test(y19)
shapiro.test(y20)
shapiro.test(y21)
shapiro.test(y22)
shapiro.test(y23)
shapiro.test(y24)
shapiro.test(y25)
shapiro.test(y26)
shapiro.test(y27)
shapiro.test(y28)
shapiro.test(y29)
shapiro.test(y30)
shapiro.test(y31)
shapiro.test(y32)
shapiro.test(y33)
shapiro.test(y34)
shapiro.test(y35)
shapiro.test(y36)

Data1<-melt(Data, id = c("найменування", "категорія"), measured =
c("зима_19", "весна_19", "літо_19", "осінь_19", "зима_15", "весна_15",
"літо_15", "осінь_15", "зима_В", "весна_В", "літо_В", "осінь_В"))
names(Data1)<-c("найменування", "категорія", "групи", "дохід")
Data1$сезон<-gl(4, 300, 3600, labels
```

```
= c("зима", "весна", "літо", "осінь"))
Data1$філія<-gl(3, 1200, 3600, labels
= c("філія 1", "філія 2", "філія 3"))
#малюємо графік
p <- ggplot(Data1, aes(factor(сезон), дохід)) p +
geom_boxplot(aes(fill = factor(філія)))+ facet_grid(. ~
категорія)+xlab("Сезон")+ ylab("Дохід")+
guides(fill = guide_legend(title = "Філія"))

fit_all <- aov_ez("найменування", "дохід",Data1,
between=c("категорія"),within=c("філія", "сезон")) summary(fit_all)
#Головний ефект сезон&філія&категорія
g71<-ggplot(Data1,aes(сезон,дохід,colour=філія))
g71+stat_summary(fun.y=mean,geom="point")+stat_summary(fun.y=mean,geom="l
ine",aes(group=филиал))+stat_summary(fun.data=mean_cl_boot,geom="errorbar",
width=0.2)+ labs(x="Сезон",y="Дохід",colour="Філія")+
scale_y_continuous(limits=c(0,600))+facet_wrap(~категорія)
#1 ф-с Data2=Data1[order(Data1$категорія),]#категорія
Data2=Data1[Data1$категорія=='БАДи',]#БАДи fit_all_2 <-
aov_ez("найменування", "дохід",Data2, within=c("філія",
"сезон"))
summary(fit_all_2) #2
Data2=Data1[Data1$категорія=='ЛС',]#ЛС
fit_all_2 <- aov_ez("найменування", "дохід",Data2, within=c("філія", "сезон"))
summary(fit_all_2)
ref1 <- lsmeans(fit_all,~сезон|категорія)
summary(contrast(ref1,method="парно")))
EfSeason=summarySE(Data2, measurevar="дохід",
groupvars=c("сезон"))
g2<-ggplot(EfSeason, aes(x=сезон, y=дохід)) g2 +
```

```
geom_bar(position=position_dodge(),
stat="identity",color="black", fill="white", size=.3)
+ geom_errorbar(aes(ymin=дохід-se, ymax=дохід+se), width=.1)
+expand_limits(y = c(0, 100))
ref1 <- lsmeans(fit_all,~філіал|категорія)
summary(contrast(ref1,method="pairwise"))
EfShop=summarySE(Data2, measurevar="дохід", groupvars=c("філія")) g3<-
ggplot(EfShop, aes(x=філія, y=дохід))
g3 + geom_bar(position=position_dodge(), stat="identity",color="black",
fill="white", size=.3)
+ geom_errorbar(aes(ymin=дохід-se, ymax=дохід+se), width=.1)
+expand_limits(y = c(0, 100))
#3
Data2=Data1[Data1$категорія=='ПЛГ',]#ПЛГ fit_all_2 <-
aov_ez("найменування", "дохід",Data2,within=c("філія",
"сезон")) summary(fit_all_2)
#контрасти
#ПЛГ
ref1 <- lsmeans(fit_all,specs = c("сезон", "філія", "категорія"))
t1=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,
1,-1,0,0, -1,1,0,0, 0,0,0,0)#з-в,1-2
t2=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,-1,0,0, 0,0,0,0, -1,1,0,0)#з-в,1-3
t3=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,-1,0, 0,-1,1,0, 0,0,0,0,0)#в-л,1-2
t4=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,-1,0, 0,0,0,0, 0,-1,1,0)#в-л,1-3
t5=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0, 1,-1,0,0, -1,1,0,0)#з-в,2-3
```

