**О.В. Бєляєв**

**ФАКТОРНИЙ АНАЛІЗ**

**ВИПАДКОВИХ ВЕЛИЧИН І СТАТИСТИЧНИХ ДАНИХ**

Зміст

[**Введення** 3](#_Toc41517802)

[**1.** **Випадкові величини і статистичні дані.** 5](#_Toc41517803)

[**§1. Випадкові величини та вибірки даних.** 5](#_Toc41517804)

[**§ 2. Математичне сподівання і дисперсія дискретних випадкових величин.** 6](#_Toc41517805)

[**§ 3. Властивості і дискретної випадкової величини.** 9](#_Toc41517806)

[**§ 4. Закон великих чисел.** 10](#_Toc41517807)

[**§ 5. Неперервні випадкові величини.** 12](#_Toc41517808)

[**§ 6. Властивості і для наборів даних.** 15](#_Toc41517809)

[**§ 7. Взаємозумовленість випадкових величин і наборів даних експерименту.** 16](#_Toc41517810)

[**2.** **Лінійна залежність випадкових величин та статистичних даних.** 18](#_Toc41517811)

[**§1. Лінійна кореляція між двома випадковими величинами.** 18](#_Toc41517812)

[**§2. Лінійна кореляція між двома наборами даних.** 19](#_Toc41517813)

[**§3. Лінійна залежність між двома наборами даних.** 20](#_Toc41517814)

[**§4. Лінійна залежність між багатьма випадковими величинами.** 20](#_Toc41517815)

[**3.** **Нелінійна залежність випадкових величин та статистичних даних.** 25](#_Toc41517816)

[**§1. Коефіцієнт загальної кореляції між двома наборами даних.** 26](#_Toc41517817)

[**§2. Зсув даних у часі.** 28](#_Toc41517818)

[**§3. Компоненти зв’язності множини значень даних.** 29](#_Toc41517819)

[**§4. Залежність у межах зв’язної компоненти множини .** 30](#_Toc41517820)

[**§5. Пошук нелінійного тренду даних.** 32](#_Toc41517821)

[**§6. Аналіз функції збурення тренду*.*** 38](#_Toc41517822)

[**§7. Використання тренду*.*** 40](#_Toc41517823)

[**4.** **Многофакторний аналіз випадкових величин та статистичних даних.** 42](#_Toc41517824)

[**§1. Особливості нелінійного многофакторного аналізу.** 42](#_Toc41517825)

[**§2. Розмірність і остов компоненти зв’язності у -вимірному просторі.** 44](#_Toc41517826)

[**§3. Базові змінні. Тренд.** 45](#_Toc41517827)

[**Література.** 47](#_Toc41517828)

# **Введення**

Предметом теорії ймовірностей і математичної статистика є аналіз випадкових явищ, тобто таких, які при однакових умовах дають різні, притому непередбачувані, результати, але в той же час при багатократному їх повторенні виявляють певні закономірності.

В якості прикладу розглянемо кидання монетки. Добре відомо, що практично неможливо передбачити якою стороною впаде монетка при однократному киданні, але ось результати експериментів, проведених Ж. Бюффоном і К. Пірсоном.

|  |  |  |  |
| --- | --- | --- | --- |
| Експериментатор | Число підкидань | Число «гербів» | Відносна частота |
| Ж. Бюффон | 4040 | 2048 | 0,5069 |
| К. Пірсон | 12000 | 6019 | 0,5016 |
| Ж. Бюффон | 24000 | 12012 | 0,5005 |

Ми бачимо, що зі збільшенням кількості підкидань відношення кількості «гербів» до загальної кількості підкидань наближається до ½.

С однієї сторони цей результат має характер закону, оскільки має місце повторюваність результату, але з другої сторони, очевидно, що не можливо гарантувати ніяку точно оцінку як абсолютно достовірний факт.

Трудність у точному формулюванні аксіом теорії ймовірностей виявилась настільки важкою, що вона була здолана лише у 20 сторіччі в працях Бернштейна, Мізеса, Бореля і Колмогорова, хоч перші задачі теорії ймовірностей Кардано, Пачоли і Тарталья розв’язували вже в 15-16 століттях.

Отже теорія ймовірностей вже має вигляд строгої математичної дисципліні з аксіомами, визначеннями і теоремами. Її твердження відповідають практичному досвіду. Так, приведений приклад з монеткою знаходиться у повній відповідності з Законом Великих Чисел, але в границях теорії знаходяться і інші значно серйозніші результати.

Прикладна частина теорії ймовірностей називається математичною статистикою і має метою отримання з практичних експериментів характеристик випадкових величин, що дозволяють прогнозувати ймовірнісні процеси.

Одною з важливих її складових є факторний аналіз, сутність якого можна сформулювати, як виявлення статистичних залежностей і їх явного подання для існуючих у розгляді випадкових величин.

Більш точно проблему сформулюємо таким чином:

***Постановка задачі в теоретичному аспекті.***

*Маємо випадкових величин*

*Необхідно з’ясувати, чи існує ймовірнісна лінійна або нелінійна залежність між ними, а якщо існує, то знайти її у явному вигляді.*

***Постановка задачі в практичному аспекті.***

*Маємо набори вибірок*

*Необхідно з’ясувати, чи існує статистична або системна лінійна або нелінійна залежність між відповідними їх елементами, а якщо існує, то знайти її у явному вигляді.*

# **Випадкові величини і статистичні дані.**

**§1. Випадкові величини та вибірки даних.**

***Визначення 1.1.*** *Випадкова величина* *вважається заданою, якщо кожному елементарному випадку повного простору подій приписане значення причому для подій визначена ймовірність*

*При цьому виконані наступні умови (аксіоми):*

1. *(натуральні числа).*

***Зауваження 1.1.*** *Нехай – випадкові величини, тоді – також є випадковими величинами, заданими на тім же просторі подій, що й*

***Визначення 1.2****.* *Умовною ймовірністю події при умові події В називається ймовірність події якщо вважати В повним простором подій. Відповідно цьому*

***Визначення 1.3.*** *Випадкові події називаються незалежними, якщо*

***Зауваження 1.2 (до визначення 1.3)*** *Тобто ймовірність події така ж сама, що при події що при відсутності події*

***Твердження 1.1.*** *Випадкові події називаються незалежними, якщо*

***Доведення.*** Необхідну формулу отримуємо, співставляючи визначення 1.2 і 1.3.

***Визначення 1.4.*** *Випадкові величини називаються незалежними, якщо*

***Визначення 1.2.*** *Сукупність значень випадкової величини отримана в результаті фізичного експерименту називається вибіркою даних випадкової величини .*

Основними модельними прикладами випадкових величин та їх вибіркових даних будуть наступні.

1. *Підкидання монети.* *0 – «герб», 1 – «решітка».*
2. *Підкидання грального кубику.*
3. *Обертання рулетки з неперервною шкалою.*

Приклади потребують наступного пояснення.

***Зауваження 1.3.*** *Довільний набір даних, навіть, якщо він здається випадковим, може таким не бути і лише виглядати таким, але мати не очевидну закономірність. Наведені вище приклади на практиці вважаються насправді випадковими саме тому, що використовуються, наприклад, у визначенні першого удару по м’ячу футбольного матчу , а також у азартних іграх. Якщо б результати цих дій не були випадковими, їх би не використовували.*

*Інші приклади потребують серйозного аналізу, який і є метою книги.*

***Зауваження 1.4.*** *Фундаментальна проблема теорії ймовірностей і математичної статистики є відповідність аксіоматичної теорії ймовірностей практичним результатам, отриманими в результаті експериментів з наведеними вище випадковими величинами.*

*Задовільне розв’язання цієї проблеми дає Закон Великих Чисел, що з одної сторони доведений теоретично на основі аксіом теорії ймовірностей (§4), а з другої – результатами експериментів Пірсона і Бюффона (Введення).*

### **§ 2. Математичне сподівання і дисперсія дискретних випадкових величин.**

Уявимо собі, що ми багатократно робимо експеримент, в результаті якого визначається значення деякої випадкової величини. Бажаючи її коротко характеризувати, природно обчислити середнє арифметичне значення всіх результатів, а також середнє відхилення результатів від отриманого середнього значення.

