

**МІНІСТЕРСТВО НАУКИ І ОСВІТИ УКРАЇНИ**  
**ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ім. І.І. МЕЧНИКОВА**

Варбанець П.Д., Якімова Н.А.

## **ЛІНГВОСТАТИСТИКА**

Методичний посібник для практичних занять

Одеса

2014

Друкується за рішенням Вченої Ради ІМЕМ ОНУ

від 19 вересня 2013 року, протокол №1

укладачі: д. ф.-м. н. Варбанець П.Д., к. ф.-м. н. Якімова Н.А.

рецензенти: д. ф.-м. н. Леонов Ю.Г.

к. ф.-м. н. Покась С.М.

# ЗМІСТ

Передмова.....	3
1. Імовірнісні простори.....	5
1.1. Класичне визначення ймовірності.....	5
1.2. Урнова схема.....	6
1.3. Скінченна схема з неоднаково можливими результатами.....	7
1.4. Контрольні запитання.....	9
2. Умовні ймовірності.....	11
2.1. Повна ймовірність.....	11
2.2. Формула Байєса.....	11
2.3. Послідовності випробувань.....	13
2.4. Контрольні запитання.....	14
3. Числові характеристики випадкової величини.....	15
3.1. Вибіркове середнє і вибіркова дисперсія.....	15
3.2. Критерій нормальності розподілу.....	17
3.3. Критерій однорідності.....	21
3.4. Розподіл середнього арифметичного значення.....	23
3.5. Контрольні запитання.....	25
4. Лінгвістичні гіпотези.....	26
4.1. Критерій Стьюдента.....	26
4.2. Критерій Ван дер Вардена.....	28
4.3. Контрольні запитання.....	29
5. Вивчення залежності лінгвістичних ознак.....	30
5.1. Кореляційна залежність.....	30
5.2. Лінійна кореляція.....	31
5.3. Кореляційні відносини.....	35
5.4. Парціальна кореляція.....	36
5.5. Критерії кореляційного аналізу.....	36
5.5.1. Критерій вірогідності залежності ознак $X$ і $Y$ .....	36
5.5.2. Критерій значимості розходження кореляційних відносин.....	37
5.6. Регресійний аналіз.....	37
5.7. Контрольні запитання.....	39
Список літератури.....	40
Додаток 1.....	41
Додаток 2.....	42
Додаток 3.....	43
Додаток 4.....	44
Додаток 5.....	47

## ПЕРЕДМОВА

Мова являє собою, як прийнято говорити в сучасному мовознавстві, деяке системно-структурне утворення. Окремі підсистеми мови називають *рівнями*, які представлені відповідними одиницями – фонемами, морфемами, лексемами, синтагмами (реченнями).

Оскільки одиниці кожного рівня мови перебувають в ієрархічній залежності від одиниць вищестоящего рівня, то зрозуміло, що, наприклад, число похідних слів у тій або іншій мові буде залежати від кількості афіксів з дериваційним значенням, а кількість морфем - від кількості фонем. У той же час кількість фонем у різних мовах не збігається. Ці прості приклади показують, що мова характеризується певними якісними й кількісними ознаками [6].

*Якісний* аналіз мови являє собою його категоризацію, тобто виділення в мові певних класів явищ, об'єднаних певними якісними ознаками. Цими явищами (категоріями) можуть бути одиниці мови (фонема, морфема, лексема), граматичні категорії, граматичні способи (афіксація, словоскладання, редуплікація й т.д.), типи слів (знаменні, службові; вульгаризми, діалектизми; архаїзми, неологізми й т.д.), типи речень (складні, прості; сурядні, підрядні й т.д.). Однак будь-яка категоризація, тобто якісний аналіз мови, нерозривно пов'язана із *квантифікацією* мови, тобто його кількісним аналізом. Таким чином, стає очевидним, що мова поряд з якісними ознаками володіє й кількісними. Ще більшою мірою має кількісні ознаки мова і її письмове втілення - текст.

У сучасній науці розрізняють так звані «добре організовані системи» і «погано організовані (дифузійні) системи» [9, стор.7]. До добре організованих систем належить, наприклад, рух планет. Завдяки чіткій упорядкованості цієї системи стає можливим точно обчислити й заздалегідь пророчити, наприклад, час сонячного затемнення. До погано організованих систем належить інтелектуальна діяльність людини, а разом з нею і мовне поведіння, тобто використання мови. Уважається, що найбільш ефективними методами вивчення погано організованих систем є методи математичної статистики.

Таким чином, мова може бути дослідженою за допомогою якісних і кількісних методів. Залежно від цілей і задач, які ставить перед собою лінгвіст при вивченні явищ мови й мовлення, у здійснюваному дослідженні можуть застосовуватися або якісні, або кількісні методи аналізу, або й ті, і інші рівною мірою, або переважно перші або другі. Можуть виникнути також задачі, (особливо при аналізі тексту), які не можуть бути вирішені інакше, як за допомогою кількісних методів.

У самому математичному апараті, точніше, у сукупності математичних методів можна умовно розрізнити кількісні й некілісні методи [10]. За допомогою некілісних методів (теорія множин, теорія алгоритмів, математична логіка) доцільно вивчати, насамперед, систему мови. Цей розділ науки називається *комбінаторна лінгвістика*. За допомогою кількісних методів (насамперед, теорія ймовірностей і математична статистика) доцільно досліджувати мову (текст). Цей другий напрямок називають *квантитативною лінгвістикою*. Якщо розглядати лінгвостатистику як одну зі складових частин квантитативної лінгвістики, стає очевидним, що між лінгвостатистикою і квантитативною (математичною) лінгвістикою існує помітна різниця, тому що кількість об'єктів і набір методів, за допомогою яких ці об'єкти вивчаються в лінгвостатистиці, значно вужче, ніж у математичній лінгвістиці в цілому.

За допомогою квантитативних методів у цей час досліджуються всі мовні підсистеми, тобто всі рівні мови й мовлення. Є спроби яким-небудь чином систематизувати лінгвостатистичні дослідження, тобто виділити ті галузі мовознавства, де квантитативні методи успішно застосовуються або можуть бути ефективно використані. Приведемо деякі сфери використання математичних методів у мовознавстві [6].

1. Дослідження фонетичного ладу мови (частота зустрічальності звуків, букв, букво- і словосполучень і т.д.).
2. Дослідження лексичного складу мови й частоти зустрічальності слів у тексті. Найважливішою задачею в цій галузі є створення частотних словників.
3. Дослідження авторського й функціонального стилів. Сюди ж можна віднести дослідження, пов'язані із установленням авторства твору. Можливості застосування статистичних методів для дослідження властивостей тексту безмежні, а потреби вивчення різних одиниць тексту надзвичайно великі. Властивості й відмінні риси текстів можуть бути досліджені з

урахуванням різних параметрів і ознак: хронологічні, тематичні, жанрові, гендерні, соціально-статусні, вікові, лексичні, морфологічні, синтаксичні й інші кількісні характеристики тексту й уживаних у ньому одиниць.

4. Кількісна характеристика різних одиниць мови, вивчення довжини слова в різних текстах і різних мовах, вивчення довжини складів, морфем, речень.
5. Дослідження швидкості (темпів) і закономірностей розвитку й зміни мови - насамперед його лексичного складу.
6. Типологічне (порівняльне) вивчення різних мов і їхніх підсистем.
7. Квантитативне дослідження діалектології.
8. Квантитативний аналіз даних, отриманих за допомогою психолінгвістичних експериментів (асоціативний експеримент і семантичний диференціал Ч. Остуда).
9. Квантитативний аналіз семантичних і формальних відносин у реконструйованому за допомогою порівняльно-історичного методу лексичному складі прамови.
10. Дослідження семантики мови - парадигматичних і синтагматичних відносин у лексиці, синонімії, полісемії й інших явищах.
11. Перекладознавство. Є роботи, у яких кількісні методи використовуються при зіставленні двох мов при перекладі.
12. Методика викладання іноземних мов.

# 1. ІМОВІРНІСНІ ПРОСТОРИ

## 1.1. Класичне визначення ймовірності

*Достовірною* називається подія, що обов'язково відбудеться при здійсненні певного комплексу умов. Відповідно, *неможливою* називається подія, що при заданому комплексі умов не відбудеться ніколи. *Випадковою* називається така подія, що при заданому комплексі умов може як відбутися, так і не відбутися [1]. Міра можливості здійснення такої події і є її *ймовірність*. Достовірною й неможливою події можуть розглядатися як крайні окремі випадки випадкових подій.

Випадкові події будемо позначати великими латинськими буквами  $A, B, C, \dots$ . Достовірною подією позначається буквою  $\Omega$ , неможливою -  $\emptyset$ . Уведемо тепер деякі відносини між подіями. Дві події  $A$  і  $B$  *несумісні*, якщо настання одного з них виключає настання іншого. *Сума подій*  $A$  і  $B$  – це така третя подія  $C=A+B$ , що відбувається тоді, коли настає або подія  $A$ , або подія  $B$ , або вони обидві одночасно. *Добуток подій*  $A$  і  $B$  – це така третя подія  $C=AB$ , що настає тоді, коли відбуваються й подія  $A$ , і подія  $B$ . Подія  $\bar{A}$  *протилежна* події  $A$ , якщо вона несумісна з подією  $A$  і разом з нею утворює достовірну подію, тобто  $\bar{A}+A=\Omega$ .

Однією з моделей з скінченним числом результатів є класична імовірнісна схема. У цій схемі визначення ймовірності ґрунтується на рівній можливості кожного з скінченного числа результатів. Таке визначення виникло на основі перших спроб вирахування шансів в азартних іграх. Так, у випадку із гральною кісткою при однократному киданні однакова можливість випадання кожної із шести граней, на які нанесені цифри 1, 2, 3, 4, 5, 6. Позначимо ці однаково можливі результати або *елементарні події* через  $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6$ . Природно, що шанс здійснитися не одному результату, а одному із двох, наприклад, або  $\omega_1$ , або  $\omega_2$ , у два рази більше. Міркуючи таким чином, можна визначити шанси здійснення будь-якої складеної події, що складається з декількох елементарних.

У загальному випадку, коли є  $n$  однаково можливих елементарних подій  $\omega_1, \dots, \omega_n$ , *ймовірність* будь-якої складеної події  $A$ , що складається з  $m$  елементарних подій  $\omega_{i_1}, \dots, \omega_{i_m}$ , визначається як відношення кількості елементарних подій, що сприяють події  $A$ , до загальної кількості елементарних подій, тобто

$$P(A) = \frac{m}{n} \quad (1.1)$$

Наприклад, у випадку із гральною кісткою ймовірність події  $A$ , що полягає у випаданні парної кількості вічок (тобто  $A = \{\omega_2, \omega_4, \omega_6\}$ ), дорівнює  $P(A) = 3/6 = 1/2$ , тому що в подію  $A$  входять три елементарних події, а загальна кількість елементарних подій дорівнює шести.

Із класичного визначення ймовірностей, зокрема, випливає, що ймовірність повної події  $\Omega$ , що включає всі  $n$  елементарних подій, дорівнює  $P(\Omega) = n/n = 1$ . Але тоді повна подія  $\Omega$ , що складається в появі будь-якого із усього набору елементарних подій  $\omega_1, \omega_2, \dots, \omega_n$ , і є достовірною подією, тому що воно обов'язково відбувається. Тому ймовірність достовірної події дорівнює одиниці.

Якщо події розглядати як підмножини множини елементарних подій, то відносини між подіями, введені вище, можна інтерпретувати як співвідношення між множинами. Несумісні події – це такі події, які не містять спільних елементів. Сума  $A+B$  і добуток  $AB$  – це відповідно їхнє об'єднання  $A \cup B$  і перетинання  $A \cap B$ . Протилежна подія  $\bar{A}$  – це доповнення  $A$ . Запис  $A \subset B$  означає, що в  $B$  містяться всі елементарні події з  $A$ , і можуть міститися елементарні події, що не входять в  $A$ . Якщо  $A \subset B$  і  $B \subset A$ , то  $A=B$ .

У випадку класичного визначення ймовірності справедлива наступна теорема додавання ймовірностей.

**Теорема 1.1 (додавання ймовірностей).** *Формулювання.* Якщо дві складені події  $A = \{\omega_{i_1}, \dots, \omega_{i_m}\}$  і  $B = \{\omega_{j_1}, \dots, \omega_{j_k}\}$  є несумісними, то ймовірність об'єднаної події  $C = A \cup B$  дорівнює сумі ймовірностей цих двох подій.

Подія  $\bar{A}$  називається *протилежною* стосовно  $A$ , якщо до неї входять всі елементарні події,

що не входять в  $A$ . Іншими словами,  $A$  і  $\bar{A}$  - це такі несумісні події, які разом утворюють достовірну подію, тобто  $A \cup \bar{A} = \Omega$ . З теореми додавання випливає, що  $P(\Omega) = P(A) + P(\bar{A}) = 1$ , тому  $P(\bar{A}) = 1 - P(A)$ . Звідси, зокрема, випливає, що ймовірність неможливої події  $\emptyset$ , що є протилежною стосовно достовірної події  $\Omega$ , дорівнює нулю.

## 1.2. Урна схема

Класична схема, незважаючи на всю свою обмеженість, придатна для вирішення ряду суцільно практичних задач. Розглянемо, наприклад, деяку сукупність елементів об'єму  $N$ . Це можуть бути вироби, кожне з яких є придатним або бракованим. Подібного роду ситуації описуються урною схемою: в урні є  $N$  куль, з них  $M$  білих,  $(N - M)$  чорних.

Наприклад, уявимо собі, що є тільки руйнуючі засоби контролю кожного виробу на придатність (наприклад, сірника). У такому випадку не можна обстежити всю партію виробів, а тільки частину її. Отже, з урни, що містить  $N$  куль, у якій перебуває невідома кількість  $M$  білих куль, витягається вибірка об'єму  $n$ . Така процедура називається *вибіркою без повернення*. Необхідно визначити ймовірність того, що у вибірці буде виявлено  $m$  білих куль. Це задача на застосування класичного визначення ймовірності. Справді, в описаній ситуації кожна вибірка не має переваги стосовно будь-якої іншої, тобто всі вони однаково можливі. Підрахуємо кількість всіх можливих вибірок об'єму  $n$  з  $N$  елементів. Як відомо з комбінаторики, кількість способів, за допомогою яких можна вибрати  $n$  елементів із загальної їхньої кількості  $N$ , дорівнює числу сполучень із  $N$  по  $n$ , тобто  $C_N^n = \frac{N!}{n!(N-n)!}$ , де  $N! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot N$ . Таким чином, загальна кількість результатів дорівнює  $C_N^n$ .

З'ясуємо, скільки результатів із загальної кількості елементарних результатів сприяє події  $A$ , тобто наявності у вибірці об'єму  $n$  білих куль у кількості  $m$ . Кількість способів, якими можна з  $M$  білих куль витягти  $m$  штук, дорівнює  $C_M^m$ , а кількість способів вибрати з  $(N - M)$  чорних куль  $(n - m)$  штук дорівнює  $C_{N-M}^{n-m}$ . Тому кількість результатів, сприятливих події  $A$ , дорівнює  $C_M^m \cdot C_{N-M}^{n-m}$  і, отже, її ймовірність, що дорівнює відношенню кількості сприятливих результатів до їхньої загальної кількості, така:

$$P(A) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} = \frac{C_n^m C_{N-n}^{n-m}}{C_N^m} = P_{M,N}(m, n). \quad (1.2)$$

**Приклад 1.1.** Нехай є партія, що складається з 500 виробів, у якій два бракованих. Яка ймовірність у вибірці з 5 виробів не виявити жодного бракованого?

Рішення. Скористаємося формулою (1.2). Маємо

$$P_{498,500}(5, 5) = \frac{C_{498}^5 C_2^0}{C_{500}^5} = \frac{498! \cdot 2!}{493! \cdot 5! \cdot 0! \cdot 2!} = \frac{500!}{495! \cdot 5!} = 0.98$$

Розглянемо тепер, наприклад, урну з кулями, вибірка куль із якої відбувається послідовно по одній кулі, і при цьому щораз фіксується номер кулі, а сама куля повертається знову до урни. Така процедура називається *вибіркою з поверненням* [2]. У цьому випадку ймовірність події  $A$ , обчислена аналогічним способом, дорівнює

$$P(A) = C_n^m \frac{M^m (N - M)^{n-m}}{N^n} = C_n^m \left( \frac{M}{N} \right)^m \left( 1 - \frac{M}{N} \right)^{n-m}.$$

Говорити про ймовірності як про міри можливості здійснення випадкової події  $A$  має сенс тільки при здійсненні певного комплексу умов. При зміні умов зміниться й ймовірність. Так, якщо до комплексу умов при якому вивчалася ймовірність  $P(A)$ , додати нову умову, що полягає в появі події  $B$ , то одержимо інше значення ймовірності  $P(A/B)$  – умовну ймовірність події  $A$  за умови, що відбулася подія  $B$ . Ймовірність  $P(A)$ , на відміну від умовної, називається *безумовною*.

Виведемо *формулу умовної ймовірності*. Нехай подіям  $A$  і  $B$  сприяють  $m$  і  $k$  елементарних результатів з  $n$ . Тоді, згідно (1.1), їхні безумовні ймовірності дорівнюють  $\frac{m}{n}$  й  $\frac{k}{n}$  відповідно. Нехай подія  $A$  за умови, що подія  $B$  відбулася, сприяє  $r$  елементарних результатів. Тоді, згідно (1.1), умовна ймовірність події  $A$  дорівнює  $P(A/B) = \frac{r}{k}$ . Розділивши й чисельник, і знаменник на  $n$ , одержимо формулу умовної ймовірності:

$$P(A/B) = \frac{r/n}{k/n} = \frac{P(A \cap B)}{P(B)}, \quad (1.3)$$

оскільки подія  $A \cap B$  відповідає  $r$  результатів і, отже,  $\frac{r}{n}$  - його безумовна ймовірність.

Подія  $A$  називається *незалежною від  $B$* , якщо її умовна ймовірність дорівнює безумовній, тобто  $P(A/B) = P(A)$ . При цьому з формули (1.3) одержуємо

$$P(A \cap B) = P(A) \cdot P(B), \quad (1.4)$$

тобто властивість незалежності взаємна й для незалежних подій імовірність їхнього одночасного здійснення дорівнює добутку їхніх ймовірностей. Формула (1.3), записана у вигляді

$$P(A \cap B) = P(A) \cdot P(B), \quad (1.5)$$

називається *формулою множення для залежних подій*, а формула (1.4) – *теоремою множення для незалежних подій*.

Наприклад, у експерименті із гральною кісткою нехай подія  $A$  полягає у випаданні числа вічків, що ділиться на три, тобто  $A = \{\omega_3, \omega_6, \dots\}$ , а подія  $B$  – у випаданні парного числа вічків, тобто  $B = \{\omega_2, \omega_4, \omega_6, \dots\}$ . Тоді  $A \cap B = \omega_6$  і по формулі умовної ймовірності (1.3) одержуємо, що

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}.$$

Але  $P(A) = 2/6 = 1/3$ . Тому  $P(A/B) = P(A)$ , тобто події  $A$  і  $B$  незалежні.

Взаємність незалежності подій означає, що якщо подія  $A$  не залежить від події  $B$ , то й подія  $B$  не залежить від події  $A$ . Тоді при  $P(A) > 0$  і з огляду на формулу (1.5) маємо:

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A/B) \cdot P(B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B)$$

### 1.3. Скінченна схема з неоднаково можливими результатами

Обмеженість класичного визначення ймовірності, зокрема, закладена в однаковій можливості результатів. Дійсно, навіть невелике ускладнення практичної ситуації негайно ввійде в суперечність із однаковою можливістю, що може розглядатися, скоріше, як окремий випадок більш загальної ситуації.

Розглянемо, наприклад, стрілянину по круговій мішені. Елементарними результатами тут є влучення в те або інше кільце кругової мішені. Влучення в мале внутрішнє коло оцінюється в 10 вічок, у навколишнє його кільце – в 9 вічок, у наступне – в 8 вічок й т.д., у саме зовнішнє кільце – 1 вічко, невлучення в кругову мішень – 0 вічка. Таким чином, є 11 елементарних подій  $\omega_{10}, \omega_9, \dots, \omega_1, \omega_0$ . Для кожного стрільця певного класу є свої певні стійкі шанси (імовірності) вибити за один постріл ту або іншу кількість вічок  $p_{10}, p_9, \dots, p_1, p_0$ . Ці події, загалом кажучи, неоднаково можливі. Наприклад, для майстрів спорту, очевидно, виключена подія  $\omega_0$ , тому  $p_0 = 0$ , тобто відразу виключається однакова можливість.

