МІНІСТЕРСТВО НАУКИ І ОСВІТИ УКРАЇНИ

ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ім. І.І. МЕЧНИКОВА

Варбанець П.Д., Якімова Н.А.

ЛІНГВОСТАТИСТИКА

Курс лекцій

Одеса

2013

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. Якимова Н.А.

рецензенти: д. ф.-м. н. Леонов Ю.Г.

к. ф.-м. н. Покась С.М.

ЗМІСТ

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 вищестоящого 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 (фонеми, морфеми, лексеми), 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 затьмарення. До погано організованих систем належить інтелектуальна діяльність людини, а разом з нею і мовне поводження, 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 апарату, точніше, у сукупності математичних методів можна умовно розрізнити кількісні й некількісні методи [10].За допомогою некількісних 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 лінгвостатистикой 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 використаються 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*. Достовірна й неможлива події можуть розглядатися як крайні окремі випадки випадкових подій.

Випадкові події будемо позначати великими латинськими буквами $A, B, C, \ldots \ldots$ Достовірна подія позначається буквою Ω, неможлива - $\varnothing$. 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 наступає або подія A, або подія B, або вони обидві одночасно. *Добуток подій A і B* – це така третя подія $C=AB$, що наступає тоді, коли відбуваються й подія A, і подія B. Подія $\overline{A}$ *протилежна* події A, якщо вона несумісна з подією A и разом з нею утворює достовірну подію, тобто $\overline{A}+A=\Omega$.

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 ω_1, або ω_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 $\omega_1, \ldots, \omega_n$, *імовірність* будь-якої складеної події A, що складається з m елементарних подій $\omega_{i_1}, \ldots, \omega_{i_m}$, визначається як відношення кількості елементарних подій, що сприяють події A, до загальної кількості елементарних подій, тобто

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

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 $\boldsymbol{P}(A)=3/6=1/2$, 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 Ω, що включає всі n елементарних подій, дорівнює $\boldsymbol{P}(\Omega)=n/n=1$. Але тоді повна подія Ω, що складається в появі будь-якого із усього набору елементарних подій $\omega_1, \omega_2, \ldots, \omega_n$, 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, уведені вище, можна інтерпретувати як співвідношення між множинами. Несумісні події – це такі події, які не містять спільних елементів. Сума $A+B$ и добуток AB – це відповідно їхнє об'єднання $A\cup B$ і перетинання $A\cap B$. Протилежна подія $\overline{A}$ - 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.

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

Доказ. Дійсно, імовірності подій A і B дорівнюють $\frac{m}{n}$ й $\frac{k}{n}$ відповідно. Подія $C=A\cup B=\{\omega_{i_1}$, …, ω_{i_m}, ω_{j_1}, …, $\omega_{j_k}\}$ містить $m+k$ елементарних подій, тому що за умовою теореми серед елементарних подій $\{\omega_{i_1}, \ldots, \omega_{i_m}\}$ немає жодного, котре входило б у набір $\{\omega_{j_1}, \ldots, \omega_{j_k}\}$. Тому, відповідно до класичного визначення, його ймовірність дорівнює

$$\boldsymbol{P}(C)=\frac{m+k}{n}=\frac{m}{n}+\frac{k}{n}=\boldsymbol{P}(A)+\boldsymbol{P}(B).$$

Теорему доведено.

Подія $\overline{A}$ називається *протилежною* стосовно A, якщо до неї входять всі елементарні події, що не входять в A. Іншими словами, A і $\overline{A}$ - це такі несумісні події, які разом утворюють достовірну подію, тобто $A\cup\overline{A}=\Omega$. З теореми додавання випливає, що $\boldsymbol{P}(\Omega)=\boldsymbol{P}(A)+\boldsymbol{P}(\overline{A})=1$, тому $\boldsymbol{P}(\overline{A})=1-\boldsymbol{P}(A)$. Звідси, зокрема, випливає, що ймовірність неможливої події $\varnothing$, що є протилежною стосовно достовірної події Ω, дорівнює нулю.

1.2. Урнова схема

Класична схема, незважаючи на всю свою обмеженість, придатна для вирішення ряду сугубо 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 урновою 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 $N!=1\cdot 2\cdot 3\cdot\ldots\cdot N$. 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:

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

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

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

Говорити про ймовірності як про міри можливості здійснення випадкової події A має сенс тільки при здійсненні певного комплексу умов. При зміні умов зміниться й імовірність. Так, якщо до комплексу умов при якому вивчалася ймовірність $\boldsymbol{P}(A)$, додати нову умову, що полягає в появі події B, то одержимо інше значення ймовірності $\boldsymbol{P}(A/B)$ – умовну ймовірність події A за умови, що відбулася подія B. Імовірність $\boldsymbol{P}(A)$, 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 події A за умови, що подія B відбулася, сприяє r елементарних результатів. Тоді, згідно (1.1), умовна ймовірність події A дорівнює $\boldsymbol{P}(A/B)=\frac{r}{k}$. Розділивши й чисельник, і знаменник на n, одержимо формулу умовної ймовірності:

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

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

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

$$\boldsymbol{P}(A\cap B)=\boldsymbol{P}(A)\cdot\boldsymbol{P}(B), \tag{1.4}$$

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

$$\boldsymbol{P}(A\cap B)=\boldsymbol{P}(A)\cdot\boldsymbol{P}(A/B), \tag{1.5}$$

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

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

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

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 ω_{10}, ω_9, …, ω_1, ω_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, …, p_1, p_0. Ці події, загалом кажучи, неоднаково можливі. Наприклад, для майстрів спорту, очевидно, виключена подія ω_0, тому $p_0=0$, тобто відразу виключається однакова можливість.

Скінченна схема з неоднаково можливими результатами визначається в такий спосіб. Є скінченний набір елементарних подій $\Omega=\{\omega_1, \omega_2, \ldots, \omega_n\}$, і для кожної елементарної події ω_i задана його ймовірність p_i, $0\le p_i\le 1$, причому $\sum_{i=1}^{n} p_i=1$. Імовірність будь-якої складеної події $A=\{\omega_{i_1}, \ldots, \omega_{i_m}\}$ визначається як сума ймовірностей вхідних у нього елементарних подій:

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

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.

<u>Теорема 1.2</u>. <u>Формулювання</u>. Для двох несумісних подій A і B, що є підмножинами Ω, $\boldsymbol{P}(A \cap B)=\boldsymbol{P}(A)+\boldsymbol{P}(B)$.

<u>Доказ</u>. Нехай $A=\{\omega_{i_1}, \ldots, \omega_{i_m}\}$, $B=\{\omega_{j_1}, \ldots, \omega_{j_k}\}$. Згідно (1.6),

$$\boldsymbol{P}(A)=\sum_{l=1}^{m} p_{i_l}\ , \quad \boldsymbol{P}(B)=\sum_{q=1}^{k} p_{j_q}\ .$$

Оскільки A і B несумісні, вони не мають спільних елементарних подій і, отже, $C=A\cup B=\{\omega_{i_1}, \ldots, \omega_{i_m}, \omega_{j_1}, \ldots, \omega_{j_k}\}$. На підставі (1.6) маємо

$$\boldsymbol{P}(C)=\sum_{l=1}^{m} p_{i_l} + \sum_{q=1}^{k} p_{j_q} = \boldsymbol{P}(A)+\boldsymbol{P}(B).$$

Теорему доведено.