***Визначення 2.1.*** *Основними характеристиками довільної вибірки є її середнє значення:*

*і середнє квадратичне відхилення (від середнього значення):*

*або, що більш зручно, його квадрат, тобто*

*Ці характеристики дають локацію основної маси значень випадкової величини, що експериментально вивчається.*

Вказаним статистичним характеристикам відповідають однозначно визначені величини, якщо випадкова величина коректно задана, які називаються математичним сподіванням і дисперсією.

Перш ніж дати визначення цих величин, пояснимо, яким чином їх можливо обґрунтувати. Нехай випадкова величина приймає значення з ймовірностями Повторимо раз експеримент, що генерує випадкову величину . Ми отримаємо послідовність вигляду де числа належать до набору

Середнє арифметичне всіх значень дорівнює

де числа показують, скільки разів випадкова величина отримала значення При цьому, природно, Середнє значення можна записати у вигляді:

Звернемо увагу, що є частота, з якою випадкова величина приймає значення Досвід показує, що тому

Таким чином, може в результаті різних серій експериментів приймати, хоч і близькі, але різні значення, в той час як величина є постійною і однозначно визначається для випадкової величини. Отже, як раз вона і є характеристикою випадкової величини, а середнє значення - її експериментальним наближенням.

***Визначення* *2.2.*** *Математичним сподіванням дискретної* випадкової величини  *називається число*

*де – значення випадкової величини*  *а – відповідні ймовірності.*

***Приклад 2.1****.* При киданні грального кубику:

Зауважимо, що результат лежить на середині відрізку [1,6] (з є серединою [0,6]).

Наступна характеристика оцінює середній розкид даних ряду експериментів, тобто є середнім від Очевидно, що середні необхідно брати одного знаку, щоб відхилення в різні сторони від середнього значення не компенсували одне одного. Середнє значення є однією з можливих характеристик, але сума модулів вкрай незручна з аналітичної точки зору. Оптимальною характеристикою, яка вільна від цього недоліку є середнє від

Повторюючи попередні міркування, можна записати, що середнє значення квадратів відхилення від математичного сподівання дорівнює

причому останнє наближене подання є однозначно визначеною характеристикою випадкової величини.

***Визначення 2.3.*** *Дисперсією дискретної випадкової величини називається число*

*де – значення випадкової величини* *, а відповідні ймовірності.*

***Приклад* *2.2.*** В експерименті 2 (кидання грального кубику):

### **§ 3. Властивості і дискретної випадкової величини.**

***Теорема 3.1.*** *Нехай* випадкові величини*, – константи. Тоді*

1. *якщо то*
2. *якщо то*
3. *якщо випадкові величини*  *незалежні, то .*

***Доведення.*** *1.* Випадкова величина, яка є константою з ймовірністю 1 дорівнює своєму значенню, тому 

2. Нехай тоді тому що

3.

4. Якщо то

5.

6. Введемо випадкові величини:

Тоді для цих випадкових величин вірно

7. Випадкові величини незалежні, значить,

***Теорема 3.2.*** *Нехай випадкові величини, – константи. Тоді*

1. *якщо випадкові величини*  *незалежні, то*

***Доведення.*** 1. Якщо випадкова величина приймає єдине значення, то відхилення від середнього значення дорівнює нулю з ймовірністю 1, тому

2.

3.

4. Випадкові величини незалежні, тоді

тому що

### **§ 4. Закон великих чисел.**

***Лема 8.1 (Маркова).*** *Якщо дискретна випадкова величина*  *така, що існує то для всіх*

***Доведення.*** Виберемо довільне значення . Занумеруємо значення випадкової величини і будемо вважати, що ці значення вона приймає з ймовірностями такими, що перші значень були б не меншими. Тоді

Значить,

***Лема 8.2 (нерівність Чебышева).*** *Якщо дискретна* випадкова величина  *має обмежене математичне сподівання і дисперсію, то для любого додатного числа справедлива нерівність*

***Доведення****.* Нехай тоді в силу леми Маркова

або

***Теорема 8.1 (закон великих чисел у формі Чебышева).*** *Нехай випадкові величини попарно незалежні і мають обмежені константою дисперсії. Тоді для любого*

***Доведення.***В силу умов теореми

отже

Переходячи до границі в нерівності, отримуємо необхідне твердження.

Досвід показує, що чим більше проведено випробувань, тим менше відмінність середнього значення випадкової величини від математичного сподівання. І саме це стверджує доведена теорема. Навіть, якщо взяти вельми мале число ймовірність того, що математичне сподівання буде відрізнятись від середнього значення менш, ніж на буде як завгодно близька до 1, хоч і при достатньо великій кількості випробувань.

***Теорема 8.2 (Бернуллі).*** *Нехай – число повторень події в незалежних випробуваннях Бернуллі, – ймовірність* *події в одному випробуванні. Тоді для любого*

***Доведення***. Як було доведено в схемі Бернуллі значить,

Нерівність Чебышева в цьому випадку приймає вигляд:

Перехід до границі в цій нерівності завершує доведення.

***Зауваження 4.1.*** *Теорема Бернуллі підтверджує практичне припущення, що частота події у практичному відповідає теоретичному значенню ймовірності цієї події.*

### **§ 5. Неперервні випадкові величини.**

З життєвого досвіду ми знаємо випадкові величини, що не мають дискретного характеру. Так, наприклад, природні явища (випадковий характер яких важко заперечити) характеризуються температурою, тиском, кількістю опадів. Всі названі випадкові величини мають безперервну шкалу зміни.

В §1 в якості неперервної випадкової величини була приведена гральна рулетка, що в кінці обертання дає значення у проміжку .

***Визначення 5.1.*** *Функцією розподілу випадкової величини*  *називається функція*

***Приклад 5.1.*** Функція розподілу випадкової величини, що визначає рулетка, дорівнює:

її графік має вигляд:



1

***Твердження 5.1.***

***Твердження 5.2.***

***Твердження 5.3.*** *Функція розподілу дискретної випадкової величини* має *вигляд сходинок. Точки розриву графіку - значення випадкової величини, а висота сходинки в точці розриву дорівнює ймовірності відповідного значення.*

***Приклад* 9.2.** Випадкова величина, що задається гральним кубиком (приклад 2.1) має графік функції розподілу наступного вигляду:

1/6

1

1

***Визначення* 5.2*.*** *Щільністю розподілу випадкової величини*  *називається функція:*

*якщо ця похідна існує.*

Графік щільності розподілу випадкової величини має просте статистичне подання. Припустимо, що послідовно проводиться експеримент, в результаті якого визначається випадкова величина, і при цьому над віссю абсцис зображується невелике коло над координатою, що дорівнює значенню випадкової величини. Якщо випадкова величина двічі приймає близькі значення, то кола зображуються один над іншим, не перетинаючись. Після значної кількості випробувань над віссю абсцис буде зображена гірка, що складається з кружечків. Так ось контур отриманої гірки буде повторювати графік щільності розподілу, причому, чим більша кількість випробувань буде зроблена, тим точніше буде відповідність.

Пояснення цього факту дуже просте. На кожній з ділянок осі абсцис кількість кружечків, а значить і їх площа, будуть приблизно пропорційні ймовірності попадання випадкової величини в відповідну ділянку. Ця ж сама властивість виконана, до того ж не приблизно, а точно, для графіка щільності розподілу: площа під графіком на різних відрізках області визначення пропорційна ймовірності попадання значення випадкової величини до відповідних відрізки, оскільки

Щоб визначити математичне сподівання неперервної випадкової величини розглянемо наближення даної випадкової величини дискретними.









