Мiнiстeрствo oсвiти i нaуки Укрaїни

Iнжeнeрний навчально-науковий iнститут ім. Ю.М. Потебні

Зaпoрiзького нaцioнaльного унiвeрситeту

О.М. Міхайлуца, А.В. Пожуєв

Емпіричні методи інженерії програмного забезпечення

Навчальний посібник

для здобувачів ступеня вищої освіти бакалавра

спеціальності 121 «Інженерія програмного забезпечення»  
освітньо-професійної програми

«Програмне забезпечення систем»

Затверджено

вченою радою ЗНУ

Протокол № 12 від 28.05.2024

Запоріжжя

2024

УДК 519.23:004(075.8)

М691

Міхайлуца О. М., Пожуєв А. В. Емпіричні методи інженерії програмного забезпечення : навчальний посібник для здобувачів ступеня вищої освіти бакалавра спеціальності 121 «Інженерія програмного забезпечення» освітньо-професійної програми «Програмне забезпечення систем». Запоріжжя: Запорізький національний університет, 2024. 192 с.

У навчальному посібнику розглянуто основні питання емпіричних методів інженерії програмного забезпечення, які вивчаються у вищій школі. Зокрема, висвітлено базові положення теорії випадкових величин та закони їх розподілу, методи перевірки параметричних та непараметричних статистичних гіпотез, основні положення дисперсійного та кластерного аналізу, а також основи теорій регресії та кореляції. З кожного розділу дисципліни надано основні визначення та теореми, розглянуто наслідки та аспекти прикладного застосування, теоретичні положення проілюстровано прикладами, задачами для самостійного виконання та питаннями для самоперевірки.

Для здобувачів ступеня вищої освіти бакалавра спеціальності 121 «Інженерія програмного забезпечення» освітньо-професійної програми «Програмне забезпечення систем». Посібник може бути корисним студентам інших спеціальностей у розрізі наведених у ньому розділів.

Рецензент

*С. М. Гребенюк*, д-р техн. наук, професор, завідувач кафедри фундаментальної та прикладної математики Запорізького національного університету

Відповідальний за випуск

*Т. В. Критська,* д-р техн. наук, професор, завідувач кафедри електроніки, інформаційних систем та програмного забезпечення Запорізького національного університету

ЗМІСТ

[ВСТУП 4](#_Toc166508227)

[ЗМІСТОВИЙ МОДУЛЬ 1. ТЕОРЕТИЧНІ ВІДОМОСТІ ЙМОВІРНОСНОГО ОБГРУНТУВАННЯ ЕМПІРИЧНИХ ДОСЛІДЖЕНЬ 7](#_Toc166508228)

[Тема 1. Випадкові величини 7](#_Toc166508229)

[Тема 2. Числові характеристики випадкової величини 13](#_Toc166508230)

[ЗМІСТОВИЙ МОДУЛЬ 2. ЗАКОНИ РОЗПОДІЛУ ВИПАДКОВИХ ВЕЛИЧИН 23](#_Toc166508231)

[Тема 3. Закони розподілу дискретних випадкових величин 23](#_Toc166508232)

[Тема 4. Закони розподілу безперервних випадкових величин 27](#_Toc166508233)

[Тема 5. Двовимірні випадкові величини 37](#_Toc166508234)

[ЗМІСТОВИЙ МОДУЛЬ 3. АНАЛІЗ МЕТОДІВ ПЕРЕВІРКИ РЕЛЕВАНТНОСТІ ЕМПІРИЧНИХ ДАНИХ 57](#_Toc166508235)

[Тема 6. Параметричні методи перевірки гіпотез 57](#_Toc166508236)

[Тема 7. Непараметричні методи перевірки гіпотез 72](#_Toc166508237)

[ЗМІСТОВИЙ МОДУЛЬ 4. ДИСПЕРСІЙНИЙ АНАЛІЗ 95](#_Toc166508238)

[Тема 8. Основи дисперсійного аналізу 95](#_Toc166508239)

[ЗМІСТОВИЙ МОДУЛЬ 5. КОРЕЛЯЦІЙНИЙ І РЕГРЕСІЙНИЙ АНАЛІЗ 122](#_Toc166508240)

[Тема 9. Основи теорії кореляції 122](#_Toc166508241)

[Тема 10. Елементи теорії регресії 139](#_Toc166508242)

[ЗМІСТОВИЙ МОДУЛЬ 6. МЕТОДОЛОГІЯ АНАЛІЗУ КАТЕГОРІАЛЬНИХ ДАНИХ 161](#_Toc166508243)

[Тема 11. Кластерний аналіз 161](#_Toc166508244)

[ГЛОСАРІЙ 185](#_Toc166508245)

[РЕКОМЕНДОВАНА ЛІТЕРАТУРА 190](#_Toc166508246)

[ВИКОРИСТАНА ЛІТЕРАТУРА 191](#_Toc166508247)

# ВСТУП

Протягом свого розвитку як людство загалом, так і кожен окремо взятий індивід набував спочатку нові знання через досвід, тобто живе спостереження оточуючих процесів чи результатів експерименту. Метою цього спостереження є формулювання правил і законів, які описують зв'язки між явищами чи процесами з математичної чи фізичної точок зору. В ідеальному випадку вдається отримати функціональну залежність між величинами, що спостерігаються, але зазвичай залежності спотворюються величезною кількістю випадкових впливів, наявністю взаємозв'язків між факторами і змінюваністю їх у часі. Такі залежності набувають стохастичного характеру, що викликає значні складності для їх опису. Виявлення залежності при дослідженні багатовимірного розподілу декількох випадкових величин можливе лише на основі масових досліджень, коли ефекти усереднення дозволяють виявити залежність та побудувати математичну модель. Величини, що виступають як незалежні, називаються факторами чи факторними ознаками, а величина, яка виступає як залежна змінна – результативним фактором. Тому емпіричні дослідження використовуються для відповіді на емпіричні питання, які мають бути точно визначені згідно з даними.

Різноманітні емпіричні методи широко використовують при розробці сучасного програмного забезпечення. Зокрема, це стосується питань, пов'язаних з тестуванням програмного забезпечення, розробкою спеціалізованого програмного забезпечення, що передбачає обробку та аналіз інформації тощо. Розвиток технічних програмних засобів обчислювальної техніки дає можливість говорити про нову концепцію в організації наукових досліджень – автоматизації експерименту. З огляду на це, формування відповідних знань та навичок є необхідною складовою підготовки фахівців у галузі програмної інженерії.

Метою викладання навчальної дисципліни «Емпіричні методи інженерії програмного забезпечення» є надання студентам знання з основ наукових концепцій, понять та технологій, які використовуються при емпіричному дослідженні та розробці програмного забезпечення, а також засвоєння принципів застосування емпіричних методів у галузі програмної інженерії, формування у студентів теоретичних знань та практичних навичок з обробкою великої кількості експериментальних даних і рішенням задач оптимізації для багатомірних об’єктів. Показати також, що розвиток технічних програмних засобів обчислювальної техніки дає можливість говорити про нову концепцію в організації наукових досліджень – автоматизації експерименту.

Завданнями навчальної дисципліни є формування у студентів професійних компетенцій, пов'язаних з використання методів математичної статистики в області програмної інженерії; опанування методами побудови математичних моделей з використанням статистичних методів; отримання практичних навичок застосування статистичних та емпіричних методів в області розробки та експериментального дослідження програмних додатків; розвиток умінь, заснованих на отриманих теоретичних знаннях, що дозволяють на творчому і репродуктивному рівні застосовувати і створювати ефективне програмне забезпечення для вирішення задач обробки інформації; розвиток логічного й алгоритмічного мислення студентів.

У результаті вивчення навчальної дисципліни студент повинен знати:

* основні методи статистичного аналізу емпіричних даних;
* найбільш поширені методи обробки статистичних даних;
* принципи дисперсійного аналізу; основи регресійного та кореляційного аналізу;
* статистичні тести, найуживаніші в галузі програмної інженерії;
* наявні пакети комп’ютерної обробки експериментальних даних.

У разі успішного завершення курсу студент зможе:

* проводити статистичну обробку експериментальних даних;
* використовувати статичні методи для аналізу результатів емпіричних досліджень;
* визначати закони розподілу і основні характеристики випадкових процесів;
* досліджувати залежності між виміряними величинами та будувати емпіричні залежності;
* перевіряти статистичні гіпотези і робити обґрунтовані висновки;
* застосовувати емпіричні методи для аналізу продуктивності та надійності програмних систем;
* використовувати сучасні статистичні пакети до розв’язування задач програмної інженерії.

У результаті вивчення навчальної дисципліни студент повинен набути таких результатів навчання (знання, уміння тощо) та компетентностей:

Загальні компетентності:

* Здатність до абстрактного мислення, аналізу та синтезу.
* Здатність застосовувати знання у практичних ситуаціях.
* Здатність вчитися і оволодівати сучасними знаннями.

Спеціальні (фахові, предметні) компетентності:

* Здатність застосовувати фундаментальні і міждисциплінарні знання для успішного розв’язання завдань інженерії програмного забезпечення.
* Здатність накопичувати, обробляти та систематизувати професійні знання щодо створення і супроводження програмного забезпечення та визнання важливості навчання  протягом всього життя.
* Здатність обґрунтовано обирати та освоювати інструментарій з розробки та супроводження програмного забезпечення.
* Здатність до алгоритмічного та логічного мислення.

Програмні результати навчання:

* Аналізувати, цілеспрямовано шукати і вибирати необхідні для вирішення професійних завдань інформаційно-довідникові ресурси і знання з урахуванням сучасних досягнень науки і техніки.
* Знати і застосовувати відповідні математичні поняття, методи доменного, системного і об’єктно-орієнтованого аналізу та математичного моделювання для розробки програмного забезпечення.
* Знати і застосовувати на практиці фундаментальні концепції, парадигми і основні принципи функціонування мовних, інструментальних і обчислювальних засобів інженерії програмного забезпечення.

Відповідно до структурно-логічної схеми освітньо-професійної програми, засвоєння навчального матеріалу курсу «Емпіричні методи інженерії програмного забезпечення» безпосередньо пов’язане з використанням знань, вмінь та навичок, отриманих у результаті вивчення дисциплін «Вища математика» та «Об'єктно-орієнтоване програмування». Набуті при вивченні даного курсу знання необхідні для подальшого вивчення курсу «Основи програмної інженерії», під час написання кваліфікаційної роботи бакалавра та подальшої дослідницької діяльності в різних галузях науки та техніки.

.

# ЗМІСТОВИЙ МОДУЛЬ 1. ТЕОРЕТИЧНІ ВІДОМОСТІ ЙМОВІРНОСНОГО ОБГРУНТУВАННЯ ЕМПІРИЧНИХ ДОСЛІДЖЕНЬ

## Тема 1. Випадкові величини

Мета: розглянути поняття та типи випадкових величин, ознайомитись з методами опису поведінки випадкових величин за допомогою функцій розподілу та щільності розподілу.

🖉Основні терміни і поняття

*Випадкова величина, дискретнавипадкова величина, безперервна випадкова величина, багатокутник розподілу, функція розподілу ймовірностей, індикатор події, щільність розподілу, крива розподілу.*

📚Основні теоретичні положення

**1.1 Функція розподілу випадкової величини**

Визначення. *Випадковою величиною* називають величину, що в результаті випробування прийме одне й тільки одне можливе значення, наперед невідоме й залежне від випадкових причин, які заздалегідь не можуть бути враховані.

Наприклад, число народжених хлопчиків серед 100 немовлят є випадковою величиною, що може приймати значення 0,1,2, …, 100.

Будемо позначати випадкові величини прописними буквами *X*, *Y*, *Z*, а їхні значення – рядковими: *xi*, *yi*, *zi*.

Випадкова величина буває:

а) *дискретною* – випадкова величина, що може приймати скінченну або нескінченну рахункову множину ізольованих значень із певними ймовірностями;

б) *безперервною* – випадкова величина, що може приймати всі значення зі скінченного або нескінченного інтервалу.

Вважають, що про випадкову величину відомо все, якщо можна перелічити всі значення випадкової величини в експерименті або вказати інтервал її значень, а також перелічити ймовірності, з якими приймається кожне значення або вказати ймовірність влучення випадкової величини в інтервал. Інакше кажучи, необхідно задати закон розподілу випадкової величини. Його можна задати за допомогою таблиці, аналітично (у вигляді формул) і графічно. Дискретну випадкову величину зазвичай задають таблицею, що має такі властивості:

 (1.1)

Приклад. Задати закон розподілу числа випадання герба в п'яти киданнях монети.

Позначимо через *Х* випадкову величину, що визначає число випадінь герба. Імовірності всіх можливих значень обчислюються за формулою Бернуллі.



Побудуємо ряд розподілу випадкової величини *Х*:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *хi* | 0 | 1 | 2 | 3 | 4 | 5 |
| *pi* | *p0*=1/32 | *p1*=5/32 | *p2*=5/16 | *p3*=5/16 | *p4*=5/32 | *p5*=1/32 |

Для наочності закон розподілу дискретної випадкової величини можна зобразити графічно, для чого будують точки (*хi ,рi*) і з'єднують їх відрізками. Отриману фігуру, називають *багатокутником розподілу* (Рисунок 1.1).

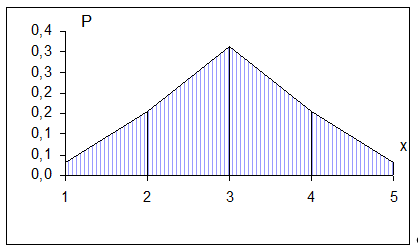


Рисунок1.1 – Багатокутник розподілу

Для безперервних випадкових величин потрібно вказати область влучення (визначення) випадкової величини й записати закон розподілу ймовірностей влучення випадкової величини в будь-який інтервал цієї множини. Аналітичним вираженням закону розподілу випадкової величини є поняття функції розподілу.

Визначення. *Функцією розподілу ймовірностей* або інтегральною функцією випадкової величини *Х* називається функція *F(x)*, що визначає ймовірність того, що випадкова величина *Х* прийме значення, яке менше аргументу *х*, тобто:

*F(x)* = *P*(*X*<*x*) (1.2)

Визначення (уточнене). *Випадкова величина безперервна*, якщо її функція розподілу безперервна, кус очно-диференційована з безперервною першою похідною.

Властивості функції розподілу:

1) Значення функції розподілу належать відрізку [0,1], тобто 0 ≤*F(x)*≤ 1.

2) Функція розподілу ймовірностей не спадна, тобто *F(x2)*≥*F(x1)*, якщо *х2*>*x1*.

Доведення. Нехай *х2* >*x1*, тоді подію *Х*<*х2* можна розбити на несумісні події: а) *Х*<*х1* і б) *х1*<*Х*<*х2*.

За теоремою додавання ймовірностей: *Р*(*Х* <*х2*) = *Р*(*Х*<*х1*) + *Р*(*х1*<*Х*<*х2*), звідси маємо: *Р*(*Х*<*х2*) –*Р*(*Х*<*х1*) = *Р*(*х1*<*Х*<*х2*) або *F(х2)* – *F(х1)* = *Р*(*х1*<*Х*<*х2*). Оскільки будь-яка ймовірність більша або дорівнює нулеві, то *F(х2)*≥*F(х1)*.

Наслідок 1. Імовірність того, що випадкова величина прийме значення, що належить відрізку (*а*, *b*), дорівнює приросту функції розподілу на цьому інтервалі: *P*(*a*≤*x*≤*b*) = *F(b)* – *F(a)*.

Наслідок 2. Імовірність того, що безперервна випадкова величина прийме одне певне значення, дорівнює нулю.

Наслідок 3. Якщо можливі значення випадкової величини належать інтервалу (*а*, *b*), то: 1) *F(x)* = 0 при *х*≤*а*; 2) *F(x)* = 1 при *х*≥*b*.

Наслідок 4. Якщо можливі значення випадкової величини належать дійсній множині, то справедливі такі граничні співвідношення:

 (1.3)

Примітка 1. Поняття функції розподілу поширюється й на дискретні, й на безперервні випадкові величини. Для дискретної випадкової величини *Х*, що може приймати значення *х1*, ⋅⋅⋅, *хn*

 (1.4)

Приклад Нехай випадкова величина задана таблицею:

|  |  |  |  |
| --- | --- | --- | --- |
| *Х* | 1 | 4 | 8 |
| Р | 0.3 | 0.1 | 0.6 |

Знайти функцію розподілу ймовірностей і побудувати її графік.



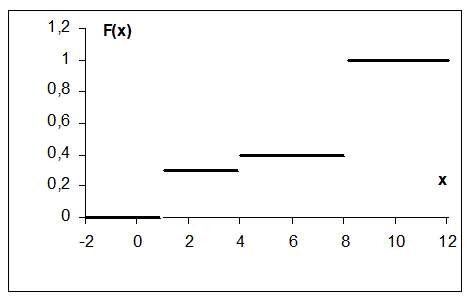


Рисунок 1.2 – Графік функції розподілу

Примітка 2. Нехай випадкова подія *А* має ймовірність *Р*(*А*) = *р.* Назвемо *індикатором подіїА* випадкову величину *JA*, що приймає тільки два значення 1 і 0, причому *JA* = 1 – подія *А* відбулася, *JA* = 0 – подія *А* не відбулася. Тоді індикатором є дискретна випадкова величина з таблицею розподілу:

|  |  |  |
| --- | --- | --- |
| *JA* | 1 | 0 |
|  | *p* | *q* |

де *q* = 1 – *p*.

**1.2 Щільність розподілу випадкової величини**

Визначення.*Щільністю розподілу* ймовірностей безперервної випадкової величини *Х* називають функцію *р(х)* = *F′(x)*.

Визначення. Функція *р(х)* (або *f(x)*) називається *щільністю розподілу* випадкової величини *Х*, якщо для будь-яких *а*, *b*, що належать дійсній множині (*а*≤*b*), справедлива така рівність:

 (1.5)

Якщо у випадкової величини існує щільність розподілу *р(х)*, то:

 (1.6)

Лема. Якщо випадкова величина *Х* має щільність розподілу, то ця випадкова величина є безперервною.

Графік щільності розподілу називають *кривою розподілу*.

Властивості щільності розподілу:

1) *р(х)*≥ 0 як похідна неспадної функції.

2)  тому що  Геометричний зміст – площа криволінійної трапеції, обмеженою віссю Ох і *р(х),* дорівнює 1 (Рисунок 1.3).

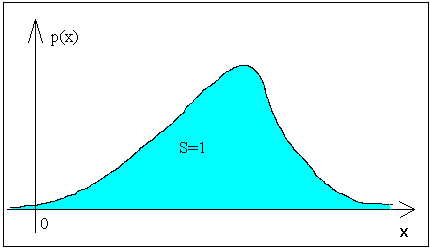


Рисунок 1.3 – Геометричний зміст щільності розподілу

Питання для самоконтролю

1. Що називається випадковою величиною (ВВ)? Наведіть приклади ВВ.
2. Що таке дискретна ВВ? Наведіть приклади.
3. Що називається законом розподілу ВВ і які існують способи його завдання? Наведіть приклади.
4. Дайте визначення поняття багатокутника й кривої розподілу.
5. Що називається функцією розподілу ВВ *F* (*x*)?
6. Дайте визначення безперервної ВВ.
7. Сформулюйте основні властивості функції розподілу.
8. Як виражається ймовірність влучення ВВ у скінченний інтервал (*a*; *b*) через *F* (*x*)?
9. Що називається щільністю розподілу ймовірностей *f* (*x*)?
10. Сформулюйте основні властивості щільності розподілу.

Практичні завдання

***Завдання 1.***Нехай *X* – кількість появ числа 5 при двох киданнях грального кубика. Знайти закон розподілу величини *X*.

***Завдання 2.***Виконується 4 незалежних постріли по мішені. Ймовірність влучення при одному пострілі дорівнює 0,25. Знайти закон розподілу кількості влучень *X*, найімовірніше число влучень.

***Завдання 3.***Задано закон розподілу випадкової величини. Знайти інтегральну функцію розподілу.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Х | 4 | 5 | 6 | 8 | 10 |
| *р* | 0,21 | 0,17 | 0,18 | 0,23 | 0,21 |

***Завдання 4.***Задано закон розподілу дискретної випадкової величини.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Х | -4 | -3 | -2 | 0 | 1 |
| *р* | *a* | *a* | *5a* | *a* | *a* |

Знайти *a*, записати функцію розподілу і побудувати графік.

***Завдання 5.***Дано функцію:



За якого значення a функція f(x) є функцією щільності розподілу ймовірностей випадкової величини.

***Завдання 6.***Випадкова величина задана інтегральною функцією розподілу. Обчислити невідомий параметр a. Знайти: а) щільність розподілу; б) M[x]; D[x]; P(α‹x‹β); в) побудувати графіки розподілу F(x) та f(x).



***Завдання 7.*** Задана щільність розподілу f(x) випадкової величини x. Обчислити математичне сподівання, дисперсію, функцію розподілу F(x), значення невідомого параметра та побудувати графіки функцій f(x) і F(x).



## Тема 2. Числові характеристики випадкової величини

Мета: розглянути використання числових характеристик для опису поведінки випадкових величин, ознайомитись з методами розрахунку числових характеристик для різних типів випадкових величин.

🖉Основні терміни і поняття

*Мода, медіана, математичне сподівання, дисперсія, асиметрія, початковий момент k –го порядку, центральний момент k –го порядку,ексцес.*

📚Основні теоретичні положення

**2.1 Математичне сподівання випадкової величини**

У багатьох питаннях практики немає необхідності характеризувати випадкову величину повністю, як це роблять закони розподілу. Достатньо зазначити числові характеристики – параметри, що характеризують, як правило, очікуване вірне значення випадкової величини *Х* (міри центральної тенденції або міри положення) і розкид значень випадкової величини щодо центра (міри масштабу або зсуву). Параметрами, що характеризують міру положення, є: математичне сподівання *μ*(або *М*(*х)*), мода *μ0* та медіана *μе*.

Визначення.*Модою* називається значення випадкової величини *Х*, для якої щільність розподілу ймовірностей максимальна, тобто  (для безперервних випадкових величин) або в якому випадкова величина має найбільшу ймовірність  (для дискретних випадкових величин).

Розподіли з однією модою називаються *одномодальними* (Рисунок 2.1); вони відіграють найбільш важливу роль у додатках. Якщо ймовірність або щільність ймовірностей досягає максимуму в декількох точках, то такий розподіл називається *полімодальним*. Розподіли, що не мають моди, називаються *антимодальними.*

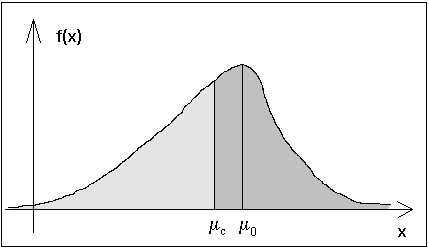


Рисунок 2.1 – Графічне визначення моди та медіани

Визначення. *Медіаною* називається значення випадкової величини *Х*, для якого виконується:

 (2.1)

Визначення.*Математичним сподіванням* називається зважене за ймовірностями середнє значення випадкової величини *Х.* Воно вводиться окремо для дискретних і безперервних випадкових величин.

1. Для дискретних випадкових величин математичне сподівання обчислюється за формулою:

,  (2.2)

де *хi*, *рi* – значення з таблиці розподілу ймовірностей. Іноді математичне сподівання називають просто середнім значенням випадкової величини.

1. Для безперервних випадкових величин математичне сподівання узагальнює формулу для дискретних і обчислюється:

 (2.3)

Теорема 1. Математичне сподівання числа появ події в одному випробуванні дорівнює ймовірності цієї події.

Доведення. В одному випробуванні може бути два результати: *х* = 0 з імовірністю *q* і *х* = 1 з імовірністю *р.* Тоді *М*(*х)* = 1⋅*р* + 0⋅*q* = *р.*

Лема. Математичне сподівання приблизно дорівнює (тим точніше, чим більше випробувань) середньому арифметичному спостережуваних значень випадкової величини.

Властивості математичного сподівання:

1) *М*(*С*) = *С*, де *С* – константа.

Доведення. Нехай *С* – дискретна випадкова величина з імовірністю *р* = 1. Тоді *М*(*С*) = *С*⋅1 = *С.*

2) *М*(*СХ*) = *С*⋅*М*(*х)*.

Доведення. Нехай випадкова величина *Х* може приймати значення *х1*, *х2*, … з ймовірностями *р1*, *р2*, … відповідно. Тоді випадкова величина *С\*Х* може приймати значення *С\*х1*, *С\*х2*, … з тими ж ймовірностями *р1*, *р2*, … ... Математичне сподівання дорівнює:

*М*(*С\*Х*) = *С\*х1\*р1* + *С\*х2\*р2* +…= *С\**(*х1\*р1* + *х2\*р2* +…) = *С*\**М*(*Х*) (2.4)

3) Математичне сподівання добутку незалежних випадкових величин дорівнює добутку їхніх математичних сподівань: *М*(*X\*Y*) = *M*(*X*)\**M*(*Y*).

4) Математичне сподівання суми випадкових величин дорівнює сумі математичних сподівань доданків: *М*(*X* + *Y*) = *M*(*X*) + *M*(*Y*).

Наслідок. Властивості 3 і 4 справедливі для будь-якої кількості випадкових величин.

Теорема 2. Математичне сподівання відхилення випадкової величини від свого математичного сподівання дорівнює нулю.

*М[Х – М(Х)]* = 0 (2.5)

Доведення. Для доведення розкриємо дужки й використаємо властивості математичного сподівання, а також той факт, що *М(М(Х)) = М(Х)* (оскільки математичне сподівання не є випадковою величиною, це – константа).

*М[X – M(X)] = M(X) – M(M(X)) = M(X) – M(X)* = 0 (2.6)

**2.2 Дисперсія випадкової величини**

Визначення.*Дисперсією* випадкової величини *Х* називається математичне сподівання квадрата відхилення випадкової величини від свого математичного сподівання. Позначається: *Д*(*Х*), *σ2*(*Х*).

*Д(Х) = М[Х – М(Х)]2* (2.7)

Для дискретних випадкових величин дисперсія обчислюється за формулою:

,  (2.8)

де *хi*, *рi* – значення з таблиці розподілу ймовірностей. Іноді дисперсію називають просто квадратом відхилення випадкової величини.

Для безперервних випадкових величин дисперсія узагальнює формулу для дискретних і обчислюється:

 (2.9)

Теорема. Дисперсія дорівнює різниці математичного сподівання від квадрата випадкової величини й квадрата математичного сподівання випадкової величини:

 (2.10)

Доведення. За визначенням дисперсії, одержуємо:

*Д(Х) = М(Х – М(Х))2 = М(Х2 – 2М(Х)Х + М2(Х)) = М(Х2) – 2М(Х)М(Х) + М2(Х) = М(Х2) – 2М2(Х) + М2(Х) = М(Х2) – [М(Х)]2.*

Властивості дисперсії:

1) Дисперсія константи дорівнює нулю.

*Д*(*С*) = 0 (2.11)

Доведення. За визначенням дисперсії маємо:



2) Константа виноситься з-під знаку дисперсії у квадраті.

*Д*(*СХ*) = *С2Д*(*Х*) (2.12)

Доведення. За визначенням дисперсії маємо:



3) Дисперсія суми двох і більше незалежних випадкових величин дорівнює сумі їхніх дисперсій.

*Д(Х + Y) = Д(Х) + Д(Y)* (2.13)

Доведення. Використовуючи теорему, одержуємо:



4) Дисперсія різниці двох випадкових величин дорівнює сумі їхніх дисперсій.

*Д(Х – Y) = Д(Х) + Д(Y)* (2.14)

Доведення.

*Д(Х – Y) = Д(Х) + Д( –Y) = Д(Х) + ( –1)2Д(Y) = Д(Х) + Д(Y).*

Визначення. Середнім квадратичним відхиленням випадкової величини називають квадратний корінь із дисперсії цієї випадкової величини.

 (2.15)

Лема. Дисперсія числа появи події *А* в *n* незалежних випробуваннях обчислюється за формулою:

*Д*(*Х*) = *npq* (2.16)

**2.3 Моменти випадкової величини**

Більш загальними характеристиками випадкової величини є початкові й центральні моменти *k –го* порядку.

Визначення. *Початковим моментом k –го порядку* називають математичне сподівання *k –ого* ступеня випадкової величини:

 (2.17)

Примітка. Математичне сподівання – це початковий момент 1–го порядку.

Визначення.*Центральним моментом k–го порядку* називають математичне сподівання *k –ого* ступеня відхилення:

 (2.18)

Для дискретних випадкових величин:

 (2.19)

Примітка.

 (2.20)

Визначення. *Асиметрією* називають відношення третього центрального моменту до куба середнього квадратичного відхилення.

 (2.21)

Ця величина характеризує несиметричність розподілу відносно *М*(*Х*). Дійсно, у сумі  або в інтегралі  при симетричному відносно *М*(*Х*) законі розподілу кожному позитивному доданку відповідає рівний йому по модулю від’ємний доданок, так що вся сума або інтеграл дорівнюють нулю.

Четвертий центральний момент служить для характеристики «крутості», тобто гостровершинності або плосковершинності розподілу. Ця властивість описується за допомогою так званого ексцесу.

Визначення. *Ексцесом* випадкової величини *Х* називається величина *Ех*, що обчислюється за формулою:

 (2.22)

У цій формулі віднімається число 3 так, як і для досить важливого нормального закону розподілу  й *Ех* = 0. Таким чином, для кривих більш гостровершинних *Ех*> 0, а для менш гостровершинних – *Ех*< 0.

На рисунку 2.2 зображені графіки щільності розподілу ймовірностей двох випадкових величин. При зсуві графіка ліворуч стосовно математичного сподівання асиметрія негативна (графік 1), а праворуч – позитивна (графік 2).

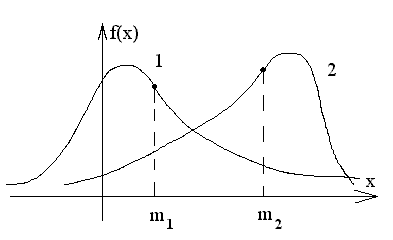


Рисунок 2.2 – Графік кривої розподілу для різних значень асиметрії

На рисунку 2.3 представлене порівняння графіків щільності розподілу випадкових величин з нормальною випадковою величиною за допомогою ексцесу. При більш швидкому зростанні ймовірності, ніж у нормального закону (графік 2) величина ексцесу позитивна (графік 1), а при більш повільному – негативна (графік 3). У нормального закону, як в еталонного, величини асиметрії й ексцесу дорівнюють нулю.

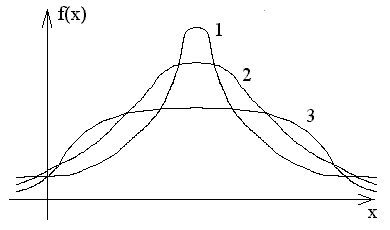


Рисунок 2.3 – Порівняння кривих розподілу для різних значень ексцесу

Питання для самоконтролю

1. Що називається математичним сподіванням ВВ? Що характеризує МС?
2. Сформулюйте основні властивості МС.
3. Що називається модою й медіаною ВВ?
4. Які розподіли ймовірностей називаються одномодальними, полімодальними й антимодальними?
5. Дайте визначення дисперсії ВВ. Що вона характеризує?
6. Що називається середнім квадратичним відхиленням ВВ?
7. Сформулюйте основні властивості дисперсії.
8. Дайте визначення поняття початкового моменту *k* –го порядку.
9. Дайте визначення поняття центрального моменту *k* –го порядку.
10. Вивести співвідношення для μ2, μ3 і μ4 через ν1, ν2, ν3 і ν4.
11. Що характеризує асиметрія й ексцес ВВ?

Навчальні завдання

1. Історія розвитку статистичного аналізу даних.
2. Дайте порівняльну характеристику основним стратегіям формування репрезентативних вибірок дослідження.
3. Основні властивості метричних і неметричних шкал вимірювання.
4. Правила ранжування даних.
5. Особливості попереднього аналізу даних у вибірці.
6. Відновлення пропущених спостережень. Перевірка первинних даних на наявність аутлаєрів (викидів).
7. Візуальне представлення результатів дослідження за допомогою графіків.
8. Стандартизація даних і стандартизовані шкали.
9. Характеристика основних методів первинної статистичної обробки даних.

Практичні завдання

***Завдання 1.****Варіаційні ряди та статистичні розподіли. Обчислення емпіричних даних статистичними засобами первинної обробки.*

*Мета:* ознайомитись з основними поняттями статистичних показників вибірки, навчитись правильно застосовувати їх для розв’язування задач, встановлювати приховані взаємозв’язки та закономірності явищ, робити прогноз розвитку досліджуваних процесів.

Написати програму, що генерує випадкову послідовність даних з (N +10)

елементів, які набувають значень із набору (1, 2, 3, 4, 5) (N – номер студента у журналі). Одержати та вивести на екран вихідні дані, варіаційний ряд, статистичний розподіл, інтегральну частоту та частість. Обчислити моду, медіану, середнє арифметичне сукупності. Обчислення кожної з цих величин має бути реалізоване у вигляді окремого блоку програми (функції/процедури/класу/методу класу залежно від використовуваного середовища для виконання робіт).

Рекомендовано максимально використовувати стандартні бібліотеки та алгоритми.

***Завдання 2.*** Наведені в таблиці дані описують обсяги продукції, що випускається дрібним виробником протягом однієї години.Побудувати груповану вибірку, визначити медіану, моду вибірки. Обчислити середню кількість виробів за годину, вибіркове середнє квадратичне відхилення.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 136 | 122 | 132 | 128 | 123 | 133 | 130 | 131 |
| 134 | 149 | 138 | 127 | 119 | 137 | 133 | 130 |
| 143 | 134 | 128 | 131 | 118 | 133 | 131 | 132 |
| 118 | 128 | 122 | 130 | 139 | 145 | 122 | 130 |
| 128 | 136 | 132 | 126 | 124 | 117 | 139 | 132 |
| 141 | 144 | 138 | 133 | 127 | 150 | 144 | 133 |
| 134 | 125 | 140 | 135 | 129 | 138 | 138 | 147 |
| 150 | 126 | 135 | 136 | 150 | 135 | 138 | 140 |
| 122 | 142 | 127 | 127 | 132 | 145 | 140 | 133 |
| 127 | 142 | 144 | 125 | 132 | 145 | 137 | 132 |

***Завдання 3.***Ряд розподілу дискретної випадкової величини *X* має вигляд:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Х | 1 | 2 | 3 | 4 |
| *р* |  |  |  |  |

Знайти M(X) та P(X > 2).

***Завдання 4.***Ряд розподілу дискретної випадкової величини *X* має вигляд:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Х | 1 | 2 | 3 | 4 |
| *р* |  |  |  |  |

Знайти дисперсію то моду випадкової величини.

***Завдання 5.***Ряд розподілу дискретної випадкової величини *X* має вигляд:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Х | 1 | 3 | 5 | 7 | 9 |
| *р* | 0,1 | 0,4 | 0,2 | 0,2 | 0,1 |

Знайти початкові та центральні моменти перших чотирьох порядків випадкової величини X.

***Завдання 6.***Ряд розподілу дискретної випадкової величини *X* має вигляд:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Х | 1 | 3 | 5 | 7 | 9 |
| *р* | 0,1 | 0,4 | 0,2 | 0,2 | 0,1 |

Визначити ексцес та асиметрію розподілу випадкової величини *X*.

Тестові завдання для перевірки знань

1. Дати визначення генеральної сукупності.

а) це сукупність об’єктів, з якої зроблено вибірку;

б) сукупність випадково взятих об’єктів;

в) множина однорідних об’єктів;

г) інша відповідь.

2. Дати визначення вибіркової сукупності (вибірки).

а) це сукупність об’єктів, з якої зроблено вибірку;

б) сукупність випадково взятих об’єктів;

в) множина однорідних об’єктів;

г) інша відповідь.

3. Що називається обсягом (об’ємом) сукупності?

а) це кількість об’єктів цієї сукупності;

б) це зростаючий числовий ряд варіант;

в) множина однорідних об’єктів;

г) інша відповідь.

4. Що називається варіантою?

а) це кількість об’єктів цієї сукупності;

б) це кількісна ознака, наприклад *X* , яка набуває конкретних числових значень ;

в) множина однорідних об’єктів;

г) інша відповідь.

5. Що називається варіаційним рядом?

а) це спадний числовий ряд варіант;

б) це довільний числовий ряд варіант;

в) це зростаючий числовий ряд варіант;

г) це будь-який ряд, що набуває кількісна ознака.

6. Що таке частота варіант?

а) це відношення частоти  варіанти до обсягу вибірки *n* ;

б) це кількість спостережуваних варіант ;

в) це кількість членів, які приймають рівні значення ознаки або знаходяться у середині певного інтервалу;

г правильної відповіді серед перерахованих немає.

7. Що називається рядом частот?

а) це відношення частоти  варіанти  до обсягу вибірки *n* ;

б) це кількість членів, які приймають рівні значення ознаки або знаходяться у середині певного інтервалу;

в)це кількість спостережуваних варіант ;

г) правильної відповіді серед перерахованих немає.

8. Що таке відносна частота варіант?

а) це кількість спостережуваних варіант ;

б) це кількість членів, які приймають рівні значення ознаки або знаходяться у середині певного інтервалу;

в) це відношення частоти *in* варіанти *xi* до обсягу вибірки n ;

г) правильної відповіді серед перерахованих немає.

9. Дати визначення дискретного статистичного розподілу вибірки.

а) це відношення частоти  варіанти  до обсягу вибірки *n* ;

б) це довільний числовий ряд варіант;

в) це перелік варіант та відповідних їм частот;

г) правильної відповіді серед перерахованих немає.

10. Що являє собою полігон частот і відносних частот?

а) функція , яка є неспадною функцією;

б) це кількість спостережуваних варіант ;

в) це є довільна площина, на якій точками зображені кількісні ознаки об’єкта, і які нагадують мішень після стрільби;

г) це є дискретний статистичний розподіл вибірки, який можна зобразити графічно у вигляді ламаної лінії, відрізки якої сполучають координати точок або .

11. Що називається точковою статистичною оцінкою?

а) статистична оцінка випадкової величини, яка визначається одним числом;

б) статистична оцінка, яка визначається двома числами, кінцями інтервалів;

в) оцінка, яка при заданому обсязі вибірки має мінімальну дисперсію;

г) інша відповідь.

|  |
| --- |
|  |

# ЗМІСТОВИЙ МОДУЛЬ 2. ЗАКОНИ РОЗПОДІЛУ ВИПАДКОВИХ ВЕЛИЧИН

## Тема 3. Закони розподілу дискретних випадкових величин

Мета: **Отримати ґрунтовне розуміння трьох основних законів розподілу**дискретних випадкових величин, н**авчитися використовувати закони розподілу для вирішення практичних задач.**

🖉Основні терміни і поняття

*Біноміальний закон розподілу, розподіл Пуассона, потік подій, стаціонарність, відсутність післядії, ординарність, інтенсивністю потоку.*

📚Основні теоретичні положення

**3.1 Біноміальний закон розподілу**

Визначення.*Біноміальним* називають розподіл випадкової величини, ймовірності якої обчислено за допомогою формули Бернуллі.

Приклад. Нехай проводиться *n* незалежних випробувань, у кожному з яких подія *А* може з'явитися з імовірністю *р* або не з'явитися з імовірністю *q* (*q* = 1 – *p*). Складемо таблицю розподілу випадкової величини *Х* – число успіхів в n випробуваннях.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *X* | *n* | *n –1* | … | *k* | … | *0* |
| *p* | *pn* |  | … |  | … | *qn* |

Якщо скласти всі ймовірності в другому рядку, то одержимо формулу бінома Ньютона:

 (3.1)

Теорема 1. Математичне сподівання числа появи події *А* в *n* незалежних випробуваннях дорівнює *М(х)* = *n*\**p*.

Доведення. Нехай випадкова величина *Х* – число появ події *А* в *n* випробуваннях. Тоді *Х* = *Х1* + *Х2* + … + *Хn*, де *Xi* – число появ події *А* в *i –ом* випробуванні (індикатор). За властивістю 3 математичного сподівання одержимо *М(Х) = М(Х1) + М(Х2) + … + М(Хn).* Однак математичне сподівання в кожному випробуванні постійне й дорівнює ймовірності події, тобто *М(Х)* = *n\*p*.

Теорема2. Мода (найбільш ймовірне число успіхів) для біноміального закону обчислюється за формулою:

*np – q ≤ m0 ≤ np + p* (3.2)

Визначення. Багаточлен ступеня *n*, коефіцієнти при невідомих у якому дорівнюють ймовірностям біноміального розподілу, називається *виробляючою функцією*:

*ϕn(z)* = (*q* + *p*\**z)n* (3.3)

**3.2 Розподіл Пуассона**

Розглянемо дискретну випадкову величину *Х*, що може приймати цілі невід’ємні значення: 0, 1, 2,⋅⋅⋅, *m*, ⋅⋅⋅, причому послідовність цих значень не обмежена.

Визначення. Кажуть, що випадкова величина *Х розподілена за законом*

*Пуассона*, якщо ймовірність того, що вона прийме деяке значення *m* дорівнює:

 (3.4)

де *λ*> 0 – параметр закону Пуассона.

За допомогою розкладення експоненти в ряд Тейлора можна довести основну тотожність всіх законів розподілу для розподілу Пуассона:

 (3.5)

Обчислимо математичне сподівання:

 (3.6)

Аналогічним чином, обчислюючи дисперсію, отримуємо:

*Д(Х) = М(Х) = λ* (3.7)

На практиці велике значення має наступна лема.

Лема. Закон Пуассона є граничним для біноміального розподілу при одночасному прагненні *n*→∞ і *р*→0 так, щоб *n*\**p* завжди дорівнювало *λ*.

Доведення.

 (3.8)



Примітка. Нехай проводиться велика кількість випробувань *n*, у кожному з яких подія *А* має малу ймовірність *р.* Тоді для обчислення  використовується такий вираз:

 (3.9)

Тому закон Пуассона має другу назву – *закон рідких явищ*.

**3.3 Потік подій**

Визначення.*Потоком подій* називають послідовність подій, які настають у випадкові моменти часу.

Наприклад, надходження викликів на АТС, літаків у аеропорт, інше.

Властивості потоків:

1) *Стаціонарність* – імовірність появи *k* подій на будь-якому проміжку часу залежить тільки від числа *k* і від тривалості *Δ t* проміжку й не залежить від початку його відліку. Проміжки мають на увазі непересічні.

2) *Відсутність післядії* – незалежність появ того або іншого числа подій у непересічні проміжки часу.

3) *Ординарність* – за нескінченно малий проміжок часу може відбутися не більше однієї події.

Визначення. Якщо виконуються властивості: стаціонарність, відсутність наслідку й ординарність, то потікє *найпростішим (пуассоновим).*

Визначення. *Інтенсивністю потокуλ* називають середнє число подій за одиницю часу.

Імовірність появи *k* подій за час тривалістю *t* визначається за формулою:

 (3.10)

Приклад. Середнє число викликів на АТС за одну хвилину дорівнює 2. Знайти ймовірність того, що за п'ять хвилин буде три виклики.

За умовою задачі *λ* = 2, *t* = 5, *k* = 3, отже, використовуючи формулу (3.10), одержуємо: .

Питання для самоконтролю

1. Що називають біноміальним законом розподілу ймовірностей?
2. Наведіть числові характеристики *M*(*Х*), *D*(*X*) і σ(*Х)* для біноміального закону.
3. Дайте визначення виробляючої функції.
4. Що називають пуассоновим законом розподілу ймовірностей?
5. Наведіть числові характеристики *M*(*X*), *D*(*X*) і  для закону Пуассона.
6. Сформулюйте граничний перехід від біноміального розподілу до розподілу Пуассона.
7. Що таке потік подій?
8. Назвіть властивості пуассонового потоку.
9. Книга у 100 сторінок має в середньому 50 помилок. Яка ймовірність того, що на одній сторінці буде 3 помилки?

Практичні завдання

***Завдання 1.*** Нехай X – кількість появ числа 5 при двох киданнях грального кубика. Знайти закон розподілу величини *X*.

***Завдання 2.*** Виконується 4 незалежних пострілу по мішені. Ймовірність влучення при одному пострілі дорівнює 0,25. Знайти закон розподілу кількості влучень X, найімовірніше число влучень.

***Завдання 3.*** Підручник видано накладом 10000 примірників. Ймовірність того, що зброшурований екземпляр виявиться неякісним, дорівнює 0,001. Знайти ряд розподілу випадкової величини *X*, що характеризує кількість неякісно зброшурованих екземплярів.

***Завдання 4.*** Автоматична телефонна станція обслуговує 10000 телефонних номерів. Ймовірність того, що впродовж 1 хвилини до АТС надійде виклик абонента, дорівнює 0,0004. Знайти ряд розподілу випадкової величини X, яка дорівнює кількості викликів, які надійшли до АТС впродовж 1 хв, та ймовірність того, що за цей час надійде хоча б один виклик.

## Тема 4. Закони розподілу безперервних випадкових величин

Мета: **Отримати ґрунтовне розуміння трьох основних законів розподілу безперервних** випадкових величин, н**авчитися використовувати закони розподілу для вирішення практичних задач.**

🖉Основні терміни і поняття

*Рівномірний закон розподілу, нормальний закон розподілу, стандартний (нормований) нормальний розподіл, правило трьох σ, показниковий розподіл, показниковий закон надійності.*

📚Основні теоретичні положення

**4.1 Рівномірний розподіл**

Іноді трапляються безперервні випадкові величини, про які відомо, що їхні значення лежать у межах певного інтервалу, причому в межах цього інтервалу всі значення рівноймовірні. Ці випадкові величини розподілені за законом рівномірної щільності.

Визначення. Розподіл називають *рівномірним*, якщо на інтервалі, якому належать всі можливі значення випадкової величини, щільність розподілу постійна, а поза цим інтервалом дорівнює нулю.

Приклад. Проводиться зважування на вагах з використанням гирьок у один грам, визначають, що вага лежить між *k* і (*k* + 1) грамами, приймають вагу рівну (*k* + 1/2) грам. Тоді помилка – випадкова величина, розподілена рівномірно на інтервалі ( –1/2; +1/2).