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 $\Omega=\{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$, загалом кажучи, містить зліченну кількість елементарних подій. Для кожної елементарної події задана її ймовірність $p_i=\boldsymbol{P}(\omega_i)$, $0\le p_i\le 1$, причому $\sum_{i=1}^{\infty} p_i =1$. Імовірність будь-якої скінченної або зліченної підмножини $A\subset\Omega$ множини елементарних подій Ω дорівнює сумі ймовірностей елементарних подій, що її складають, тобто якщо $A=\bigcup_{l=1}^{\infty}\omega_{i_l}$, то $\boldsymbol{P}(A)=\sum_{l=1}^{\infty} p_{i_l}$. 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 $A=\{\omega_{i_1}, \ldots, \omega_{i_m}\}$ і $B=\{\omega_{j_1}, \ldots, \omega_{j_l}, \ldots, \omega_{j_k}\}$ такі, що $\omega_{i_1}=\omega_{j_1}, \ldots, \omega_{i_l}=\omega_{j_l}$, $l\le m, k$. Інакше кажучи, $A\cap B=\{\omega_{i_1}, \ldots, \omega_{i_l}\}$. Тоді

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

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

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:

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

Таким чином, ми знову одержуємо ту ж формулу умовної ймовірності, що й у класичній схемі. Незалежність подій визначається аналогічно класичній схемі.

1.4. Вирахування подій

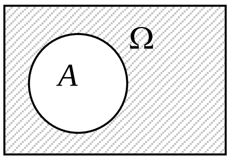
Одним з основних понять теорії ймовірностей є простір елементарних подій Ω і події як деякі підмножини цього простору. У загальному випадку простір Ω може бути будь-якої природи: як скінченним, так і нескінченним, як дискретним, так і безперервним. Розглянемо, наприклад, стрілянину по мішені. Якщо нас цікавить тільки сам факт влучення в мішень, то елементарними результатами служать $\omega_1=1$ (влучення в мішень) і $\omega_0=0$ (невлучення в мішень). Якщо важливо влучення в окремі області мішені (області розрізняються з погляду уразливості реальної мети), то елементарними подіями можуть бути $\omega_{10}=10$, $\omega_9=9$, …, $\omega_1=1$ (відповідають кількості вічків, приписаних влученню в певну область) і $\omega_0=0$ (невлучення в мішень). Нарешті, якщо істотно важливим є знання, у яку саме крапку щита, на якому зображена мішень, відбулося влучення, то довільний елементарний результат $\omega=\{x, y\}$ являє собою координати крапки влучення, а простір елементарних подій – це множина крапок щита.

Таким чином, нехай є простір елементарних подій Ω будь-якої природи. Будемо розглядати як події підмножини A, B, C цього простору. У такому випадку дії над подіями стають діями над підмножинами.

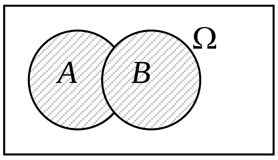
Подія A відбувається тоді й тільки тоді, коли відбувається одне з елементарних подій ω, з яких складається A. Несумісність подій A і B означає, що вони не мають жодної спільної елементарної події, тобто не перетинаються. Подія $A+B$ еквівалентна об'єднанню $A\cup B$, тобто такій множині елементарних подій, які входять або в A, або в B, а подія AB – перетинанню $A\cap B$, тобто множині елементарних подій, які є спільними для A і B. Подія, що включає всі елементарні події, тобто збігається із простором елементарних подій Ω, є достовірною. Звідси можна зробити висновок, що для будь-якої події $A\cap\Omega=A$. Порожньою є подія, що не містить жодного елемента. Звідси випливає, що дві події A і B несумісні, якщо $A\cap B=\varnothing$. Очевидно, що $A\cup\varnothing=A$. *Теоретико-множинною різницею* двох подій A і B називається така подія $A\backslash B$, що містить ті елементарні події, що належать A, які не входять у B. Звідси випливає, що подія $\overline{A}$, протилежна події A, є $\Omega\backslash A$. Справді, $A\cap(\Omega\backslash A)=\varnothing$ і $A\cup(\Omega\backslash A)=\Omega$, тобто дійсно $\overline{A}=\Omega\backslash A$. Зокрема, $\overline{\Omega}=\varnothing$, $\overline{\varnothing}=\Omega$. Іноді використається симетрична різниця подій $C=A\oplus B$, що представляє собою таку подію, у яку входять ті елементарні події, які входять в A або в B, але не входять у їхнє перетинання $A\cap B$. У такий спосіб, ця операція може бути представлена за допомогою вже уведених операцій у такий спосіб:

$$C=A\oplus B=(A\backslash B)\cup(B\backslash A)=(A\cap\overline{B})\cup(\overline{A}\cap B).$$

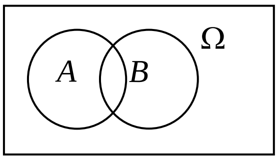
Для кращого розуміння операцій над подіями – підмножинами – звичайно використають умовні графічні зображення, представляючи достовірну подію Ω як прямокутник, а інші події – як кола. Тоді уведені вище операції над подіями можуть бути представлені у вигляді діаграм Венна [1, стор. 19; 3. стор. 639], де результати операцій зображені у вигляді заштрихованих фігур. Операція доповнення зображена на мал. 1.1, об'єднання – на мал. 1.2, перетинання – на мал. 1.3, різниці – на мал. 1.4, симетричної різниці – на мал. 1.5.



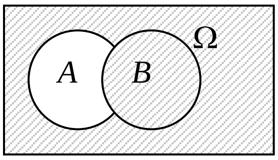
Мал. 1.1



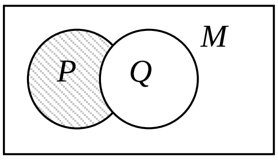
Мал. 1.2



Мал. 1.3



Мал. 1.4



Мал. 1.5

Дії над подіями, зокрема операції об'єднання (додавання) і перетинання (множення), аналогічні додаванню й множенню чисел. Ці операції

комутативні: $A\cup B=B\cup A$, $A\cap B=B\cap A$;

асоціативні: $(A\cup B)\cup C=A\cup(B\cup C)$, $(A\cap B)\cap C=A\cap(B\cap C)$;

дистрибутивні: $(A\cup B)\cap C=(A\cap C)(B\cap C)$.

Інша подібність із діями над числами полягає в тому, що для операції перетинання роль нуля й одиниці при множенні чисел виконують множини Ω і $\varnothing$, тому що $\Omega\cap A=A\cap\Omega=A$ і $\varnothing\cap A=A\cap\varnothing=\varnothing$. Разом з тим теоретико-множинні рівності $A\cup A=A$ і $A\cap A=A$ і їм подібні показують, що повної аналогії немає.

