МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний технічний університет
«Харківський політехнічний інститут»

МЕТОДИЧНІ ВКАЗІВКИ
до виконання лабораторної роботи

«Виявлення взаємозв'язків в статистичних даних»
з дисципліни «Інтелектуальний аналіз даних»

для студентів другого рівня підготовки спеціальностей
122 «Комп'ютерні науки», 124 «Системний аналіз»

Затверджено
редакційно-видавничою
радою університету,
протокол № 1 від 16.02.2023 р.

Харків
НТУ «ХПІ»
2023

Методичні вказівки до виконання лабораторної роботи «Виявлення взаємозв’язків в статистичних даних» з дисципліни «Інтелектуальний аналіз даних» для студентів другого рівня підготовки спеціальностей 122 «Комп’ютерні науки», 124 «Системний аналіз» / уклад. О. С. Мельников. – Харків : НТУ «ХПІ», 2023. – 22 с.

Укладач О. С. Мельников

Рецензент М. А. Гринченко

Кафедра системного аналізу та інформаційно-аналітичних технологій

ВСТУП

Математичним підґрунтям всіх дисциплін, пов'язаних з аналізом даних, є теорія ймовірностей. Тому для кращого розуміння подальшого матеріалу доцільно нагадати деякі основні концепції з цієї дисципліни.

Хоча перші роботи з теорії ймовірностей відносяться до XVII віка, строге визначення поняття «ймовірність» було надано радянським математиком А. Н. Колмогоровим лише в 1929 р. Воно базується на теорії множин та теорії міри і є занадто формальним для прикладних досліджень. З точки зору інтелектуального аналізу даних найбільш доречною є частотна інтерпретація ймовірності, в якої вона асоціюється з відносною частотою спостереження певної події в наявних даних.

Дуже важливу роль в інтелектуальному аналізі даних відіграє поняття умовної ймовірності. На ньому базуються алгоритми байєсівської класифікації, пошуку асоціативних правил тощо. Також на базі умовних ймовірностей формалізуються загальні поняття залежних та незалежних випадкових величин. Отже, для опанування подальшого матеріалу дисципліни потрібно добре знайомство з цими та іншими базовими поняттями теорії ймовірностей.

Метою даної лабораторної роботи є оновлення знань щодо умовних ймовірностей, випадкових величин і зв'язків між ними та застосування цього апарату для ідентифікації залежностей в статистичних даних.

1. ТЕОРЕТИЧНІ ОСНОВИ

1.1 Випадкові події

Основним в теорії ймовірностей є поняття випадкового експерименту, результати якого неможливо передбачити заздалегідь. Можливі наслідки такого експерименту називаються випадковими подіями. Вважається, що є можливість повторювати експеримент велику кількість разів. Приклади випадкових експериментів наведені в Табл. 1.

Випадкові події зазвичай позначають великими латинськими або грецькими буквами. Подія Ω, яка настає при кожній реалізації експерименту, називається достовірною. Подія $Ж$, яка не може настати ні при одній реалізації експерименту, називається неможливою.

Таблиця 1. Приклади випадкових експериментів

Випадковий експеримент	**Пов'язані події**
Кидання монети	Випадає цифра, випадає герб
Кидання грального кубика	Випадає 4 очка Випадає парна кількість очок Випадає число очок менше 4
Очікування потягу в метро	Час очікування не перевищує двох хвилин Час очікування складає від однієї до трьох хвилин

Із кожної подією A можна пов'язати подію, яка полягає у тому, що A не настає. Цю подію називають протилежною до A і позначають $\overline{A}$.

Сумою двох подій A і B (позначається $A+B$) називається така подія, яка полягає в настанні принаймні однієї із цих подій.

Добутком двох подій A і B (позначається AB) називається така подія, яка полягає в тому, що обидві події відбуваються одночасно.

Дві події A і B називаються несумісними, якщо їх сумісне настання неможливе: $A \cap B = \varnothing$.

Подія A спричиняє подію B (B є наслідком A), якщо із того, що подія A настала випливає, що настає подія B, тобто $A \subset B$.

Сукупність подій $A_1,...,A_n$ утворюють повну групу, якщо одна і тільки одна із цих подій в результаті експерименту обов'язково настає: $A_1 \cup ... \cup A_n = \Omega,\ A_i \cap A_j = \varnothing,\ i \neq j$.

Подія ω називається елементарною, якщо для довільної події A вона спричиняє або подію A, або $\overline{A}$. Тобто елементарні події є найпростішими наслідками випадкового експерименту.

Множина $\Omega=\{\omega\}$ всіх елементарних подій називається простором елементарних подій. Випадкові події розглядаються як підмножини простору елементарних подій.

1.2 Ймовірності випадкових подій

З точки зору інтелектуального аналізу даних найбільш корисною є так звана частотна інтерпретація ймовірності.

Нехай в однакових умовах проводиться серія із n випадкових експериментів, у кожному із яких може настати деяка подія A. Якщо $n(A)$

– число експериментів, у яких подія A настала, то відношення $\nu(A) = n(A)/n$ називається відносною частотою настання події A.