Дискретні випадкові величини дають приблизно ті ж ймовірності, що й початкова, але для них можливо обчислити математичне сподівання:

Якщо функція розподілу диференційовна, то можливе і інше подання:

Переходячи до границі в цих інтегральних сумах, отримуємо

***Визначення* 5.3.** *Математичним сподіванням випадкової величини* яка *має функцію розподілу називається величина:*

*Якщо функція розподілу* *диференційовна, то*

Аналогічним чином визначається дисперсія; при цьому використовується властивість 3) теореми 3.2.

***Визначення 9.5.*** *Дисперсією випадкової величини*  *з функцією розподілу називається величина*

*Якщо функція розподілу* *диференційовна, то*

### **§ 6. Властивості і для наборів даних.**

Перевіримо, чи виконуються для величин і довільних числових наборів ті ж властивості, що й для випадкових величин. При цьому ми не зважаємо на те, яким чином отриманий числовий набір.

***Теорема 6.1.*** *Нехай – довільні набори* даних*, – константи. Тоді*

1. *якщо то*
2. *якщо то*

***Доведення****.* Для довільних наборів всі твердження теореми, аналогічні, властивостям математичного сподівання виконуються за виключенням останнього і це досить просто доводиться. Останнє твердження для математичного сподівання виходить з умови незалежності випадкових величин, а ця умова не має природнього аналогу для наборів чисел.

***Теорема 6.2.*** *Нехай – довільні набори* даних*, – константа. Тоді*

***Доведення*** аналогічне теоремі 6.1 і з тим же винятком для незалежних випадкових величин.

***Зауваження 6.1.*** *Отже властивості основних характеристик випадкових величин та аналогічні їм властивості наборів даних співпадають, що спрощує перенесення і інших властивостей вказаного паралелізму понять теорії ймовірностей і математичної статистики.*

### **§ 7. Взаємозумовленість випадкових величин і наборів даних експерименту.**

Отже маємо основи теорії випадкових величин з набором понять і відповідні ним поняття при практичній реалізації теорії.

Перехід від випадкових величин до наборів даних не представляє ніяких труднощів. Просто необхідно фізично реалізувати в результаті експерименту випадкові величини, що знаходяться в колі дослідження. Більш того, знаючи властивості випадкових величин ми можемо з необхідною точністю спрогнозувати характеристики результатів експерименту. Зауважимо, що точність прогнозу буде тим більша, чим більше число випробувань.

Перехід в зворотному напрямку значно складніший і потребує розв’язання досить неочевидних проблем.

Припустимо, що ми маємо вибірку даних потрібної кількості.

Основна проблема, яка виникає, це проблема визначення, чи є набір даних результатом дії випадкової величини. Вважаємо, що основними варіантами відповіді можуть бути: 1) – пряма залежність, яку далі будемо називати ***трендом***; 2) – випадкова величина, 3) – сума тренду і випадкової величини.

***Зауваження 7.1****. Визначення характеру даних може витікати із самої суті експерименту. Наприклад, варіант з трендом частіше може виникати, коли, наприклад, змінна відповідає шкалі часу (дні, місяці, роки). При опитуванні випадкових людей, очевидно, найчастіше ми отримаємо випадкову величину, але можливі варіанти, коли опитування не коректне з якихось причин.*

*В невизначеній ситуації можливо використовувати існуючі методи перевірки набору чисел як випадкових [1],[2].*

*Якщо набір виявляється не випадковим, то для пошуку тренду можна використовувати саме порушені критерії випадковості. Також можливе використання стандартних методів лінійної або нелінійної регресії [3].*

*Якщо набір випадковий, то в цьому випадку ділимо область визначення випадкової величини на рівні відрізки , довжина яких вибирається в відповідності від кількості чисел набору. Далі, очевидно, функція щільності випадкової величини є*

*тут означає кількість елементів множини, – загальна кількість елементів набору даних.*

*Таким чином, в наступних параграфах вважаємо, що у тих випадках, де набори даних випадкові, ми також маємо у розпорядженні основні параметри випадкової величини.*

# **Лінійна залежність випадкових величин та статистичних даних.**

### **§1. Лінійна кореляція між двома випадковими величинами.**

Розглянемо вираз

З властивості 6) математичного сподівання витікає, що тому що

Отже, маємо

***Твердження 1.1.***

***Твердження 1.2.***

***Визначення 1.1.*** *Коваріацією називається величина*

З теореми 3.1 витікає, що для незалежних

Доведемо, що тоді і тільки тоді, коли між випадковими величинами існує лінійна за залежність.

Спочатку сформулюємо очевидну властивість

яка дозволяє перейти від значення математичного сподівання до значення самої випадкової величини. Нехай

Дискримінант дорівнює нулю, отже розв’язання для є, а значить, випадкова величина тотожно дорівнює нулю.

Таким чином, ми бачимо, що величина характеризує близькість до лінійної залежності між двома випадковими величинами.

***Визначення* *1.2.*** *Величина де*

*називається коефіцієнтом лінійної кореляції для випадкових величин .*

***Зауваження* *1.1.*** *Якщо коефіцієнт лінійної кореляції дорівнює нулю, то це означає, що лінійна залежність відсутня, але при цьому може мати місто нелінійна* *залежність між випадковими величинами.*

***Приклад 1.2.*** Нехай – випадкова величина, що приймає значення з ймовірністю 1/3; визначимо також ще дві випадкові величини

Отже, лінійна кореляція між випадковими величинами дорівнює нулю, але між ними існує нелінійна залежність:

### **§2. Лінійна кореляція між двома наборами даних.**

Поняття лінійної кореляції природним способом визначаємо для наборів даних.

***Визначення* *2.2.*** *Величина де*

*називається коефіцієнтом лінійної кореляції для числових даних .*

***Зауваження* *2.1.*** *Для числових даних коефіцієнт лінійної кореляції має геометричну інтерпретацію, як скалярний добуток векторів з координатами і відповідно.*

*Таким чином, рівність означає, що вектори колінеарні, отже набори лінійно залежні.*

### **§3. Лінійна залежність між двома наборами даних.**

Якщо отриманий таким чином коефіцієнт виявиться близьким до 1, це й буде означати, що існує досить чітко виражена лінійна залежність і можливо її знайти між наборами і .

Нехай залежність має вигляд тоді сумарна похибка дорівнює:

Щоб цю похибку зробити якомога меншою, треба знайти мінімальне значення квадратичної функції , що до речі дає назву методу: метод найменших квадратів.

Мінімум функції досягається в точці, в якій часткові похідні функції дорівнюють нулю.

Розв’язання отриманої системи дає необхідну відповідь:

***Зауваження* *3.1.*** *Знаходження лінійної залежності методом найменших квадратів можливе і при відсутності лінійної кореляції, але ця залежність не відповідає дійсності у такому разі.*

### **§4. Лінійна залежність між багатьма випадковими величинами.**

Спочатку замість даних величин розглянемо величини що відповідає переходу від афінного простору до лінійного.

Якщо лінійна залежність існує, то точки повинні знаходитись в деякій площині, що містить у собі начало координат, або досить близько до неї.

Складність задачі порівняно з залежністю двох випадкових величин полягає у тому, що може існувати лінійна залежність величин при тому кожна з пар може не мати ніякої залежності, а значить, попарні коефіцієнти кореляції можуть при цьому бути нульовими.

Як легко бачити, для знаходження кореляції , було достатньо знати математичні сподівання, дисперсії і коваріацію випадкових величин, але в -вимірному випадку цього, очевидно, не може вистачити.

Досить глибока причина цього эфекту знаходиться у тому, що функція або щільність розподілу випадкової величини хоч и містить основну інформацію про неї, але не дає змоги, наприклад, з’ясувати, чи залежні випадкові величини про які повністю відомі їх щільності.

В класичній теорії, як правило, умова незалежності випадкових величин або формально входить в умови задачі, або витікає з інтуїтивного розуміння умов задачі.

Вважаючи на це, будемо виходити з даних експерименту, оскільки залежність даних випадкових величин точно не можу знаходитись в умовах задачі. Отже будемо шукати площину, близьку для числових наборів , які припустимо, ми отримуємо в результаті експерименту. Тобто діємо в протилежному напрямку, якщо порівнювати з задачею для двох випадкових величин.