Дії над подіями важливі не самі по собі, а як засіб визначення ймовірності одних подій через імовірності інших подій. У випадку скінченної або зліченної теоретико-імовірнісної схеми, розглянутої вище, як подія розглядалася будь-яка підмножина скінченного або зліченного простору елементарних подій Ω і ймовірність події визначалася як сума ймовірностей вхідних у нього елементарних подій. Якщо ж простір Ω неперервний, то має місце континуум елементарних результатів. Спроба вважати подією будь-яку підмножину безперервного простору Ω сполучена з великими труднощами. Тому в загальному випадку доводиться мати справу не з усіма підмножинами простору Ω, а лише з певним класом, замкнутим щодо операцій об'єднання й перетинання. Клас підмножин простору Ω, замкнутий щодо операцій доповнення, об'єднання й перетинання, в якому в той же час містяться множини $\varnothing$ і Ω, називається *полем*. Будемо позначати поле буквою S. Мінімальне поле S_0 складається з повної й порожньої множини $S_0=\{\varnothing, \Omega\}$. Дійсно, $\varnothing$ і Ω входять у цей клас, а результатами операцій об'єднання, доповнення й перетинання над цими множинами знову служать дані множини: $\Omega\cup\varnothing=\Omega$, $\Omega\cap\varnothing=\varnothing$, $\overline{\Omega}=\varnothing$, $\overline{\varnothing}=\Omega$. Іншим, більш змістовним прикладом поля подій служить клас із чотирьох подій $S=\{\varnothing, A, \overline{A}, \Omega\}$. Дійсно,

$$\varnothing\cup A=A,\ \varnothing\cup\overline{A}=\overline{A},\ \Omega\cup\varnothing=\Omega,\ \varnothing\cap A=\varnothing,\ \varnothing\cap\overline{A}=\varnothing,$$

$$\Omega\cap\varnothing=\varnothing,\ A\cup\overline{A}=\Omega,\ A\cup\Omega=\Omega,\ A\cap\overline{A}=\varnothing,\ A\cap\Omega=A,$$

$$\overline{A}\cup\Omega=\Omega,\ \overline{A}\cap\Omega=\overline{A},\ \overline{\varnothing}=\Omega,\ \overline{\overline{A}}=A,\ \overline{\Omega}=\varnothing.$$

1.5. Аксіоматична побудова теорії ймовірностей.

При аксіоматичній побудові теорії ймовірностей вихідним «матеріалом» служать простір елементарних подій Ω і виділений у ньому клас підмножин, що утворює поле подій S. Будова простору Ω і класу S визначається конкретною областю додатка.

Імовірністю називається числова функція, визначена на полі подій S, що володіє наступними властивостями.

Аксіома 1. Для будь-якої події $A\in S$ $\boldsymbol{P}(A)\geq 0$.

Аксіома 2. Імовірність достовірної події дорівнює одиниці $\boldsymbol{P}(\Omega)=1$.

Аксіома 3. Імовірність об'єднання двох несумісних подій дорівнює сумі ймовірностей цих подій, тобто для будь-яких двох несумісних подій $A\in S$, $B\in S$, $A\cap B=\varnothing$ $\boldsymbol{P}(A\cup B)=\boldsymbol{P}(A)+\boldsymbol{P}(B)$.

Скінченна схема задається скінченною множиною елементарних подій $\Omega=\{\omega_1, \omega_2, \ldots, \omega_n\}$ і ймовірностями кожної з них $0\le p_i\le 1$, причому $\sum_{i=1}^{n} p_i=1$. Імовірність будь-якої події A, що є підмножиною Ω, тобто $A=\{\omega_{i_1}, \ldots, \omega_{i_m}\}$, визначається як сума ймовірностей вхідних у неї елементарних подій, тобто по формулі (1.6)

Клас S всіх підмножин Ω утворює поле. Тепер переконаємося, що скінченна схема задовольняє вимозі аксіоми 1. Для цього виберемо довільну подію A, що є підмножиною Ω. Так як $A=\{\omega_{i_1}, \ldots, \omega_{i_m}\}$, то, відповідно до скінченної схеми, $\boldsymbol{P}(A)=\sum_{l=1}^{m} p_{i_l}$, $0\le p_{i_l}\le 1$, $l=\overline{1,m}$, тому $\boldsymbol{P}(A)\ge 0$, тобто умова аксіоми 1 виконується. Умова аксіоми 2 виконується, тому що $\Omega=\{\omega_1, \omega_2, \ldots, \omega_n\}$ і на підставі (1.6) $\boldsymbol{P}(\Omega)=\sum_{i=1}^{n} p_i=1$. Умова аксіоми 3 також виконується, тому що вона являє собою зміст теореми додавання для скінченної схеми. Таким чином, скінченна схема являє приклад об'єкта, для якого виконується система аксіом теорії ймовірностей.

Як приклад імовірнісної схеми з безперервним простором елементарних подій розглянемо схему з геометричними ймовірностями. Нехай простором елементарних подій служить множина крапок деякої області G, що має площу на площині. Як події будемо розглядати підмножини A, B, C, … цієї області, що мають площу. Клас цих підмножин також утворює поле. При цьому ймовірність будь-якої події A (підмножини, що має площу $mes(A)$), можна задати в такий спосіб:

$$\boldsymbol{P}(A)=\frac{mes(A)}{mes(G)}.$$

Описана схема також задовольняє аксіомам теорії ймовірностей. Аналогічно можна побудувати геометричні ймовірності в будь-якому скінченновимірному просторі.

У багатьох випадках виконання аксіоми 3 потрібно в розширеному варіанті. Аксіома 3 постулює додавання ймовірностей для скінченної кількості несумісних подій, тоді як у розширеному варіанті мова йде про зліченну кількість несумісних подій.

<u>Аксіома 3’</u>. Якщо $A_i\in S$, $i=1, 2, \ldots$, $A_i\cap A_j=\varnothing$, $i\ne j$, то $\boldsymbol{P}(\bigcup_{i=1}^{\infty} A_i)=\sum_{i=1}^{\infty} P(A_i)$.

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:

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

Аксіоми теорії ймовірностей лише постулюють існування ймовірностей для всіх подій, що утворюють поле S, і задають певні правила дії з ймовірностями. Експериментальне ж визначення ймовірності будь-якої події $A\in S$ може бути здійснене в результаті випробувань, виконуваних при певному тому ж самому комплексі умов. Із усього вищесказаного випливає, що певна теоретико-імовірнісна схема задається трійкою $\{\Omega, S, \boldsymbol{P}\}$, тобто конкретним простором елементарних подій, конкретним набором підмножин Ω, що утворюють поле, а також конкретним завданням ймовірностей на множинах поля. Набір цих трьох компонентів називається *імовірнісним простором* [1, 4]. Імовірність $\boldsymbol{P}$ на $\{\Omega, S\}$ називається *розподілом ймовірностей на* Ω.