В багатьох випадках при проведенні різних серій із великої кількості експериментів відносні частоти події для цих серій наближаються до певного числа. Ця закономірність називається властивістю статистичної стійкості відносних частот. Таким чином, з кожною випадковою подією можна пов'язати деяке стале число, яке і вважається ймовірністю випадкової події.

Означення. Число, до якого наближається відносна частота події A при зростанні числа експериментів, називається ймовірністю події A і позначається $P(A)$.

На практиці, при великій кількості експериментів, за ймовірність наближено приймають відносну частоту.

Із даного означення випливають такі властивості ймовірності:

1) $0 \leq P(A) \leq 1$;

2) $P(\Omega) = 1$;

3) $P(\varnothing) = 0$;

4) якщо $A \cap B = \varnothing$, то $P(A+B)=P(A)+P(B)$.

Розрахунок ймовірностей подій зручно проілюструвати за допомогою діаграм Венна (рис. 1).

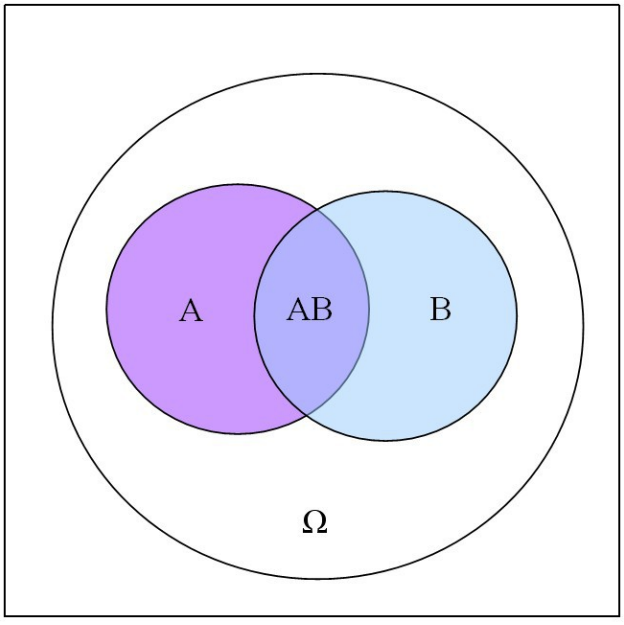


Рис. 1. Діаграма Венна для подій A і B.

Зокрема, з діаграми негайно випливає формула для визначення ймовірності суми двох подій:

$$P(A+B)=P(A)+P(B)-P(AB).$$

1.3 Умовні ймовірності

Якщо при обчисленні ймовірності події A не накладається ніяких умов, крім тих, якими визначається випадковий експеримент, то ймовірність $P(A)$ називають безумовною. Але часто необхідно обчислити ймовірності подій при додатковій умові, що настала деяка інша подія B.

Ймовірність події A, обчислена за припущенням, що подія B уже настала, називається умовною ймовірністю події A за умови B і позначається $P(A\,|\,B)$.

Розглянемо приклад знаходження умовної ймовірності в класичній моделі. Позначимо через n_A, n_B, n_{AB} кількість елементарних подій, що спричиняють відповідно події A, B, AB. Тоді

$$P(A)=\frac{n_A}{n}; P(B)=\frac{n_B}{n}; P(AB)=\frac{n_{AB}}{n}.$$

Якщо подія B вже настала, то змінюються умови експерименту і у новому (умовному) експерименті число можливих наслідків буде рівне n_B – числу елементарних подій, що спричиняють подію B, а подію A будуть спричиняти тільки ті елементарні події, які спричиняють AB. Тому

$$P(A\,|\,B)=\frac{n_{AB}}{n_B}=\frac{n_{AB}/n}{n_B/n}=\frac{P(AB)}{P(B)}.$$

Аналогічно отримаємо

$$P(B\,|\,A)=\frac{n_{AB}}{n_A}=\frac{n_{AB}/n}{n_A/n}=\frac{P(AB)}{P(A)}.$$

З останніх двох формул випливають два способи обчислення добутку подій A і B:

$$P(AB)=P(A\,|\,B)P(B)=P(B\,|\,A)P(A),$$

а також формула для зв’язку між двома умовними ймовірностями:

$$P(A\,|\,B)=\frac{P(B\,|\,A)P(A)}{P(B)}.$$

Формула відома як теорема Байєса.

Приклад. Нехай при киданні грального кістки стало відомо, що випало більше 2 очок. Знайти ймовірність того, що випало 6 очок.

Позначимо через x кількість очок, що випали. Визначимо подію A як «$x = 6$», а подію B – як «$x > 2$». Тоді за формулою шукана ймовірність дорівнює

$$P(x = 6 \mid x > 2) = \frac{P(x = 6, x > 2)}{P(x > 2)} = \frac{P(x = 6)}{P(x > 2)} = \frac{1/6}{4/6} = \frac{1}{4}.$$

Часто формулу надають в дещо іншій формі. Нехай подія A може настати із однією із подій $H_1,...,H_n$, що утворюють повну групу подій. Із попарної несумісності подій $H_1,...,H_n$ випливає, що події $AH_1,...,AH_n$ також несумісні. Тому

$$P(A) = \sum_{i=1}^{n} P(AH_i) = \sum_{i=1}^{n} P(A \mid H_i)P(H_i).$$

Формула називається формулою повної ймовірності. Поєднавши цю формулу з формулою , отримаємо:

$$P(H_i \mid A) = \frac{P(A \mid H_i)P(H_i)}{P(A)} = \frac{P(A \mid H_i)P(H_i)}{\sum_{i=1}^{n} P(A \mid H_i)P(H_i)}.$$

В аналізі даних події $H_1,...,H_n$ називаються гіпотезами, а $P(H_i)$ – апріорною ймовірністю гіпотези H_i. Умовна ймовірність $P(H_i \mid A)$ називається апостеріорною ймовірністю гіпотези H_i.