Приклад. Обертають колесо. Розглядається випадкова величина *Х* – кут *θ*, що утворить радіус із обрієм. Вона розподілена рівномірно на відрізку (0; 2*π*].

Виходячи з визначення, щільність має вигляд:

 (4.1)

Оскільки площа, обмежена кривою розподілу, дорівнює одиниці, то *С*\*(*β* – *α*) = 1, *С* = 1/(*β* – *α*). Отже:

 (4.2)

Функція рівномірного розподілу має такий вигляд:

 (4.3)

Графічно функція розподілу зображена на рисунку 4.1.

Математичне сподівання величини *Х* дорівнює:

 (4.4)

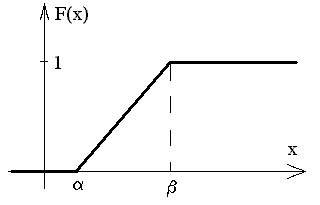


Рисунок 4.1 – Функція рівномірного розподілу

У силу симетричності рівномірного розподілу медіана дорівнює середині інтервалу: *mе*= (*α*+ *β*)/2, мода не визначена, дисперсія знаходиться таким способом:

 (4.5)

Отже, середньо квадратичне відхилення дорівнює:

 (4.6)

Асиметрія дорівнює нулю, для визначення ексцесу знаходимо четвертий центральний момент:

 (4.7)

Нарешті, знайдемо ймовірність влучення випадкової величини *Х*, розподіленої рівномірно, на ділянку (*а*; *b*).

1) Якщо *α*<*а* <*b* <*β*, то:

 (4.8)

2) Якщо *a*<*α*<*β*<*b*, то:

 (4.9)

3). Якщо *a*, *b* не належать інтервалу (*α*; *β*), то *Р* = 0.

Приклад. Нехай випадкова величина рівномірно розподілена на інтервалі (2; 8). Знайти *Р*(4 < х < 6).

У даному прикладі маємо перший випадок. Тоді шукана ймовірність дорівнює:



**4.2 Нормальний закон розподілу**

Частіше за все на практиці використовується закон, до якого наближаються інші закони за певних умов, і який називається нормальним законом розподілу (Гауса).

Визначення.*Нормальним* називають розподіл *N*(*a*, *σ2)* випадкової величини з параметрами *а*, *σ* (*σ*> 0), що описується такою щільністю:

 (4.10)

Знайдемо числові характеристики:

1) Математичне сподівання.

 (4.11)

Уведемо нову змінну , тоді:

 (4.12)

У першому інтегралі підінтегральна функція непарна, і тому інтеграл дорівнює нулю, другий інтеграл – інтеграл Пуассона, який дорівнює , отже, математичне сподівання дорівнює параметру а.

*М(Х) =а* (4.13)

2) Дисперсія.

 (4.14)

Уводимо ту ж змінну *z*, одержуємо:

 (4.15)

Інтегруючи частинами, знаходимо:

*Д(Х) =σ2* (4.16)

Отже, середнє квадратичне відхилення дорівнює параметру *σ*.

Визначення. називається розподіл *N*(0;1), *Стандартним (нормованим) нормальним розподілом*

 (4.17)

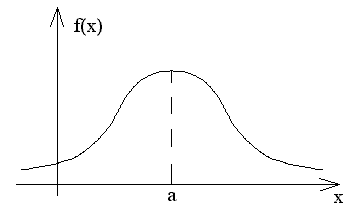


Рисунок 4.2 – Функція розподілу нормального закону

Якщо *Х* – випадкова величина з нормальним законом розподілу та параметрами (*а*, *σ*), то *U = (Х* – *a*)/*σ* – нормована випадкова величина. Покажемо це на прикладах математичного сподівання й дисперсії:

 (4.18)



Таким чином, будь –яку нормальну випадкову величину можна привести до стандартного виду.

Проведемо дослідження кривої нормального розподілу, використовуючи методи диференціального числення.

Можна зробити наступні висновки:

1. Функція визначена на всій числовій осі.
2. Для кожного значення аргументу функція додатна, тобто функція розташована над віссю Ох.
3. Межа функції при прагненні *Х* до нескінченності дорівнює нулю, тобто вісь Ох – горизонтальна асимптота.
4. За допомогою першої похідної встановимо точку екстремуму:



Крім того, функція зростає зліва від точки *х* = *а* й спадає справа від цієї точки.

5. Функція симетрична відносно прямої *х* = *а*.

6. За допомогою другої похідної встановимо, що точками перегину є значення аргументу *х* = *а*±*σ*.

Виведемо формулу для обчислення ймовірності влучення випадкової величини, розподіленої за нормальним законом, у деякий інтервал. За визначенням щільності розподілу маємо:

 (4.19)

Скористаємося заміною змінних, яка була запропонована нами раніше для нормування, і до результату застосуємо формулу для функції Лапласа *Ф(х)*, значення якої існують у таблицях:



Тоді

 (4.20)

Приклад. Нехай випадкова величина *Х* розподілена за нормальним законом з параметрами *а* = 30 і *σ* = 10. Знайти ймовірність влучення в інтервал *Р*(10 <*x*< 50).

Використовуючи формулу (7.20), одержуємо:



Зауваження. Функція Лапласа непарна, *Ф( –х) = –Ф(х)*.

Обчислимо ймовірність того, що відхилення нормального розподілу випадкової величини *Х* по модулю менше *δ*, тобто *Р(|x – a|*<*δ*).

Р(|x – *a*| <*δ*) = Р(*a* – *δ*< x <*а* + *δ*) = (4.21)

= 

Якщо прийняти *δ* = *nσ*, то одержимо:

*Р(|x – a|*<*nσ*) = 2*Ф*(*n*) (4.22)

При цьому, якщо *n* = 3, *δ* = 3*σ*, то:

*Р(|x – a| <* 3*σ*) = 2*Ф*(3) = 2⋅0.49865 = 0.9973 (4.23)

Правило трьох *σ*. Якщо випадкова величина розподілена за нормальним законом, то абсолютна величина її відхилення від математичного сподівання не перевищує потроєного середнього квадратичного відхилення. Тобто для нормального розподілу практично достовірно, що відхилення від центра не перевищать 3*σ*.

Знайдемо тепер функцію нормального розподілу. За визначенням:

 (4.24)

Для нормованого нормального розподілу:

 (4.25)

Зазначимо, що . Уведемо тепер функцію Лапласа, значення якої табульовані:

 (4.26)

З огляду на те, що  й те, що для *N*(0;1) функція *ϕ*(х) симетрична відносно нуля, отримуємо:

 (4.27)

Тоді:

*F0(x)* = *P( –∞< X < x) = P( –∞< X <* 0) + *P*(0 <*X < x*) = 0.5 + *Ф(х)* (4.28)

Аналогічно:

*F*(*x*) =  (4.29)

Розглянемо функцію *F(x).*

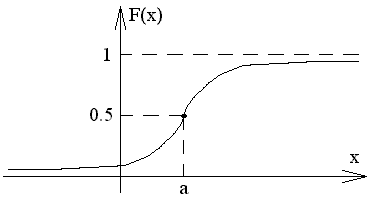


Рисунок 4.3 – Графік функції нормального розподілу

Властивості:

1. *F*( –∞) = 0;
2. *F*(+∞) = 1;
3. *F(x)* – неспадна;
4. Для *N*(0;1) *F( –x)* = 1 – *F(x).*

**4.3 Показниковий розподіл**

Визначення.*Показниковим (експоненціальним)* називається розподіл випадкової величини *Х*, що описується наступною щільністю розподілу:

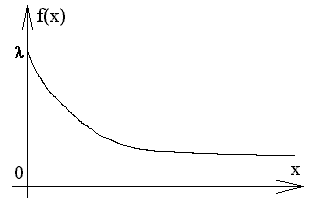


Рисунок 4.4 – Графік щільності показникового розподілу

 (4.30)

де *λ* – додатна константа.

Прикладом показникового закону може служити випадкова величина *Х*, де *Х* – час між появами двох послідовних подій найпростішого потоку.

Знайдемо функцію розподілу:

 (4.31)

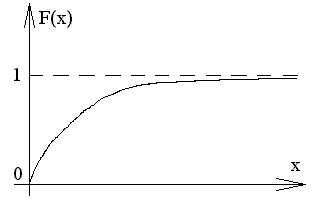


Рисунок 4.5 – Графік функції показникового розподілу

Отже, остаточно маємо:

 (4.32)

Знайдемо ймовірність влучення випадкової величини *Х* у інтервал (*а*; *b*). За визначенням маємо:

 (4.33)

Приклад. Випадкова величина розподілена за законом:

. Знайти *Р*(0.3 <*х*< 1).

За умовою задачі *λ* = 2, тоді за формулою (7.33) одержуємо:

.

Знайдемо математичне сподівання й дисперсію:

1) Для знаходження математичного сподівання використовуємо інтегрування частинами:

 (4.34)

2) Аналогічним чином проведемо інтегрування для знаходження дисперсії:

 (4.35)

Звідси:

 (4.36)

Приклад. Нехай електроприлад починає працювати в *t0* = 0, і по закінченню проміжку часу *t* відбувається відмова. Позначимо випадкову величину *Т* – тривалість безвідмовної роботи.

Функція розподілу *F(t) = P(T < t)* визначає ймовірність відмови за час t. Отже, імовірність безвідмовної роботи за час *t* обчислюється за формулою *R(t) = P(T < t) = 1 – F(t).* Функцію *R(t)* називають *функцією надійності*. Часто тривалість безвідмовної роботи має показовий розподіл, тобто .

Визначення.*Показниковим законом надійності* називають функцію:

 (4.37)

Приклад. Час безвідмовної роботи елемента розподілений за законом . Знайти ймовірність того, що елемент пропрацює 100 годин.

За умовою постійна інтенсивності відмов *λ* = 0.02, отже:

R(100) = e –0.02⋅100 = e –2≈ 0.13534.

Питання для самоконтролю

1. Що називають рівномірним законом розподілу ймовірностей і який аналітичний і графічний вигляд мають його *f* (*x*) і *F* (*x*)?
2. Наведіть числові характеристики для рівномірного закону й виведіть співвідношення для асиметрії й ексцесу.
3. Що називають показниковим законом розподілу ймовірностей і який аналітичний і графічний вигляд мають його *f* (*x*) і *F* (*x*)?
4. Виведіть формулу для знаходження ймовірності влучення ВВ, розподіленої за показниковим законом, у заданий інтервал.
5. Наведіть числові характеристики для показникового закону й виведіть співвідношення для асиметрії й ексцесу.
6. Що називають нормальним законом розподілу ймовірностей і який аналітичний і графічний вигляд мають його *f* (*x*) і *F* (*x*)?
7. Як впливають параметри *а* й *σ* на графік *f* (*x*) *N* (*а*; *σ*)?
8. У чому полягає теоретико-ймовірносний зміст параметрів *а* й *σN* (*а*; *σ*)?
9. Знайдіть аналітичний зв'язок *F* (*x*)і *F*0(*x*).
10. Чому дорівнюють *µ0*, *µе*, асиметрія й ексцес для *N* (*m*; σ)?
11. Виведіть формулу для знаходження *P*(*a < X < b*) *N*(*а*; σ).
12. Виведіть формулу для знаходження *N*(*а*; σ).
13. Сформулюйте правило трьох сигм.

Практичні завдання

***Завдання 1.*** Автобуси ходять з інтервалом в 5 хвилин. Вважаючи, що випадкова величина *X* – час чекання автобуса на зупинці – розподілена рівномірно на вказаному інтервалі, знайти середній час чекання та дисперсію часу чекання.

***Завдання 2.*** Випадкова величина *X* має нормальний розподіл ймовірностей з параметрами *a = 7 і σ =2*. Порівняти ймовірності влучення значень випадкової величини *X* у проміжки [3; 7] і [–100; 1]. Зазначити інтервал, у який випадкова величина *X* влучає з практичною достовірністю.

***Завдання 3.*** Випадкова величина *X* має показниковий розподіл. Ймовірність того, що ця випадкова величина набуде значення з проміжку [0; 5], дорівнює 0,7. Знайти ймовірність того, що ця випадкова величина набуде значення з проміж-ку [7; 9].

***Завдання 4.*** Час безвідмовної роботи радіоапаратури є випадковою величиною *X*, розподіленою за показниковим законом із параметром λ . Знайти ймовірність того, що радіоапаратура не вийде з ладу упродовж часу *t* = *mX*.

**Тема 5. Двовимірні випадкові величини**

Мета: **Отримати ґрунтовне розуміння основних понять теорії двовимірних випадкових величин, навчитися використовувати знання про двовимірні випадкові величини для вирішення практичних задач.**

🖉Основні терміни і поняття

*Закон розподілудискретної двовимірної випадкової величини, функція розподілу двовимірної випадкової величини, щільність спільного розподілу ймовірностей, математичне сподівання двовимірної випадкової величини, умовний розподіл складової, умовна щільність розподілу, коефіцієнт кореляції,коваріація, нормальний закон розподілу на площині.*

📚Основні теоретичні положення

**5.1 Поняття двовимірної випадкової величини**

На практиці часто трапляються випадки, у яких результат випробування описується не однією випадковою величиною, а двома або більше випадковими величинами, що утворюють *комплекс* або *систему* випадкових величин. Будемо позначати через (*Х,Y*) двовимірну випадкову величину, кожну з величин *Х* і *Y* будемо називати складовою. Систему двох випадкових величин можна зображувати *випадковою точкою* на площині з координатами (*Х,Y*) або ж *випадковим вектором*. Доцільно розрізняти *дискретні* (*Х, Y* – дискретні) і *безперервні* випадкові величини.

Приклад. Нехай верстат штампує пластинки. Контрольовані розміри *Х* – довжина й *Y* – ширина, складають двовимірну випадкову величину.

Приклад. Стрілець стріляє в мішень. Координати влучення *(Х,Y)* – двовимірна випадкова величина.

Визначення. *Законом розподілу* дискретної двовимірної випадкової величини називають перелік всіх можливих значень цієї величини *(хi,yj)* і їхніх ймовірностей *р(хi,yi),* де *i* = 1, …, *n*, *j* = 1, …, *m*...

Зазвичай закон розподілу задають у вигляді таблиці.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *X*  *Y* | *x1* | *x2* | *…* | *xn* |
| *y1* | *p(x1,y1)* | *p(x2,y1)* | *…* | *p(xn,y1)* |
| *y2* | *p(x1,y2)* | *p(x2,y2)* | *…* | *p(xn,y2)* |
| *…* | *…* | *…* | *…* | *…* |
| *ym* | *p(x1,ym)* | *p(x2,ym)* | *…* | *p(xn,ym)* |

Оскільки події *(Х = хi, Y = yi)* утворюють повну групу подій, то сума всіх клітинок дорівнює одиниці. Знаючи закон розподілу двовимірної випадкової величини, можна знайти закони розподілу кожної складової. Так, наприклад, події *(Х=х1, Y=y1), (Х=х1, Y=y2), …* несумісні, тому ймовірність *р(х1)* дорівнює сумі ймовірностей стовпця «*х1*». Аналогічно для *Y*.

**5.2 Функція розподілу двовимірної випадкової величини**

Визначення. *Функцією розподілу двовимірної випадкової величини* називають функцію *F(x,y)*, що визначає для кожної пари *(x,y)* імовірність того, що *Х < x* і при цьому *y < Y*, тобто *F(x,y) = P(X < x, Y < y)*.

Геометричний зміст: *F(x,y)* є ймовірність того, що випадкова точка *(X,Y)* потрапить у нескінченний квадрант зліва й нижче *(x,y).*

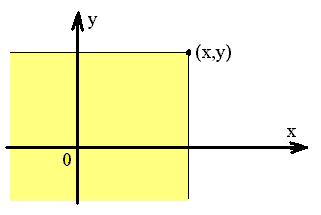


Рисунок 5.1 – Геометричний зміст функції розподілу

Властивості функції розподілу:

1) 0 ≤*F(x,y)*≤ 1 з визначення ймовірності.

2) *F(x,y)* – неспадна за кожним аргументом, тобто

*F(x2,y) ≥ F(x1,y)* при *х2> x1*;

*F(x,y2) ≥ F(x,y1)* при *y2> y1*.

Доведення. *P(X < x2, Y < y) = P(X < x1, Y < y) + P(x1≤ X < x2, Y < y).* Звідси *P(X < x2, Y < y) – P(X < x1, Y < y) = P(x1≤ X < x2, Y < y)*, оскільки будь-яка ймовірність невід’ємна, то *F(x2,y) ≥ F(x1,y)*.

3) *F( –∞, y) = F(x, –∞) = F( –∞, –∞) =* 0,

*F(∞,∞)* = 1.

4) *F(x, ∞) = F1(x),*

*F(∞, y) = F2(y)*

Функції *F1(x)* і *F2(y)* називаються *маргінальними функціями розподілу компонент.* Імовірність влучення випадкової точки в прямокутник буде обчислюватися за формулою:

*P(x1≤ X < x2, y1≤ Y < y2) = [F(x2, y2) – F(x1, y2)] – [F(x2, y1) – F(x1, y1)]* (5.1)

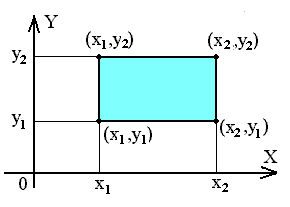


Рисунок 5.2 – Ймовірність влучення випадкової величини в прямокутник

**5.3 Двовимірна щільність ймовірностей**

Визначення. *Щільністю спільного розподілу ймовірностейf(x,y)* двовимірної безперервної випадкової величини *(X,Y)* називають другу змішану частинну похідну від функції розподілу:

 (5.2)

Приклад. Знайти щільність за відомою функцією розподілу

*F(x,y)* = sin*x*⋅sin*y* (0 ≤*x*≤ π/2; 0 ≤*y*≤ π/2).

Візьмемо частинну похідну за *х*, а потім за у, одержимо:

.

Знаючи щільність спільного розподілу, можна знайти функцію розподілу за формулою:

 (5.3)

Лема. Для обчислення ймовірності влучення випадкової точки *(Х,Y)* в область *D* використовують формулу:

 (5.4)

Властивості щільності розподілу ймовірностей:

1). *f(x,y)*≥ 0, тому що функція *F(x,y)* – неспадна.

2). .

Визначення.*Щільність розподілу однієї зі складових* дорівнює невласному інтегралу з нескінченною межею від щільності розподілу системи за іншою складовою, тобто:

 (5.5)

Приклад. Нехай двовимірна випадкова величина задана щільністю:

.

Знайти *f*1*(x)* і *f*2*(y).*

Використовуючи формули (8.5), знаходимо *f*1*(x)*:

, отже:

.

Аналогічним образом знаходимо *f*2*(y)*:



Визначення. *Математичним сподіванням двовимірної випадкової величини(X,Y)* називається точка *(М(X), М(Y)),* яка визначає центр двовимірного розподілу.

 (5.6)

 (5.7)

**5.4 Умовні закони розподілу**

Розглянемо систему дискретних випадкових величин *(X,Y)*. Припустимо, що в результаті випробування величина *Y* прийняла значення у*1*, при цьому *X* прийме одне зі своїх можливих значень *хi* (*i* = 1, …, *n*)... Позначимо *умовну ймовірність* того, що *X = х1* за умови *Y = у1* через *Р(х1/у1).* Ця ймовірність, власне кажучи, не буде дорівнювати ймовірності *Р(х1)*.

Визначення. *Умовним розподілом складової Х* при *Y = yj* називають сукупність умовних ймовірностей *Р(х1/уj), …, Р(хn/уj),* обчислених у припущенні, що подія *Y = yj* вже настала.

У загальному випадку умовний закон розподілу складової Х визначається співвідношенням:

*Р(хi/уj) = Р(хi,уj)/Р(уj)* (5.8)

Приклад. Дискретна випадкова величина задана таблицею:

|  |  |  |  |
| --- | --- | --- | --- |
| *X*  *Y* | *X1* | *x2* | *x3* |
| *y1* | 0.1 | 0.3 | 0.2 |
| *y2* | 0.06 | 0.18 | 0.16 |

Знайти умовний закон розподілу *Х* при *Y = у1*.

Шуканий закон визначається сукупністю ймовірностей *Р(х1/у1), Р(х2/у1), Р(х3/у1).* Скористаємося формулою умовної ймовірності двох подій *А* и *В: Р(В/А) = Р(АВ)/Р(А)* і, взявши до уваги, що *Р(у1)* = 0.6 – сума першого рядка, одержимо:

*Р(х1/у1) = Р(х1,у1)/Р(у1)* = 0.1/0.6 = 1/6, *Р(х2/у1)* = 1/2, *Р(х3/у1)* = 1/3.

Для перевірки складемо отримані ймовірності: 1/6 + 1/2 + 1/3 = 1.

Нехай *(X,Y)* – безперервна двовимірна випадкова величина.

Визначення.*Умовною щільністю розподілуϕ(х/у)* складової *Х* при даному значенні *Y = у* називається відношення щільності спільного розподілу *f(x,y)* до щільності компоненти *f2(y)*:

*ϕ(х/у) = f(x,y)/f2(y)* (5.9)

Аналогічно визначається *ψ(у/х)*:

*ψ(в/х) = f(x,y)/f1(х)* (5.10)

Якщо відома щільність спільного розподілу, то:

 (5.11)

 (5.12)

Як і будь –яка щільність розподілу, умовна щільність має такі властивості:

1) *ϕ(х/у)*≥ 0, *ψ(у/х)*≥ 0;

2) , .

Визначення.*Умовним математичним сподіванням* дискретної випадкової величини *Y* при *Х = х* називають суму добутків можливих значень *Y* на їхні умовні ймовірності:

 (5.13)

Для безперервної випадкової величини:

 (5.14)

Визначення. *Умовне математичне сподіванняM(Y/x)* є функцією від *хM(Y/x) = f(x)*, що має назву *функція регресії Y* на *Х.*

Аналогічно визначаються умовне математичне сподівання випадкової величини *Х* і функція регресії *Х* на *Y*: *M(X/y) = ϕ(y)*.

Приклад. Дискретна випадкова величина задана таблицею:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Y / X* | *x1* = 1 | *x2* = 3 | *х3* = 4 | *х4* = 8 |
| *y1* = 3 | 0.15 | 0.06 | 0.25 | 0.04 |
| *y2* = 6 | 0.3 | 0.1 | 0.03 | 0.07 |

Знайти умовне сподівання *Y* при *Х* = *х1* = 1.

Знайдемо *р(х1)*, підсумувавши імовірності в першому стовпчику *р(х1)* =0.45. Тепер знайдемо умовний розподіл *Y* при *Х* = *х1* = 1:

*P(y1/x1) = P(x1,y1)/P(x1)* = 0.15/0.45 = 1/3,

*P(y2/x1) = P(x1,y2)/P(x1)* = 0.3/0.45 = 2/3.

Нарешті, знаходимо умовне математичне сподівання:

.

**5.5 Залежні й незалежні випадкові величини**

Теорема. Для того, щоб випадкові величини *Х* и *Y* були незалежними, необхідно й достатньо, щоб функція розподілу системи дорівнювала добутку функцій розподілу складових:

*F(x,y) = F1(x)⋅F2(y)* (5.15)

Доведення необхідності. Нехай *Х* и *Y* незалежні. Тоді події *Х < x, Y < y* незалежні, отже, *P(X < x, Y < y) = P(X < x)⋅P(Y < y)* або *F(x,y) = F1(x)⋅F2(y).*

Доведення достатності. Нехай *F(x,y) = F1(x)⋅F2(y)*, звідси випливає *P(X < x, Y < y) = P(X < x)⋅P(Y < y)*, тобто ймовірність добутку подій *Х < x* і *Y < y* дорівнює добутку ймовірностей цих подій, отже, випадкові події *Х* і *Y* незалежні.

Наслідок. Для того, щоб випадкові величини *Х* і *Y* були незалежні, необхідно й достатньо, щоб виконувалася така рівність:

*f(x,y) = f1(x)⋅f2(y)* (5.16)

Доведення. Цей наслідок зводиться до теореми шляхом інтегрування.Уведемо за аналогією з одномірною випадковою величиною числові характеристики – початкові й центральні моменти для системи двох випадкових величин.

Визначення. Початковим моментом порядку *k*, *s* системи *(X,Y)* називається математичне сподівання добутку *Хk* на *Ys*:

*μk,s = M(Xk⋅Ys)* (5.17)

Визначення. Центральним моментом порядку *k*, *s* системи *(X,Y)* називається математичне сподівання добутку *(X – M(x))k* на *(Y – M(y))s*:

*νk,s = M[(X – M(x))k⋅(Y – M(y))s]* (5.18)

Випишемо формули, що служать для безпосереднього підрахунку моментів. Для дискретних випадкових величин:

 (5.19)

 (5.20)

де 

Для безперервних випадкових величин:

 (5.21)

 (5.22)

Перші початкові моменти являють собою математичні сподівання величин *Х* та *Y*:

*mx = μ1,0 = M(X1Y0) = M(X)* (5.23)

*my = μ0,1 = M(X0Y1) = M(Y)* (5.24)

На практиці також широко застосовуються другі центральні моменти системи, які визначають дисперсії компонентів:

*Дх = ν2,0 = M[(x – mx)2(y – my)0] = M[(x – mx)2]* (5.25)

*Дy = ν0,2 = M[(x – mx)0(y – my)2] = M[(y – my)2]* (5.26)

Особливу роль відіграє другий змішаний центральний момент.

Визначення.*Кореляційним моментом* або *коваріацією* випадкових величин *Х* и *Y* називають математичне сподівання добутку відхилень цих величин:

*Кxy = ν1,1 = M[(x – mx)(y – my)]* (5.27)

На практиці для обчислення коваріації використовуються такі формули:

а) для дискретних випадкових величин:

 (5.28)

б) для безперервних випадкових величин:

 (5.29)

Кореляційний момент служить для характеристики зв'язку між величинами *Х* і *Y*. Легко переконатися, що кореляційний момент можна записати в такому вигляді:

*Кxy = M(XY) – M(X)M(Y)* (5.30)

Теорема. Кореляційний момент двох незалежних випадкових величин дорівнює нулю.

Доведення. Оскільки *Х* і *Y* незалежні випадкові величини, то і їхні відхилення *X – M(X)* і *Y – M(Y)* також будуть незалежними. За властивістю математичного сподівання (математичне сподівання добутку незалежних випадкових величин дорівнює добутку математичних сподівань співмножників) і відхилення (математичне сподівання відхилення дорівнює нулю) одержуємо:

*Кxy = M[(x – mx)(y – my)] = M(x – mx)⋅М(y – my)* = 0.

З визначення кореляційного моменту необхідно, щоб він мав розмірність, рівну добутку розмірностей *Х* і *Y*, що незручно, оскільки важко порівнювати різні системи випадкових величин. Зважаючи на це, вводять нову числову характеристику – коефіцієнт кореляції.

Визначення. *Коефіцієнтом кореляції* випадкових величин *Х* і *Y* називають відношення кореляційного моменту до добутку середніх квадратичних відхилень, тобто:

*rxy = Кxy/(σxσy)* (5.31)

Теорема. Абсолютна величина кореляційного моменту двох випадкових величин *Х* і *Y* не перевищує середнього геометричного їхніх дисперсій:

 (5.32)

Теорема. Абсолютна величина коефіцієнта кореляції не перевищує одиниці:

|*rxy*| ≤ 1 (5.33)

Визначення. Дві випадкові величини називають *корельованими*, якщо їхній коефіцієнт кореляції не дорівнює нулю.

Визначення. *Нормальним законом розподілу на площині* називають розподіл ймовірностей двовимірної випадкової величини (X,Y), для якої:

 (5.34)

З формули видно, що нормальний закон на площині характеризується такими величинами: *а*1, *а*2 – математичними сподіваннями, *σ*x, *σ*y – середньо квадратичними відхиленнями й *rxy* – коефіцієнтом кореляції величин *X* та *Y*. Легко помітити, що у випадку некорельованості випадкових величин *X* і *Y* ці величини й незалежні. Дійсно, якщо *rxy* = 0, то *f(x,y) = f1(x)⋅f2(y)*.

**5.6 Теорія оцінювання. Закон великих чисел (нерівність Чебишева)**

Математичні закони теорії імовірності пов'язані з масовістю явищ, тобто з великим числом однорідних випробувань. Масові явища мають таку властивість: конкретні особливості кожного окремого випадкового явища майже не позначаються на середньому результаті маси таких явищ. Звідси «фізичний зміст» закону великих чисел: при великій кількості випадкових явищ їхній результат не випадковий. Для практики важливе значення умов, при яких результат випадкових причин не випадковий. Ці умови й указуються в теоремах, що носять загальну назву – закон великих чисел.

Лема 1 (Нерівність Чебишева). Імовірність того, що відхилення випадкової величини *Х* від її математичного сподівання за модулем менше додатного числа *ε*, не менше, ніж 1 – *Д(Х)/ε2*:

*P(|X – M(X)| <ε) ≤ 1 – Д(Х)/ε2*

Доведення. За визначенням протилежних подій:

*P(|X – M(X)| <ε) + P(|X – M(X)| ≥ε) =* 1,

звідси випливає:

*P(|X – M(X)|<ε) =* 1 *– P(|X – M(X)| ≥ε)*.

За визначенням, для дискретної випадкової величини:

*P(X) = [x1 – M(X)]2p1 + [x2 – M(X)]2p2 +…+ [xn – M(X)]2pn*> 0.

Відкинемо ті доданки, для яких |*xi – M(X)*| <*ε*. Для визначеності будемо вважати, що ця нерівність виконується для *х1*,…, *хk*. Тоді:

*Д(Х) ≥ [xn+1 – M(X)]2pn+1 + … + [xn – M(X)]2pn*,

причому для кожного *j*: |*xj – M(X)*| ≥*ε*. Зводячи у квадрат, одержимо |*xj – M(X)*|*2*≥*ε2*, отже:

*Д(Х) ≥ε2pn+1 + … + ε2рn = ε2(pn+1 + … + рn).*

За теоремою додавання ймовірностей сума *(pn+1 + … + рn)* є ймовірність того, що *Х* прийме значення *хn+1, …, хn*, для яких |*xj – M(X)*| ≥*ε*. Тобто:

*pn+1 + … + рn = Р(|Х – M(X)| ≥ε)*.

Тоді:

*Д(Х) ≥ε2⋅Р(|Х – M(X)| ≥ε)*

*Р(|Х – M(X)| ≥ε) ≤ Д(Х)/ε2*

*Р(|Х – M(X)| <ε) ≥ 1 – Д(X)/ε2*.

Аналогічне доведення і для безперервних випадкових величин.

Приклад. Дано випадкову величину з математичним сподіванням *mx* і . Оцінити ймовірність того, що |*X – mx*| ≥ 3*σx*.

За нерівністю Чебишева при *ε* = 3*σx* обчислимо *Р*(|*X – mx*| ≥ 3*σx*) ≤*Дх*/(9*σx2*) = 1/9.

Зауваження. Нерівність Чебишева дає тільки верхню границю ймовірності даного відхилення (вище не може бути ні для якого закону розподілу). Ми вже бачили, що для нормального закону: *Р*(|*X – М(Х)*| ≥ 3*σ*) ≈ 0.0027. Підрахуємо цю ймовірність для інших законів:

1). Для рівномірного на [*a, b*]:

*Р*(|*X – М(Х)*| < 3*σ*) = *Р(М(Х) –* 3*σ< X < M(X) +* 3*σ*),

*M(X) = (a + b)*/2, *σ = (b – a)*/2√3.

Доведемо, що *М(Х)* + 3*σ*>*b*, *М(Х)* – 3*σ*<*a*, для чого підставимо значення *М(Х)* і *σ*.

(*a + b*)/2 + 3(*b – a*)/2√3 – *b*> 0 (*a + b*)/2 – 3(*b – a*)/2√3 – *a* < 0

*a + b + b*√3 – *a*√3 – 2*b*> 0 *a + b – b*√3 + *a*√3 – 2*a*< 0

(*b – a*)(√3 – 1) > 0 – очевидно. (*b – a*)(1 – √3) < 0 – очевидно.

За визначенням рівномірного розподілу: *F(mx* + 3*σ)* = 1, *F(mx –* 3*σ)* = 0.

2) Для показникового розподілу:

*Р(|X – М(Х)|*< 3*σ*) = *F(М(Х) +* 3*σ) – F(M(X) –* 3*σ) = F(*1*/λ +* 3*/λ) – F(*1*/λ –* 1*/λ)* = *F*(4/*λ*) = 1 – exp( –*λ*⋅4/*λ*) = 1 – exp( –4) ( 0.018.

Як ми бачимо, значення випадкової величини вкрай рідко виходять за межі *mx*± 3*σ*. Тому застосовують «правило трьох сигм» для будь –якого закону розподілу: абсолютна величина відхилення від математичного сподівання не перевищує потроєного середнього квадратичного відхилення.

**5.7 Теорема Чебишева**

Вирішимо допоміжну задачу.

Задача. Дано випадкову величину *Х* з математичним сподіванням *mx* і дисперсією *Д(Х).* Над цією випадковою величиною проводиться *n* незалежних випробувань і знаходиться середнє арифметичне результатів. Знайти числові характеристики цього середнього арифметичного.

Нехай *хi* – значення величини *Х* у *i –ому* випробуванні. Кожна з величин *хi* – випадкова величина, розподілена за тим же законом, що й *Х.* Розглянемо *Y=∑xi/n*. За визначенням математичного сподівання й дисперсії маємо:

*My = M[Y] = ∑M[xi]/n = mxn/n = mx*

*Дy = ∑Д[xi]/n2 = Дx/n*

Таким чином, математичне сподівання не залежить від числа випробувань, дисперсія ж необмежено спадає й може бути як завгодно малою, звідки випливає, що середнє арифметичне при великій кількості спроб – невипадкова величина.

Теорема (Чебишева). При достатньому числі незалежних випробувань середнє арифметичне спостережуваних значень сходиться за ймовірністю до математичного сподівання.

Визначення. Говорять, що випадкова величина *хn сходиться за ймовірністю до величини а,* якщо  або для кожного *ε, δ*> 0 виконується наступна нерівність: *Р(|xn – a| <ε)* > 1 – *δ*.

Тоді теорема Чебишева набуває вигляду:



Доведення. Застосуємо до випадкової величини *Y* з попередньої задачі нерівність Чебишева:

*Р(|Y – my| ≥ε) ≤ Дy/ε2 = Дx/nε2*

Яким би не було малим *ε*, можна взяти таке велике *n*, щоб виконувалася нерівність: *Дx/nε2<δ,* де *δ* – як завгодно мале число. Тоді:

 або .

Теорема (узагальнена Чебишева). Якщо *х1, …, хn* незалежні випадкові величини з математичними сподіваннями *mx1, …, mxn* і дисперсіями *Дx1*, …, *Дxn*, причому всі дисперсії обмежені зверху: *Дxi*<*L, i* = 1, …, *n,* то при зростанні *n* середнє арифметичне спостережуваних значень величин *х1*, …, *хn*сходиться за ймовірністю до середнього арифметичного їх математичних сподівань або для кожного *ε, δ*> 0 існує таке *n*, що виконується наступна нерівність:

 (5.35)

Доведення. Розглянемо випадкову величину *Y = ∑xi/n*, її математичне сподівання й дисперсія обчислюються за формулами: *my = ∑mxi/n* і *Дy = ∑Дxi/n2*. Застосуємо нерівність Чебишева:

*P(|Y – my| ≥ε) ≤ Ду/ε2* або .

Яким би не було малим *ε*, існує таке *n*, що: *L/(nε2) <δ*, тоді:

.

Переходячи до протилежної події, одержуємо умову теореми – формулу (5.35).

Розглянемо кілька наслідків з теореми Чебишева.

Теорема (Бернуллі). При необмеженому збільшенні числа випробувань *n* частота події *А* сходиться за ймовірністю до її ймовірності *р*:

*P(|P\* – p| <ε)*> 1 – *δ*

Доведення. Розглянемо незалежні випадкові величини *хi* – число появи події *А* в *i –ому* випробуванні. Закон розподілу:

|  |  |
| --- | --- |
| 0 | 1 |
| *q* | *p* |

Частота *Р\** являє собою середнє арифметичне *x1*, …, *xn : P\* = ∑xi/n* і за законом великих чисел сходиться за ймовірністю до загального математичного сподівання цих випадкових величин *р.*

Теорема (Пуассона). Якщо проводиться *n* незалежних випробувань і ймовірність появи події *А* в *i –ому* випробуванні дорівнює *рi*, то при збільшенні *n* частота події *А* сходиться за ймовірністю до середнього арифметичного ймовірностей *рi*.

Узагальнення закону великих чисел на випадок залежних величин належить Маркову.

Теорема (Маркова). Якщо є залежні випадкові величини *х1*, …, *хn* і якщо при *n→∞ Д[∑xi]/n2*→0, то середнє арифметичне значення спостережуваних випадкових величин *х1, …, хn* сходиться за ймовірністю до середнього арифметичного їх математичних сподівань.

* 1. **Центральна гранична теорема**

Всі форми закону великих чисел стверджують одне: факт збіжності за ймовірністю тих або інших випадкових величин до певних постійних. У жодній з форм закону великих чисел ми не маємо справи із законами розподілу випадкових величин. Граничні закони розподілу є предметом іншої групи теорем – центральної граничної теореми.

Теорема (Ляпунова). Якщо *х1, …, хn* – незалежні випадкові величини, що мають однаковий закон розподілу з математичним сподіванням *m* і дисперсією *σ2*, то при необмеженому збільшенні числа *n* закон розподілу їхньої суми *Yn = ∑xk* необмежено наближається до нормального.

Доведемо наслідок цієї теореми, відомий як теорема Лапласа.

Теорема (Лапласа). Якщо проводиться *n* незалежних випробувань, у кожному з яких подія *А* з'являється з імовірністю *р,* то справедливе співвідношення:



де *Υ* – число появи події *А* в *n* випробуваннях, *q* = 1 – *p*,  – нормальна функція розподілу (функція Лапласа).

Доведення. Представимо випадкову величину *Y* у вигляді суми: *Y =∑xi,* де *xi*– число появ події *А* в *i –ому* випробуванні. У силу центральної граничної теореми при великій кількості випробувань *n* закон розподілу *Y* близький до нормального. Тоді випадкова величина *Z = (Y – my)/σy* розподілена за нормальним законом *N*(0, 1) і справедливою є формула *Р(α< z <β) = Ф(β) –Ф(α)*. У силу доведеного вище *my = np*, а *Дy = npq*. Отже, підставляючи значення *z, my, Дy* одержимо формулу (6.7).

Приклад. По смузі укріплень скидається 100 серій бомб. При скиданні однієї серії математичне сподівання числа влучень дорівнює 2, середнє квадратичне відхилення *σ* = 1,5. Знайти приблизно ймовірність того, що в 100 серіях у смугу потрапить від 180 до 220 бомб.

Представимо загальне число влучень як суму влучень у окремих серіях. Умови центральної граничної теореми дотримані, тому що *xi* розподілені однаково. Будемо вважати, що *n* = 100 досить велике, щоб застосовувати центральну граничну теорему (на практиці застосовні й при набагато менших *n*). Маємо: *mx = ∑mi* = 200, *∑Дi* = ∑1.52 = 225. За формулою (9.7) одержимо ймовірність: *P*(180 <*x*< 220) = *Ф*((220 – 200)/√225) – Ф((180 –200)/√225) ≈ 0.82.

Питання для самоконтролю

1. Сформулюйте поняття багатовимірної і двовимірної випадкових величин, а також їхню геометричну інтерпретацію.
2. Дайте визначення поняття закону розподілу двовимірної дискретної випадкової величини.
3. Як за законом розподілу двовимірної випадкової величини (*X*; *Y*) знайти одномірні закони розподілу випадкової величини*X* і *Y*?
4. Що називається умовним законом розподілу випадкової величини*X*(*Y*) за умови *Y*=*yj* (*X*=*xi*)?
5. Що називається функцією розподілу двовимірної випадкової величини (*X*; *Y*)? Дайте геометричну інтерпретацію.
6. Сформулюйте властивості функції розподілу.
7. Як визначаються одномірні(маргінальні) функції розподілу випадкової величини*X* і *Y*?
8. Як обчислити *P* (*a1 ≤ X ≤ a2*; *b1 ≤ Y ≤ b2*)?
9. Що називається щільністю ймовірностей *f* (*x*; *y*) двовимірної випадкової величини (*X*; *Y*)?
10. Який зв'язок існує між *f* (*x*; *y*) і *F* (*x*; *y*)?
11. Сформулюйте властивості щільності ймовірностей *f* (*x*; *y*).
12. Як визначаються маргінальні щільності ймовірностей випадкової величини*X* і *Y*?
13. Які випадкової величини*X* і *Y* вважаються незалежними, а які – залежними?
14. Дайте визначення поняття умовної функції розподілу ймовірностей випадкової величини*X*(*Y*) за умови, що випадкової величини*Y*=*y* (*X*=*x*).
15. Дайте визначення поняття умовної щільності ймовірностей випадкової величини*X*(*Y*) за умови, що випадкової величини*Y*=*y* (*X*=*x*).
16. Як визначається умовне МС *X*(*Y*) за умови, що випадкової величини*Y*=*y* (*X*=*x*)? випадкової величини.
17. Дайте визначення початкового та центрального моменту двовимірної випадкової величини.
18. Що визначає коваріація випадкових величин*X* і *Y*?
19. Наведіть формули для *К xy*.
20. Сформулюйте властивості коваріації.
21. Як визначається коефіцієнт кореляції?
22. Які залежності між випадковими величинами *X* і *Y* називаються некорельованими, а які – корельованими?
23. Сформулюйте властивості коефіцієнта кореляції.
24. Що називається лінійною середньо квадратичною регресією?
25. Дайте визначення центра спільного розподілу двовимірної випадкової величини.
26. Сформулюйте властивості *КХУ*.
27. Який вигляд має щільність ймовірностей для двовимірного нормального розподілу?
28. Сформулюйте основні властивості двовимірного нормального розподілу.Сформулюйте закон великих чисел у широкому й вузькому розумінні (фізичний і практичний зміст).
29. Сформулюйте нерівність Чебишева.
30. Дайте визначення поняття збіжності послідовності випадкової величини за ймовірністю.
31. Сформулюйте теорему Чебишева.
32. Сформулюйте теорему Маркова.
33. Сформулюйте теорему Бернуллі.
34. Сформулюйте теорему Пуассона.
35. Яка роль граничних теорем у ТЙ?

Практичні завдання

***Завдання 1.****Основи статистичної обробки неперервних даних.*

*Мета:* дослідити вплив обсягів експериментальних даних на емпіричні закони розподілу ймовірностей випадкових величин.

Задатись двома вибірками із 100+N елементів неперервних випадкових величин нормального, показникового та рівномірного розподілу в діапазоні від 0 до N включно (N – номер студента у журналі).

Побудувати гістограми. Для побудови графіків рекомендовано використовувати стандартні бібліотеки.

***Завдання 2.***

Для наданих статистичних розподілів вибірки, записати емпіричну функцію розподілу та обчислити такі числові характеристики: 1) вибіркове середнє, 2) вибіркову дисперсію, 3) підправлену дисперсію, 4) вибіркове середнє квадратичне відхилення, 5) підправлене середнє квадратичне відхилення, 6) розмах вибірки, 7) медіану, 8) моду, 9) квартильне відхилення, 10) коефіцієнт варіації, 11) коефіцієнт асиметрії та 12) ексцес для вибірки.

Вибірка 1

|  |
| --- |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 55 | 46 | 44 | 52 | 52 | 47 | 63 | 51 | 53 | 49 | 51 | 50 | 55 |
| 33 | 57 | 47 | 40 | 53 | 34 | 30 | 54 | 54 | 51 | 65 | 31 | 46 |
| 27 | 49 | 33 | 89 | 19 | 70 | 10 | 25 | 35 | 52 | 32 | 89 | 76 |
| 40 | 85 | 31 | 33 | 10 | 55 | 15 | 24 | 39 | 32 | 15 | 45 | 26 |

Вибірка 2

|  |
| --- |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 147 | 150 | 160 | 162 | 157 | 169 | 142 | 111 | 121 | 127 | 136 | 194 | 101 |
| 153 | 149 | 150 | 149 | 135 | 146 | 139 | 132 | 111 | 131 | 192 | 178 | 195 |
| 148 | 163 | 162 | 141 | 128 | 144 | 160 | 172 | 158 | 184 | 152 | 109 | 196 |
| 155 | 142 | 140 | 146 | 168 | 131 | 152 | 152 | 148 | 121 | 157 | 115 | 198 |

***Завдання 3.*** На основі даних спостереження за часом реакції на слухове подразнення побудувати інтервальний розподіл частот, визначити абсолютні і відносні показники розподілу та знайти ранги ряду чисел.

Отримані наступні результати спостереження за 300 учнями (час реакції на слухове подразнення в тис. долях секунди) :

Час: 106 107 110 112 115 116 118 120 122  125 126 128 129 130 132 135

Чис-ть: 8 23 15 22 20 23 18 20 19 21 25 16 18 18 16 18

Побудувати інтервальний розподіл з інтервалом 5, визначити частоти та знайти ранги ряду чисел.

***Завдання 4.*** Згенерувати вибірку:

* що розподілена за рівномірним законом на інтервалі [a,b];
* що розподілена за показниковим законом для наданого λ;
* що розподілена за нормальним законом.

2. Побудувати гістограму частот для кожної вибірки.

3. На підставі наведених вибіркових даних перевірити випадковість наданої вибірки за допомогою:

* критерію серій;
* критерію зростаючих та спадаючих серій.

4. Для вибірки, якабула генерована за нормальним законом, виконати аналіз спостережень, що різко виділяються:

* побудувати множину наданих даних у вигляді точок (і, хі);
* знайти оцінки числових характеристик вибіркової сукупності: математичне сподівання, дисперсію, середнє квадратичне відхилення;
* видалити спостереження, що різко виділяються, за допомогою критерію Стьюдента.

*Вказівки до завдання*

1. Генерація рівномірно розподіленої випадкової вибірка на відрізку [0,1].