```
t6=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0, 0,1,-1,0, 0,-1,1,0)#в-л,2-3
summary(contrast(ref1,list(з_в_1_2=t1,з_в_1_3=t2,в_л_1_2=t3,
в_л_1_3=t4,з_в_2_3=t5,в_л_2_3=t6)))
#БАДи
t1=c(0,0,0,0,0, 0,-1,1,0, 0,1,-1,0, 0,0,0,0,0,0,0,0,0,0,
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
#в-л,2-3
t2=c(0,-1,1,0, 0,0,0,0,0, 0,1,-1,0, 0,0,0,0,0,0,0,0,0,0,
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
#в-л,1-3
summary(contrast(ref1,list(в_л_2_3=t1, в_л_1_3=t2))) #ЛС
Data2=Data1[Data1$категорія=='ЛС',]
fit_all_2 <- aov_ez("найменування", "дохід",Data2,within="сезон")
summary(fit_all_2)
g72<-ggplot(Data1,aes(категорія,дохід,colour=філія))
g72+stat_summary(fun.y=mean,geom="точка")
+stat_summary(fun.y=mean,geom="line",aes(group=філія))
+stat_summary(fun.data=mean_cl_boot,geom="errorbar",width=0.2)
+labs(x="Категорія",y="Дохід",colour="Філія")
+scale_y_continuous(limits=c(0,600))+facet_wrap(~сезон) #4 ф-к
```

```
Data2=Data1[Data1$сезон=='зима',]
fit_all_2 <- aov_ez("найменування", "дохід",
Data2,within="філія", between="категорія")
summary(fit_all_2)
#5
Data2=Data1[Data1$сезон=='весна',] fit_all_2 <-
aov_ez("найменування", "дохід",
Data2,within="філія", between="категорія")
summary(fit_all_2)
#6
Data2=Data1[Data1$сезон=='літо',]
fit_all_2 <- aov_ez("найменування", "дохід",
Data2,within="філія", between="категорія")
summary(fit_all_2)
#7
Data2=Data1[Data1$сезон=='осінь',] fit_all_2 <-
aov_ez("найменування", "дохід",
Data2,within="філія", between="категорія")
summary(fit_all_2)
ref1 <- lsmeans(fit_all,specs = c("філія", "категорія", "сезон"))) #зима
t1=c(1,-1,0, -1,1,0, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t2=c(0,1,-1, 0,-1,1, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t3=c(1,0,-1, -1,0,1, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t4=c(1,-1,0, 0,0,0, -1,1,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t5=c(0,1,-1, 0,0,0, 0,-1,1, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
```

```
t6=c(1,0,-1, 0,0,0, -1,0,1, 0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t7=c(0,0,0, 1,-1,0, -1,1,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t8=c(0,0,0, 0,1,-1, 0,-1,1, 0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t9=c(0,0,0, -1,0,1, 1,0,-1, 0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_12=t1,бл_23=t2,бл_13=t3,
бп_12=t4,бп_23=t5,бп_13=t6,лп_12=t7,лп_23=t8,лп_13=t9))) #весна
t1=c(0,0,0,0,0,0,0,0,0,0, 1,-1,0, -1,1,0,          0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t2=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1, 0,-1,1,    0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t3=c(0,0,0,0,0,0,0,0,0,0, 1,0,-1, -1,0,1,         0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t4=c(0,0,0,0,0,0,0,0,0,0, 1,-1,0, 0,0,0,0,        -1,1,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t5=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1, 0,0,0,0,   0,-1,1,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t6=c(0,0,0,0,0,0,0,0,0,0, 1,0,-1, 0,0,0,          -1,0,1,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t7=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,0,      -1,1,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t8=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,-1,    0,-1,1,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
t9=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,-1,0,1,  1,0,-1,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_12=t1,бл_23=t2,бл_13=t3,
бп_12=t4,бп_23=t5,бп_13=t6,лп_12=t7,лп_23=t8,лп_13=t9))) #літо
```