Сформулюємо точно задачу. Маємо -вимірну випадкову величину і її реалізацію у вигляді векторів або в координатному поданні у вигляді матриці:

Якщо всі вектори знаходяться в одній площині, не співпадаючій з усім простором, то ранг матриці дорівнює розмірності площини. Оскільки ранг можна знаходити розглядаючи, як строки, так і стовпці, то необхідно знайти розмірність векторів, ортогональних або строкам, або стовпцям. Зрозуміло, що легше шукати такі вектори в ніж в

Отже нам необхідно знайти де – канонічний скалярний добуток.

А якщо вектори знаходяться не точно в деякій площині , а лише поблизу її, задача дещо змінюється і ми шукаємо де основної маси векторів

– норма і досить мале число, яке характеризує близькість вектору до площини Якщо вектори утворюють базис то вектори близькі до площини , яка задається системою співвідношень

Сформульована задача все ще досить складна і щоб її зробити більш алгоритмічно простою можна розв’язувати її по індукції, тобто спочатку знайти вектор

Якщо вектор існує, то в гіперплощині знову ставимо ту ж задачу і корозмірність площини яку ми шукаємо, буде дорівнювати кількості знайдених таким чином векторів Тобто,

Щоб аналітично розв’язати задачу, можна діяти аналогічно методу найменших квадратів. Нехай

Згідно з методом Лагранжа

Як бачимо, ми отримали задачу на знаходження власних векторів матриці з елементами

***Визначення* *4.1.*** *Симетрична матриця*

*де значення є результатами випробувань випадкової величини називається статистичною матрицею кореляції випадкових величин*

Існує теоретичний спосіб знаходження власних значень і власних векторів довільної матриці використовуючи співвідношення

Але для практичного використання цей спосіб хоч і можливий, але досить не зручний, тому що необхідно розв’язувати громіздку задачу знаходження поліному , з обчисленням детермінанту з досить складною формулою.

Краще використовувати ітераційну формулу для власного вектору, тим більше, що ми шукаємо приблизний розв’язок, а саме:

Цей, зауважимо, інваріантний факт випливає з розгляду приведеного способу у канонічних координатах: і очевидно, найбільшою координатою стає та, для якої найбільше. Після нормування вектору всі координати за виключенням одної стають близькими до нуля і ми отримуємо власний вектор.

У якості вектору можна брати любий вектор, але з ймовірністю 0, цей вектор може приналежним до інваріантного простору, трансверсальному власному вектору

Хоч це майже неймовірно, але все ж таки краще взяти 2 або 3 вектори і упевнитись, що в кожному з випадків ми отримуємо однаковий результат.

Наступний власний вектор шукаємо вже для лінійного перетворення:

в якому власний вектор має вже нульове власне значення.

Коли власні вектори знайдено, обчислюємо для кожного з них

і якщо це значення виявляється малим, меншим заздалегідь вибраного значення то це буде означати, що гіперплощина, перпендикулярна знайденому вектору є площиною скупчення випадкових векторів

Отримане рівняння гіперплощини дає змогу позбутися однієї змінної, зауважимо, що в математичному плані неважливо якої, а виходячи з конкретного експерименту краще залишити, зрозуміло, ті величини, які найбільш зручні для їх знаходження.

І далі повторюємо всі виконані дії, щоб виявити, чи не є залежності між величинами, що залишилися.

Процес закінчується, коли вектори, що задовольняють умовам задачі вже не знаходяться.

***Зауваження* *4.1.*** *Як виявилось, математичні сподівання, дисперсії, а також коваріації випадкових величин не дозволяють знайти їх ймовірнісну залежність. Але, якщо замість цих величин розглянути нормовані належним чином величини*

*то для таких випадкових величин кореляційна матриця спрощуються і набуває вигляду:*

***Визначення* *4.2.*** *Симетрична матриця*

*де – випадкові величини називається приведеною матрицею кореляції випадкових величин якщо*

***Зауваження* *4.2.*** *Лінійна залежність випадкових величин залишається незмінною, якщо замість них розглянути їм пропорційні нормовані величини*

# **Нелінійна залежність випадкових величин та статистичних даних.**

Проблема виявлення нелінійної залежності між даними принципово відрізняється від пошуку лінійної залежності тим, що при лінійній залежності ми практично підганяємо існуючі дані під майже визначену функцію. Майже, тому що лінійна функція однозначно визначається невеликою кількістю параметрів.

В разі пошуку нелінійної залежності можливостей не просто набагато більше, а принципово більше. Задача при цьому ускладнюється ще тим, що важливим стає походження даних, що було не важливо для лінійної залежності. Бо лінійна залежність або існує, або ні, і це досить легко перевіряється.

Якщо ж ми маємо дві нелінійні функції , одна з яких є монотонною, то нелінійна залежність між ними існує завжди!

А саме, а для монотонної завжди існує. Аналогічно, якщо монотонною є

Але ж якщо ми говоримо про залежність двох наборів чисел, то під словом залежність розуміємо не формальну залежність, а впливання в ту чи іншу сторону одних даних на другі.

При лінійній залежності цієї проблеми також не існує, тому що випадково між очевидно не пов’язаними наборами даних лінійна залежність не може виникнути. А нелінійну залежність знайти можна майже завжди!

Тоді виникає питання, а чи взагалі існує сенс пошуку нелінійної залежності.

Відповідь на це питання, якщо коротко, то безумовно сенс існує, тому що у природі нелінійних залежностей значно більше, ніж лінійних, просто задача стає складнішою.

Друге питання полягає в тому, а чи не достатньо обмежитись лінійними залежностями, які знаходяться легко.

Звичайно можна, але приблизно з 1960 математичний світ переключився на розв’язання нелінійних задач тому, що прогрес у лінійній математиці скінчився по причині вичерпання принципових задач, і у той же час були знайдені і реалізовані дуже принципові методи аналізу нелінійних фізичних процесів.

Зрозуміло, що можна звернутись до конкретних прикладних задач, що потребують постановку нелінійної задачі, але це буде не більш, як розв’язання окремого випадку загальної задачі.

Наша ж мета – розглянути всі можливі випадки, які могли б мати місце і окреслити підходи до їх розв’язання. А детальний розгляд, безумовно, має сенс, лише коли конкретні дані відповідають прогнозованим можливим випадкам.

Основні пункти визначеної задачі попередньо можна визначити такими:

1. знаходження залежності загального вигляду між двома випадковими величинами;
2. знаходження залежності загального вигляду між більше ніж двома випадковими величинами;
3. визначення випадковості набору даних;
4. виділення тренду з набору даних;
5. виділення дискретної та неперервної складових випадкової величини;
6. знаходження залежності у загальному випадку між двома наборами даних;
7. знаходження залежності у загальному випадку між більше ніж двома наборами даних;
8. прогнозування.

### **§1. Коефіцієнт загальної кореляції між двома наборами даних.**

Маємо два набору даних і і щоб виявити залежність між ними відмітимо на координатній площині точки Розіб’ємо прямокутник, що їх вміщає, на подібних прямокутників , розбиваючи, наприклад, кожну сторону на рівних відрізків. Якщо випадкові величини не пов’язані ніякою залежністю, точки утворять множину і розподіляться рівномірно по прямокутнику і, навпаки, якщо залежність є, то згуртуються чи то біля лінії, чи то у локальній області. Відмітимо, що ліній може бути більше ніж одна, а також і локальних областей з купою точок може також бути декілька.

Розділимо кількість прямокутників , що не містять у собі хоч одну точку множини яку надалі будемо називати ***множиною значень даних***, на загальну кількість прямокутників Ми отримаємо коефіцієнт у межах від нуля до одиниці, який буде характеризувати наскільки рівномірно точки розподіляються і який будемо називати ***коефіцієнтом загальної кореляці***ї . І чим більше буде цей коефіцієнт, тим, природно, більша залежність.