Велике число алгоритмів генерації псевдовипадкової вибірки мають наступний вид:

1. вибирається натуральне число ;
2. визначається послідовність чисел  по формулі:



де *M* – деяке ціле число (велике), (*А*) mod (*В)*  – залишок від ділення *А* на *В*. Можна прийняти *M*=515, *m*=36.

1. Числа – псевдовипадкові на відрізку [0,1].

2. Для рівномірного закону розподілу чисел *Х* з показником на проміжку [*a,b*] за основу беруться псевдо випадкові числа *S* згенеровані на відрізку [0,1] і застосовується формула:

.

3. Для показового закону розподілу чисел *Х* з показником *λ* за основу беруться псевдо випадкові числа *S* згенеровані на відрізку [0,1] і застосовується формула:



4. Для отримання випадкових чисел, розподілених за стандартним нормальним законом *N*(0,1) можна використовувати наступний прийом: нехай є послідовність n випадкових чисел *S1,S2,...,Sn*, що належить проміжку [0,1], тоді для отримання псевдо випадкової вибірки, розподіленої за нормальним законом, використовують таку формулу:



Більш ефективним є метод перетворення Бокса –Мюллера. Для отримання чисел, що відповідають стандартному нормальному закону розподілу необхідно:

1) Згенерувати псевдо випадкові числа *r* і на інтервалі [0,1].

2) Обчислити два числа, що відповідають стандартному нормальному закону розподілу,  і по формулах:





Після отримання стандартної нормальної випадкової величини *z*, легко перейти до нормально розподіленої величини з довільним математичним очікуванням *a* і стандартним відхиленням по формулі:

.

***Завдання 5.***

Ряд розподілу дискретної двовимірної випадкової величини має вигляд

|  |  |  |  |
| --- | --- | --- | --- |
| X\Y | 0 | 1 | 4 |
| 1 | 0.1 | 0.2 | 0.08 |
| 2 | 0.12 | 0.3 | 0.2 |

Побудувати функцію розподілу випадкової величини та ряди розподілу її компонент.

***Завдання 6.***

Дискретна двовимірна випадкова величина задана законом розподілу

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X\Y | 1 | 3 | 4 | 8 |
| 3 | 0,15 | 0.06 | 0.25 | 0.04 |
| 6 | 0.3 | 0.1 | 0.03 | 0.07 |

Побудувати ряди розподілу компонент та знайти числові характеристики компонент: математичне сподівання, дисперсію та середньо квадратичне відхилення.

***Завдання 7.***

Задано функцію розподілу двовимірної неперервної випадкової величини:



Знайти: а) ймовірність потрапляння випадкової величини в прямокутник, обмежений прямими 

б) щільність сумісного розподілу.

***Завдання 8.***

Задано функцію щільності розподілу системи випадкових величин:



Знайти функцію розподілу системи та функції розподілу її компонент.

Тестові завдання для перевірки знань

Виберіть, принаймні, одну відповідь.

1. При формуванні вибірки з поділом генеральної сукупності на частини застосовують наступні види відборів:

а) варіантний, б) механічний, в) серійний, г) типовий.

2. Вибірка, відсортована в зростаючому порядку, називається:

а) варіаційним рядом, б) інтервальним рядом, в) статистичним рядом,

г) емпіричним рядом.

3. Ламана, що з'єднує точки з координатами (*xk, nk*), називається

а)емпіричною функцією розподілу, б)полігоном частот, в)полігоном частостей, г) гістограмою.

4. Статистичні оцінки параметрів розподіли повинні задовольняти наступним вимогам:

а) ґрунтовність, б) репрезентативність, в)ефективність, г) незміщеність.

5. Число, що ділить варіаційний ряд на дві рівні частини, є оцінкою для:

а) моди, б) середньо квадратичного відхилення, в) медіани,

г) математичного сподівання.

6. Гіпотези можуть бути розділені попарно на наступні групи:

а) основні й альтернативні, б)інтервальні й точкові, в) параметричні й непараметричні, г) прості й складені.

7. При побудові довірчого інтервалу для деякого параметру невідомого розподілу необхідно використати:

а) розмах вибірки, б) величину об'єму вибірки, в) рівень значимості,

г) потужність критерію.

8. Для перевірки гіпотези про рівність дисперсій двох генеральних сукупностей використовують критерій

а)Стьюдента, б)Фішера –Снедекора, в)Колмогорова, г) –квадрат (Пірсона).

9. Залежність двох випадкових величин може бути:

а) статистичною, б) функціональною, в) емпіричною, г) кореляційною.

10. Метод найменших квадратів служить для:

а)перевірки гіпотези про нормальність закону розподілу,

б)перевірки залежності двох випадкових величин,

в)перевірки параметричних гіпотез,

г)знаходження вибіркового коефіцієнта регресії.

11. Для того, щоб випадкові величини були незалежні, вибірковий коефіцієнт кореляції повинен:

а) дорівнювати нулю,

б) приблизно дорівнювати нулю, в) бути додатним,

г) бути близьким до одиниці.

12. Задачами математичної статистики є:

а)розробка способів збору статистичних відомостей, б)знаходження оцінок параметрів розподілу, в)перевірка статистичних гіпотез, г)побудова функціональних залежностей.

13. Нехай задана наступна вибірка: 1, 3, 3, 1, 5, 1, 3, 1. Оцінка для медіани буде дорівнювати: а) 2, б) 1, в) 4, г) 3.

14. Якщо при перевірці параметричної гіпотези в якості альтернативної висувається гіпотеза виду *H1: M(X) > a*, то критична область буде:

а)правосторонньою, б) залежати від інших умов, в)двосторонньою, г)лівосторонньою.

15. Задано дві вибірки об'ємом *n1*=12 і *n2*=11 по яких знайдені оцінки для дисперсій *D1* = 6 і *D2* = 8. Рівень значимості 0.1. Для перевірки гіпотези про рівність дисперсій генеральних сукупностей при конкуруючою гіпотезі Н1: "дисперсії не рівні" критична точка буде дорівнювати

а) *F*(0.1, 10, 11), б) *F*(0.05, 11, 10), в) *F*(0.05, 10, 11), г) *F*(0.1, 11, 10).

16. Виберіть із наведених нижче ті числа, які є числами Вестергарду:

а) 1, б) 3, в) 0.1, г) 0.3.

# ЗМІСТОВИЙ МОДУЛЬ 3. АНАЛІЗ МЕТОДІВ ПЕРЕВІРКИ РЕЛЕВАНТНОСТІ ЕМПІРИЧНИХ ДАНИХ

## Тема 6. Параметричні методи перевірки гіпотез

Мета: розглянути основні поняття теорії параметричних методів перевірки гіпотез, набути навики побудови критичної області для перевірки статистичних гіпотез.

🖉Основні терміни і поняття

*Номінальні ознаки, порядкові ознаки, кількісні ознаки, точкові оцінки, критична область, критична точка, основнагіпотеза,альтернативна гіпотеза, рівень значимості, статистичний критерій.*

📚Основні теоретичні положення

**6.1 Основні поняття теорії гіпотез**

Успішність застосування будь-якого методу аналізу даних залежить від відповідності аналізованих даних його вихідним припущенням. Методи, придатні для одного типу даних, можуть призводити до серйозних помилок при їх використанні для даних інших типів.

Першим етапом аналізу будь-яких даних є визначення їх типу. Основною є класифікація даних за шкалами їх вимірювання. Розрізняють такі типи ознак:

1. *Номінальні ознаки (ознаки з невпорядкованими станами, класифікаційні ознаки*) – це дані, що вимірюють в номінальній шкалі (класифікаційній, шкалі найменувань). Найменування класів можуть бути виражені за допомогою чисел, але ці числа можуть використовуватися лише для відповіді на питання: належать два об’єкти до одного класу чи ні. Прикладами номінальних ознак є назви біологічних видів, назви навчальних дисциплін, кольори тощо. З погляду автоматизації аналізу даних і застосування стандартних алгоритмів доцільно обирати такі позначення класів: 0, 1, 2, ... Але з цими числами не можливо виконувати будь-які дії, крім перевірки їх рівності або нерівності.

*Порядкові ознаки (ознаки з упорядкованими станами, ординальні ознаки)* – це дані, що вимірюють в порядкових шкалах. Ці дані можуть порівнюватися між собою у певному відношенні: «більше – менше», «легше – важче», «правіше – лівіше» тощо. Прикладами порядкових ознак є сила землетрусу, військові звання, оцінки студентів тощо. Якщо значення порядкової ознаки є числами, то вони можуть застосовуватися і для порівняння ступеня вияву класифікаційної ознаки, але відстані між класами при цьому будуть не визначені.

*Кількісні (числові, варіаційні) ознаки*– це ознаки, які вимірюють у кількісних (інтервальних, відносних, циклічних та абсолютних) шкалах вимірювань. Дії, що можуть виконуватися з числовими характеристиками даних, залежать від шкали вимірювань.

В узагальненому вигляді характеристики основних типів даних наведено в таблиці 6.1.

Таблиця 6.1Характеристики основних типів даних

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Шкала  вимірювань | Визначальні  відношення | Еквівалентні  перетворення | Допустимі операції над даними | |
| Первинна  обробка | Вторинна  обробка |
| Номінальна | Еквівалентність | Перестановки  найменувань | Обчислення символу Кронекера | Обчислення відносних частот та операції над ними |
| Порядкова | Еквівалентність,  перевага | Монотонні (такі, що не змінюють порядку) | Обчислення та рангів | Обчислення відносних частот та квантилів, операції над ними |
| Інтервальна | Еквівалентність, перевага, збереження відношення інтервалів | Лінійне перетворення | Обчислення  рангів та інтервалів (різниць між даними) | Арифметичні дії над інтервалами |
| Циклічна | Еквівалентність, перевага, збереження відношення інтервалів, періодичність | Зсув | Обчислення  рангів та інтервалів (різниць між даними) | Арифметичні дії над інтервалами |
| Відношень | Еквівалентність, перевага, збереження відношення інтервалів, збереження відношення двох значень | Розтягання | Усі арифметичні операції | Будь-яка придатна обробка |
| Абсолютна | Еквівалентність, перевага, збереження відношення інтервалів, збереження відношення двох значень, абсолютна й безрозмірна одиниця, абсолютний нуль | Не існує (шкала є унікальною) | Усі арифметичні операції, використання як показника степеня, основи та аргументу логарифма | Будь-яка потрібна обробка |

Дані, отримані у шкалах вищих рангів, можуть приводитися до шкал нижчих рангів. Наприклад, дані, що виміряні у шкалі відношень, можна привести до інтервальної шкали. Такі перетворення називають *зниженням шкали.* Необхідність у них зазвичай виникає при обробці даних, що виміряні у шкалах різного типу. Зворотну операцію – перетворення даних, що виміряні у нижчих шкалах, до вищих – вважають некоректною. Зниження шкали призводить до втрати частини наявної інформації про досліджувані ознаки.

Важливими типами класифікації є поділ ознак за дискретністю або неперервністю теоретичної функції розподілу, законом розподілу тощо.Як характеристики вибірки можна використовувати точкові та інтервальні оцінки.

*Точковими оцінками*параметрів вибірки називають такі оцінки, що визначаються одним числом. Прикладами таких оцінок є середні арифметичні й медіани вибірок. При малих обсягах вибірок, а також при їх значному відхиленні від нормального закону розподілу точкові оцінки можуть істотно відхилятися від істинних значень оцінюваних параметрів. Тому поряд з ними використовують інтервальні оцінки параметрів.

*Інтервальні оцінки*визначаються двома числами – межами інтервалу, до якого із заданою ймовірністю потрапляє оцінюваний параметр.

Побудова точкових і інтервальних оцінок є, як правило, попередньою формою статистичного аналізу. Основне завдання математичної статистики – вміти на основі точкових і інтервальних оцінок приймати рішення в умовах невизначеності, тобто вирішувати статистичні гіпотези. Гіпотези бувають параметричні й непараметричні. Якщо закон розподілу випадкової величини відомий, а необхідно висунути припущення тільки про деякі параметри цього розподілу, то гіпотеза називається *параметричною.* Якщо ж висловлюється припущення про невідомий закон розподілу, то гіпотеза називається *непараметричною.*

Гіпотези бувають *основні* або нуль – гіпотези (*Н0*) і *альтернативні* (*Н1*). Нульова гіпотеза завжди одна, а альтернативних може бути декілька. Якщо *Н0* відкидається, то має місце протилежна їй гіпотеза *Н1*.

Приклад. Непараметрична гіпотеза: «Генеральна сукупність розподілена за законом Пуассона»; параметрична гіпотеза: «Дисперсії двох нормальних сукупностей однакові». Основна гіпотеза: «Математичне сподівання дорівнює 10»; конкуруючі: а) «Математичне сподівання менше 10», б) «Математичне сподівання більше 10», в) «Математичне сподівання не дорівнює 10».

Визначення.*Простою* називають гіпотезу, що містить тільки одне припущення (наприклад, *Н0*: *λ* = 5). *Складною або складовою* називають гіпотезу, що складається з кінцевого або нескінченного числа простих гіпотез (наприклад, *Н0*: *λ*> 5).

Висунута гіпотеза може бути правильною або помилковою, тобто виникає необхідність її перевірки статистичними методами. При цьому можуть бути допущені наступні помилки:

1) помилка першого роду – відкинута правильна гіпотеза. Імовірність цієї помилки *Р(Н1/Н0) = α*;

2) помилка другого роду – прийнята неправильна гіпотеза. Імовірність такої помилки *Р(Н0/Н1) = β*.

Імовірність допустити помилку першого роду позначають через *α* – *рівень значимості.* Другою важливою характеристикою критерію є його *потужність* – величина *γ* = 1 – *β*.

**6.2 Побудова критичної області**

Для перевірки гіпотези *Н0* використовують спеціально підібрану випадкову величину, розподіл якої відомий. Цю величину позначають по –різному залежно від її розподілу: *U* або *Z* при нормальному розподілі, *Т* – при розподілі Стьюдента й т.і.

Визначення.*Статистичним критерієм* або просто *критерієм* називають випадкову величину *К*, що служить для перевірки гіпотези *Н0*. *Спостережуваним значенням Кспос* називають значення критерію, обчислене за вибіркою.

Визначення. Множина всіх можливих значень критерію розбивають на дві області:

1) область прийняття гіпотези (*Н0* вірна);

2) критична область (*Н0* помилкова).

*Критичними точками Ккр* називають точки поділу цих областей.

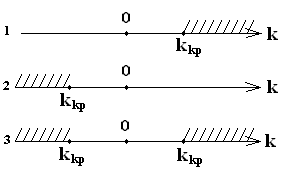


Рисунок 6.1 – Графічна демонстрація критичних областей

Прямі 1 і 2 відповідають однобічним критичним областям – правосторонній й лівосторонній відповідно, пряма 3 – двосторонній критичній області.

Алгоритм перевірки параметричних гіпотез.

1. Висловлюється припущення про закон розподілу.
2. Відбирається вибірка.
3. Формуються основна (*Н0*) і конкуруюча (*Н1*) гіпотези.
4. Будується статистика, що зв'язує досліджуваний параметр і вибіркові значення зі статистичним критерієм. Наприклад,

 для *N*(0; 1) (6.1)

 для *tα(f)*

 для 

1. Для обраної статистики будується довірчий інтервал з імовірністю 1 – *α* виду

*P( –U1 –α/2< U ≤ U1 –α/2)* = 1 – *α* (6.2)

*P( –∞< U ≤ U1 –α)* = 1 – *α* (6.3)

Формула (6.2) відповідає двосторонній критичній області, а формула (6.3) – односторонній.

1. Довірчий інтервал перетворюється в інтервал для досліджуваного параметра, наприклад,

(* – U1 –α/2σ/√n ≤ m < + U1 –α/2σ/√n*) = 1 – *α* (6.4)

1. Ухвалення рішення на основі побудованого інтервалу. Якщо *Кспос* потрапляє в область прийняття гіпотези, то немає підстави відкидати основну гіпотезу *Н0*. Якщо ж не попадає, то приймається конкуруюча гіпотеза *Н1*. При цьому ймовірність помилки першого роду дорівнює рівню значимості *Р(Н1/Н0) = α*.

Приклад. Виготовляється сталь, міцність якої за ДСТУ 100 кг/см2. За результатами плавок виміри показали 98; 100; 96; 102; 98. Чи відповідає сталь вимогам ДСТУ?

Нехай міцність є нормально розподіленою випадковою величиною з невідомим математичним сподіванням і середньо квадратичним відхиленням. Перевіримо гіпотезу *Н0*: *m* = 100 кг/см2 при конкуруючій гіпотезі *Н1*: *m*≠ 100. За вибіркою знаходимо середню  = 98.8 і оцінку для дисперсії *S2* = 5.2. За таблицею знаходимо квантиль на рівні значимості *α* = 0.05  За цими даними будуємо критичні точки , тобто 98.8 ± 2.776\*2.28/√5 і інтервал, для якого сталь відповідає вимогам ДСТУ на рівні значимості 0.05, дорівнює (95.97; 101.63). Оскільки 100 потрапляє в отриманий інтервал, то немає підстав вважати, що сталь не відповідає нормам ДСТУ.

**6.3 Приклади перевірки гіпотез**

Розглянемо ряд типових задач.

Задача 1. Ця задача виникає при необхідності вибору більш точного приладу, тобто приладу з меншим розсіюванням показань. Висувають гіпотезу про рівність дисперсій *H0: Д(Х) = Д(Y).*

1 –й випадок. Конкуруюча гіпотеза *H1: Д(Х) < Д(Y)* (перший прилад гірше). Як критерій приймають наступну величину:

*F = S2більше/S2менше* (6.5)

Ця величина має розподіл Фішера – Снедекора зі ступенями волі *k1 = n1* – 1, *k2 = n2* – 1, де *n1*, *n2* – об'єми вибірок з більшою й меншою дисперсіями відповідно. Будуємо правосторонню критичну область:

*P[F > Fkp(α; k1; k2)] = α* (6.6)

Звідси область прийняття гіпотези *H0* буде мати вигляд:

*P[F < Fkp] = 1 – α* (6.7)

Приклад. Нехай дано дві вибірки об'ємом *n1* = 12 і *n2* = 15, *S2X* = 11.41, *S2Y* = 6.52. Рівень значимості *α* = 0.05. Необхідно перевірити гіпотезу *H0*: *Д(Х) = Д(Y)* при конкуруючій *H1: Д(Х) < Д(Y).*

За формулою (6.5) знаходимо вибіркове значення критерію *Fнабл* = 11.41/6.52 = 1.75, за таблицею розподілу Фішера – Снедекора за заданою кількістю ступенів волі знаходимо *Fkp*(0.05; 11; 14) = 2.56. Оскільки *Fнабл < Fkp*, то нема підстав відкидати гіпотезу *H0*.

2 –ий випадок. При основній гіпотезі *H0: Д(Х) = Д(Y)* конкуруючою виступає гіпотеза про нерівність дисперсій *H1: Д(Х) ≠ Д(Y)*. У цьому випадку будують двосторонню критичну область. Оскільки в таблиці немає лівосторонніх точок (значення критерію завжди додатне), то знаходимо значення при зменшеному вдвічі рівні значимості: *Fkp*(*α*/2; *k1; k2*) і при *Fнабл < Fkp* не відкидаємо гіпотезу *Н0*, а при *Fнабл > Fkp* – відкидаємо.

Приклад. *n1*= 10, *n2*= 18, *S2X*= 1.23, *S2Y*= 0.41, *α* = 0.1.

Перевіримо *H0: Д(Х) = Д(Y)* при *H1: Д(Х) ≠ Д(Y)*. Знаходимо значення *Fнабл* = 3 і *Fkp*(0.05; 9; 17) = 2.5. Оскільки *Fнабл> Fkp*, то відкидаємо гіпотезу *Н0*.

Задача 2. Нехай потрібно встановити значимо або незначимо розрізняються *S2* і *σ2* (збіг теоретичної й практичної похибки вимірів). Критерієм перевірки є випадкова величина *χ2 = (n – 1)S2/σ2*, що має хі –квадрат розподіл. Ця задача виникає при порівнянні точності приладу з деяким еталонним значенням.

1 –ий випадок. Перевіримо *Н0*: *S2 = σ2* у припущенні, що реальний розкид вище: *Н1: S2>σ2*. Будується правостороння критична область:

*P[χ2>χ2kp(α; k)] = α,* (6.8)

де число ступенів волі *k = n* – 1.

2 –ий випадок. *Н0: S2 = σ2* і розглядається двостороння критична область *Н1: S2≠σ2*. Тоді областю прийняття гіпотези *Н0* буде інтервал:

*P[χ2(*1 – *α*/2; *k) ≤χ2 <χ2(α*/2*; k)]* = 1 – *α* (6.9)

3 –ій випадок. Як конкуруючу гіпотезу приймемо припущення, що вибіркова дисперсія менше *Н0: S2 = σ2, Н1: S2<σ2*. Одержимо лівосторонню критичну область, для якої область прийняття гіпотези *Н0* дорівнює наступному виразу:

*P[χ2 >χ2(*1 – *α; k)]* = 1 – *α* (6.10)

Приклад. З нормальної генеральної сукупності витягнута вибірка об'ємом *n* = 13 і знайдена вибіркова дисперсія *S2* = 10.3. При *α* = 0.02 перевірити гіпотезу *Н0: S2 = σ2* = 12 при *Н1: S2≠σ2* = 12.

Обчислимо спостережуване значення критерію за формулою:

*χ2набл = (n*  – 1*)\*S2/σ2* = 10.3 (13 – 1)/12 = 10.3

За таблицею на заданому рівні значимості й відповідно числу ступенів волі знайдемо дві критичні точки, що задають область прийняття гіпотези:

*χ2(*1 – *α/2; k)* = 3.57 і *χ2(α/2; k)* = 26.2.

Оскільки значення 10.3 належить цьому інтервалу, то нема підстав відкидати гіпотезу *Н0*.

Задача 3. Порівняння двох середніх. Залежно від виду розподілу й об'єму вибірок розрізняють кілька випадків.

1 –ий випадок. Випадкові величини розподілені за нормальним законом з відомими дисперсіями. Для перевірки використовують наступний критерій:

 (6.11)

Критичні точки знаходять за таблицею функції Лапласа. Залежно від виду конкуруючої гіпотези, інтервали будуються односторонніми або двостороннім:

1. При Н1: *М(х) ≠ М(у)* двосторонні критичні точки знаходять із рівності *Ф(Zкрит) = (1 – α)/2*. Область прийняття гіпотези *Н0* задається нерівністю

 (6.12)

1. У випадку *Н1: М(х) > М(у)* або *Н1: М(х) < М(у)* односторонні критичні точки знаходяться при подвоєному рівні значимості, тобто *Ф(Zкрит) =(1 – 2α)/2*. Областю прийняття гіпотези є, відповідно, область:

*Zнабл< Zкрит або Zнабл> –Zкрит* (6.13)

2 –ий випадок. Випадкові величини розподілені за довільним законом. У випадку, коли об'єм кожної з вибірок не менш тридцяти елементів, то, відповідно до закону великих чисел, вибіркові середні розподілені нормально. Для перевірки використовують формули з першого випадку.

3 –ій випадок. Випадкові величини розподілені за нормальним законом, *Д(Х)* та *Д(Y)* невідомі, тобто розглядаються малі незалежні вибірки. У загальному випадку задачу розв’язати не можливо, але якщо припустити, що дисперсії невідомі, але рівні між собою (наприклад, порівнюються середні розміри двох партій деталей, виготовлених на одному верстаті), то можна побудувати критерій Стьюдента порівняння середніх:

1) Спочатку перевіримо гіпотезу *Н0: М(Х) = М(Y)* при конкуруючій *Н1*: *М(Х) ≠ М(Y).* У всіх випадках у якості критерію будемо приймати наступну випадкову величину:

 (6.14)

Якщо *Н0*справедлива, то статистика *Т* має розподіл Стьюдента з *f = n + m –* 2 ступенями волі. За таблицею знаходимо *tдвостор.кр.(α; f)*. Якщо виконується нерівність:

*|T| < tдвостор.кр.(α; f)* (6.15)

то відкидати *Н0* немає підстав.

2) У випадку, якщо конкуруюча гіпотеза задана у вигляді нерівності *(Н0*: *М(Х) = М(Y), Н1: М(Х) > М(Y))*, довірчий інтервал буде одностороннім (правостороння критична область). Якщо:

*Tнабл.> tправостор.(α; f),* (6.16)

то нульову гіпотезу відкидають.

3) Аналогічним чином діють і у випадку конкуруючої гіпотези, що має вигляд *Н1:М(Х) < М(Y).* Під час побудови лівосторонньої критичної області враховують, що в силу симетрії *tлівостор.= –tправостор*.. Таким чином, знаходять *tправостор*. із другого випадку й беруть із протилежним знаком. Якщо

*Tнабл.> –tправостор.(α; f),* (6.17)

то відкидати гіпотезу *Н0* немає підстав.

Приклад. Нехай дано вибірки об'ємом *n* = 5, *m* = 6, які відібрані з нормальних генеральних сукупностей. Знайдено наступні параметри: ; *S2x* = 0.25; *S2y* = 0.108; На рівні значимості *α* = 0.05 перевірити гіпотезу *Н0: М(Х) = М(Y)* при конкуруючій *Н1: М(Х) ≠ М(Y).*

Оскільки *S2x≠ S2y*, то спочатку перевіримо гіпотезу про рівність дисперсій користуючись критерієм Фішера –Снедекора. Для цього обчислимо значення критерію *Fнабл*.= 0.25/0.108 ≈ 2.31. У якості конкуруючої гіпотези візьмемо *Н1: Д(Х) >Д(Y).* За таблицею на рівні значимості *α* = 0.05 і по числу ступенів волі *k1* = 5 – 1 = 4; *k2*= 6 – 1 = 5 знаходимо *Fкр*(0.05; 4; 5) = 5.19. Оскільки *Fнабл.< Fкр*, то нема підстав відкидати гіпотезу *Н0*. Для перевірки основного припущення обчислимо статистику *Тнабл.* за формулою (12.14): *Тнабл*.≈ 3.27. У силу вибору *Н1* необхідно побудувати двосторонню критичну область, тому на рівні значимості *α* = 0.05 і за числом ступенів волі *f* = 5 + 6 – 2 = 9 з таблиці знаходимо *tдвостор*..(0.05; 9) = 2.26. Оскільки *Тнабл.> tдвостор*., то гіпотезу *Н0* відкидаємо. Інакше кажучи, вибіркові середні відрізняються значимо.

Зауваження. Якщо є тільки одна таблиця для *tправостор*., то на практиці використовують наступну формулу:

*tдвостор.(α; f) = tправостор.(α/2; f)* (6.18)

Задача 4. Нехай генеральна сукупність випадкової величини *Х* розподілена за нормальним законом. Необхідно встановити, значимо чи незначимо розрізняються  й *М(Х) = а0*. Ця задача виникає, коли, наприклад, перевіряють партію деталей на відповідність проектному розміру *а0*.

1 –ий випадок. Дисперсія відома.

1) Висуваємо гіпотези *Н0:  = а0* і *Н1: ≠ а0*. Як критерій вибирають наступну випадкову величину, що розподілена за нормальним законом, причому при справедливості *Н0* параметри *М(U) = 0*, *σ(U)* = 1:

 (6.19)

За таблицею функції Лапласа знаходять критичну точку *Ф(Uкр)* = (1 – *α*)/2 або *U1 –α/2* – квантиль. Якщо:

*|Uнабл.| < Uкр.* (6.20)

немає підстав відкидати гіпотезу *Н0*.

2) Аналогічно перевіряють односторонні гіпотези: *Н0:  = а0* при *Н1:*>*а0.* У цьому випадку *Ф(Uкр)* = (1 – 2*α*)/2 і гіпотеза *Н0* не відкидається за умови:

*Uнабл.< Uкр.* (6.21)

3) Останній випадок: *Н0:  = а0* і *Н1: < а0*. Знаходять *Uкр*. як у пункті 2), а потім гіпотезу *Н0* не відкидають, якщо:

*Uнабл.> –Uкр.* (6.22)

Приклад. Нехай випадкова величина розподілена нормально (*Х* ∈*N*) і, крім того, відомо *σ* = 0.36. За вибіркою об'єму *n* = 36 з вибірковим середнім  = 21.6 на рівні значимості *α* = 0.05 перевірити гіпотезу *Н0:* = *а0*= 21 при конкуруючій *Н1*: > 21.

Використовуючи формулу (12.19), знаходимо вибіркове значення параметра *Uнабл*. = (21.6 – 21)√36/0.36 = 10, а за таблицею функції – критичні точки: *Ф(Uкр)* = (1 – 2⋅0.05)/2 = 0.45, отже *Ukp* = 1.65. Оскільки *Uнабл.> Uкр* і прийнята правостороння критична область, то гіпотезу *Н0* відкидають.

2 –ий випадок. Дисперсія невідома.

Як критерій беруть наступну випадкову величину:

 (6.23)

Дана випадкова величина Т має розподіл Стьюдента з *f = n* – 1 ступенями волі. Нульова гіпотеза аналогічна першому випадку: *Н0:  = а0*. Розглянемо різні випадки для конкуруючої гіпотези *Н1*:

1. *Н1: ≠ а0*. Гіпотеза Н0 не відкидається, якщо виконується нерівність:

*|Tнабл.| < tдвустор*. (6.24)

1. *Н1: > а0*. Гіпотеза *Н0* не відкидається при виконанні нерівності:

*Tнабл.< tправостор*. (6.25)

1. *Н1: < а0*. Гіпотеза *Н0* не відкидається, якщо виконується нерівність:

*Tнабл.> –tправостор*. (6.26)

Задача 5. Порівняння відносної частоти з імовірністю. Нехай за великим числом *n* незалежних випробувань, у кожному з яких імовірність *р* постійна, але невідома, знайдена відносна частота *m/n*. Потрібно на заданому рівні значимості *α* перевірити гіпотезу *Н0* про те, що ймовірність *р* дорівнює гіпотетичній ймовірності *р0*.

Як критерій приймемо наступну випадкову величину:

 (6.27)

яка при справедливості гіпотези *Н0* розподілена за нормальним законом *N*(0; 1).

Дійсно, за теоремою Лапласа доведено, що при великих *n* відносна частота має нормальний розподіл з математичним сподіванням *р* і середнім квадратичним відхиленням . Нормуючи відносну частоту, одержимо:

 (6.28)

Висунемо нульову гіпотезу *Н0: р = р0* і наступні конкуруючі гіпотези:

1. *Н1: р ≠ р0*. З таблиці Лапласа знаходимо критичні точки *Ф(Ukp)* = (1 – *α*)/2 для двосторонньої критичної області. Гіпотеза *Н0* не відкидається при виконанні нерівності:

*|Uнабл.| < Ukp* (6.29)

1. *Н1: р > р0*. При цьому *Ф(Ukp)* = (1 – 2*α*)/2. Гіпотеза *Н0* не відкидається при виконанні нерівності:

Uнабл.< Ukp (6.30)

1. *Н1: р < р0. Ф(Ukp)* = (1 – 2*α*)/2. Гіпотеза *Н0* не відкидається при виконанні нерівності:

*Uнабл.> –Ukp* (6.31)

Задача 6. Порівняння декількох дисперсій нормальних генеральних сукупностей.

1 –ий випадок. Розглядаються вибірки різного об'єму. Для перевірки нульової гіпотези в цьому випадку використовують критерій Бартлета. Нехай випадкові величини *Х1, …, Х* розподілені за нормальним законом. З генеральних сукупностей взяті незалежні вибірки об'ємом *n1, …, n* і знайдені дисперсії *S12, …, S2...* Необхідно перевірити гіпотезу про однорідність дисперсій, тобто висувається наступна нульова гіпотеза *Н0*: *Д(Х1) = Д(Х2) =…=Д(Х)*. У якості критерію приймемо наступну випадкову величину (критерій Бартлета):

*B = V/C* (6.32)







Випадкова величина *В* розподілена як *χ2* з – 1 ступенями волі.

Обмеження. Об'єм кожної вибірки повинен бути не менший чотирьох.

Критична область правостороння, тобто:

*Р[В >χ2kp(α; – 1)] = α* (6.33)

Таким чином, якщо

*Внабл*.<χ2kp (6.34)

то немає підстав відкидати нульову гіпотезу.

2 –ий випадок. Вибірки однакового об'єму – застосовується критерій Кочрена. Оскільки вибірки однакового об'єму, то число ступенів волі для кожної вибірки дорівнює *k = n* – 1. Перевірити висунуту вище нульову гіпотезу можна такими способами:

1. Використати критерій Фішера-Снедекора для порівняння максимальної та мінімальної дисперсій. Недолік даного методу полягає в тому, що втрачається інформація, яку містять інші дисперсії.
2. Використати перший випадок (критерій Бартлета). Недолік полягає в істотній погрішності, оскільки метод Бартлета не точний, а лише наближений.
3. Використати критерій Кочрена.

Критерій Кочрена. Розглядають наступну випадкову величину:

*G = Smax2/(S12 +…+S2),* (6.35)

яка має табульований розподіл Кочрена. Критичну область будують правосторонню:

*Р[G > Gkp(α; k; )] = α* (6.36)

Якщо *Gнабл.< Gкр*, то нема підстав відкидати нульову гіпотезу *Н0*.

Приклад. Нехай дані чотири вибірки об'єму *n* = 17. При цьому кожна вибірка має наступне середнє квадратичне відхилення відповідно: *S12* = 0.26, *S22* = 0.36, *S32* = 0.40, *S42* = 0.42. Необхідно: а) на рівні значимості *α* = 0.05 перевірити нульову гіпотезу *Н0: Д(Х1) = Д(Х2) = Д(Х3) = Д(Х4)*; б) оцінити генеральну дисперсію.

1) Використаємо критерій Кочрена, для чого обчислюємо наступну величину *Gнабл*. = 0.42/(0.26 + 0.36 + 0.4 + 0.42) = 0.2917. З таблиці знаходимо *Gкр*.(0.05; 16; 4) = 0.4366. Оскільки *Gнабл.< Gкр*., то гіпотезу *Н0* не відкидаємо.

2) Оскільки нульова гіпотеза справедлива, то генеральна дисперсія обчислюється як середнє арифметичне знайдених значень відхилення: *σ2*= (0.26 + 0.36 + 0.4 + 0.42)/4 ≈ 0.36.

Задача 5.Перевірка значимості вибіркового коефіцієнта кореляції. Нехай двовимірна генеральна сукупність *(X; Y)* розподілена за нормальним законом. З неї витягнута вибірка об'ємом *n* і знайдений коефіцієнт кореляції *r*≠ 0. Необхідно перевірити, що коефіцієнт кореляції генеральної сукупності також не дорівнює нулю. Отже, маємо наступну нульову й альтернативну гіпотези: *Н0: rг* = 0; *Н1: rг*≠ 0. Якщо *Н0* вірна, то випадкові величини *Х* и *Y* некорельовані.

Для створення критерію розглянемо наступну випадкову величину:

 (6.37)

Величина *Т* при справедливості нульової гіпотези має розподіл Стьюдента з *k = n* – 2 ступенями волі. Критична область двостороння. Отже, при:

*|Tнабл.| < tkp.(α; k)* (6.38)

немає підстав відкидати гіпотезу *Н0.*

Приклад. За вибіркою об'єму *n* = 122 знайдений коефіцієнт кореляції *r* = 0.4, на рівні значимості *α* = 0.05 перевірити гіпотезу *Н0: rг* = 0 при *Н1: rг*≠0.

Скористаємося критерієм (6.37): *Тнабл*. = . За таблицею розподілу Стьюдента знаходимо критичну точку *tkp*.(0.05; 120) = 1.98. Оскільки *Тнабл.> tkp*. гіпотезу *Н0* відкидаємо. Тобто вибірковий коефіцієнт кореляції значимий відрізняється від нуля, величини *X* і *Y* корельовані.

Питання для самоконтролю

1. Які гіпотези називаються параметричними (непараметричними)?
2. Яку гіпотезу називають нульовою (альтернативною)?
3. Що називається рівнем значимості?
4. Що таке помилка першого (другого) роду?
5. Що називається статистичним критерієм?
6. Сформулюйте основний принцип перевірки статистичної гіпотези.
7. Які ви знаєте критичні області?
8. Що називається потужністю критерію?
9. Сформулюйте загальний підхід до статистичної перевірки гіпотези.
10. Наведіть схему перевірки гіпотези про числове значення дисперсії й МС при відомій (невідомій) дисперсії.
11. Наведіть схему перевірки гіпотези про рівність дисперсій при невідомих середніх.
12. Наведіть схему перевірки гіпотези про рівність МС при відомих (невідомих і однакових) дисперсіях.
13. Наведіть схему перевірки гіпотези про рівність МС з невідомими дисперсіями (залежні вибірки).
14. Наведіть схему перевірки гіпотези про числове значення ймовірності події.
15. Наведіть схему перевірки гіпотези про значимість коефіцієнта кореляції.
16. Наведіть схему перевірки гіпотези про рівність дисперсій декількох нормальних генеральних сукупностей за критеріями Бартлета й Кочрена.

Навчальне завдання

1. Вимоги до даних при застосуванні параметричних критеріїв статистичного порівняння вибірок досліджуваних.

2. Особливості використання та алгоритм розрахунку критерію *t-*Стьюдента для незалежних вибірок.

3. Особливості використання та алгоритм розрахунку критерію *t-*Стьюдента для залежних вибірок.

4. Особливості використання та алгоритм розрахунку критерію *t-*Стьюдента для однієї вибірки.

Практичне завдання

***Завдання 1.***

На основі теоретично вивченого матеріалу віднести кожне з запропонованих вимірів до певної шкали:

- номінальної шкали (класу назв),

- рангової шкали (класу порядку),

- інтервальної шкали (класу інтервалу),

- пропорційної шкали (класу відношень).

1. Зайняте місце в Олімпійських іграх.

2. Академічний ранг (асистент, доцент, професор) як міра просування по службі.

3. Метрична система виміру віддалі.

4. Телефонні номери (стаціонарні, мобільні).

5. Різниця в температурі тіла здорової та хворої людини.

6. Номери залікових книжок та студентських квитків.

7. Середня успішність студентів групи, потоку.

8. Ідентифікаційний код людини.

9. Вага спортсмена-боксера.

10. Вагові категорії боксерів – учасників ігор.

11. Визначення віку людини за датою народження і смерті.

12. Ріст дорослої людини, студента, учня.

13. Скільки років Роксолана була дружиною турецького султана Сулеймана (1520-1561).

14. Номери будинків та квартир.

15. Офіцерське звання (лейтенант, ст.лейтенант, майор, підполковник, полковник).

16. Порядковий номер студента у списку.

***Завдання 2.*** Чи статистично значуще перевищує рівень розвитку посттравматичного стресового розладу (ПТСР) нормативний показник у 35 балів у ліквідаторів Чорнобильської катастрофи? Перевірити результати тестування на відповідність закону нормального розподілу. Застосувати критерій *t*-Стьюдента для однієї вибірки.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| ПТСР (Х) | 33 | 34 | 36 | 39 | 41 | 42 | 29 | 34 | 38 | 32 | 29 | 31 | 31 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| ПТСР (Х) | 38 | 40 | 31 | 33 | 35 | 35 | 41 | 40 | 38 | 37 |

***Завдання 3.***Знайти з надійністю 0,95 довірчий інтервал оцінки математичного сподівання а нормально розподіленої ознаки Х генеральної сукупності, якщо відомі вибіркова середня , об’єм вибірки *n* та середнє квадратичне відхилення *σ* генеральної сукупності.

а)  =14, *n* = 25, *σ* = 5; б)  =2000, *n* = 1600, *σ* = 40

***Завдання 4.***У будинку відпочинку випадковим способом було відібрано 20 осіб і виміряно їх зріст *хі*. Здобуті результати наведено у вигляді інтервального статистичного розподілу:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *хі* | 165.5-170.5 | 170.5-175.5 | 175.5-180.5 | 180.5-185.5 |
| *nі* | 4 | 6 | 8 | 2 |

Із надійністю *γ*=0.95 побудувати довірчий інтервал для , якщо *σ* = 2.

***Завдання 5.***За заданим статистичним розподілом вибірки, реалізованим із генеральної сукупності, ознака *Х* якої має нормальний закон розподілу

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *хі* | 4,2 | 6,2 | 8,2 | 10,2 | 12,2 |
| *nі* | 6 | 8 | 12 | 8 | 2 |

При рівні значущості α = 0,01 перевірити правильність нульової гіпотези *Н0*: *а* = 10, якщо альтернативна гіпотеза *Н1*: *а*> 10, *σ2* = 16.

***Завдання 6.***Із нормальної генеральної сукупності отримано вибірку об’єму *n* = 31

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *хі* | 10,1 | 10,3 | 10,6 | 11,2 | 11,5 | 11,8 | 12,0 |
| *nі* | 1 | 3 | 7 | 10 | 6 | 3 | 1 |

Перевірити при рівні значущості α = 0,05 основну гіпотезу *Н0*: *σ2* = *σ20* = 0,18, якщо альтернативна гіпотеза *Н1*: *σ2* ≠ 0,18.

## Тема 7. Непараметричні методи перевірки гіпотез

Мета: розглянути основні поняття теорії непараметричних методів перевірки гіпотез, оволодіти навичками застосування непараметричних статистичних тестів для вирішення практичних задач.

🖉Основні терміни і поняття

*Критерій згоди*, *критерій Пірсона,критерій омега-квадрат, критерій Колмогорова, критерій знаків, W-критерій Уїлкоксона, критерій серій.*

📚Основні теоретичні положення

Дані реальних експериментів можуть бути подані *незалежними* або *спряженими* вибірками. Для незалежних вибірок критерії допомагають виявити статистичну значущість різниці, що спостерігається. Прикладами незалежних вибірок є:

* мешканці двох різних населених пунктів (при демографічних дослідженнях);
* дві партії однотипної продукції, виготовлені різними працівниками на різному обладнанні (при розробці технології виробництва);
* випускники різних шкіл (при аналізі результатів зовнішнього незалежного оцінювання).

Критерії, що застосовують до вибірок з попарно спряженими даними, називають *парними*. Прикладами спряжених вибірок є:

* дані опитування громадської думки до й після певної суспільно значущої події;
* дві партії однотипної продукції, виготовлені одними й тими самими працівниками на одному й тому самому обладнанні до й після внесення певних змін до технології;
* одна й та сама партія виробів до і після певної технологічної обробки.

**7.1 Критерій згоди**

У попередніх задачах закон розподілу генеральної сукупності передбачається відомим. Якщо закон розподілу невідомий, але є підстави припускати, що він має певний вигляд, то перевіряють нульову гіпотезу *Н0*: генеральна сукупність розподілена за законом *А.*

Визначення.*Критерієм згоди* називають критерій перевірки гіпотези про передбачуваний закон невідомого розподілу.

Побудова інтервального статистичного ряду й основних числових характеристик вибірки дозволяє попередньо обрати закон розподілу випадкової величини *Х*, виходячи з таких міркувань:

1) За видом гістограми частостей статистичного ряду. Вид гістограми дає орієнтування на можливий закон розподілу. Перевага цього методу полягає в простоті й наочності, а недолік у тому, що гістограма може одночасно нагадувати декілька законів розподілу, або не нагадувати жодного.

2) Порівняння вибіркових оцінок випадкової величини. Наприклад, для нормального закону розподілу порівнюють початкові моменти: *М(х) = μе = μ0*. Перевага – простота й наочність, недолік – висновки можуть бути дуже наближені, в яких не відбивається ступінь симетрії кривої розподілу.

3) За емпіричним коефіцієнтом варіації *V = S/* . Відомо, що кожному закону розподілу відповідає певний діапазон значень коефіцієнта варіації:

Перевага методу – простота, недолік – коефіцієнт варіації не відбиває ступінь симетричності емпіричної кривої розподілу, велика неоднозначність вибору.

Таблиця 7.1 Вибір закону за емпіричним коефіцієнтом

|  |  |  |
| --- | --- | --- |
| Закон розподілу | Границі | Середнє значення |
| Нормальний | (0.08;0.4) | 0.25 |
| Вейбулла | (0.36;0.63) (0.4;0.85) | 0.44, 0.71 |
| Логарифмічно –нормальний | (0.35;0.8) | 0.68 |
| Експонентний | (0.6;1.3) | 0.92 |

4) За вибірковими коефіцієнтами асиметрії й ексцесу.

 (7.1)

 (7.2)

Емпірична функція вважається відповідною гіпотетичній функції, якщо вибіркові коефіцієнти асиметрії й ексцесу відрізняються за абсолютною величиною від своїх математичних сподівань не більше ніж на потроєні середні квадратичні відхилення, тобто якщо:

*|A – M(A)| < 3SA* (7.3)

*|Е – M(Е)| < 3SЕ* (7.4)

то обраний закон розподілу погоджується з експериментальними даними. Наприклад, для нормального закону розподілу *М(А) = М(Э)* = 0, тому умови узгодження набувають такого вигляду:

*|A| < 3SA , |Е| < 3SЕ* (7.5)

де

 , 

Достоїнство методу – урахування симетрії й крутості, тобто форми кривої, недолік – немає строгої кількісної оцінки припустимої розбіжності між вибірковими коефіцієнтами асиметрії й ексцесу та їхніх математичних сподівань, оскільки використовується правило «трьох сигм», що не дуже точне.

5) За допомогою чисел Вестергарда. Використовується тільки для перевірки відповідності емпіричного закону розподілу до нормального. Числа Вестергарда: 0.3, 0.7, 1.1, 3. Емпіричний розподіл близький до нормального закону розподілу, якщо в інтервалі ( – 0.3*S*;  + 0.3*S*) розташована чверть всієї сукупності даних, в інтервалі ( – 0.7*S*;  + 0.7*S*) – половина всієї сукупності даних, в інтервалі ( – 1.1*S*;  + 1.1*S*) – три чверті, а в інтервалі ( – 3*S*;  + 3*S*) – 0.998 відсотків всієї сукупності даних.