Приклад. Припустімо, що тест на COVID-19 має чутливість 99% та специфічність 99%. Тобто, цей тест даватиме 99% правильних позитивних результатів для тих, хто хворий на COVID, і 99% правильних негативних результатів для тих, хто ні. Припустімо, що в даний конкретний час 0.5% від загальної кількості людей хворіють на COVID. Якщо для випадково вибраної особи перевірка виявляється позитивною, то якою є ймовірність, що вона дійсно хвора на COVID?

Позначимо наявність хвороби через *yes*/*no*, а результати тесту через +/–. За формулою

$$P(yes \mid +) = \frac{P(+ \mid yes)P(yes)}{P(+)} = \frac{P(+ \mid yes)P(yes)}{P(+ \mid yes)P(yes) + P(+ \mid no)P(no)},$$

$$P(yes\,|+) = \frac{0{,}99 \cdot 0{,}005}{0{,}99 \cdot 0{,}005 + 0{,}01 \cdot 0{,}995} \approx 33.2\%\,.$$

Отже, навіть якщо індивідуальна перевірка дає позитивний результат, то ймовірніше, що людина не хворіє на COVID.

Цей несподіваний результат виникає тому, що кількість «здорових» є дуже великою у порівнянні з кількістю хворих. Таким чином, кількість хибних позитивних результатів (0.995%) переважує кількість правильних позитивних результатів (0.495%).

На конкретних цифрах, якщо перевірено 1000 осіб, то слід очікувати 995 здорових і 5 хворих на COVID. Із 995 здорових очікується $0.01 \times 995 \simeq 10$ хибних позитивних результатів. Із 5 хворих очікується $0.99 \times 5 \simeq 5$ правильних позитивних результатів. Отже, із 15 позитивних результатів лише 5, тобто близько 33%, є істинними.

На базі формули вводиться одне із найважливіших понять теорії ймовірностей – поняття незалежності подій.

Події A і B називаються незалежними, якщо

$$P(AB) = P(A)P(B).$$

З формули безпосередньо випливає, що для незалежності подій A і B необхідно і достатньо, щоб виконувалась одна із наступних умов:

$$P(A\,|\,B) = P(A);\; P(B\,|\,A) = P(B).$$

1.4. Випадкові величини та їх характеристики

Випадковою величиною X називається величина, значення якої залежить від певної випадкової події. Прикладом може бути кількість очок, що випала під час кидання гральної кістки, або час роботи електричної лампочки до перегорання. В першому з цих прикладів випадкова величина може прийняти одне із скінченної (або зліченної) множини можливих значень; такі випадкові величини називають дискретними. В другому прикладі випадкова величина може прийняти будь-яке значення з певного діапазону числової осі, тобто множина можливих значень не є скінченною. Такі випадкові величини називають неперервними.

У подальшому будемо позначати випадкові величини великими літерами, а невипадкові – малими.

Законом розподілу випадкової величини X називається співвідношення, яке встановлює зв'язки між її можливими значеннями та ймовірностями отримання таких значень.

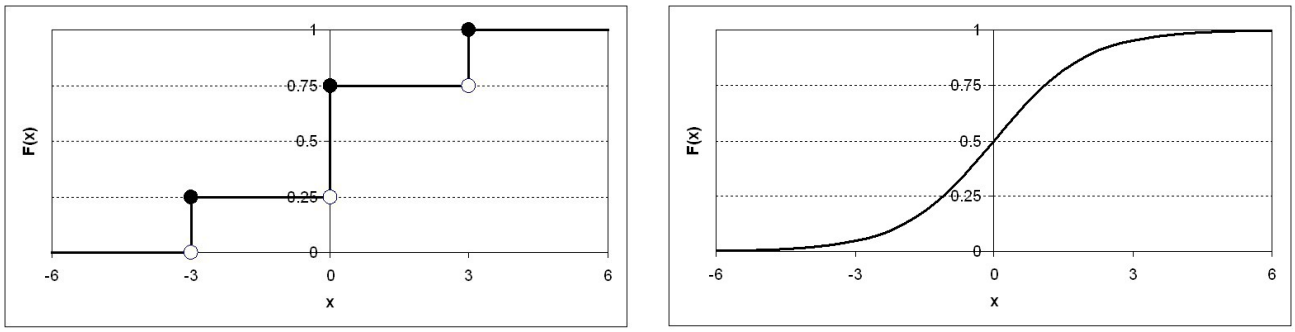
Найбільш повну інформацію про поведінку випадкової величини надає її функція розподілу (також відома як інтегральна або кумулятивна функція розподілу). Вона визначається як ймовірність того, що випадкова величина X прийме значення, яке не перевищатиме заданого рівня x:

$$F(x) = P\{X < x\}.$$

Із визначення функції розподілу безпосередньо випливають її властивості:

1. $0 \le F(x) \le 1 \; \forall x \in (-\infty, \infty)$.
2. $F(-\infty) = 0$.
3. $F(\infty) = 1$.
4. $F(x)$ монотонно не зменшується за своїм аргументом.