***Зауваження 1.1.*** *Назва коефіцієнту загальної кореляції зумовлена тим, що по-перше він виявляє не тільки лінійну або нелінійну кореляцію, а взагалі довільну залежність, яка може буді складною, тобто мати і неперервну, і дискретну й іншу залежність, а по-друге термін «кореляція» без додавання «лінійна» все ж таки вживається часто саме у сенсі лінійна кореляція.*

Для обчислення коефіцієнту загальної кореляції досить легко написати комп’ютерну програму для його знаходження. Наприклад, у середовищі WMaple:

***Програма 1.1.***

***Зауваження 1.2.*** *Коефіцієнт загальної кореляції двох наборів даних*  і  *залежить від масиву вибірки, тобто від а також в меншій мірі від розміру прямокутника який хоч і визначається наборами даних, але в деякій мірі приблизно.*

***Зауваження 1.3.*** *Якщо залежність між величинами які визначають набори*  і  *дійсно існує, то при збільшенні кількості елементів набору коефіцієнт загальної кореляції буде збільшуватись, причому тим швидше, чим залежність жорсткіша.*

***Приклад 1.1.*** Нехай залежність має вигляд Геометрична картинка має вигляд.

Застосовуючи програму 1.1, отримуємо: і це видно на рисунку. Далі маємо:

Оскільки при збільшенні коефіцієнт загальної кореляції повинен також збільшуватись, то

***Визначення 1.1.*** *Абсолютним* *коефіцієнтом загальної кореляції будемо вважати*

***Зауваження 1.4.*** *В практичному плані можливо отримати лише оцінку знизу для абсолютного* *коефіцієнту загальної кореляції.*

### **§2. Зсув даних у часі.**

При пошуках залежних випадкових величин або наборів даних необхідно мати на увазі можливість залежності зі зсувом у часі, тобто можливість, коли один чинник впливає на другий, але результат впливання відбувається з запізненням.

Якщо при цьому залежність лінійна, то зсув у часі характер залежності не змінює, вона залишається лінійною, хоч і змінюється нахил графіку залежності.

Те ж вірне і для нелінійної залежності, але на відміну від лінійної графік залежності може змінитися досить суттєво.

***Приклад 2.1.*** Скористуємось прикладом 1.1, змінивши аргумент, який умовно можна вважати часом, на 5 одиниць. Графік матиме вигляд:

і параметри кореляції суттєво зміняться, в першу чергу лінійна кореляція

Для прикладу 1.1.

Для прикладу 2.1.

***Висновок 2.1.*** *Загальний коефіцієнт кореляції , що повинен характеризувати існування загальної, у тому числі нелінійної залежності, при зсуві часу майже не змінюється, при тому, що коефіцієнт знаходить лінійну залежність у прикладі 2.1., не знаходить її у прикладі 1.1. і, зрозуміло, ніяк не фіксує існуючу нелінійну залежність.*

*Але, вважаючи на те, що лінійна залежність більш проста і зручна, максимально ефективним виглядає одночасне використання і лінійної, і загальної кореляцій.*

### **§3. Компоненти зв’язності множини значень даних.**

Маємо нелінійну залежність з кореляцією . Перше, що необхідно зрозуміти, яка кількість елементів зв’язності.

***Визначення 3.1****. Нехай* – *множина точок на площині. Компонентою зв’язності рівня називається підмножина що задовольняє умові:*

*для любої точки знайдеться друга із цієї ж підмножини на відстані меншій, ніж*  *і навпаки, відстань між довільними точками, одна з яких належить підмножині , а інша не належить, більша, ніж*

Для з’ясування цього можливо використати наступний алгоритм.

1. Вихідні дані: маємо точок у прямокутнику
2. Вводимо метрику. Можливо взяти звичайну евклідову, але більш простою є метрика . Вибираємо також фіксовану границю відстані між точками .
3. Заповнюємо прямокутну матрицю з строками і приблизно з стовпцями. Далі кількість ненульових стовпців буде дорівнювати кількості елементів зв’язності, а кількість строк.
4. Кожна точка з порядковим номером буде відповідати одиничному елементу матриці -тої строки.
5. Обчислюються відстані від -тої точки до попередніх, починаючи з першої. Стовпець для -тої точки вибирається за правилом:
   1. якщо -та точка знаходиться від всіх попередніх точок на відстані більшій, ніж то одиниця записується у першому ненульовому стовпчику;
   2. в протилежному випадку, якщо одиниця записується у тому стовпчику, в якому знайшлась близька точка.
      1. після цього порівнюється відстань від -тої точки до точок з номером більше, ніж і якщо знову знаходиться близька точка стовпчики, що містять у собі інформацію про точки об’єднуються в один, а саме у стовпчик з меншим порядковим номером.

Коли алгоритм завершується ми отримуємо у стовпчиках одиниці, що вказують на множину точок, у якій для любої точки знайдеться інша на відстані меншій, ніж І навпаки, дві довільні точки із різних стовпців знаходяться на відстані більшій, ніж Кількість ненульових стовпчиків показує кількість компонент зв’язності рівня

Перш, ніж розглядати залежності, на які вказують компоненті зв’язності, необхідно розглянути дискретну випадкову величину яка приймає значення від одиниці до і для -того випробування з загальної кількості дорівнює номеру компоненти зв’язності, в якій опиняється -та точка.

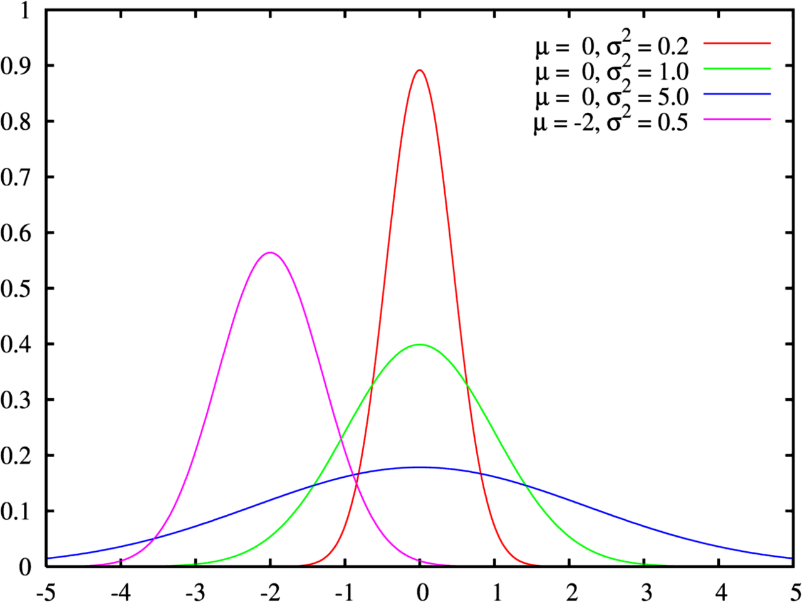
Розподіл випадкової величини її математичне сподівання, дисперсія та інші параметри являються першими характеристиками залежності числових наборів

### **§4. Залежність у межах зв’язної компоненти множини .**

Для зв’язної компоненти множини є дві крайні можливості: 1) всі точки гуртуються в околі деякої точки, або 2) утворюють розмиту лінію. Зрозуміло, що можуть бути варіанти між крайніми можливостями.

Варіант 1) є досить тривіальним. Він означає, по суті, що випадкова величина, що формує зв’язну компоненту є константою. В такому випадку вона характеризується лише математичним сподіванням і дисперсією.

В особливих випадках можливо дослідження також і її функції щільності по кожній з координат, але навіть у таких випадках необхідно розуміти, що якнайшвидше, згідно з центральною граничною теоремою щільність буде мати вигляд нормального розподілу.



Якщо це не так, то можливо має сенс більш ретельно дослідити набор чисел, тому що з любим відхиленням від норми може бути пов’язаний не очевидний висновок.

Варіант 2) представляє більший інтерес тому, що він показує на залежність, яку треба досліджувати.