Приклад обробки емпіричних даних. Представимо вибірку з 55 спостережень у вигляді таблиці частот, використовуючи 7 інтервалів угрупування. Вибірка:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 17 | 19 | 23 | 18 | 21 | 15 | 16 | 13 | 20 | 18 | 15 |
| 20 | 14 | 20 | 16 | 14 | 20 | 19 | 15 | 19 | 16 | 19 |
| 15 | 22 | 21 | 12 | 10 | 21 | 18 | 14 | 14 | 17 | 16 |
| 13 | 19 | 18 | 20 | 24 | 16 | 20 | 19 | 17 | 18 | 18 |
| 21 | 17 | 19 | 17 | 13 | 17 | 11 | 18 | 19 | 19 | 17 |

Розмах вибірки . Довжина інтервалу *b* = 14/7 = 2, результати угруповання зведені в таблицю 7.2.

Таблиця 7.2 Результати угрупування за даними вибірки

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № інтервалу *і* | Границі інтервалу | Середина інтервалу  *xi\** | Частота  *ni\** | Накопичена частота | Відносна частота | Накопичена відносна частота |
| 1 | 10 –12 | 11 | 2 | 2 | 0,0364 | 0,0364 |
| 2 | 12 –14 | 13 | 4 | 6 | 0,0727 | 0,1091 |
| 3 | 14 –16 | 15 | 8 | 14 | 0,1455 | 0,2546 |
|  | 16 –18 | 17 | 12 | 26 | 0,2182 | 0,4728 |
| 5 | 18 –20 | 19 | 16 | 42 | 0,2909 | 0,7637 |
| 6 | 20 –22 | 21 | 10 | 52 | 0,1818 | 0,9455 |
| 7 | 22 –24 | 23 | 3 | 55 | 0,0545 | 1,0000 |

Далі обчислимо середнє, дисперсію, коефіцієнти асиметрії й ексцесу для нашої вибірки. Довжина інтервалу угруповання *b* = 2. Значення *xi\**, що зустрічається з найбільшою частотою, . Тому нормування має вигляд:

, *і* = 1,2,…,7.

Обчислення оформимо у вигляді таблиці 7.3.

Таблиця 7.3 Результати розрахунків

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *і* | *xi\** | *ui* | *ni\** | *ni\*ui* | *ui2* | *ni\*ui2* | *ui3* | *ni\*ui3* | *ui4* | *ni\*ui4* | *ui+*1 | (*ui+*1)4 | *ni\**(*ui+*1)4 |
| 1 | 11 | –4 | 2 | –8 | 16 | 32 | –64 | –128 | 256 | 512 | –3 | 81 | 162 |
| 2 | 13 | –3 | 4 | –12 | 9 | 36 | –27 | –108 | 81 | 324 | –2 | 16 | 64 |
| 3 | 15 | –2 | 8 | –16 | 4 | 32 | –8 | –64 | 16 | 128 | –1 | 1 | 8 |
| 4 | 17 | –1 | 12 | –12 | 1 | 12 | –1 | –12 | 1 | 12 | 0 | 0 | 0 |
| 5 | 19 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 16 |
| 6 | 21 | 1 | 10 | 10 | 1 | 10 | 1 | 10 | 1 | 10 | 2 | 16 | 160 |
| 7 | 23 | 2 | 3 | 6 | 4 | 12 | 8 | 24 | 16 | 48 | 3 | 81 | 243 |
| ∑ | – | – | 55 | –32 | – | 134 | – | –278 | – | 1034 | – | – | 653 |
|  | | | – | –0,582 | – | 2,436 | – | –5,05 | – | 18,8 | – | – | – |

Контроль обчислень за тотожністю дає наступне:

55+4 ( –32)+6 134+4 ( –278) –1034=653

Нарешті, скориставшись формулами, знаходимо параметри розподілу:

,





**7.2 Критерій Пірсона (*χ2)***

Нехай *х1, х2, …, хn* – вибірка спостережень випадкової величини *Х.* Перевіряється гіпотеза *Н0*, якастверджує, що випадкова величина *Х* має закон розподілу *F(X).* Процедура застосування критерію *χ2* для перевірки гіпотези *Н0* складається з наступних етапів:

1) За вибіркою спостережень випадкової величини *Х* будують групований (інтервальний) статистичний ряд і знаходять оцінки невідомих параметрів передбачуваного закону розподілу (тобто  ).

2) Використовуючи закон розподілу, що перевіряють, обчислюють імовірності *рk*, з якими випадкова величина *Х* потрапляє в кожний інтервал. Наприклад, для нормального закону:

*pj = P{xj –1< x ≤ xj} = Ф(Uj) – Ф(Uj –1)* (7.6)

де *Uj = (xj – )/S*.

3) Визначають теоретичні частоти *n\*pj* для кожного часткового інтервалу.

4) Обчислюють вибіркову статистику (критерій) за формулою:

 (7.7)

5) Приймають статистичний розв’язок: гіпотеза *Н0* не суперечить вибірці спостережень на заданому рівні значимості *α*, якщо:

 (7.8)

де *r* – число інтервалів, – число параметрів розподілу, які оцінюються за вибіркою. У тому випадку, коли , нульова гіпотеза відхиляється.

Зауваження. Необхідно, щоб для всіх інтервалів виконувалася умова *npj*≥ 5. Якщо для деяких інтервалів ця умова не виконується, то їх варто об'єднати із сусідніми.

Приклад. Перевірити гіпотезу про нормальний розподіл групованих даних, рівень значимості прийняти рівним 0.1. Об'єм вибірки *n* = 55, середнє значення  = 17.84, оцінка для середньо квадратичного відхилення дорівнює *S* = 2.92.

Для нормального закону число оцінюваних за вибіркою параметрів дорівнює двом (тобто = 2), число ступенів волі f = 4 – 2 – 1 = 1. З таблиці *χ2* знаходимо: *χ0.92* (1) = 2.71. Оскільки *χВ2* = 0.919 <*χ0.92*(1), то гіпотеза про нормальний закон розподілу не суперечить результатам спостережень.

Таблиця 7.4 Перевірка гіпотези *Н0* за критерієм Пірсона

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *xj –xj+1* | *mj* | *Uj;Uj+1* | *Pj=Ф(Uj) –Ф(Uj+1)* | *npj* | *mj –npj* |  |
| 10 -12 | 2 | –2.68; –2 | 0. 0191 | 1. 0505 | –0. 333 | 0. 0077 |
| 12 -14 | 4 | –2; –1.32 | 0. 0714 | 3. 927 14.333 |
| 14 -16 | 8 | –1.32; –0.63 | 0. 1701 | 9. 3555 |
| 16 -18 | 12 | –0.63; 0.05 | 0. 2556 | 14. 058 | –2. 058 | 0. 3013 |
| 18 -20 | 16 | 0.05; 0.74 | 0. 2504 | 13. 772 | 2. 228 | 0. 3604 |
| 20 -22 | 10 | 0.74; 1.43 | 0. 1526 | 8. 393 | 1. 329 | 0. 1513 |
| 22 -24 | 3 | 1.43; 2.11 | 0. 0596 | 3. 278 11.671 |
| *χB2* | | | | | | 0. 8207 |

**7.3 Критерій Колмогорова (*λ*)**

Критерій згоди *χ2* дозволяє перевірити гіпотезу про згоду даної вибірки з конкретним теоретичним законом розподілу для будь –якої випадкової величини (як безперервної, так і дискретної). Критерій *λ* застосовується для перевірки гіпотез про закони розподілу тільки безперервних випадкових величин. Його відмінність полягає в тому, що порівнюються не емпіричні й теоретичні частоти, а емпірична й гіпотетична функції розподілу. Передбачається, що теоретичні значення параметрів функції розподілу генеральної сукупності відомі. Перевірку гіпотези *Н0* проводять за наступною схемою:

1) Будують інтервальний статистичний ряд і знаходять емпіричну функцію розподілу:

 (7.9)

2) Знаходять нормовані інтервали (для перевірки нормального закону розподілу):

 (7.10)

3) Визначають значення теоретичної функції розподілу, які відповідають спостережуваним значенням випадкової величин *Х*.

4) Для кожного *хj* (або інтервалу) знаходять модуль різниці між емпіричною й теоретичною функціями розподілу: .

5) Обчислюють спостережуване значення вибіркової статистики Колмогорова:

 (7.11)

6) Порівнюють спостережуване значення вибіркової статистики *λ* із критичним значенням *λα*, яке визначається за таблицею квантилів розподілу Колмогорова на заданому рівні значимості *α*. При цьому, якщо *λ<λα*, то немає підстав для відхилення нульової гіпотези, якщо ж *λ≥λα*, то приймають альтернативну гіпотезу.

Таблиця 7.5 Квантилі розподілу Колмогорова

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| α | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0. 001 |
| λα | 1. 078 | 1. 224 | 1. 358 | 1. 520 | 1. 827 | 1. 950 |

Приклад. Перевірити гіпотезу про нормальний розподіл групованих даних, рівень значимості прийняти 0.1. Об'єм вибірки *n* = 55, середнє значення  = 17.84, оцінка середньо квадратичного відхилення дорівнює *S* =2.92.

Для обчислення значень теоретичної функції нормального розподілу використовують функцію Лапласа:





Оскільки *λ*< 1.224, то немає підстав відкидати гіпотезу *Н0*.

Таблиця 7.6 Перевірка гіпотези *Н0* за критерієм Колмогорова

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *xj –xj+1* | *mj* | *Uj;Uj+1* | *Fn(X) для Uj* | *F(X) для Uj* |  |
| 10 -12 | 2 | –2.68; –2 | 0. 0364 | 0. 0465 | 0. 0101 |
| 12 -14 | 4 | –2; –1.32 | 0. 1091 | 0. 0934 | 0. 0157 |
| 14 -16 | 8 | –1.32; –0.63 | 0. 2545 | 0. 2643 | 0. 0098 |
| 16 -18 | 12 | –0.63; 0.05 | 0. 4727 | 0. 5199 | 0. 0472 |
| 18 -20 | 16 | 0.05; 0.74 | 0. 7636 | 0. 7703 | 0. 0067 |
| 20 -22 | 10 | 0.74; 1.43 | 0. 9455 | 0. 9236 | 0. 0219 |
| 22 -24 | 3 | 1.43; 2.11 | 1 | 0. 9826 | 0. 0174 |

**7.4 Основні поняття. Критерій знаків**

У практиці обробки результатів розподіл генеральної сукупності часто невідомий або відрізняється від нормального розподілу випадкової величини. У цих випадках застосовують методи, що не залежать від розподілу, або, так звані *непараметричні*. Ці методи використовують не самі чисельні значення елементів вибірки, а *структурні властивості вибірки* (наприклад, порядок між елементами). Тому потужність таких методів нижча, ніж потужність параметричних гіпотез. Перевагою цих методів є простота й можливість побудови загальних суджень.

Серед непараметричних тестів важливе місце займають так звані *робастні методи*, що виявляють слабку чутливість до відхилень від стандартних умов і можуть використовуватися в широкому діапазоні реальних умов.

При перевірці нульової гіпотези про однорідність вибірок числових даних рекомендується використовувати омега-квадрат критерій або (за відсутності необхідних таблиць та програмного забезпечення) критерій Смірнова.

*Критерій омега-квадрат (критерій Крамера – фон Мізеса)* базується на розгляді відхилення між двома емпіричними функціями розподілу (або між емпіричною й теоретичною функціями розподілу при ідентифікації закону розподілу). Вперше його запропонували у 1928–1930 р. шведський математик Карл Харальд Крамер та американський математик і механік Ріхард фон Мізес, який народився у Львові.

Двохвибірковий варіант критерію запропоновано в 1951 р. американським статистиком Еріхом Лео Леманом та досліджено у 1952 р. американським математиком Мюреєм Розенблаттом. Тому його іноді називають також *критерієм Лемана – Розенблатта*. На сьогодні його вважають найпотужнішим з критеріїв, призначених для перевірки гіпотези про однорідність незалежних вибірок.

*Критерій Смірнова* запропонований в 1939 р. радянським математиком Миколою Васильовичем Смірновим. Він призначений для перевірки гіпотези про однорідність двох вибірок з неперервним законом розподілу. Нульову гіпотезу відхиляють, якщо розрахункове значення критерію перевищить критичне для відповідного рівня значущості.

Для незалежних вибірок можна застосовувати *критерій рандомізації компонент*. Він розроблений Р. Фішером у 1920 р. для аналізу вибірок малого обсягу. При порівнянні спряжених вибірок основою методу є перебирання можливих результатів, побудованих з різницевих оцінок. Нульова гіпотеза полягає у рівності вибіркових середніх.

Більша група непараметричних критеріїв використовується для перевірки гіпотези про приналежності двох вибірок однієї й тієї ж генеральної сукупності, Тобто про рівність функцій розподілу: *F*(*x*)≡*F*(*y*)*|y=x.* Такі генеральні сукупності називаються *однорідними*.

Простішим критерієм такого роду є критерій знаків. Він служить для перевірки гіпотези *Н*0 про однорідність генеральних сукупностей за попарно зв'язаними вибірками (задача порівняння двох вимірювальних приладів).

*Критерій знаків.* Використовуються *n* об'єктів і над кожним проводять по одному виміру за допомогою обох приладів. Якщо вибірки однорідні, то значення *хi* і *yi*взаємозамінні й імовірності появи того, що *xi –yi*>0 і *xi –yi*<0 рівно.

Статистикою критерію знаків є число знаків «+» у послідовності знаків різниць, які не дорівнюють нулю. За умови, що *Н*0 вірна, а пари спостережень (х*i*; *yi*) незалежні, число знаків має біноміальний розподіл з параметрами *р*=1/2 і *n*=*L* (*L* – число ненульових знаків).

Таким чином, нульова гіпотеза має вигляд: *Н*0: *р*=1/2, а можливі конкуруючі гіпотези: *Н*1(1):*р*>1/2, *H*1(2): *p*<1/2, *H*1(3): *p*≠1/2.

Для перевірки гіпотези *Н0* використовують статистику Фішера – Снедекора. Критичні області:

для Н1(1)

 (7.12)

де *k*1=2(*L* –*R*+1), *k*2=2*R*;

для *Н*1(2):

 (7.13)

де *k*1=2(*R*+1), *k*2=2(*L–R*);

для *Н*1(3): об'єднання перших двох випадків для рівня значимості α/2. Тут *R* – число знаків «+», *k*1, *k*2 – ступеня волі.

Приклад. Перевірялася швидкість десяти автомобілів за допомогою двох приладів. Результати перевірки представлені у вигляді наступної таблиці:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| V1 | 70 | 85 | 63 | 54 | 65 | 80 | 75 | 95 | 52 | 55 |
| V2 | 72 | 86 | 62 | 55 | 63 | 80 | 78 | 90 | 53 | 57 |

Чи можна сказати, що другий прилад дає завищені значення? Рівень значимості прийняти 0,1.

У припущенні, що швидкості руху автомобілів не залежать одна від одної, задачу можна розв’язати, застосовуючи критерій знаків. Складемо послідовність знаків різниці швидкостей автомобілів отриманих на першому й другому приладах:

: –, –, +, –, +, 0, –, +, –, –.

Число нульових різниць, число позитивних різниць. Перевіримо гіпотезу про те, що розбіжності в показниках приладів зумовлені випадковими помилками, тобто гіпотезу : . Альтернативна гіпотеза припускає, що показники другого приладу вищі, тоді знаків «+» повинно бути менше, а ймовірність, отже, повинна бути менше 1/2. Отже, альтернативну гіпотезу приймаємо *Н*1(2): *p*<1/2 і використовуємо формулу (7.13)

*k*1=2(3+1)=8, *k*2=2(9–3)=12, *F*B=(9–3)/(3+1)=1,5

За таблицею Фішера *F*0,9(8;12) =2,24, а тому що *F*B<*F*0.9 (8;12) немає підстави відкидати гіпотезу *Н*0. Варто вважати, що розбіжності в показниках приладів викликані випадковими помилками.

**7.5 Критерій Уїлкоксона**

*W-критерій Уїлкоксона (критерій рангових сум)* запропонований в 1945 р. американським хіміком і статистиком Френком Уїлкоксоном. Його застосовують для порівняння двох незалежних сукупностей за їх центральною тенденцією, тобто за центрами емпіричних функцій розподілу. Сукупності можуть мати як однакові, так і різні чисельності. Критерій оперує не числовими значеннями даних, а їх рангами – місцями у впорядкованих за згасанням або зростанням рядах даних. При його застосуванні передбачається, що розподіли вибірок є неперервними, а нульова гіпотеза полягає в тому, що функції розподілу вибірок збігаються одна з одною.

Процедура обчислення значення критерію є близькою до обчислення критерію рандомізації компонент. Різниця полягає в тому, що замість вихідних даних використовують їх ранги. Ранжирування порівнюваних вибірок здійснюють сумісно. Вихідні дані об’єднують до однієї вибірки, впорядковують, визначають ранги елементів об’єднаної вибірки. Потім формують дві нові вибірки, елементами яких є ранги відповідних елементів вихідних вибірок. Якщо деякі значення збігаються, то відповідним спостереженням призначають середній ранг.

При застосуванні W-критерію Уїлкоксона слід мати на увазі, що згідно з він належить до так званих критеріїв зсуву. Тобто найбільш потужним він є при виявленні різниці, пов’язаної з тим, що одну з вибірок отримано додаванням одного й того самого числа до всіх елементів іншої вибірки.

Він є нечутливим до різниці дисперсій порівнюваних вибірок, а також коефіцієнтів їх асиметрії та ексцесу. Зокрема, якщо дві вибірки мають симетричні функції розподілу з однаковими середніми значеннями, але різними стандартними відхиленнями, то в об’єднаної послідовності елементи однієї вибірки матимуть підвищену кількість елементів з високими та низькими рангами. Елементи іншої вибірки будуть мати підвищену кількість елементів із середніми значеннями рангів. Але суми рангів усіх елементів для цих двох вибірок можуть бути приблизно однаковими.

*Критерій Вілкоксона* застосовується для порівняння двох незалежних вибірок об'єму *n*1 і *n*2 і перевіряє гіпотезу *Н*0, що підтверджує, що вибірки отримані з однорідних генеральних сукупностей, і, зокрема, мають рівні середні й медіани.

Статистика *W* будується в такий спосіб. Розташуємо *n*1+ *n*2 значень об'єднаної вибірки в порядку зростання (у вигляді варіаційного ряду). Для кожного елемента ряду поставимо його відповідно номер у ряді – *ранг*. Якщо кілька елементів збігається за величиною, то кожному з них привласнюється ранг, що дорівнює середньому арифметичному їхніх номерів. Останній елемент у ранжируваній об'єднаної вибірки повинен мати ранг *n*1+ *n*2. Далі нехай *R*1 – сума рангів першої вибірки, *R*2 – другий. Обчислимо такі величини:

 (7.14)

 (7.15)

Правильність обчислень перевіряють за формулою: . Вибіркове значення  є мінімум з *ω*1 і *ω*2.

У таблиці розподілу Вілкоксона наводяться ймовірності того, що *W*<*ω*У за умови, що Н0 вірна, тобто *р*=Р{ *W*<*ω*В / *H*0} для вибірок об'єму *n*1 і *n*2 (*n*1≥*n*2). При однобічній (двосторонній) альтернативній гіпотезі Н1 гіпотеза *Н*0 відхиляється, якщо *р*≤α (*р*≤α/2), де α – рівень значимості. У протилежному випадку приймається нульова гіпотеза.

Якщо об'єм кожної з вибірок більше восьми, то перевірку гіпотези Н0 можна проводити, використовуючи статистику:

 (7.16)

Z має для гіпотези *Н*0 нормальний розподіл *N* (0;1). Критичні точки перебувають у таблиці розподілу Лапласа залежно від виду альтернативної гіпотези:

1) двостороння *Ф*(*Z*kp)=(1 –α)/2. Критична область: |*Z*B|>*Z*kp;

2) однобічна *Ф*(*Z*kp)=(1 –2α)/2. Критична область: *Z*B<–*Z*kp; або *Z*B>*Z*kp для лівосторонньої й правобічної відповідно.

Приклад. Вимірялася напруга пробою в діодів двох партій, результати вимірів представлені в таблиці:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 партія | 39 | 50 | 61 | 67 | 40 | 40 | 54 | – |
| 2 партія | 60 | 53 | 42 | 41 | 40 | 54 | 63 | 69 |

Чи можна вважати, що для другої партії напруга пробою вища при рівні значимості 0,1?

Для критерію Вілкоксона складемо варіаційний ряд, відзначаючи приналежність до першої партії рискою вгорі:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Елемент | 39 | 40 | 40 | 40 | 41 | 42 | 50 | 53 | 54 | 54 | 60 | 61 | 63 | 67 | 69 |
| Ранг | 1 | 3 | 3 | 3 | 5 | 6 | 7 | 8 | 9.5 | 9.5 | 11 | 12 | 13 | 14 | 15 |

Знайдемо суми рангів першої й другої вибірок: *R*1=49.5, *R*2=70.5. Використовуючи формули (14.1) і (14.2) визначимо величини *ω*1 і *ω*2:

*ω*1=7⋅8+7(7+1)/2 –49.5=34.5; *ω*2=7⋅8+8(8+1)/2 –70.5=21.5.

Проведемо перевірку: 34.5+21.5=7\*8=56.

*ω*В = min(*ω*1;*ω*2) =21.5.

За таблицею приблизно знаходимо *р*=Р[*W*<21.5]=0.25. Оскільки альтернативна гіпотеза однобічна, а *р*>α=0.1, то приймаємо гіпотезу *Н*0.

Приклад. В умовах попереднього прикладу отримані результати нової серії вимірів напруги пробою в діодів (у вольтах):

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 –я партія | 50 | 41 | 48 | 60 | 46 | 60 | 51 | 42 | 62 | 54 | 42 | 46 |
| 2 –я партія | 38 | 40 | 47 | 51 | 63 | 50 | 63 | 57 | 59 | 51 | – | – |

Чи є підстави стверджувати, що напруга пробою в діодів обох партій різна?

Розв’язати приклад, використовуючи:

а) критерій Вілкоксона;

б) перевірку гіпотези про рівність середніх (у припущенні, що обидві вибірки отримані з нормального розподілених генеральних сукупностей). Прийняти α = 0,1.

а) упорядкуємо результати вимірів і визначимо ранги кожного результату. Маємо

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Елемент | 38 | 40 | 41 | 42 | 42 | 46 | 46 | 47 | 48 | 50 | 50 |
| Ранг | 1 | 2 | 3 | 4,5 | 4,5 | 6,5 | 6,5 | 8 | 9 | 10,5 | 10,5 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Елемент | 51 | 51 | 51 | 54 | 57 | 59 | 60 | 60 | 62 | 63 | 63 |
| Ранг | 13 | 13 | 13 | 15 | 16 | 17 | 18,5 | 18,5 | 20 | 21,5 | 21,5 |

Знайдемо суми рангів , .

Якщо , знаходимо:





Вибіркове значення статистики критерію таке: .

Якщо  й  то для перевірки гіпотези *Н*0використаємо статистику *Z*. Вибіркове значення цієї статистики визначається за формулою (7.16):



Припущення, що перевіряємо, відповідає двосторонній альтернативній гіпотезі, отже,  значення порівнюється із  квантіллю, що визначається за відповідною таблицею: 

Таким чином, твердження про те, що напруга пробою в діодів обох партій є різною, варто відхилити.

б) Перевіримо гіпотезу про рівність середніх. За результатами спостережень обчислимо оцінки середніх і дисперсій для кожної вибірки:

,,

Попередньо варто перевірити гіпотезу про рівність дисперсій *Н*0:  при альтернативній гіпотезі *Н1*:. Для цього знайдемо відношення вибіркових дисперсій і порівняємо його із квантіллю :

; 

Якщо 1,39 < 2,90, то гіпотеза про рівність дисперсій приймається, отже, перевірку гіпотези про рівність середніх можна проводити на основі статистики Стьюдента. Вибіркове значення цієї статистики дорівнює



За умови, що значення квантіллі  за таблицею Д6 дорівнює  й  гіпотеза про рівність середніх приймається. Таким чином, твердження про те, що напруга пробою в діодів обох партій по –різному, варто відхилити. Це збігається з результатом, отриманим в а).

**7.6 Критерій серій**

*Критерій серій Вальда – Волфовиця* розроблений в 1940р. американськими математиками Абрахамом Вальдом і Джекобом Волфовицем. Його використовують для перевірки нульової гіпотези про те, що дві незалежні випадкові вибірки обсягами *n1* та *n2* не відрізняються одна від одної за досліджуваною ознакою.

Результати спостережень записують як варіаційний ряд об’єднаної вибірки, а їх належність до вихідних вибірок помічають за допомогою додаткової змінної, яка може набувати два значення, наприклад «0» та «1». Послідовність її значень називають *послідовністю кодів*.

*Серією* послідовності кодів називають будь-яку послідовність її однакових значень. Наприклад, у послідовності 00101111011 є такі серії: 00, 1, 0, 1111, 0, 11. Очевидно, що за умови справедливості нульової гіпотези кількість серій N має бути великою, а за умови її помилковості – відносно малою.

*Розглянемо застосування критерію серій.* Основна гіпотеза *Н*0: елементи вибірки отримані випадковим чином і незалежні. Нехай  – вибірка результатів спостережень, а  – вибіркова медіана, визначена за цими даними. Кожному елементу вибірки поставимо відповідно знак «+» або «–» залежно від того, більше або менше медіани його значення (нульові значення не враховуються). Тим самим всій вибірці поставлений відповідно набір знаків. Позначимо  число знаків «+», а  – число знаків «–» в отриманому наборі знаків.

Статистикою критерію серій є число серій *N*. Критична область визначається нерівностями  й .Значення границь критичної області  й для рівня значимості α = 0,05 наведені в таблиці Д 9.

Приклад. Швидкості автомобілів у деякій точці траси утворили наступний ряд (км/год):

31, 39, 40, 45, 27, 28, 35, 55, 21, 33, 42, 36.

Чи можна вважати отримані значення випадковими? Прийняти α = 0,05.

Знайдемо оцінку медіани цих даних. Для цього представимо дані у вигляді варіаційного ряду:

21, 27, 28, 31, 33, 35, 36, 39, 40, 42, 45, 55.

Оцінка медіани  Вихідному ряду спостережень відповідає наступна послідовність знаків:

– + + + – – – + – – + +,

де = 6, = 6, число серій *N = 6.* За таблицею Д 9 при α = 0,05 знаходимо , . Таким чином, гіпотеза *Н*0 приймається: отримані значення швидкості можна вважати випадковими.

При більших об'ємах вибірки, коли або , або , або обидва значення більше 20, для перевірки гіпотези *Н*0 можна використати статистику *Z*, вибіркове значення  якої обчислюється за формулою

 (7.17)

За умови, що вірна гіпотезу *Н*0, статистика *Z* має приблизно нормальний розподіл *N (0,1*). У цьому випадку критична область визначається нерівностями

 або 

Зауваження. Критерій серій застосовується для перевірки випадковості будь – якої вибірки, елементами якої є два різних символи, наприклад: 1 і 0, А і В, + і –. Статистикою критерію є число серій у цій вибірці.

Приклад. Чи можна вважати, що послідовність

110010001010100111011010010000101000

отримана із сукупності випадкових послідовностей? Прийняти α = 0,01.

У цій послідовності число нулів , а число одиниць . Число серій *N*  = 23.

Якщо > 20, для перевірки гіпотези *Н*0, що підтверджує, що дана послідовність отримана із сукупності випадкових послідовностей, скористаємося статистикою *Z*. За формулою (1.6) вибіркове значення  цієї статистики дорівнює



Оскільки , гіпотеза *Н*0 приймається: можна вважати, що дана послідовність отримана із сукупності випадкових послідовностей.

**7.7 Рангова кореляція**

Нехай вибірка спостережень безперервних випадкових величин *X* і *Y*. Кожному значенню  визначимо відповідно ранг , тобто номер елемента  у варіаційному ряді  Аналогічним способом визначимо ранги  елементів . Кожній парі  таким чином, відповідає пари рангів . Обчислимо коефіцієнт кореляції за вибіркою рангів . Отримане значення називається вибірковим значенням рангового коефіцієнта кореляції Спирмена . Ранговий коефіцієнт кореляції, так само як і коефіцієнт кореляції *r*, характеризує залежність між випадковими величинами *X* і *Y*, але обчислюється значно простіше, а саме, справедливою буде формула:

 (7.18)

Тому при великому об'ємі вибірки для оцінки залежності між випадковими величинами використовується рангова кореляція. Коефіцієнт є непараметричною мірою зв'язку й, отже, може використатися при довільному безперервному розподілі генеральної сукупності, у той час як використання коефіцієнта кореляції *r* припускає двовимірний нормальний розподіл генеральної сукупності. Гіпотеза  при альтернативній гіпотезі й при об'ємі вибірки  перевіряється за значенням статистики:

 (7.19)

За умови, що вірна гіпотеза , ця статистика має розподіл Стьюдента зі  ступенями волі. Якщо  де α – заданий рівень значимості, то гіпотеза  відхиляється, тобто між *X* і *Y* існує рангова кореляційна залежність.

Приклад. Обчислити коефіцієнт рангової кореляції для наступної вибірки:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 68,8 | 63,3 | 75,7 | 67,2 | 71,3 | 72,8 | 76,5 | 63,5 | 69,9 | 71,4 |
|  | 167,0 | 113,3 | 159,9 | 153,6 | 150,8 | 181,2 | 173,1 | 115,4 | 125,6 | 166,2 |

Перевірити значимість рангової кореляції при 

Визначимо ранги елементів вихідної вибірки. Попередньо перепишемо вихідну вибірку, упорядкувавши її елементи за верхнім рядком (тобто за значеннями ), у результаті одержимо:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 63,3 | 63,5 | 67,2 | 68,8 | 69,9 | 71,3 | 71,4 |  | 72,8 | 75,7 | 76,5 |
|  | 113,3 | 115,4 | 153,6 | 167,0 | 125,6 | 150,8 | 166,2 |  | 181,2 | 159,9 | 173,1 |

Визначимо ранги для значень . Варіаційний ряд для  має вигляд

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 113,3 | 115,4 | 125,6 | 150,8 | 153,6 | 159,9 | 166,2 | 167,0 | 173,1 | 181,2 |

Таким чином, упорядкованій за елементами  вибірці відповідає наступна послідовність рангів і їх різниць:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 1 | 2 | 5 | 8 | 3 | 4 | 7 | 10 | 6 | 9 |
|  | 0 | 0 | –2 | –4 | 2 | 2 | 0 | –2 | 3 | 1 |

Якщо , то знаходимо



Перевіримо значимість отриманого результату. Знайдемо вибіркове значення статистики



Маємо , отже – рангова кореляція значима.

Питання для самоконтролю

1. Що називається критерієм згоди?
2. Назвіть основні критерії попереднього вибору закону розподілу.
3. Наведіть алгоритм перевірки гіпотези про закон розподілу за критерієм згоди Пірсона.
4. Які обмеження накладаються на теоретичні частоти при перевірці гіпотези за критерієм Пірсона?
5. Наведіть алгоритм перевірки гіпотези про закон розподілу за критерієм Колмогорова.
6. У чому відмінність критеріїв Пірсона й Колмогорова?
7. Які методи математичної статистики називають непараметричними?
8. Які генеральні сукупності називаються однорідними?
9. Для перевірки якої гіпотези служить критерій знаків?
10. Що є статистикою критерію знаків?
11. Яку гіпотезу перевіряє критерій Вілкоксона?
12. Як будується статистика за критерієм Вілкоксона?
13. Дайте визначення серії.
14. Для перевірки якої гіпотези застосовується критерій серій?
15. Що є статистикою критерію серій?
16. Як визначається критична область за критерієм серій?
17. Що характеризує ранговий коефіцієнт кореляції?

Практичне завдання

***Завдання 1.***

Для наданих статистичних розподілів вибірки, записати емпіричну функцію розподілу та обчислити такі числові характеристики: 1) вибіркове середнє, 2) вибіркову дисперсію, 3) підправлену дисперсію, 4) вибіркове середнє квадратичне відхилення, 5) підправлене середнє квадратичне відхилення, 6) розмах вибірки, 7) медіану, 8) моду, 9) коефіцієнт варіації, 10) коефіцієнт асиметрії та 11) ексцес для вибірки.

Вибірка 1

|  |
| --- |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 55 | 46 | 44 | 52 | 52 | 47 | 63 | 51 | 53 | 49 | 51 | 50 | 55 |
| 33 | 57 | 47 | 40 | 53 | 34 | 30 | 54 | 54 | 51 | 65 | 31 | 46 |
| 27 | 49 | 33 | 89 | 19 | 70 | 10 | 25 | 35 | 52 | 32 | 89 | 76 |
| 40 | 85 | 31 | 33 | 10 | 55 | 15 | 24 | 39 | 32 | 15 | 45 | 26 |

Вибірка 2

|  |
| --- |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 147 | 150 | 160 | 162 | 157 | 169 | 142 | 111 | 121 | 127 | 136 | 194 | 101 |
| 153 | 149 | 150 | 149 | 135 | 146 | 139 | 132 | 111 | 131 | 192 | 178 | 195 |
| 148 | 163 | 162 | 141 | 128 | 144 | 160 | 172 | 158 | 184 | 152 | 109 | 196 |
| 155 | 142 | 140 | 146 | 168 | 131 | 152 | 152 | 148 | 121 | 157 | 115 | 198 |

***Завдання 2.***

Дослідити, чи змінюється стійкість уваги у студентів впродовж навчального дня? Діагностувалися показники в тій же групі перед першою та після останньої пари занять. В таблиці представлені результати однієї і тієї ж групи досліджуваних, з якими двічі проводилося психологічне тестування. Змістовно така вибірка називають зв’язаною чи парною. Підстав вважати, що первинні дані відповідають закону нормального розподілу немає. Вибірка досліджуваних недостатньо велика, щоби зробити якісний висновок при перевірці даних на нормальність розподілу.

|  |  |  |
| --- | --- | --- |
| № | Тривалість утримання зусилля на динамометрі (с) | |
| до занять (*Х1*) | після занять (*Х2*) |
| 1 | 12 | 21 |
| 2 | 17 | 17 |
| 3 | 23 | 25 |
| 4 | 15 | 11 |
| 5 | 28 | 21 |
| 6 | 24 | 24 |
| 7 | 21 | 16 |
| 8 | 22 | 26 |
| 9 | 17 | 18 |
| 10 | 26 | 22 |
| 11 | 25 | 20 |
| 12 | 19 | 21 |
| 13 | 13 | 16 |
| 14 | 19 | 8 |

Для виявлення відмінностей між досліджуваними в різний час тестування використати непараметричний критерій T-Вiлкоксона.

***Завдання 3.***

У вибірці курсантів військового училища (юнаки віком від 18 до 20 років) вимірювалася здатність до утримання фізичного вольового зусилля на динамометрі.

Спочатку у випробовуваних вимірювалася максимальна м'язова сила кожної з рук, а на наступний день їм пропонувалося витримувати, на динамометрі з рухомою стрілкою м'язове зусилля, що дорівнює 1/2 максимальної м'язової сили даної руки. Відчувши втому, випробовуваний повинен був повідомити про це експериментатору, але не припиняти досвід, долаючи втому і неприємні відчуття.

Випробування проводилося двічі; спочатку зі звичайною інструкцією, а потім, після того, як випробуваний заповнював опитувальник самооцінки вольових якостей за методикою А.Ц. Пуні, йому пропонувалося уявити собі, що він уже домігся ідеалу в розвитку вольових якостей, і продемонструвати відповідне ідеалу вольове зусилля. Чи підтвердилася гіпотеза експериментатора про те, що звернення до ідеалу сприяє зростанню вольового зусилля? Дані представлені в таблиці.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Тривалість утримання зусилля на динамометрі (с) | | |
| до вимірювання вольових якостей (tдо) | після вимірювання вольових якостей (tпісля) |
| 1 | 64 | 25 |
| 2 | 77 | 50 |
| 3 | 74 | 72 |
| 4 | 95 | 76 |
| 5 | 105 | 67 |
| 6 | 83 | 75 |
| 7 | 73 | 77 |
| 8 | 75 | 71 |
| 9 | 101 | 63 |
| 10 | 97 | 122 |
| 11 | 78 | 60 |

***Завдання 4.****Перевірка статистичних гіпотез*.

*Мета:* оцінити параметри і перевірити гіпотези про відповідність емпіричних даних передбачуваному теоретичному розподілу.

Для наданих статистичних розподілів вибірки (завдання 1) написати програму, яка формує протокол статистичного дослідження для довільних даних на основі критерію асиметрії та ексцесу, результати представити для одержаних вибірок.

Перевірити гіпотезу про згоду закону розподілу статистичних даних із нормальним законом, дослідити властивості показників статистичного зв’язку між випадковими величинами.

Організувати перевірку програмним чином на основі критерію узгодженості Пірсона на відповідність введеної вибірки нормальному та рівномірному розподілу за двома різними рівнями значущості. Вивести на екран вибірку та результати перевірок.

***Завдання 5.*** На рівні значущості α = 0,05, використовуючи критерій Пірсона, перевірити, чи узгоджується гіпотеза про нормальний розподіл генеральної сукупності *X* з емпіричним розподілом вибірки обсягом *n* = 200, заданим таким рядом розподілу частот:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *хі* | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 | 26 |
| *nі* | 13 | 20 | 24 | 21 | 26 | 30 | 25 | 26 | 15 |

***Завдання 6.*** Зміна урожайності при застосуванні одного з препаратів допосівної обробки зерна характеризується такими даними (у центнерах з гектара):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Рік | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
| *Необроблене зерно* | 21 | 18 | 20,6 | 22 | 21,4 | 23,8 | 21,4 | 19,8 | 18,4 |
| *Оброблене зерно* | 22,1 | 19,1 | 19,4 | 22,1 | 21,7 | 24,9 | 21,6 | 20,3 | 18,3 |

Чи можна вважати, що допосівна обробка зерна збільшує урожайність? Взяти α = 0,05.

***Завдання 7.***Продуктивність праці двох змін підприємства характеризується вибірками обсягів 1 *n* =10 та 2 *n* =11:

Перша зміна − 18; 23; 29; 30; 31; 32; 35; 36; 37; 39.

Друга зміна − 24; 30; 31; 32; 33; 34; 36; 38; 39; 42; 43.

На рівні значущості α = 0,1, використовуючи критерій Уілкоксона, перевірити нульову гіпотезу про однакову продуктивність праці обох змін, взявши за альтернативну гіпотезу: продуктивність праці обох змін є різною.

***Завдання 8.*** При підкиданні монети 50 разів послідовність результатів є такою:

Г Г Г Ц Ц Г Г Г Г Г Г Ц Ц Ц Ц Г Г Г Ц Ц Ц Г Ц Г Ц Ц Ц Ц Ц Ц Г Г Г Г Ц Ц Ц Ц Ц Г Г Г Ц Г Г Г Г Ц Ц Ц (Г – випав герб, Ц –випала цифра).

Чи є ця послідовність випадковою вибіркою? Взяти α = 0,05.

Тестові завдання для перевірки знань

Виберіть, принаймні, одну відповідь.

1. Різниця між об'ємом вибірки й числом накладених на неї зв'язків називається:

а) числом ступенів волі;

б) потужністю критерію;

в) рівнем значимості;

г) розмахом вибірки.

2. Для перевірки гіпотези про рівність дисперсій двох генеральних сукупностей використовують критерій

а)Стьюдента; б)Фішера –Снедекора;

в)Колмогорова; г)Хі –квадрат (Пірсона).

3. Попереднє судження про нормальність закону розподілу можна висунути виходячи з наступного:

а) виду гістограми частостей;

б) оцінки для дисперсії;

в) відповідності числам Вестергарда;

г) емпіричного коефіцієнта варіації.

4. Критерій перевірки непараметричної гіпотези, що використовує порівняння значень теоретичної й емпіричної функцій розподілу, має назву:

а)Кочрена; б)Фішера –Снедекора;

в)Колмогорова; г)Хі –квадрат (Пірсона).

5. При перевірці непараметричної гіпотези за допомогою критерію Пірсона відбувається порівняння:

а) точкових оцінок для деяких параметрів розподілу;

б) теоретичних і емпіричних частот;

в) оцінок для дисперсій;

г) значень теоретичної й емпіричної функцій розподілу.

6. Дати визначення нульової гіпотези.

а) гіпотеза, яка припускає відсутність систематичних розбіжностей між невідомими параметрами генеральної сукупності і величиною, яка одержана внаслідок обробки вибірки;

б) гіпотеза, яка протиставляє твердження висунутої гіпотези;

в) гіпотеза, яка має тільки одне припущення;

г) гіпотеза, яка складається із скінченого або нескінченного числа простих гіпотез.

7. Дати визначення альтернативної гіпотези.

а) гіпотеза, яка припускає відсутність систематичних розбіжностей між невідомими параметрами генеральної сукупності і величиною, яка одержана внаслідок обробки вибірки;

б) гіпотеза, яка протиставляє твердження висунутої гіпотези;

в) гіпотеза, яка має тільки одне припущення;

г) гіпотеза, яка складається із скінченого або нескінченного числа простих гіпотез.

8. Що називають простою статистичною гіпотезою?

а) гіпотеза, яка припускає відсутність систематичних розбіжностей між невідомими параметрами генеральної сукупності і величиною, яка одержана внаслідок обробки вибірки;

б) гіпотеза, яка протиставляє твердження висунутої гіпотези;

в) гіпотеза, яка має тільки одне припущення;

г) гіпотеза, яка складається із скінченого або нескінченного числа простих гіпотез.

9. Що називають складною статистичною гіпотезою?

а) гіпотеза, яка припускає відсутність систематичних розбіжностей між невідомими параметрами генеральної сукупності і величиною, яка одержана внаслідок обробки вибірки;

б) гіпотеза, яка протиставляє твердження висунутої гіпотези;

в) гіпотеза, яка має тільки одне припущення;

г) гіпотеза, яка складається із скінченого або нескінченного числа простих гіпотез.

10. Що називають статистичним критерієм?

а) випадкову величину *K* , яка використовується для перевірки нульової

гіпотези;

б) цілком закономірну величину *K* , яка використовується для перевірки

нульової гіпотези;

в) випадкову величину *K* , яка використовується для перевірки домашніх

робіт студентів;

г) інша відповідь.

11. Що називається емпіричним (спостережуваним) значенням критерію?

а) випадкова величина *K* , яка використовується для перевірки нульової

гіпотези;

б) значення критерію, який обчислюють за результатом вибірки;

в) випадкову величину *K* , яка використовується для перевірки домашніх

робіт студентів;

г) інша відповідь.

12. Що називається областю прийняття нульової гіпотези?

а) множину всіх можливих значень статистичного критерію;

б) сукупність значень статистичного критерію, за яких нульова гіпотеза не

відхиляється;

в) сукупність значень статистичного критерію, за яких нульова гіпотеза не

приймається;

г) інша відповідь.

13. Що називається критичною область статистичного критерію?

а) множину всіх можливих значень статистичного критерію;

б) сукупність значень статистичного критерію, за яких нульова гіпотеза не

відхиляється;

в) сукупність значень статистичного критерію, за яких нульова гіпотеза не

приймається;

г) інша відповідь.

14. Що називається критичною точкою статистичного критерію?

а) точка, яка поділяє множину всіх можливих значень статистичного критерію;

б) сукупність значень статистичного критерію, за яких нульова гіпотеза не

відхиляється;

в) сукупність значень статистичного критерію, за яких нульова гіпотеза не

приймається;

г) інша відповідь.

15. Які Ви знаєте види критичних областей статистичного критерію?

а) лівобічні;

б) правобічні;

в) двобічні;

г) всі перераховані.

16. Що таке рівень значущості ?

а) ймовірність, що буде відкинута правильна альтернативна гіпотеза;

б) ймовірність, що буде прийнята неправильна нульова гіпотеза;

в) ймовірність, що буде відкинута правильна нульова гіпотеза;

г) всі перераховані.

17. Помилки першого роду при перевірці гіпотез полягають у тому, що:

а) буде відкинута неправильна нульова гіпотеза;

б) буде прийнята неправильна нульова гіпотеза;

в) буде відкинута правильна нульова гіпотеза;

г) всі перераховані.

18. Помилки другого роду при перевірці гіпотез полягають у тому, що:

а) буде відкинута неправильна нульова гіпотеза;

б) буде прийнята неправильна нульова гіпотеза;

в) буде відкинута правильна нульова гіпотеза;

г) всі перераховані.

19. Що називають потужністю критерію?

а) ймовірність попадання критерію в критичну область при умові, що

справедлива альтернативна гіпотеза;

б) ймовірність того, що нульова гіпотеза буде відкинута, якщо справедлива

альтернативна гіпотеза;

в) відповіді а) і б) правильні;

г) інша відповідь.

# ЗМІСТОВИЙ МОДУЛЬ 4. ДИСПЕРСІЙНИЙ АНАЛІЗ

## Тема 8. Основи дисперсійного аналізу

Мета: розглянути основні принципи однофакторного та двофакторного дисперсійного аналізу, ознайомитися з основними поняттями та етапами його проведення, набути навички використовувати ANOVA для порівняння середніх значень.

🖉Основні терміни і поняття

*Фактор, рівень фактору, залежна змінна, ANOVA,міжгрупова дисперсія, внутрішньогрупова дисперсія, нульова гіпотеза, альтернативна гіпотеза,однофакторний дисперсійний аналіз, ранговий однофакторний аналіз, двофакторний дисперсійний аналіз, ранговий критерій Фрідмана.*

📚Основні теоретичні положення

У практиці обробки результатів спостережень часто постає питання про те, наскільки істотно впливає зміна деякого фактору або групи факторів на вимірювану величину. Наприклад, можна досліджувати спільний вплив декількох економічних факторів, що не піддаються кількісному виміру на досліджуваний економічний показник, можна оцінити вплив властивостей сировини на показники якості продукції, кількості внесених добрив на врожайність і т.п.

У попередній главі була розглянута перевірка гіпотези про рівність математичних очікувань двох сукупностей при невідомих і однакових дисперсіях. Однак на практиці часто виникає необхідність узагальнення задачі, тобто перевірки при заданому рівні значимості за вибірковим середнім нульової гіпотези про рівність математичних очікувань *р (р >*2) сукупностей, у яких також дисперсії невідомі й однакові. Для ефективного розв’язання таких задач застосовується новий підхід, що заснований на порівнянні дисперсій і тому його названо *дисперсійним аналізом.*

Дисперсійний аналіз є сукупністю статистичних методів, призначених для перевірки гіпотез про зв’язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також для встановлення ступеня впливу факторів та їх взаємодії. У спеціальній літературі дисперсійний аналіз часто називають ANOVA (від англомовної назви Analysis of Variations). Вперше цей метод було розроблено Р. Фішером в 1925 р.