2. УМОВНІ ЙМОВІРНОСТІ

2.1. Повна ймовірність

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

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

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

<u>Доказ</u>. Події повної групи A_1, A_2, …, A_n попарно несумісні, тому попарно несумісні і їхні добутки (перетинання) з подією B, тобто події $B \cap A_i$, $B \cap A_j$ при $i \neq j$ несумісні. Так як подія B, з урахуванням (2.1), можна представити у вигляді $B=\bigcup_{i=1}^{n}(B \cap A_i)$, то, застосувавши до цього розкладання події B аксіому додавання ймовірностей, маємо

$$\boldsymbol{P}(B)=\boldsymbol{P}\left(\bigcup_{i=1}^{n}(B \cap A_i)\right)=\sum_{i=1}^{n} P(B \cap A_i).$$

Використовуючи формулу множення ймовірностей (1.5) для кожного доданка, остаточно одержуємо

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

Теорему доведено.

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, $\boldsymbol{P}(A_i)>0$, $B \in \bigcup_{i=1}^{n} A_i$:

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

2.2. Формула Байєса

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

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

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

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

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

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

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

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

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

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: $\{\omega_0, \omega_1\}$, ω_0 – невдача, ω_1 – успіх, при цьому $A=\{\omega_1\}$, $\overline{A}=\{\omega_0\}$. Множина елементарних результатів для n випробувань складається вже з 2^n елементарних подій $\omega_{i_1,i_2,\ldots,i_n}=\{i_1, \ldots, i_n\}$, кожне з яких відповідає конкретному результату випробувань, при цьому набір індексів $i_1, \ldots, i_n$ 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 $\omega_{i_1,i_2,\ldots,i_n}$, при цьому $(i_1, \ldots, i_n)$ – конкретна послідовність нулів і одиниць, що відповідає послідовності невдач або успіхів у кожному з n індивідуальних випробувань, наприклад (A_{i_1}, A_{i_2}, …, A_{i_n}). Тоді з незалежності друг від друга результатів окремих випробувань одержуємо

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

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

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

І, отже, по аксіомі додавання ймовірностей може бути визначена ймовірність будь-якої події, що складається з декількох елементарних подій. Зокрема, якщо нас цікавить імовірність $\boldsymbol{P}_n(m)$ того, що в n випробуваннях відбулося m успіхів, те її визначаємо як суму ймовірностей елементарних подій, що характеризуються m успіхами. Імовірність такого елементарного результату, згідно (2.3), дорівнює $p^m q^{n-m}$. Отже, для знаходження ймовірності $\boldsymbol{P}_n(m)$ 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. Остаточно одержуємо

$$\boldsymbol{P}_n(m)=C_n^m\ p^m q^{n-m}. \tag{2.4}$$

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

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

3. ВИПАДКОВА ВЕЛИЧИНА

3.1. Визначення випадкової величини й її функція розподілу

Явища, які відбуваються в природі, є результатом дії внутрішніх і (або) зовнішніх факторів. Якщо зміна явища відбувається під впливом одного фактору, то цей вплив можна до деякої міри визначити, і результат зміни розглядається як детермінований цим фактором. Але часто кінцевий результат явища, що змінюється, викликаний дією багатьох причин, визначення яких сполучено з великими труднощами й не завжди є доцільним. Дія комплексу причин щораз викликає певні зміни в кількісних характеристиках розглянутого явища. Причому сутність явища не змінюється. Такі коливання в значеннях кількісної характеристики є випадковими. Вони викликані дією множини причин, кожна з яких може бути несуттєвою для природи даного явища в цілому або для окремих його якостей. При багаторазовому повторенні тих самих умов експерименту (при дії того самого комплексу причин) у змінах значень кількісної характеристики ознаки (явища) можна виявити деяку закономірність. Ця закономірність, що є присутньою у випадкових числах, буде відбивати ту якісну сторону (якісна ознака) явища, що найбільш яскраво проявляється у всіх експериментах (випробуваннях).

Кількісна характеристика розглянутої ознаки (явища), значення якої залежить від випадку, називається випадковою величиною [5]. Отже, *випадковою величиною* називається функція $X=X(\omega)$, визначена на множині елементарних подій Ω, $\omega\in\Omega$. У випадку скінченної або рахункової Ω випадковою величиною є будь-яка функція, визначена на Ω. У загальному випадку функція $X(\omega)$ повинна бути такою, щоб для будь-яких x подія $A=\{\omega: X(\omega)<x\}$, що полягає в тому, що випадкова величина X попадає в інтервал $(-\infty, x)$, належало полю подій S і, таким чином, для будь-якої такої події була визначена ймовірність $\boldsymbol{P}(A)=\boldsymbol{P}\{X<x\}$.

Випадкові величини позначаються прописними латинськими буквами X, Y, Z, …, а значення випадкової величини на елементарній події – маленькими латинськими буквами x, y, z, … ...Якщо множина елементарних подій скінченна, то випадкову величину можна задати, перелічивши її значення на всіх елементарних подіях. Однак у багатьох задачах немає необхідності розглядати випадкові величини як функції від елементарної події, а досить лише знати ймовірності будь-яких подій, пов'язаних з випадковою величиною, тобто закон розподілу випадкової величини. Говорять, що *закон розподілу випадкової величини* X заданий, якщо для будь-якої множини дійсних чисел B, що є об'єднанням або перетинанням скінченного або рахункового числа проміжків, задана ймовірність $\boldsymbol{P}\{\omega: X(\omega)\in B\}$ події, що полягає в тому, що $X(\omega)$ прийме значення із цієї множини [1].

Випадкові величини діляться на безперервні й дискретні. Значення, прийняті безперервною випадковою величиною, повністю заповнюють один або кілька числових інтервалів. Значення дискретної випадкової величини розташовані дискретно, тобто суцільно не заповнюють ніякий числовий інтервал. Наприклад, нехай досліджується величина частотного інтервалу мелодійного завершення фрази (частотний інтервал виражається відношенням максимуму частоти основного тону до мінімуму частоти основного тону розглянутого мовного відрізка). Очевидно, що ця величина є випадковою і її значеннями можуть бути будь-які числа, більші одиниці. Таким чином, величина частотного інтервалу є безперервною випадковою величиною. Якщо ж вивчається кількість слів у різних фразах даної мови, то ця величина також буде випадковою, але її значеннями можуть бути тільки цілі позитивні числа, які, природно, не заповнюють ніякого числового інтервалу. Тому ця випадкова величина буде дискретною. Таким чином, випадкова величина дискретна, якщо існує скінченна або рахункова множина чисел x_1, x_2, x_3, … таких, що $\boldsymbol{P}\{X=x_n\}=p_n\geq 0$, $n=1, 2, 3, \ldots$, $p_1+p_2+p_3+\ldots=1$... Закон розподілу дискретної випадкової величини X визначений, якщо відомі всі x_n і ймовірності $p_n=\boldsymbol{P}\{X=x_n\}$ такі, що $p_1+p_2+p_3+\ldots=1$... Якщо скласти таблицю, у верхньому рядку якої помістити значення дискретної випадкової величини, а в нижньої – відповідні ймовірності, то одержимо *ряд розподілу випадкової величини*. Сума ймовірностей, записаних у другому рядку таблиці, повинна дорівнювати одиниці. Ряд розподілу задає закон розподілу дискретної випадкової величини.

Закон розподілу випадкової величини X заданий, якщо можна визначити ймовірність події $\{\omega: X(\omega)\in A\}$ для будь-якої множини A, що є об'єднанням або перетинанням довільних проміжків.