Приклади функцій розподілу для дискретних і неперервних випадкових величин наведені на рис. 2.



а *б*

Рис. 2. Приклади функцій розподілу дискретних (*а*) та неперервних (*б*) випадкових величин

Якщо відома функція розподілу, то ймовірність влучання випадкової величини в інтервал $(a, b]$ визначається як

$$P\{a < X \le b\} = F(b) - F(a).$$

Для дискретних випадкових величин часто більш зручною формою опису є ряд розподілу. Це таблиця, яка задає перелік можливих значень випадкової величини та ймовірності отримання таких значень:

X	x_1	x_2	…	x_i	…

,

P	p_1	p_2	…	p_i	…

де

$$p_i = f(x_i) = P\{X = x_i\},\ i = 1,2,...$$

За наявністю ряду розподілу, інтегральну функцію розподілу для дискретних випадкових величин можна отримати як

$$F(x) = \sum_{y \le x} f(y).$$

Для неперервних випадкових величин аналогом ряду розподілу є функція густини ймовірності. Це більш складна конструкція, яка вимагає знання математичного аналізу. В цій та в переважній більшості наступних лабораторних робіт ми будемо мати справу тільки з дискретними випадковими величинами. Тому питання, пов'язані з неперервними випадковими величинами, залишаються читачеві для самостійної роботи. Викладення цих питань можна знайти в будь-якому ґрунтовному посібнику з теорії ймовірностей, наприклад в [1].

Функція розподілу надає повну інформацію про поведінку випадкової величини, але для визначення такої функції в реальних умовах потрібен великий обсяг статистичної інформації, яку може бути складно отримати. Більш стисло можна охарактеризувати випадкову величину за допомогою чисельних характеристик, найбільш розповсюдженими з яких є математичне сподівання та дисперсія.

Математичне сподівання випадкової величини X характеризує її середнє значення. Воно визначається як

$$M[X] = \mu_x = \sum_{x:\, f(x)>0} xf(x).$$

Аналогічно визначається математичне сподівання деякої функції від випадкової величини $h(X)$:

$$M[h(X)] = \sum_{x:\, f(x)>0} h(x)f(x).$$

Для лінійної функції $h(X) = a + bX$, де a,b – довільні константи, з формули випливає, що

$$M[a + bX] = a + bM[X].$$

Дисперсія випадкової величини X характеризує її розкид навколо середнього значення. Вона визначається як

$$D[X] = \sigma_x^2 = M[(X - M[X])^2] = M[X^2] - (M[X])^2.$$

Якщо підставити $Y = a + bX$ в формулу , то після спрощення отримаємо наступну важливу властивість дисперсії:

$$D[a + bX] = b^2 D[X];$$

зокрема, $D[a] = 0$.

Квадратний корінь від дисперсії називається середньоквадратичним відхиленням випадкової величини *X*:

$$\sigma_x = \sqrt{D[X]}.$$

Середньоквадратичне відхилення часто використовується для грубої оцінки діапазону можливих значень випадкової величини.

1.5. Системи випадкових величин

Сумісний розподіл двох випадкових величин *X* та *Y* може бути заданий через інтегральну функцію як

$$F(x, y) = P\{X \text{ J } x, Y \text{ J } y\}.$$

або через таблицю значень як

$$f(x, y) = P\{X = x, Y = y\}.$$

Граничний розподіл випадкових величин *X* та *Y* отримується шляхом підсумовування за значеннями іншої змінної:

$$f_X(x) = \text{e}_y f(x, y);\ f_Y(y) = \text{e}_x f(x, y).$$

Умовний розподіл випадкової величини *Y* при відомому значенні $X = x$ визначається як

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}.$$

Умовний розподіл випадкової величини є звичайним розподілом ймовірностей і для нього можна розрахувати будь-які чисельні характеристики. Зокрема, умовне математичне сподівання випадкової величини *Y* при відомому значенні випадкової величини $X = x$ визначається як

$$M[Y \mid X = x] = \text{e}_{i=1}^{n} y_i P\{Y = y_i \mid X = x\}.$$

Всі ці поняття природним чином узагальнюються на випадок сумісного розподілу трьох та більше випадкових величин.

1.6. Визначення залежностей між випадковими величинами

Випадкові величини X та Y є незалежними, якщо знання однієї із змінних ніяк не впливає на розподіл іншої, тобто $f(y \mid x) = f_Y(y);\ f(x \mid y) = f_X(x).$ Застосувавши ці рівняння до формули , отримаємо наступне правило:

$f(x,y) = f_X(x) f_Y(y)$ Ы X, Y є незалежними.