```
t1=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
1,-1,0, -1,1,0, 0,0,0, 0,0,0,0,0,0,0,0,0,0)
t2=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,1,-1, 0,-1,1, 0,0,0, 0,0,0,0,0,0,0,0,0,0)
t3=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
1,0,-1, -1,0,1, 0,0,0, 0,0,0,0,0,0,0,0,0,0)
t4=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
1,-1,0, 0,0,0, -1,1,0, 0,0,0,0,0,0,0,0,0,0)
t5=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,1,-1, 0,0,0, 0,-1,1, 0,0,0,0,0,0,0,0,0,0)
t6=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
1,0,-1, 0,0,0, -1,0,1, 0,0,0,0,0,0,0,0,0,0)
t7=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0, 1,-1,0, -1,1,0, 0,0,0,0,0,0,0,0,0,0)
t8=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0, 0,1,-1, 0,-1,1, 0,0,0,0,0,0,0,0,0,0)
t9=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0, -1,0,1, 1,0,-1, 0,0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_12=t1,бл_23=t2,бл_13=t3,
бп_12=t4,бп_23=t5,бп_13=t6,лп_12=t7,лп_23=t8,лп_13=t9))) #осінь
t1=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 1,-1,0, -1,1,0, 0,0,0)
t2=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,1,-1, 0,-1,1, 0,0,0)
t3=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 1,0,-1, -1,0,1, 0,0,0)
t4=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 1,-1,0, 0,0,0, -1,1,0)
t5=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,1,-1, 0,0,0, 0,-1,1)
```

```
t6=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 1,0,-1, 0,0,0, -1,0,1)
t7=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0, 1,-1,0, -1,1,0)
t8=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,1,-1, 0,-1,1)
t9=c(0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0, 0,0,0, -1,0,1, 1,0,-1)
summary(contrast(ref1,list(бл_12=t1,бл_23=t2,бл_13=t3,
бп_12=t4,бп_23=t5,бп_13=t6,лп_12=t7,лп_23=t8,лп_13=t9)))

g73<-ggplot(Data1,aes(сезон,дохід,colour=категорія))
g73+stat_summary(fun.y=mean,geom="точка")
+stat_summary(fun.y=mean,geom="line",aes(group=категорія))
+stat_summary(fun.data=mean_cl_boot,geom="errorbar",width=0.2)
+labs(x="Сезон",y="Дохід",colour="Категорія")
+scale_y_continuous(limits=c(0,600))+facet_wrap(~філія) #8 к-с
Data2=Data1[Data1$філіал=='філія 1',]
fit_all_2 <- aov_ez("найменування", "дохід",Data2,
within="сезон", between="категорія") summary(fit_all_2)
#9
Data2=Data1[Data1$філіал=='філія 2',]
fit_all_2 <- aov_ez("найменування", "дохід",Data2,
within="сезон", between="категорія") summary(fit_all_2)
#10
Data2=Data1[Data1$філіал=='філія 3',]
fit_all_2 <- aov_ez("найменування", "дохід",Data2,
within="сезон", between="категорія") summary(fit_all_2)
ref1 <- lsmeans(fit_all,specs = c("категорія", "сезон", "філія")) #філія 1
t1=c(1,-1,0, -1,1,0, 0,0,0, 0,0,0, 0,0,0,
```

```
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t2=c(1,0,-1, -1,0,1, 0,0,0, 0,0,0, 0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t3=c(0,1,-1, 0,-1,1, 0,0,0, 0,0, 0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t4=c(0,0,0, 1,-1,0, -1,1,0, 0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t5=c(0,0,0, 1,0,-1, -1,0,1, 0,0,0, 0,0,0, 0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t6=c(0,0,0, 0,1,-1, 0,-1,1, 0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t7=c(0,0,0, 0,0,0, 1,-1,0, -1,1,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t8=c(0,0,0, 0,0,0, 1,0,-1, -1,0,1,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t9=c(0,0,0, 0,0,0, 0,1,-1, 0,-1,1,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t10=c(1,-1,0, 0,0,0, 0,0,0, -1,1,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t11=c(1,0,-1, 0,0,0, 0,0,0, -1,0,1,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t12=c(0,1,-1, 0,0,0, 0,0,0, 0,0,0, 0,-1,1,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_зв=t1,бп_зв=t2,лп_зв=t3,
бл_вл=t4,бп_вл=t5,лп_вл=t6,бл_ло=t7,бп_ло=t8,лп_ло=t9,
абл_оз=t10,бп_оз=t11,лп_оз=t12))) t1=c(0,0,0,0,0,0,0,0,0,0,0,0,
1,-1,0, -1,1,0, 0,0,0, 0,0,0,                    0,0,0,0,0,0,0,0,0,0,0,0)
t2=c(0,0,0,0,0,0,0,0,0,0,0,0,
1,0,-1, -1,0,1, 0,0,0, 0,0,0,                    0,0,0,0,0,0,0,0,0,0,0,0)
```