У свою чергу, будь-які такі множини A можна одержати за допомогою операцій об'єднання, перетинання й доповнення із множини всіляких інтервалів виду $(-\infty, x)$. Отже, і події $\{\omega: X(\omega)\in A\}$ можна одержати за допомогою відповідних операцій над подіями із усіляких подій виду $\{\omega: X(\omega)<x\}$. По визначенню випадкової величини для будь-яких x подія $\{\omega: X(\omega)<x\}$ належить полю подій S. Отже, і події виду $\{\omega: X(\omega)\in A\}$ також належать S і для них визначена ймовірність. Таким чином, щоб задати закон розподілу довільної випадкової величини X, досить знати для будь-яких x імовірності подій $\{\omega: X(\omega)<x\}$.

Функція розподілу $F_X(x)$ випадкової величини X визначається формулою

$$F_X(x)=\boldsymbol{P}\{\omega: X(\omega)<x\}.$$

Цю рівність звичайно записують коротше у вигляді $F_X(x)=\boldsymbol{P}\{X<x\}$. Для простоти в тих випадках, коли це не може привести до неточності, пишуть $F(x)$ замість $F_X(x)$.

Розглянемо *властивості функції розподілу*.

1. Функція розподілу приймає значення із проміжку $[0, 1]$: $0\le F(x)\le 1$. Доказ. Дана властивість випливає з того, що функція розподілу – це ймовірність події $\{X<x\}$, а значення ймовірності будь-якої події ненегативно й не перевищує одиниці.
2. Імовірність того, що випадкова величина прийме значення з напівінтервалу $[x_1, x_2)$, дорівнює різниці $F(x_2) - F(x_1)$:

   $$\boldsymbol{P}\{x_1\le X<x_2\}=F(x_2) - F(x_1).$$

   Доказ. Представимо подію, що полягає в тому, що випадкова величина прийме значення, менше x_2, у вигляді суми неспільних подій $\{\omega: X(\omega)<x_2\}=\{\omega: X(\omega)<x_1\}\cup\{\omega: x_1\le X(\omega)<x_2\}$. За аксіомою додавання, імовірність суми неспільних подій дорівнює сумі ймовірностей цих подій:

   $$F(x_2)=\boldsymbol{P}\{\omega: X(\omega)<x_2\}=\boldsymbol{P}\{\omega: X(\omega)<x_1\}+\boldsymbol{P}\{\omega: x_1\le X(\omega)<x_2\}=F(x_1)+\boldsymbol{P}\{\omega: x_1\le X(\omega)<x_2\},$$

   звідки й треба доказувана властивість.
3. Функція розподілу – неубутна функція, тобто $F(x_2)\ge F(x_1)$, якщо $x_2>x_1$. Доказ. Нехай $x_2>x_1$. Повторюючи наведені вище викладення й з огляду на те, що ймовірність будь-якої події ненегативна, одержуємо

   $$F(x_2)=F(x_1)+\boldsymbol{P}\{\omega: x_1\le X(\omega)<x_2\}\ge F(x_1).$$

4. $\boldsymbol{P}\{X\ge x\}=1 - F(x)$. Доказ. Подія $\{\omega: X(\omega)\ge x\}$ є протилежною події $\{\omega: X(\omega)<x\}$ і, отже, $\boldsymbol{P}\{X\ge x\}=1 - \boldsymbol{P}\{X<x\}=1 - F(x)$.
5. Якщо $x\to\infty$, то $F(x)\to 1$.
6. Якщо $x\to -\infty$, то $F(x)\to 0$.
7. Функція розподілу безперервна ліворуч, тобто $\lim\limits_{\Delta\to+0} F(x\quad\Delta)=F(x)$.

Функцією розподілу може бути будь-яка функція, що задовольняє властивостям 1, 3, 5, 6 і 7, тобто безперервна ліворуч у кожній точці x монотонна неубутна функція, множиною значень якої є проміжок (0, 1), що включає або не включає граничні точки 0 і 1.

3.2. Розподіл дискретних випадкових величин

Дискретні випадкові величини приймають скінченну або рахункову множину значень. Нехай X – дискретна випадкова величина, що приймає значення $x_1<x_2<x_3<\ldots$с ймовірностями p_1, p_2, p_3, …, $p_1+p_2+p_3+\ldots=1$. Якщо по осі абсцис відкласти x_1, x_2, x_3, …, а по осі ординат - відповідні ймовірності p_1, p_2, p_3, … і з'єднати сусідні точки відрізками, то одержимо *багатокутник розподілу випадкової величини X.* Багатокутник розподілу – це графічне зображення ряду розподілу дискретної випадкової величини. Розглянемо функцію розподілу $F(x)$ дискретної випадкової величини X. Якщо $x\le x_1$, то $F(x)=\boldsymbol{P}\{X<x\}=0$, тому що в цьому випадку подія $\{X<x\}$ є неможливою. Якщо $x_1<x\le x_2$, то подія $\{X<x\}$ наступить тоді й тільки тоді, коли наступить подія $\{X=x_1\}$, тому $F(x)=\boldsymbol{P}\{X<x\}=\boldsymbol{P}\{X=x_1\}=p_1$. Якщо $x_2<X\le x_3$, то подія $\{X<x\}$ дорівнює сумі подій $\{X=x_1\}$ і $\{X=x_2\}$ і

$$F(x)=\boldsymbol{P}\{X<x\}=\boldsymbol{P}\{X=x_1\}+\boldsymbol{P}\{X=x_2\}=p_1+p_2.$$

Аналогічно, якщо $x_i<X<x_i+1$, то $F(x)=p_1+p_2+\ldots+p_i$. Таким чином, функція розподілу дискретної випадкової величини дорівнює $F(x)=\sum_{x_i<x} p_i$, де $p_i=\boldsymbol{P}\{X=x_i\}$ і підсумовування виробляється по тим i, для яких $x_i<x$.

Функція розподілу дискретної випадкової величини постійна на проміжках $(-\infty, x_1]$, $(x_1, x_2]$, $(x_i, x_{i+1}]$, … . У точках x_1, x_2, x_3, … функція розподілу має стрибки, що дорівнюють ймовірності того, що випадкова величина прийме відповідне значення.

Приклад 3.1. Знайти функцію розподілу випадкової величини, що дорівнює кількості випадань «герба» при киданні двох монет.