Скінченна схема з неоднаково можливими результатами визначається в такий спосіб. Є скін-



ченний набір елементарних подій  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ , і для кожної елементарної події  $\omega_i$  задана його ймовірність  $p_i$ ,  $0 \leq p_i \leq 1$ , причому  $\sum_{i=1}^n p_i = 1$ . Ймовірність будь-якої складеної події  $A = \{\omega_{i_1}, \dots, \omega_{i_m}\}$  визначається як сума ймовірностей вхідних у нього елементарних подій:

$$P(A) = \sum_{l=1}^m p_{i_l}. \quad (1.6)$$

Ця схема є узагальненням класичної схеми. Справді, якщо повернутися до випадку однакової можливості й приписати кожній елементарній події ймовірність  $\frac{1}{n}$ , то формула (1.6) приводить до класичного визначення ймовірності.

У випадку скінченної схеми також має місце теорема додавання.

**Теорема 1.2. Формулювання.** Для двох несумісних подій  $A$  і  $B$ , що є підмножинами  $\Omega$ ,  $P(A \cup B) = P(A) + P(B)$ .

Однак, як відомо, не завжди події є несумісними. Вони також можуть бути залежними. Тому в загальному випадку для будь-яких (як несумісних, так і залежних) подій  $A$  і  $B$  має місце наступна формула додавання:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Точно так само, як скінченна схема з неоднаково можливими результатами є узагальненням класичної скінченної схеми з однаково можливими результатами, дискретна схема з нескінченною кількістю неоднаково можливих подій, у свою чергу, є узагальненням скінченної схеми.

У дискретній схемі множина  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\}$ , загалом кажучи, містить зліченну кількість елементарних подій. Для кожної елементарної події задана її ймовірність  $p_i = P(\omega_i)$ ,  $0 \leq p_i \leq 1$ , причому  $\sum_{i=1}^{\infty} p_i = 1$ . Ймовірність будь-якої скінченної або зліченної підмножини  $A \subset \Omega$  множини елементарних подій  $\Omega$  дорівнює сумі ймовірностей елементарних подій, що її складають, тобто якщо  $A = \bigcup_{l=1}^{\infty} \omega_{i_l}$ , то  $P(A) = \sum_{l=1}^{\infty} p_{i_l}$ . Якщо ж  $A = \bigcup_{l=1}^m \omega_{i_l}$ , то має місце (1.6).

У скінченній схемі, як і в класичній, можна вивести формулу умовної ймовірності. Розглянемо події  $A = \{\omega_{i_1}, \dots, \omega_{i_m}\}$  і  $B = \{\omega_{j_1}, \dots, \omega_{j_l}, \dots, \omega_{j_k}\}$  такі, що  $\omega_{i_1} = \omega_{j_1}, \dots, \omega_{i_l} = \omega_{j_l}$ ,  $l \leq m, k$ . Інакше кажучи,  $A \cap B = \{\omega_{i_1}, \dots, \omega_{i_l}\}$ . Тоді

$$P(A) = \sum_{\mu=1}^m p_{i_{\mu}}, \quad P(B) = \sum_{q=1}^k p_{j_q} > 0, \quad P(A \cap B) = \sum_{\mu=1}^l p_{i_{\mu}}.$$

Нехай подія  $B$  відбулася. Тому має місце нова скінченна схема з  $k$  результатами,  $k \leq n$ , отже, сума ймовірностей повного набору цих нових результатів повинна дорівнювати одиниці, а вона, відповідно до первісної схеми, дорівнює  $P(B) = \sum_{q=1}^k p_{j_q}$ .

Щоб забезпечити рівність суми ймовірностей елементарних подій одиниці, уведемо нові ймовірності результатів:

$$\tilde{p}_{j_q} = \frac{p_{j_q}}{P(B)}, \quad \sum_{q=1}^k \tilde{p}_{j_q} = \frac{\sum_{q=1}^k p_{j_q}}{P(B)} = 1.$$

У рамках нової схеми (тобто за умови, що відбулася подія  $B$ ) визначаємо ймовірність події  $A$ :

$$P(A/B) = \sum_{q=1}^k \tilde{p}_{j_q} = \frac{\sum_{q=1}^l p_{j_q}}{P(B)} = \frac{P(A \cap B)}{P(B)}.$$

Таким чином, ми знову одержуємо ту ж формулу умовної ймовірності, що й у класичній схемі. Незалежність подій визначається аналогічно класичній схемі. Розглянемо найпростіші приклади схеми послідовних випробувань як ілюстрацію скінченної схеми з неоднаково можливими результатами.

**Приклад 1.2.** Система контролю виробів складається із двох незалежних перевірок, виконуваних одночасно. Виріб приймається, якщо він пройшло обидві перевірки. У результаті кожної перевірки бракований виріб приймається з ймовірностями  $\alpha_1, \alpha_2$  відповідно. Знайти ймовірність того, що буде прийнятий бракований виріб.

*Рішення.* Якщо на вхід системи контролю надійшов бракований виріб, то можливі наступні чотири елементарних результати:  $\omega_1 = \{0, 0\}$ ,  $\omega_2 = \{0, 1\}$ ,  $\omega_3 = \{1, 0\}$ ,  $\omega_4 = \{1, 1\}$ , де 0 означає, що виріб визнаний бракованим, а 1 – що виріб визнаний придатним. Випробування незалежні, тому одержуємо наступні значення ймовірностей елементарних результатів  $\omega = \{i_1, i_2\}$ :

$$\begin{aligned} p_1 &= p(\omega_1) = P\{i_1=0, i_2=0\} = P\{i_1=0\} \cdot P\{i_2=0\} = (1 - \alpha_1)(1 - \alpha_2), \\ p_2 &= p(\omega_2) = P\{i_1=0, i_2=1\} = P\{i_1=0\} \cdot P\{i_2=1\} = (1 - \alpha_1)\alpha_2, \\ p_3 &= p(\omega_3) = P\{i_1=1, i_2=0\} = P\{i_1=1\} \cdot P\{i_2=0\} = \alpha_1(1 - \alpha_2), \\ p_4 &= p(\omega_4) = P\{i_1=1, i_2=1\} = P\{i_1=1\} \cdot P\{i_2=1\} = \alpha_1\alpha_2. \end{aligned}$$

Сума ймовірностей елементарних подій повинна дорівнювати одиниці. Дійсно,

$$P(\Omega) = \sum_{l=1}^4 p(\omega_l) = (1 - \alpha_1)(1 - \alpha_2) + (1 - \alpha_1)\alpha_2 + \alpha_1(1 - \alpha_2) + \alpha_1\alpha_2 = 1.$$

Відповідно до умов задачі й сформованій схемі ймовірність прийняти бракований виріб – це ймовірність елементарної події  $\omega_4$ , що полягає в тому, що й перша, і друга перевірки визнають бракований виріб придатним. Тому шукана ймовірність дорівнює  $p_4 = \alpha_1\alpha_2$ .

**Приклад 1.3.** В умовах приклада 1.2 задані ймовірності  $\beta_1, \beta_2$  відбракувати придатний виріб у результаті першої й другої перевірок відповідно. Знайти ймовірність того, що буде відбракований придатний виріб.

*Рішення.* Якщо на вхід системи контролю надійшло придатний виріб, то можливі ті ж самі чотири елементарних результати, однак їхні ймовірності будуть іншими. Знову скористаємося незалежністю випробувань, тоді одержимо наступні ймовірності елементарних результатів:

$$\begin{aligned} \tilde{p}_1 &= p(\omega_1) = P\{i_1=0, i_2=0\} = P\{i_1=0\} \cdot P\{i_2=0\} = \beta_1\beta_2, \\ \tilde{p}_2 &= p(\omega_2) = P\{i_1=0, i_2=1\} = P\{i_1=0\} \cdot P\{i_2=1\} = \beta_1(1 - \beta_2), \\ \tilde{p}_3 &= p(\omega_3) = P\{i_1=1, i_2=0\} = P\{i_1=1\} \cdot P\{i_2=0\} = (1 - \beta_1)\beta_2, \\ \tilde{p}_4 &= p(\omega_4) = P\{i_1=1, i_2=1\} = P\{i_1=1\} \cdot P\{i_2=1\} = (1 - \beta_1)(1 - \beta_2). \end{aligned}$$

Подія, що полягає в тому, що відбракованим є придатний виріб, містить у собі елементарні події  $\omega_1, \omega_2, \omega_3$ . Тому шукана ймовірність дорівнює

$$\tilde{p}_1 + \tilde{p}_2 + \tilde{p}_3 = \beta_1\beta_2 + \beta_1(1 - \beta_2) + (1 - \beta_1)\beta_2 = \beta_1 + \beta_2 - \beta_1\beta_2.$$

#### 1.4. Контрольні запитання

1. Що таке випадкова подія?
2. Яка подія є достовірною, а яка неможливою?
3. Які події є несумісними?
4. Що таке добуток подій?
5. Як визначається ймовірність складеної події?

6. Сформулюйте теорему додавання ймовірностей.
7. В чому полягає урнова схема?
8. Які події є незалежними?
9. Наведіть формулу умовної ймовірності.
10. Сформулюйте теорему множення для незалежних подій.
11. У чому полягає скінченна схема з неоднаково можливими результатами?
12. Перелічить основні дії над подіями та їх властивості.
13. Що таке поле подій?
14. Сформулюйте аксіоми теорії ймовірностей.
15. Що таке імовірнісний простір?
16. Що таке розподіл ймовірностей?

## 2. УМОВНІ ЙМОВІРНОСТІ

### 2.1. Повна ймовірність

Події  $A_1, A_2, \dots, A_n$  утворюють *повну групу подій*, якщо вони попарно несумісні й разом утворюють достовірну подію, тобто  $A_i \cap A_j = \emptyset, i \neq j, \bigcup_{i=1}^n A_i = \Omega$ .

**Теорема 2.1 (про формулу повної ймовірності).** Формулювання. Якщо події  $A_1, A_2, \dots, A_n, P(A_i) > 0$ , утворюють повну групу подій, то ймовірність події  $B$  може бути подана як сума добутків безумовних ймовірностей подій повної групи на умовні ймовірності події  $B$ :

$$P(B) = \sum_{i=1}^n P(A_i) \cdot P(B/A_i). \quad (2.1)$$

Вимога, що полягає в тому, що події  $A_i$  утворюють повну групу подій, може бути замінена більш слабкою: події  $A_i$  попарно не перетинаються,  $B \in \bigcup_{i=1}^n A_i$ . Крім того, на основі аксіоми 3' (аксіоми зліченної адитивності) теорему повної ймовірності можна поширити й на зліченну множину попарно непересічних подій  $A_i, P(A_i) > 0, B \in \bigcup_{i=1}^{\infty} A_i$ :

$$P(B) = \sum_{i=1}^{\infty} P(A_i) \cdot P(B/A_i).$$

**Приклад 2.1.** У трьох партіях деталей, що надійшли на склад, відсоток придатних становить відповідно 89%, 92% і 97%, а загальна кількість деталей у партіях співвідноситься як 1:2:3. Необхідно визначити ймовірність випадкового вибору непридатної деталі із всіх трьох партій.

Рішення. Позначимо через  $A_1, A_2, A_3$  події, що полягають в тому, що обрана навмання деталь належить відповідно до першої, другої й третьої партій. Так як ці події утворюють повну групу подій, то  $P(A_1) + P(A_2) + P(A_3) = 1$ . За умовою  $P(A_1) : P(A_2) : P(A_3) = 1 : 2 : 3$ , отже,  $P(A_1) = \frac{1}{6}, P(A_2) = \frac{1}{3},$

$P(A_3) = \frac{1}{2}$ . Розглядаючи вибір непридатної деталі як випадкову подію  $B$ , ймовірності такої події за умови, що деталь вибирається з першої, другої й третьої партій, відповідно мають значення:  $P(B/A_1) = 0.11, P(B/A_2) = 0.08, P(B/A_3) = 0.03$ . Ймовірність випадкового вибору непридатної деталі із всіх трьох партій визначається по формулі повної ймовірності:

$$P(B) = P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) + P(A_3) \cdot P(B/A_3) = \frac{1}{6} \cdot 0.11 + \frac{1}{3} \cdot 0.08 + \frac{1}{2} \cdot 0.03 = 0.06.$$

### 2.2. Формула Байєса

На основі комутативності операції перетинання множин  $A \cap B = B \cap A$  можна записати  $P(A \cap B) = P(B \cap A)$  або  $P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$  [3]. Це співвідношення справедливо й для випадку, коли під  $A$  розуміється деяка подія  $A_k$  з повної групи подій  $A_1, A_2, \dots, A_n$ , тобто  $P(A_k) \cdot P(B/A_k) = P(B) \cdot P(A_k/B)$ , звідки

$$P(A_k/B) = P(A_k) \frac{P(B/A_k)}{P(B)}.$$

Підставляючи сюди  $P(B)$  по формулі повної ймовірності (2.1) одержуємо *формулу Байєса*:

$$P(A_k/B) = \frac{P(A_k) \cdot P(B/A_k)}{\sum_{i=1}^n P(A_i) \cdot P(B/A_i)}. \quad (2.2)$$

По цій формулі можна обчислити ймовірності подій  $A_i$ ,  $i = \overline{1, n}$ , за умови, що відбулася подія  $B$ , якщо відомі ймовірності  $P(A_i)$  і  $P(B/A_i)$ . Очевидно, що одержувані при цьому умовні ймовірності  $P(A_i/B)$  задовольняють співвідношенню  $\sum_{i=1}^n P(A_i/B) = 1$ .

Ймовірності  $P(A_i)$  подій  $A_i$  називають *апостеріорними ймовірностями*, тобто ймовірностями подій до виконання експерименту, а умовні ймовірності цих подій  $P(A_i/B)$  – *апостеріорними*, тобто уточненими в результаті експерименту, результатом якого послужила поява події  $B$  [1, 2].

**Приклад 2.2.** На підприємстві виготовляють вироби певного виду на трьох потокових лініях. На першій лінії виробляється 20% виробів від усього об'єму їхнього виробництва, на другій - 30%, на третьої - 50%. Кожна з ліній характеризується відповідно наступними відсотками придатності виробів: 95%, 98% і 97%. Потрібно визначити ймовірність того, що навмання взятий виріб, випущений підприємством, виявиться бракованим, а також ймовірності того, що цей бракований виріб зроблений на першій, другій і третій лініях.

*Рішення.* Позначимо через  $A_1, A_2, A_3$  події, що полягають в тому, що навмання взятий виріб зроблений на першій, другій і третій лініях. Відповідно до умов задачі,  $P(A_1)=0.2$ ,  $P(A_2)=0.3$ ,  $P(A_3)=0.5$  і ці події утворюють повну групу подій, тому що вони попарно несумісні й  $P(A_1)+P(A_2)+P(A_3)=1$ . Позначимо через  $B$  подію, що полягає в тому, що навмання взятий виріб виявився бракованим. Відповідно до умов задачі,  $P(B/A_1)=0.05$ ,  $P(B/A_2)=0.02$ ,  $P(B/A_3)=0.03$ . Використовуючи формулу повної ймовірності, одержуємо

$$P(B)=P(A_1) \cdot P(B/A_1)+P(A_2) \cdot P(B/A_2)+P(A_3) \cdot P(B/A_3)=0.05 \cdot 0.2+0.02 \cdot 0.3+0.03 \cdot 0.5=0.031,$$

тобто ймовірність того, що навмання взятий виріб виявиться бракованим, дорівнює 3.1%.

Апостеріорні ймовірності того, що навмання взятий виріб виготовлений відповідно на першій, другій і третій лініях, рівні 0.2, 0.3 і 0.5. Потім був виконаний експеримент, у результаті якого навмання взятий виріб виявився бракованим. Визначимо тепер апостеріорні ймовірності того, що цей виріб виготовлений на першій, другій і третій лініях. По формулі Байєса маємо:

$$P(A_1/B)=\frac{P(A_1) \cdot P(B/A_1)}{\sum_{i=1}^3 P(A_i) \cdot P(B/A_i)}=\frac{0.05 \cdot 0.2}{0.031}=\frac{10}{31},$$

$$P(A_2/B)=\frac{P(A_2) \cdot P(B/A_2)}{\sum_{i=1}^3 P(A_i) \cdot P(B/A_i)}=\frac{0.02 \cdot 0.3}{0.031}=\frac{6}{31},$$

$$P(A_3/B)=\frac{P(A_3) \cdot P(B/A_3)}{\sum_{i=1}^3 P(A_i) \cdot P(B/A_i)}=\frac{0.03 \cdot 0.5}{0.031}=\frac{15}{31}.$$

Таким чином, ймовірності того, що навмання взятий виріб, що виявився бракованим, виготовлено на першій, другій і третій лініях, відповідно рівні 0.322, 0.194 і 0.484.

Формула множення ймовірностей (1.5) може бути поширена на випадок довільної скінченної кількості подій  $P(A_1 \cap A_2 \cap \dots \cap A_n)=P(A_1) \cdot P(A_2/A_1) \cdot \dots \cdot P(A_n/A_1 \cap A_2 \cap \dots \cap A_{n-1})$ .

Події  $A_1, A_2, \dots, A_n$  *незалежні в сукупності*, якщо для будь-якої їхньої підмножини

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k})=P(A_{i_1}) \cdot P(A_{i_2}) \cdot \dots \cdot P(A_{i_k}).$$

Якщо ця умова виконується тільки для  $k=2$ , то такі події називаються *попарно незалежними*. З незалежності в сукупності випливає попарна незалежність, а з попарної незалежності не випливає незалежність у сукупності.

### 2.3. Послідовності випробувань

Нехай проводиться скінченне число  $n$  послідовних незалежних випробувань, у кожному з яких може відбутися певна подія – успіх – або наступить протилежна подія – невдача. Така послідовність випробувань називається *схемою Бернуллі*. У схемі Бернуллі одному випробуванню відповідає множина елементарних результатів, що складається із двох елементарних подій:  $\{\omega_0, \omega_1\}$ ,  $\omega_0$  – невдача,  $\omega_1$  – успіх, при цьому  $A = \{\omega_1\}$ ,  $\bar{A} = \{\omega_0\}$ . Множина елементарних результатів для  $n$  випробувань складається вже з  $2^n$  елементарних подій  $\omega_{i_1, i_2, \dots, i_n} = \{i_1, \dots, i_n\}$ , кожне з яких відповідає конкретному результату випробувань, при цьому набір індексів  $i_1, \dots, i_n$  являє собою конкретну послідовність нулів і одиниць, що відповідає результатам випробувань на кожному кроці.

Якщо задані ймовірності успіху й невдачі в окремому випробуванні  $p_1 = p, p_0 = 1 - p = q$ , то можна визначити ймовірність будь-якого елементарного результату в  $n$  випробуваннях. Дійсно, розглянемо будь-який елементарний результат  $\omega_{i_1, i_2, \dots, i_n}$ , при цьому  $(i_1, \dots, i_n)$  – конкретна послідовність нулів і одиниць, що відповідає послідовності невдач або успіхів у кожному з  $n$  індивідуальних випробувань, наприклад  $(A_{i_1}, A_{i_2}, \dots, A_{i_n})$ . Тоді з незалежності друг від друга результатів окремих випробувань одержуємо

$$P\{\omega_{i_1, i_2, \dots, i_n}\} = P(\omega_{i_1})P(\omega_{i_2}) \dots P(\omega_{i_n}) = p_{i_1} p_{i_2} \dots p_{i_n} = pq \dots p.$$

Таким чином, якщо загальний елементарний результат включає  $m$  успіхів і  $n - m$  невдач, то його ймовірність дорівнює

$$P(\omega_{i_1, i_2, \dots, i_n}) = p^m q^{n-m} \quad (2.3)$$

І, отже, по аксіомі додавання ймовірностей може бути визначена ймовірність будь-якої події, що складається з декількох елементарних подій. Зокрема, якщо нас цікавить ймовірність  $P_n(m)$  того, що в  $n$  випробуваннях відбулося  $m$  успіхів, те її визначаємо як суму ймовірностей елементарних подій, що характеризуються  $m$  успіхами. Ймовірність такого елементарного результату, згідно (2.3), дорівнює  $p^m q^{n-m}$ . Отже, для знаходження ймовірності  $P_n(m)$  треба визначити кількість елементарних подій, що характеризуються  $m$  успіхами, тобто встановити, скількома способами можуть бути на  $n$  місць розставлені  $m$  одиниць (інші  $n - m$  місць займаються нулями). Але це аналогічно тому, що з  $n$  елементів треба вибрати (позначити)  $m$  елементів. Кількість таких вибірок, як відомо, дорівнює кількості сполучень із  $n$  по  $m$ , тобто  $C_n^m$ . Остаточню одержуємо

$$P_n(m) = C_n^m p^m q^{n-m}. \quad (2.4)$$

Сума біноміальних ймовірностей, що вийшли, дорівнює одиниці:

$$\sum_{m=0}^n P_n(m) = \sum_{m=0}^n C_n^m p^m q^{n-m} = (p+q)^n = 1^n = 1.$$

У випадку урнної схеми можна уявити собі, що здійснюється вибірка об'єму  $n$  з урни не відразу, а послідовно куля за кулею. У результаті приходимо до схеми послідовних випробувань, однак на відміну від схеми Бернуллі тут результати наступних випробувань уже залежать від результатів попередніх. Так, якщо ймовірність на першому кроці витягти білу кулю дорівнює  $\frac{M}{N}$ , то

умовна ймовірність витягти білу кулю на другому кроці дорівнює  $\frac{M-1}{N-1}$ , якщо на першому кроці

витягнута біла куля, і  $\frac{M}{N-1}$ , якщо на першому кроці витягнута чорна куля. Але у випадку, коли

генеральна сукупність велика, тобто  $N \rightarrow \infty$ , урнову схему можна замінити схемою Бернуллі. На практиці це означає, що при об'ємі вибірки, істотно меншому об'єму генеральної сукупності, можна замість ймовірностей урнної схеми приблизно використати відповідні ймовірності схеми Бернуллі., тобто при  $n \ll N$

$$P_{M,N}(m, n) \approx C_n^m \left( \frac{M}{N} \right)^m \left( \frac{N-M}{N} \right)^{n-m}.$$

#### 2.4. Контрольні запитання

1. Що таке повна група подій?
2. Сформулюйте теорему про формулу повної ймовірності.
3. Наведіть формулу Байєса.
4. Які ймовірності називають апіорними, а які апостеріорними?
5. Які події є незалежними у сукупності?
6. Яка послідовність випробувань називається схемою Бернуллі?