Це правило виконується також і для інтегральної функції розподілу:

$F(x,y) = F_X(x) F_Y(y)$ Ы X, Y є незалежними.

Приклад. Нехай X, Y – результати двох кидань монети; $X,Y = 1$ якщо випав герб і 0 в іншому випадку. Нехай $Z = X + Y$. Можливі 4 комбінації значень X та Y: 00,01,10,11, ймовірність кожної з яких складає 0,25.

Сумісний розподіл X та Y може бути заданий таблицею

X \ Y	Y=0	Y=1	$f_X(x)$
X=0	0,25	0,25	0,5
X=1	0,25	0,25	0,5
$f_Y(y)$	0,5	0,5	1

Для всіх значень x та y $f(x,y) = f_X(x) f_Y(y)$. Отже, X та Y є незалежними.

Розглянемо тепер сумісний розподіл X та Z.

X \ Z	Z=0	Z=1	Z=2	$f_X(x)$
X=0	0,25	0,25	0	0,5
X=1	0	0,25	0,25	0,5
$f_Y(y)$	0,25	0,5	0,25	1

$f(1,0) = 0$ № $f_X(1) f_Z(0) = 0,5$ г $0,25$. Отже, X та Z є залежними.

Умовне математичне сподівання $M[Y \mid X = x]$ називається також регресією Y на X. Якщо $M[Y \mid x] = M[Y]$ при будь-якому значенні X, то випадкові величини Y на X називають незалежними за математичним

очікуванням.

Коваріацією випадкових величин X та Y називається величина

$$Cov(X,Y) = M[(x - M[x])(y - M[y])] = M[xy] - M[x]M[y].$$

Зауважимо, що за попереднім визначенням $Cov(X,X) = D[x]$.

Якщо X та Y незалежні, то

$$Cov(X,Y) = M[(x - \mu_x)(y - \mu_y)] = M[x - \mu_x]M[y - \mu_y] = 0.$$

Зворотне невірно: якщо $Cov(X,Y) = 0$ (в цьому випадку величини X та Y називаються некорельованими), то вони можуть бути залежними.

Сенс коваріації полягає у наступному. Якщо між X та Y є позитивний зв'язок, то великим значенням змінної X найчастіше будуть відповідати великі значення змінної Y. Тоді обидва співмножники в формулі будуть мати однаковий знак і коваріація буде позитивною. Аналогічно, від'ємне значення коваріації свідчить про наявність зворотного зв'язку між змінними.

На жаль, абсолютне значення коваріації нічого не говорить про силу такого зв'язку. Для цього використовується поняття кореляції.

Кореляція випадкових величин X та Y визначається як

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Можна довести, що $-1 \leq \rho_{XY} \leq 1$. Чим ближче до одиниці абсолютне значення кореляції, тим тіснішим є зв'язок між змінними X та Y.

Якщо $Y = a + bX$, тобто Y є лінійною функцією від X, то

$$\rho_{XY} = \frac{cov(a + bX, X)}{\sqrt{D[X]D[Y]}} = \frac{bD[X]}{\sqrt{D[X]D[Y]}} = b\frac{\sigma_X}{\sigma_Y}.$$

Якщо до того ж змінні X та Y будуть стандартизовані так, що $\sigma_X = \sigma_Y = 1$ (цього можна досягти шляхом ділення на σ_X, σ_Y), то коефіцієнт кореляції буде співпадати з коефіцієнтом b. Отже, кореляція є показником лінійного зв'язку між змінними.

Таким чином, можна виділити три ступеня незалежності між випадковими величинами X та Y.

1. Незалежність: $f(x,y) = f_X(x) f_Y(y)$.
2. Незалежність за математичним сподіванням: $M[Y \mid X] = M[Y]$.
3. Відсутність кореляції: $\rho_{XY} = 0$.

Перше визначення є «найсильнішим», а третє – «найслабшим». Тобто, з визначення 1 випливає визначення 2, а з нього, у свою чергу – визначення 3. Зворотні твердження в загальному випадку не є вірними.

Визначення взаємозв'язків між статистичними даними є центральним питанням інтелектуального аналізу даних.

Виявлення залежності між випадковими змінними в найбільш загальній формі є дуже складною задачею. Для дискретних випадкових величин корисним інструментом для пошуку таких залежностей є таблиці спряженості.

Таблиця спряженості (факторна таблиця) – це засіб представлення спільного розподілу двох змінних, призначений для дослідження зв'язку між ними. Рядки таблиці відповідають значенням однієї змінної, стовпці – значенням іншої. На перетині рядка та стовпця вказується частота спільної появи відповідних значень двох ознак. Сума частот по рядку називається маргінальною частотою рядка; сума частот за стовпцем – маргінальною частотою стовпця. Сума маргінальних частот надає граничний розподіл змінних, що утворюють рядки та стовпці таблиці. У таблицях спряженості можуть використовуватись як абсолютні, так і відносні частоти. Відносні частоти можуть розраховуватися стосовно: а) маргінальної частоти за рядком; б) маргінальної частоти за стовпцем; в) до обсягу вибірки.

Приклад. В таблиці 2 наведені деякі результати другого туру президентських виборів 2019 року в Україні по обраним областям і по країні в цілому за даними, наведеними в [10].