Рішення. Залежно від результату досвіду, ця кількість може виявитися рівною 0, 1 і 2. Виберемо в якості елементарних наступні події: ω_1=ГГ={на першій монеті випав «герб», на другій монеті – теж «герб»}, ω_2=РГ={на першій монеті випала «решка», на другій монеті – «герб»}, ω_3=ГР={на першій монеті випав «герб», на другій монеті – «решка»}, ω_4=РР={на першій монеті випала «решка», на другій монеті – теж «решка»}. Кількість X «гербів, що випали», є функцією від елементарної *події* $X(\omega1)$=X(ГГ)=2, $X(\omega_2)$=X(РГ)=1, $X(\omega_3)$=X(ГР)=1, $X(\omega_4)$=X(РР)=0. Результат експерименту є випадковим, тому значення X також є випадковим. Функція $X=X(\omega_i)$ – випадкова величина, визначена на множині $\Omega=\{\omega_1, \omega_2, \omega_3, \omega_4\}$. Якщо монети симетричні, то елементарні події ω_1, ω_2, ω_3, ω_4 однаково можливі й тому природно в математичній моделі експерименту вважати ймовірності $\boldsymbol{P}(\omega_i)$ рівними ¼ , i=1, 2, 3, 4. Знаючи ймовірності $\boldsymbol{P}(\omega_i)$, можна знайти ймовірність того, що випадкова величина X прийме те або інше значення. У наведеній таблиці поміщені елементарні події ω_i, імовірності $\boldsymbol{P}(\omega_i)$ і значення випадкової величини X, рівної кількості «гербів, що випали».

Таблиця 3.1

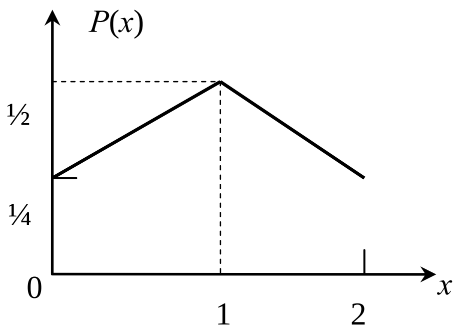
i	1	2	3	4
ω_i	ГГ	РГ	ГР	РР
$\boldsymbol{P}(\omega_i)$	¼	¼	¼	¼
$X(\omega_i)$	21	1	1	0

Слід зазначити, що випадкова величина X може приймати на різних елементарних подіях ті самі значення. Так, у розглянутому прикладі $X(\omega_2)=X(\omega_3)=1$. Подія, що полягає в тому, що випадкова величина X прийме значення, рівне нулю, дорівнює {РР}. Імовірність цієї події дорівнює ¼. Подія $\{\omega_i: X(\omega_i)=1\}$ дорівнює події {РГ, ГР}. Її ймовірність дорівнює $\boldsymbol{P}\{X=1\}=\boldsymbol{P}$\{РГ, ГР\}=½. Нарешті, $\boldsymbol{P}\{X=2\}=\boldsymbol{P}$\{ГГ\}=¼. При побудові математичної моделі в цьому випадку можна було б об'єднати події РГ і ГР в одну й у якості елементарних розглядати події {0}, {1} і {2}, що складаються в тім, що кількість «гербів, що випали», дорівнює нулю, одиниці або двійці відповідно: {0}={РР}, {1}={ГР, РГ}, {2}={ГГ}. Щоб задати імовірнісний простір і випадкову величину на «звуженому» просторі елементарних подій, у цьому випадку досить задати лише можливі значення випадкової величини й відповідні ймовірності, з якими ці значення приймаються. Вони наведені в таблиці 3.2, де x_i – значення випадкової величини, $p_i=\boldsymbol{P}\{X=x_i\}$, i=1, 2, 3.

Таблиця 3.2

x_i	0	1	2
p_i	¼	½	¼

На мал. 3.1 наведений багатокутник розподілу розглянутої випадкової величини.

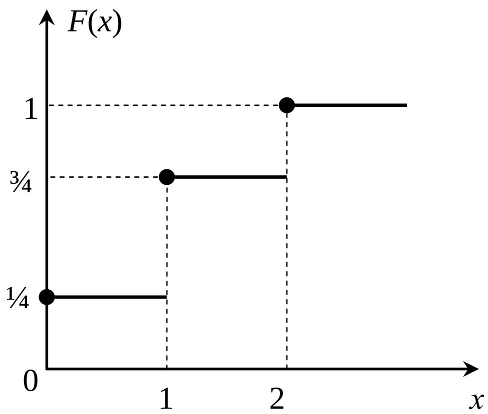


Мал.3.1

Знайдемо тепер функцію розподілу цієї випадкової величини. Функція розподілу при $x \le 0$ дорівнює нулю. У точці $x=0$ вона має стрибок, рівний ¼, у точці $x=1$ – стрибок, рівний ½, у точці $x=2$ – стрибок, рівний ¼. Між цими точками функція розподілу постійна. При $x>2$ функція $F(x)=1$. Таким чином,

$$F(x)=\begin{cases} 0, & \text{якщо } x \le 0, \\ 1/4, & \text{якщо } 0 < x \le 1, \\ 3/4, & \text{якщо } 1 < x \le 2, \\ 1, & \text{якщо } x > 2. \end{cases}$$

Графік функції $F(x)$ зображений на мал. 3.2.



Мал. 3.2.

Приведемо тепер закони розподілу дискретних випадкових величин.

3.2.1. *Рівномірний розподіл на множині* {1, 2, ..., *n*}

Випадкова величина X, що приймає цілочисельні значення від 1 до n, має рівномірний розподіл, якщо $\boldsymbol{P}\{X=m\}=1/n$, $m=1, 2, \ldots, n$.. Багатокутник розподілу випадкової величини x являє собою відрізок прямої, паралельний осі абсцис. Кінці відрізка мають координати $(1, 1/n)$ і $(n, 1/n)$.

3.2.2. *Гіпергеометричний розподіл*

Випадкова величина X має гіпергеометричний розподіл, якщо

$$\boldsymbol{P}\{X=m\}=\frac{C_M^m C_{N-M}^{n-m}}{C_N^n},\ m=0, 1, 2, \ldots, \min(n, M).$$

Гіпергеометричний розподіл має місце, наприклад, у наступній задачі. Нехай партія з N виробів M виробів 1-го сорту, а інші $N - M$ виробів – 2-го сорту. Із цієї партії витягають для контролю n виробів. Потрібно знайти закон розподілу випадкової величини X, рівної кількості виробів 1-го сорту серед обраних n виробів. Відповідно до (1.2) X має гіпергеометричний розподіл.

3.2.3. *Геометричний розподіл*

Випадкова величина X має геометричний розподіл, якщо $\boldsymbol{P}_m=\boldsymbol{P}\{X=m\}=q^m p$, $m=0, 1, 2, \ldots$, $0<p<1$, $q=1-p$.

Підсумувавши нескінченно убутну геометричну прогресію, легко переконатися в тому, що $\sum_{m=0}^{\infty} P_m=1$: $\sum_{m=0}^{\infty} P_m=\sum_{m=0}^{\infty} pq^m=p\dfrac{1}{1-q}=1$.

Геометричний розподіл має випадкова величина X, що дорівнює кількості випробувань Бернуллі до першого успіху з імовірністю успіху в одиничному випробуванні p.