### 3. ЧИСЛОВІ ХАРАКТЕРИСТИКИ ВИПАДКОВОЇ ВЕЛИЧИНИ

#### 3.1. Вибіркове середнє й вибіркова дисперсія

Для вивчення випадкової величини, що характеризує деяке явище (наприклад, мелодія фрази, наголос, темп проголошення й т.д.), проводиться експеримент. Кожне випробування (результат) експерименту доставляє дослідникові певне значення випадкової величини. Сукупність значень випадкової величини, отриманих у ході експерименту, називається *випадковою вибіркою*. Нехай випадкова величина  $X$  у ході експерименту, що складається з  $n$  випробувань, прийняла ряд значень  $x_1, x_2, \dots, x_n$  (не обов'язково всі  $x$  різні між собою). Тоді говорять, що  $(x_1, x_2, \dots, x_n)$  є випадковою вибіркою для  $X$  об'єму  $n$ . Кожна випадкова вибірка несе в собі деяку якісну й кількісну інформацію про досліджуване явище. Щоб зробити цю інформацію більш компактною й доступною для огляду, використовуються деякі числові характеристики випадкової вибірки, які, природно, розглядаються як наближення до відповідних характеристик самої випадкової величини  $X$ . До таких числових характеристик належать вибіркове середнє арифметичне значення й вибіркова дисперсія. Іноді слово «вбіркове» опускають.

*Вибіркове середнє арифметичне* значення – це число, навколо якого розташовуються (по обидві сторони) всі значення вибірки. У загальному виді для вибірки  $(x_1, x_2, \dots, x_n)$  маємо:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Щоб визначити характер розташування вибіркових значень щодо свого середнього арифметичного значення  $\bar{x}$ , обчислюється *вбіркова дисперсія*. Її звичайно позначають через  $S^2$ . Для її одержання необхідно від кожного вибіркового значення відняти середнє арифметичне значення, результати піднести до квадрата, а потім скласти й отриману суму розділити на  $(n - 1)$ , де  $n$  – об'єм вибірки, тобто

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

Зміст вибіркової дисперсії як міри розсіювання (розкиду) полягає в наступному: якщо є дві випадкові вибірки того самого об'єму  $n$  (або приблизно однакових об'ємів) і з тим самим середнім арифметичним значенням  $\bar{x}$ , то та вибірка, у якої  $S^2$  менше, має значення, більш компактно розташовані відносно  $\bar{x}$ . Як правило, така вибірка виникає при вивченні більш однорідного матеріалу.

Корінь квадратний з вибіркової дисперсії, тобто  $S$ , називається *вбірковим середнім квадратичним відхиленням*.

**Приклад 3.1.** Нехай метою експерименту є вивчення тривалості наголошених голосних в англійській мові, обмірюваної в мсек. У результаті десяти випробувань була отримана випадкова вибірка: 206, 110, 136, 150, 200, 164, 170, 178, 140, 166. Очевидно, що об'єм вибірки дорівнює 10. Щоб одержати середнє арифметичне значення для цієї вибірки, необхідно всі знайдені числа скласти й суму розділити на 10. Таким чином, позначаючи вибіркове середнє через  $\bar{x}$ , будемо мати

$$\bar{x} = \frac{206 + 110 + 136 + 150 + 200 + 164 + 170 + 178 + 140 + 166}{10} = \frac{1620}{10} = 162 \text{ (мсек)}.$$

$$S^2 = \frac{1}{9} ((206-162)^2 + (110-162)^2 + (136-162)^2 + (150-162)^2 + (200-162)^2 + (164-162)^2 + (170-162)^2 + (178-162)^2 + (140-162)^2 + (166-162)^2) = \frac{7728}{9} = 858.67.$$

$$S = \sqrt{858.67} = 29.3.$$



Розглянутий приклад показує, що визначення  $\bar{x}$  й  $S^2$  пов'язане з великою кількістю обчислень, що збільшуються разом з об'ємом вибірки. Однак процес обчислень може бути істотно спрощений без особливої втрати точності. Для цього проводять *угруповання* вибірових значень. Нехай є вибірка об'єму  $n$  ( $x_1, x_2, \dots, x_n$ ). Можна вважати, що числа  $x_1, x_2, \dots, x_n$  розташовані в порядку зростання (цього завжди можна домогтися). Вибираємо деяке ціле число  $k$ , таке, що  $10 \leq k \leq 20$ , складаємо різницю  $x_n - x_1$  і визначаємо число  $h$  по формулі

$$h = \frac{x_n - x_1}{k}.$$

Тепер можна вказати інтервали групування:  $(x_1, x_1+h)$ ,  $(x_1+h, x_1+2h)$ , ...,  $(x_1+(n-1)h, x_n)$ . Кожне зі значень нашої випадкової вибірки  $x_1, x_2, \dots, x_n$  попадає в один із зазначених вище інтервалів. Так, якщо деяке вибірове значення  $x_l$  задовольняє нерівності  $x_1+lh \leq x_l < x_1+(l+1)h$ , то воно належить до інтервалу  $(x_1+lh, x_1+(l+1)h)$ . Позначимо через  $n_i$  кількість значень вибірки, що потрапили в інтервал  $(x_1+(i-1)h, x_1+ih)$ . Якщо деяке вибірове значення потрапило на границю двох інтервалів, то воно належить до правого інтервалу. Числа  $n_i$  називаються *частотами відповідних інтервалів*. Тепер для визначення  $\bar{x}$  й  $S^2$  діють по наступному алгоритму:

1. Фіксується якийсь із побудованих інтервалів (звичайно беруть інтервал з максимальним значенням частоти  $n_i$ ).
2. Обчислюється середина зафіксованого інтервалу. Нехай це буде  $x_0$ .
3. Всі інтервали нумеруються в зростаючому порядку від 1 до  $k$  (у нас рівно  $k$  інтервалів).
4. З номера кожного інтервалу віднімається номер зафіксованого інтервалу (ці різниці з відповідними знаками позначаються через  $c_i = N_i - N_0$ ).
5. Далі користуються формулами:

$$\bar{x} = x_0 + \frac{1}{n} \left( \sum_{i=1}^k n_i c_i \right) \cdot h.$$

$$S^2 = \left[ \sum_{i=1}^k n_i c_i^2 - \frac{1}{n} \left( \sum_{i=1}^k n_i c_i \right)^2 \right] \cdot h^2 \cdot \frac{1}{n-1}.$$

Тут  $n_i c_i$  – добуток частоти  $n_i$  інтервалу на відповідне йому число  $c_i$ ,  $n$  – об'єм вибірки.

Природно очікувати, що інтервальна розбивка приводить до деяких помилок в обчисленні вибірового середнього й вибірової дисперсії. Щоб нівелювати цю погрішність, можна внести виправлення на групування в  $\bar{x}$  і  $S^2$ . Виявляється, що виправлень потребує тільки величина  $S^2$ . З урахуванням виправлення (її вираження зазначене Шеппардом і дорівнює  $-\frac{h^2}{12}$ ) маємо вираження для  $S^2$ :

$$S^2 = \left[ \sum_{i=1}^k n_i c_i^2 - \frac{1}{n} \left( \sum_{i=1}^k n_i c_i \right)^2 \right] \cdot \frac{h^2}{n-1} - \frac{h^2}{12}.$$

Звичайно виправлення Шеппарда становить менш 5% від виправленого  $S^2$ , а тому її варто вносити тільки для вибірок досить великого об'єму, коли є впевненість, що вибірова дисперсія добре наближає справжню дисперсію.

**Приклад 3.2.** При вивченні тривалості наголошених голосних в англійській мові була отримана вибірка, що містить 699 значень. Обчислення тривалості здійснювалося в мсек, і після групування отриманих даних були заповнені стовпці в таблиці 3.1.

Таблиця 3.1

№ п/п	Інтервали	$n_i$	$c_i=N_i-N_0$	$n_i c_i$	$n_i c_i^2$
	1	2	3	4	5
1	80-100	2	-6	-12	72
2	100-120	5	-5	-25	125
3	120-140	32	-4	-128	512
4	140-160	59	-3	-177	531
5	160-180	71	-2	-142	284
6	180-200	92	-1	-92	92
7	200-220	127	-0	0	0
8	220-240	77	1	77	77
9	240-260	66	2	132	264
10	260-280	67	3	201	603
11	280-300	36	4	144	576
12	300-320	31	5	155	775
13	320-340	15	6	90	540
14	340-360	12	7	84	588
15	360-380	7	8	56	448
	Сума	$\sum n_i$		$\sum n_i c_i$	$\sum n_i c_i^2$

З таблиці видно, що  $k=15$ ,  $h=20$ . Зафіксуємо інтервал з № 7 і обчислимо числа  $c_i$  – стовпець № 3. Для одержання чисел стовпця № 4 перемножуються відповідні елементи другого й третього стовпців. Останній (п'ятий) стовпець отримується множенням елементів стовпця № 2 на квадрати чисел стовпця № 3 (піднесення у квадрат виробляється звичайно усно). Складаючи елементи другого, четвертого й п'ятого стовпців, відповідно одержуємо значення для  $n = \sum n_i = 699$ ,  $\sum_{i=1}^k n_i c_i = 363$ ,

$$\sum_{i=1}^k n_i c_i^2 = 5487, \text{ звідки}$$

$$\bar{x} = 210 + \frac{363}{699} \cdot 20 = 210 + 10.4 \approx 220,$$

$$S^2 = \left[ 5487 - \left( \frac{363}{699} \right)^2 \right] \cdot \frac{20^2}{698} = 3000,$$

$$S = \sqrt{3000} = 54.77 \approx 55.$$

При рішенні було враховано, що серединою інтервалу № 7 є число 210. Так як  $h=20$ , то уточнене значення  $S^2$  дорівнює

$$S^2 = 3000 - \frac{400}{12} \approx 2967,$$

$$S = \sqrt{2967} \approx 54.5.$$

### 3.2. Критерій нормальності розподілу

Одним з законів розподілу випадкової величини, що найбільш часто зустрічаються на практиці, є нормальний закон розподілу. Графік функції нормального розподілу симетричний відносно  $a$ . На практиці  $a$  знаходить своє вираження в середньому арифметичному значенні  $\bar{x}$ , а  $\sigma^2$  – у вибірковому середньому  $S^2$ . Як ми вже відзначали, значення випадкової величини з нормальним законом розподілу розташовуються по обидві сторони від генерального середнього  $a$ , і чим далі ці значення відділені від  $a$ , тим вони менш імовірні. Крім того, чим менше  $\sigma^2$ , тим швидше гілки

кривої наближаються до осі абсцис, тобто тим менш імовірні значення випадкової величини, що сильно ухиляються від  $a$  [5]. У силу того, що середнє арифметичне значення вибірки  $\bar{x}$  так само, як і вибіркова дисперсія  $S^2$ , є наближеннями до  $a$  і  $\sigma^2$ , то можна чекати, що характер розташування у вибірці значень випадкової величини (іноді значення називають варіантами), розподіленої за нормальним законом, буде мати зазначені властивості. Нормальна випадкова величина з генеральним середнім  $a$  й дисперсією  $\sigma^2$  має наступну важливу властивість: з імовірністю  $\beta$  варіанти вилучені від  $a$  на відстань, не більше  $d_\beta\sigma$ , де  $d_\beta$  можна визначити з таблиці 3.2.

Таблиця 3.2

$\beta$	0.687	0.900	0.950	0.980	0.990	0.999
$d_\beta$	1.000	1.645	1.960	2.326	2.576	3.291

З таблиці видно, що практично (з імовірністю, більшою 0.99) всі значення нормальної випадкової величини вилучені від  $a$  на відстань, що не перевершує  $3\sigma$ . Це твердження зветься *правилом трьох сигм*.

Нормально розподілені випадкові величини звичайно відповідають явищам, що є результатом багатьох одночасно діючих факторів, дія кожного з яких не є переважаючим над іншими. Така ситуація виникає при вивченні явища, що протікає в однорідних умовах. Так, наприклад, акустичні характеристики наголошеного складу доцільно досліджувати на однорідному лінгвістичному матеріалі за допомогою «однорідної» групи дикторів. Але умову однорідності не завжди можна витримати. Крім того, природа явища іноді буває мало вивчена, і тому важко визначити ступінь впливу окремих факторів на кількісні характеристики даного явища. До того ж мовне явище на рівні мовлення являє собою неоднорідний матеріал. Очевидно, не можна очікувати, що досліджуване явище обов'язково визначить нормально розподілену випадкову величину. Тому на практиці виникає задача визначення закону розподілу.

Визначення закону розподілу по вибірці розпадається на три етапи: формулювання виду закону розподілу, знаходження параметрів цього розподілу й перевірка згоди вихідної вибірки із прийнятим (сформульованим) законом розподілу. Однією з передумов для формулювання виду закону розподілу є емпіричний розподіл. Для побудови емпіричного розподілу в отриманій вибірці проводять інтервальну розбивку й складають таблицю й на її підставі будують гістограму емпіричного розподілу.

**Приклад 3.3.** Для умови прикладу 3.2 маємо таблицю розподілу 3.3.

Таблиця 3.3

Інтервали	Частоти ( $n_i$ )	Відносні частоти ( $n_i/n$ )	Очікувані теоретичні ймовірності
80-100	2	2/699	
100-120	5	5/699	
120-140	32	32/699	
140-160	59	59/699	
160-180	71	71/699	
180-200	92	92/699	
200-220	127	127/699	
220-240	77	77/699	
240-260	66	66/699	
260-280	67	67/699	
280-300	36	36/699	
300-320	31	31/699	
320-340	15	15/699	
340-360	12	12/699	
360-380	7	7/699	
$n = \sum n_i = 699$			

Цій таблиці відповідає гістограма

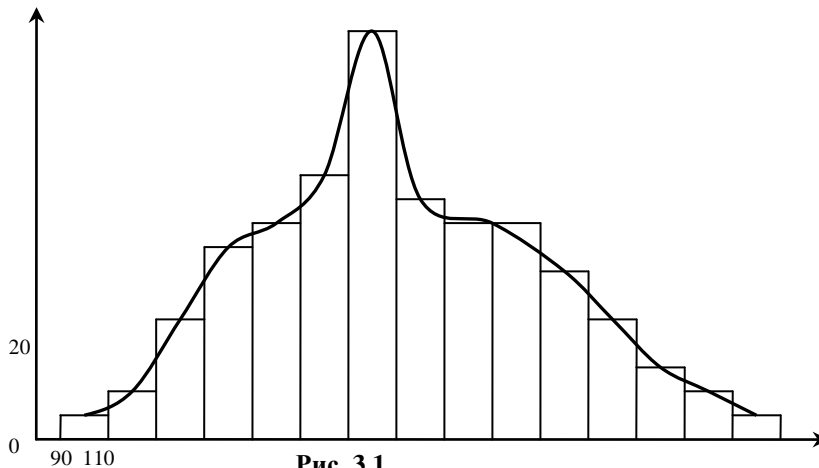


Рис. 3.1

Приблизно через середини верхніх сторін прямокутників гістограми проводиться плавна крива. Отримана крива, хоч і віддалено, нагадує криву нормального розподілу. Тому природно припустити, що досліджувана вибірка взята з нормальної генеральної сукупності, параметри якої  $a$  й  $\sigma^2$  оцінюються на основі вибірки.

У загальному випадку, якщо із приводу гістограми було зроблене припущення про вид

теоретичного розподілу й були оцінені його параметри, то залишається порівняти теоретичний закон розподілу з вибіркою. Для такого порівняння необхідно вибрати критерій перевірки гіпотези про погодженість емпіричного й теоретичного розподілу. Такі критерії називаються *критеріями згоди*. Всі критерії згоди складаються за однаковою схемою. Для цього вибирається деякий параметр, закон розподілу якого відомий, і який має досить великою «чутливістю», тобто відмінність закону, що перевіряє, від дійсного істотно позначалося б на значеннях цього параметра. Найбільшою поширеністю користується *критерій Пірсона*, роль параметра в якому грає величина

$$\chi^2 = n \sum_{i=1}^k \frac{\left( \frac{n_i}{n} - p_i \right)^2}{p_i}, \quad (3.1)$$

де  $n$  – об'єм вибірки,  $k$  – кількість інтервалів розбивки,  $\left( \frac{n_i}{n} \right)$  – відносні частоти інтервалів,  $p_i$  – теоретичні ймовірності цих інтервалів (у припущенні про вірність передбачуваного теоретичного розподілу).

Уведений параметр  $\chi^2$  використовується для перевірки (на підставі вибірки) гіпотези про те, що справжній закон розподілу збігається з передбачуваним. Цю гіпотезу позначають через  $H_0$  і називають *нульовою гіпотезою*. Нульова гіпотеза  $H_0$  відкидається, якщо ймовірність спостерігати дану вибірку в умовах гіпотези  $H_0$  дуже мала. На практиці прийнято вважати, що якщо ця ймовірність менше 0.05, то гіпотезу  $H_0$  відкидають. Число 0.05 називається *рівнем значимості*. Іноді замість 0.05 беруть 0.01. Параметр  $\chi^2$  визначає критерій перевірки гіпотези  $H_0$ . Його називають *критерієм «хі-квадрат»*. Застосування критерію засноване на тім, що для обраного рівня значимості визначають табличне  $\chi_{0.05}^2$  або  $\chi_{0.01}^2$  й порівнюють зі значенням  $\chi^2$ , обчисленим по формулі (3.1). Щоб знайти табличне  $\chi^2$ , варто звернутися до таблиці розподілу  $\chi^2$  (див. Додаток, табл. 2) для значення  $f = k - r - 1$ , де  $k$  – кількість інтервалів групування, використаних при обчисленні  $\chi^2$ , а  $r$  – кількість параметрів теоретичного закону розподілу, оцінених на основі даної вибірки. Число  $f$  називається *числом ступенів волі*. Для випадку, коли передбачуваний теоретичний закон розподілу є нормальним,  $f = k - 3$ .

Для застосування критерію згоди  $\chi^2$  при перевірці нульової гіпотези  $H_0$ : «дана вибірка взята з генеральної сукупності з нормальним законом розподілу» - проводиться обчислення в наступному порядку:

1. Нехай вирішене провести інтервальну розбивку спостережень вибірки на  $k$  інтервалів і нехай  $y_1, y_2, \dots, y_{k+1}$  – границі всіх інтервалів.
2. Визначаються вибіркові  $\bar{x}$  й  $S^2$ .
3. Обчислюються різниці  $y_1 - \bar{x}, y_2 - \bar{x}, \dots, y_{k+1} - \bar{x} \dots$
4. Ці різниці нормуються, тобто обчислюються величини

$$u_1 = \frac{y_1 - \bar{x}}{S}, u_2 = \frac{y_2 - \bar{x}}{S}, \dots, u_{k+1} = \frac{y_{k+1} - \bar{x}}{S}.$$

5. По таблиці знаходимо величини (див. Додаток, табл.1)  $\Phi(u_1), \Phi(u_2), \dots, \Phi(u_{k+1})$ .
6. Складаємо різниці  $p_1 = |\Phi(u_1) - \Phi(u_2)|, \dots, p_k = |\Phi(u_k) - \Phi(u_{k+1})|$ .
7. Обчислюємо  $\tilde{n}_i = p_i n$ .
8. Визначаємо значення  $\chi^2$  по формулі

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}.$$

Ця формула дає те ж значення  $\chi^2$ , що й формула (3.1), але вона більш зручна для обчислень. До того ж вона свідчить про правильність проведеної інтервальної розбивки. Вважається, що «очікувані частоти»  $\tilde{n}_i$  не повинні бути менше п'яти. У протилежному випадку кілька середніх інтервалів поєднують в один, тобто тут уже не обов'язково піклуватися про рівність довжин інтервалів (таке об'єднання треба проводити вже після обчислення  $\bar{x}$  й  $S^2$ ).