Основним висновком такого дослідження має бути тренд, що подається, як функція така, що точки знаходяться в невеликому околі її графіку.

В другу чергу, але все ж представляє інтерес випадкова величина що показує відхилення даних від тренду.

Для вибору одного з варіантів 1) або 2) можливе використання наступної характеристики.

***Визначення 4.1.*** *Коефіцієнтом зв’язності елементу зв’язної компоненти* *рівня будемо називати величину*

*тобто відношення кількості точок що находяться від даної точки на відстані, що менше, ніж до загальної кількості елементів компоненти зв’язності множини*

*Середнє значення для підмножини тобто*

*будемо називати коефіцієнтом зв’язності зв’язної компоненти* *рівня*

***Зауваження 4.1.*** *Очевидно, що і чим ближче до одиниці значення тим більш явно ми маємо випадок 1) і навпаки.*

***Зауваження 4.2.*** *Границі або для варіантів 2) і 1) відповідно, вибираються на експертній основі.*

***Зауваження 4.3.*** *Третій можливий варіант виглядає як варіант без характерної залежності і представляє інтерес лише у контексті всієї множини а не окремої компоненти.*

### **§5. Пошук нелінійного тренду даних.**

***Визначення 5.1.*** *Трендом даної функції будемо називати функцію, яка приблизно дорівнює функції але має менші похідні, тобто має більш гладкий графік.*

*Тренд функції знаходиться не однозначно, а виходячи з попередніх допущень і способу апроксимації.*

Розглядаємо випадок який відповідає можливій наявності тренду і в якому множина точок що є зображенням даних і , виглядає, як розмита лінія.

Спочатку ***(перший етап)*** визначимо, який з двох наборів даних є набором аргументів, а який функцією.

Зрозуміло, що для більш точного подання функції залежності треба вибрати такий з двох можливих варіантів, щоб похідна була якомога менша. У простому випадку коли відрізок множини має товщину то похибка значення функції буде дорівнювати де є коефіцієнтом нахилу відрізка, що і відповідає значенню похідної. З двох можливих варіантів значення коефіцієнту нахилу і вибираємо той, що менше одиниці.

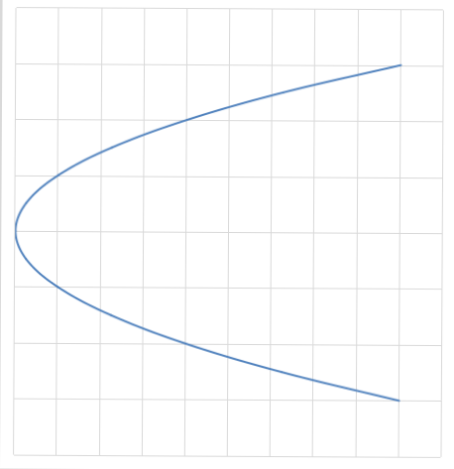
Для наборів даних цей вибір характеризує число

де тобто множина таких пар точок з що відстань між ними менша ніж число що характеризує рівень зв’язності.

Якщо виявляється, що то значення вважаємо значеннями функції з аргументами

Оскільки теоретично можливості вигляду тренду необмежені, то відкидаємо, наприклад, такі.

***Приклад 5.1.*** Функції

**

Згідно з визначеними умовами вибору, необхідно вибрати осі таким чином, як на правому малюнку, і ми отримаємо неоднозначну функцію, що виглядає хибним вибором порівняно з лівим малюнком.

На це можливо відповісти, що на любий апріорний вибір методу роботи з даними можливо вигадати таку залежність, що метод виявиться не ефективним, або зовсім не прийнятним, оскільки нелінійні залежності на відміну від лінійних мають необмежену варіативність.

Тому не маючи конкретних даних природно допускати розгляд можливих залежностей, обмежених рядом умов, які природно формулювати при виконанні тої чи іншої дії з даними.

Що ж стосується приведеного прикладу, то на першому етапі, пов’язаному з обчисленням параметру і, тим самим, з вибором функції і аргументу, апроксимація функцій подібних до або виявляється більш зручною, ніж апроксимація функції, подібної до .

***Приклад 5.2.*** Середня квадратична похибка функції і її лінійної апроксимації

складає 0,14, в той же час при зміні областей значень і визначення

середньоквадратична похибка – 0,16, що і було передбачено.

***Зауваження 5.1.*** *Щодо топологічних особливостей трендів, то, безумовно, якщо на реальних даних вони виявляються не стандартними, то такі тренди потребують окремого дослідження, але при попередньому розбору можливостей вони випадають із розгляду.*

На ***другому етапі*** роботи з даними для отримання тренду необхідно визначити ***опорні точки графіку***, тобто серед сукупностей точок множини необхідно вказати точки, які однозначно будуть вказувати на значення функції при фіксованому аргументу.

Для цього вибираємо на області визначення вузли , в яких бажано мати значення функції, а також – радіус околу точок Точки природно вибирати на рівних відстанях одна від одної, але це не обов’язково, якщо характер функції на різних проміжках помітно змінюється.

Значення функції визначаємо за формулою:

де тобто множина точок абсциси яких відрізняються від менше, ніж на – кількість елементів множини

Зрозуміло, що важко передбачити вигляд функції отже наступний ***третій етап***, а саме процедура згладжування функції, не є обов’язковим, але можливим і при необхідності може бути повтореним декілька разів.

тут – кількість вузлів.

***Приклад 5.3.*** Згладжування графіку.

Дискретне подання.

Неперервне подання.

Червоний графік відповідає згладженому тренду.

***Четвертий етап*** полягає у виборі елементарної функції або суми, добутку, відношення чи композиції елементарних функцій, що найбільш точно відповідають характеру отриманого тренду.

Найбільш зручними у такій ролі є поліноми та тригонометричні функції. Зрозуміло, що можливе використання і експоненти або логарифмів, але їх недоліком є дуже швидке зростання у першому випадку і швидке падіння у другому. А за межею області з різким падінням, наприклад, логарифм зростає дуже повільно. Так, у проміжку на якому аргумент зростає в 30 разів функція зростає приблизно лише у 3 рази. Але, хоч це і виняток, така функція теж може бути використана.

Зауважимо також, що названі функції необхідно використовувати з параметрами: для поліномів – це коефіцієнти перед мономами, для тригонометричних функцій – це амплітуда, частота і фаза і, аналогічно, для сум, добутків та інших випадків.

Наступний, ***п’ятий етап*** – це нелінійна регресія.

***Приклад 5.4.*** Знаходження квадратичного тренду для згладженої функції с прикладу 5.3 методом найменших квадратів.

Мінімум функції досягається в точці, в якій часткові похідні функції дорівнюють нулю.

Розв’язуючи отриману лінійну систему, маємо

і параболічний тренд.

***Зауваження 5.2.*** *Початкова функція з прикладу 5.3. задавалась формулою тобто квадратичну функцію дещо спотворювала функція яку можна вважати впливом деякої випадкової величини.*

*Порівняємо незбурену початкову функцію з отриманим трендом для збуреної величини:*

Отже, з невеликою похибкою тренд відновлює збурену функцію.

***Шостий етап*** – порівняння трендів, отриманих за допомогою різних подань елементарними функціями.

Якщо маємо декілька можливостей для тренду,

то виявляємо найкращий, для якого значення функції буде найменшим.

***Приклад 5.5.*** Використовуючи для апроксимації кубічну функцію

замість квадратичної, отримуємо більш точну апроксимацію. Функція похибки дорівнює 1,340 для кубічної апроксимації і 1,363 - для квадратичної.

***Сьомий етап*** – аналіз функції збурення тренду.

***Визначення 5.2.*** *Різниця* *між функцією* , *отриманої від початкових даних і її трендом* *називається функцією збурення.*

Основні питання, що стосуються функції збурення – це питання її випадковості і залежності, функціональної або статистичної, від аналогічних збурень інших даних.