Таблиця 2. Деякі результати другого туру президентських виборів в Україні в 2019 р., тис. осіб

Область Кандидат	Львівська	Харківська	…	Усього
Зеленський	548	1126	…	13542
Порошенко	676	146	…	4522
Усього	1224	1272	…	18064

Перетворимо цю таблицю в таблицю сумісного розподілу змінних «область мешкання виборця» та «вибір кандидата». Для цього поділимо дані таблиці на загальну кількість врахованих бюлетенів (18 064 тис.). Результати зведено в табл. 3.

Дані в останньому стовпчику табл. 3 надають безумовний граничний розподіл голосів за кандидатами, а дані в нижньому рядку – безумовний

граничний розподіл виборців за областю мешкання. Наприклад, в Харківській області мешкає біля 7% від загальної кількості громадян, які взяли участь в виборах.

Таблиця 3. Імовірнісна інтерпретація результатів другого туру президентських виборів в Україні в 2019 р.

Область Кандидат	Львівська	Харківська	…	Усього
Зеленський	0.030	0.062	…	0.750
Порошенко	0.037	0.008	…	0.250
Усього	0.068	0.070	…	1.000

Для того, щоб визначити вплив регіону на електоральні уподобання, знайдемо умовні ймовірності обрання того чи іншого кандидату з врахуванням області мешкання виборця. Для цього поділимо дані табл. 3 на граничний розподіл виборців за областями (нижній рядок таблиці). Результати зведено в табл. 4.

Таблиця 4. Умовні ймовірності вибору кандидатів з врахуванням області мешкання виборця

Область Кандидат	Львівська	Харківська
Зеленський	0.448	0.885
Порошенко	0.552	0.115
Усього	1.000	1.000

Як можна побачити з табл. 4, регіональні відмінності суттєво впливають на вибір кандидата. Так, шанси голосування за Порошенко в Львівській області майже в 5 разів перевищували аналогічні показники в Харківській області.

Спираючись на дані табл. 3, можна спробувати спрогнозувати місце мешкання виборця виходячи з його електоральних симпатій. Для цього в якості умовної змінної слід взяти обраного кандидата. Результати, отримані шляхом поділу даних табл. 3 на граничні ймовірності обрання кандидатів (останній стовпчик), зведені в табл. 5.

Таблиця 5. Умовні ймовірності мешкання в певній області виходячи із обраного кандидата

Область Кандидат	Львівська	Харківська	…	Усього
Зеленський	0.040	0.083	…	1.000
Порошенко	0.149	0.032	…	1.000

Таким чином, якщо про людину відомо, що в другому турі президентських виборах вона голосувала за Порошенко, то непогані шанси на те, що вона зі Львівської області.

Нарешті, визначимо, як виглядали би результати виборів, якби електоральні уподобання виборців не залежали би від місця їх мешкання. В цьому випадку кількість голосів за кожного кандидата визначалась би як (гранична ймовірність голосування за певного кандидата) × (гранична ймовірність мешкання в певній області) × (загальна кількість виборців).

Відповідні результати наведені в табл. 6.

Таблиця 6. Очікувані результати виборів в припущенні незалежності електоральних уподобань від місця мешкання

Область Кандидат	Львівська	Харківська
Зеленський	918	954
Порошенко	306	318
Усього	1224	1272

Легко бачити, що значення в табл. 6 суттєво відрізняються від реальних результатів, наведених в табл. 2. Наступне питання полягає в тому, наскільки суттєвою є ця різниця. Для цього використовуються різні статистичні тести, наприклад, критерій Фішер або критерій згоди Пірсона (хі-квадрат). Детальніше про ці тести можна дізнатися, наприклад, в [7].

Таблиці спряженості можуть використовуватись також і для аналізу неперервних даних, але для цього кількісні шкали попередньо повинні бути згруповані в інтервали.

2. ЗАВДАННЯ НА ЛАБОРАТОРНУ РОБОТУ

До лабораторної роботи додається файл Titanic.xls, який містить різноманітну інформацію про підмножину пасажирів, які перебували на борту корабля «Титанік» під час його фатального рейсу 1912 року. Стовпці таблиці містять такі дані:

1) чи вижив пасажир (0–ні, 1–так) (Survived);

2) клас каюти (від 1 до 3) (Pclass);

3) ім'я пасажира (Name);

4) стать пасажира (Sex);

5) вік пасажира (Age);

6) кількість братів, сестер та подружжя на борту (Siblings/Spouses);

7) кількість дітей та/або батьків на борту (Children/Parents);

8) ціна квитка (у фунтах стерлінгів) (Fare).

Використовуючи цю інформацію, знайдіть наведені нижче величини.

1. Загальну ймовірність виживання.

2. Умовну ймовірність виживання для кожного класу каюти. Визначити, чи існує залежність між класом каюти та ймовірністю виживання.

3. Умовну ймовірність виживання для чоловіків та жінок. Визначити, чи існує залежність між статтю пасажиру та ймовірністю виживання.

4. Умовну ймовірність виживання для пасажирів, які мали ознаку, вказану в таблиці варіантів (табл. 7). Визначити, чи існує залежність між ймовірністю виживання та цією ознакою.