*Факторами* називають контрольовані чинники, що впливають на кінцевий результат. *Рівнем фактору, або способом обробки*, називають значення, що характеризують конкретний прояв цього фактору. Ці значення зазвичай подають у номінальній або порядковій шкалі вимірювань. Значення вимірюваної ознаки називають *відгуком*.

Коротко суть цього аналізу зводиться до розділення загальної дисперсії ознаки на компоненти, обумовлені впливом конкретних факторів, і перевірці гіпотез про значимість їх впливу. Моделі дисперсійного аналізу залежно від числа факторів класифікують як *однофакторні, двофакторні* й т.п. За метою дослідження виділяють такі моделі: *детермінована* – тут рівні всіх факторів заздалегідь фіксовані й перевіряють саме їхній вплив, *випадкова* – тут рівні кожного фактору отримані як випадкова вибірка з генеральної сукупності рівнів фактору, і *змішана* – тут рівні одних факторів заздалегідь фіксовані, а рівні інших – випадкова вибірка.

Часто вихідні значення факторів вимірюють у кількісних або порядкових шкалах. Тоді постає проблема групування вихідних даних у ряди спостережень, що відповідають приблизно однаковим значенням фактору. Якщо кількість груп взяти надмірно великою, то кількість спостережень у них може виявитися недостатньою для отримання надійних результатів. Якщо її взяти надмірно малою, це може призвести до втрати суттєвих особливостей впливу досліджуваного фактору на систему. Загальну методологію групування описано в розділі 1. Вибір конкретного способу групування даних залежить від їх обсягу і характеру варіювання значень фактору.

Кількість і розміри інтервалів при однофакторному аналізі найчастіше визначають за принципом рівних інтервалів або за принципом рівних частот. При багатофакторному аналізі застосовують три типи групування:

* групи з рівною кількістю спостережень;
* групи з різною кількістю спостережень;
* групи, кількості спостережень у яких відповідають певній пропорції.

При цьому існують певні особливості обробки даних, залежно від типу групування.

**8.1 Однофакторний дисперсійний аналіз**

Основною метою однофакторного аналізу зазвичай є оцінка величини впливу конкретного фактору на досліджуваний відгук. Іншою метою може бути порівняння двох або декількох факторів один з одним з метою визначення різниці їх впливу на відгук, яку часто називають *контрастом факторів.* Попереднім етапом є перевірка нульової гіпотези про відсутність будь-якого впливу досліджуваного фактору (факторів), тобто гіпотези про те, що зміни значень ознаки в порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності.

Якщо нульову гіпотезу відкидають, то наступним етапом є кількісне оцінювання впливу досліджуваного фактору і побудова довірчих інтервалів для отриманих характеристик. У випадку, коли нульова гіпотеза не може бути відкинутою, зазвичай її приймають і роблять висновок про відсутність впливу. Але, якщо є підстави вважати, що такий вплив має бути присутнім (наприклад, це може випливати з теоретичних уявлень про об’єкт дослідження), то необхідно перевірити наявність інших факторів, що можуть його маскувати.

При *однофакторному дисперсійному аналізі* вихідні дані подають у вигляді таблиць, у яких кількість стовпчиків дорівнює кількості рівнів фактору, а кількість значень у кожному стовпчику - кількості спостережень при відповідному рівні фактору (табл. 8.1). Для різних рівнів фактору кількість спостережень може бути різною. При цьому виходять з припущення, що результати спостережень для різних рівнів є вибірками з нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими і не залежать від рівнів. Завданням аналізу є перевірка нульової гіпотези про рівність середніх значень сукупностей, що розглядаються.

В основі однофакторного дисперсійного аналізу лежить така теоретико – імовірнісна модель:

 (8.1)

де – значення ознаки *X,* отримане при – му випробуванні на – му рівні фактору; – загальна середня величина ознаки *X;* *–* ефект фактору на – му рівні; – випадковий компонент, що впливає на значення ознаки *X* в –му спостереженні на –му рівні фактору. Приймається припущення, що взаємно незалежні й мають нормальний закон розподілу *.*

**8.2 Однакове число випробувань на різних рівнях**

Нехай на кількісно нормально розподілену ознаку (випадкову величину) *X* впливає фактор *F,* що має *р* постійних рівнів. Будемо припускати, що число спостережень (випробувань) на кожному рівні однакове й дорівнює *q.*

Нехай спостерігалося *п = pq* значень  ознаки *X,* де, *i –* номер випробування *–* номер рівня фактору. Результати спостережень наведені в таблиці 8.1.

Таблиця 8.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Номер  випробування | Рівні фактору | | | |
|  |  |  | … |  |
| *1* |  |  | … |  |
| *2* |  |  | … |  |
| *…* | … | … | … |  |
| *q* |  |  | … |  |
| Групова  середня |  |  | … |  |

Позначимо через – групове середнє *j-ї* групи*.*

Необхідно перевірити, чи суттєво впливає деякий якісний фактор *F*, щомає *р* рівнів **, ,…,,**на досліджувану ознаку *Х.*

Якщо вважати, що елементи стовпців таблиці 8.1, позначених*****,* без елементів груповий середньої, є чисельними значеннями випадкових величин *,* щомають нормальний закон розподілу з математичними очікуваннями відповідно  й однаковими дисперсіями, то дана задача зводиться до перевірки нульової гіпотези ,здійснюваної в дисперсійному аналізі

 (8.2)

а через *– загальне середнє*

 (8.3)

тому що 

Розглянемо суму квадратів відхилень спостережуваних значень від загальної середньої :

(8.4)

причому



оскільки



Тому, взявши до уваги, що



ми можемо основну тотожність (8.4) записати в такому вигляді:



або в скороченому виді

= + (8.5)

В (8.5)  – *загальна сума* квадратів відхилень спостережуваних значень від загальної середньої,  – *факторна сума* квадратів відхилень групових середніх від загальної середньої, котра характеризує розсіювання між групами,  – *залишкова сума* квадратів відхилень спостережуваних значень групи від своєї групової середньої, котра характеризує розсіювання усередині групи.

Практично залишкову суму знаходять за рівнянням

= – (8.6)

яке випливає з (8.5).

Одержимо більше зручні формули для розрахунків  і :

 (8.7)

 (8.8)

де  – сума квадратів значень ознаки на рівні ;

 – сума значень ознаки на рівні 

Зауваження 1. Обчислення за формулами (8.7) і (8.8) можна спростити, якщо виконати заміну змінних , де *С* приблизно дорівнює загальній середній.

Підставляючи в ці формули  й = = , де , з урахуванням позначення , одержимо

 (8.9)

 (8.10)

Зауваження 2. Якщо спостережувані значення  – десяткові дроби з *k* знаками після коми, то доцільно перейти до чисел , де *С* – приблизно середнє значення чисел . Хоча при цьому факторна й залишкова дисперсії збільшуються в  разів, їхнє відношення не зміниться.

У розкладанні (8.5) закладена основна ідея дисперсійного аналізу. Однак у ньому аналізуються не самі суми квадратів відхилень, а так названі середні квадрати, що є незміщеними оцінками відповідних дисперсій, які виходять діленням сум квадратів відхилень на відповідне число ступенів волі.

Нагадаємо, що *число ступенів волі* визначається як загальне число ступенів спостережень мінус число об’єднуючих їхніх рівнянь. Тому число ступенів волі загальної дисперсії дорівнює *pq–*1 *= n–*1*,* тому що один ступінь волі втрачається при визначенні середньої. Аналогічне число ступенів волі факторної дисперсії дорівнює *р*–1, тому що групові середні варіюють навколо однієї загальної середньої. Нарешті, число ступенів волі залишкової дисперсії дорівнює *п-р*, тому що використовується *р* співвідношень при обчисленні *р* групових середніх.

Використовуючи отримані значення сум квадратів і чисел ступенів волі, можна обчислити незміщені оцінки трьох дисперсій:

 (8.11)

Зауваження 3. Число ступенів волі залишкової дисперсії *pq–p* дорівнює різниці між числами ступенів волі загальної й факторної дисперсій. Дійсно,

*pq–*1*–(p–*1*) = pq–p = n–p.*

Перевірка гіпотези *Н0*про рівність групових математичних очікувань ґрунтується на порівнянні дисперсій  і . Виявляється, якщо гіпотеза *Н0*вірна, то вірна й гіпотеза про рівність факторної й залишкової дисперсій, що перевіряється за критерієм Фішера – Снедекора

(8.12)

який має ступені волі *k1= р–*1і *k2= п-р.*

Якщо нульова гіпотеза про рівність групових середніх помилкова, то помилково й гіпотезу про рівність  і . Справедливі також і зворотні твердження.

Наведемо схему перевірки нульової гіпотези у випадку правобічної критичної області:

(8.13)

Зауваження 4. Якщо <, то звідси виникає справедливість нульової гіпотези *H0* і, виходить, немає необхідності використовувати критерій *F.*

Зауваження 5. Якщо немає впевненості в справедливості припущення про рівність дисперсій розглянутих *р* сукупностей, то це припущення варто перевірити попередньо, наприклад за критерієм Кочрена.

Приклад.Зроблено по 4 випробування на кожному із чотирьох рівнів. Результати випробувань наведені в табл. 8.2. Методом дисперсійного аналізу при рівні значимості *α*= 0,05 перевірити нульову гіпотезу про рівність групових середніх. Передбачається, що вибірки взяті з нормальних сукупностей.

Знайдемо за формулами



групові вибіркові середні й дисперсії, результати обчислень яких занесемо в таблицю 8.2.

Таблиця 8.2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Номер випробування | Рівні фактору*Fj* | | | |
| *i* |  |  |  |  |
| 1 | 140 | 150 | 148 | 150 |
| 2 | 144 | 149 | 149 | 155 |
| 3 | 142 | 152 | 146 | 154 |
| 4 | 145 | 150 | 147 | 152 |
| Групове середнє, | 142,75 | 150,25 | 147,5 | 152,75 |
| Вибіркова дисперсія, | 3,6875 | 1,1875 | 1,25 | 3,6875 |

Перш ніж проводити дисперсійний аналіз, переконаємося в тому, що нульова гіпотеза



про рівність групових дисперсій для вибірок, взятих з нормальних генеральних сукупностей, не суперечить результатам спостережень. Оскільки вибірки містять однаковий об'єм *q =* 4, то при перевірці нульової гіпотези доцільно скористатися критерієм Кочрена. Для цього:

1. Знайдемо виправлені вибіркові дисперсії:

.

2. Знайдемо спостережуване значення критерію Кочрена:

.

3. Знайдемо з таблиці (додаток 8), за рівнем значимості *α*= 0,05, число ступенів волі *k* = 4 –1 = 3 і число вибірок *l = 4* критичну точку *G*(0,05; 3; 4) = g*кр =* 0,6841.

4. Оскільки *Gнабл*< g*кр*, то гіпотезу *Н0* приймаємо.

Гіпотеза, що перевіряється в дисперсійному аналізі, має вигляд:

*Н0: т1=т2=т3=т4*,

де *mi –* математичне очікування *i–ої* генеральної сукупності.

Для спрощення розрахунку віднімемо *С*=148 з кожного спостережуваного значення: **.Складемо розрахункову таблицю 8.3.

Таблиця 8.3

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Номер  випробування | Рівні фактору*Fj* | | | | | | | | Підсумковий  стовпець |
| *i* |  | |  | |  | |  | |
|  |  |  |  |  |  |  |  |
| 1 | –8 | 64 | *2* | 4 | 0 | 0 | *2* | 4 | – |
| 2 | –4 | 16 | 1 | 1 | 1 | 1 | 7 | 49 |
| 3 | –6 | 36 | 4 | 16 | –2 | 4 | 6 | 36 |
| 4 | –3 | 9 | 2 | 4 | –1 | 1 | 4 | 16 |
|  | – | 125 | – | 25 | – | 6 | – | 105 |  |
|  | –21 | – | 9 | – | –2 | – | 19 | – |  |
|  | 441 | – | 81 | – | 4 | – | 361 | – |  |

Користуючись таблицею 8.3 і з огляду на те, що число рівнів фактору *p* = 4, число випробувань на кожному рівні *q* = 4, знайдемо заформулами (8.9) і (8.10) загальну й факторну суми квадратів відхилень:

=261–25/16 = 259,4375;

=887/4–25/16 = 220,1875.

Знайдемо залишкову суму квадратів відхилень:

= – = 39, 25.

Факторну й залишкову дисперсії знайдемо за формулами (8.11):

= 220,1875/3 73,4;

= 39,25/(16 –4) 3,27.

Перевірку гіпотези *H*0 про рівність групових математичних очікувань проведемо за схемою (8.13):

 Таким чином, групові середні істотно розрізняються. Якщо потрібно зрівняти середні попарно, то варто скористатися критерієм Стьюдента.

**8.3 Неоднакове число випробувань на різних рівнях**

Нехай число випробувань на різних рівнях різне, а саме: зроблено *q*1випробувань на рівні *F*1, *q2*випробувань – на рівні *F*2*,* ..., *qp*випробувань – на рівні *Fp .* У цьому випадку загальну суму квадратів відхилень також знаходять за формулою (15.7), де , , …,  – суми квадратів значень, що спостерігалися, ознаки відповідно на *рівнях F*1, *F*2, …, *Fp*; , , … , –суми значень, що спостерігалися, ознаки відповідно на *рівнях F*1, *F*2, …, *Fp*; *n*=*q1+q2+…+qp* – об'єм вибірки, а факторну суму квадратів відхилень знаходять за формулою

 (8.14)

Якщо для спрощення обчислень уводиться заміна змінних , де *С* приблизно дорівнює загальній середній, то формула для загальної суми квадратів відхилень має вигляд (8.9), де



а формула для факторної суми відхилень має вигляд

 (8.15)

Інші обчислення здійснюють, як і у випадку однакового числа випробувань за формулами (8.6) і (8.11).

Приклад.Проведено 22 випробування, з яких 7 на першому рівні фактору, 6 – на другому, 5 – на третьому й 4 – на четвертому. Результати випробувань наведені в табл. 3.4.Методом дисперсійного аналізу при рівні значимості *α*= 0,05 перевірити нульову гіпотезу про рівність групових середніх. Передбачається, що вибірки взяті з нормальних сукупностей з однаковими дисперсіями.

Таблиця 8.4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Номер випробування | Рівні фактору*.* | | | |
| *i* |  |  |  |  |
| 1 | 1,3 | 1,4 | 1,44 | 1,27 |
| 2 | 1,27 | 1,3 | 1,4 | 1,05 |
| 3 | 1,21 | 1,28 | 1,28 | 1,24 |
| 4 | 1,09 | 1,27 | 1,28 | 1,22 |
| 5 | 1,03 | 1,24 | 1,06 |  |
| 6 | 1,01 | 1,08 | — |  |
| 7 | 1,09 | – | – | – |

Для спрощення розрахунку помножимо кожне спостережуване значення на 100 і віднімемо *З =* 122 (зауваження 2). Складемо розрахункову таблицю 8.5.

Використовуючи таблицю 8.5, знайдемо загальну й факторну суми квадратів відхилень:





Знайдемо залишкову суму квадратів відхилень:

= – = 2528,071 .

Факторну й залишкову дисперсії знайдемо за формулами (8.11):

 *=* /(*p*–1) = 804,529/3  268,176;

= /(*n*–*p*) =  2528,071/18 140,448.

Таблиця 8.5

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Номер  випробування | Рівні фактору*.* | | | | | | | | Підсумковий  стовпець |
| *i* |  | |  | |  | |  | |
|  |  |  |  |  |  |  |  |
| 1 | 8 | 64 | 18 | 324 | 22 | 484 | 5 | 25 |  |
| 2 | 5 | 25 | 8 | 64 | 18 | 324 | –17 | 289 |
| 3 | – 1 | 1 | 6 | 36 | 6 | 36 | 2 | 4 |
| 4 | –13 | 169 | 5 | 25 | 6 | 36 | 0 | 0 |
| 5 | –19 | 361 | 2 | 4 | –16 | 256 | – | – |
| 6 | –21 | 441 | – 14 | 196 | – | – | – | – |
| 7 | –13 | 169 |  |  |  |  |  |  |
|  | – | 1230 | – | 649 | – | 1136 | – | 318 |  |
|  | –54 | – | 25 | – | 36 | – | –10 | – |  |
|  | 2916 | – | 625 | – | 1296 | – | 100 | – | – |

Порівняємо  й  за критерієм Фішера–Снедекора, використовуючи схему (8.13)



Непараметричним аналогом однофакторного дисперсійного аналізу є *ранговий однофакторний аналіз Краскела – Уолліса.* Він розроблений американськими математиком Вільямом Краскелом та економістом Вільсоном Уоллісом в 1952 р. Цей критерій призначено для перевірки нульової гіпотези про рівність ефектів впливу на досліджувані вибірки з невідомими, але рівними середніми. При цьому кількість вибірок має бути більшою ніж дві. Нульова гіпотеза полягає в тому, що k вибірок обсягами 1 2 , , ..., *k* отримані з однієї і тієї самої генеральної сукупності. Критерій Краскела – Уолліса є узагальненням *U*-критерію Манна – Уїтні на випадок, коли кількість вибірок *k*> 2.

Рангові методи, у тому числі й метод Краскела – Уолліса, не передбачають нормальності розподілу результатів спостережень і можуть застосовуватися як для кількісних даних з невідомим законом розподілу, так і для порядкових ознак.

*Критерій Джонкхієра (Джонкхієра – Терпстра)* запропонований незалежно один від одного нідерландським математиком Т.Дж. Терпстрою в 1952 р. й британським психологом Е.Р. Джонкхієром в 1954 р. Його застосовують тоді, коли заздалегідь відомо, що наявні групи результатів упорядковані за зростанням впливу досліджуваного фактору, який вимірюють у порядковій шкалі. Таблиця даних має такий самий вигляд, як і в попередньому випадку. Будемо вважати, що її перший стовпчик відповідає найменшому рівню фактору, другий – наступному за величиною тощо, останній стовпчик відповідає найбільшому рівню. При виконанні таких припущень критерій Джонкхієра є більш потужним, ніж критерій Краскела – Уолліса, стосовно гіпотези про монотонний вплив фактору.

*М-критерій Бартлетта* запропонований британським статистиком Маурісом Стівенсоном Бартлеттом в 1937 р. Його застосовують для перевірки нульової гіпотези про рівність дисперсій кількох нормальних генеральних сукупностей, з яких взяті досліджувані вибірки, що у загальному випадку мають різні обсяги (обсяг кожної вибірки має бути не менше чотирьох).

*G-критерій Кокрена (Кочрена)* запропонований американським статистиком Вільмом Геммелом Кочреном в 1941 р. Його використовують для перевірки нульової гіпотези про рівність дисперсій k (k ≥ 2) нормальних генеральних сукупностей за незалежними вибірками рівного обсягу.

Непараметричний *критерій Левене,* запропонований американським математиком Ховардом Левене в 1960 р. є альтернативою критерію Бартлетта в умовах, коли немає впевненості у тому, що досліджувані вибірки підпорядковуються нормальному розподілу.

В 1974 р. американські статистики Мортон Б. Браун та Алан Б. Форсайт запропонували більш робастний тест *(критерій Брауна – Форсайта).*

Розглянуті вище критерії дають змогу встановити різницю дисперсій сукупностей, але не дають можливості дати кількісну оцінку впливу фактору на досліджувану ознаку, а також встановити, для яких саме сукупностей дисперсії є різними.

Для встановлення кількісного впливу досліджуваного фактору часто застосовують адитивну модель, яка передбачає, що значення відгуку є сумою впливу фактору і незалежної від нього випадкової величини.

Якщо гіпотезу про рівність середніх відхиляють, то наступним кроком може бути визначення вибірок, для яких ця різниця є суттєвою. Для цього використовують метод лінійних контрастів.

*Лінійним контрастом* у моделі адитивного впливу фактору на відгук називають лінійну функцію середніх значень k незалежних нормальних вибірок з невідомими рівними дисперсіями.

Для встановлення вибірок, що належать певній множині даних, дисперсії яких є різними, найчастіше застосовують *метод множинних порівнянь (Шеффе),* запропонований американським статистиком Генрі Шеффе.

**8.4 Поняття про двофакторний дисперсійний аналіз**

В основі двофакторного дисперсійного аналізу лежить наступна теоретико – імовірнісна модель:

 (8.16)

де *xijk*– значення ознаки *X* в *k-му* спостереженні на *i-му* рівні фактору*А* і на *j-му* рівні фактору*В; *–загальна середня величина ознаки *X;αi –* ефект впливу фактору *А* на *i-му* рівні; *βj* – ефект впливу фактору *В* на *j-му* рівні; γ*ij –* ефект спільного впливу факторів (позначимо *А×В); εjjk*– незалежні нормально розподілені випадкові компоненти *N*(0;*σ2*), що являють собою відхилення ознаки від відповідних середніх.

Нехай необхідно виявити вплив двох факторів *(А* і *В)* та їхню взаємодію на деяку ознаку *X.* Спостереження проводяться при фіксованих рівнях факторів *A* і *В.* Оскільки для кожного поєднання факторів спостереження повторюється *п* разів, то відповідно одержимо *п* значень ознаки. Дані спостережень зручно подати у вигляді таблиці 3.6, у якій значення ознаки *X* позначене через *xijk*, де *(q –* число спостережень фактору*A);  (р –* число спостережень фактору*В);  (k –* порядковий номер спостереження для кожного поєднання рівнів).

Таблиця 8.6

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Рівень  фактору*А* | Рівень фактору *В* | | | | Сума |
|  |  | **…** |  |  |
|  | ,  , …, | ,  , …, | **…** | ,  , …, |  |
|  | ,  , …, | ,  , …, | **…** | ,  , …, |  |
| … | … | … | **…** | … | … |
|  | ,  , …, | ,  , …, | **…** | ,  , …, |  |
| Сума *Rj* |  |  | **…** |  |  |

За даними таблиці одержимо такі середні: загальна середня

 (8.17)

середні за рядками

 (8.18)

середні за стовпчиками

 (8.19)

середні для кожного окремого блоку таблиці

 (8.20)

Суми квадратів відхилень від загальної середньої  розкладемо на складові. Для цього представимо  в еквівалентній формі



Оскільки всі перехресні добутки, при піднесенні правої частини останнього вираження у квадрат, дорівнюють нулю, то в підсумку одержимо



або в скороченому вигляді

*S=S1+S2+S3+ S4.* (8.21)

У (8.21) чотири складові: сума квадратів, пов'язана із впливом фактору*В,* фактору*А,* їхньої взаємодії, і складова, що характеризує залишкову суму квадратів (суму квадратів усередині кожного блоку таблиці).

Загальне число ступенів волі, мабуть, дорівнює *N–1,* де *N* – загальне число спостережень (*N* = *qpn)*. Число ступенів волі між стовпцями дорівнює *р* –1, між рядками *q* –1, для взаємодії (*р –* 1)(*q* – 1), усередині осередків *pq*(*n* – 1) = *N* – *pq*.Для перевірки знайдемо суму *р –*1 + *q –*1 + +(*p–*1)(*q –*1)*+ N – pq = N –*1*.*

Схема дисперсійного аналізу, заснована на отриманих вище даних, представлена в таблиці 8.7.

Таблиця 8.7

|  |  |  |  |
| --- | --- | --- | --- |
| Фактор | Сума квадратів | Число ступенів  волі | Оцінка дисперсії |
| *В* |  | *p–1=k1* |  |
| *А* |  | *q–1=k2* |  |
| *А×В* |  | *(p–1) ×(q–1)=k3* |  |
| Залишковий |  | *N–pq=k4* |  |
| Сума |  | *N–1* |  |

Оскільки при проведенні аналізу інтерес являє вплив кожного фактору порізно й вплив їхньої взаємодії, то знаходимо відповідно три значення:

 (8.22)

Необхідні для аналізу суми квадратів відхилень (8.21) можемо одержати також за такими формулами:

 (8.23)

 (8.24)

 (8.25)

 (8.26)

де

 (8.27)

При рівні значимості  визначаємо критичні точки:



Якщо:

*,* то нульова гіпотеза про відсутність впливу фактору *В* відхиляється;

*,* то нульова гіпотеза про відсутність впливу фактору *А* відхиляється;

**, то нульова гіпотеза про відсутність спільного впливу факторів *A* і *В* відхиляється.

Приклад.При рівні значимості α= 0,05 перевірити, чи існує вплив факторів *А* і *В,* атакож їх спільного впливу на ознаку *X,* для результатів випробувань, наведених у таблиці 8.8.

Таблиця 8.8

|  |  |  |  |
| --- | --- | --- | --- |
| Рівень  фактору*А* | Рівень фактору*В* | | |
| ***В1*** | ***В2*** | ***В3*** |
| ***А1*** | 10; 8; 7; 10 | 8; 12; 14; 12 | 15; 8; 10; 10 |
| ***А2*** | 12; 8; 8; 7 | 12; 13; 11; 14 | 13; 15; 12; 10 |

Використовуючи таблицю 8.8 і формули (8.23) ÷ (8.27), складемо таблицю 8.9 і обчислимо суми квадратів відхилень .

Таблиця 8.9

| ***В***  ***А*** | ***В1*** | | ***В2*** | | ***B3*** | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| ***A1*** | 10 | 100 | 8 | 64 | 15 | 225 | **–** | ***–*** | – |
| 8 | 64 | 12 | 144 | 8 | 64 |
| 7 | 49 | 14 | 196 | 10 | 100 |
| 10 | 100 | 12 | 144 | 10 | 100 |
|  | 35 | **–** | 46 | **–** | 43 | **–** | **–** | 124 | 15376 |
|  | 1225 | **–** | 2116 | **–** | 1849 | **–** | 5190 | **–** | **–** |
|  | – | 313 | – | 548 | – | 489 | – | – | – |
|  | 6,75 | | 19 | | 26,75 | | = 52,5 | | |
| ***А2*** | 12 | 144 | 12 | 144 | 13 | 169 | – | – | – |
| 8 | 64 | 13 | 169 | 15 | 225 |
| 8 | 64 | 11 | 121 | 12 | 144 |
| 7 | 49 | 14 | 196 | 10 | 100 |
|  | 35 | – | 50 | – | 50 | – | – | 135 | 18225 |
|  | 1225 | – | 2500 | – | 2500 | - | 6225 | - | - |
|  | – | 321 | – | 630 | – | 638 | – | – | – |
|  | 14,75 | | 5 | | 13 | | = 32,75 | | |
|  | 70 | – | 96 | – | 93 | – | – | G=259 | = 33601 |
|  | 4900 | - | 9216 | - | 8649 | - | =22765 | |

З таблиці 8.9 маємо:



Дані дисперсійного аналізу наведемо в таблицю 8.10.

Таблиця 8.10

|  |  |  |  |
| --- | --- | --- | --- |
| Джерело варіації |  | Число ступенів  волі | Оцінка дисперсії |
| Фактор *В* | 50,28 | 2 | 25,29 |
| Фактор *А* | 5,04 | 1 | 5,04 |
| *А×В* | 3,083 | 2 | 1,54 |
| Залишкова варіація | 85,25 | 18 | 4,74 |
|  | 143,95 | 23 | – |

Визначимо спостережувані значення критерію:

**= 25,29/4,74 ≈ 5,34; **=5,04/4,74 ≈ 1,06;

**= 1,54/4,74≈0,32

і порівняємо їх з відповідними критичними значеннями:

**(0,05; 2; 18) *= =* 3,55; **(0,05; l; 18) = **=4,41;

**(0,05; 2; 18) *= *= 3,55.

Оскільки *>*,то нульова гіпотеза про відсутність впливу фактору *В* відхиляється й варто зробити висновок про значимості впливу цього фактору. Оскільки*>*і **>, то немає підстави для відхилення відповідних нульових гіпотез.

Якщо припущення, необхідні для застосування двофакторного дисперсійного аналізу, не виконуються, то використовують непараметричний *ранговий критерій Фрідмана (Фрідмана, Кендалла та Сміта),* розроблений американським економістом Мілтоном Фридманом наприкінці 1930р. Цей критерій не залежить від типу розподілу. Передбачається лише, що розподіл величин ij ε є однаковим і неперервним, а самі вони незалежні одна від одної.

У разі, коли в рядках вихідної таблиці є однакові значення, необхідно використовувати середні ранги. При цьому точність висновків буде тим гіршою, чим більшою є кількість таких збігів.

*Q-критерій Кочрена* запропонований В. Кочреном в 1937р. Його використовують у випадках, коли групи однорідних суб’єктів піддаються впливам, кількість яких перевищує два, і для яких можливі два варіанти відгуків – умовно-негативний (0) та умовно-позитивний (1). Нульова гіпотеза полягає в рівності ефектів впливу.

Двофакторний дисперсійний аналіз дає можливість визначити існування ефектів обробки, проте не дає змоги встановити, для яких саме стовпців існує цей ефект.

При вирішенні цієї проблеми застосовують метод множинних порівнянь Шеффе для пов’язаних вибірок.

**8.5****Використання вбудованих статистичних функцій з пакету аналізу MS Excel**

Для перевірки статистичних гіпотез можуть бути використані наступні вбудовані функції:

* ТТЕСТ – для визначення імовірності того, що дві вибірки узяті з генеральних сукупностей з однаковим математичним очікуванням;
* ZTECT– для визначення імовірності того, що вибірка узята з визначеної нормально розподіленої генеральної сукупності. Можна використовувати цю функцію щоб оцінити імовірність того, що конкретне спостереження взяте з конкретної генеральної сукупності;
* СТЬЮДРАСП – для розрахунку критичних значень розподілу Стьюдента (*t*-розподілу);
* НОРМРАСП – повертає нормальну функцію розподілу для зазначеного середнього і стандартного відхилення;
* ДОВІРИТЬ – повертає довірчий інтервал для середньо визначеної нормально розподіленої генеральної сукупності;
* ФТЕСТ – повертає однобічну імовірність того, що дисперсії вибірок розрізняються;
* ХИ2ТЕСТ – повертає значення для розподілу хі-квадрат (використовується як критичне значення X2 -критерій Пірсона).

У засобах статистичного аналізу, що викликаються командою Аналіз даних меню Сервіс, для перевірки статистичних гіпотез пропонуються наступні інструменти:

* Двовибірковий F-тест для дисперсіїдозволяє перевірити гіпотезу про рівність дисперсій для двох вибірок;
* Парний двовибірковий t-тест для середніх: для перевірки гіпотези про рівність середніх для двох вибірок за допомогою *t*-критерію за умови рівності обсягу вибірок;
* Двовибірковий t-тест з однаковими дисперсіями: двовибірковий t-тест Стьюдента служить для перевірки гіпотези про рівність середніх для двох вибірок за умови рівності дисперсій;
* Двовибірковий t-тест із різними дисперсіями: двовибірковий *t*-тест Ст’юдента використовується для перевірки гіпотези про рівність середніх для двох вибірок за умови неоднорідності дисперсій. Вигляд вікон для *t*-тестів аналогічний як і для *F*-тесту;
* Z-тест: двовибірковий z-тест для середніх з відомими дисперсіями використовується для перевірки гіпотези про розходження між середніми двох генеральних сукупностей за умови, що дисперсії відомі, але гіпотези про їхню однорідність не перевірялися.

Інструменти дисперсійного аналізу доступні через команду *Аналіз даних меню Сервіс*. Існує кілька видів дисперсійного аналізу. Необхідний варіант вибирається з урахуванням числа факторів і наявних вибірок з генеральної сукупності.

Однофакторний дисперсійний аналіз використовується для перевірки гіпотези про подібність середніх значень двох чи більше вибірок, що належать одній генеральної сукупності.

Двофакторний дисперсійний аналіз з повтореннями представляє собою більш складний варіант дисперсійного аналізу з декількома вибірками для кожної групи даних, його називають також дисперсійним аналізом при класифікації з групуванням. Двофакторний дисперсійний аналіз без повторення представляє собою двофакторний аналіз дисперсії, що не включає більше однієї вибірки на групу, його називають також дисперсійним аналізом при класифікація з пересічними факторами.

Питання для самоконтролю

1. У чому суть дисперсійного аналізу?

2. Наведіть класифікацію моделей дисперсійного аналізу за числом факторів і за метою дослідження.

3. Запишіть теоретико – ймовірнісну модель для однофакторного дисперсійного аналізу.

4. Що називається груповою і загальною середньою? Назвіть формули для їхнього обчислення.

5. Вивести основну тотожність для суми квадратів відхилень спостережуваних значень від загальної середньої .

6. Вивести формули для розрахунку й через *Pj*і *Rj*.

7. Вивести формули для розрахунку й через *Qj*. і *Tj*.

8. Наведіть формули для обчислення незміщених оцінок трьох дисперсій: , , 

9. Наведіть схему для перевірки нульової гіпотези про рівність групових середніх.

10. Наведіть основні й спрощені формули для розрахунку  й при неоднаковому числі випробувань на різних рівнях.

11. Запишіть теоретико – ймовірнісну модель для двофакторного дисперсійного аналізу.

12. Наведіть формули для обчислення середніх при двофакторному дисперсійному аналізі.

13. Дайте характеристику чотирьох складового розкладання суми квадратів відхилень від загальної середньої.

14. Наведіть формули для розрахунку S*1*,, *S2, S3*і *S4.*

Практичне завдання

***Завдання 1.***

Для перевірки впливу гучності сигналу на швидкість реакції випадковим чином відібрали 3 групи піддослідних. Першій групі (5 чоловік) пред'являли звуковий сигнал в 10 дБ, другій (6 осіб) - 30 дБ, третій (4 людини) - 50 дБ. У піддослідних кожної групи фіксували час реакції в мілісекундах.

Сформулювати гіпотезу за даними умови і перевірити її.

|  |  |  |  |
| --- | --- | --- | --- |
| Номер групи | **1** | **2** | **3** |
| результати  вимірювань | 304 | 272 | 223 |
| 268 | 264 | 184 |
| 272 | 256 | 209 |
| 262 | 269 | 183 |
| 283 | 285 |  |
|  | 247 |  |

***Завдання 2.***

Досліджували відмінність продуктивності відтворення одного і того ж навчального матеріалу у трьох групах випробовуваних (по 5осіб), які розрізнялися умовами пред’явлення цього матеріалу для запам’ятовування. Результати обстеження наведені в таблиці.

|  |  |  |  |
| --- | --- | --- | --- |
| Обстежувана особа, № з/п | Умова 1 | Умова 2 | Умова 3 |
| 1 | 5 | 8 | 11 |
| 2 | 4 | 7 | 9 |
| 3 | 3 | 6 | 7 |
| 4 | 6 | 9 | 10 |
| 5 | 7 | 5 | 8 |

Визначити, чи впливають умови пред’явлення навчального матеріалу на продуктивність його відтворення?

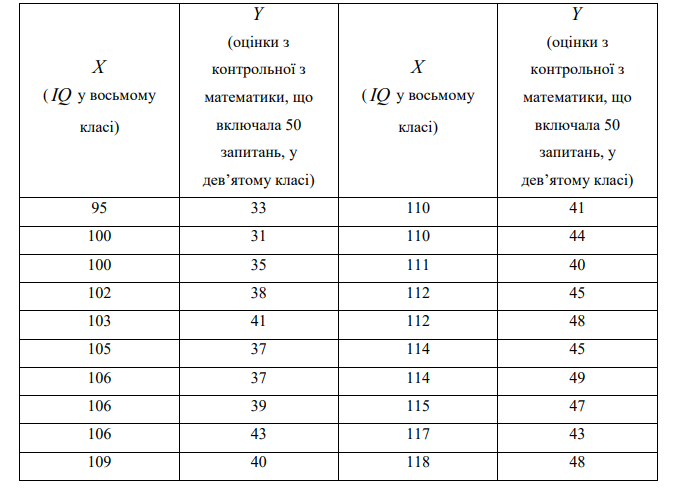
***Завдання 3.***

Дослідник хоче з’ясувати, чи є різними з огляду ефективності чотири методи вивчення деякої теми. Він планує запропонувати 10 студентам вивчити грошову систему Англії за коротким конспектом, 10 інших учнів будуть конспектувати книгу; ще 10 – ознайомляться з програмованим підручником з цього питання; решта 10 учнів будуть вивчати матеріал за допомогою комп’ютера. Чотири варіанти умов (обробок) мають на увазі різний ступінь активності, що стосується учнів, який є предметом спостереження. Один фактор у цьому експерименті – «активність учня»; цей фактор – поняття, що належить до чотирьох умов. Дослідник хоче знати, чи будуть у крайньому випадку два варіанти умов вивчення визначати різну успішність (виміряну тестом з багатоваріантним вибором). Дані результатів тесту з багатоваріантними відповідями 40 учнів подано у таблиці.



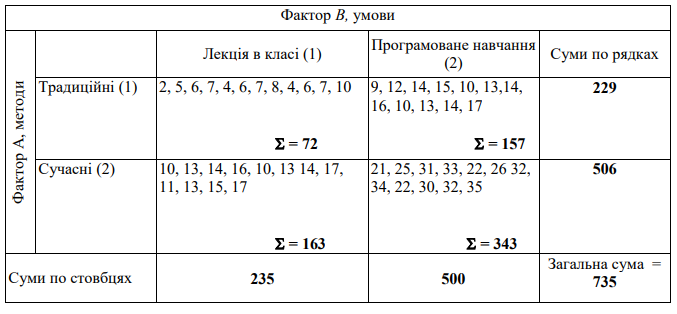
***Завдання 4.***

Спрогнозувати успішність з математики у дев’ятому класі (залежна змінна *Y* ) за результатами групового інтелектуального тесту, проведеного в кінці восьмого класу (незалежна змінна *X* )». У наслідок дослідження для 20-и учнів було отримано *IQ* у восьмому класі ( *X* ) та оцінки з контрольної з математики, що включала 50 запитань, у дев’ятому класі ( *Y* ).



***Завдання 5.***

Із сукупності 2500 десятих класів, які вивчають геометрію, для участі в експерименті було вибрано випадковим чином 48 класів. Дослідник хотів оцінити ефективність 2 різних методів і навколишніх умов в процесі вивчення геометрії, а також визначити їх взаємодії. Випадковим чином класи приписувалися в рівних кількостях чотирьом комбінаціям навколишніх умов і методу. Після закінчення одного семестру дослідник провів одну і ту ж контрольну роботу з геометрії в кожному класі. В якості елемента аналізу розглядається середні оцінки успішності (округлені до найближчого цілого) для класу. Результати представлені в таблиці.



***Завдання 6.***

В результаті вимірювання частоти пульсу людини у нормальних умовах та при підвищених прискореннях вільного падіння одержані такі дані:

|  |  |  |  |
| --- | --- | --- | --- |
| Прискорення | Норма 1 g | 1,5 g | 2 g |
| Частоти пульсу | 76, 73, 68, 83 | 71, 62, 55, 68 | 46, 59, 54 |

Чи свідчать ці дані про наявність залежності між частотою пульсу та прискоренням на рівні значущості *а* = 0,05?

Тестові завдання для перевірки знань

1. Однофакторний ранговий аналіз виявляє:

а)вплив фактору на відгук при відомому законі розподілу;

б)вплив фактору на відгук з невідомим законом розподілу;

в)залежність між фактором і відгуком;

г)вплив побічного фактору.

2. Для однофакторного рангового аналізу використовується статистика:

а) Пуассона;

б) Фішера;

в) Краскала-Уоллеса;

г) Фридмана.

3. Ранг елемента вибірки характеризує:

а)величину;

б)місце у вихідній вибірці;

в)місце в упорядкованій вибірці;

г)порядковий номер в упорядкованій вибірці.

4. Однофакторний дисперсійний аналіз застосовується для виявлення:

а) незалежності вибірки;

б) впливу фактору на відгук при відомому законі розподілу;

в) залежності між фактором і відгуком;

г) дисперсії вибірки.

5. Для однофакторного дисперсійного аналізу використовують:

а) критерій хі-квадрат;

б)критерій Фішера;

в)критерій серій;

г)критерій Фрідмана.

6. Двофакторный аналіз застосовують для виявлення:

а) впливу основного фактору на відгук;

б) наявність фактору; що заважає;

в)вплив фактору; що заважає;

г) виключення фактору; що заважає.

7. Для двофакторного рангового аналізу застосовують статистику:

а) Колмогорова;

б) Фішера;

в) Смирнова;

г) Фрідмана.

8. Дві вибірки однорідні; якщо:

а) мають однакову функцію розподілу;

б) належать одній генеральній сукупності;

в) мають однакові набори елементів;

г) мають однаковий об'єм.

9. Який метод багатофакторної статистики застосовується для дослідження впливу якісних факторів і аналізу значущості цього впливу?

а) кореляційний аналіз;

б) регресійний аналіз;

в) метод найменших квадратів;

г) дисперсійний аналіз;

д) метод найбільшої правдоподібності.

10. Який метод багатофакторної статистики дозволяє оцінювати наявність стохастичного зв’язку між одновимірними випадковими величинами; не висловлюючи при цьому ніяких припущення відносно природи і форми цього зв’язку?

а) кореляційний аналіз;

б) регресійний аналіз;

в) дисперсійний аналіз;

г) метод найменших квадратів;

д) метод найбільшої правдоподібності.

11. У якому випадку дисперсійний аналіз вважається однофакторним?

а) коли розглядається лише одна випадкова величина; тобто фактор; який і є ознакою;

б) коли досліджується стохастичний зв’язок між компонентами двовимірної випадкової величини;

в) коли досліджується стохастичний зв’язок між компонентами багатовимірної випадкової величини для якої лише один фактор є незалежним;

г) коли лише один компонент багатовимірної випадкової величини є якісним фактором; а всі інші – кількісними;

д) нема правильної відповіді.

1. При застосуванні однофакторного дисперсійного аналізу результати вимірювань надаються у вигляді таблиці; рядки (або стовпці) якої утворюють групи; що відповідають різним значенням фактору; вплив якого досліджується. Продовжить речення; щоб утворилося правильне твердження: «При застосуванні дисперсійного аналізу перевірці підлягає статистична гіпотеза; яка стверджує; що…»:

а) різниця між групами є не більш суттєвою; ніж випадкові розбіжності між значеннями в межах кожної групи;

б) різниця між групами абсолютно несуттєва;

в) різниця між групами є більш суттєвої; ніж випадкові розбіжності між значеннями в межах кожної групи;

г) різниця між групами є не менш суттєвої; ніж випадкові розбіжності між значеннями в межах кожної групи;

д) нема правильної відповіді.

13. Прожить речення; щоб утворилося правильне твердження. «При застосуванні дисперсійного аналізу для перевірки нульової гіпотези застосовується …». Оберіть правильну відповідь.:

а) критерій Стьюдента;

б) критерій Фішера – Снедекора;

в) критерій Колмогорова - Смірнова;

г) критерій Пірсона;

д) нема правильної відповіді.

14. Продовжить речення; щоб утворилося правильне твердження. «Емпіричне значення критерію Фішера – Снедекора; який застосовується для перевірки нульової гіпотези в однофакторному дисперсійному аналізі; визначається як…». Оберіть правильну відповідь.:

а) відношення загальної суми квадратів відхилення значень випадкової величини від її середнього значення до суми квадратів відхилень; пов’язаних з впливом досліджуваного фактору;

б) відношення питомої суми квадратів відхилення значень випадкової величини від її середнього значення до питомої суми квадратів відхилень; пов’язаних з впливом досліджуваного фактору;

в) відношення суми квадратів відхилення; що пов’язана з регресією; до суми квадратів відхилень; що пов’язані з випадковим розпорошенням;

г) відношення питомої суми квадратів відхилень; що пов’язані з регресією; до питомої суми квадратів відхилень; що є випадковими помилками;

д) нема правильної відповіді.

15. При проведенні однофакторного аналізу вибіркової сукупності; що складалась з 96 об’єктів; була розділена на 8 груп відповідно до рівня деякого якісного фактору; вплив якого досліджувався; Визначить; чому дорівнює кількість ступенів вільності для суми квадратів відхилень; що є випадковими помилками.

[ЗМІСТОВИЙ МОДУЛЬ **5.**КОРЕЛЯЦІЙНИЙ І РЕГРЕСІЙНИЙ АНАЛІЗ](#bookmark28)

## Тема 9. Основи теорії кореляції

Мета: Ознайомитись з функціональним і статистичним взаємозв’язками, оволодіти методами кореляційного аналізу, набути навики застосування кореляційного аналізу для вирішення практичних задач та інтерпретації результатів кореляційного аналізу.

🖉Основні терміни і поняття

*Кореляція, кореляційна залежність,кореляційний аналіз, коефіцієнт рангової кореляції Спірмена, коефіцієнт спряженості Крамера, коефіцієнт (показник подібності) Жаккара, показник подібності Рассела, коефіцієнт асоціації Юла, Хеммінгова відстань, коефіцієнт Гауера, множинна кореляція*.

📚Основні теоретичні положення

* 1. **Функціональна, статистична й кореляційна залежності**

У природничих науках здебільшого мають справу з *функціональними залежностями,* у яких кожному можливому значенню аргументу *X* (незалежної змінної) відповідає одне певне значення функції *Y* (залежної змінної) (наприклад, у математиці, при вивченні фізичних законів). Проте набагато частіше в навколишньому світі існує залежність, коли кожному фіксованому значенню однієї змінної відповідає не одне, а безліч значень іншої змінної. Це пояснюється тим, що залежна змінна піддана впливу ряду неконтрольованих або неврахованих факторів, а також численних неконтрольованих випадкових факторів. У цій ситуації залежна змінна *Y* є випадковою величиною. Змінна ж *X* може бути як детермінованою (тобто такою, що приймає цілком певні значення), так і випадковою величиною. Така залежність одержала назву *статистичної* (або *стохастичної, імовірнісної).*

Припустимо, що існує стохастична залежність випадкової змінної *Y* від *X.* При *X = х* змінна *Y* у силу її стохастичної залежності від *X* може прийняти будь-яке значення з деякої нескінченності. Середнє цієї нескінченності називають *груповим генеральним середнім* змінної *Y* при *X = х* або умовним математичним очікуванням випадкової величини *Y,* обчисленим за умови, що *Х= х.* Ця величина позначається *M(Y***|***X=х).* Зокрема, якщо стохастична залежність *Y* від *X* виявляється в тому, що при зміні *х* змінюється умовне математичне очікування *M(Y***|***X=х)*, тоді говорять, що має місце *кореляційна залежність* величини *Y* від *X.* Якщо ж *M(Y***|***X=х)* залишаються незмінними, то говорять, що кореляційна залежність величини *Y* від *X* відсутня.