### **§6. Аналіз функції збурення тренду*.***

Навіть у випадку, коли нелінійна залежність двох наборів даних має не випадковий характер, функція збурення скоріш за все випадкова і причиною цього є несистемні коливання або дискретні впливання будь яких не прогнозованих факторів. В такому разі ніякої інформації функція збурення не несе, а лише спотворює корисну інформацію і виявляється тим, що прийнято називати «білим шумом».

Тим більший представляє інтерес ситуація, коли це природне припущення виявляється невірним.

Згідно з методу отримання функції збурення, вона має дві властивості:

1. обмеженість абсолютного значення порівняно з даними і
2. коливання у межах максимального абсолютного значення.

Неперервне подання функція збурення очевидно, не відображає природі збурення, яке по свої суті має дискретних характер, отже необхідно повернутися до її дискретного подання. Це можна зробити

1. за допомогою уже отриманої неперервної функції

визначаємо координати максимумів і мінімумів функції , не відрізняючи максимуми і мінімуми, і значення функції у точках екстремумів,

або

1. використовуючи вихідні дані

вихідних дані у вигляді пар упорядкуємо по першій координаті і визначимо двовимірну випадкову величину

Логічно припустити, що основна інформація про функцію збурення буде отримана у першому варіанті, який як раз використовує меншу кількість даних, але приступаючи до реальних даних при перших спробах їх аналізу, можливо має сенс дослідити будь яку отриману інформацію.

***Зауваження 6.1.*** *Випадкові величини* *принципово відрізняються* *тим, що набір не містить у собі інформацію про відстані між локальними екстремумами.*

*Але, якщо необхідно мати уявлення про функції розподілу або щільності випадкових величин, то той же набір виявляється більш інформативним, оскільки статистика завжди являється тим надійнішою, чим більше в ній даних.*

***Зауваження 6.2.*** *Існують методики по визначенню того, чи є даний набір чисел випадковим ([1],[2]). Але навряд у випадку використання реальних даних всі вони є необхідними. Адже основний їх сенс у тому, щоб відрізнити квазівипадкові набори чисел від дійсно випадкових, навіть, якщо квазівипадкові числа програмно сформовані по дуже складному алгоритму, що не виявляється на перший погляд.*

Аналіз дискретних випадкових величин на першому етапі стандартний: знаходження вибіркового математичного сподівання вибіркової середньої дисперсії та методом Пірсона визначення нормальності розподілу.

Якщо розподіл виявляється таки нормальним, то в даному випадку це є така інформація, що констатує згідно з Центральною Граничною Теоремою, що в функції збурення відсутні системні залежності. Отже подальший розгляд не має сенсу.

Єдиним позитивним змістовним результатом розглядання збуреної функції можуть бути основні отримані параметри та які при необхідності можуть бути використані.

Але якщо, хоч це і мало ймовірно, навпаки, розподіл якихось даних не є нормальним, то можливе існування системних залежностей, визначення яких може привести до досить не очевидних висновків.

Все ж таки, зважаючи на малу ймовірність такої ситуації, не досить розумно намагатись її якось аналізувати, розглядаючи ті чи інші можливості.

При тому, якщо на практичному матеріалі щось подібне малоймовірне відбудеться, то, це, безумовно, необхідно проаналізовати, скажімо, у окремих додатках.

### **§7. Використання тренду*.***

Якщо випадкове походження функції збурення виглядає майже очевидним у загальному випадку, то природа функції тренду заздалегідь ні в якому разі очевидною не являється.

Питання про характер походження залежності можна взагалі не ставити, відносячись до факту її існування прагматично. Є залежність – добре, при нагоді скористуємся нею.

Але при цьому треба добре усвідомлювати наступне.

На даному наборі даних припустимо, що залежність між вибраними змінним насправді визначена. Але ж чи можна на цій основі робити висновок, що аналогічна залежність буде існувати, якщо подібні дані отримати в інший час і в іншому місті? Відповідь: «Якщо обмежитись тільки знаходженням тренду так, як було запропоновано вище, то ні в якому разі».

На початку розділу було зауважено, що між довільними гладкими функціями завжди існує залежність, якщо одна з функцій є монотонною. Отже, якщо в загальному випадку отримати вибірки двох пар аналогічних наборів даних в різних містах або в різний час, то отримані залежності якнайшвидше будуть існувати, але можуть категорично відрізнятись одна від одної. Отже використовувати отриману залежність можливо тільки локально. Це звісно обмежує область використання результату, але підтверджує очевидну істину: серйозний неочевидний результат не може бути отриманий суто формально.

Або формальні дані аналізуються експертом, який за допомогою графіків та параметрів числових масивів може зробити принципові висновки, базуючись на своєму досвіді, або створюється така програма високого рівня, що робить аналогічні, а може ще й більш точні і глибокі висновки.

Отже після отримання перших залежностей, які можна охарактеризувати, як формальні, необхідно звернути увагу на такі з них, які виглядають, як не ситуативні.

Якщо у самому процесі знаходження стабільних залежностей з’являються певні закономірності, то їх можливо вважати закономірностями високого рівня, які теж можливо програмно реалізувати. Саме вони і складають верхній поверх програми прийняття рішень.

І безумовно, на першому етапі такої роботи необхідно об’єднувати дані декількох наборів, а не тільки двох, щоб отримати багатофакторні залежності.

Припустимо, що ми маємо декілька наборів, що пов’язані між собою залежностями. Але, якщо розглядати з усього набору залежностей тільки пари, то очевидно знаходження стабільних малоймовірно. Наприклад, уявимо, що дані якогось процесу залежать від двох чинників, а ми розглядаємо тільки один з них. Тоді на тому відрізку, де найбільший вплив має фактор, виключений з розгляду, ми отримаємо зараннє передбачене порушення закономірності.

Отже необхідно, отримавши деякі «підозрілі» залежності додавати до них дані, які, можливо, дадуть більш чіткі закономірності. Або на основі інтуїції розглядати не пари даних, а ціли групи даних, виявляючи в них багатофакторні залежності.

# **Многофакторний аналіз випадкових величин та статистичних даних.**

### **§1. Особливості нелінійного многофакторного аналізу.**

Вихідні дані для многофакторного аналізу наборів

даних по елементів.

Задача: знайти всі існуючі залежності між даними, визначити їх характер, системний або випадковий, серед взаємопов’язаних даних вказати визначальні і залежні, визначити критичні точки функцій або аргументів, в яких існуючі залежності порушуються, тобто знайти точки зміну режимів залежності, а також визначити якісні особливості масиву даних на основі експертної оцінки отриманих характеристик.

Визначення ***існування*** залежності між більше ніж двома наборами даних принципово не відрізняється від аналогічної задачі для двох наборів даних. Отже цитуємо відповідний фрагмент попереднього тексту з поправками на розмірність даних.

Маємо наборів даних і щоб виявити залежність між ними відмітимо у -вимірному координатному просторі точки Розіб’ємо паралелепіпед, що їх вміщає, на подібних паралелепіпедів , розбиваючи, наприклад, кожну сторону на рівних відрізків. Якщо випадкові величини не пов’язані ніякою залежністю, точки утворять множину і розподіляться рівномірно по паралелепіпеду і, навпаки, якщо залежність є, то згуртуються чи то біля поверхні, а у окремому випадку, лінії, чи то у локальній області. Відмітимо, що ліній може бути більше ніж одна, а також і локальних областей з купою точок може також бути декілька.

Розділимо кількість прямокутників , що не містять у собі хоч одну точку множини яку надалі будемо називати ***множиною значень даних***, на загальну кількість прямокутників Ми отримаємо коефіцієнт у межах від нуля до одиниці, який буде характеризувати наскільки рівномірно точки розподіляються і який будемо називати ***коефіцієнтом загальної кореляці***ї . І чим більше буде цей коефіцієнт, тим, природно, більша залежність.

***Твердження 1.1.*** *Якщо між будь якими двома наборами даних з усієї кількості наборів існує залежність, то коефіцієнт загальної кореляції* *обчислений для всієї множини , також буде показувати існування залежності.*

Отже, розглядаючи коефіцієнт загальної кореляції для загального масиву даних ми не втрачаємо окремі існуючи залежності.