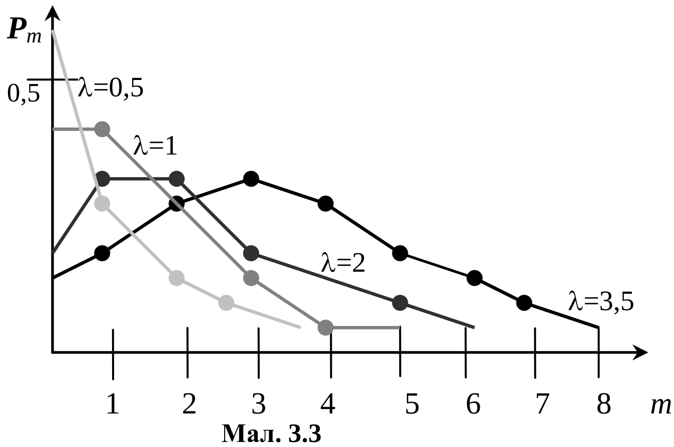
3.2.4. *Розподіл Пуассона*

Випадкова величина X має розподіл Пуассона з параметром λ, якщо

$$\boldsymbol{p}_m=\boldsymbol{P}\{X=m\}=\frac{\lambda^m}{m!}e^{-\lambda},\ m=0, 1, 2, \ldots, \lambda>0.$$

Скориставшись розкладанням e^{λ} у ряд Маклорена $e^{\lambda}=\sum_{m=0}^{\infty}\dfrac{\lambda^m}{m!}$, одержуємо

$$\sum_{m=0}^{\infty} p_m=\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}e^{-\lambda}=e^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}=e^{-\lambda}e^{\lambda}=1.$$



Мал. 3.3

Багатокутники розподілу випадкових величин, що мають розподіл Пуассона з різними значеннями параметра λ, зображені на мал. 3.3.

По цьому типу розподіляються, наприклад, частоти слів з великою кількістю значень (у словнику) [6].

3.2.5. *Біноміальний розподіл*

Випадкова величина X, значення якої дорівнюють цілим числам від 0 до n, має біноміальний розподіл, якщо $\boldsymbol{P}\{X=m\}$ задається формулою Бернуллі (2.4): $\boldsymbol{P}\{X=m\}=\boldsymbol{P}_n(m)=C_n^m p^m q^{n-m}$, $m=0, 1, 2, \ldots, n$, $0<p<1$, $q=1-p$. Ряд розподілу випадкової величини X має такий вигляд:

m	0	1	2	…	k	…	n
$\boldsymbol{P}_n(m)$	q^n	$C_n^1 pq^{n-1}$,	$C_n^2 p^2 q^{n-2}$		$C_n^k p^k q^{n-k}$		p^n

По формулі Бернуллі можна знайти ймовірність m успіхів в n випробуваннях Бернуллі з імовірністю успіху p і невдачі $1-p=q$. Таким чином, випадкова величина, що дорівнює кількості успіхів в n випробуваннях Бернуллі, має біноміальний закон розподілу.

При великих значеннях n обчислення ймовірностей $\boldsymbol{P}_n(m)$ по формулі Бернуллі стає скрутним. Однак у ряді випадків вдається замінити формулу Бернуллі підходящою наближеною асимптотичною формулою. Зокрема, якщо n завелике, а ймовірність успіху p мала, то

$$\boldsymbol{P}_n(m)\approx\frac{(np)^m}{m!}e^{-np}.$$

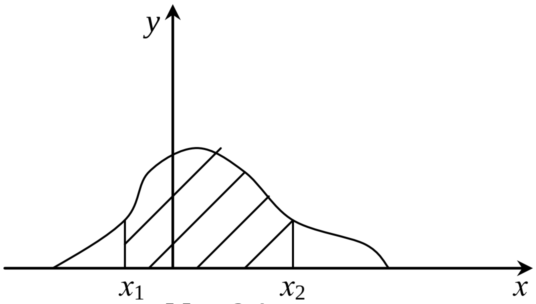
Якщо при великих n імовірність p успіху у випробуванні Бернуллі близька до одиниці, то ймовірність невдачі q близька до нуля. У такому випадку пуассоновське наближення можна використати для обчислення ймовірності кількості невдач.

3.3. Розподіл безперервних випадкових величин

Множина значень безперервної випадкової величини незліченно й звичайно являє собою деякий проміжок, скінченний або нескінченний. Безперервна випадкова величина визначена на незліченній множині елементарних подій Ω.

Нехай множина елементарних подій Ω збігається із множиною дійсних чисел $\Omega=\{x: x\in(-\infty, \infty)\}$ і нехай випадковими подіями з поля подій S є всілякі скінченні чи рахункові об'єднання. Імовірність на полі подій S задається за допомогою ненегативної функції $\varphi(x)$, інтегрувальної на будь-якому скінченному або нескінченному проміжку й такий, що $\int_{-\infty}^{\infty}\varphi(x)dx=1$. Якщо подія A дорівнює об'єднанню непересічних проміжків, то ймовірність $\boldsymbol{P}(A)$ дорівнює сумі інтегралів від $\varphi(x)$ по цих проміжках: $\boldsymbol{P}(A)=\int_A\varphi(x)dx$. Визначена в такий спосіб на S імовірність задовольняє всім аксіомам теорії ймовірностей. Описаний імовірнісний простір називається *безперервним імовірнісним простором*.

Не всяка випадкова величина, визначена на незліченній множині Ω, є безперервною. Випадкова величина X називається *безперервною*, якщо існує ненегативна функція $p_X(x)$ така, що при будь-яких x функцію розподілу $F_X(x)$ можна представити у вигляді: $F_X(x)=\int_{-\infty}^{x}p_X(y)dy$. Будемо розглядати тільки такі випадкові величини, для яких $p_X(x)$ безперервна скрізь, крім, бути може, скінченного числа точок.



Мал. 3.4

Функція $p_X(x)$ називається *щільністю розподілу ймовірностей* або *щільністю розподілу*. Якщо з контексту ясно, про яку випадкову величину мова йде, то пишуть $p(x)$ замість $p_X(x)$. З визначення випливає, що в точках безперервності щільність розподілу дорівнює похідній функції розподілу: $p(x)=F'(x)$. З визначення також випливає, що щільність розподілу $p_X(x)$ визначає закон розподілу випадкової величини X.

Щільність розподілу випадкової величини має наступні <u>*властивості*</u>:

1. Для будь-яких $x_1<x_2$ $\boldsymbol{P}\{x_1\le X<x_2\}=\int_{x_1}^{x_2}p(x)dx$.

Імовірність $\boldsymbol{P}\{x_1\le X<x_2\}$ дорівнює площі фігури, обмеженої прямими $x=x_1$, $x=x_2$, $y=0$ і щільністю розподілу $y=p(x)$. На мал. 3.4 зображена щільність розподілу ймовірностей $p(x)$ і заштрихована площа фігури, що дорівнює ймовірності $\boldsymbol{P}\{x_1\le X<x_2\}$.

2. Інтеграл по всій числовій прямій від щільності розподілу ймовірностей дорівнює одиниці, тобто $\int_{-\infty}^{\infty}p(x)dx=1$.
3. Імовірність того, що безперервна випадкова величина прийме конкретне значення, дорівнює нулю, тобто $\boldsymbol{P}\{X=a\}=0$.

Із властивості 3 випливає, що для безперервних випадкових величин

$$\boldsymbol{P}\{x_1 \le X < x_2\} = \boldsymbol{P}\{x_1 \le X \le x_2\} = \boldsymbol{P}\{x_1 < X < x_2\}.$$

Щільністю розподілу може бути будь-яка ненегативна функція, інтеграл від якої по всій числовій осі дорівнює одиниці.