Таблиця 7. Таблиця варіантів

Варіант	Ознака
1	Age≤10
2	10<Age≤20
3	20<Age≤30
4	30<Age≤40
5	40<Age≤50
6	50<Age≤60
7	60<Age≤70
8	Age>70
9	Siblings/Spouses>1
10	Siblings/Spouses=0
11	Parents/Children=0
12	Parents/Children>1

5. Середню вартість квитка для усіх пасажирів а також окремо для тих, хто вижив і тих, хто загинув. Також обчисліть її стандартне відхилення.

6. Умовну та граничну ймовірність виживання залежно від класу та статі пасажира. Тобто заповніть таку таблицю:

	Стать

		F	M	Усі
Клас	1			
	2			
	3			
	Усі			

елементи якої містять ймовірність виживання для відповідних комбінацій рядка та стовпця. Як виглядала би ця таблиця, якби ймовірність виживання не залежала від статі та класу кають?

3. КОМЕНТАРІ ДО ВИКОНАННЯ ЛАБОРАТОРНОЇ РОБОТИ

Робота розрахована на 2 академічних години.

Робота не потребує навичок програмування і розрахована на виконання в програмному середовищі Microsoft Excel. За бажанням роботу можна виконувати із використанням звичайних мов програмування для наукових та інженерних розрахунків, таких як Matlab, MathCAD, Gauss, Python, R тощо. В таких випадках до звіту з виконання лабораторної роботи обов'язково слід додати лістинг розробленої програми. Подальші коментарі передбачають використання Excel.

Для оцінки умовних ймовірностей подій, які розглядаються в лабораторній роботі, потрібно підрахувати кількість записів бази даних, які задовольняють певним вимогам. Для цього в Excel існує багато способів. Мабуть, найпростішим є використання функції `COUNTIF`.

Функція `COUNTIF` має наступний синтаксис:

```
COUNTIF(range, criterion)
```

`range` – діапазон, в якому слід підрахувати клітинки;

`criterion` – критерій у формі числа, виразу або тексту, який визначає, які клітинки слід підраховувати. Наприклад, критерій може мати вигляд `30`, `"<30"`, `"male"`, тощо.

Приклад. На рис.3 наведено фрагмент набору даних titanic.xls.

Для цього фрагменту:

`COUNTIF(A2:A15,1)` поверне кількість пасажирів, що вижили (7);

	A	B	C	D	E	H
1	Survived	Pclass	Name	Sex	Age	Fare
2	0	3	Mr. Owen Harris Braund	male	22	7.25
3	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings	female	38	71.28
4	1	3	Miss. Laina Heikkinen	female	26	7.93
5	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35	53.10
6	0	3	Mr. William Henry Allen	male	35	8.05
7	0	3	Mr. James Moran	male	27	8.46
8	0	1	Mr. Timothy J McCarthy	male	54	51.86
9	0	3	Master. Gosta Leonard Palsson	male	2	21.08
10	1	3	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	female	27	11.13
11	1	2	Mrs. Nicholas (Adele Achem) Nasser	female	14	30.07
12	1	3	Miss. Marguerite Rut Sandstrom	female	4	16.70
13	1	1	Miss. Elizabeth Bonnell	female	58	26.55
14	0	3	Mr. William Henry Saundercock	male	20	8.05
15	0	3	Mr. Anders Johan Andersson	male	39	31.28

Рис. 3. Фрагмент набору даних titanic.xls.

`COUNTIF(B2:B15,”>1”)` поверне кількість пасажирів, що подорожували другим або третім класом (10);

`COUNTIF(D2:D15,”female”)` поверне кількість жінок (7).

Для того, щоб підрахувати кількість записів, які задовольняють складним умовам, можна створити стовпчик з результатами обчислення такої умови. При цьому можуть стати в нагоді функції `AND` та `OR`.

Функція `AND(log_expr1;[log_expr2];...)` повертає значення `TRUE`, якщо всі аргументи мають значення `TRUE`; повертає значення `FALSE`, якщо хоча б один аргумент має значення `FALSE`.

Функція `OR(log_expr1;[log_expr2];...)` повертає значення `TRUE`, якщо хоча б один із аргументів має значення `TRUE`; повертає `FALSE`, якщо всі аргументи мають значення `FALSE`.

Наприклад, для того, щоб підрахувати кількість дорослих чоловіків, які подорожували першим класом, можна в клітинці `G2` ввести формулу `AND(B2=1,D2=”male”,E2>=18)` і скопіювати її в клітинки `G3:G15`. Тоді функція `COUNTIF(G2:G15,TRUE)` поверне шукану кількість (1).

В версіях Excel 2019 та вище додано функцію `COUNTIFS`, яка дозволяє перевіряти декілька умов одночасно. Вона дещо спрощує наведену вище процедуру.

Також можна скористатися інструментом "Зведена таблиця"/ "Pivot Table" з меню "Дані"/"Data". Зведена таблиця – це потужний інструмент обробки даних, що служить для їх узагальнення. Детальну інформацію щодо можливостей зведених таблиць можна знайти у довідці Excel та на відповідних інтернет ресурсах.