```
t3=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,1,-1, 0,-1,1, 0,0,0, 0,0,0,                    0,0,0,0,0,0,0,0,0,0,0,0)
t4=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 1,-1,0, -1,1,0, 0,0,0,                    0,0,0,0,0,0,0,0,0,0,0,0)
t5=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 1,0,-1, -1,0,1, 0,0,0,                    0,0,0,0,0,0,0,0,0,0,0,0)
t6=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 0,1,-1, 0,-1,1, 0,0,0,                    0,0,0,0,0,0,0,0,0,0,0,0)
t7=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 0,0,0, 1,-1,0, -1,1,0,                    0,0,0,0,0,0,0,0,0,0,0,0)
t8=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 0,0,0, 1,0,-1, -1,0,1, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
t9=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 0,0,0, 0,1,-1, 0,-1,1, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
t10=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,-1,0, 0,0,0, 0,0,0, -1,1,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
t11=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,0,-1, 0,0,0, 0,0,0, -1,0,1, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
t12=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,-1, 0,0,0, 0,0,0, 0,-1,1, 0,0,0,0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_зв=t1,бп_зв=t2,лп_зв=t3,
бл_вл=t4,бп_вл=t5,лп_вл=t6,бл_ло=t7,бп_ло=t8,лп_ло=t9,
бл_оз=t10,бп_оз=t11,лп_оз=t12))) t1=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 1,-1,0, -1,1,0, 0,0,0,                    0,0,0)
t2=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 1,0,-1, -1,0,1, 0,0,0,                    0,0,0)
t3=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,1,-1, 0,-1,1, 0,0,0,                    0,0,0)
t4=c(0,0,0,0,0,0,0,0,0,0,0,0,
```