Обчислене  $\chi^2$  порівнюємо з табличним  $\chi_{0.05}^2$  (або  $\chi_{0.01}^2$ ). Якщо виявиться, що  $\chi^2 > \chi_{0.05}^2$ , то гіпотеза  $H_0$  відкидається з рівнем значимості 0.05. Рівень значимості  $\alpha$  (звичайно  $\alpha=0.05$  або  $\alpha=0.01$ ) означає, що ймовірність одержання даної вибірки в умовах гіпотези  $H_0$  менше  $\alpha$ , тобто гіпотеза  $H_0$  слабко погоджується з даною вибіркою, а тому повинна бути відкинута. Якщо ж  $\chi^2 < \chi_{0.05}^2$ , то гіпотеза  $H_0$  приймається. Точніше кажучи, ця гіпотеза не відкидається, і наступні експерименти можуть підтвердити її або спростувати. Особливо важливо продовжувати експеримент, якщо обчислене значення  $\chi^2$  задовольняє нерівності  $\chi_{0.05}^2 < \chi^2 < \chi_{0.01}^2$ , тому що дослідник (з обережності) може вибрати рівень значимості  $\alpha=0.01$ , а тому не відкидає гіпотезу  $H_0$ , але наведена нерівність ставить під сумнів його висновок.

**Приклад 3.4.** Розглядаючи наведену вище гістограму, можна побачити, що в центрі її перебувають найбільш частотні варіанти, а ближче до країв – відносно рідкі варіанти. По цих ознаках гістограма нагадує криву щільності нормального розподілу. Використовуючи вибіркове середнє  $\bar{x}=220$  і середнє квадратичне відхилення  $S=55$ , за допомогою критерію  $\chi^2$  можна перевірити гіпотезу  $H_0$ : акустична характеристика – тривалість наголошеного складу в англійській мові – підкоряється нормальному закону розподілу з параметрами  $a=220$  і  $\sigma=55$ . Емпіричне значення  $\chi^2$  (з його допомогою перевіряється гіпотеза  $H_0$ ) обчислюється відповідно до зазначеної схеми. Результати проміжних обчислень наведені в таблиці 4.4.

Опишемо процес заповнення стовпців цієї таблиці. Було ухвалено рішення використати 15 інтервалів групування ( $k=15$ ), причому початок першого інтервалу дорівнює 80, а довжина інтервалів дорівнює 20. У стовпці № 1 зазначені середини обраних інтервалів, у стовпці № 2 – кінці відповідних інтервалів, у стовпці № 3 – відхилення кінців інтервалів від вибіркового середнього  $\bar{x}=220$ . У стовпці № 4 записані числа  $u_i$ , що є нормованими відхиленнями кінців інтервалів. Ці числа отримуються діленням відповідних чисел стовпця № 3 на  $S=55$  (з урахуванням знака). Стовпець № 5 містить значення  $\Phi(u_i)$ , які знаходять по таблиці значень функції  $\Phi$  (див. Додаток, табл.1). Так як для  $u_1 = -2.54$  маємо  $\Phi(-2.54)=0.006$ , а для  $u_{10}=0.73$  маємо  $\Phi(0.73)=0.767$ . Числа стовпця № 6 відповідають серединам інтервалів і дорівнюють різницям значень функції  $\Phi$  для кінців інтервалу (з більшого значення  $\Phi$  віднімається менше значення). Числа стовпця № 6 і є теоретичні («очікувані») частоти  $\tilde{n}_i$ . Вони отримуються множенням чисел стовпця № 6 на об'єм вибірки  $n$  (у нашому випадку  $n=699$ ). У стовпці № 8 зазначені емпіричні частоти  $n_i$ . Вони відбивають результат вибірки. Ці числа взяті зі стовпця № 2 таблиці 3.1.

Таблиця 3.4

Середини інтервалів	Границі інтервалів	Відхилення границь інтервалів від $\bar{x}$	Нормиров. відхилення границь інтервалів $u_i$	$\Phi(u_i)$	$p_i =  \Phi(u_i) - \Phi(u_{i+1}) $	$\tilde{n}_i = n_i p_i$	$n_i$
1	2	3	4	5	6	7	8
90	80	- 140	- 2.54	0.006	0.009	6.9	2
110	100	- 120	- 2.18	0.015	0.020	13.9	5
130	120	- 100	- 1.81	0.035	0.038	26.6	32
150	140	- 80	- 1.45	0.073	0.065	45.4	59
170	160	- 60	- 1.09	0.138	0.095	66.3	71
190	180	- 40	- 0.73	0.233	0.126	88.0	92
210	200	- 20	- 0.36	0.359	0.141	98.5	127
230	220	0	0	0.500	0.141	98.5	77
250	240	20	0.36	0.641	0.126	88.0	66
270	260	40	0.73	0.767	0.095	66.3	67
290	280	60	1.09	0.862	0.064	45.4	36
310	300	80	1.45	0.926	0.039	26.6	31
330	320	100	1.81	0.965	0.020	13.9	15
350	340	120	2.18	0.985	0.009	6.3	12
370	360	140	2.54	0.994	0.004	2.8	7
	380	160	2.91	0.998			

Тепер обчислюємо  $\chi^2$ . Для цього із чисел стовпця № 8 віднімаються числа стовпця № 7, отримані різниці зводяться у квадрат, а потім діляться на числа стовпця № 7. Результати підсумуються. Тому маємо:

$$\begin{aligned} \chi^2 &= \frac{(2-6.9)^2}{6.9} + \frac{(5-13.9)^2}{13.9} + \frac{(32-26.6)^2}{26.6} + \frac{(59-45.4)^2}{45.4} + \frac{(71-66.3)^2}{66.3} + \frac{(92-88)^2}{88} + \\ &+ \frac{(127-98.5)^2}{98.5} + \frac{(77-98.5)^2}{98.5} + \frac{(66-88)^2}{88} + \frac{(67-66.3)^2}{66.3} + \frac{(36-45.4)^2}{45.4} + \\ &+ \frac{(31-26.6)^2}{26.6} + \frac{(15-13.9)^2}{13.9} + \frac{(12-6.3)^2}{6.3} + \frac{(7-2.8)^2}{2.8} = \\ &= \frac{24}{6.9} + \frac{79.3}{13.9} + \frac{29.2}{26.6} + \frac{207.4}{45.4} + \frac{22.1}{66.3} + \frac{16}{88} + \frac{756.2}{98.5} + \frac{462.2}{98.5} + \frac{484}{88} + \frac{0.5}{66.3} + \frac{88.4}{45.4} + \\ &+ \frac{18.4}{26.6} + \frac{1.2}{13.9} + \frac{32.5}{6.3} + \frac{17.6}{2.8} = 47.2 \end{aligned}$$

Тепер по таблиці розподілу  $\chi^2$  знаходимо  $\chi_{05}^2$  й  $\chi_{01}^2$ , що відповідають  $f=15-3=12$ . Маємо  $\chi_{05}^2=21.0$ ,  $\chi_{01}^2=26.2$ . Оскільки  $\chi^2=47.2 > 26.2$  (тобто  $\chi_{01}^2$ ), то гіпотеза  $H_0$  відкидається. Таким чином, припущення про нормальний закон розподілу значень тривалості не підтвердилося.

### 3.3. Критерій однорідності

Цінність критерію  $\chi^2$  (для перевірки гіпотез) полягає в тому, що з його допомогою можна досліджувати якісні ознаки, які характеризуються якимись відтінками, але не мають кількісного вираження. Так, при вивченні мелодії інтонаційного типу необхідно враховувати комунікативне навантаження фраз, їхню емоційну насиченість, граматичний лад, ритмічну структуру, індивідуальні особливості голосу мовця. Якщо ні для одного із цих факторів не створено спеціальних умов, то досліджуване явище буде підкорятися нормальному закону. На практиці часто доводиться мати справу з явищами, коли для супутніх цьому явищу факторів можуть, незалежно від бажання дослідника, створюватися сприятливі для цього фактору умови. Оскільки всяке явище вивчається в

часі, то іноді ці умови можуть стати значущими. Тому перед дослідником може іноді виникати проблема підбора однорідного матеріалу для вивчення явища.

Наприклад, вивчалася частота основного тону голосного з головним наголосом в емоційних реченнях, що виражають подив. Аудиторський аналіз свідчив про чітке розходження позитивного й негативного відтінків емоції при прослуховуванні фраз у контексті. Однак коли експериментальні фрази були вичленовані з контексту, показання аудиторів виявилися плутаними, а іноді й суперечливими. Результати аудіювання дозволяють висунути гіпотезу про те, що не існує акустичних ознак розходження позитивного й негативного відтінків емоції подиву, принаймні, такий параметр, як частота основного тону, не розрізняє цих відтінків. Інакше кажучи, матеріал про значення частоти основного тону, отриманий окремо для фраз із позитивними й негативними відтінками емоції подиву, є однорідним, тобто вибірки по цих відтінках емоції можна об'єднати. Висунути гіпотезу можна перевірити за допомогою критерію  $\chi^2$ . Процес обчислення  $\chi^2$  у цьому випадку розглянемо на наступному прикладі.

**Приклад 3.5.** Значення частоти основного тону голосного з головним наголосом в емоційних реченнях, що виражають подив, зазначені в таблиці 3.5.

Таблиця 3.5.

Інтервали (частота основного тону)	Частоти для положит. відтінків емоції ( $n_{1i}$ )	Частоти для отрицат. відтінків емоції ( $n_{2i}$ )	$n^{(i)}$	$\frac{n_1 n^{(j)}}{N}$	$\frac{n_2 n^{(j)}}{N}$
1	2	3	4	5	6
0.80 – 1.20	7	2	9	6.7	2.3
1.20 – 1.60	16	3	19	14.1	4.8
1.60 – 2.00	24	7	31	23.0	8.0
2.00 – 2.40	65	26	91	67.6	23.4
2.40 – 2.80	15	6	21	15.6	5.4
$n_i$	$n_1=127$	$n_2=44$	$N=171$		

Стовпці цієї таблиці заповнюються в такий спосіб: у результаті експерименту були отримані значення частоти основного тону (як для позитивних, так і для негативних відтінків емоції подиву), які були розподілені в п'ятьох інтервалах (0.80 – 1.20), (1.20 – 1.60), (1.60 – 2.00), (2.00 – 2.40), (2.40 – 2.80). Ці інтервали зазначені в стовпці № 1. У стовпці № 2 записані числа фраз, що спостерігалися, з позитивним відтінком емоції подиву, склад з головним наголосом яких має значення частоти основного тону у відповідному інтервалі. Таким же чином заповнюється стовпець № 3, що відповідає негативним відтінкам емоції подиву. Стовпець № 4 – це сума відповідних елементів стовпців № 2 і № 3. Число  $N$  – сума всіх елементів стовпця № 4,  $n_1$  – сума елементів стовпця № 2,  $n_2$  – сума елементів стовпця № 3. Ясно, що  $N=n_1+n_2$ . У цьому прикладі  $N=171$ ,  $n_1=127$ ,  $n_2=44$ . Елементи стовпця № 5 (і відповідно стовпця № 6) отримуються як результат перемножування відповідних елементів стовпця № 4 на  $n_1$  (відповідно, на  $n_2$ ) і ділення результату на  $N$ .

Тепер можна приступитися до обчислення  $\chi^2$ . Для цього з кожного елемента стовпця № 2 віднімаємо відповідний елемент стовпця № 5, отримані різниці підносимо до квадрата, а результат ділимо на елемент стовпця № 5. У такий же спосіб вчиняємо з парами зі стовпців № 3 і № 6. Отримані числа підсумуємо. Сума і є значення  $\chi^2$ . Отже,

$$\chi^2 = \frac{(7-6.7)^2}{6.7} + \frac{(2-2.3)^2}{2.3} + \frac{(16-14.1)^2}{14.1} + \frac{(3-4.8)^2}{4.8} + \frac{(24-23)^2}{23} + \frac{(7-8)^2}{8} + \frac{(65-67.6)^2}{67.6} + \frac{(26-23.4)^2}{23.4} = 1.61.$$

Число ступенів волі  $f=k-1$ , де  $k$  – кількість інтервалів групування. У нашому прикладі  $f=4$ . З таблиці 2 Додатку знаходимо  $\chi_{05}^2=9.488$  і  $\chi_{01}^2=13.30$ . Оскільки  $\chi^2=1.61 < 9.488 = \chi_{05}^2$ , то гіпотеза  $H_0$  приймається, тобто позитивні й негативні відтінки емоції подиву однаково впливають на частоту основного тону складу з головним наголосом.

У загальному випадку, щоб побудувати критерій однорідності, припустимо, що досліджуване явище характеризується набором  $k$  взаємовиключних один одного якостей  $P_1, \dots, P_k$  (у прикладі 3.5 якості характеризувалися інтервалами групування). Нехай є  $l$  вибірок відповідно об'ємів  $n_1, \dots, n_l$  (у прикладі 4.5  $n=2$ ). Позначимо через  $n_{ij}$  кількість елементів в  $j$ -й вибірці, що володіють якістю  $P_j$ . Очевидно, що  $n_i = n_{i,1} + \dots + n_{i,k}$ . Позначимо ще  $n^{(j)} = n_{1,j} + \dots + n_{l,j}$ ;  $N = n_1 + \dots + n_l$ . Значення  $\chi^2$  обчислюється по формулі

$$\chi^2 = \sum_{i,j} \frac{\left( n_{ij} - \frac{n_i n^{(j)}}{N} \right)^2}{\frac{n_i n^{(j)}}{N}}$$

Тут підсумовування ведеться по всіляких парах чисел  $(i, j)$   $i=1, \dots, k, j=1, \dots, k$ . Це значення  $\chi^2$  порівнюється з табличним  $\chi_{05}^2$  або  $\chi_{01}^2$ , обчисленим для  $f=(k-1)(n-1)$ . Для випадку  $n=2$  ми, мабуть, маємо випадок, уже розібраний у прикладі 3.5.

### 3.4. Розподіл середнього арифметичного значення

Вибіркове середнє  $\bar{x}$  визначається на підставі деякої вибірки, отриманої в результаті експерименту. Тому як сама вибірка, так і вибіркове середнє  $\bar{x}$ , є випадковими величинами. Якщо зробити кілька вибірок, тобто кілька незалежних друг від друга експериментів, то їх вибіркові середні  $\bar{x}_1, \dots, \bar{x}_k$  можна розглядати як вибірку з деякої генеральної сукупності. Закон розподілу значень деякої лінгвістичної ознаки може бути відмінний від нормального, але вибіркові середні мають розподіл, близький до нормального. Це твердження стає практично достовірним, якщо об'єми вибірок, для яких визначалися  $\bar{x}$ , більші 30.

Як було відзначено вище, з великою ймовірністю (більшою 0.99) значення нормально розподіленої випадкової величини віддалені від її генерального середнього на відстань, меншу  $3\sigma$ . Тому, визначивши вибіркове середнє  $\bar{x}$ , можна з імовірністю 0.99 вказати інтервал (його серединою буде  $\bar{x}$ ), у якому втримується генеральне середнє  $\hat{x}$ . Але для цього треба знати середнє квадратичне  $\sigma$  нормального розподілу, якому підкоряється  $\bar{x}$ .

Як оцінку для  $\sigma$  беруть величину  $\frac{S}{\sqrt{n}}$ , де  $S$  – середнє квадратичне відхилення для вибірки, по якій знайдено  $\bar{x}$ , а  $n$  – об'єм цієї вибірки. Таким чином, з імовірністю 0.99 справжнє середнє  $\hat{x}$  перебуває в інтервалі  $(\bar{x} - 3 \frac{S}{\sqrt{n}}, \bar{x} + 3 \frac{S}{\sqrt{n}})$ . Якщо можна задовольнятися меншою ймовірністю, то границі інтервалу, що містить  $\hat{x}$ , можна звужити. У загальному випадку ці границі мають вигляд  $(\bar{x} - d_\beta \frac{S}{\sqrt{n}}, \bar{x} + d_\beta \frac{S}{\sqrt{n}})$ , де  $d_\beta$  вибирається з таблиці 4.2 на підставі заданого значення ймовірності.

**Приклад 3.6.** У прикладі 3.2 було знайдено  $\bar{x}=220$ ,  $S=55$ , причому  $n=699$ . Тому  $\frac{S}{\sqrt{n}} = \frac{55}{\sqrt{699}} = 2.1$ . Це значить, що з імовірністю, не меншою 0.99, справжнє середнє значення тривалості перебуває в межах  $220 \pm 6.3$ , тобто між 213.7 і 226.3.

Іноді при порівнянні двох вибірок на предмет їхньої репрезентативності використовується параметр  $E$ , що обчислює по формулі

$$E = \frac{3 \cdot S}{x \sqrt{n}} \cdot 100\%.$$

Уважається, що, якщо дві вибірки взяті з однієї й тієї ж генеральної сукупності (тобто отримані в результаті вивчення того самого явища, що протікає в однорідних умовах), то вибірка, для



якої значення  $E$  менше, є більш представницькою. Інакше кажучи, у такій вибірці краще проявляється ознака досліджуваного явища.

Вибірка вважається *представницькою* серед подібних їй вибірок однієї й тієї ж генеральної сукупності, якщо її значення  $E$  менше 15%. У протилежному випадку до оцінок параметрів розподілу ( $\mu$  і  $\sigma$ ), що даються цією вибіркою, варто ставитися з обережністю.

**Приклад 3.7.** Розглянемо вибірки, вивчені в прикладах 3.1 і 3.2. Позначимо їхні середні значення відповідно через  $\bar{x}_1$  і  $\bar{x}_2$ , а середні квадратичні відхилення відповідно через  $S_1$  і  $S_2$ . Для вибірки із приклада 3.1  $\bar{x}_1=162$ ,  $S_1=29.3$ , звідки випливає, що

$$\frac{S_1}{\sqrt{n_1}} = \frac{29.3}{\sqrt{10}} \approx 9.1,$$

$$(\bar{x}_1 \pm 3 \cdot S_1) = 162 \pm 29.3,$$

$$E_1 = \frac{3 \cdot 29.3}{162 \sqrt{10}} \cdot 100\% = 17\%.$$

Для вибірки із приклада 3.2  $\bar{x}_2=220$ ,  $S_2=55$ . Отже,

$$\frac{S_2}{\sqrt{n_2}} = \frac{55}{\sqrt{699}} = 2.08,$$

$$(\bar{x}_2 \pm 3 \cdot S_2) = 220 \pm 165,$$

$$E_2 = \frac{3 \cdot 55}{220 \sqrt{699}} \cdot 100\% = 2.8\%.$$

Так як  $E_1=17\%>15\%$ , то відповідно до прийнятої границі обчислене  $\bar{x}_1$  не можна прийняти, а відповідну вибірку вважати репрезентативною. У той же час, значення  $\bar{x}_2$  цілком відповідає обраному критерію. Таким чином, можна прийняти, що середнє арифметичне значення тривалості голосних під наголосом в англійській мові укладена в межах 214 – 226 мсек. Границі визначені з імовірністю 0.99.

Погодженість розподілу  $\bar{x}$  з нормальним законом дозволяє при будь-якому розподілі варіант із певною ймовірністю визначити границі для  $\bar{x}$ . А це означає, що кожна репрезентативна вибірка (тобто при  $E>15\%$ ) може мати надійну кількісну характеристику у вигляді  $\bar{x} \pm 3S$ . Дана характеристика добре відбиває якісну сторону досліджуваного явища, а це може бути використане при опису й порівнянні ознак різної природи. Крім того, оцінка  $\bar{x}$  застосовується при визначенні помилки виміру. Допустимо, необхідно зняти з інтеннограми значення частоти основного тону. Природно, що при вимірі частоти основного тону цілком може бути допущена якась помилка, величина якої визначається границями довірчого інтервалу  $X$ .