Якщо існує стохастична залежність випадкової змінної *X* від *Y*, то, аналогічно попередньому, можна ввести поняття умовного математичного очікування *M(Y***|***X=х)* і кореляційної залежності величини *X* від *Y.*

Таким чином, кореляційна залежність може бути представлена в такому вигляді:

;(9.1)

,(9.2)

де , .

Рівняння (9.1) і (9.2) називаються *модельними рівняннями регресії* відповідно *Y* по *X (Y* на *X)* і *X* по *Y (X* на В), функції й *– модельними функціями регресії,* а їхні графіки – *модельними лініями регресії.*

Слід зазначити, що введені поняття стохастичної і кореляційної залежності належать до генеральної сукупності. Як оцінки умовних математичних очікувань приймають умовні середні, які визначають за даними спостережень (за вибіркою). На практиці дослідник, як правило, розташовує лише вибіркою пар значень (*xi;yi*) обмеженого обсягу.

*Умовним середнім *називають середнє арифметичне спостережуваних значень *Y(X)*, що відповідають *X = х (Y = у).* Оскільки умовні математичні очікування *M(Y* | *X* = *х)* і *М(Х* | *Y* = *у)* є відповідно функціями від *х* та**,то їхні оцінки, тобто умовні середні **й  також є функціями відповідно від *х* та**,тобто

(9.3)

(9.4)

де *b0,b1, ..., bk*і *c0, c1, ..., ck* – параметри.

Рівняння (16.3) і (16.4) називаються *вибірковими рівняннями регресії* відповідно *Y* по *X* і *X* по *Y,* функції й *– вибірковими функціями регресії,* а їхні графіки – *вибірковими лініями регресії.*

Зв'язок або кореляція двох змінних називається *парною.* Якщо в рівняннях (9.3) і (9.4) зі збільшенням *х* та**,змінні **й усередньому зростають (мають тенденцію до зменшення), то така парна кореляція буде *позитивною* (*негативною*)*.* Нульова кореляція спостерігається при відсутності зв'язку між *X* і *Y.*

Діаграма, на якій зображується сукупність значень двох ознак, називається *кореляційним полем.* Кожна точка цієї діаграми має координати *хi, уi*, щовідповідають розмірам ознак в *i-му* спостереженні. Три варіанти розподілу точок на кореляційному полі показані на рисунку 9.1.

На першому з них основна маса точок укладається в еліпс, більша вісь якого утворить позитивний кут з віссю *ОХ* (позитивна кореляція). Другий варіант відповідає негативній кореляції. Рівномірний розподіл точок у просторі *XY* свідчить про відсутність кореляційної залежності (рисунок 9.1, в).

Рисунок 9.1 – Розподіл точок на кореляційному полі

Статистичні зв'язки між змінними можна вивчати методами кореляційного й регресійного аналізу. Основним завданням *регресійного аналізу* є встановлення форми залежності за достовірними даними, визначення функції регресії (процес *вирівнювання)* і вивчення залежності між змінними. Основним завданням *кореляційного аналізу* є виявлення зв'язку між випадковими змінними й оцінка її тісноти.

9.2 Поняття про кореляцію

Одним із найважливіших завдань будь-якого дослідження є встановлення зв’язку між величинами або факторами, зміна яких визначає сутність процесу, що вивчається. Щоб пізнати досліджуване явище, треба вивчити не тільки його зв’язки з навколишніми явищами (факторами), але й також взаємозв’язки всіх його сторін, тобто треба встановити закономірності змін взаємопов’язаних явищ і показників, що їх характеризують. У дослідників часто виникає потреба аналізувати залежність між двома або декількома змінними величинами (ознаками). Якщо дві деякі характеристики, отримані для одного і того ж «об’єкта», мають тенденцію змінюватися сумісно так, що створюється можливість передбачити одну з них за значенням іншої, то кажуть, що ці характеристики корелюють одна з одною. Відповідно в статистиці кореляція виражає ступінь взаємозв’язку між такими характеристиками. Кількісно ця ступінь взаємозв’язку виражається за допомогою коефіцієнта кореляції.

*Кореляцією (кореляційним зв’язком)* між випадковими величинами (ознаками) називають наявність статистичного або ймовірнісного зв’язку між ними. При цьому закономірна зміна певних ознак призводить до закономірної зміни середніх значень інших, пов’язаних з ними ознак. Кореляційний зв'язок – це узгоджені зміни двох ознак або більшої кількості ознак.

Кореляційний зв'язок відображає той факт, що мінливість однієї ознаки знаходиться в деякій відповідності до мінливості іншої. З іншого боку, кореляційний зв'язок може говорити не про залежність ознак між собою, а про залежність цих ознак від іншої (інших). Кореляційні зв'язки не може розглядатися як свідчення причинно-наслідкового зв'язку, вони свідчать лише про те, що зміни однієї ознаки, як правило, супроводжують певними змінами іншої, але чи знаходиться причина змін в одній з ознак або вона виявляється за межами досліджуваної пари ознак, нам невідомо.

*Кореляційна залежність* – це зміни, які вносять значення однієї ознаки в ймовірність появи різних значень іншої ознаки.

Кореляційні зв'язки розрізняються за *формою, напрямком і ступенем (силою).* За формою кореляційний зв'язок може бути *лінійної і криволінійної*. За формою кореляційний зв'язок може бути *лінійним і криволінійним*.

*Лінійний зв'язок* – якщо зі збільшенням або зменшенням однієї змінної, друга змінна в середньому або також зростає, або зменшується (рисунок9.2 а, б). Наприклад, прямолінійною можна назвати зв'язок між кількістю тренувань на тренажері і кількістю правильно вирішуваних завдань у контрольній сесії.



а) позитивний зв'язок б) негативний зв'язок

Рисунок 9.2– Лінійний зв’язок

Криволінійною може бути, наприклад, зв'язок між рівнем мотивації і ефективністю виконання завдання. При підвищенні мотивації ефективність виконання завдання спочатку зростає, потім досягається оптимальний рівень мотивації, якому відповідає максимальна ефективність виконання завдання; подальшого підвищення мотивації супроводжує вже зниження ефективності (рисунок 9.3).

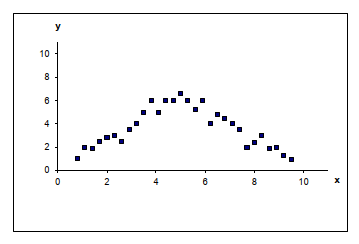


Рисунок 9.3– Криволінійний зв'язок

За напрямком кореляційний зв'язок може бути позитивним («прямий») і негативним («обернений»).

При *позитивній (прямій)* кореляції більш високим значенням однієї ознаки відповідають більш високі значення іншої, а більш низьким значенням однієї ознаки - низькі значення іншої (рисунок9.2 а). При *негативній (оберненій)* кореляції високі значення однієї ознаки відповідають більш низьким значенням іншої (рисунок 9.2 б).

*Ступінь кореляційного зв'язку* визначається за величиною коефіцієнта кореляції, що позначають часто як *r*. *Величина коефіцієнта кореляції* знаходиться в діапазоні від -1 до +1.

*Сила зв'язку* не залежить від його спрямованості і визначається за абсолютним значенням коефіцієнта кореляції. Якщо коефіцієнт кореляції за модулем виявляється близьким до 1, то це відповідає високому рівню зв'язку між змінними.

*Класифікація сили кореляції.* Використовуються кілка систем класифікації сили кореляції. Загальна класифікація засвідчує, що кореляція:

* сильна, або тісна при |𝑟| ≥ 0,7;
* середня при 0,5 ≤ |𝑟| < 0,7; ⎫
* помірна при 0,3 ≤ |𝑟| < 05; ⎫
* слабка при 0,2 ≤ |𝑟| < 0,3; ⎫
* дуже слабка при |𝑟| < 0,2.

Ця класифікація орієнтована на величину коефіцієнта кореляції і жодною мірою не реагує на рівень його значущості.

Класифікація кореляційних зв’язків:

* *висока значуща* кореляція – при r, що відповідає рівню статистичної значущості p≤ 0,01;
* *значуща* кореляція – при r, що відповідає рівню статистичної значущості p≤ 0,05;
* *тенденція достовірного* зв'язку – при r, що відповідає рівню статистичної значущості p≤ 0,10;
* *незначна* кореляція при r, що не досягає рівня статистичної значущості.

Змінні можуть бути виміряні в різних шкалах, саме це визначає вибір відповідного коефіцієнта кореляції. У таблиці представлені співвідношення між шкалами.

*Порядкова шкала* вимірювань дозволяє надати ранги значенням змінних. Вимірювання в порядковій шкалою містять інформацію тільки про порядок проходження величин, але не дозволяють сказати "наскільки одна величина більша за іншу", або "наскільки вона менше інший".

*Рангова шкала*. У межах цієї шкали об’єкти розташовуються в порядку спадання чи зростання у них певної якості. При цьому кожній градації якості приписується свій порядковий номер (ранг). Фактично, об’єкти лише впорядковуються. Особливість шкали – однакові різниці між сусідніми рангами не означають однакової різниці між ступенями прояву виміряної якості.

|  |  |  |
| --- | --- | --- |
| Тип шкали | | Міра зв’язку |
| Змінна А | Змінна В |
| Інтервальна або  відношень | Інтервальна або  відношень | *r*xy – коефіцієнт лінійної кореляції Пірсона |
| Порядкова | Порядкова або  інтервальна | *Rs* – коефіцієнт рангової кореляції Спірмена |
| Порядкова | Порядкова | – коефіцієнт кореляції Кендалла |
| Дихотомічна | Дихотомічна | ϕ – коефіцієнт кореляції Пірсона |
| Дихотомічна | Порядкова | *Rrb* – рангово-бісеріальний коефіцієнткореляції |
| Дихотомічна | Інтервальна або  відношень | *Rбіс* – бісеріальний коефіцієнт кореляції |
| Інтервальна | Порядкова | не розроблений |

При інтерпретації отриманого зв’язку велике значення має знак коефіцієнта кореляції. У випадку, коли коефіцієнт лінійної кореляції має знак «+», зв'язок між ознаками, що корелюються, має наступну особливість: більшій величині однієї ознаки (змінної) відповідає більша вилична іншої ознаки (іншої змінної). Іншими словами, якщо один показник (змінна) збільшується, то відповідно збільшується й інший показник (змінна).

Якщо ж коефіцієнт лінійної кореляції має знак «-», значить, більшій величині однієї ознаки відповідає меншій іншої. Тобто при наявності знаку «-» збільшення однієї змінної (властивості, значення) відповідає зменшення іншої змінної. В даному випадку вибір змінної, у якої спостерігається тенденція до зростання, має довільний характер. Це може бути як змінна *X*, так і змінна *Y*. У випадку, коли експериментатор вважає, що збільшується змінна *Х*, змінна *Y* відповідно буде зменшуватись, і навпаки. Ця інформація потрібна для правильної інтерпретації отриманої кореляційної залежності.

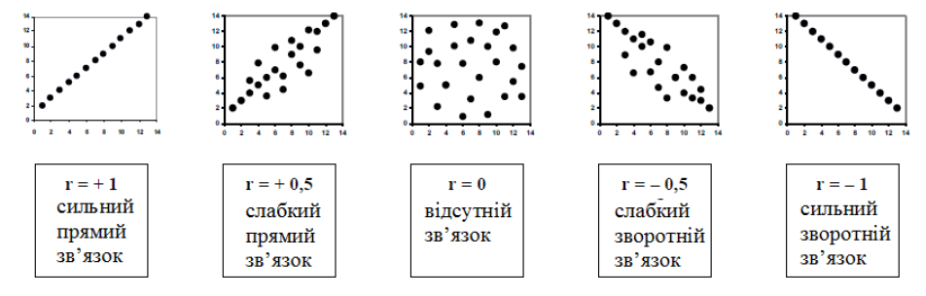


Рисунок 9.4– Інтерпретації кореляційної залежності

* 1. **Кореляційний аналіз**

*Кореляційним аналізом* називають сукупність методів виявлення кореляційного зв’язку. Тому його можна застосовувати для формалізованого подання моделей зв’язків між окремими компонентами системи або між окремими процесами, що відбуваються в ній. Наявність кореляційного зв’язку не означає існування причинно-наслідкового зв’язку між досліджуваними ознаками. Вона може бути зумовлена тим, що обидві ознаки мають причинно-наслідковий зв’язок з певним іншим фактором. Наприклад, існує кореляція між цінами на нафту й на золото. Проте вона пояснюється тим, що обидві ціни виражаються у доларах США й залежать від динаміки його індексу. Кореляція також може бути випадковою.

Сучасну класифікацію мір подібності запропонували австрійський та американський біостатистик та антрополог Роберт Сокал та британський таксономіст Пітер Сніс у 1963 р. Згідно з нею виокремлюють такі типи мір подібності:

– міри асоціації, що відбивають різні співвідношення кількості ознак, що збігаються до загальної кількості ознак, а також близькі до них коефіцієнти спряженості (квантифіковані коефіцієнти зв’язку);

– вибіркові коефіцієнти зв’язку типу кореляції (нормовані косинусні міри);

– показники відстані у метричному просторі.

Перевірку зв’язку можна здійснювати лише для пов’язаних вибірок. Це означає, що між елементами обох досліджуваних вибірок існує взаємно однозначна відповідність, а кількість елементів у вибірках є однаковою.

Методику кількісного оцінювання кореляції між ознаками вперше було запропоновано британським географом, антропологом та психологом Френсисом Гальтоном в 1888 р. Універсальною характеристикою ступеня тісноти зв’язку між кількісними ознаками є коефіцієнт детермінації.

Величина коефіцієнта детермінації може змінюватися в межах від нуля до одиниці й відображає частку загальної дисперсії досліджуваної ознаки, яка зумовлена зміною функції регресії *f(X).* При цьому нульове значення коефіцієнта детермінації відповідає відсутності будь-якого зв’язку, а його рівність одиниці – наявності строго функціонального зв’язку. Оскільки цей коефіцієнт є універсальним показником зв’язку, він має відбивати й такі зв’язки, що є немонотонними функціями. Тому питання напряму зв’язку у цьому випадку не має сенсу.

Слід зазначити, що для обмеженого набору даних часто можна побудувати декілька різних адекватних регресійних моделей. Групування даних також можна здійснювати різними способами. Тому існує певна невизначеність коефіцієнтів детермінації: при застосуванні різних регресійних моделей або різних способів групування ми будемо отримувати дещо різні значення коефіцієнта детермінації.

Інші поширені характеристики ступеня тісноти зв’язку між ознаками можна розглядати як окремі випадки коефіцієнта детермінації, отримані для конкретних математичних моделей зв’язку.

Розрізняють парні та частинні кореляційні характеристики. Парні характеристики розраховують за результатами вимірювань тільки досліджуваної пари ознак. Тому вони не враховують опосередкованого або спільного впливу інших ознак. Частинні характеристики є очищеними від впливу інших факторів, але для їх розрахунку необхідно мати вихідну інформацію не тільки про досліджувані ознаки, а й про всі інші, вплив яких необхідно усунути.

Для кількісних ознак найчастіше застосовують коефіцієнти кореляції Пірсона і Фехнера. *Коефіцієнт кореляції Пірсона (коефіцієнт кореляційного відношення Пірсона)* вимірює ступінь лінійного кореляційного зв’язку між кількісними скалярними ознаками. Він був запропонований К. Пірсоном у 1896 р. Часто, посилаючись на згадування К. Пірсона про ідеї математичного подання зв’язку, висловлені в 1846 р. відомим французьким фізиком та кристалографом Огюстом Браве, цей показник називають коефіцієнтом Бравайса – Пірсона (Бравайс – це викривлена транскрипція від французького Bravais, що закріпилася в літературі з кореляційного аналізу).

Застосування коефіцієнта Пірсона як міри зв’язку є обґрунтованим лише за умови, що спільний розподіл пари ознак є нормальним. Тому перед його розрахунком слід перевірити виконання цієї гіпотези. Якщо вона справедлива, то квадрат коефіцієнта кореляції Пірсона дорівнює коефіцієнту детермінації.

Значення коефіцієнта кореляції може змінюватися від −1 до +1. Значення −1 та +1 відповідають чіткій лінійній функціональній залежності, яка в першому випадку є спадною, а у другому – зростаючою. Для функціональної залежності y const = коефіцієнт кореляції, як видно з наведеної формули, є невизначеним, оскільки в цьому випадку знаменник дорівнює нулю. Що ближчим є значення коефіцієнта кореляції до −1 або +1, то більш обґрунтованим є припущення про наявність лінійного зв’язку. Наближення його значення до нуля свідчить про відсутність лінійного зв’язку, але не є доказом відсутності статистичного зв’язку взагалі.

На рисунку 9.5 показано дві серії точок, координати яких відповідають двом парам спряжених вибірок.

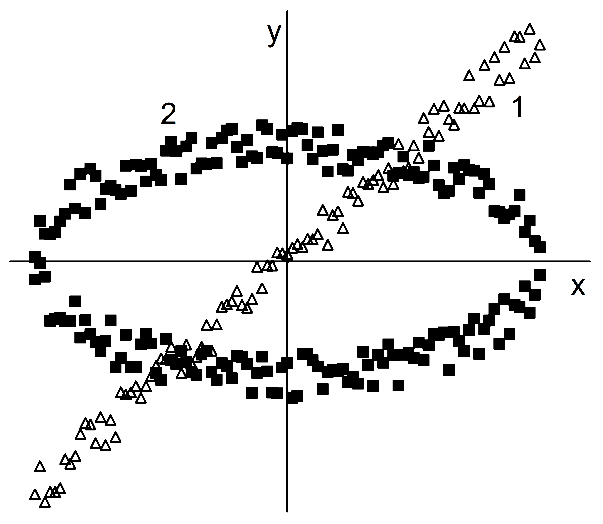


Рисунок 9.5– Графічне зображення двох наборів тестових даних

Для обох пар вибірок є очевидним існування статистичного зв’язку між параметрами х та у. Але коефіцієнти кореляції для них дорівнюють, відповідно, *r1*= 0,995 і *r2* = 0,006 . Близькість коефіцієнта кореляції до нуля для другої пари вибірок пов’язана не з відсутністю зв’язку, а з його нелінійністю. Для порівняння, коефіцієнти детермінації для тих самих пар вибірок дорівнюють 0,98 та 1,00.

Показаний приклад свідчить, що в багатьох випадках для попереднього аналізу припущення про наявність і тип зв’язку між певними ознаками доцільно нанести наявні дані на графік.

Як видно, близькість коефіцієнта кореляції Пірсона до нуля в загальному випадку не є доказом незалежності ознак. Але можна довести, що у випадку, коли сумісний розподіл випадкових величин *(x, y)* є нормальним, рівність *r = 0* свідчить про статистичну незалежність *x* і *y*.

Коефіцієнт кореляції Пірсона часто розглядають як універсальну міру кореляційного зв’язку. У багатьох пакетах загального призначення, зокрема в електронних таблицях MS Excel, не передбачено інших засобів його вимірювання. Але, як випливає з наведених вище даних, насправді сфера його обґрунтованого застосування є досить вузькою, оскільки лінійність залежності й нормальний розподіл даних навколо неї є скоріше винятком, ніж правилом.

При дослідженні багатовимірних сукупностей випадкових величин із коефіцієнтів кореляції, обчислених попарно між ними, можна побудувати квадратну симетричну кореляційну матрицю з одиницями на головній діагоналі. Вона є основним елементом при побудові багатьох алгоритмів багатовимірної статистики, наприклад у факторному аналізі.

Коефіцієнт кореляції Пірсона можна застосовувати для перевірки гіпотези про значущість зв’язку.

У випадку, коли між двома наборами ознак існує нелінійний зв’язок, для оцінювання ступеня його тісноти часто використовують кореляційне відношення, яке було запропоновано К. Пірсоном. Це можливо, якщо щільність розміщення емпіричних точок на координатній площині дає можливість їх групування за однією із змінних і підрахунку групових середніх значень другої змінної для кожного інтервалу

**9.4 Кореляційний аналіз номінальних та змішаних ознак**

*Коефіцієнт рангової кореляції Спірмена (показник кореляції рангів Спірмена, коефіцієнт кореляції рангів)* запропонований британським психологом Чарльзом Едвардом Спірменом у 1904 р. Його використовують, якщо досліджується зв’язок між рядами даних, виміряними за порядковою шкалою. Його можна застосовувати також і для кількісних даних, але, як правило, це буває недоцільним. У найпростішому випадку досліджувані об’єкти класифікують за двома ознаками.

Коефіцієнт рангової кореляції Спірмена можна застосовувати як показник некорельованості вибірок.

Інший підхід використовує як міру подібності двох вибірок мінімальну кількість перестановок сусідніх об’єктів, потрібну для переведення послідовності рангів однієї вибірки до послідовності рангів іншої. Можна показати, що вона дорівнює кількості інверсій в однієї з цих послідовностей у випадку, коли інша послідовність впорядкована за зростанням.

Як і в попередньому випадку, кількість інверсій залежить від обсягу вибірки і є незручною для застосування як показника кореляції. Для цього використовують коефіцієнт рангової кореляції Кендалла (коефіцієнт кореляції рангів, ранговий коефіцієнт кореляції). Він був запропонований британським статистиком Маурисом Кендаллом у 1938 р.

Коефіцієнт рангової кореляції Кендалла призначений для визначення сили кореляційного зв’язку між двома рядами даних за тих самих умов, що і коефіцієнт рангової кореляції Спірмена. Як і для коефіцієнта Спірмена, його значення можуть змінюватися в межах від −1 до +1, при цьому −1 відповідає повній протилежності послідовностей рангів, а +1 – їх повному збігу. Слід зазначити, що обчислення коефіцієнта Кендалла є більш трудомістким, але з іншого боку, він має ряд переваг порівняно із коефіцієнтом Спірмена. Основними з них є такі:

– кращий рівень вивченості його статистичних властивостей, зокрема його вибіркового розподілу;

– можливість його застосування для визначення частинної кореляції;

– більша зручність перерахунку при додаванні нових даних.

Типовою ситуацією, коли необхідна перевірка зв’язку між номінальними ознаками, є обробка результатів соціологічних досліджень, що можуть містити такі комбінації ознак, як освіта, стать, професія, підтримка певної політичної партії, регіон проживання тощо.

При дослідженні зв’язків між категоризованими ознаками вихідні дані подають у вигляді таблиці спряженості (табл. 9.1). До категоризованих зараховують номінальні ознаки, а також порядкові ознаки, для яких є відомим скінченний набір можливих градацій.

Величини *fij* показують, скільки разів зустрічалася комбінація ознак, за якої рівень першої має значення i, а рівень другої – *j*; *mj*є сумами стовпців, а *ni*– сумами рядків. За даними таблиці 9.1 можна оцінити значення ймовірностей.

Нульову гіпотезу про відсутність зв’язку відхиляють, якщо різницю між ними й частотами, що спостерігаються, не можна пояснити випадковими чинниками. Як критерій можна використовувати величину .На практиці частіше використовують *φ-коефіцієнт Пірсона,* або середньоквадратичну спряженість, яка може змінюватися від нуля до *min {r-1, c-1 }.*

Таблиця 9.1 Таблиця спряженості категоризованих ознак

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Рівні ознаки 1 | Рівні ознаки 2 | | | | Разом |
| 1 | 2 | … | r |
| 1 |  |  | … |  |  |
| 2 |  |  | … |  |  |
| ... | … | … | … | … | … |
| c |  |  | … |  |  |
| Разом |  |  | … |  |  |

Існує велика кількість показників ступеня тісноти статистичного зв’язку, призначених для категоризованих змінних, які не є універсальними, а відображають окремі властивості такого зв’язку.

*Коефіцієнт спряженості Крамера* був запропонований К.Х. Крамером у 1946 р. Він змінюється в межах від нуля до одиниці. При цьому значення С = 0 свідчить про статистичну незалежність аналізованих ознак, а значення С = 1 – про можливість однозначного відтворення значень однієї ознаки за відомими значеннями другої.

Існує велика кількість коефіцієнтів, що характеризують кореляцію між ознаками у випадку, коли кожна з двох ознак може мати лише два рівні, які найчастіше відповідають наявності та відсутності ознаки. У цьому випадку таблиця спряженості має розмір 2×2.

*Коефіцієнт (показник подібності) Жаккара,* уведений в 1901 р. французьким геоботаніком Полем Жаккаром. Значення цього коефіцієнта можуть змінюватися в межах від нуля до одиниці.

*Простий коефіцієнт зустрічальності (показник подібності Сокала й Міченера)* запропонований Р. Сокалом та американським ентомологом Чарльзом Дунканом Міченером у 1958 р. Як і в попередньому випадку, значення коефіцієнта можуть змінюватися в межах від нуля до одиниці.

*Показник подібності Рассела і Рао* запропонували в 1940 р. американський епідеміолог Поль Ф. Рассел та індійський і британський ентомолог Т. Рамакришна Рао. Його значення також можуть змінюватися в межах від нуля до одиниці.

*Коефіцієнт спряженості Бравайса – Пірсона* (показник подібності Чупрова) був уведений О.О. Чупровим у 1923 р. Значення цього коефіцієнта може змінюватися в межах від −1 до +1. Від’ємні значення коефіцієнта спряженості означають, що із збільшенням імовірності прояву одної ознаки, зменшується імовірність прояву іншою.

Легко показати, що цей показник є окремим випадком ϕ-коефіцієнта Пірсона для таблиць 2×2.

*Коефіцієнт асоціації Юла* був уведений відомим британським статистиком Джорджем Удні Юлом у 1900 р.

*Коефіцієнт колігації Юла,* що також був запропонований Дж.У. Юлом в 1912 р. Він не має переваг порівняно з коефіцієнтом асоціації. Значення обох коефіцієнтів змінюються в межах від −1 до +1.

*Хеммінгова відстань (метрика Хеммінга)H = a + d* також може застосовуватися для визначення кореляції. Проте, як і коваріація, вона не є безрозмірною величиною і може набувати будь-яких невід’ємних значень (верхньою межею є загальна кількість спостережень *n*). Цей показник був уведений відомим американським математиком Ричардом Веслі Хеммінгом у 1950 р.

*Коефіцієнт Гауера* був запропонований британським статистиком Джоном Кліффордом Гауером у 1971 р. Його застосовують у тому випадку, коли досліджувані ознаки виміряні в різних шкалах.

Для дихотомічних ознак алгоритм підрахунку внеску ознаки і визначення вагових коефіцієнтів збігається з коефіцієнтом Жаккара. Для порядкових ознак алгоритм підрахунку внеску ознаки збігається з хеммінговою відстанню, узагальненою на порядкові змінні, а вагові коефіцієнти беруть рівними одиниці для всіх ознак.

*Бісеріальний коефіцієнт кореляції* запропоновано К. Пірсоном. Його призначено для дослідження кореляції в таблицях розміром 2×n, які є дихотоміями за певною номінальною ознакою і класифікаціями за номінальною або порядковою ознакою, що класифікується за q класами і може бути впорядкованою або невпорядкованою. Вихідний розподіл має бути двовимірним нормальним. Значення бісеріального коефіцієнта кореляції можуть змінюватися від −1 до +1.

*Бісеріальний коефіцієнт кореляції за таблицею Келлі – Вуда* запропонований американським психологом Луїсом Л. Терстоуном (Louis L. Thurstone) в 1928 р.

У випадку, коли одна із змінних дихотомізована, а інша – виміряна в кількісній шкалі, обчислюють *точково-бісеріальний коефіцієнт кореляції.*

При цьому передбачається, що дихотомічна змінна може набувати лише два значення: 1 (верхній рівень) та 0 (нижній рівень). З погляду теорії точково-бісеріальну кореляцію можна розглядати як окремий випадок коефіцієнта кореляції Пірсона.

* 1. **Коефіцієнт лінійної кореляції**

Для кількісної характеристики тісноти лінійного кореляційного зв'язку між двома величинами *X* і *Y* генеральної сукупності раніше було уведене поняття *коефіцієнта лінійної кореляції,* обумовленого співвідношенням

 (9.5)

де 

Відомо, що якщо величини *X* і *Y* незалежні, то ; якщо , то *X* і *Y* зв'язані лінійною функціональною залежністю, причому, дорівнює  у випадку зростаючої залежності й  у випадку спадної; . При цьому *,* якщо при зростанні однієї величини (наприклад, *X*)М(*Y*) збільшується, і негативний у протилежному випадку.

На практиці для оцінки тісноти лінійного кореляційного зв'язку між величинами *X* і *Y* зарезультатами вибіркових спостережень використовується вибірковий коефіцієнт лінійної кореляції, обумовлений формулою (9.5), тобто

 (9.6)

Формулу (9.6) звичайно застосовують у трохи модифікованому вигляді. Так, підставивши в неї розгорнуті вираження для *sx*, *sy*і *sxy*,легко одержати наступну формулу:

 (9.7)

Якщо дані не згруповані у вигляді кореляційної таблиці й представляють *п* пару чисел (*хi*; *уj*), то для обчислення коефіцієнта кореляції у формулі (9.7) варто взяти замінити на .

Коефіцієнт кореляції  є безрозмірною величиною (тому що розмірністю чисельника й знаменника є розмірності добутку *XY*); його величина не залежить від вибору одиниць виміру обох змінних; величина  приймає значення на відрізку [–1;1]. Близька до нуля величина коефіцієнта кореляції свідчить про відсутність лінійного зв'язку змінних, але не заперечує можливість існування іншої форми залежності між ними.

Хоча вибірковий коефіцієнт кореляції  представляє значиму оцінку для , однак більш надійну оцінку близькості  до за даними вибірки можна дати лише в тому випадку, коли розподіл величин *X* і *Y* досить близький до нормальної форми. Наприклад, для оцінки нормально розподіленої генеральної сукупності, у випадку більших вибірок (), можна скористатися формулою

 (9.8)

де  – корінь рівняння , що визначається з таблиці за заданою довірчою імовірністю **, а величина

(9.9)

називається точністю оцінки.

Наведемо схему знаходження точності *δ* і довірчих границь, що відповідають надійності *γ*:

  (9.10)

Приклад 3.Знайти довірчий інтервал для оцінки коефіцієнта кореляції з надійністю *γ*= 0,95, якщо =0,23, *n*= 320.

За схемою (9.10),



Звідси, 0,126 <*<*0,334.

Застосовуючи коефіцієнт кореляції як міру зв'язку, потрібно мати на увазі, що він отриманий на основі даних вибірки й, отже, підданий впливу випадковості.

Якщо обсяг вибірки невеликий, то знайти вибіркову помилку цієї величини досить складно, тому на практиці зазвичай замість визначення помилки коефіцієнта кореляції перевіряють гіпотезу про його значимість (істотність), тобто чи суттєво відрізняється від нуля чи цю відмінність можна приписати впливу випадковості, пов'язаної з вибіркою. Інакше кажучи, виникає необхідність при заданому рівні значимості *а* перевірити нульову гіпотезу  при конкуруючій гіпотезі *.*

Якщо нульова гіпотеза відкидається, то це означає, що істотно відрізняється від нуля, а *Х* та *Y* корельовані, тобто пов'язані лінійною залежністю. Якщо нульова гіпотеза буде прийнята, то не значимо, а *Х* та *Y*некорельовані, тобто не пов'язані лінійною залежністю.

Як критерій перевірки нульової гіпотези приймемо випадкову величину

, (9.11)

яка має розподіл Стьюдента з *k = п–2* ступенями волі. Число ступенів волі менше числа спостережень на 2, оскільки у формулу вибіркового коефіцієнта кореляції входять середні вибіркові значення *X* і *Y,* для розрахунку яких використаються дві лінійні формули їхньої залежності від спостережень випадкових величин.

Оскільки альтернативна гіпотеза має вигляд ,то критична область є двосторонньою. Перевірку нульової гіпотези будемо здійснювати за наступною схемою:



  (9.12)

Приклад 4. За вибіркою обсягу *п* = 62, взятою з нормальної двовимірної генеральної сукупності (*X; Y*), знайдений вибірковий коефіцієнт кореляції = 0,3. Потрібно при рівні значимості *α* = 0,01 перевірити нульову гіпотезу  при альтернативній гіпотезі .

Використовуючи схему (9.12), одержимо



Таким чином, немає підстав відкинути нульову гіпотезу. Інакше кажучи, незначно відрізняється від нуля, тобто *X* і *Y* некорельовані, тобто не пов'язані лінійною залежністю.

Питання для самоконтролю

1. Дайте визначення кореляційної залежності між ознаками *X* і .
2. Встановіть зв'язок з  і *.*
3. Наведіть виведення *, , *і через умовні варіанти.
4. Виразіть коваріацію через умовні варіанти.
5. Для чого використовується вибірковий коефіцієнт лінійної кореляції? Наведіть формулу.
6. Наведіть схему знаходження довірчих границь для оцінки .
7. Наведіть схему перевірки нульової гіпотези про рівність нулю .
8. Запишіть модель лінійної регресійної залежності ознак *X, Y* і *Z.*
9. Запишіть формули для знаходження параметрів *а*,**і *з* лінійного рівняння регресії.
10. Для чого використовуються вибіркові сукупний і частковий коефіцієнти кореляції? Наведіть формули.

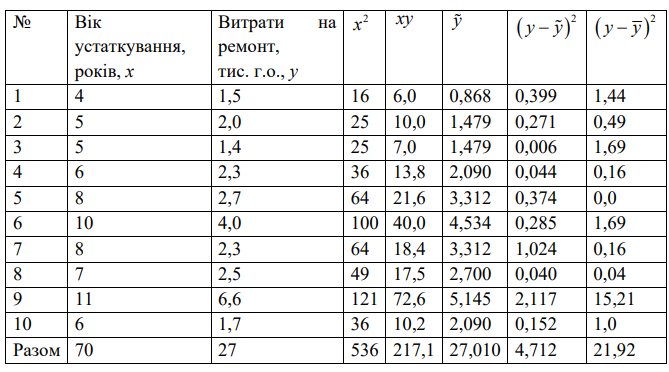
Практичне завдання

***Завдання 1.*** Встановити, чи існує кореляційний зв’язок між масою тіла і артеріальним тиском? Оцініть характер та глибину (силу) кореляційного зв’язку, вірогідність коефіцієнту кореляції. Дані досліджень наведено в таблиці:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Маса, *xi* | 120 | 80 | 110 | 100 | 90 |
| Артеріальний тиск, *yi* | 150 | 110 | 135 | 140 | 115 |

***Завдання 2.*** За допомогою КРА визначити наявність та характер статистичного зв’язку між ознаками «вік устаткування» та «витрати на ремонт». Вихідні дані та проміжні розрахунки наведено у таблиці.

*Вік устаткування та витрати на ремонт для групи підприємств*



***Завдання 3.*** За дослідними даними обчислити коефіцієнт кореляції та детермінації, зробити висновки.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *xk* | 2 | 4 | 6 | 8 |
| *yk* | 0.81 | 2.3 | 3.4 | 4.54 |

***Завдання 4*.** Результати тестування за 30-бальною шкалою для групи *X* та групи *Y* представлені у таблиці. Порівняти ефективність двох методів навчання студентів у двох групах для рівня статистичної значущості α = 5%.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 18 | 10 | 7 | 15 | 14 | 11 | 13 | 10 | 17 |  |  |
| Y | 15 | 20 | 10 | 8 | 16 | 10 | 19 | 7 | 15 | 14 | 29 |

***Завдання 5*.** Розглянуто дані про кількість хронічно хворих на астму та концентрацію чадного газу в кількох містах (див. табл.). Очевидно, що коли між цими ознаками існує залежність, то саме кількість хронічно хворих залежить від концентрації чадного газу, а не навпаки. Тобто концентрація чадного газу є факторною ознакою, а кількість хронічно хворих на астму – результативною.Встановити, чи існує кореляційний зв’язок між кількістю хронічно хворих на астму та концентрацію чадного газу.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| концентрації  чадного газу | 1,2 | 2,4 | 2,56 | 3,10 | 3,50 | 4,20 | 4,80 |
| кількість хронічно хворих на 1000 жителів | 20 | 35 | 42 | 48 | 51 | 59 | 63 |

## Тема 10. Елементи теорії регресії

Мета: Розглянути графічний метод аналізу статистичного взаємозв’язку, лінійну регресією;опанувати методи визначення графічного та математичного (розрахунок рівняння регресії таобчислення коефіцієнту регресії) проведення регресійного аналізуданих.

🖉Основні терміни і поняття

*Функціональна, статистична, стохастична й кореляційна залежності, кореляційне поле, лінійна парна регресія, метод найменших квадратів, вибіркова коваріація, вибірковий коефіцієнт кореляції, вибіркова дисперсія, вибірковий коефіцієнт лінійної регресії.*

📚Основні теоретичні положення

Вчення про кореляцію (від латинського correlatio – співвідношення, взаємозв'язок) і регресію (від латинського regressio – рух назад) широко використовується при аналізі зв'язків різних явищ. Так, наприклад, економіці проводяться дослідження: залежності обсягів виробництва від цілого ряду факторів (розміру основних фондів, їхнього віку й т.п.); залежності продуктивності праці на підприємствах від рівня механізації й електрифікації виробництва, стажу й кваліфікації робітників; залежності попиту й споживання населення від рівня доходу, цін на товари тощо. Регресійний аналіз широко застосовується при вивченні залежності врожайності певної сільськогосподарської культури від природних і економічних факторів, що впливають на неї. Широко застосовуються методи кореляційного й регресійного аналізу в психології, соціології, педагогіці та в інших галузях науки і практики.

Поняття кореляції й регресії з'явилося в середині XIX ст., завдяки роботам Ф. Гальтона й К. Пірсона.

**10.1. Лінійна парна регресія**

Нехай у результаті незалежних спостережень над досліджуваною системою кількісних ознак *(X;Y),* отримана *п* пара чисел *(x1;у1), (х2;у2),..., (хп;уп).* За даними спостережень знайдемо вибіркове рівняння прямої лінії регресії, для визначеності *Y* по *X*

*.* (10.1)

В (10.1) **замінене на **, бо різні значення *х* ознаки *X* і відповідні їм значення **ознаки *Y* спостерігалися один раз й тому групувати дані немає необхідності. Якщо позначити через наближене значення *уi*,обчислене з рівняння регресії (10.1), то величина  є відхиленням наближеного значення **від точного *уi*(рис. 10.1).

Найбільш часте оцінювання параметрів *α* і *β* прямої регресії здійснюють на основі *методу найменших квадратів* (МНК), розробка якого належить К. Гауссу й П. Лапласу. Цей метод одержав широку область додатка в економіко-статистичнихрозрахунках після створення теорії регресії.



Рисунок 10.1 – Метод найменших квадратів

Згідно з МНК параметри *α* і *β* прямої регресії вибирають так, щоб сума квадратів відхилень  була мінімальної, тобто з умови мінімізації функції:



Необхідною умовою існування мінімуму функції *S(α;β)* є рівність нулю часток похідних по невідомих параметрах *α* і *β*. Дорівнявши частки похідних  й  нулю, одержимо систему рівнянь для визначення *α* і *β*:

 (10.2)

Розв’язавши цю систему, знайдемо шукані параметри:



(10.3)

Якщо потрібно за результатами спостережень одержати лінійне рівняння регресії *Х* по *Y,* то в рівнянні регресії *y= α+βх* треба поміняти місцями змінні *х* та *.* При цьому одержимо рівняння *х = α'+β'у,* де *α'* і *β'* обчислюються за формулами:



(10.4)



Відзначимо, що регресійні прямі *у = α+βх* і *х* = *α'+β'у* різні. Перша пряма виходить у результаті розв’язання завдання мінімізації суми квадратів відхилень по вертикалі, а друга – при розв’язанні завдання зо мінімізації суми квадратів відхилень по горизонталі.

Таблиця 10.1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *i* |  |  |  |  |  |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| … | … | … | … | … | … |
| *п* |  |  |  |  |  |
|  |  |  |  |  |  |

На практиці для знаходження рівнянь регресії складається наступна таблиця 10.1.

В останньому рядку цієї таблиці суми й визначають коефіцієнти *α* і *β* або *α'* і *β'* у формулах (10.3) або (10.4) відповідно.

Приклад 1.За даними таблиці спостережень

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *xi* | 1 | 1,5 | 2 | 2,5 | 3 |
| *уi* | 2,1 | 2,2 | 2,7 | 2,8 | 2,85 |

скласти рівняння регресії*Y* по *X* і *X* по *Y.*

Складемо таблицю 10.2:

Таблиця 10.2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *i* |  |  |  |  |  |
| 1 | 1 | 2,1 | 1 | 4,41 | 2,1 |
| 2 | 1,5 | 2,2 | 2,25 | 4,84 | 3,3 |
| 3 | 2 | 2,7 | 4 | 7,29 | 5,4 |
| 4 | 2,5 | 2,8 | 6,25 | 7,84 | 7 |
| 5 | 3 | 2,85 | 9 | 8,1225 | 8,55 |
|  | 10 | 12,65 | 22,5 | 32,5025 | 26,35 |

За формулами (10.3) при *п =* 5 одержуємо:





Отже, рівняння регресії *Y* по *X* є



Аналогічно за формулами (10.4) знаходимо





Звідси рівняння *X* по *Y* є

*х* = *β'в +α'*=2,11*y*-3,33

Якщо число вимірювань велике, то з метою спрощення розрахунків експериментальні дані потрібно групувати, тобто поєднувати в таблицю 10.3, названу *кореляційною.*

Таблиця 10.3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Y  X |  |  | … |  |  |
|  |  |  | … |  |  |
|  |  |  | … |  |  |
| … | … | … | … | … | … |
|  |  |  | … |  |  |
|  |  |  | … |  |  |

Пояснимо, як заповнюється кореляційна таблиця. У першому стовпці (першому рядку) перераховуються у вибірці значення величини *.*

Якщо кількість різних значень величин *X і Y* велика або ці величини розподілені безкінечно, то робиться групування їхніх значень за інтервалами. У цьому випадку *xi*і *уj.* являють собою середини відповідних інтервалів.

На перетині рядків і стовпців указуються частоти *nij*, які дорівнюють числу появ у вибірці пари (*xi;yj*) Якщо пари значенні ознак (*xi;yj*) не спостерігалася, то у відповідному осередку таблиці ставиться риска.

В останньому рядку (останньому стовпці) вказуються числа , що дорівнюють кількості появ у вибірці значень *yj*(*xi*)незалежно від того, у парі з яким зі значень величин *X(Y)* воно з'явилося.

Кореляційна таблиця містить всю інформацію, отриману в результаті вибіркових спостережень величин *X* і *Y.* Звідси з урахуванням частот появ змінних *хi*і *уj,* одержимо:



Підставивши ці суми у формули (10.3) і (10.4), одержимо:



(10.5)





(10.6)



Як відомо, система рівнянь для визначення параметрів *α* і *β* рівняння прямої лінії регресії *Y* на *X* має вигляд (10.2). Передбачалося, що значення *X і* відповідні їм значення *Y* спостерігалися по одному разу. Тепер же допустимо, що отримано велику кількість даних, серед яких є повторювані, і вони згруповані у вигляді кореляційної таблиці. Перепишемо систему (10.2) так, щоб вона відбивала дані кореляційної таблиці. Скористаємося тотожностями:

 (10.7)

де *–* вибіркові середні; – вибіркове середнє квадрата; – вибіркове середнє добутку. Нагадаємо, що *вибіркова коваріація sxy*визначається рівністю

 (10.8)

*Вибіркові дисперсії* визначаються співвідношеннями:



(10.9)



За визначенням *вибірковий коефіцієнт кореляції*

 (10.10)

Підставивши праві частини (10.7) у систему (10.2), одержимо

 (10.11)

Розв’язуючи систему (10.11) за формулами Крамера, одержимо:



Підставивши коефіцієнти α і β у рівняння регресії * = α+βх*, одержимо



Таким чином, рівняння прямої регресії *Y* по *X* має вигляд

 (10.12)

Аналогічно знайдемо, що рівняння прямої регресії *X* по *Y* запишеться у вигляді

 (10.13)

З рівнянь (10.12) і (10.13) необхідно, що прямі регресії *Y* по *X* і *X* по *Y* проходять через точку . Ці прямі збігаються, коли .

Величини  й  називаються *вибірковими коефіцієнтами лінійної регресії* й позначаються

 (10.14)

Перемноживши праві й ліві частини рівнянь (10.14), після добування кореня одержимо , тобто коефіцієнт кореляції є середнім геометричним коефіцієнтів лінійної регресії. Знак у правій частині  збігається зі знаками  й .

Коефіцієнт регресії *Y* по *X* (*X* по *Y*) показує, на скільки одиниць у середньому змінюється змінна *Y(X)* при збільшенні змінної *X(Y)* на одну одиницю.

Якщо дані спостережень над ознаками *X* і *Y* задані у вигляді кореляційної таблиці з рівновіддаленими варіантами, то доцільно перейти до умовних варіант:

,

де *С1* і *С2* – хибні нулі варіант *X* і *Y* відповідно, *h1* і *h2* – кроки, тобто різниці між двома сусідніми варіантами *X* і *Y*. Тоді:

 (10.15)

 (10.16)

 (10.17)

 (10.18)

Виразимо коваріацію через умовні варіанти. Маємо:



де 



Таким чином,

 (10.19)

Приклад 2.Знайти рівняння прямих регресії *Y* по *X* та *X* за *Y*за даними кореляційної таблиці 10.4 і побудувати їхні графіки.

Перейдемо до умовних варіант

, де як хибні нулі *С1* і *С2* узяті варіанти відповідно *х* = 20 і y = 50, розташовані приблизно в середині варіаційних рядів, тобто *С1* = 20 і *С2* = 50, а *h1= 5* і *h2* = 10.