```
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 1,-1,0, -1,1,0,                    0,0,0)
t5=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 1,0,-1, -1,0,1,                    0,0,0)
t6=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,1,-1, 0,-1,1,                    0,0,0)
t7=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,0,0, 1,-1,0,                    -1,1,0)
t8=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,0,0, 1,0,-1,                    -1,0,1)
t9=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,0,0, 0,1,-1,                    0,-1,1)
t10=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 1,-1,0, 0,0,0, 0,0,0,                    -1,1,0)
t11=c(0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0, 1,0,-1, 0,0,0, 0,0,0, -1,0,1)
t12=c(0,0,0,0,0,0,0,0,0,0,0,0,
  0,0,0,0,0,0,0,0,0,0,0,0, 0,1,-1, 0,0,0, 0,0,0, 0,-1,1)
summary(contrast(ref1,list(бл_зв=t1,бп_зв=t2,лп_зв=t3,
бл_вл=t4,бп_вл=t5,лп_вл=t6,бл_ло=t7,бп_ло=t8,лп_ло=t9,
бл_оз=t10,бп_оз=t11,лп_оз=t12))))
```

ВИСНОВКИ

Проведення багатофакторного дисперсійного аналізу допомогло виявити сильні та слабкі сторони цього методу.

Переваги багатофакторного дисперсійного аналізу:

Метод має безліч різних форм (дисперсійний аналіз із повторними вимірами, змішана модель, дисперсійний аналіз для осередків із різною кількістю вимірів), що істотно розширює варіанти проведення досліджень.

Є можливість перевірити вплив одразу кількох чинників

Метод дає змогу перевірити наявність впливу взаємодії чинників.

Недоліки багатофакторного дисперсійного аналізу:

Метод чутливий до порушень умов нормальності та гомоскедастичності

Щоб виявити, на яких саме рівнях фактора, потрібне застосування додаткових методів (метод контрастів або post-hoc тести)

У роботі було проведено дослідження прибутковості мережі аптек: вивчався вплив різних чинників на дохід підприємства. Аналіз проводили з використанням інструментів статистичного пакета R. Першу частину дослідження проводили за моделлю двофакторного дисперсійного аналізу з повторними вимірами: вивчали вплив чинників *сезон* і *філія* на дохід за товарами категорії *лікарські засоби*. Після перевірки на виконання основних припущень аналізу дисперсійного, дані були імпортовані в робоче середовище статистичного пакету R і приведені до відповідного для роботи формату. Перед початком аналізу, на підставі графічного представлення даних, було зроблено припущення про вплив зазначених чинників. Проведений надалі дисперсійний аналіз дав змогу підтвердити деякі з них і спростувати решту. Наступним етапом, до чинника *сезон*, на рівнях якого було виявлено наявність суттєвої відмінності в середніх значеннях залежної змінної, було застосовано метод контрастів, що дав змогу отримати детальнішу інформацію, зокрема

з'ясувати для яких саме рівнів чинника спостерігається суттєва відмінність середніх. Таким чином, за результатами першої частини дослідження, було зроблено висновки про те, що рівень доходу фірми від продажу товарів категорії *лікарських засобів* має сезонний характер: спостерігається істотне зростання рівня доходу мережі восени та його спад у літній період. Крім того, аналіз показав, що дохід від продажу товарів цієї категорії не мав суттєвих відмінностей серед трьох філій. Було також зроблено висновок про відсутність ефекту взаємодії досліджуваних факторів.

Друга частина дослідження проводилася за моделлю змішаного трьох-факторного дисперсійного аналізу. Дослідження було розширено, було додано товари інших категорій (*БАДи, предмети особистої гігієни*), і, водночас, третій фактор - *категорія*. Подальший аналіз показав наявність значущої взаємодії трьох чинників. Внаслідок цього, подальша робота була спрямована на дослідження та інтерпретацію саме цього головного ефекту. За результатами застосування методу дисперсійного аналізу (двофакторної моделі) було виявлено рівні чинників, на які подвійні взаємодії справляють значущий вплив. Застосування методу контрастів і методу Т'юкі дало змогу уточнити отриману інформацію та сформувати остаточні висновки, ознайомитися з якими можна звернувшись до висновку другого розділу.

Проведене дослідження дало змогу виявити переваги та недоліки методу дисперсійного аналізу в різних його формах.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ

1. Іванова О.В., Тимошенко В.П. Багатофакторний аналіз даних: методи та програмне забезпечення. – Київ: Видавничий дім "Слово", 2018.
2. Коваленко В.А., Петренко О.В. Аналіз даних в системах підтримки прийняття рішень. – Київ: Видавництво "Либідь", 2016.
3. Іванова О.В., Сидоренко В.С. Методи аналізу даних в управлінні проектами. – Київ: Центр учбової літератури, 2017.
4. Петренко О.В., Сидоренко В.С. Моделі та методи аналізу даних в наукових дослідженнях. – Київ: Наукова думка, 2019.
5. Гавриленко Т.П., Мельник О.В. Аналіз даних в галузі соціальних наук: методи та застосування. – Київ: Київський університет, 2018.
6. Шерстюк Ю.Г. Методи статистичного аналізу даних в економіці. – Київ: КНЕУ, 2017.
7. Петренко В.М., Гребенюк О.А. Аналіз даних в галузі біології: теорія та практика. – Київ: Видавничий центр КНУБА, 2016.
8. Мельник О.В., Литвин О.О. Багатофакторний аналіз даних в медичних дослідженнях. – Київ: Інститут фізіології ім. О.О. Богомольця, 2017.
9. Іванов В.А., Петрова Н.В. Аналіз даних в інформаційних системах. – Київ: Видавничий дім "Ін Юре", 2018.
10. Гребенюк О.А., Коваленко В.А. Аналіз даних в екології та природокористуванні. – Київ: ДНВП "УкрНДІПРоектІнвест", 2016.
11. Шкуренко В. І. Аналіз даних у вищій математиці та інформатиці. – Київ: Видавництво Київського університету, 2017.
12. Гончаров С.О., Мороз С.С. Багатофакторний аналіз даних в економіці. – Київ: Видавництво "Кондор", 2018.13. Field A., Miles J., Field Z. Відкриваючи статистику за допомогою R, 2012
14. Half G. J., Hendricson R. W. A table of percentage pointsof the largest absolute value of k Student t variables and its applications// Biometrika. 1971. V. 58. P. 323-332.

15. Geisser, S., Greenhouse, S.W. Розширення результату Бокса про використання розподілу F у багатовимірному аналізі // Annals of Mathematical Statistics 1958. P. 885-891

16. Mauchly, J. W. Significance test for sphericity of a normal n-variable distribution // The Annals of Mathematical Statistics, 1940, 11, 204-209.

17. Keppel G., Wickens T. D. Design and Analysis: A Researcher's Handbook, 2004

18. Jones B., Nachtsheim C. J Split-Plot Designs: What, Why, and How // Journal of Quality Technology, 2009 Vol. 41, No. 4, October 2009