**Приклад 3.8.** Була обмірювана частота основного тону мелодійної кривої інтеннограми фрази. Для визначення помилки виміру було взято десять однослівних фраз і зроблено по двадцять вимірів максимуму частоти основного тону голосного з головним наголосом кожного із цих слів. Були визначені середні арифметичні значення частоти основного тону по двадцяти вимірам. Таке усереднення робить закон розподілу середніх значень частоти основного тону розглянутих голосних близьким до нормального, що дозволяє користуватися правилом трьох сигм для визначеного інтервалу, у якому лежить справжнє середнє значення частоти основного тону (усереднення для основних акустичних параметрів рекомендується проводити по вибірках об'єму не менш 10). Результати обчислень представлені в таблиці 3.6.

Таблиця 3.6

№ досвіду	Слово	Голосний	Середнє значення частоти основного тону (по 20 вимірам)	Дисперсія в досвідах
1	2	3	4	5
1	'abstract	æ	300	1.29
2	'accent	»	272	6.18
3	'affix	»	271	1.32
4	'conflict	)	274	4.11
5	'contest	»	275	1.30
6	'contract	»	292	6.35
7	'decrease	[i:]	271	3.89
8	'detail	»	275	1.44
9	'discus	[I]	272	5.84
10	'increase	»	270	1.27
			2772	33.01

Суму елементів стовпця № 3 ділимо на 10 (на кількість досвідів) і одержуємо середнє значення частоти основного тону голосних під наголосом однослівних фраз. Отже,  $\bar{x}=277.2$ . Із суми елементів стовпця № 4 витягаємо квадратний корінь і ділимо на 10, що дає середнє квадратичне відхилення. У нашому випадку  $S=0.57$ . Тепер довірчий інтервал, побудований за правилом трьох сигм, має вигляд  $277.2 \pm 3 \cdot 0.57 = 277.2 \pm 1.71$ . Відповіднє значення  $E$  дорівнює

$$E = \frac{3 \cdot 0.57}{277.2 \sqrt{10}} \cdot 100\% = 0.2\%$$

Таким чином, наша вибірка є репрезентативною, а виходить, зазначений довірчий інтервал цілком придатний для вживання.

### 3.5. Контрольні запитання

1. Що таке випадкова вибірка?
2. Як обчислюється вибіркєве середнє?
3. Як обчислюється вибіркєва дисперсія?
4. В який спосіб проводиться угруповання вибіркєвих значень?
5. Що таке частота інтервалу?
6. Як і коли вноситься поправка Шеппарда?
7. Сформулюйте правило трьох сигм.
8. Що показує критерій згоди Пірсона?
9. Яка величина грає роль параметра в критерії Пірсона?
10. Як висувається нульова гіпотеза?
11. Як обчислюється число ступенів волі?
12. Що показує критерій однорідності?
13. Що таке репрезентативність вибірки?
14. Який закон розподілу має середнє арифметичне значення?
15. Як обираються границі довірчого інтервалу?
16. За яким критерієм досліджується репрезентативність вибірки?

## 4. ЛІНГВІСТИЧНІ ГІПОТЕЗИ

### 4.1. Критерій Стьюдента

При вивченні показників лінгвістичних характеристик велике значення має наступна постановка питання: деякий лінгвістичний показник визначений по двох різних вибірках. Як правило, середнє значення цього показника в одній вибірці відрізняється від середнього значення його в іншій вибірці. Чи істотно це розходження?

У тих випадках, коли закон розподілу цього показника близький до нормального, відповідь на поставлене питання дає критерій Стьюдента. Вимогу нормальності можна опустити, якщо об'єми порівнюваних вибірок досить великі (більше 30). Сутність критерію Стьюдента проілюструємо на наступному прикладі.

**Приклад 4.1.** Для двох груп дикторів (група **A** – чоловіки, група **B** – жінки) зафіксовані значення сумарної енергії голосного у двоскладових словах англійської мови з наголосом на початковому складі. Результати були піддані інтервальному угрупованню й були побудовані дві таблиці.

Таблиця 4.1

#### Група A

Інтервали	Частоти ( $n_i$ )	$x_i$	$n_i x_i$	$n_i x_i^2$
180-210	2	-4	-8	32
210-240	4	-3	-12	36
240-270	5	-2	-10	20
270-300	6	-1	-6	6
300-330	8	0	0	0
330-360	6	1	6	6
360-390	4	2	8	16
390-420	3	3	9	27
420-450	2	4	8	32
	40		-7	175

#### Група B

Інтервали	Частоти ( $n_j$ )	$y_j$	$n_j y_j$	$n_j y_j^2$
120-140	1	-5	-5	25
140-160	1	-4	-4	16
160-180	3	-3	-9	27
180-200	3	-2	-6	12
200-220	4	-1	-4	4
220-240	5	0	0	0
240-260	5	1	5	5
260-280	3	2	6	12
280-300	3	3	9	27
300-320	2	4	8	32
320-340	2	5	10	50
	32		+10	210

З таблиці маємо:

$$\bar{x} = 315 + \frac{(-7)}{40} \cdot 30 = 310; \quad \bar{y} = 230 + \frac{10}{32} \cdot 20 = 236,$$

де  $\bar{x}$  - середнє значення сумарної енергії для групи **A**, а  $\bar{y}$  - середнє значення для групи **B**.

Розходження вибірових середніх (310 і 236) ще не доводить, що групи **A** і **B** істотно різні з точки зору впливу на значення сумарної енергії. Висувається гіпотеза  $H_0$ : розходження вибірових

середніх  $\bar{x}$  і  $\bar{y}$  є несуттєвим, випадковим, тобто справжні середні значення сумарної енергії для груп **A** і **B** однакові. Для перевірки цієї гіпотези обчислюємо  $s_{x-y}^2$  по формулі:

$$s_{x-y}^2 = \sqrt{\frac{\sum(x_i - \bar{x})^2 + \sum(y_j - \bar{y})^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}},$$

Де  $n_1$  і  $n_2$  – об'єми відповідно вибірок у групах **A** і **B**, крім того суми  $\sum(x_i - \bar{x})^2$  і  $\sum(y_j - \bar{y})^2$  мають той же зміст, що й при обчисленні дисперсій вибірок відповідно для груп **A** і **B**. Тому

$$\sum(x_i - \bar{x})^2 = \left(175 - \frac{(-7)}{40}\right) \cdot 30^2 = 156420,$$

$$\sum(y_j - \bar{y})^2 = \left(210 - \frac{(10)}{32}\right) \cdot 20^2 = 82800.$$

Тому

$$s_{x-y}^2 = \sqrt{\frac{156420 + 82800}{40 + 32 - 2} \cdot \frac{40 + 32}{40 \cdot 32}} = 13.9.$$

Тепер обчислюємо

$$t = \frac{|\bar{x} - \bar{y}|}{s_{x-y}} = \frac{310 - 236}{13.9} = 5.33.$$

По таблиці розподілу Стьюдента (див. Додаток 3) визначаємо  $t_{05}$  для числа ступенів волі  $f = n_1 + n_2 - 2$ . У розглянутому прикладі  $t_{05} = 2.00$ . Далі необхідно порівняти обчислене значення  $t$  з табличним  $t_{05}$ . **Критерій Стьюдента затверджує**: якщо  $t > t_{05}$ , то гіпотеза  $H_0$  відкидається. Причому ймовірність того, що вона помилкова, не менша 0.95, тобто в 95% вона не вірна. Якщо ж  $t \leq t_{05}$ , то гіпотеза  $H_0$  приймається. У розглянутому прикладі  $t = 5.33 > 2.00 = t_{05}$ , а тому припущення про відсутність розходжень у впливі дикторів груп **A** і **B** на сумарну енергію помилково.

У розглянутому прикладі вибірки для груп **A** і **B** були незалежні одна від одної, і для перевірки гіпотези  $H_0$  (у загальному випадку гіпотеза  $H_0$  формулюється так: розходження в середніх значеннях двох вибірок несуттєво) була використана величина

$$t = \frac{|\bar{x} - \bar{y}|}{s_{x-y}}.$$

Критерій Стьюдента в лінгвістиці часто використовується для аналізу лексики, наприклад для вирішення питання про розходження частот уживання слів [11, стор.68-69]. Критерій Стьюдента може бути використаний не тільки як показник ступеня розходження між частотами слів, але і як показник ступеня їхньої близькості. З його допомогою можна виміряти відстань між словами (парно). Він також може бути використаний і при дослідженні функціональних стилів [12]. Критерій Стьюдента в сполученні із правилом трьох сигм може також використатися для дослідження середньої довжини слова й речення, лексичних, морфологічних і синтаксичних властивостей тексту [6, стор.134]. Дослідження морфологічного рівня тексту подано, зокрема, у роботі [13], у якій вивчена частота зустрічальності різних частин мови української мови в трьох функціональних стилях. Показано, що за допомогою критерію Стьюдента можна встановити ступінь розходження між художнім, публіцистичним і науковим стилями В більшості випадків середні частоти зустрічальності різних частин мови істотно відрізняються одна від одної й, таким чином, середня частота вживання частин мови може бути одним з диференціальних ознак кожного стилю.

Дуже важливим є один окремий випадок застосування критерію Стьюдента. Нехай, наприклад, вивчається енергетична неповноцінність переднаголосних і післянаголосних складів трискладових слів української мови. Висловлено припущення, що в межах того самого слова розхо-

дження немає (це гіпотеза  $H_0$ ). Для перевірки цієї гіпотези не слід окремо вивчати переднаголосні й післянаголосні склади, а досить розглянути пари: переднаголосний-післянаголосний склади. Знайти різницю їх енергетичних наповненостей і перевірити, чи значно середнє значення цих різниць відрізняється від нуля [5]. Такий метод перевірки гіпотези  $H_0$  іноді називають *методом порівнянь*. Він зручний для математичної обробки (замість  $2n$  спостережуваних значень обробляється тільки  $n$  чисел). Але застосовуючи метод парних порівнянь, слід дотримуватися певних умов - пари утворюються в процесі дослідження як результат дії двох постійних факторів, розходження в дії яких цікавить дослідника. Не можна довільно зіставляти пари тільки на тій підставі, що є дві вибірки однакових об'ємів.

Використання критерію Стьюдента має велике значення у фонетиці при відборі однорідного матеріалу для експериментів. Так, наприклад, досліджуючи енергетичні характеристики наголошених складів на відміну від ненаголошених, можна зштовхнутися з тим фактом, що група ненаголошених включає дві підгрупи – переднаголосних і післянаголосних. Причому позиційні умови можуть впливати на акустичні характеристики цих складів. Отримані середні значення для переднаголосних і післянаголосних складів трохи відрізняються, але чи настільки істотно це розходження, що можна говорити про дві різні сукупності, затверджувати неможливо без відповідного статистичного аналізу. Експериментатор висуває гіпотезу  $H_0$  про неістотність розходження вибірових середніх переднаголосних і післянаголосних складів і перевіряє її, використовуючи критерій Стьюдента. Підтвердження гіпотези  $H_0$  дає можливість надалі розглядати переднаголосні й післянаголосний голосні як одну сукупність. Критерій Стьюдента також часто використовується в експериментальній фонетиці, якщо за даними аудиторського аналізу немає істотних розходжень у сприйнятті окремих підгруп явища, і експериментатор хоче об'єднати їх у більші групи.

#### 4.2. Критерій Ван дер Вардена

Критерій Ван дер Вардена, або критерій  $X$ , належить до непараметричних критеріїв, при застосуванні яких немає необхідності обчислювати статистичні параметри вибірки (середнє, дисперсію). Подібно критерію Стьюдента, критерій  $X$  рекомендується застосовувати при вирішенні питання про істотність розбіжностей порівнюваних вибірок варіант, тобто про значимості тієї або іншої ознаки досліджуваної лінгвістичної одиниці. Незважаючи на схожість розв'язуваних задач, обидва критерії розрізняються умовами застосування. Застосування критерію  $X$ , як і будь-якого непараметричного критерію, не вимагає знання функції розподілу варіант. Цей критерій може застосовуватися при малій кількості варіант, тобто в тих випадках, коли не можна вирішити питання про функції розподілу вихідних даних. При великій кількості варіант ( $n \geq 30$ ) критерій Ван дер Вардена діє аналогічно критерію Стьюдента й тому в цих випадках застосовується саме критерій Стьюдента.

**Приклад 4.2.** Досліджується вплив ритміки фрази на значення її фізичних характеристик, зокрема, на величину частотного діапазону. Досліджуються фрази наступних ритмічних структур: з початковим головним наголошеним складом і з кінцевим головним наголошеним складом однієї й тієї ж інтонаційної одиниці. Значення частотного діапазону першої групи фраз позначимо як прояв випадкової величини  $x$ , другої групи – через  $y$ . У результаті експерименту (були записані німецькі емоційні фрази) отримані наступні значення частотного діапазону:

$x$ : 1.55, 1.61, 1.73, 1.80, 1.92, 2.09, 2.33, 2.44;  $n_1=8$ ;

$y$ : 1.33, 1.51, 1.66, 1.70, 1.82, 1.88, 2.20;  $n_2=7$ .

Розглянемо сукупність чисел, що складається зі значень  $x$  і  $y$ , і розташуємо числа цієї сукупності в зростаючому порядку. Тим самим кожному розглянутому числу приписується порядковий номер. Для спрощення розрахунків припустимо, що серед всіх значень,  $x$  і  $y$ , немає однакових. У розглянутому прикладі саме така ситуація. Розташування чисел і приписування їм порядкових номерів  $r$  зручно вести за допомогою таблиці 5.2:

Таблиця 4.2

$x$			1.55	1.64			1.73	1.80			1.92	2.09		2.33	2.44
$y$	1.33	1.61			1.66	1.70			1.82	1.88			2.20		
$r$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Таким чином, числам першої групи, тобто числам, що відповідають випадковій величині  $x$ , приписані номери 3, 4, 7, 8, 11, 12, 14, 15. Кількість чисел у сукупності  $x$  (відповідно  $y$ ) позначене через  $n_1$  (відповідно  $n_2$ ) і покладається  $n=n_1+n_2$ . У розглянутому прикладі  $n_1=8$ ,  $n_2=7$ ,  $n=15$ . Складаємо послідовність дробів  $\frac{r}{n+1}$ , де  $r$  пробігає всі номери, приписані числам із сукупності  $x$ . У розглянутому прикладі це дробі  $\frac{3}{16}, \frac{4}{16}, \frac{7}{16}, \frac{8}{16}, \frac{11}{16}, \frac{12}{16}, \frac{14}{16}, \frac{15}{16}$ . Для кожної дробі по таблиці функції  $\Psi$  (Додаток, табл. 4) знаходимо  $\Psi(\frac{r}{n+1})$ , а потім складаємо суму  $X=\sum_r \Psi(\frac{r}{n+1})$ . У розглянутому прикладі

$$\begin{aligned} X &= \Psi\left(\frac{3}{16}\right) + \Psi\left(\frac{4}{16}\right) + \Psi\left(\frac{7}{16}\right) + \Psi\left(\frac{8}{16}\right) + \Psi\left(\frac{11}{16}\right) + \Psi\left(\frac{12}{16}\right) + \Psi\left(\frac{14}{16}\right) + \Psi\left(\frac{15}{16}\right) = \\ &= (-0.89) + (-0.67) + (-0.16) + (0.00) + (0.49) + (0.67) + (1.15) + (1.53) = (-1.72) + (3.84) = 2.12. \end{aligned}$$

Тепер варто звернутися до таблиці критерію  $X$  (Додаток, табл. 5). По числах  $n$  і  $n_1 - n_2$  (якщо  $n_1 < n_2$ , то по числах  $n$  і  $n_2 - n_1$ ) знаходимо  $X_{05}$ . Якщо обчислене значення  $X$  перевершує  $X_{05}$ , то робиться висновок про існування розходження середніх значень сукупностей  $x$  і  $y$  (точніше, робиться висновок про приналежність двох вибірок до різних генеральних сукупностей). У протилежному випадку, тобто коли  $X \leq X_{05}$ , приймається гіпотеза про відсутність розходжень у сукупностях  $x$  і  $y$ . Іноді, з метою обережності, при виконанні нерівності  $X > X_{05}$  порівнюють  $X$  з  $X_{01}$ . Якщо  $X > X_{01}$ , то, безумовно, робиться висновок про розходження генеральних сукупностей, з яких витягнуті вибірки. Якщо ж  $X_{05} < X < X_{01}$ , то ситуація вважається сумнівною, і для остаточного висновку варто продовжувати експеримент.

Для розглянутого приклада  $X_{05}=3.24$ , і значить  $X < X_{05}$ . Тому гіпотеза  $H_0$  приймається. У такий спосіб, розходження в значеннях частотного діапазону двох ритмічних структур є несуттєвим і, отже, можна з великою ймовірністю припустити, що ритміка фрази в подібних реченнях не впливає на величину їхнього частотного діапазону. Природно, щоб даний висновок мав більш загальний характер, необхідно досліджувати різні ритмічні структури на фразах з різним інтонаційним оформленням.

Практика застосування критеріїв розходження показує, що критерій Ван дер Вардена має більшу чутливість до розрізнення, і отримані висновки завжди погодяться або можуть бути пояснені лінгвістичним аналізом. Умовою для застосування критерію  $X$  є вимога, щоб різниця кількостей порівнюваних варіант не перевищувала 5 (тобто  $|n_1 - n_2| \leq 5$ ).

### 4.3. Контрольні запитання

1. Чим визначається постановка лінгвістичної задачі?
2. Коли використовується критерій Стьюдента?
3. В чому полягає метод порівнянь?
4. До якої групи критеріїв належить критерій Ван дер Вардена?
5. Коли може бути застосований критерій Ван дер Вардена?

## 5. ВИВЧЕННЯ ЗАЛЕЖНОСТІ ЛІНГВІСТИЧНИХ ОЗНАК

### 5.1. Кореляційна залежність

Для встановлення загальних законів, по яких протікають лінгвістичні явища, необхідно не тільки проаналізувати кожний з компонентів якого-небудь явища, але представити детальну кількісну і якісну характеристику різних зв'язків, що існують між ознаками, що цікавлять дослідника процесу, і описати результати впливу даних зв'язків на досліджувані процеси [5].

Прояв однієї ознаки перебуває в тісному зв'язку з багатьма іншими ознаками досліджуваного явища. Внутрішня структура явища характеризується багатопрічинним зв'язком. Зведення складної системи відносин до більш простих її видів, виділення тих зв'язків, які є основними в досліджуваній функції мовного явища - це ті задачі, які повинні бути вирішені експериментатором на початковому етапі дослідження.

Теоретично всі зв'язки, з узагальнення яких виникають закони науки, можна розділити на два види – функціональна залежність і кореляційна залежність. При *функціональній залежності* будь-якому фіксованому значенню однієї ознаки відповідає строго визначене, завжди те саме, значення іншої ознаки. При *кореляційній залежності* фіксованому значенню однієї ознаки можуть відповідати кілька значень іншої ознаки, причому до проведення експерименту це відповідне значення другої ознаки дослідникові невідомо. Причина цього полягає в тому, що в жодному експерименті, загалом кажучи, не можна повністю врахувати вплив другорядних факторів. У лінгвістичних явищах всі зв'язки, як правило, кореляційні. І в тих випадках, коли лінгвістичне явище описується рядом ознак (факторів), що мають якісне вираження, характер і ступінь залежності ознак можна вивчити за допомогою кореляційного аналізу.

Кореляційний зв'язок може розглядатися з погляду його «тісноти» і «форми». Під *тісністю* кореляційного зв'язку розуміється сила впливу досліджуваних ознак однієї на одну. По *тісноті* кореляція може бути *слабкою*, *середньою* й *сильною*. *Форма* кореляційного зв'язку показує, як у середньому змінюються значення однієї ознаки при зміні іншої. За *формою* кореляція може бути *прямолінійною (лінійною)* або *криволінійною*. Прямолінійна форма кореляційного зв'язку виникає тоді, коли рівним змінам першої ознаки відповідають рівні (у середньому) по величині й за знаком зміни другої ознаки.

**Приклад 5.1.** Нехай є дві ознаки,  $X$  і  $Y$ , що характеризують одне й те ж лінгвістичне явище. У результаті експерименту були отримані наступні значення:

$x$	4	6	9	10	12
$y$	20	28	30	35	37

**Повна пряма  
кореляція**

З отриманих даних видно, що зі збільшенням кожного значення ознаки  $X$  збільшується кожне значення ознаки  $Y$ . Залежність, при якій зі збільшенням (або зменшенням) кожного значення однієї ознаки збільшується (або зменшується) кожне значення іншої ознаки, називається *повною*.

В іншому випадку ознаки  $X$  і  $Y$  одержали такі значення:

$x$	4	6	9	10	12
$y$	30	20	35	28	37

**Часткова пряма  
кореляція**

Наведені дані показують, що зі збільшенням значень ознаки  $X$  збільшується не кожне значення ознаки  $Y$ . Залежність називається *частковою*, коли в середньому зі збільшенням (або зменшенням) значень однієї ознаки збільшуються (зменшуються) значення іншої ознаки. Повна й часткова кореляція за своєю формою є *прямою*, якщо збільшення (зменшення) значень однієї ознаки спричиняє збільшення (зменшення) значень іншої ознаки. У тому випадку, коли зі збільшенням (зменшенням) значень однієї ознаки зменшуються (збільшуються) значення іншої ознаки, кореляція називається *зворотною*.