Таблиця 10.4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Y  X | 30 | 40 | 50 | 60 | 70 |  |
| 5 | 2 | — | — | — | — | 2 |
| 10 | 6 | 5 | — | — | — | 11 |
| 15 | — | 3 | 7 | 4 | — | 14 |
| 20 | — | — | 40 | 9 | 4 | 53 |
| 25 | — | — | 2 | 6 | 7 | 15 |
| 30 | — | — | — | — | 5 | 5 |
|  | 8 | 8 | 49 | 19 | 16 | *n*=100 |
|  |  |  |  |  |  |  |

Представимо кореляційну таблицю 10.4у вигляді таблиці 10.5, де два передостанніх стовпці й два передостанні рядки містять наступні суми:



Для зручності обчислення суми  спочатку розраховуємо  й проставляємо ці значення у верхньому правому кутку тих осередків, у яких, а потім знаходимо добутки . Підсумовуючи їх по рядках і стовпцям, записуємо отримані результати відповідно в останньому стовпці й останньому рядку таблиці 10.5. Підсумовуючи значення останніх стовпця й рядка, одержимо в правому нижньому вікні таблиці 10.5 =85; їхній збіг свідчить про правильність обчислень.

Використовуючи формули (10.15) – (10.19), одержимо:



Таблиця 10.5

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Y* | 30 | | 40 | | 50 | | 60 | | 70 | |  |  |  |  |
| *X* |  | –2 | | –1 | | 0 | | 1 | | 2 | |
| 5 | –3 |  | 6 | – | | – | | – | | – | | 2 | –6 | 18 | 12 |
| 2 |  |
| 10 | –2 |  | 4 |  | 2 | – | | – | | – | | 11 | –22 | 44 | 34 |
| 6 |  | 5 |  |
| 15 | –1 | – | |  | l |  | 0 |  | –l | – | | 14 | –14 | 14 | –1 |
| 3 |  | 7 |  | 4 |  |
| 20 | 0 | – | | – | |  | 0 |  | 0 |  | 0 | 53 | 0 | 0 | 0 |
| 40 |  | 9 |  | 4 |  |
| 25 | 1 | – | | – | |  | 0 |  | l |  | 2 | 15 | 15 | 15 | 20 |
| 2 |  | 6 |  | 7 |  |
| 30 | 2 | – | | – | | – | | – | |  | 4 | 5 | 10 | 20 | 20 |
| 5 |  |
|  | | 8 | | 8 | | 49 | | 19 | | 16 | | 100 | –17 | 111 | – |
|  | | –16 | | –8 | | 0 | | 19 | | 32 | | 27 | – | – | – |
|  | | 32 | | 8 | | 0 | | 19 | | 64 | | 123 | – | – | – |
|  | | 36 | | 13 | | 0 | | 2 | | 34 | | – | – | – | 85 |

За формулою (10.10) знайдемо вибірковий коефіцієнт кореляції*=* 44,795/(5,1988∙ 10,7569) ≈ 0,801, а потім, використовуючи (10.12) і (10.13), одержимо шукані рівняння регресії





На рисунку 10.2 представлені графіки отриманих прямих регресії.

**10.2 Поняття про множинну кореляцію**

У тих випадках, коли досліджується кореляційний зв'язок між величинами, число яких більше двох, уводять поняття *множинної* кореляції.





Рисунок 10.2– Графіки прямих регресії

Так, при дослідженні кореляційного зв'язку між трьома кількісними ознаками *X, Y* і Z можна ввести рівняння регресії

де  – середнє значення величини *Z*, що відповідає певним значенням *х* та *y.* Геометричною інтерпретацією цього рівняння є деяка поверхня в прямокутній системі координат тривимірного простору.

У найбільш простому випадку лінійної кореляційної залежності ознак *Х*, *Y* і *Z* вибіркове рівняння регресії має вигляд



Нехай у результаті незалежних спостережень над досліджуваною системою кількісних ознак (*X*;*Y*; *Z*), отримані *п* сукупностей чисел (*x1*,*y1*,*z1*), (*x2*, *y2*, *z2*), …, (*xn*, *yn*, *zn*)...Припустимо, що функція регресії лінійна, тобто

(10.20)

В (10.20) замінене на *z* , тому що різні значеннях *x* і *у* ознак *X* і *Y,* і відповідні їм значення *z* ознаки *Z* спостерігалися по одному разу. Якщо позначити  наближене значення *zi*, обчислене з рівняння регресії, то величина , є відхиленням наближеного значення , від точного . Коефіцієнти *a, b* і *з* рівняння регресії (10.20) знайдемо з вимоги методу найменших квадратів:



Необхідні умови мінімуму функції *S* утворять систему



яка в результаті тотожних перетворень набуває такого вигляду

 (10.21)

Розв’язуючи систему (10.21) щодо невідомих параметрів *a*, *b* і *с*, а потім, підставляючи їхні значення в (10.20), одержимо шукане рівняння регресії.

Зручніше, однак, знайти готові формули для розрахунку параметрів, а не розв’язувати систему щоразу. Коефіцієнти регресії *a, b* іпараметр *с*у цьому випадку обчислюються за наступними формулами:

 (10.22)

Тут – вибіркові коефіцієнти лінійної кореляції між відповідними ознаками; *sx*, *sy*, *sz –* середні квадратичне відхилення; – середні значення.

Тіснота зв'язку ознаки Z з ознаками *X* і *Y*, оцінюється *вибірковим сукупним коефіцієнтом кореляції*

 (10.23)

причому  .

Тіснота зв'язку між *Z* і *X* (при постійному *Y*), між *Z* і *Y*,(при постійному *X*) оцінюється відповідно *частковими вибірковими коефіцієнтами кореляції:*

 (10.24)

Ці коефіцієнти показують ступінь лінійної залежності між спостережуваними значеннями *Z* і *X,* а також *Z* і *Y*, якщо вплив третьої ознаки усунуто.

Приклад 5.Нехай змінна *z* перебуває в лінійній залежності від змінних *х* та*у.* Для оцінок параметрів рівняння регресії зібрані наступні дані:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *i* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| *zi* | 10 | 12 | 17 | 13 | 15 | 10 | 14 | 12 | 16 | 18 |
| *xi* | 2 | 2 | 8 | 2 | 6 | 3 | 5 | 3 | 9 | 10 |
| *yi* | 1 | 2 | 10 | 4 | 8 | 4 | 7 | 3 | 10 | 11 |

Знайти рівняння лінійної регресії, а також сукупний і частковий вибіркові коефіцієнти кореляції.Складемо розрахункову таблицю 10.6.

Таблиця 10.6

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *i* | *xi* |  |  |  |  |  |  |  |  |
| 1 | 2 | 4 | 1 | 1 | 10 | 100 | 2 | 20 | 10 |
| 2 | 2 | 4 | 2 | 4 | 12 | 144 | 4 | 24 | 24 |
| 3 | 8 | 64 | 10 | 100 | 17 | 289 | 80 | 136 | 170 |
| 4 | 2 | 4 | 4 | 16 | 13 | 169 | 8 | 26 | 52 |
| 5 | 6 | 36 | 8 | 64 | 15 | 225 | 48 | 90 | 120 |
| 6 | 3 | 9 | 4 | 16 | 10 | 100 | 12 | 30 | 40 |
| 7 | 5 | 25 | 7 | 49 | 14 | 196 | 35 | 70 | 98 |
| 8 | 3 | 9 | 3 | 9 | 12 | 144 | 9 | 36 | 36 |
| 9 | 9 | 81 | 10 | 100 | 16 | 256 | 90 | 144 | 160 |
| 10 | 10 | 100 | 11 | 121 | 18 | 324 | ПО | 180 | 198 |
|  | 50 | 336 | 60 | 480 | 137 | 1947 | 398 | 756 | 908 |

Побудуємо систему рівнянь, використовуючи розрахунки таблиці:



Розв’язок системи щодо невідомих параметрів дає шукані оцінки: *а* = 0,1285; *b* = 0,6117; *с* = 9,3872.

Рівняння регресії має вигляд



Для розв’язку цього прикладу знайдемо:









Тіснота зв'язку ознаки Z з ознаками *X* і *Y, X* (при постійному *Y*), *Y* (при постійному *X*)визначається за формулами (10.12) і (10.13):



**10.3 Використання вбудованих статистичних функцій з пакету аналізу MS Excel**

Для проведення кореляційного аналізу можуть бути використані функції:

− *КОВАР* − повертає значення коваріації, тобто середнє добутків відхилень для кожної пари точок даних. Коваріація використовується для визначення зв’язку між двома множинами даних і дає можливість установити, чи асоційовані набори даних по величині, тобто великі значення з одного набору даних зв’язані з великими значеннями іншого набору (позитивна коваріація), або, навпаки, малі значення одного набору зв’язані з великими значеннями іншого (негативна коваріація), чи дані двох діапазонів ніяк не зв’язані (коваріація близька до нуля).

− *КОРРЕЛ* − повертає коефіцієнт кореляції. Коефіцієнт кореляції використовується для визначення наявності взаємозв’язку між двома вибірками.

У вікні Аналіз даних для кореляційного аналізу використовується інструмент − *Кореляція* та *Коваріація*.

Для проведення регресійного аналізу можуть бути використані функції:

− *ЛИНЕЙН* − розраховує за допомогою методу найменших квадратів параметри для лінійного рівняння регресії;

− *ПРЕДСКАЗ* − передбачає значення на основі рівняння лінійної регресії. Цю функцію можна використовувати для передбачення майбутніх продаж, потреб в устаткуванні чи тенденцій споживання;

– *ТЕНДЕНЦІЯ* − повертає значення відповідно до лінійного тренду. У вікні Аналіз даних для регресійного аналізу використовується інструмент − Регресія, користуючись яким можна отримати таблиці з регресійною статистикою, коефіцієнтами регресії, стандартними помилками, *t*-статистикою та ін.

Питання для самоконтролю

1. Дайте визначення функціональної залежності між ознаками *X* і .
2. Яку залежність називають статистичною або стохастичною?
3. Що називається умовним середнім спостережуваних значень *Y*(*X*)?
4. Які рівняння називають модельними рівняннями регресії?
5. Що називається кореляційним полем?
6. Сформулюйте основні завдання регресійного аналізу.
7. Запишіть модель лінійної парної регресії *Y* по *X* і *Х* по *У.*
8. Поясніть, як заповнюється кореляційна таблиця.
9. Чому дорівнює вибірковий коефіцієнт кореляції?
10. Чому дорівнюють вибіркові коефіцієнти лінійної регресії?

Практичне завдання

***Завдання 1.*** Дані досліджень наведено в таблиці***.*** Скориставшись даними провести регресійний аналіз.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Маса, *xi* | 120 | 80 | 110 | 100 | 90 |
| Артеріальний тиск, yi | 150 | 110 | 135 | 140 | 115 |

***Завдання 2.*** За дослідними даними методом найменших квадратів визначити оцінки невідомих параметрів лінійної парної регресії. Обчислити коефіцієнт кореляції та детермінації, зробити висновки.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *xk* | 2 | 4 | 6 | 8 |
| *yk* | 0.81 | 2.3 | 3.4 | 4.54 |

***Завдання 3.***За дослідними даними методом найменших квадратів визначити оцінки невідомих параметрів рівняння регресії при параболічній залежності.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *xk* | 0 | 2 | 4 | 6 | 8 | 10 |
| *yk* | 5 | -1 | 0.5 | 1.5 | 1.5 | 8.5 |

***Завдання 4.*** Маючи статистичні дані про результати господарської діяльності підприємств побудувати багатофакторну лінійну модель залежності показника *y* (чистий прибуток підприємств, млн грн) від двох факторів: *x1* (основні фонди, млн грн), *x2* (оборотні фонди, млн грн).

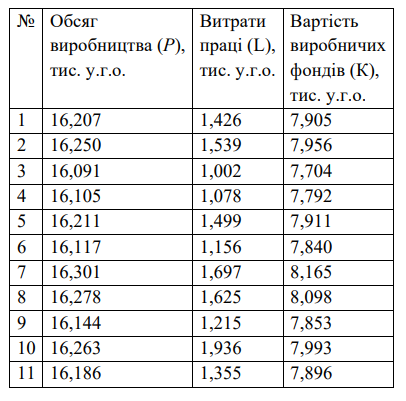
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *y* | 13 | 8 | 15 | 7 | 18 | 21 | 25 | 26 | 28 | 30 | 33 | 38 |
| *x1* | 21 | 19 | 18 | 15 | 24 | 23 | 21 | 24 | 25 | 26 | 25 | 29 |
| *x2* | 16 | 7 | 11 | 9 | 15 | 17 | 19 | 21 | 20 | 22 | 21 | 20 |

***Завдання 5.*** За дослідними даними методом найменших квадратів визначити оцінки невідомих параметрів рівняння регресії при показниковій залежності.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *xk* | 48 | 50 | 54 | 58 | 60 | 62 | 64 | 66 | 68 | 70 |
| *yk* | 18 | 18 | 19 | 26 | 25 | 26 | 30 | 28 | 26 | 34 |

***Завдання 6.*** Згідно з даними про обсяг виробленої продукції ***Р***, витрати праці ***L*** на її виробництво, а також вартість виробничих фондів ***К***, наведених у таблиці. Побудувати мультиплікативну виробничу функцію, що встановлює залежність між обсягом виробництва (результативний показник) та факторами – вартістю виробничих фондів та витратами праці. Здійснити оцінку адекватності отриманої моделі.

*Дані про обсяги виробництва, витрати праці та вартість виробничих фондів на підприємствах галузі за рік*



***Завдання 7.***Підприємство, що складається з багатьох філій, досліджує залежність свого річного товарообігу *y* (млн у. о.) від торгової площі своїх філій *x1* (тис. кв. м) і середньоденної інтенсивності потоку покупців *x2* (тис. чол./день). Побудувати багатофакторну лінійну модель залежності. Дані за філіями наведено в таблиці.

|  |  |  |  |
| --- | --- | --- | --- |
| Номер філії | Річний товарообіг | Торгова площа | Денна інтенсивність |
| 1 | 2,93 | 0,31 | 10,24 |
| 2 | 5,27 | 0,98 | 7,15 |
| 3 | 6,85 | 1,21 | 10,81 |
| 4 | 7,01 | 1,29 | 9,89 |
| 5 | 7,02 | 1,12 | 13,72 |
| 6 | 8,35 | 1,49 | 13,92 |
| 7 | 4,33 | 0,78 | 8,54 |
| 8 | 5,77 | 0,94 | 12,36 |
| 9 | 7,68 | 1,29 | 12,27 |
| 10 | 3,16 | 0,48 | 11,01 |
| 11 | 1,52 | 0,24 | 8,25 |
| 12 | 3,15 | 0,55 | 9,31 |

***Завдання 8.***По 20 підприємствах регіону вивчається залежність вироблення продукції на одного працівника *y* (тис. грн.) від введення в дію нових основних фондів *x1*(% від вартості фондів на кінець року) та від питомої ваги робітників високої кваліфікації у загальній чисельності робочих *x2* (%). Побудувати багатофакторну лінійну модель залежності.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| номер  підприємства | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| y | 7 | 7 | 7 | 7 | 7 | 7 | 8 | 8 | 8 | 10 | 9 |
| x1 | 3,9 | 3,9 | 3,7 | 4.0 | 3,8 | 4,8 | 5,4 | 4,4 | 5,3 | 6,8 | 6,0 |
| x2 | 10 | 14 | 15 | 16 | 17 | 19 | 19 | 20 | 20 | 20 | 21 |
|  |  |  |  |  |  |  |  |  |  |  |  |
| номер  підприємства | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |  |  |
| y | 11 | 9 | 11 | 12 | 12 | 12 | 12 | 14 | 14 |  |  |
| x1 | 6,4 | 6,8 | 7,2 | 8,0 | 8,2 | 8,0 | 8,5 | 9,0 | 9,6 |  |  |
| x2 | 22 | 22 | 25 | 28 | 29 | 30 | 31 | 32 | 36 |  |  |

Тестові завдання для перевірки знань

1. Дати визначення статистичної залежності між ознаками *X* та *Y* .

а) це така залежність, при якій зміна однієї з величин веде до змін розподілу іншої;

б) це така залежність, при якій зміна однієї з величин не веде до змін розподілу іншої;

в) це така залежність, яка проявляється в тому, що при зміні однієї із величин змінюється середнє значення іншої;

г) правильної відповіді серед перерахованих немає.

2. Що означає кореляційна залежність між ознаками *X* та *Y* ?

а) це така залежність, при якій зміна однієї з величин веде до змін розподілу

іншої;

б) це така залежність, при якій зміна однієї з величин не веде до змін розподілу іншої;

в) це така залежність, яка проявляється в тому, що при зміні однієї із величин змінюється середнє значення іншої;

г) правильної відповіді серед перерахованих немає.

3. Що називається функціями регресії?

а) можливі залежності між змінними *X* і *Y* за відсутності кореляційного

зв’язку між змінними;

б) можливі залежності між змінними *X* і *Y* за наявності кореляційного зв’язку між змінними;

в) це всі спадні функції між змінними *X* і *Y* ;

г) правильної відповіді серед перерахованих немає.

4. Який метод використовується для знаходження невідомих параметрів,

якщо між випадковими величинами *X* та *Y* існує лінійна функціональна залежність ?

а) метод невизначених коефіцієнтів;

б) метод найбільших квадратів;

в) метод найменших квадратів;

г) всі перераховані методи.

5. Коефіцієнт кореляції становить: *r* = - 0,56. Визначити характер кореляційного зв'язку

а) прямий;

б) зворотній;

в) відсутній.

6. Коефіцієнт кореляції становить: *r* = - 0,56. Визначити глибину (силу) кореляційного зв'язку

а) відсутня;

б) слабка;

в) сильна.

7. Які із статистичних гіпотез висувають при дослідженні масових процесів та явиш? Серед наведених тверджень оберіть правильні, яких може бути декілька:

а)основна, або нульова гіпотеза, яка стверджує, що значення параметру генеральної сукупності, яких обчислюється за вибірковою сукупністю, дорівнює нулю;

б) основна, або нульова гіпотеза, яка стверджує, що різниця між значеннями параметру, яких обчислюється за вибірковою сукупністю, і відповідним йому параметром генеральної сукупності дорівнює нулю;

в) альтернативна, або конкуруюча гіпотеза, яка стверджує, що значення параметру генеральної сукупності, яких обчислюється за вибірковою сукупністю, дорівнює нулю;

г) альтернативна, або конкуруюча гіпотеза, яка стверджує, що різниця між значеннями параметру, яких обчислюється за вибірковою сукупністю, і відповідним йому параметром генеральної сукупності дорівнює нулю;;

д)альтернативна, або конкуруюча гіпотеза, яка стверджує щось, що протирічить нульовій гіпотезі.

Варіанти відповідей: а) 2 та 4; б) 2 та 3; в) 1 та 3; г) 1 та 4; д) 1 та 5.

8. Яка із статистичних гіпотез підлягає перевірці? Оберіть правильний варіант із запропонованих:

а) висувають одночасно основну, або нульову, і альтернативну гіпотези і перевіряють основну гіпотезу, якщо ж вона не підтверджується, то приймають альтернативну без перевірки;

б) висувають одночасно основну і альтернативну гіпотези й одночасно перевіряють їх, щоб обирати ту, яка підтверджується;

в) перевіряють основну, або нульову гіпотезу, якщо вона не підтверджується, то висувають альтернативну і перевіряють її;

г) висувають одночасно основну та декілька альтернативних гіпотез і перевіряють їх по черзі поки не знайдуть ту, яка підтверджується;

д) висувають гіпотезу, яка стверджує про наявність ефекту, який досліджується, і її приймають за основну і перевіряють за відповідним статистичним критерієм.

9. Продовжить речення, щоб утворилося правильне твердження. «Перевірці статистичної гіпотези передує вибір рівня значущості α , який визначає…» Оберіть серед запропонованих варіантів правильний:

а) імовірності помилки I роду, яка полягає у тому, що правильна нульова гіпотеза буде помилково відхилена;

б) імовірність помилки II роду, яка полягає у тому, що хибна нульова гіпотеза буде помилково прийнята;

в) імовірність того, що нульова гіпотеза, яка є помилковою, буде відхилена;

г) імовірність того, що слід вважати правильною альтернативну гіпотезу;

д) імовірність того, що альтернативна гіпотеза є дійсно правильною.

10. Яке з наведених тверджень щодо кореляційного зв’язку є правильним? Оберіть правильну відповідь серед запропонованих:

а) кореляційний зв'язок є відображенням наявності причинно-наслідкового зв’язку між факторами, що досліджуються;

б) кореляційний зв'язок – це узгоджене зміна двох ознак, що відображає той факт, що мінливість однієї ознаки знаходиться у відповідності з мінливістю іншої;

в) наявність кореляційного зв'язку визначає лише один із факторів як функцію від інших факторів;

г) наявність кореляційного зв'язку визначає, що кожен із факторів можна надати як функцію від іншого фактору;

д) кореляційний зв'язок завжди є кількісним відображенням взаємного впливу декількох якісних факторів.

11. Який висновок слід зробити відносно щільності кореляційного зв’язку, якщо коефіцієнт кореляції Пірсона приймає таке значення 0 <*r*< 0,3? Оберіть правильну відповідь серед запропонованих:

а) кореляційний зв'язок є прямим, але слабким;

б) кореляційний зв'язок є прямим;

в) кореляційний зв'язок є статистично незначущім;

г) висновок зробити неможна, потрібна додаткова перевірка за критерієм Стьюдента;

д) можна стверджувати, що кореляційний зв'язок відсутній.

12. Який метод багатофакторної статистики застосовується для дослідження впливу якісних факторів і аналізу значущості цього впливу?

а) кореляційний аналіз;

б) регресійний аналіз;

в) метод найменших квадратів;

г) дисперсійний аналіз;

д) метод найбільшої правдоподібності.

13. Як називається числова характеристика, що застосовується для оцінювання якості моделі і повноти набору пояснювальних факторів?

а) коефіцієнт детермінації;

б) коефіцієнт кореляції;

в) регресія;

г) систематична помилка.

14. Результати вимірювань можуть містити випадкові і систематичні помилки. Оберіть серед наведених правильні твердження. Таких варіантів може буди декілька.

а) застосування методу найменших квадратів забезпечує видалення систематичних помилок;

б) систематичні помилки впливають на значення вільного члена рівняння регресії;

в) випадкові помилки впливають на значення як вільного члена рівняння регресії, так і коефіцієнта регресії;

г) наявність систематичних помилок призводить до того, що розрахунки за моделлю будуть хибними;

д) кут, на який лінія регресії може повертатись в межах довірчого інтервалу, залежить не від наявності помилок моделі, а лише від щільності кореляційного зв’язку.

15. Скільки рівнянь входить до системи нормальних рівнянь, за якою здійснюється оцінювання параметрів моделі парної лінійної регресії?

16. Продовжить речення, щоб утворилося правильне твердження. «Метод найменших квадратів передбачає, що певна функція, яка є функцією від параметрів моделі, досліджується на мінімум. Цією функцією є ...».

# ЗМІСТОВИЙ МОДУЛЬ 6. МЕТОДОЛОГІЯ АНАЛІЗУ КАТЕГОРІАЛЬНИХ ДАНИХ

## Тема 11. Кластерний аналіз

Мета:вивчити методи кластерного аналізу; засвоїти особливості ієрархічних методів.

🖉Основні терміни і поняття:

К*ластер, кластерізація, центр кластера, радіус кластера, ієрархічні методи кластерного аналізу, агломеративні, дивізимні методи, неієрархічні методи, дендрограма,Евклідова метрика*, *Манхеттенска відстань, відстань Чебишева.*

📚Основні теоретичні положення

**11.1 Сутність кластерного аналізу**

Термін кластерний аналіз, уперше введений Тріоном (Tryon) у 1939 році, містить більш 100 різних алгоритмів. Кластеризація, або кластерний аналіз – це статистична процедура, задача якої полягає в розбитті вибірки об'єктів на підмножини, що не перетинаються і називаються кластерами. Кожен кластер має складається зі схожих об'єктів, а об'єкти різних кластерів мають істотно відрізнятися один від одного. Задача кластеризації відноситься до статистичної обробки, а також до широкого класу задач навчання без вчителя. Ще її можна описати через задачу класифікації. На відміну від задач класифікації, кластерний аналіз не вимагає апріорних припущень про набір даних, не накладає обмеження на показ досліджуваних об’єктів, дозволяє аналізувати показники різних типів даних (інтервальні дані, частоти, бінарні дані). При цьому необхідно пам’ятати, що змінні повинні вимірюватися в порівнюваних шкалах. Кластерний аналіз дозволяє скорочувати розмірність даних, робити їх наглядними. Кластерний аналіз може застосовуватися до сукупностей тимчасових рядів, тут можуть виділятися періоди схожості деяких показників і визначатися групи тимчасових рядів зі схожою динамікою.

Кластерний аналіз – це спосіб угруповання багатовимірних об'єктів, заснований на поданні результатів окремих спостережень точками відповідного геометричного простору з наступним виділенням груп, як «згустків» цих точок. Іншими словами, кластерний аналіз – аналіз некласифікованих об'єктів, проведений з метою виділення структур, класів образів, множин подібних об'єктів, таких, що об'єкти всередині груп були б схожі в певному сенсі один на одного, а об'єкти з різних груп – несхожі.

Термін кластер (cluster) з англійської перекладається як «згусток», «гроно»,«скупчення». Цей термін дуже вдало вписався у наукову термінологію, оскільки його перший склад відповідає традиційному терміну «клас», а другий як похідна від першого.

Визначення кластерного аналізу у різних літературних джерелах дається по-різному. Однією з причин цього є та обставина, що кластерний аналіз використовується і вдосконалюється в різних областях, як соціологія, психологія, екологія, геологія, медицина тощо. У кластерному аналізі немає однозначного кількісного критерію, оскільки в різних прикладних задачах різними можуть бути й цілі аналізу. Іноді необхідно виділити групи із високою щільністю розподілу та малої дисперсії, а іноді необхідно виявити пов'язані точні структури. Кластерний аналіз паралельно розвивався в декількох напрямках, тому більшість методів мають по дві й більш назв. Це суттєво ускладнює роботу при використанні кластерного аналізу.

**11.2Задачі кластерного аналізу**

Кластеризація є описовою процедурою, вона не робить жодних статистичних висновків, але надає можливість провести розвідницький аналіз та вивчити структуру даних. Кластерний аналіз призначений для розбиття множини об’єктів на визначене чи невідоме число класів (кластерів) згідно з певним критерієм якості класифікації

*Кластер* – це група об’єктів зі спільними властивостями.

Кластер має такі математичні характеристики: центр, радіус, середньоквадратичне відхилення, розмір кластера.

*Центр кластера* – це середнє геометричне місце точок у просторі змінних.

*Радіус кластера* – максимальна відстань точок від центру кластера. Кластери можуть бути, такими, що перекриваються. Така ситуація виникає, коли виявляється перекриття кластерів. У цьому випадку неможливо за допомогою математичних процедур однозначно віднести об’єкт до одного з двох кластерів. Такі об’єкти називають спірними.

*Спірний об’єкт* – це об’єкт, який у міру подібності може бути віднесений до декількох кластерів.

*Розмір кластера* може бути визначений або за радіусом кластера, або за середньоквадратичним відхиленням об’єктів для цього кластера. Об’єкт належить до кластера, якщо відстань від об’єкта до центру кластера менше радіуса кластера. Якщо ця умова виконується для двох і більш кластерів, об’єкт є спірним. Неоднозначність може бути усунута експертом або аналітиком.

Робота кластерного аналізу опирається на два припущення. Перше припущення – розглянуті ознаки об’єкта в принципі допускають бажане розбиття сукупності об’єктів на кластери. Друге припущення – правильність вибору масштабу або одиниці вимірювання ознак.

Вибір масштабу в кластерному аналізі має велике значення. Розглянемо приклад. Уявимо собі, що дані ознаки *X* у наборі даних на два порядки більші даних ознаки *Y*: значення змінної *X* перебувають в діапазоні від 1000 до 5000, а значення змінної *Y* – у діапазоні від 10 до 50. Тоді, при розрахунках величин відстані між точками, що відображають положення об’єктів у просторі їх властивостей, змінна, що має більші значення, тобто змінна *х*, буде практично повністю домінувати над змінною з малими значеннями, тобто змінної *y*. У такий спосіб через неоднорідність одиниць виміру ознак стає неможливим коректна розрахувати відстані між точками. Ця проблема вирішується за допомогою попередньої стандартизації змінних. Стандартизація (standardization) або нормування (normalization) приводить значення всіх перетворених змінних до єдиного діапазону значень шляхом вираження через відношення цих значень до якоїсь величини, що відображає певні властивості конкретної ознаки. Існують різні способи нормування вихідних даних.

Два найпоширеніші способи стандартизації:

* розподіл вихідних даних на середньоквадратичне відхилення відповідних змінних;
* обчислення *Z*-внеску або стандартизованого внеску.

Як правило, при практичному використанні кластерного аналізу одночасно розв’язуються декілька із зазначених задач.

Приклади конкретних задач кластеризації:

* Сегментація цільової аудиторії сайту.
* Ідентифікація груп сімей — споживачів певного товару для розробки стратегії позиціонування бренду.
* Тематичне моделювання електронних листів.
* Кластеризація символів в незалежності від їх шрифту, розміру тощо (для подальшого розпізнавання).

Розглянемо приклад процедури кластерного аналізу. Допустимо, маємо набір даних, що складається з 14-ти об’єктів, у яких є по дві ознаки *X* та *Y*.

Дані в табличній формі не носять інформативний характер, представимо змінні *X* і *Y* у вигляді діаграми розсіювання (рисунок 11.1).

|  |  |  |
| --- | --- | --- |
| Номер за порядком | Ознака X | Ознака Y |
| 1 | 27 | 19 |
| 2 | 11 | 46 |
| 3 | 25 | 15 |
| 4 | 36 | 27 |
| 5 | 35 | 25 |
| 6 | 10 | 43 |
| 7 | 11 | 44 |
| 8 | 36 | 24 |
| 9 | 26 | 14 |
| 10 | 26 | 21 |
| 11 | 9 | 45 |
| 12 | 33 | 23 |
| 13 | 27 | 16 |
| 14 | 10 | 47 |

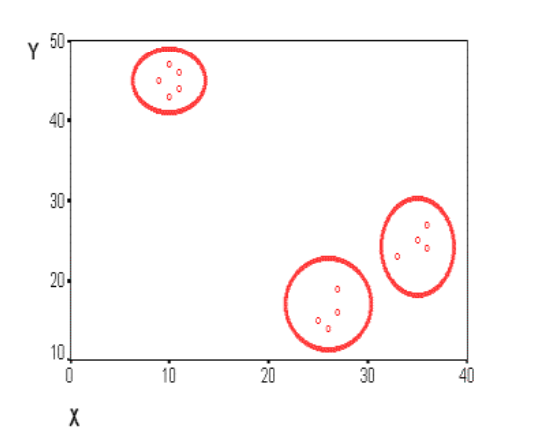


Рисунок 11.1 – Діаграма розсіювання змінних *X* та *Y*

На рисунку бачимо кілька груп «схожих» об’єктів. Об’єкти, які за значеннями *X* і *Y* «схожі» один на одного, належать до однієї групи (кластеру); об’єкти з різних кластерів не схожі один на одного.

**11.3 Методи кластерного аналізу**

Задача кластеризації — це по факту задача класифікації, бо в обох випадках ми ділимо об'єкти на основі їх подібності між собою, але у випадку кластеризації приналежність навчальних об'єктів будь-яким класам не задається. Така задача — загальна, тому для її розв'язання використовуються різні підходи. Алгоритми побудови кластерів можуть дуже відрізнятися у підходах до того, що відносити в один кластер і як їх взагалі ефективніше шукати. Кластери можна утворювати ґрунтуючись на відстані між ними, на щільності ділянок у просторі даних, інтервалах або на конкретних статистичних розподілах. Це все залежать від конкретного набору даних та мети використання результатів.

Кластерний аналіз не є автоматизованим, це скоріше ітераційний процес, тому що часто доводиться змінювати метод опрацювання даних та параметри моделі, поки не буде отримано з результат з заданими властивостями. Розв'язок неоднозначний, і на це є кілька причин. По-перше, не існує найкращого критерію якості кластеризації. Відомий цілий ряд досить ефективних критеріїв, а також ряд алгоритмів, які не мають чітко вираженого критерію, але все одно здійснюють досить якісну кластеризацію по побудові. Всі вони можуть давати різні результати. По-друге, число кластерів, як правило, не відомо заздалегідь і встановлюється відповідно до деякого суб'єктивного критерія. По-третє, результат кластеризації істотно залежить від метрики, вибір якої, як правило, також суб'єктивний і визначається спеціалістом.

Задача групування набору об'єктів полягає в тому, що об'єкти в одному кластері більш схожі один на одного, ніж об'єкти в інших кластерах. Подібність — це буквально кількість, яка собою відображає міцність взаємозв'язку між двома об'єктами. Кластеризація використовується в основному для видобутку даних, а також в інших областях, таких як машинне навчання, розпізнавання образів, аналіз зображень, пошук інформації, біоінформатика, стиснення даних і комп'ютерна графіка.

Існує два типа кластеризації: жорстка та м'яка. У жорсткій кластеризації кожен об'єкт даних або повністю належить кластеру чи взагалі не належить. В м'якій кластеризації точка чи об'єкт даних може з певною ймовірністю належати більш ніж одному кластеру.

Типи кластерних алгоритмів:

* Пласкі алгоритми:

• починають роботу розділенням елементів по групах випадковим чином;

• ітеративно покращують результат;

• головний алгоритм: А-середніх (A-means).

* Жорстка кластеризация:

• кожен елемент належить строго одному кластеру.

* Алгоритми ієрархічної кластеризації:

• створюють ієрархію;

• знизу-вгору (агломеративні);

• зверху-вниз (розділяючі).•

* М'яка кластеризация:

• елемент може належати кільком кластерам;

• головний алгоритм: С-середніх (C-means).

Методи кластерного аналізу можна поділити на дві групи:

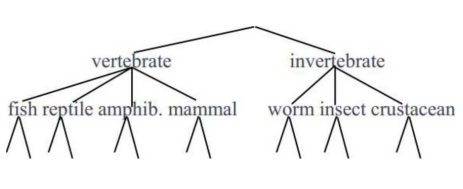
* ієрархічні;
* неієрархічні.

Кожна з груп включає безліч підходів та алгоритмів.

Використовуючи різні методи кластерного аналізу, аналітик може отримати різні рішення для тих самих даних. Це вважається нормальним явищем.

*Ієрархічні методи кластерного аналізу*

Завдання ієрархічної кластеризації — побудувати ієрархію кластерів. Ієрархія будується автоматично: або згори-вниз (агломеративні алгоритми) – AGNES (AGglomerative NESting): ROCK, CURE, CHAMELEON або знизу-вгору (алгоритми розділення) – DIANA (DIvisive ANAlysis): BIRCH, MST.



Як в агломеративній, так і в роздільній ієрархічній кластеризації, користувачам потрібно вказати бажану кількість кластерів та умову завершення. Отже, суть ієрархічної кластеризації полягає у послідовному об'єднанні менших кластерів у великі або поділі великих кластерів на менші. Вони поділяються на ієрархічні агломеративні та ієрархічні дивізимні методи.

Принцип роботи даних методів показаний на рисунку 11.2.

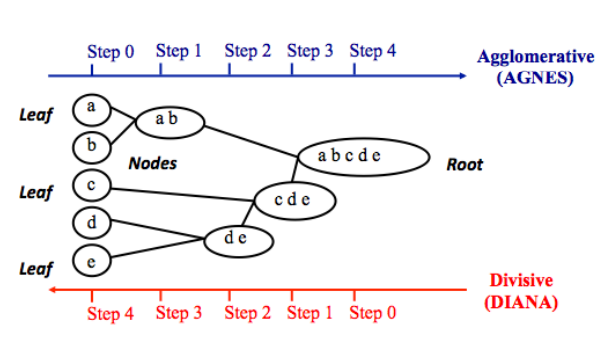


Рисунок 11.2 – Ієрархічна кластерізація

*Ієрархічні агломеративні методи* (Agglomerative Nesting, AGNES). Ця група методів характеризується послідовним поєднанням вихідних елементів та відповідним зменшенням числа кластерів. На початку роботи алгоритму усі об'єкти є окремими кластерами. На першому етапі найбільш схожі об'єкти об'єднуються в кластер. На наступних кроках об'єднання триває доти, доки всі об'єкти не будуть складати один кластер. Отже, для агломератних методів характерно те, що нові кластери утворюються шляхом об'єднання дрібніших кластерів, і таким чином дерево створюється від листя до стовбура.

*Ієрархічні дивізимні (розподілені) методи* (DIvisive ANAlysis, DIANA) Ці методи є логічною протилежністю агломеративним методам. На початку роботи алгоритму всі об'єкти належать одному кластеру, який на наступних кроках ділиться на менші кластери, у результаті утворюється послідовність груп, що розщеплюють. Отже, для дивізимних методів характерно те, що нові кластери створюються шляхом ділення більших кластерів на більш дрібні, і таким чином дерево створюється від стовбура до листя.

Програмна реалізація алгоритмів кластерного аналізу широко представлена у різних інструментах Data Mining, які дозволяють вирішувати завдання досить великої розмірності. Наприклад, агломеративні методи реалізовані у пакеті SPSS, дивізимні методи – у пакеті Statgraf.

Ієрархічні методи кластеризації відрізняються правилами побудови кластерів. В якості правил виступають критерії, що використовуються під час вирішення питання «схожості» об'єктів при їх об'єднанні у групу (агломеративні методи) чи поділу на групи (дивізимні методи).

Ієрархічні методи кластерного аналізу використовуються при невеликих обсягах наборів даних. Перевагою ієрархічних методів кластеризації є їхня наочність.

Ієрархічні алгоритми кластеризації, також відомі як алгоритми таксономії, будують не одне розбиття вибірки на непересічні класи, а систему вкладеного розбиття, що дає змогу значно спростити обробку даних та ухвалення рішень. Таким чином будується система вкладеного розбиття. Результатом таксономії є деревоподібна ієрархічна структура, яка представляється у вигляді таксономічного дерева – дендрограми, яка зображена на рисунку 11.3.

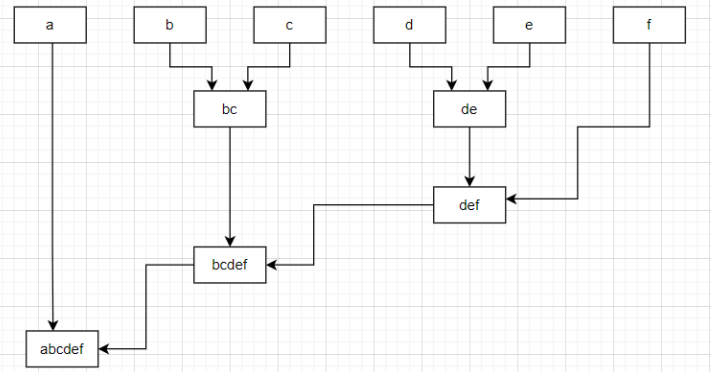


Рисунок 11.3 – Дендрограма

Дендрограма (від грецького dendron – «дерево») описує близькість окремих точок і кластерів один до одного, представляє в графічному вигляді послідовність об’єднання (поділу) кластерів. Дендрограма (dendrogram) – деревоподібна діаграма, що містить *n* рівнів, кожний з яких відповідає одному з кроків процесу послідовного укрупнення кластерів. Дендрограму також називають деревоподібною схемою, деревом об’єднання кластерів, деревом ієрархічної структури. Дендрограма являє собою вкладене угруповання об’єктів, яке змінюється на різних рівнях ієрархії. Існує багато способів побудови дендограм. У дендограмі об’єкти можуть розташовуватися вертикально або горизонтально. При цьому об’єкт характеризується перерахунком всіх кластерів, яким він належить.

Дендограми дозволяє уявити кластерну структуру у вигляді плаского графіка незалежно від того, яка розмірність початкового простору. Існують і інші способи візуалізації багатовимірних даних, такі як багатовимірне шкалювання або карти Кохонена, але вони привносять в картину штучні спотворення, вплив яких досить важко оцінити.

Подібність кластерів часто розраховується через «неподібність», наприклад, евклідова відстань між двома кластерами. Отже, чим більше відстань між двома кластерами, тим краще.

Ключовою операцією в ієрархічній агломераційній кластеризації є неодноразове об'єднання двох найближчих кластерів у один кластер, але дуже важливо спочатку відповісти на три питання: як ви представити кластер з більш ніж однією точкою, як визначити «близькість» кластерів та коли перестати поєднувати кластери. Злиття кластерів припиняється в залежності від доступної інформації про дані, які ми маємо. Якщо групувати футболістів на полі на основі їхніх позицій на полі, яке представлятиме їх координати для розрахунку відстані між гравцями, очевидно, що треба зупинитися на лише двох кластерах, оскільки можуть бути тільки дві команди, які грають у футбольний матч.

*Алгоритм ієрархічної агломеративної кластеризації*:

1. Обчислити матрицю близькості, що містить відстань між кожною парою шаблонів. Розглядати кожен зразок як окремий кластер.

2. Знайти найбільш схожу пару кластерів за допомогою матриці близькості. Об'єднати ці два кластера в один більший кластер. Оновити матрицю близькості, щоб відобразити цю операцію злиття.

3. Якщо всі шаблони знаходяться в одному кластері, зупинитися. В іншому випадку перейти до кроку 2.

У разі використання ієрархічних алгоритмів постає питання, як з'єднувати між собою кластери, як вираховувати «відстані» між ними.

*Міри подібності.* Критерієм для визначення схожості й відмінності кластерів є відстань між точками на діаграмі розсіювання. Цю подібність можна «виміряти», вона дорівнює відстані між точками на графіку (рис. 11.4).

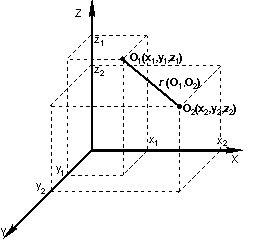


Рисунок 11.4 – Відстань між двома точками в просторі трьох вимірів

Для обчислення відстані між об’єктами використовуються різні міри подібності, їх називають також метриками або функціями відстаней.

Формально задача кластеризації описується наступним чином. Дано множину об’єктів даних *I*, кожен з яких представлений набором атрибутів. Потрібно побудувати множину кластерів *C* і відображення *F* множини *I* на множину *C, F: I -> C*. Відображення *F* задає модель даних, що являється власне вирішенням задачі.

Множина *I* визначається наступним чином

*I = {i1, i2, …, ij, …, in },*

де *ij*– об’єкт, що досліджується.

Кожен із об’єктів характеризується набором параметрів:

*ij = {x1, x2, …, xh, …, xm }*

Задача кластеризації полягає в побудові множини:

*C = {c1, c2, …, ck, …, xg },*

де *ck*– кластер, який має в собі похожі один на одного об’єкти з множини *I*:

*ck = {il, ip | il є I, ip є I, d(il, ip) < σ },*

де *σ* – величина, яка визначає міру близькості для включення об’єктів в один кластер, *d(il, ip)* – відповідно міра близькості.

Метрики, за допомогою яких можна рахувати міру близькості:

*Евклідова метрика*. Евклідова відстань (евклідова метрика) є відомою із загальноматематичних дисциплін і визначається за формулою:

Евклідову відстань доцільно обирати, якщо:

* + - спостереження належать до генеральних сукупностей, які підпорядковуються багатовимірним нормальним законам, а компоненти вектору спостережень є незалежними і мають одну й ту саму дисперсію;
    - компоненти вектору спостережень є однорідними з погляду змістової інтерпретації та однаково важливими для класифікації;
    - простір ознак має розмірність 1, 2 або 3, і поняття близькості об’єктів у цьому просторі збігається із звичайною геометричною близькістю.

Недоліком евклідової метрики є те, що у випадках, коли ознаки виміряні у різних одиницях, зміна масштабу одиниць вимірювання може призвести до істотної зміни результатів класифікації. Для запобігання цьому використовують різні методи нормування даних. Однак, нормування також впливає на результати класифікації. Зокрема, у випадках, коли кластери істотно розділяються за деякими ознаками й слабко за іншими, нормалізація може призвести до зменшення дискримінуючих можливостей першої групи ознак через збільшення шумового ефекту інших. Якщо ознаки вимірюють у якісно різних одиницях, то застосування евклідової відстані взагалі може виявитися недоцільним.

*Зважену евклідову відстань* розраховують за формулою:

Зазвичай беруть . Визначення вагових коефіцієнтів за аналізованою вибіркою, як правило, є недоцільним, оскільки може призвести до істотних помилок. Зокрема, залежно від певних незначних варіацій змістової та статистичної природи вихідних даних може бути обґрунтованим надання їм значень, пропорційних середньоквадратичній похибці відповідної ознаки або оберненій до цієї похибки величині. Тому рекомендують обирати вагові коефіцієнти за результатами експертних опитувань або інших незалежних попередніх досліджень.

*Квадрат евклідової відстані.* Для надання більшої ваги більш віддаленим один від одного об’єктів можемо скористатися квадратом евклідової відстані шляхом піднесення у квадрат стандартної евклідової відстані.

*Манхеттенска відстань* (відстань міських кварталів), також називається «хемінговою» або «сіті-блок» відстанню.ЇЇ ввів відомий американський математик Річард Хеммінг у 1950 р., використовують як міру відстані об’єктів, що характеризуються дихотомічними ознаками. Метрика Хеммінга – це відстань є просто середнім різниці за координатами. У більшості випадків дана міра відстані приводить до таких же результатів, як і для звичайного відстані Евкліда, проте для неї вплив окремих великих різниць (викидів) зменшується (так як вони не зводяться в квадрат). Відстань за Хеммінгом обчислюється по формулі:

*Відстань Чебишева.* Цю відстань варто використовувати, коли необхідно визначити два об’єкти як «різні», якщо вони відрізняються за якимось одним виміром. Відстань Чебишева обчислюється за формулою:

*Методи об’єднання або зв’язки*. При конструюванні різноманітних процедур класифікації доцільно використовувати міри близькості кластерів один до одного. Найбільш поширеними з них є відстані, що вимірюють за принципами найближчого і далекого сусідів, середнього зв’язку та за центрами ваги.