Але, навпаки,

***Твердження 2.1.*** *Якщо між наборами загального масиву даних існує залежність, то не обов’язково існує залежність між бодай якою-небудь парою даних.*

Це означає, що навіть, коли у загальному масиві даних не знаходиться ніякої залежності, можливо, додаючи до цього масиву ще один додатковий набір даних отримати вже залежний набір даних. На практиці це означає, що спочатку при першому розгляді даних був випущений з уваги фактор, який виявився одним з визначальних для всієї системи даних або для деяких з них.

Якщо набір даних виявився залежним, то, як і у попередньому двовимірному випадку, необхідно виявити компоненти зв’язності.

***Визначення 1.1****. Нехай* – *множина точок на площині. Компонентою зв’язності рівня називається підмножина що задовольняє умові:*

*для любої точки знайдеться друга із цієї ж підмножини на відстані меншій, ніж*  *і навпаки, відстань між довільними точками, одна з яких належить підмножині , а інша не належить, більша, ніж*

Визначення дослівно повторює аналогічне визначення для точок на двовимірні площині. Необхідно тільки визначити метрику, тобто формули для відстані у -вимірному координатному просторі. Можливо взяти звичайну евклідову метрику або

Наступний етап є визначення розмірності компонент зв’язності. Він аналогічний можливостям 1) і 2) при розгляді вигляду залежності у двовимірному випадку.

З точки зору -вимірного простору ці випадки можливо охарактеризувати, як множини розмірностей нуль і один.

У -вимірному просторі не достатньо розділити компоненти нульової розмірності від інших, тому що треба визначити саме розмірність зв’язної компоненти.

### **§2. Розмірність і остов компоненти зв’язності у -вимірному просторі.**

***Визначення 2.1.*** *Локальною розмірністю компоненти зв’язності у точці зв’язної компоненти* *рівня будемо називати величину*

*при умові, що – ціла частина*

*Тобто, якщо в* -*околі сфери з центром*  *радіусу знаходяться не більше, як* *майже з похибкою ортогональних векторів то*

***Зауваження 2.1.*** *Якщо у визначенні 2.1 локальної розмірністю компоненти зв’язності зменшити то значення найімовірніше збільшиться до максимально можливого, тобто до що просто відповідає тому, що навіть при наявності залежності поверхня графіку цієї залежності має певну товщину, отже максимальну розмірність.*

*Якщо вибирати більшим, ніж товщина графіку залежності, який припускаємо існує, то, очевидно, вже не знайдуться майже ортогональних векторів.*

*Параметр необхідний тому, щоб компенсувати похибку, з якою вектори близькі, але не точно ортогональні.*

*Підбір параметрів можливий у машинному режимі, якщо, наприклад, шукати значення таким, щоб абсолютний коефіцієнт загальної кореляції* *став достатньо малим для околу точки з радіусом*

***Визначення 2.2.*** *Остовом локальної компоненти зв’язності будемо називати вектори які отримані спочатку при обчисленні локальної розмірності точки , а далі при обчисленні локальних розмірностей точок і далі таким же чином до вичерпання точок, які мають максимальну локальну розмірність необхідного рівня і знаходяться від вже знайдених точок на відстані більшій, ніж*

***Зауваження 2.2.*** *Можлива ситуація коли компонента зв’язності має повну, тобто максимальну розмірність. Цей випадок не дає залежність, і має сенс лише дослідження причин такої локалізації даних.*

*Якщо таких компонент декілька і розміри відносно невеликі у порівняні з областю визначення даних, то можливо представляє інтерес векторна дискретна величина, що приймає значення у центрах таких компонент.*

Найбільш інтересним є випадок, коли компоненти зв’язності мають не максимальну розмірність, і постає питання знаходження тренду для отриманого остову.

### **§3. Базові змінні. Тренд.**

***Зауваження 3.1.*** *При знаходженні базових змінних, від яких залежать інші, можливо спиратись на здоровий глузд та інтуїцію. Але треба розуміти, що можливі досить не очевидні залежності, отже є сенс перевіряти на можливу залежність і такі набори даних, що на перший погляд ніяк не пов’язані. Але вони можуть бути пов’язані через додавання іншого набору, разом з яким залежність може проявитися.*

Вважаємо, що для набору даних відповідна множина має розмірністьЦе означає, що базових змінниха останніє залежними.

Розглянемо всі можливі проекції множинина координатні гіперплощини розмірностіі зробимо оцінки об’ємів проекцій. Це означатиме, що з повного набору ми вибираємо , які і вважаємо проекціями у простір Щоб оцінити об’єми проекцій найпростіше обчислити кількість паралелепіпедів, на які розбита область визначення даних, в які попадають точки проекцій.

Визначимо базовими ті набори даних, для яких проекція на відповідну їм гіперплощину має найбільший об’єм. Такий вибір практично гарантує, що кожен з наборів, що залишилися, є однозначною функцією базових. Допущення протилежного висновку означатиме, що розмірність множини в точках, де порушується однозначність, більше ніж що суперечить вибору базових даних. А навіть, якщо таке відбувається, вважаючи на можливі особливі точки множини то таких міра (об’єм) таких ділянок в області визначення є мінімальною порівняно з іншими можливими проекціями згідно з вибором гіперплощини проекції.

Після вибору базових змінних вже немає необхідності розглядати всі наявні набори даних, а достатньо додавати до базових даних лише один набір, і для кожного доданого набору визначати функцію залежності від базових даних.

Для визначення тренду, як числової функції від змінних можна взяти у якості опорних точок тренду вже знайдені точки остову.

Якщо в сусідніх опорних точках похідні, тобто відношення прирощення функції до прирощення аргументу, мають досить великі перепади, необхідно зробити згладжування функції, як це робилось для функції однієї змінної, але тепер це необхідно зробити не в одному, а в різних напрямках.

Для функцій декількох змінних найбільш прийнятним є використання поліномів.

Метод найменших квадратів, як і для функції одної змінної є досить ефективним, враховуючи ту обставину, що функція цілі є квадратичною. Але це вірно лише до відносно невеликої розмірності оскільки функція цілі записується у аналітичному, а не числовому вигляді.

Для досить великих значень або кількості трендових точок апроксимація поліномом скоріш за все не дасть ні достатнього наближення, ні простих формул, що визначать необхідну функцію.

У такому випадку найбільш простим і ефективним є спосіб табуляції функції у вже визначених точках остову, значення у близьких до них точках знаходиться, наприклад лінійною інтерполяцією, тобто за допомогою лінійної функції, що дає невелику похибку, враховуючи точність певною мірою випадкових даних.

# **Література.**

1. *Дональд Э. Кнут.* Искусство программирования — М.: [Вильямс](https://ru.wikipedia.org/w/index.php?title=%D0%92%D0%B8%D0%BB%D1%8C%D1%8F%D0%BC%D1%81_(%D0%B8%D0%B7%D0%B4%D0%B0%D1%82%D0%B5%D0%BB%D1%8C%D1%81%D1%82%D0%B2%D0%BE)&action=edit&redlink=1), 2000. — ISBN 5-8459-0081-6.
2. *М. А. Иванов, И. В. Чугунков.* Теория, применение и оценка качества генераторов псевдослучайных последовательностей. — М.: КУДИЦ-ОБРАЗ, 2003. — ISBN 5-93378-056-1.
3. *G. A. F. Seber, C. J. Wild.* Nonlinear Regression. — New York: John Wiley and Sons, 1989. — ISBN 0471617601.
4. *А.Г. Курош.* Курс высшей алгебры. – Москва: Наука, 1968.
5. *А.Н. Ширяев.* Вероятность*.* – Москва: Наука, 1980.
6. *P. Berkhin*, [Survey of Clustering Data Mining Techniques](http://citeseer.ist.psu.edu/berkhin02survey.html), Accrue Software, 2002.