4. КОНТРОЛЬНІ ЗАПИТАННЯ

1. Наведіть приклади випадкових експериментів.
2. В чому полягає частотна інтерпретація ймовірності?
3. Що таке простір елементарних подій?
4. Як визначаються сума та добуток подій?
5. Як визначається умовна ймовірність?
6. В чому полягає теорема Байєса?
7. Як визначається незалежність випадкових подій?
8. Розкрийте зміст понять апріорної та апостеріорної ймовірності.
9. В чому полягає формула повної ймовірності?
10. Дайте визначення поняття «випадкова величина».
11. Що називається розподілом випадкової величини?
12. Як визначається функція розподілу випадкової величини?
13. Чим розрізняються дискретні та неперервні випадкові величини?
14. Назвіть основні властивості інтегральної функції розподілу.
15. Як визначити ймовірність влучання випадкової величини у визначений інтервал числової осі?
16. Як визначається і що характеризує математичне сподівання випадкової величини?
17. Як визначається математичне сподівання функції від випадкової величини?
18. Як визначається і що характеризує дисперсія випадкової величини?
19. Як визначається і що характеризує середньоквадратичне відхилення випадкової величини?
20. Як визначається сумісний розподіл двох випадкових величин?
21. Що таке граничний розподіл ймовірностей в системі двох випадкових величин?
22. Як знайти розподіл ймовірностей випадкової величини за деякої відомої умови?
23. Коли дві випадкові величини можна вважати незалежними?
24. Дайте визначення умовного математичного сподівання випадкової величини.
25. Як визначається коваріація між двома випадковими величинами?
26. Як визначається і що характеризує коефіцієнт кореляції?
27. Назвіть три ступеня незалежності між випадковими величинами.

28. Чи можуть бути корельованими незалежні випадкові величини?
29. Чи можуть бути залежними некорельовані випадкові величини?
30. Для чого використовуються таблиці спряженості?

СПИСОК ЛІТЕРАТУРИ

1. Вентцель, Е.С. Теория вероятностей: учеб. для вузов. – 6е изд., стер. / Е.С.Вентцель. – Москва : Высш. шк., 1999. – 576 с.
2. Кушлик-Дивульська, О. І. Теорія ймовірностей та математична статистика: навч. посіб./ О. І. Кушлик-Дивульська, Н. В. Поліщук, Б. П. Орел, П. І. Штабалюк. – Київ: НТУУ «КПІ», 2014. – 212 с.
3. Мазманишвили, А. С. Теория вероятностей: учебное пособие к практическим занятиям / А. С. Мазманишвили. – Харьков: НТУ «ХПИ», 2007. – 212 с.
4. Огірко, О. І. Теорія ймовірностей та математична статистика: навчальний посібник / О. І. Огірко, Н.В.Галайко. – Львів: ЛьвДУВС, 2017. – 292 с.
5. Слюсарчук, П. В. Теорія ймовірностей та математична статистика / П. В. Слюсарчук. – Ужгород: Вид-во «Карпати», 2005. – 178 с.
6. Теорія ймовірностей та математична статистика: Частина 1. Випадкові події: Лекції і практикум [Електронний ресурс] : навч. посіб. / уклад.: І. В. Веригіна, О. В. Островська. Київ : КПІ, 2018. – 57с. URL: https://ela.kpi.ua/bitstream/123456789/23501/1/NP(T_Ym)_1.pdf (дата звернення: 12.01.2023).
7. Greene, W. H. Econometric Analysis. 8th edition / William H. Greene. – London: Pearson, 2017. – 1176 p.
8. Ozdemir, S. Principles of Data Science. 2nd edition / Sinan Ozdemir, Sunil Kakade, Marco Tibaldeschi. – Birmingham-Mumbai: Packt Publishing, 2018. – 420 p.
9. Statistical functions (reference) [Електронний ресурс]. URL: https://support.microsoft.com/en-us/office/statistical-functions-reference-624dac86-a375-4435-bc25-76d659719ffd (дата звернення: 10.01.2023).
10. Протокол Центральної виборчої комісії про результати повторного голосування з виборів Президента України 21 квітня 2019 року [Електронний ресурс]. URL: https://www.cvk.gov.ua/wp-content/uploads/2019/11/vpu_2019_protokol_cvk_30042019.pdf (дата звернення: 12.01.2023).

Навчальне видання

Методичні вказівки
до виконання лабораторної роботи
«Виявлення взаємозв’язків в статистичних даних»
з дисципліни «Інтелектуальний аналіз даних»

для студентів другого рівня підготовки спеціальностей
122 «Комп’ютерні науки», 124 «Системний аналіз»

Укладач:
МЕЛЬНИКОВ Олег Станіславович

Відповідальний за випуск Ю. І. Дорофєєв
Роботу до видання рекомендував М. І. Безменов
Комп’ютерна верстка М. І. Безменов

У авторській редакції

План 2023 р., поз. 432

Підп. до друку 17.05.2023 р. Гарнітура Таймс. Ум. друк. арк. 1,1.
Електронне видання

Видавничий центр НТУ «ХПІ».
Свідоцтво про державну реєстрацію ДК № 5478 від 21.08.2017 р.
61002, Харків, вул. Кирпичова, 2