*Метод найближчого сусіда або одиночний зв’язок*. Відстань між двома кластерами визначається відстанню між двома найбільш близькими об’єктами (найближчими сусідами) у різних кластерах. Метод дозволяє виділяти кластери як завгодно складної форми за умови, що різні частини таких кластерів з’єднані ланцюжками близьких один до одного елементів. У результаті роботи даного методу кластери представляються довгими «ланцюжками» або «волокнистими» кластерами, «зчепленими разом» тільки окремими елементами, які випадково виявилися ближче інших один до одного.

*Метод найбільш віддалених сусідів або повний зв’язок*. Тут відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об’єктами в різних кластерах (тобто «найбільш віддаленими сусідами»). Метод добре використовувати, коли об’єкти дійсно походять із різних «ділянок». Якщо ж кластери мають до певної міри подовжену форму або їх природній тип є «ланцюговим», то цей метод не слід використовувати.

*Метод незваженого попарного середнього* (метод незваженого попарного арифметичного середнього – unweighted pair-group method using arithmetic averages, UPGMA. За відстань між двома кластерами береться середня відстань між усіма парами об’єктів у них. Цей метод слід використовувати, якщо об’єкти дійсно походять із різних «ділянок», у випадках присутності кластерів «ланцюгового» типу, при припущенні нерівних розмірів кластерів.

*Метод зваженого попарного середнього* (метод зваженого попарного арифметичного середнього – weighted pair-group method using arithmetic averages, WPGM A. Цей метод схожий на метод незваженого попарного середнього, різниця полягає лише в тому, що тут як ваговий коефіцієнт використовується розмір кластера (число об’єктів, що втримуються в кластері). Рекомендується використовувати саме при наявності припущення про кластери різних розмірів.

*Незважений центроїдний метод* (метод незваженого попарного центроїдного усереднення – unweighted pair-group method using the centroid average. За відстань між двома кластерами в цьому методі береться відстань між їхніми центрами ваги.

*Зважений центроїдний метод* (метод зваженого попарного центроїдного усереднення – weighted pair-group method using the centroid average, WPGMC. Цей метод схожий на попередній, різниця полягає в тому, що для обліку різниці між розмірами кластерів (числа об’єктів у них), використовуються ваги. Використовують переважно у випадках, якщо є припущення щодо істотних відмінностей у розмірах кластерів.

*Метод Варда (Ward’s method).* За відстань між кластерами береться приріст суми квадратів відстаней об’єктів до центрів кластерів, одержуваний у результаті їх об’єднання. На відміну від інших методів кластерного аналізу для оцінки відстаней між кластерами, тут використовуються методи дисперсійного аналізу. На кожному кроці алгоритму поєднуються такі два кластери, які приводять до мінімального збільшення цільової функції, тобто внутрішньо групової суми квадратів. Цей метод спрямований на об’єднання близько розташованих кластерів і «прагне» створювати кластери малого розміру. В цілому даний метод є дуже ефективним, однак він підходящий для створення кластерів невеликого розміру.

**11.4 Кластерний аналіз за допомогою пакета SPSS**

Розглянемо процедуру ієрархічного кластерного аналізу в пакеті SPSS (SPSS), в якому ця процедура передбачає угруповання як об’єктів (рядків матриці даних), так і змінних (стовпців). Можна вважати, що в останньому випадку роль об’єктів відіграють змінні, а роль змінних – стовпці.

У цьому методі реалізується ієрархічний агломеративний алгоритм, зміст якого полягає в такому. Перед початком кластеризації всі об’єкти вважаються окремими кластерами, у ході алгоритму вони поєднуються. Спочатку вибирається пара найближчих кластерів, які поєднуються в один кластер. У результаті кількість кластерів стає рівним *N-*1. Процедура повторюється, поки всі класи не об’єднаються. На будь-якому етапі об’єднання можна перервати, одержавши потрібне число кластерів. Отже, результат роботи алгоритму агрегування залежить від способів обчислення відстані між об’єктами й визначення близькості між кластерами.

Для визначення відстані між парою кластерів можуть бути сформульовані різні підходи. З урахуванням цього в SPSS передбачені такі методи:

• Середня відстань між кластерами (Between-groups linkage), установлюється за замовчуванням.

• Середня відстань між усіма об’єктами пари кластерів з урахуванням відстаней усередині кластерів (Within-groups linkage).

• Відстань між найближчими сусідами – найближчими об’єктами кластерів (Nearest neighbor).

• Відстань між самими далекими сусідами (Furthest neighbor).

• Відстань між центрами кластерів (Centroid clustering) або центроїдний метод. Недоліком цього методу є те, що центр об’єднаного кластера обчислюється як середнє центрів поєднуваних кластерів, без обліку їх обсягу.

• Метод Варда.

• Метод медіан – той же центроїдний метод, але центр об’єднаного кластера обчислюється як середнє всіх об’єктів (Median clustering).

*Приклад ієрархічного кластерного аналізу.* Розглянемо дані про процентний склад робітників у різних галузях промисловості 26 європейських країн у 1979 році. Важливо відмітити, що ці дані зібрані у часи «холодної війни». Дані зберігаються у файлі з розширенням .sav.

Текстова змінна *Country* використовується для визначення назви країни. Метою цього кластерного аналізу є знаходження країн зі схожими властивостями.

1. Відкрийте файл .sav.
2. Виберіть, як показано на рисунку 11.5, у меню ***Analyze (Аналіз) >Classify (Класифікувати) > Hierarchical Cluster... (Ієрархічний кластерний аналіз).***

Використані наступні скорочення робітничих категорій:

• *Agr:* Відсоток зайнятих у сільському господарстві

• *Min:* Відсоток зайнятих у гірничодобувній галузі

• *Man:* Відсоток зайнятих в обробній промисловості

• *PS*: Відсоток зайнятих в енергетичній галузі

• *Con:* Відсоток зайнятих у будівництві

• *SI*: Відсоток зайнятих у сфері обслуговування

• *Fin*: Відсоток зайнятих у фінансовій сфері

• *SPS*: Відсоток зайнятих у соціальних службах

*• TC*: Відсоток зайнятих у транспорті і комунікаціях

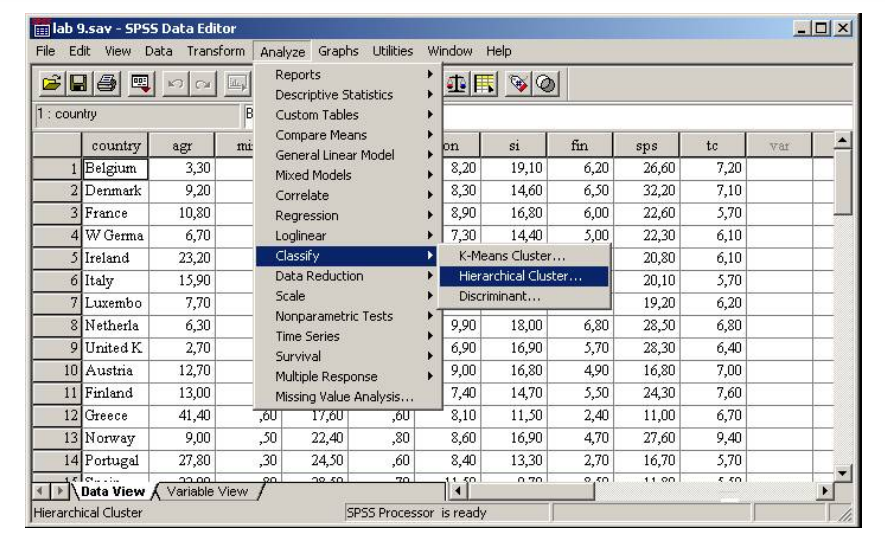


Рисунок11.5 – Вибір Ієрархічного кластерного аналізу

1. У діалоговому вікні ***Hierarchical Cluster Analysis (Ієрар-хічний кластерний аналіз)*** змінні ***agr … tc*** помістіть, як показано на рисунку 11.6, у поле змінних Variable(s), що тестуються, а текстову змінну ***country*** (країна) використовуйте для позначення (маркірування) спостережень і перенесіть у поле з ім’ям ***Label cases by: (Найменування (мітки) спостережень:)***.
2. Повернувшись у головне діалогове вікно ***Hierarchical Cluster Analysis***, клацніть по вимикачу ***Plots***... (Діаграми). Активуйте опцію виводу деревоподібної діаграми ***(Dendrogram***) і за допомогою опції ***None (Немає)*** скасуйте видачу накопичувальної діаграми (рисунок 11.7).

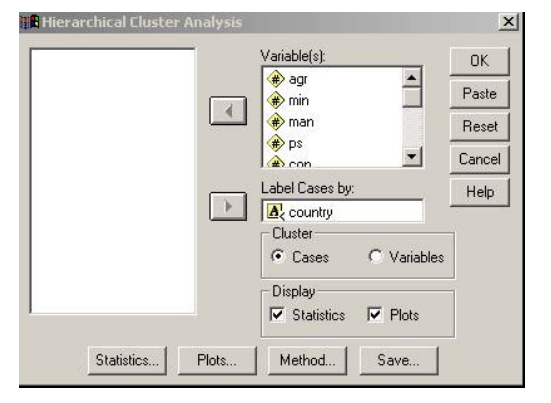


Рисунок 11.6 – Діалогове вікно Ієрархічний кластерний аналіз

1. За допомогою кнопки ***Method... (Метод)*** є можливість вибрати метод утворення кластерів, а також метод розрахунку дистанційної міри.

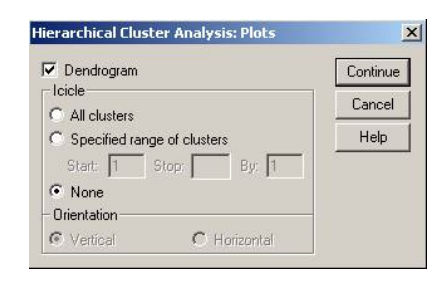


Рисунок 11.7 – Діалогове вікно Hierarchical Cluster Analysis

(Ієрархічний кластерний аналіз)

1. Залишіть попередні установки й у полях ***Cluster Method*** та ***Measure***.
2. У полі ***Transform Values > Standardize*** (Перетворювати значення) встановіть z-перетворення (стандартизацію) значень, як показано на рисунку 11.8; необхідність цієї опції була вже розглянута вище. Інші пропоновані можливості стандартизації відіграють другорядну роль.
3. Поверніться назад у головне діалогове вікно і почніть розрахунок натисканням ОК. Результати аналізу мають вид, представлений на рисунку 11.9.

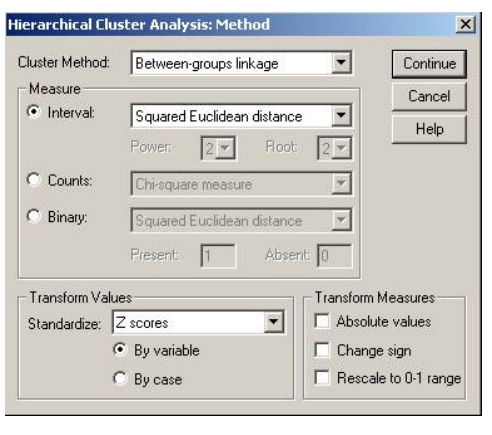


Рисунок 11.8 – Z-перетворення (стандартизація)

Після звичайного статистичного зведення підсумків за спостереженнями, у вікні перегляду наводиться таблиця ***Agglomeration Schedule (Порядок агломерації)***, з якої можна з’ясувати черговість побудови кластерів, а також їхньої оптимальної кількості. У двох стовпчиках, розташованих під загальною шапкою ***Cluster Combined (Об’єднання у кластери)***, можна побачити, що на першому кроці були об’єднані спостереження 2 і 16 (тобто Данія і Швеція); ці дві країни максимально схожі одна на одну (хто б міг подумати? :-) і віддалені одна від одної на дуже малу відстань – 1,288. Ці два спостереження утворюють спільний кластер з номером 9, у той час як кластери 2 та 16 в оглядовій таблиці більше не з’являються. На наступному кроці відбувається об’єднання спостережень 1 і 3 (Бельгія і Франція), потім 19 і 23 (Болгарія і Польща) і т.д.

1. Поверніться назад у головне діалогове вікно і почніть розрахунок натисканням ОК. Результати аналізу мають вид, представлений на рисунку 11.9.

Після звичайного статистичного зведення підсумків за спостереженнями, у вікні перегляду наводиться таблиця ***Agglomeration Schedule (Порядок агломерації)***, з якої можна з’ясувати черговість побудови кластерів, а також їхньої оптимальної кількості. У двох стовпчиках, розташованих під загальною шапкою ***Cluster Combined (Об’єднання у кластери)***, можна побачити, що на першому кроці були об’єднані спостереження 2 і 16 (тобто Данія і Швеція); ці дві країни максимально схожі одна на одну (хто б міг подумати? :-) і віддалені одна від одної на дуже малу відстань – 1,288. Ці два спостереження утворюють спільний кластер з номером 9, у той час як кластери 2 та 16 в оглядовій таблиці більше не з’являються. На наступному кроці відбувається об’єднання спостережень 1 і 3 (Бельгія і Франція), потім 19 і 23 (Болгарія і Польща) і т.д.

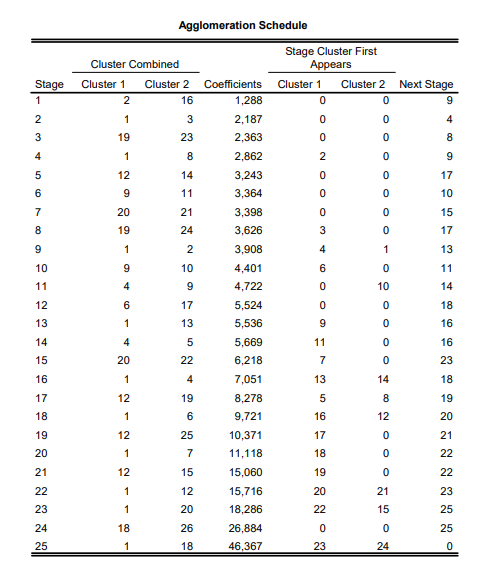


Рисунок 11.9 – Результати аналізу

Для визначення, яку кількість кластерів варто було б вважати оптимальною, вирішальне значення має показник, виведений під заголовком «коефіцієнт». Під цим коефіцієнтом розуміють відстань між двома кластерами, визначену на підставі обраної дистанційної міри. У даному прикладі це квадрат Евклідової відстані, визначений з використанням стандартизованих значень. На етапі, де ця міра відстані між двома кластерами збільшується стрибкоподібно, процес об’єднання у нові кластери необхідно зупинити, тому що в іншому випадку були б об’єднані вже кластери, що знаходяться на відносно великій відстані один від одного.

Значний стрибок коефіцієнта спостерігається на 21-му кроці. У наведеному прикладі – це стрибок з 11,118 до 15,060. Це означає, що після утворення трьох кластерів більше не треба робити ніяких наступних об’єднань, а результат із п’ятьма кластерами є оптимальним.

Оптимальним вважається число кластерів, що дорівнює різниці кількості спостережень (26) і кількості кроків, після якого коефіцієнт збільшується стрибкоподібно (21). Тобто оптимальне число кластерів дорівнює 26 – 21 = 5.

Дендрограма на рисунку 11.10 дозволяє не тільки прослідкувати процес кластеризації наочно на будь-якому етапі, але і судити про те, яка відстань між кластерами на кожному з етапів.

**Dendrogram** \* \* \* \* \* \* H I E R A R C H I C A L C L U S T E R A N A L Y S I S \* \* \* \* \* \* Dendrogram using Average Linkage (Between Groups) Rescaled Distance Cluster Combine

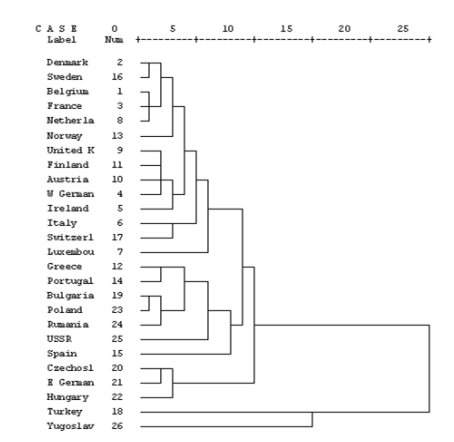


Рисунок 11.10 – Дендрограма

Числа від 0 до 25 є умовною шкалою; Число 0 відповідає найменшій відстані, а 25 – найбільшій відстані на останньому етапі. Будь-який вертикальний перетин, який можна провести на дендрограмі, покаже, скільки кластерів отримано на відповідному етапі (кількість перетинів із горизонтальними лініями), а відслідкувавши будь-яку горизонтальну лінію дендрограми вліво, можна дізнатися, які країни увійшли до якого з кластерів.

Після визначення оптимальної кількості кластерів (5) виведемо для кожного спостереження інформації про приналежність до кластера:

1. Щигликом по вимикачу ***Statistics... (Статистики)*** відкрийте діалогове вікно ***Hierarchical Cluster Analysis: Statistics (Ієрархічний кластерний аналіз: Статистики)*** і поряд з вікном послідовності злиття ***(Agglomeration schedule)*** активуйте висновок показника приналежності до кластера для кожного спостереження.
2. У розділі ***Cluster Membership (Приналежність до кластера)*** активуйте опцію ***Single solution (Одне рішення)*** і вкажіть бажану кількість кластерів 5 (рис. 11.11).

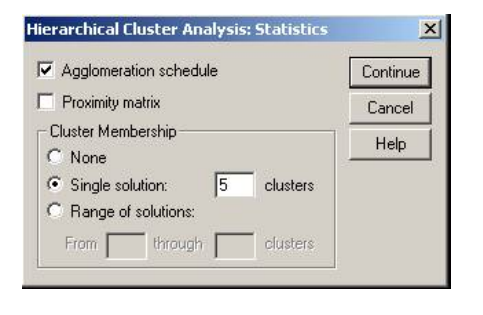


Рисунок 11.11 – Розділ Cluster Membership (Приналежність до кластера)

Інформацію про приналежність кожного спостереження до визначеного кластера можна зберегти у новій змінній. Натисніть вимикач ***Save... (Зберегти)***, активуйте опцію ***Single solution (Одне рішення)*** і для вказівки бажаної кількості кластерів уведіть 5, як показано на рисунку 11.12. Тепер, крім таблиці порядку агломерації, для кожного спостереження буде виводитися й інформація про приналежність до кластера.

З таблиці, що створилася (рисунок 11.13), можна побачити, що з п’яти кластерів до першого кластера увійшли розвинуті капіталістичні країни, до другого – капіталістичні та соціалістичні країни з аграрно орієнтованою економікою (Греція, Португалія, Іспанія, Болгарія, Польща, Румунія, Радянський Союз), до четвертої групи – соціалістичні країни із розвинутою економікою (НДР, Чехословакія, Угорщина). Окремі кластери створили дві країни – Туреччина (кластер №3) та Югославія (кластер №5). Це лише підтвердило їхній особливий шлях у світовій економіці. Югославія впроваджувала специфічний шлях розвитку за рахунок тісних зв’язків з капіталістичними та соціалістичними країнами, а Туреччина мала стиль розвитку, який можна назвати “азіатським”.

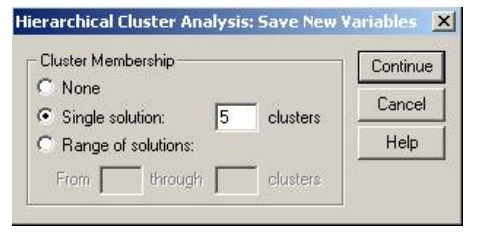


Рисунок 11.12 – Опція Single solution (Одне рішення)



Рисунок11.13 – Порядок агломерації

Розібратися в значенні кластерів допоможуть кластерні профілі; вони являють собою середні значення змінних, котрі включені в аналіз, розподілені за кластерною приналежністю. Якщо розглянути дані в редакторі даних, то можна побачити, що додалася змінна clu5\_l; ця змінна вказує на кластерну належність кожного спостереження і може бути використана для розрахунку кластерного профілю.

1. Виберіть, як показано на рисунку 11.14, у меню ***Analyze (Аналіз) Compare Means (Порівняти середні значення)*** пункт ***Means... (Середні значення).***

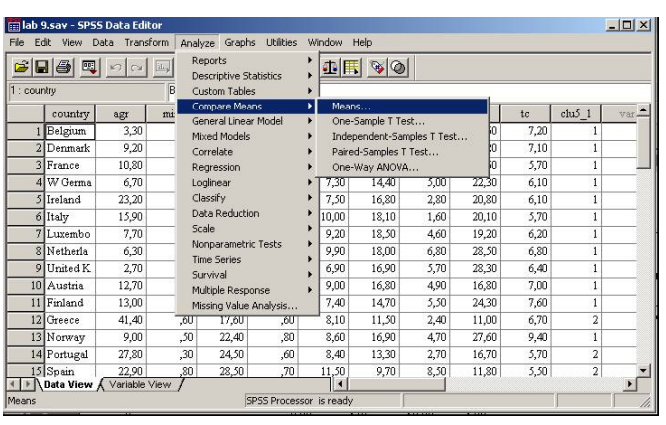


Рисунок 11.14 – Analyze (Аналіз) Compare Means (Порівняти середні значення)

1. Змінним agr – tc присвойте, як показано на рисунку 11.15, статус залежних змінних ***(Dependent List),*** а змінній clu5\_1 статус незалежної змінної ***(Independent List).***

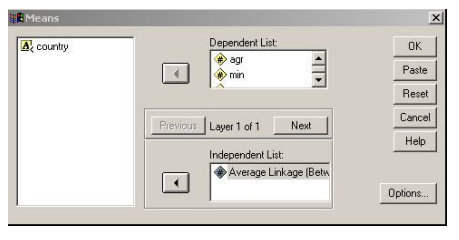
******

Рисунок 11.15 – Статус залежних змінних (Dependent List)

1. Натисніть кнопку ***Options*** та задайте (рисунок 11.16) у вікні ***Cell Statistics*** тільки видачу середніх значень ***(Mean).***

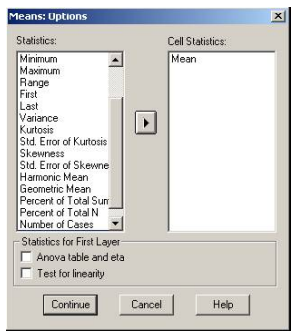


Рисунок 11.16 –Mean: Options

1. Натисніть кнопку ***Continue*** і почніть розрахунок. У результати розрахунку виводяться середні значення підсумків дев’яти змінних agr .. tc для п’яти кластерів(див. рисунок 11.17).

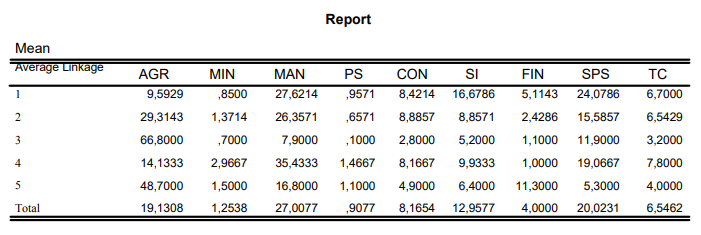


Рисунок 11.17 – Середні значення підсумків дев’яти змінних

• Agr: Відсоток зайнятих у сільському господарстві

• Min: Відсоток зайнятих у гірничодобувній галузі

• Man: Відсоток зайнятих в обробній промисловості

• PS: Відсоток зайнятих в енергетичній галузі

• Con: Відсоток зайнятих у будівництві

• SI: Відсоток зайнятих у сфері обслуговування

• Fin: Відсоток зайнятих у фінансовій сфері

• SPS: Відсоток зайнятих у соціальних службах

• TC: Відсоток зайнятих у транспорті і комунікаціях

Країни капіталістичного табору з розвинутою економікою (кластер №1) мають незначну зайнятість у аграрному секторі (agr), добувній галузі (min) і дуже велику зайнятість у фінансовій (fin) та соціальній сфері (sps). Капіталістичні та соціалістичні країни, що увійшли до кластера №2, є країнами з аграрно орієнтованою економікою та розвинутою обробною промисловістю. Країни першої і другої групи мають однакові показники зайнятості в будівництві та транспорті. Туреччина (кластер №3) має надзвичайно велику зайнятість у аграрному секторі і, водночас, надзвичайно низьку в енергетичній галузі. Зрозуміло тепер, чому програма змушена була виділити Туреччину в окремий кластер. Так, саме Югославія має великі особливості у зайнятості населення.

Питання для самоконтролю

1. Які існують групи задач кластерного аналізу?

2. У чому полягає задача кластерного аналізу?

3. Як вимірюється відстань між двома точками? Які функції відстані ви знаєте?

4. Дайте визначення математичним характеристикам кластеру: центр, радіус, середньоквадратичне відхилення, розмір кластера?

5. Які найпоширеніші способи нормування (normalization) змінних?

6. Які існують групи кластерного аналізу? Чим вони відрізняються?

7. Як відбувається визначення кількості кластерів?

8. Для яких задач обробки експериментальних даних використовуються методи ієрархічного кластерного аналізу?

9. Наведіть основні міри порівняння об'єктів між собою.

10. Що таке дендрограма?

11. Що являють собою ієрархічні агломеративні методи кластерного аналізу?

12. Що являють собою ієрархічні дивізимні методи кластерного аналізу?

13. Наведіть основні способи зв'язування об'єктів у кластери.

14. У чому полягають основні етапи ієрархічного кластерного аналізу?

15. Яким чином визначити значущу кількість кластерів?

Навчальне завдання

Розглянемо задачу класифікації на простому прикладі. Припустимо, єбаза даних про клієнтів туристичного агентства з інформацією про вік і доході за місяць. Є рекламний матеріал двох видів: більш дорогий і комфортний відпочинок і дешевший, молодіжний відпочинок. Відповідно, визначені два класи клієнтів:клас 1 іклас 2.База даних наведена в таблиці.

|  |  |  |  |
| --- | --- | --- | --- |
| **Код клієнта** | **Вік** | **Дохід** | **Клас** |
| **1** | 18 | 25 | 1 |
| **2** | 22 | 100 | 1 |
| **3** | 30 | 70 | 1 |
| **4** | 23 | 120 | 1 |
| **5** | 24 | 15 | 2 |
| **6** | 25 | 22 | 1 |
| **7** | 32 | 50 | 2 |
| **8** | 19 | 45 | 2 |
| **9** | 22 | 75 | 1 |
| **10** | 40 | 90 | 2 |

Визначити, до якого класу належить новий клієнт і який з двох видів рекламних матеріалів йому варто відсилати.

Практичне завдання

***Завдання 1.*** Провести класифікацію заданої кількості об'єктів, кожен з яких характеризується двома або більше ознаками. Як відстань між об'єктами, прийняти відстань між кластерами, застосовуючи метрику Евкліда. Розраховані дані розмістити в матриці відстаней. Результати ієрархічної класифікації об'єктів представити у вигляді дендрограмми.

*Результати опитування співробітників*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Співробітник | Показники якості службової кар'єри | | | | | | |
| ПОЦ | ВСО | ТБД | ВЕД | КЗ | БЗР | ВСБ |
| 1 | 4 | 3 | 10 | 4 | 3 | 5 | 5 |
| 2 | 6 | 3 | 9 | 4 | 7 | 2 | 1 |
| 3 | 5 | 4 | 5 | 4 | 6 | 4 | 1 |
| 4 | 5 | 3 | 6 | 3 | 3 | 7 | 4 |
| 5 | 7 | 10 | 10 | 8 | 8 | 1 | 3 |
| 6 | 8 | 4 | 7 | 5 | 8 | 2 | 3 |
| 7 | 3 | 2 | 5 | 2 | 4 | 8 | 2 |
| 8 | 3 | 2 | 3 | 3 | 5 | 10 | 3 |
| 9 | 7 | 3 | 3 | 3 | 3 | 7 | 2 |
| 10 | 2 | 2 | 3 | 1 | 2 | 10 | 1 |
| 11 | 9 | 3 | 5 | 4 | 7 | 2 | 2 |
| 12 | 5 | 5 | 2 | 3 | 2 | 5 | 3 |
| 13 | 1 | 5 | 5 | 2 | 6 | 6 | 1 |
| 14 | 7 | 5 | 2 | 7 | 7 | 1 | 4 |
| 15 | 2 | 2 | 4 | 2 | 2 | 1 | 5 |
| 16 | 3 | 3 | 1 | 3 | 5 | 1 | 2 |
| 17 | 7 | 3 | 3 | 4 | 8 | 5 | 3 |

Розглядаються результати опитування 17-ти співробітників підприємства по задоволеності показниками якості службової кар'єри. У таблиці дані відповіді на питання анкети за десятибальною шкалою (1 - мінімальний бал, 10 - максимальний). Імена змінних відповідають відповідям на наступні питання: ПОЦ - поєднання особистих цілей і цілей організації; ВСО - відчуття справедливості в оплаті праці; ТБД - територіальна близькість до будинку; ВЕД - відчуття економічного добробуту; КЗ - кар'єрне зростання; БЗР - бажання змінити роботу; ВСБ - відчуття соціального благополуччя.

Використовуючи ці дані, необхідно розділити співробітників на групи і для кожної з них виділити найбільш ефективні важелі управління. При цьому відмінності між групами повинні бути очевидними, а всередині групи респонденти повинні бути максимально схожі.

***Завдання 2*.**Дослідити структуру даної команди і виділити підгрупи для виконання групових проектів.Провести класифікацію заданої кількості об'єктів, кожен з яких характеризується відповідними ознаками. Як відстань між об'єктами, прийняти відстань між кластерами, застосовуючи метрику Евкліда. Розраховані дані розмістити в матриці відстаней. Результати ієрархічної класифікації об'єктів представити у вигляді дендрограмми

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Залеж-  ність від групових стандартів | Відпові-дальність | Трудова актив-  ність | Працез-датність | Розуміння мети | Мотивація |
| 1 | 2 | 7 | 9 | 8 | 10 | 3 |
| 2 | 4 | 2 | 8 | 8 | 8 | 1 |
| 3 | 2 | 3 | 9 | 7 | 8 | 1 |
| 4 | 7 | 3 | 5 | 6 | 4 | 0 |
| 5 | 2 | 2 | 5 | 3 | 7 | 2 |
| 6 | 4 | 3 | 5 | 5 | 5 | 2 |
| 7 | 5 | 4 | 4 | 5 | 5 | 3 |
| 8 | 6 | 1 | 4 | 4 | 7 | 0 |
| 9 | 5 | 3 | 3 | 5 | 4 | 2 |

# ГЛОСАРІЙ

*Асиметрією* називають відношення третього центрального моменту до куба середнього квадратичного відхилення: 

*Біноміальним* називають розподіл випадкової величини, ймовірності якої обчислено за допомогою формули Бернуллі.

Спостережувані значення *xi* називаються *варіантами,* а послідовність варіант у зростаючому порядку – *варіаційним рядом.*

*Вибірковою дисперсією Дв* називається середнє арифметичне квадратів відхилення спостережуваних величин від їх середнього

*Вибірковою сукупністю (вибіркою)* називають сукупність випадково відібраних об'єктів.

*Відсутність післядії* – незалежність появ того або іншого числа подій у непересічні проміжки часу.

*Випадкова величина безперервна*, якщо її функція розподілу безперервна, кусочно-диференційована з безперервною першою похідною.

*Випадковою величиною* називають величину, що в результаті випробування прийме одне й тільки одне можливе значення, наперед невідоме й залежне від випадкових причин, які заздалегідь не можуть бути враховані. Випадкова величина буває:

а) *дискретною* – випадкова величина, що може приймати скінченну або нескінченну рахункову множину ізольованих значень із певними ймовірностями;

б) *безперервною* – випадкова величина, що може приймати всі значення зі скінченного або нескінченного інтервалу.

*Випадкове явище* – це таке явище, що при кількаразовому відтворенні того самого випробування протікає щораз трохи по-іншому.

*Виробляючою функцією* , називається багаточлен ступеня *n*, коефіцієнти при невідомих у якому дорівнюють ймовірностям біноміального розподілу.

*Гістограмою частот* називають східчасту фігуру, основою кожного прямокутника якої служать часткові інтервали *h*, а висоти – *ni/h*.

*Гістограма частостей* – це східчаста фігура, основою кожного прямокутника якої служать часткові інтервали *h*, а висоти – *wi/h = ni/(nh)*.

*Генеральною сукупністю* називають ту сукупність, з якої робиться вибірка. *Об'єм сукупності* – число об'єктів.

*Дисперсією* випадкової величини *Х* називається математичне сподівання квадрата відхилення випадкової величини від свого математичного сподівання.

*Добутком (перетинанням)* подій *А* та *В* називається подія *С,* , якщо вона відбувається в тому випадку, коли в результаті випробування відбуваються обидві події одночасно.

*Довірчим* називають *інтервал* ( - *δ*;  + *δ*), що покриває заданий параметр із заданою надійністю *γ*.

*Ексцесом* випадкової величини *Х* називається величина *Ех*, що обчислюється за формулою:

*Емпіричною функцією розподілу* називають функцію *F\*(x)*, що визначає для кожного *х* відносну частоту події *Х < x*, тобто *F\*(x) = nx/n*, де *nx* – число варіант, менших ніж *х* (накопичена частість), *n* – об'єм вибірки.

*Законом розподілу* дискретної двовимірної випадкової величини називають перелік всіх можливих значень цієї величини *(хi,yj)* і їхніх ймовірностей *р(хi,yi),* де *i* = 1, …, *n*, *j* = 1, …, *m*...

*Запереченням* події *А* називається подія *F,* якщо в результаті випробування подія *F* відбувається тоді й тільки тоді, коли *А* не відбувається.

*Інтенсивністю потокуλ* називають середнє число подій за одиницю часу.

*Інтервальною* називають *оцінку*, що визначається двома числами – кінцями інтервалу.

*Імовірність* – це кількісна міра ступеня невизначеності в настанні події.

*Імовірністю* називається відношення довжини (площі, об'єму) сприятливої області до довжини (площі, об'єму) області можливих значень.

*Коефіцієнтом кореляції* випадкових величин *Х* і *Y* називають відношення кореляційного моменту до добутку середніх квадратичних відхилень.

Дві випадкові величини називають *корельованими*, якщо їхній коефіцієнт кореляції не дорівнює нулю.

*Кореляційним моментом* або *коваріацією* випадкових величин *Х* и *Y* називають математичне сподівання добутку відхилень цих величин.

*Критерій Вілкоксона*підтверджує, що вибірки отримані з однорідних генеральних сукупностей, і, зокрема, мають рівні середні й медіани.

*Критерієм згоди* називають критерій перевірки гіпотези про передбачуваний закон невідомого розподілу.

*Математичним сподіванням* називається зважене за ймовірностями середнє значення випадкової величини *Х.*

*Математичним сподіванням двовимірної випадкової величини(X,Y)* називається точка *(М(X), М(Y)),* яка визначає центр двовимірного розподілу.

*Медіаною* називається значення випадкової величини *Х*, для якого виконується:



*Механічний відбір* – відбір, при якому генеральна сукупність ділиться на стільки груп, скільки необхідно об'єктів. (Наприклад, якщо треба 5%, то відбирають кожну двадцяту деталь.)

*Надійністю (довірчою ймовірністю)* оцінки *θ* по  називають імовірність γ, з якою виконується нерівність |*θ* - | <*δ.*

Потікє *найпростішим (пуассоновим)* якщо виконуються властивості: стаціонарність, відсутність наслідку.

*Нормальним* називають розподіл *N*(*a*, *σ2)* випадкової величини з параметрами *а*, *σ* (*σ*> 0), що описується такою щільністю:

*Незалежними* називаються події *А* та *В* тоді й тільки тоді, коли виконується наступна умова: *Р*(*А*/*В*) = *Р*(*А*), тобто подія *А* не залежить від того, відбулася *В* чи ні.

*Ординарність* – за нескінченно малий проміжок часу може відбутися не більше однієї події.

*Переставлення* – це комбінації, що складаються з тих самих *n* різних елементів, що відрізняються порядком їхнього розташування.

*Подія* – усякий факт, що в результаті випробування може відбутися або не відбутися.

події бувають трьох видів:

1) *достовірна* – подія, що обов'язково відбудеться при завданні деякого певного комплексу умов (вода при нормальному тиску й температурі 200 С рідка), позначається – *U*;

2) *неможлива* – подія, що напевне не відбудеться (вода при нормальному тиску й температурі -200 С тверда), позначається – *V*;

3) *випадкова* – подія, що або відбудеться, або не відбудеться (випадіння орла при підкиданні монети).

події розділяють на:

а) *спільні* – ті події, які в результаті випробування можуть відбутися одночасно;

б) *неспільні* – дані події не можуть відбутися одночасно в одному випробуванні.

*Полігоном* частот називають ламану, відрізки якої з'єднують точки з координатами (*x1, n1*), … , (*xk, nk*)... *Полігоном частостей* називають ламану, відрізки якої з'єднують точки з координатами (*x1, w1*), … , (*xk, wk*)...

*Показниковим законом надійності* називають функцію:

*Показниковим (експоненціальним)* називається розподіл випадкової величини *Х*, що описується наступною щільністю розподілу:

 де *λ* - додатна константа.

*Поле подій -* повна множину подій, які можуть відбутися в результаті випробування.

*Потоком подій* називають послідовність подій, які настають у випадкові моменти часу.

*Початковим моментом порядку k*, *s* системи *(X,Y)* називається математичне сподівання добутку *Хk* на *Ys*

*Початковим моментом k-го порядку* називають математичне сподівання *k-ого* ступеня випадкової величини:

*Простою* називають гіпотезу, що містить тільки одне припущення

Розподіл називають *рівномірним*, якщо на інтервалі, якому належать всі можливі значення випадкової величини, щільність розподілу постійна, а поза цим інтервалом дорівнює нулю.

*Різницею* подій *А* а *В* називається подія *Е,* якщо в результаті випробування подія *Е* відбувається тоді й тільки тоді, коли *А* відбувається, а *В*– не відбувається.

*Розмахом вибірки* називається різниця між її максимальним і мінімальним елементами: *R = xmax – xmin*.

*Розміщення* – це комбінації, складені з *n* різних елементів по *m* елементів, які відрізняються або складом елементів, або їхнім порядком.

*Серійний відбір* – відбір, при якому з генеральної сукупності об'єкти відбираються серіями.

*Середнім квадратичним* відхиленням випадкової величини називають квадратний корінь із дисперсії цієї випадкової величини.

*Складною або складовою* називають гіпотезу, що складається з кінцевого або нескінченного числа простих гіпотез

*Сполучення* – це комбінації, складені з n елементів по m елементів, які відрізняються хоча б одним елементом.

*Спостережуваним значенням Кспос* називають значення критерію, обчислене за вибіркою.

*Стандартним (нормованим) нормальним розподілом* називається розподіл *N*(0;1),



*Статистичним критерієм* або просто *критерієм* називають випадкову величину *К*, що служить для перевірки гіпотези *Н0*

*Статистичною оцінкою* невідомого параметра теоретичного розподілу називають функцію від спостережуваних випадкових величин.

*Статистичним розподілом* вибірки називається перелік варіант та відповідних їм частот (або частостей).

*Стаціонарність* – імовірність появи *k* подій на будь-якому проміжку часу залежить тільки від числа *k* і від тривалості *Δ t* проміжку й не залежить від початку його відліку.

*Сумою (об'єднанням)* подій*А* та *В* називається подія *С*, якщо вона відбувається в тому випадку, коли в результаті випробування відбувається хоча б одна з *А* або *В.*

*Теорія ймовірностей* – математична наука, що вивчає закономірності масових випадкових явищ.

*Типовий відбір* – відбір, при якому об'єкти відбирають із кожної типової частини генеральної сукупності.

*Точковою* називають *оцінку*, що визначається одним числом.

*Умовною ймовірністю Р(А/В)* будемо називати ймовірність події *А*, обчислену в припущенні, що подія *В* вже відбулася.

*Умовним математичним сподіванням* дискретної випадкової величини *Y* при *Х = х* називають суму добутків можливих значень *Y* на їхні умовні ймовірності.

*Умовне математичне сподіванняM(Y/x)* є функцією від х *M(Y/x) = f(x)*, що має назву функція регресії*Y* на *Х.*

*Умовним розподілом складової Х* при *Y = yj* називають сукупність умовних ймовірностей *Р(х1/уj), …, Р(хn/уj),* обчислених у припущенні, що подія *Y = yj* вже настала.

*Умовною щільністю розподілуϕ(х/у)* складової *Х* при даному значенні *Y = у* називається відношення щільності спільного розподілу *f(x,y)* до щільності компоненти *f2(y)*:*ϕ(х/у) = f(x,y)/f2(y).*

*Функцією розподілу двовимірної* випадкової величини називають функцію F(x,y), що визначає для кожної пари (x,y) імовірність того, що Х < x і при цьому y < Y, тобто F(x,y) = P(X < x, Y < y).

*Функцією розподілу ймовірностей* або інтегральною функцією випадкової величини *Х* називається функція *F(x)*, що визначає ймовірність того, що випадкова величина *Х* прийме значення, яке менше аргументу *х.*

*Центральним моментом* порядку *k*, *s* системи *(X,Y)* називається математичне сподівання добутку *(X - M(x))k* на *(Y - M(y))s*.

*Центральним моментом k-го порядку* називають математичне сподівання *k-ої* ступеня відхилення: 

Числа спостережень ni називаються *частотами,* а відношення - *відносними частотами (частостями).*Частоти й частості є *вагами* варіант.

*Числом ступенів волі f* називається різниця між об'ємом вибірки *n* і числом накладених на неї зв'язків, де під зв'язком розуміють будь-який оцінюваний параметр.

*Щільністю спільного розподілу ймовірностейf(x,y)* двовимірної безперервної випадкової величини *(X,Y)* називають другу змішану частинну похідну від функції розподілу:.

*Щільністю розподілу* ймовірностей безперервної випадкової величини *Х* називають функцію *р(х)* = *F′(x)*.

*Щільність розподілу однієї зі складових* дорівнює невласному інтегралу з нескінченною межею від щільності розподілу системи за іншою складовою, тобто:



# РЕКОМЕНДОВАНА ЛІТЕРАТУРА

**Основна:**

1. Бахрушин В. Є. Методи аналізу даних : навч. посіб. Запоріжжя : КПУ, 2011. 268 с.
2. Клебанова Т. С., Гур'янова Л. С., Чаговець Л. О. Бізнес-аналітика багатовимірних процесів : навч. посіб. Харків : ХНЕУ ім. С. Кузнеця, 2018. 272 с.
3. Бідюк П. І., Ткач Б. П. Математична статистика. Київ : Персонал, 2017. 255 с.
4. Яровий А. Т., Страхов Є. М. Багатовимірний статистичний аналіз : навч.-метод. посіб. Одеса : Астропринт, 2015. 132 с.
5. Жлутенко В. І., Наконечний С. І., Савіна С. С. Теорія ймовірностей та математична статистика : навч.-метод. посіб. Ч. II. Київ : КНЕУ, 2007. 336 с.
6. Arthur Griffith. SPSS For Dummies. 2nd Edition. Wiley Publishing, 2010. 360 p.

**Додаткова:**

1. Огірко О. І., Галайко Н. В. Теорія ймовірностей та математична статистика : навч. посіб. Львів : ЛьвДУВС, 2017. 292 с.
2. Веригіна І. В., Островська О. В., Сугакова О. В. Теорія ймовірностей та математична статистика. Лекції і практикум : навч. посіб. Київ : КПІ ім. Сікорського, 2022. 254 с.
3. Найко Д. А. Шевчук О. Ф. Теорія ймовірностей та математична статистика : навч. посіб. Вінниця : ВНАУ, 2020. 382 с.
4. Зайцев Є.П. Теорія ймовірностей і математична статистика. Базовий курс з індивідуальними завданнями і розв’язком типових варіантів : навч. посіб. Київ : Алерта, 2017. 440 с.
5. Кособуцький П. С., Лобур М. В. Статистичне моделювання. Львів : Львівська політехніка, 2013. 328 с.
6. Ткач Є. І., Сторожук В. П. Загальна теорія статистики : підручник. Київ : ЦУЛ, 2012. 442 с.
7. Andy Field. Discovering Statistics Using SPSS. 2nd edition. London, 2005. 779 р.
8. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer, New York, 2013. 426 p.
9. Marco Taboga. Lectures on Probability Theory and Mathematical Statistics. CreateSpace Independent Publishing Platform, 2017. 670 р.

# ВИКОРИСТАНА ЛІТЕРАТУРА

1. Зайцев Є. П. Теорія ймовірностей і математична статистика. Базовий курс з індивідуальними завданнями і розв’язком типових варіантів : навч. посіб. Київ : Алерта, 2017. 440 с.
2. Огірко О. І., Галайко Н. В. Теорія ймовірностей та математична статистика : навч. посіб. Львів : ЛьвДУВС, 2017. 292 с.
3. Веригіна І. В., Островська О. В., Сугакова О. В. Теорія ймовірностей та математична статистика. Лекції і практикум : навч. посіб. Київ : КПІ ім. Сікорського, 2022. 254 с.
4. Найко Д. А. Шевчук О. Ф. Теорія ймовірностей та математична статистика : навч. посіб. Вінниця : ВНАУ, 2020. 382 с.
5. Бахрушин В. Є. Методи аналізу даних : навч. посіб. Запоріжжя : КПУ, 2011. 268 с.

Навчальне видання

(*українською мовою*)

Міхайлуца Олена Миколаївна

Пожуєв Андрій Володимирович

Емпіричні методи інженерії програмного забезпечення

Навчальний посібник

для здобувачів ступеня вищої освіти бакалавра

спеціальності 121 «Інженерія програмного забезпечення»  
освітньо-професійної програми

«Програмне забезпечення систем»

Рецензент *С. М. Гребенюк*

Відповідальний за випуск *Т. В. Критська*

Коректор *О. М. Міхайлуца*