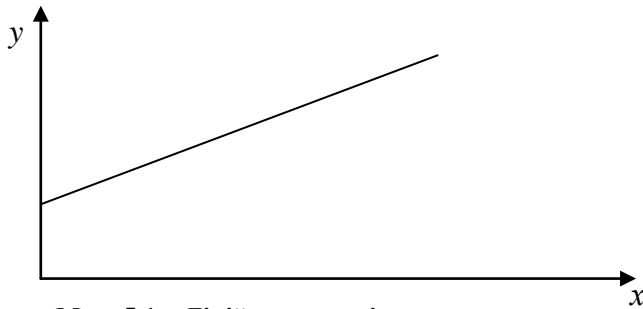
<i>x</i>	4	6	9	10	12
<i>y</i>	37	35	30	28	20

**Повна зворотна  
кореляція**

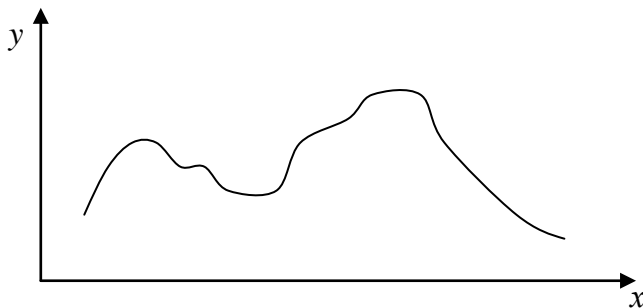
<i>x</i>	4	6	9	10	12
<i>y</i>	37	28	35	20	30

**Часткова зворотна  
кореляція**

При лінійній залежності розташування значень ознак  $X$  і  $Y$  можна подати графічно уздовж однієї прямої. Ця пряма називається *лінією регресії*. Залежність вважається *лінійною*, якщо рівним збільшенням значень однієї ознаки відповідають більш-менш рівні зміни того самого знака іншої ознаки. Якщо рівним збільшенням значень однієї ознаки відповідають різні зміни як за знаком, так і по величині, іншої ознаки, то така залежність називається *криволінійною*.



Мал. 5.1 – Лінійна залежність двох ознак



Мал. 6.2 – Криволінійна залежність двох ознак

Кореляційному аналізу завжди повинен передувати лінгвістичний аналіз досліджуваних ознак. Якщо в результаті лінгвістичного аналізу виявлено, що між двома ознаками якого-небудь лінгвістичного явища існує зв'язок, то застосування кореляційного аналізу дає можливість виявити ступінь і характер даного зв'язку, або відсутність зв'язку між даними ознаками. При цьому необхідно враховувати вплив інших ознак на досліджувані. У цьому випадку, коли ці додаткові ознаки дуже впливають (цей висновок робиться тільки на основі лінгвістичного аналізу), то дані кореляційного аналізу не будуть відбивати реальну картину залежності досліджуваних двох ознак. У таких випадках застосовується парціальна кореляція.

Кореляційна залежність двох ознак не має двостороннього характеру, тобто некомутативна. Точніше кажучи, якщо вплив 1-ї ознаки на 2-гу має певну міру й форму зв'язку,

то тіснота й форма кореляційного зв'язку між 2-ю і 1-ю ознаками може бути іншою. Але в тих випадках, коли кореляційний зв'язок між двома ознаками є лінійним, ступені впливу однієї ознаки на іншу однакові. Цим пояснюється те, що надалі велику увагу буде приділено саме лінійному кореляційному зв'язку.

## 5.2. Лінійна кореляція

Припустимо, що є значення двох ознак  $X$  і  $Y$  якогось лінгвістичного явища. У кожному прояві явища обидві ці ознаки приймають певне числове значення. У результаті численних спостережень було помічено, що ознаки  $X$  і  $Y$  взаємозалежні. Зв'язок може прийняти різний вид, але ми вправі припустити найпростішу залежність між ними – прямолінійну (лінійну). Лінійний зв'язок є найпростішим й досить часто зустрічається. Мірою тісноти лінійного зв'язку служить *коефіцієнт кореляції*  $r$ , обумовлений формулою

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Тут, як звичайно,  $\bar{x}$  (відповідно  $\bar{y}$ ) позначає середнє арифметичне значення ознаки  $x$  за спостереженнями  $x_1, x_2, \dots, x_n$  (відповідно для  $y$ ).



**Приклад 5.2.** Лінгвістичний аналіз показав, що між компонентами словесного наголосу – інтенсивністю, тривалістю й сумарною енергією існують складні зв'язки, і що роль цих компонентів у створенні ефекту наголошеності складу не однакова. Виникає задача про об'єктивне вивчення ступеня й характеру зв'язку між інтенсивністю, тривалістю й сумарною енергією методами математичної статистики. Для визначення кореляційного зв'язку між тривалістю й інтенсивністю наголошеного голосного кінцевого складу був проведений експеримент із 50 спостережень. Експериментальний матеріал складався зі слів того самого звукового складу й однорідної акцентної структури. Були обчислені показники основних акустичних характеристик – інтенсивності, тривалості й сумарної енергії наголошених і ненаголошених складів. Позначимо тривалість через  $X$  й інтенсивність через  $Y$ . Якщо за результатами експерименту ми в змозі визначити значення цих ознак у кожному експерименті (спостереженні), то ми одержимо сукупність пар  $(x_1, y_1), \dots, (x_n, y_n)$ . Тут  $n$  – кількість спостережень в експерименті, а пари  $(x_i, y_i)$  означають, що в  $i$ -тому спостереженні ознака  $X$  прийняла значення  $x_i$ , а ознака  $Y$  – значення  $y_i$ . Розрахунки для обчислення коефіцієнта кореляції наведені в таблиці 5.1, де  $X$  – тривалість наголошеного голосного кінцевого складу,  $Y$  – інтенсивність наголошеного голосного кінцевого складу.

Таблиця 5.1

№	1		2		3		4
	Значення ознаки		Відхилен. від серед. арифм.		Квадрати відхилень від серед. арифм.		Добутки відхилень
	$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	3	4	5	6	7	8
1	350	17	+110	+3.5	12100	12.25	+385
2	320	17	+80	+3.5	6400	12.25	+280
3	360	16	+120	+2.5	14400	6.25	+300
4	360	15	+120	+1.5	1440	2.25	+180
5	340	17	+100	+3.5	10000	12.25	+350
6	240	10.5	0	- 3.0	0	9.0	0
7	200	9.5	- 40	+4.0	160	16.0	+160
8	260	14.5	+20	- 1.0	400	1.0	+20
9	200	10.5	- 40	- 3.0	1600	9.0	+120
10	240	11	0	- 2.5	0	6.25	0
11	290	16.5	- 50	+3.0	2500	9.0	- 150
12	220	19.5	- 20	+6.0	400	36.0	- 120
13	200	16	- 40	+2.5	1600	6.25	- 100
14	220	10	- 20	- 3.5	400	12.25	+70
15	300	13	+60	- 0.5	3600	0.25	- 30
16	230	16	- 10	+2.5	100	6.25	- 25
17	270	15	+30	+1.5	900	2.25	+45
18	300	15.5	+60	- 2.0	3600	4.0	+120
19	300	19.5	+60	+6.0	3600	36.0	+360
20	270	19.5	+30	+6.0	900	36.0	+180
21	250	20	+10	+6.5	100	42.25	+65
22	250	15	+10	+1.5	100	2.25	- 15
23	310	16	+70	+2.5	4900	6.25	+175
24	320	14.5	+80	+1.0	6400	1.0	+80
25	280	11	+40	- 2.5	1600	6.25	- 100
26	230	15	- 10	+1.5	100	2.25	- 15
27	250	14	+10	+0.5	100	0.25	+5
28	280	19	+40	+5.5	1600	30.25	+220
29	230	13	- 10	- 0.5	100	0.25	+5
30	270	17	+30	+3.5	900	12.25	+105

31	250	16	+10	+2.5	100	6.25	+2.5
32	200	16	- 40	+2.5	1600	6.25	- 100
33	180	14	- 60	+0.5	3600	0.25	- 30
34	240	10	0	- 3.5	0	12.25	0
35	200	14	- 40	+0.5	1600	0.25	- 20
36	270	13	+30	- 0.5	900	0.25	- 15
37	170	11	- 70	- 2.5	4900	6.25	+175
38	190	13	- 50	- 0.5	2500	0.25	+25
39	130	3.5	- 110	- 10.0	12100	100.0	+1100
40	130	4	- 110	- 9.0	12100	90.25	+1054
41	150	7	- 90	- 6.5	8100	42.25	+585
42	150	3	- 90	- 10.5	8100	110.25	+945
43	220	9.5	- 20	- 4.0	400	16.0	+80
44	220	17.5	- 20	- 4.0	400	16.0	+80
45	240	14	0	+0.5	0	0.25	0
46	240	16	0	+2.5	0	6.25	0
47	2400	18	0	+4.5	0	20.25	0
48	180	7	- 60	- 6.5	3600	42.25	+390
49	180	9	- 60	- 4.5	3600	20.25	+270
50	180	6	- 60	- 7.5	3600	56.25	+450
<b>Разом</b>	<b>12000</b>	<b>674.5</b>			<b>161600</b>	<b>890.25</b>	<b>7545</b>

$$r = \frac{7545}{\sqrt{161600 \cdot 890}} = \frac{7545}{11928} = 0.63.$$

З наведеного приклада видно, що безпосередній підрахунок  $r$  являє собою досить громіздку процедуру, особливо при великих  $n$ . Значення  $r$  також можна визначити, використовуючи метод групування даних у кореляційних ґратах, що істотно скорочує кількість необхідних обчислень. До того ж, кореляційні ґрати дозволяють проводити попередній статистичний аналіз характеру кореляційної залежності. Для побудови кореляційних ґрат діємо в такий спосіб: по спостережуваним значенням ознаки  $X$  (відповідно,  $Y$ ), тобто по числах  $x_1, x_2, \dots, x_n$  (відповідно,  $y_1, y_2, \dots, y_n$ ) визначаємо інтервали групування. Рекомендується, щоб кількість інтервалів групування по кожній ознаці перебувала в межах від 8 до 15 (для кожної ознаки  $X$  і  $Y$  своя кількість інтервалів). Не рекомендується розбивати отримані дані на занадто малу кількість інтервалів, тому що це може привести до зменшення точності вимірів тісноти зв'язку між ознаками.

**Приклад 5.3.** Для умови приклада 5.2 обчислимо коефіцієнт кореляції з використанням кореляційних ґрат. Для цього всі отримані виміри ознаки  $X$  (тривалості) розбиваємо на 8 інтервалів, а всі виміри ознаки  $Y$  (інтенсивності) – на 9 інтервалів. Отримані дані наведені в таблиці 5.2.

З наведеної таблиці видно, що крім кореляційних ґрат, вона містить ще ряд додаткових рядків і стовпців. Але насамперед розглянемо самі кореляційні ґрати (на числа кліток, що розташовані у дужках, поки не обертаємо уваги). Видно, що хоча заповнені клітки й не лежать строго на діагоналі, що з'єднує лівий верхній і правий нижній кути таблиці, все-таки зосереджені біля цієї діагоналі. Тому природно очікувати, що кореляційний зв'язок між тривалістю й інтенсивністю близький до лінійного (причому зі збільшенням тривалості збільшується й інтенсивність). Розглянемо тепер зміст додаткових рядків і стовпців таблиці 5.2.

Перший стовпець (позначений  $n_x$ ) складається із чисел, що дорівнюють сумі чисел, зазначених у заповнених клітках відповідного рядка кореляційних ґрат. Підраховуємо, скільки разів значення ознаки  $X$  у кожному з інтервалів зустрічається в кожному з інтервалів ознаки  $Y$ . Так, перше число стовпця  $n_x$ , тобто число 4, означає, що значенню тривалості з інтервалу 130 – 160 відповідає 4 спостереження в експерименті, а саме, три зі значенням інтенсивності в інтервалі 3 – 5 і одне в інтервалі 7 – 9.

Стовпець і рядок №2 (позначені відповідно  $c_x$  і  $c_y$ ) являють собою штучні шкали. Для одержання елементів другого стовпця вибирають інтервал ознаки  $X$ , приписують йому значення 0, а іншим послідовно значення – 1, – 2 і т.д. (якщо ці інтервали передують зафіксованому) і значення

1, 2, 3 і т.д. (для наступних за зафіксованим інтервалів). Звичайно за нуль приймають одну із центральних частот (переважно максимальну). У розглянутому прикладі значення 0 приписане інтервалу 250 – 280.

Третій стовпець (позначений як  $n_x c_x$ ) складається з попарних добутоків елементів 1-го й 2-го стовпців (з урахуванням знаків).

Четвертий стовпець (позначений як  $n_x c_x^2$ ) складається з попарних добутоків елементів 1-го стовпця на квадрати елементів 2-го стовпця. Всі добутки беруться зі знаком «+».

Аналогічним образом будуються елементи чотирьох додаткових рядків.

Щоб одержати елементи 5-го стовпця (рядка), спочатку визначають числа в круглих дужках усередині кореляційних грат. Ці числа є добутками відповідних елементів 2-го стовпця й 2-го рядка (з урахуванням знаків). Так, в 1-му рядку кореляційних грат під числом 3 у дужках розташоване число 12, що є результатом добутку числа 3 на число 4. Тепер для одержання елементів стовпця № 5 (позначеного як  $\Sigma$ ) складаються попарні добутки чисел із кліток відповідного рядка кореляційних грат на відповідні числа в круглих дужках (з урахуванням знака). Результати добутоків відповідного рядка складаються. Так, число 42 у стовпці № 5 дорівнює  $3 \cdot 12 + 6 \cdot 1$ . Аналогічно знаходять елементи рядка № 5.

Таблиця 5.2

X	Y									$n_x$	$c_x$	$n_x c_x$	$n_x c_x^2$	$\Sigma_1$
	3-5	5-7	7-9	9-11	11-13	13-15	15-17	17-19	19-21					
130-160	3 (12)		1 (6)							4	- 3	- 12	36	42
160-190		1 (6)	1 (4)	1 (2)	1 (10)	1 (- 2)				5	- 2	- 10	20	10
190-220				2 (1)		2 (- 2)	3 (- 2)			7	- 1	- 7	7	- 6
220-250				4 (0)	1 (0)	2 (0)	3 (0)	2 (0)	1 (0)	13	0	0	0	0
250-280						3 (1)	3 (2)	1 (3)	2 (4)	9	1	9	9	20
280-310					1 (0)	1 (2)	1 (4)		2 (8)	5	2	10	20	22
310-340						1 (3)	1 (6)	1 (9)		3	3	9	27	18
340-370							2 (8)	2 (12)		4	4	16	64	40
$n_y$ (1)	3	1	2	7	3	10	13	6	5	50		15	183	146
$c_y$ (2)	- 4	- 3	- 2	- 1	0	1	2	3	4					
$n_y c_y$ (3)	- 12	- 3	- 4	- 7	0	10	26	18	20	48				
$n_y c_y^2$ (4)	48	9	8	7	0	10	52	54	80	268				
$\Sigma_2$	36	6	10	4	0	4	26	36	24	146				

Тепер для визначення коефіцієнта кореляції обчислюємо:

а) число, що дорівнює сумі елементів стовпця № 1 (перевіркою потрібно переконатися, що таке ж значення  $n$  дає сума елементів рядка № 1 (тобто рядка  $n_y$ ). У розглянутому прикладі  $n=50$ ;

б) суму елементів стовпця № 3 позначимо через  $X_1$ , а суму елементів рядка № 3 – через  $Y_1$ . У нас  $X_1=15$ ,  $Y_1=48$ ;

в) суму елементів стовпця № 4 позначимо через  $X_2$ , а суму елементів рядка № 4 – як  $Y_2$ . У нас  $X_2=183$ ,  $Y_2=288$ ;

г) суму елементів стовпця № 5 позначається через  $Z$ , у нас  $Z=146$ . Таке ж значення  $Z$  повинне вийти (і це повинне служити як перевірка) як сума елементів рядка № 5.

Обчислимо тепер  $\Sigma_{xy}=Z - \frac{X_1 Y_1}{n}$ ,  $\Sigma_{xx}=X_2 - \frac{X_1^2}{n}$ ,  $\Sigma_{yy}=Y_2 - \frac{Y_1^2}{n}$ . Звідси

$$r = \frac{\Sigma_{xy}}{\sqrt{\Sigma_{xx} - \Sigma_{yy}}}$$

Для розглянутого приклада

$$\Sigma_{xy} = 146 - \frac{15 \cdot 48}{50} = 146 - 14.4 = 131.6;$$

$$\Sigma_{xx} = 183 - \frac{15^2}{50} = 183 - 4.5 = 178.5;$$

$$\Sigma_{yy} = 268 - \frac{48^2}{50} = 268 - 46 = 222.$$

Отже,

$$r = \frac{131.6}{\sqrt{178.5 \cdot 222}} = 0.66.$$

Для того, щоб зробити висновок про кореляційну залежність ознак  $X$  і  $Y$  варто врахувати, що коефіцієнт кореляції  $r$  може приймати значення від  $-1$  до  $+1$ , причому негативне значення для  $r$  говорить про те, що залежність між  $X$  і  $Y$  зворотна (тобто з ростом однієї ознаки інша зменшується), а при позитивних значеннях  $r$  зв'язок прямий (з ростом однієї ознаки збільшується й інша). Крім того, чим ближче  $r$  до  $\pm 1$ , тим сильніше лінійний зв'язок між  $X$  і  $Y$ , а коли  $r$  близько до нуля, то між  $X$  і  $Y$  лінійний зв'язок відсутній. Однак при цьому цілком можливо, що між  $X$  і  $Y$  має місце кореляційний зв'язок, і навіть сильний, але криволінійний.

### 5.3. Кореляційні відносини

Будь-який вид зв'язку двох ознак можна звести до прямолінійної або криволінійної форми залежності. Тіснота зв'язку ознак як при прямолінійній, так і при криволінійній кореляції, визначається так званими кореляційними відносинами. На відміну від коефіцієнта кореляції, кореляційні відносини виражають однобічний зв'язок ознак як у випадку лінійної, так і у випадку криволінійної кореляції, тобто залежність  $X$  від  $Y$ , позначувану через  $e_{x/y}$ , і залежність  $Y$  від  $X$ , позначувану через  $e_{y/x}$ . Величина  $e$  завжди позитивна. Вона змінюється від 0 до 1. Якщо вона приймає значення, близьке або рівне одиниці, то говорять про тісний зв'язок (або повну кореляцію) досліджуваних ознак, і, навпаки, значення  $e$ , близьке або рівне нулю, свідчить про відсутність якої-небудь залежності між досліджуваними ознаками.

Для визначення  $e_{x/y}$ , і  $e_{y/x}$  також використовують кореляційну таблицю, але при цьому для зручності обчислень додають ще два стовпці № 6 і № 7 і два рядки № 6 і № 7. Щоб одержати елементи 6-го стовпця (позначеного як  $y_x$ ), варто помножити числа відповідного рядка кореляційних грат на відповідні елементи 2-го рядка (позначеної як  $c_y$ ). Сума цих попарних добутків і дає значення відповідного елемента стовпця № 6. Елементи 7-го стовпця (7-го рядка) дорівнюють квадратам елементів 6-го стовпця (6-го рядка), діленим на відповідні елементи 1-го стовпця (1-го рядка). Позначимо через  $S_1$  (відповідно  $S_2$ ) суму всіх елементів 7-го стовпця (відповідно, 7-го рядка) і обчислимо  $S_x = X_2 - S_2$ ,  $S_y = Y_2 - S_1$ . Значення  $X_2$  і  $Y_2$  вже були знайдені при обчисленні коефіцієнта кореляції  $r$ . З урахуванням того, що кількість інтервалів розбивки ознаки  $X$  дорівнює  $k$ , а кількість інтервалів розбивки ознаки  $Y$  дорівнює  $l$ , одержуємо формули для обчислення кореляційних відносин:

$$e_{x/y} = \sqrt{1 - \frac{n-1}{n-k} \cdot \frac{S_x}{\Sigma_{xx}}}, \quad e_{y/x} = \sqrt{1 - \frac{n-1}{n-l} \cdot \frac{S_y}{\Sigma_{yy}}}.$$

Якщо об'єм вибірки, тобто  $n$ , невеликий (наприклад, менше 30), то можна користуватися спрощеними формулами для  $e_{x/y}$  і  $e_{y/x}$ :

$$e_{x/y} = \sqrt{1 - \frac{S_x}{\Sigma_{xx}}}, \quad e_{y/x} = \sqrt{1 - \frac{S_y}{\Sigma_{yy}}}.$$

**Приклад 5.4.** Обчислимо кореляційні відносини для ознак  $X$  і  $Y$  в умовах приклада 5.2. Для цього необхідно таблицю 5.2 доповнити ще двома стовпцями й ще двома рядками, значення елементів яких обчислюється наведеним способом. Так, наприклад, перший елемент 6-го стовпця дорівнює  $3 \cdot (-4) + 1 \cdot (-2) = -14$ , другий елемент дорівнює  $1 \cdot (-3) + 1 \cdot (-2) + 1 \cdot (-1) + 1 \cdot 0 + 1 \cdot 1 = -5$ , і т.д. Аналогічно визначаються елементи 6-го рядка. Так, його 4-й елемент дорівнює  $1 \cdot (-2) + 2 \cdot (-1) + 4 \cdot 0 = -4$ . Отже, у цілому таблиця 5.2 доповнюється такими стовпцями й строками:

6-й стовпець: - 14; - 5; 6; 14; 20; 11; 6; 10.

7-й стовпець: 49; 5; 5.1; 44.5; 24.2; 12; 6.

6-й рядок: 9; - 2; - 5; - 4; 0; 4; 13; 12; 6.

7-й рядок: 27; 4; 12.5; 2.3; 0; 1.6; 13; 24; 7.2.

Далі знаходимо:

$$S_1=179.8, S_2=91.6, S_x=183 - 91.6=91.4, S_y=268 - 179.8=88.2,$$

$$e_{x/y} = \sqrt{1 - \frac{50-1}{50-8} \cdot \frac{91.6}{178.5}} = 0.63, \quad e_{y/x} = \sqrt{1 - \frac{50-1}{50-9} \cdot \frac{88.2}{222}} = 0.72.$$

Отримані значення кореляційних відносин дозволяють затверджувати, що кожна з ознак  $X$  і  $Y$  (тобто тривалість і інтенсивність) досить сильно впливають одна на одну.

#### 5.4. Парціальна кореляція

У більшості випадків при лінгвістичних дослідженнях характер кореляції між двома ознаками повністю або частково залежить від того, що дві досліджувані ознаки або одна з них залежать, у свою чергу, від якої-небудь третьої ознаки. Для того, щоб точно оцінити зв'язок між двома ознаками, необхідно виключити залежність від третьої ознаки. Це можна зробити, зокрема, при постійному значенні третьої ознаки. Але на практиці це здійснити дуже важко, практично неможливо. Однак можна спробувати виключити залежність від третьої ознаки, зіставивши її з кожною з основних ознак. Коефіцієнт кореляції між двома ознаками  $X$  і  $Y$  при постійних значеннях третьої ознаки  $Z$  називається *парціальним коефіцієнтом кореляції*. Парціальний коефіцієнт кореляції може бути обчислений тільки у випадку строго лінійного зв'язку всіх трьох ознак за допомогою наступної формули:

$$r_{xy/z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

#### 5.5. Критерії кореляційного аналізу

##### 5.5.1. Критерій вірогідності залежності ознак $X$ і $Y$ .

Часто виникає необхідність визначити, чи дійсно залежить  $X$  від  $Y$ , або отримане по одній вибірці значення є чисто випадковим. Інакше кажучи, чи можна на підставі одного лише значення  $e_{x/y}$ , близького до одиниці, затверджувати, що в тій же генеральній сукупності ознака  $X$  у тому ж ступені залежить від ознаки  $Y$ . Щоб можна було поширити висновок за межі вибірки, перевіряють отримане значення  $e_{x/y}$  за допомогою формули

$$K = \frac{e_{x/y}}{m(e_{x/y})},$$

де  $m(e_{x/y}) = \frac{1 - e_{x/y}^2}{\sqrt{n}}$  - середня помилка обчислення. Нульова гіпотеза затверджує незалежність ознак

$X$  і  $Y$ . Відповідно до критерію, при  $K > 3$   $H_0$  відкидається, при  $K \leq 3$   $H_0$  приймається. Подібним же чином перевіряється вірогідність висновку щодо залежності ознаки  $Y$  від  $X$ .

**Приклад 5.5.** В умовах прикладів 5.2 і 5.3 перевіримо вірогідність залежності ознаки  $X$  від  $Y$  і ознаки  $Y$  від  $X$ .

$$m(e_{x/y}) = \frac{1 - e_{x/y}^2}{\sqrt{n}} = \frac{0.596}{7.07} = 0.08, K = \frac{e_{x/y}}{m(e_{x/y})} = \frac{0.63}{0.08} = 8 > 3.$$

Отже, гіпотеза  $H_0$  відкидається з імовірністю 0.99, тобто ознака  $X$  залежить від  $Y$ .

$$m(e_{y/x}) = \frac{1 - e_{y/x}^2}{\sqrt{n}} = \frac{0.475}{7.07} = 0.07, K = \frac{e_{y/x}}{m(e_{y/x})} = \frac{0.72}{0.07} = 10.3 > 3.$$

Отже, і в цьому випадку гіпотеза  $H_0$  відкидається з імовірністю 0.99, тобто варто визнати, що ознака  $Y$  залежить від  $X$ .

### 5.5.2. Критерій значимості розходження кореляційних відносин

Часто виникає необхідність знати, чи можна вважати значення  $e_{x/y}$  і  $e_{y/x}$  рівними (тобто їхні розходження є чисто випадковим фактором) або необхідно встановити, що їхні розходження значимі, а виходить, можна затверджувати, що одне з них завжди більше іншого в генеральній сукупності. Отже, нульова гіпотеза затверджує, що  $e_{x/y} = e_{y/x}$ . Для перевірки  $H_0$  використовується критерій

$$K = \frac{(e_{x/y} - e_{y/x})^2}{m^2(e_{x/y}) + m^2(e_{y/x})}.$$

Якщо  $K > 9$ , то гіпотеза  $H_0$  відкидається з імовірністю 0.99. Якщо  $K \leq 9$ , то з тією же ймовірністю гіпотеза  $H_0$  приймається.

**Приклад 5.6.** Для розглянутих умов  $K = \frac{(0.72 - 0.63)^2}{0.08^2 + 0.07^2} = \frac{0.09^2}{0.0113} = 0.71 < 9$ . Отже, гіпотеза  $H_0$  приймається, тобто  $e_{x/y} = e_{y/x}$ .

## 5.6. Регресійний аналіз

З кореляційного аналізу відомо, як перевірити наявність залежності між ознаками й установити тісноту цього зв'язку. Кореляційні відносини дають нам ступінь одностороннього зв'язку ознак, але не вказують, наскільки змінюється одна ознака зі зміною іншої. Часто буває необхідно довідатися, як поводить ся ознака  $X$  із збільшенням (або зменшенням) значення ознаки  $Y$ . На це питання дає відповідь регресійний аналіз. Залежно від форми зв'язку ознак розрізняють: лінійну й криволінійну регресію. Найпростішою формою зв'язку є лінійна. При строго функціональній лінійній залежності двох ознак  $X$  і  $Y$  їхня форма зв'язку геометрично виразиться однією прямою лінією. При кореляційній лінійній залежності (коли коефіцієнт кореляції, загалом кажучи, відмінний від  $\pm 1$ ) однією прямою лінією цей зв'язок виразити не вдається. Однак за допомогою двох прямих можна показати, як кількісні зміни однієї ознаки впливають на кількісні зміни іншої ознаки. Ці дві прямі називаються *лініями регресії*. У загальному випадку, тобто при криволінійному кореляційному зв'язку, можна за допомогою двох кривих описати вплив однієї ознаки на іншу. Будемо розглядати тільки той випадок, коли регресійний зв'язок виражається за допомогою прямих ліній регресії. Перша з них називається *лінією регресії  $Y$  по  $X$*  и показує, як змінюється  $Y$  при зміні  $X$ . Її рівняння має вигляд

$$y - \bar{y} = b_{x/y}(x - \bar{x}).$$

Щоб визначити  $x$ ,  $y$ ,  $b_{x/y}$  варто звернутися до кореляційної таблиці 5.2. Тоді

$$\bar{x} = c_x + \frac{\sum n_x c_x}{n} d_x, \quad \bar{y} = c_y + \frac{\sum n_y c_y}{n} d_y, \quad b_{x/y} = \frac{\sum xy}{\sum xx} \cdot \frac{d_y}{d_x}$$

Тут  $c_x$  (відповідно  $c_y$ ) – середина інтервалу для ознаки  $X$  (відповідно  $Y$ ), якому приписане значення

0, а  $d_x$  (відповідно  $d_y$ ) – довжини інтервалів групування ознаки  $X$  (відповідно  $Y$ ). Для розглянутого приклада 5.3 маємо:

$$\bar{x} = 235 + \frac{15}{50} \cdot 30 = 244; \quad \bar{y} = 12 + \frac{48}{50} \cdot 2 \approx 14; \quad b_{x/y} = \frac{131.6}{178.5} \cdot \frac{2}{30} = 0.05.$$

Таким чином, рівняння прямої лінії регресії  $Y$  по  $X$  має вигляд:

$$y - 14 = 0.05(x - 244).$$

Коефіцієнт 0.05 показує, що при збільшенні (зменшенні) значення ознаки  $X$  на одиницю ознака  $Y$  збільшується (зменшується) на 0.05.

Аналогічно лінія регресії ознаки  $X$  по  $Y$  задається рівнянням

$$x - \bar{x} = b_{y/x}(y - \bar{y}).$$

Значення  $\bar{x}$  й  $\bar{y}$  визначаються по наведеним вище формулах, а значення  $b_{y/x}$  задається співвідношенням

$$b_{y/x} = \frac{\sum_{xy}}{\sum_{yy}} \cdot \frac{d_x}{d_y}.$$

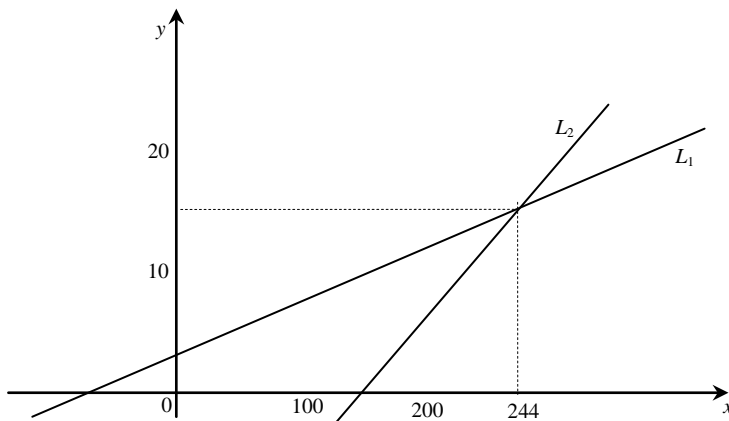
Для приклада 5.3 маємо:

$$b_{y/x} = \frac{131.6}{222} \cdot \frac{30}{2} = 8.9.$$

Тоді рівняння прямої лінії регресії  $X$  по  $Y$  приймає вид:

$$x - 244 = 8.9(y - 14).$$

На графіку ці прямі мають вигляд:



Мал. 5.3 –  $L_1$  – лінія регресії  $Y$  по  $X$ ;  $L_2$  – лінія регресії  $X$  по  $Y$

Практичне застосування цих ліній регресії таке, що якщо дослідник задається деяким значенням ознаки  $X$  (наприклад, у розглянутому випадку задамося значенням тривалості наголошеного голосного 200 мсек), то в середньому варто очікувати, що ознака  $Y$  прийме значення ординати відповідної точки прямої  $L_1$  (у розглянутому прикладі  $Y \approx 11.8$  мм). І навпаки, якщо задаватися значенням ознаки  $Y$ , то розглядаючи пряму  $L_2$ , одержуємо середнє очікуване значення ознаки  $X$ . Так, при інтенсивності 7 мм тривалість у середньому дорівнює 181.7 мсек.

Однак варто помітити, що користуватися графіком ліній регресії в зазначеному вище змісті можна тільки в певних межах. Іншими словами, задаватися значенням одного з ознак можна тільки в межах найменшого - найбільшого зі значень цієї ознаки, отриманих у даній вибірці. Так, лінією регресії  $L_1$  із приклада 5.2 можна користуватися, задаючись значеннями ознаки тривалості тільки з інтервалу 130 - 370.

При обчисленні коефіцієнтів  $b_{y/x}$  і  $b_{x/y}$  можуть вийти й негативні значення, і знак мінус у відповідного коефіцієнта говорить лише про те, що зі збільшенням однієї ознаки інша зменшується. У всьому іншому зміст ліній регресії аналогічний уже описаному. Величина кута між лініями регресії  $L_1$  і  $L_2$  характеризує тісноту зв'язку ознак  $X$  і  $Y$  – чим сильніше залежність ознак, тим менше кут між прямими. Коли цей кут близький до  $90^\circ$ , то практично лінійний зв'язок між ознаками відсутній. У цьому випадку доцільно побудувати емпіричну криву регресії  $Y$  по  $X$  (і аналогічно  $X$  по

У). Процес побудови емпіричної лінії регресії виглядає в такий спосіб. Нехай  $(x_1, y_1), \dots, (x_n, y_n)$  –  $n$  пар отриманих значень ознак  $X$  і  $Y$ , і нехай  $X_1, \dots, X_k$  – всі різні значення  $X$  серед чисел  $x_1, \dots, x_n$  (деякі з  $x_i$  можуть збігатися, так що  $k \leq n$ ). Якщо значення  $X_1$  зустрілося серед чисел  $x_1, \dots, x_n$  рівно  $n_1$  раз, то йому відповідає  $n_1$  значень  $Y$  серед чисел  $y_1, \dots, y_n$ . Позначимо через  $\bar{Y}_1$  середнє значення цих значень  $Y$ . Одержуємо пари  $(X_1, \bar{Y}_1)$ . Аналогічно знаходять пари  $(X_2, \bar{Y}_2), \dots, (X_k, \bar{Y}_k)$ . Пари точок  $(X_1, \bar{Y}_1), \dots, (X_k, \bar{Y}_k)$  зображуємо на малюнку, а потім побудовані точки з'єднуємо прямими відрізками. Отримана ламана лінія і є емпіричною лінією регресії  $Y$  по  $X$ .

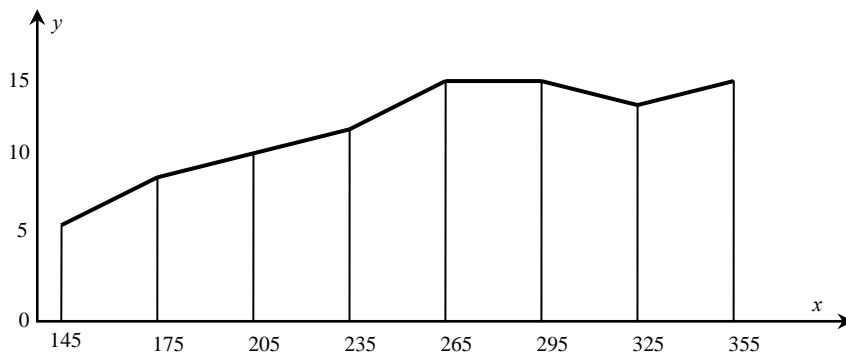
Якщо в прикладі 5.2 середини інтервалів прийняти за справжні значення, які приймають ознаки  $X$  і  $Y$  у вибірці, то одержимо ряд значень:  $X_1=145, X_2=175, X_3=205, X_4=235, X_5=265, X_6=295, X_7=325, X_8=355$ . Числа  $n_1, \dots, n_8$  по суті є числами з першого стовпця кореляційної таблиці 5.2. Щоб знайти  $\bar{Y}_1$ , урахуємо, що  $X_1=145$  зустрілося три рази в парі з  $Y=4$  (середина інтервалу 3 – 5) і один раз у парі з  $Y=8$  (середина інтервалу 7 – 9). Звідси

$$\bar{Y}_1 = \frac{3 \cdot 4 + 1 \cdot 8}{4} = 5.$$

Аналогічно знаходимо:

$$\bar{Y}_2=10, \bar{Y}_3=13.7, \bar{Y}_4=14.1, \bar{Y}_5=16.4, \bar{Y}_6=16.4, \bar{Y}_7=16, \bar{Y}_8=17.$$

Тепер зображуємо крапки  $(X_i, \bar{Y}_i)$  і будуємо ламану.



Мал.. 5.4 - Емпірична лінія регресії

Коли кількість побудованих точок досить велика, їх вдається з'єднати плавною кривою, що дає уявлення про теоретичну лінію регресії.

Варто помітити, що навіть коли залежність ознак  $X$  і  $Y$  не є лінійною, можна з певним ступенем наближення (хоча б з метою якісного аналізу) використати прямі лінії регресії.

## 5.7. Контрольні запитання

1. Чим кореляційна залежність відрізняється від функціональної?
2. Якою може бути кореляція за тісністю?
3. Якою може бути кореляція за формою?
4. Яка кореляція є повною, а яка частковою?
5. Яка кореляція є прямою, а яка зворотною?
6. За якою формулою обчислюється коефіцієнт лінійної кореляції?
7. Яким чином інтерпретується числове значення коефіцієнту кореляції?
8. Що таке кореляційні грати?
9. Що виражають кореляційні відносини?
10. Яка кореляція називається парціальною?
11. Які вам відомі критерії кореляційного аналізу?
12. Як побудувати лінії регресії?
13. Що можна дослідити за допомогою ліній регресії?



## Додаток 1

### Значення функції нормального розподілу $\Phi(u)$

<i>u</i>	0	1	2	3	4	5	6	7	8	9
<b>-0.0</b>	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
<b>-0.1</b>	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
<b>-0.2</b>	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
<b>-0.3</b>	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
<b>-0.4</b>	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
<b>-0.5</b>	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
<b>-0.6</b>	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
<b>-0.7</b>	2420	2389	2358	2327	2297	2266	2236	2206	2177	2148
<b>-0.8</b>	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
<b>-0.9</b>	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
<b>-1.0</b>	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
<b>-1.1</b>	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
<b>-1.2</b>	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
<b>-1.3</b>	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
<b>-1.4</b>	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
<b>-1.5</b>	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
<b>-1.6</b>	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
<b>-1.7</b>	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
<b>-1.8</b>	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
<b>-1.9</b>	0288	0281	0274	0268	0262	0256	0250	0244	0239	0233
<b>-2.0</b>	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
<b>-2.1</b>	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
<b>-2.2</b>	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
<b>-2.3</b>	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
<b>-2.4</b>	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
<b>-2.5</b>	0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
<b>-2.6</b>	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
<b>-2.7</b>	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
<b>-2.8</b>	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
<b>-2.9</b>	0019	0018	0017	0016	0016	0015	0015	0014	0014	0014
	<b>-3.0</b>	<b>-3.1</b>	<b>-3.2</b>	<b>-3.3</b>	<b>-3.4</b>	<b>-3.5</b>	<b>-3.6</b>	<b>-3.7</b>	<b>-3.8</b>	<b>-3.9</b>
	0013	0010	0007	0005	0003	0002	0002	0001	0001	0000

Для визначення значень  $\Phi(u)$  при  $u \geq 0$  використовують співвідношення  $\Phi(u) = 1 - \Phi(-u)$ . Так, наприклад,  $\Phi(1.65) = 1 - \Phi(-1.65) = 1 - 0.0495 = 0.9505$ .

## Додаток 2

### Критичні значення $\chi^2$

<i>f</i>	5%	15%	<i>f</i>	5%	1%	<i>f</i>	5%	1%
<b>1</b>	3.8	6.6	<b>18</b>	28.9	34.8	<b>35</b>	49.8	57.3
<b>2</b>	6.0	9.21	<b>19</b>	30.1	36.2	<b>36</b>	51.0	58.6
<b>3</b>	7.8	11.3	<b>20</b>	31.4	37.6	<b>37</b>	52.2	59.9
<b>4</b>	9.5	13.3	<b>21</b>	32.7	38.9	<b>38</b>	53.4	61.2
<b>5</b>	11.1	15.1	<b>22</b>	33.9	40.3	<b>39</b>	54.6	62.4
<b>6</b>	12.6	16.8	<b>23</b>	35.2	41.6	<b>40</b>	55.8	63.7
<b>7</b>	14.1	18.5	<b>24</b>	36.4	43.0	<b>41</b>	56.9	65.0
<b>8</b>	15.5	20.1	<b>25</b>	37.7	44.3	<b>42</b>	58.1	66.2
<b>9</b>	16.9	21.7	<b>26</b>	38.9	45.6	<b>43</b>	59.3	67.5
<b>10</b>	18.3	23.1	<b>27</b>	40.1	47.0	<b>44</b>	60.5	68.7
<b>11</b>	19.7	24.7	<b>28</b>	41.3	48.3	<b>45</b>	61.7	70.0
<b>12</b>	21.0	26.2	<b>29</b>	42.6	49.6	<b>46</b>	62.8	71.2
<b>13</b>	22.4	27.7	<b>30</b>	43.8	50.9	<b>47</b>	64.0	72.4
<b>14</b>	23.7	29.1	<b>31</b>	45.0	52.2	<b>48</b>	65.2	73.7
<b>15</b>	25.0	30.6	<b>32</b>	46.2	53.5	<b>49</b>	66.3	74.9
<b>16</b>	26.3	32.0	<b>33</b>	47.4	54.8	<b>50</b>	67.5	76.2
<b>17</b>	27.6	33.4	<b>34</b>	48.6	56.1			

Нульова гіпотеза приймається при  $\chi^2 \leq \chi_{05}^2$  й відкидається при  $\chi^2 > \chi_{01}^2$ .

## Додаток 3

### Критичні значення $t$ (критерію Стьюдента)

$f$	Рівень довіри			$f$	Рівень довіри		
	95%	99%	99.9%		95%	99%	99.9%
1	12.71	63.60		21	2.08	2.83	3.82
2	4.30	9.93	31.60	22	2.07	2.82	3.79
3	3.18	5.84	12.94	23	2.07	2.81	3.77
4	2.78	4.60	8.61	24	2.06	2.80	3.75
5	2.57	4.03	6.86	25	2.06	2.79	3.73
6	2.45	3.71	5.96	26	2.06	2.78	3.71
7	2.37	3.50	5.41	27	2.05	2.77	3.69
8	2.31	3.36	5.04	28	2.05	2.76	3.67
9	2.26	3.25	4.78	29	2.04	2.76	3.66
10	2.23	3.17	4.59	30	2.04	2.75	3.65
11	2.20	3.11	4.44	40	2.02	2.70	3.55
12	2.18	3.06	4.32	50	2.01	2.68	3.50
13	2.16	3.01	4.22	60	2.00	2.66	3.46
14	2.15	2.98	4.14	80	1.99	2.64	3.42
15	2.13	2.95	4.07	100	1.98	2.63	3.39
16	2.12	2.92	4.02	120	1.98	2.62	3.37
17	2.11	2.90	3.97	200	1.97	2.60	3.34
18	2.10	2.88	3.92	500	1.96	2.59	3.31
19	2.09	2.86	3.88	$\infty$	1.96	2.58	3.29
20	2.09	2.85	3.85				
$f$	5%	1%	0.1%	$f$	5%	1%	0.1%
	Рівень значимості				Рівень значимості		

Нульова гіпотеза приймається при  $t \leq t_{05}$  й відкидається при  $t > t_{01}$ .

## Додаток 4

Значення функції  $\Psi$ 

$p \rightarrow$ $\downarrow$	0	1	2	3	4	5	6	7	8	9
0.00	$-\infty$	- 3.09	- 2.88	- 2.75	- 2.65	- 2.58	- 2.51	- 2.46	- 2.41	- 2.37
0.1	- 2.33	- 2.29	- 2.26	- 2.23	- 2.20	- 2.17	- 2.14	- 2.12	- 2.10	- 2.07
0.02	- 2.05	- 2.03	- 2.01	- 2.00	- 1.98	- 1.96	- 1.94	- 1.93	- 1.91	- 1.90
0.03	- 1.88	- 1.87	- 1.85	- 1.84	- 1.83	- 1.81	- 1.80	- 1.79	- 1.77	- 1.76
0.04	- 1.75	- 1.74	- 1.73	- 1.72	- 1.71	- 1.70	- 1.68	- 1.67	- 1.66	- 1.65
0.05	- 1.64	- 1.64	- 1.63	- 1.62	- 1.61	- 1.60	- 1.59	- 1.58	- 1.57	- 1.56
0.06	- 1.55	- 1.55	- 1.54	- 1.53	- 1.52	- 1.51	- 1.51	- 1.50	- 1.49	- 1.48
0.07	- 1.48	- 1.47	- 1.46	- 1.45	- 1.45	- 1.44	- 1.43	- 1.43	- 1.42	- 1.41
0.08	- 1.41	- 1.40	- 1.39	- 1.39	- 1.38	- 1.37	- 1.37	- 1.36	- 1.35	- 1.35
0.09	- 1.34	- 1.33	- 1.33	- 1.32	- 1.32	- 1.31	- 1.30	- 1.30	- 1.29	- 1.29
0.10	- 1.28	- 1.28	- 1.27	- 1.26	- 1.26	- 1.25	- 1.25	- 1.24	- 1.24	- 1.23
0.11	- 1.23	- 1.22	- 1.22	- 1.21	- 1.21	- 1.20	- 1.20	- 1.19	- 1.19	- 1.18
0.12	- 1.18	- 1.17	- 1.17	- 1.16	- 1.16	- 1.15	- 1.15	- 1.14	- 1.14	- 1.13
0.13	- 1.13	- 1.12	- 1.12	- 1.11	- 1.11	- 1.10	- 1.10	- 1.09	- 1.09	- 1.09
0.14	- 1.08	- 1.08	- 1.07	- 1.07	- 1.06	- 1.06	- 1.05	- 1.05	- 1.05	- 1.04
0.15	- 1.04	- 1.03	- 1.03	- 1.02	- 1.02	- 1.02	- 1.01	- 1.01	- 1.00	- 1.00
0.16	- 0.99	- 0.99	- 0.99	- 0.98	- 0.98	- 0.97	- 0.97	- 0.97	- 0.96	- 0.96
0.17	- 0.95	- 0.95	- 0.95	- 0.94	- 0.94	- 0.93	- 0.93	- 0.93	- 0.92	- 0.92
0.18	- 0.92	- 0.91	- 0.91	- 0.90	- 0.90	- 0.90	- 0.89	- 0.89	- 0.89	- 0.88
0.19	- 0.88	- 0.87	- 0.87	- 0.87	- 0.86	- 0.86	- 0.86	- 0.85	- 0.85	- 0.85
0.20	- 0.84	- 0.84	- 0.83	- 0.83	- 0.83	- 0.82	- 0.82	- 0.82	- 0.81	- 0.81
0.21	- 0.81	- 0.80	- 0.80	- 0.80	- 0.79	- 0.79	- 0.79	- 0.78	- 0.78	- 0.78
0.22	- 0.77	- 0.77	- 0.77	- 0.76	- 0.76	- 0.76	- 0.75	- 0.75	- 0.75	- 0.74
0.23	- 0.74	- 0.74	- 0.73	- 0.73	- 0.73	- 0.72	- 0.72	- 0.72	- 0.71	- 0.71
0.24	- 0.71	- 0.70	- 0.70	- 0.70	- 0.69	- 0.69	- 0.69	- 0.68	- 0.68	- 0.68
0.25	- 0.67	- 0.67	- 0.67	- 0.67	- 0.66	- 0.66	- 0.66	- 0.65	- 0.65	- 0.65
0.26	- 0.64	- 0.64	- 0.64	- 0.63	- 0.63	- 0.63	- 0.63	- 0.62	- 0.62	- 0.62
0.27	- 0.61	- 0.61	- 0.61	- 0.60	- 0.60	- 0.60	- 0.59	- 0.59	- 0.59	- 0.59
0.28	- 0.58	- 0.58	- 0.58	- 0.57	- 0.57	- 0.57	- 0.57	- 0.56	- 0.56	- 0.56
0.29	- 0.55	- 0.55	- 0.55	- 0.54	- 0.54	- 0.54	- 0.54	- 0.53	- 0.53	- 0.53
0.30	- 0.52	- 0.52	- 0.52	- 0.52	- 0.51	- 0.51	- 0.51	- 0.50	- 0.50	- 0.50
0.31	- 0.50	- 0.49	- 0.49	- 0.49	- 0.48	- 0.48	- 0.48	- 0.48	- 0.47	- 0.47
0.32	- 0.47	- 0.46	- 0.46	- 0.46	- 0.46	- 0.45	- 0.45	- 0.45	- 0.45	- 0.44
0.33	- 0.44	- 0.44	- 0.43	- 0.43	- 0.43	- 0.43	- 0.42	- 0.42	- 0.42	- 0.42
0.34	- 0.41	- 0.41	- 0.41	- 0.40	- 0.40	- 0.40	- 0.40	- 0.39	- 0.39	- 0.39
0.35	- 0.39	- 0.38	- 0.38	- 0.38	- 0.37	- 0.37	- 0.37	- 0.37	- 0.36	- 0.36
0.36	- 0.36	- 0.36	- 0.35	- 0.35	- 0.35	- 0.35	- 0.34	- 0.34	- 0.34	- 0.33
0.37	- 0.33	- 0.33	- 0.33	- 0.32	- 0.32	- 0.32	- 0.32	- 0.31	- 0.31	- 0.31
0.38	- 0.31	- 0.30	- 0.30	- 0.30	- 0.30	- 0.29	- 0.29	- 0.29	- 0.28	- 0.28
0.39	- 0.28	- 0.28	- 0.27	- 0.27	- 0.27	- 0.27	- 0.26	- 0.26	- 0.26	- 0.26
0.40	- 0.25	- 0.25	- 0.25	- 0.25	- 0.24	- 0.24	- 0.24	- 0.24	- 0.23	- 0.23
0.41	- 0.23	- 0.23	- 0.22	- 0.22	- 0.22	- 0.21	- 0.21	- 0.21	- 0.21	- 0.20
0.42	- 0.20	- 0.20	- 0.20	- 0.19	- 0.19	- 0.19	- 0.19	- 0.18	- 0.18	- 0.18
0.43	- 0.18	- 0.17	- 0.17	- 0.17	- 0.17	- 0.16	- 0.16	- 0.16	- 0.16	- 0.15
0.44	- 0.15	- 0.15	- 0.15	- 0.14	- 0.14	- 0.14	- 0.14	- 0.13	- 0.13	- 0.13

$p \rightarrow$ $\downarrow$	0	1	2	3	4	5	6	7	8	9
0.45	- 0.13	- 0.12	- 0.12	- 0.12	- 0.12	- 0.11	- 0.11	- 0.11	- 0.11	- 0.10
0.46	- 0.10	- 0.10	- 0.10	- 0.09	- 0.09	- 0.09	- 0.09	- 0.08	- 0.08	- 0.08
0.47	- 0.08	- 0.07	- 0.07	- 0.07	- 0.07	- 0.06	- 0.06	- 0.06	- 0.06	- 0.05
0.48	- 0.05	- 0.05	- 0.05	- 0.04	- 0.04	- 0.04	- 0.04	- 0.03	- 0.03	- 0.03
0.49	- 0.03	- 0.02	- 0.02	- 0.02	- 0.02	- 0.01	- 0.01	- 0.01	- 0.01	- 0.00
0.50	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
0.51	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.05	0.05
0.52	0.05	0.05	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.07
0.53	0.08	0.08	0.08	0.08	0.09	0.09	0.09	0.09	0.10	0.10
0.54	0.10	0.10	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.12
0.55	0.13	0.13	0.13	0.13	0.14	0.14	0.14	0.14	0.15	0.15
0.56	0.15	0.15	0.16	0.16	0.16	0.16	0.17	0.17	0.17	0.17
0.57	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.19	0.20	0.20
0.58	0.20	0.20	0.21	0.21	0.21	0.21	0.22	0.22	0.22	0.23
0.59	0.23	0.23	0.23	0.24	0.24	0.24	0.24	0.25	0.25	0.25
0.56	0.25	0.26	0.26	0.26	0.26	0.27	0.27	0.27	0.27	0.28
0.61	0.28	0.28	0.28	0.29	0.29	0.29	0.30	0.30	0.30	0.30
0.62	0.31	0.31	0.31	0.31	0.32	0.32	0.32	0.32	0.33	0.33
0.63	0.33	0.33	0.34	0.34	0.34	0.35	0.35	0.35	0.35	0.36
0.64	0.36	0.36	0.36	0.37	0.37	0.37	0.37	0.38	0.38	0.38
0.65	0.39	0.39	0.39	0.39	0.40	0.40	0.40	0.40	0.41	0.41
0.66	0.41	0.42	0.42	0.42	0.42	0.43	0.43	0.43	0.43	0.44
0.67	0.44	0.44	0.45	0.45	0.45	0.45	0.46	0.46	0.46	0.46
0.68	0.47	0.47	0.47	0.48	0.48	0.48	0.49	0.49	0.49	0.49
0.69	0.50	0.50	0.50	0.50	0.51	0.51	0.52	0.52	0.52	0.52
0.70	0.52	0.53	0.53	0.53	0.54	0.54	0.54	0.54	0.55	0.55
0.71	0.55	0.56	0.56	0.56	0.57	0.57	0.57	0.57	0.58	0.58
0.72	0.58	0.59	0.59	0.59	0.59	0.60	0.60	0.60	0.61	0.61
0.73	0.61	0.62	0.62	0.62	0.63	0.63	0.63	0.63	0.64	0.64
0.74	0.64	0.65	0.65	0.65	0.66	0.66	0.67	0.67	0.67	0.67
0.75	0.67	0.68	0.68	0.68	0.69	0.69	0.70	0.70	0.70	0.70
0.76	0.71	0.71	0.71	0.72	0.72	0.72	0.73	0.73	0.73	0.74
0.77	0.74	0.74	0.75	0.75	0.75	0.76	0.76	0.76	0.77	0.77
0.78	0.77	0.78	0.78	0.78	0.79	0.79	0.80	0.80	0.80	0.80
0.79	0.81	0.81	0.81	0.82	0.82	0.82	0.83	0.83	0.83	0.84
0.80	0.84	0.85	0.85	0.85	0.86	0.86	0.87	0.87	0.87	0.87
0.81	0.88	0.88	0.89	0.89	0.89	0.90	0.90	0.90	0.91	0.91
0.82	0.92	0.92	0.92	0.93	0.93	0.93	0.94	0.94	0.95	0.95
0.83	0.95	0.96	0.97	0.97	0.97	0.97	0.98	0.98	0.99	0.99
0.84	0.99	1.00	1.00	1.01	1.01	1.02	1.02	1.02	1.03	1.03
0.85	1.04	1.04	1.05	1.05	1.05	1.06	1.07	1.07	1.07	1.08
0.86	1.08	1.09	1.09	1.09	1.10	1.10	1.11	1.11	1.12	1.12
0.87	1.13	1.13	1.14	1.14	1.15	1.15	1.16	1.16	1.17	1.17
0.88	1.18	1.18	1.19	1.19	1.20	1.20	1.21	1.21	1.22	1.22
0.89	1.23	1.23	1.24	1.24	1.25	1.25	1.26	1.26	1.27	1.28
0.90	1.28	1.29	1.29	1.30	1.30	1.31	1.32	1.32	1.33	1.33
0.91	1.34	1.35	1.35	1.36	1.37	1.37	1.38	1.39	1.39	1.40
0.92	1.41	1.41	1.42	1.43	1.43	1.44	1.45	1.45	1.46	1.47
0.93	1.48	1.48	1.49	1.50	1.51	1.51	1.52	1.53	1.54	1.55
0.94	1.55	1.56	1.57	1.58	1.59	1.60	1.61	1.62	1.63	1.64

$p \rightarrow$ $\downarrow$	0	1	2	3	4	5	6	7	8	9
0.95	1.64	1.65	1.66	1.67	1.68	1.70	1.71	1.72	1.73	1.74
0.96	1.75	1.76	1.77	1.79	1.80	1.81	1.83	1.84	1.85	1.87
0.97	1.88	1.90	1.91	1.93	1.94	1.96	1.98	2.00	2.01	2.03
0.98	2.05	2.07	2.10	2.12	2.14	2.17	2.20	2.23	2.26	2.29
0.99	2.33	2.37	2.41	2.46	2.51	2.58	2.65	2.75	2.88	3.09

## Додаток 5

Критичні значення  $X$  (Критерію Ван дер Вардена)

$n$	$n_x - n_y = 0$ або 1		$n_x - n_y = 2$ або 3		$n_x - n_y = 4$ або 5	
	5%	1%	5%	1%	5%	1%
6	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
7	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
8	2.40	$\infty$	2.30	$\infty$	$\infty$	$\infty$
9	2.48	$\infty$	2.20	$\infty$	$\infty$	$\infty$
10	2.60	3.20	2.49	3.10	2.30	$\infty$
11	2.72	3.40	2.58	3.40	2.40	$\infty$
12	2.86	3.60	2.79	3.58	2.68	3.40
13	2.96	3.71	2.91	3.64	2.78	3.50
14	3.11	3.94	3.06	3.88	3.00	3.76
15	3.24	4.07	3.19	4.05	3.06	3.88
16	3.39	4.26	3.36	4.25	3.28	4.12
17	3.49	4.44	3.44	4.37	3.36	4.23
18	3.63	4.60	3.60	4.58	3.53	4.50
19	3.73	4.77	3.69	4.71	3.61	4.62
20	3.86	4.94	3.84	4.92	3.78	4.85
21	3.96	5.10	3.92	5.05	3.85	4.96
22	4.08	5.26	4.06	5.24	4.01	5.17
23	4.18	5.40	4.15	5.36	4.08	5.27
24	4.29	5.55	4.27	5.53	4.23	5.48
25	4.39	5.68	4.36	5.65	4.30	5.58
26	4.50	5.83	4.48	5.81	4.44	5.76
27	4.59	5.95	4.56	5.92	4.51	5.85
28	4.69	6.09	4.68	6.07	4.64	6.03
29	4.78	6.22	4.76	6.19	4.72	6.13
30	4.88	6.35	4.87	6.34	4.84	6.30
31	4.97	6.47	4.95	6.44	4.91	6.39
32	5.07	6.60	5.06	6.58	5.03	6.55
33	5.15	6.71	5.13	6.69	5.10	6.64
34	5.25	6.84	5.24	6.82	5.21	6.79
35	5.33	6.95	5.31	6.92	5.28	6.88
36	5.42	7.06	5.41	7.05	5.38	7.02
37	5.50	7.17	5.48	7.15	5.45	7.11
38	5.59	7.28	5.58	7.27	5.55	7.25
39	5.67	7.39	5.65	7.37	5.62	7.33
40	5.75	7.50	5.74	7.49	5.72	7.47
41	5.83	7.62	5.81	7.60	5.79	7.56
42	5.91	7.72	5.90	7.71	5.88	7.69
43	5.99	7.82	5.97	7.81	5.95	7.77
44	6.06	7.93	6.06	7.92	6.04	7.90
45	6.14	8.02	6.12	8.01	6.10	7.98
46	6.21	8.13	6.21	8.12	6.19	8.10
47	6.29	8.22	6.27	8.21	6.25	8.18
48	6.36	8.32	6.35	8.31	6.34	8.29
49	6.43	8.41	6.42	8.40	6.39	8.37
50	6.50	8.51	6.50	8.50	6.48	8.48

## СПИСОК ЛІТЕРАТУРИ.

1. Колемаев В.А., Староверов О.В., Турундаевский В.Б. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1991. – 400с.
2. Севастьянов Б.А. Курс теории вероятностей и математической статистики. – М.: Наука, 1982. – 256с.
3. Сигорский В.П. Математический аппарат инженера. – Киев: Техника, 1975. – 768с.
4. Уилкс С. Математическая статистика. – М.: Наука, 1967. – 632с.
5. Бровченко Т.А., Варбанец П.Д., Таранец В.Г. Метод статистического анализа в фонетических исследованиях. /учебное пособие. – Одесса, 1976. – 100с.
6. Левицкий В.В. Квантитативные методы в лингвистике. – Черновцы: Рута, 2004. – 190с.
7. Коршунов Ю.М. Математические основы кибернетики. – М.: Энергия, 1980. – 424с.
8. Носенко И.А. Начала статистики для лингвистов. – М.: Высшая школа, 1981.
9. Налимов В.В. Теория эксперимента. – М.: Наука, 1971.
10. Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. – М.: Высшая школа, 1997.
11. Перебийніс В.С., Муравицька М.П., Дарчук Н.П. Частотні словники та їх використання. – Київ: Наукова думка, 1985.
12. Статистичні параметри стилів. – Київ: Наукова думка, 1967.
13. Тищенко В. Частота частин мови в різних функціональних стилях сучасної української мови// Перебийніс В.С., Муравицька М.П. Питання структурної лексикології. – Київ, 1970. – С. 215-224.
14. Мусурівська О.В. Прикметники, об'єднані семантикою фігму сучасній англійській мові. - Одеса: АКД, 1993.
15. Ковтанюк В.Р. Зіставлення лексико-семантичних мікросистем із значенням «міцний» у французькій та англійській мовах. – Донецьк: АКД, 2001.