<u>Приклад 3.2</u>. Нехай простір елементарних подій Ω є відрізком, Ω=[-1, 1], подіями є всілякі об'єднання проміжків, що належать відрізку [-1, 1]. Нехай далі ймовірність подій задається за допомогою функції $f(\omega)=½$, побудованої на всьому відрізку [-1, 1]: $\boldsymbol{P}\{A\}=\boldsymbol{P}\{\omega;\ \omega\in A\}=\int_A \frac{1}{2}d\omega$. Якщо $A=[a, b]\in[-1, 1)$, то $\boldsymbol{P}\{A\}=(b-a)/2$.

Ми визначили всі елементи імовірнісного простору $\{\Omega, S, P\}$. Задамо тепер на множині елементарних подій Ω випадкову величину. Нехай $X(\omega)=2\cdot|\omega|$ (наприклад, при $\omega=-0.4$ $X(\omega)=2\cdot 0.4=0.8$). Знайдемо функцію розподілу й щільність розподілу випадкової величини X. Так як $0\le X(\omega)\le 2$, то $F(x)=0$ для $x\le 0$ і $F(x)=1$ при $x\ge 2$. Знайдемо $F(x)$ для $x\in(0, 2)$. Множина $Ax=\{\omega: X(\omega)<x\}$ являє собою інтервал $(-x/2, x/2)$:

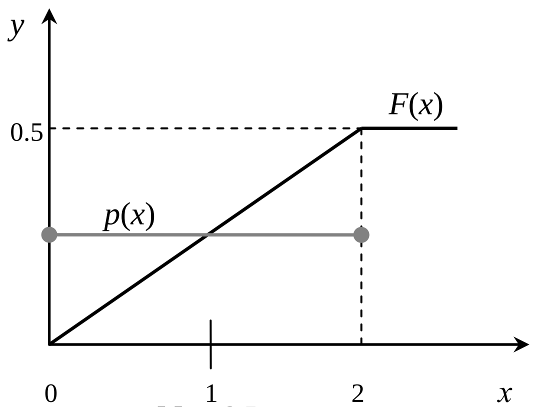
$$Ax=\{\omega: X(\omega)<x\}=\{\omega: 2\cdot|\omega|<x\}=\{\omega: -x/2<\omega<x/2\}.$$

Для $x\in(0, 2)$ функція $F(x)$ дорівнює

$$F(x)=\boldsymbol{P}\{A_x\}=\int_{A_x} f(\omega)d\omega=\int_{-\frac{x}{2}}^{\frac{x}{2}}\frac{1}{2}d\omega=\frac{1}{2}x.$$

У результаті маємо

$$F(x)=\begin{cases}0 & \text{при } x\le 0,\\ \frac{1}{2} & \text{при } 0<x\le 2,\\ 1 & \text{при } x>2.\end{cases}$$



Мал. 3.5

Графік функції $F(x)$ зображений на мал. 3.5. Функція розподілу $F(x)$ безперервна на всій числовій осі й дифференцируема скрізь, крім крапок $x=0$ і $x=2$. Отже, X – безперервна випадкова величина. Знайдемо її щільність розподілу

$$p(x)=F'(x)=\begin{cases}0 & \text{при } x<0,\\ \frac{1}{2} & \text{при } 0<x<2,\\ 1 & \text{при } x>2.\end{cases}$$

У крапках $x=0$ і $x=2$ можна покласти $p(x)$ рівної довільному числу, наприклад нулю. Графік щільності розподілу $p(x)$ також наведений на мал. 3.5.

Приведемо тепер закони розподілу безперервних випадкових величин.

3.3.1. *Рівномірний розподіл*

Безперервна випадкова величина X, що приймає значення на відрізку $[a, b]$, має рівномірний розподіл, якщо щільність розподілу $p(x)$ має вигляд

$$p(x)=\begin{cases}0 & \text{при } x\notin[a, b],\\ \frac{1}{b-a} & \text{при } x\in[a, b].\end{cases}$$

Для випадкової величини, що має рівномірний розподіл, імовірність того, що випадкова величина прийме значення із заданого інтервалу (x, x+Δ)∈[a, b], не залежить від розміщення цього інтервалу на числовій осі й пропорційна довжині цього інтервалу Δ:

$$\boldsymbol{P}\{x<X<x+\Delta\}=\int_{x}^{x+\Delta}\frac{1}{b-a}dt=\frac{\Delta}{b-a}.$$

Функція розподілу має вигляд:

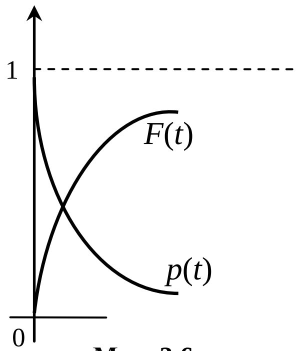
$$F(x)=\begin{cases}0 & при\ x\le a,\\ \dfrac{x-a}{b-a} & при\ a<x\le b,\\ 1 & при\ x>b.\end{cases}$$

3.3.2. *Експонентний розподіл*

Безперервна випадкова величина X, що приймає ненегативні значення, має експонентний (показовий) розподіл з параметром λ, якщо щільність розподілу випадкової величини при $x\ge 0$ дорівнює $p(x)=\lambda e^{-\lambda x}$ і при $x<0$ $p(x)=0$. Функція розподілу випадкової величини X дорівнює

$$F(x)=\int_{-\infty}^{x}p(t)dt=\begin{cases}0 & при\ x\le 0,\\ 1-e^{-\lambda x} & при\ x>0.\end{cases}$$

Графіки функції розподілу й щільності розподілу наведені на мал. 3.6.



Мал. 3.6

3.3.3. *Нормальний розподіл*

Безперервна випадкова величина X має нормальний розподіл ймовірностей з параметрами a й σ>0, якщо її щільність розподілу має вигляд [1, 4, 7]

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}.$$

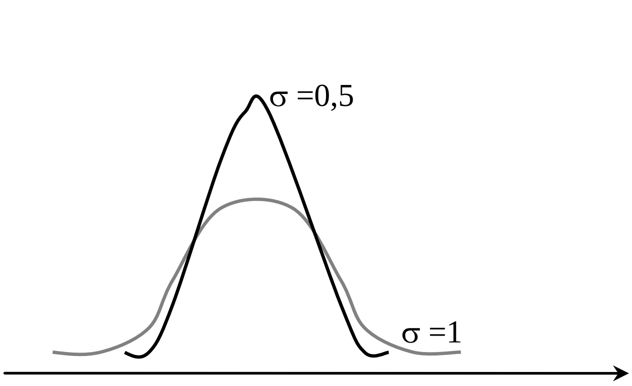


Рис. 3.7

Якщо X має нормальний розподіл, то коротко це записується у вигляді $X\sim N(a,\sigma)$.

Графіки щільності розподілу при різних значеннях параметра σ і тому самому значенні a наведені на мал. 3.7. Щільність розподілу $p(x)$ симетрична відносно прямої $x=a$ (тобто $p(a+y)=p(a-y)$). Якщо $x\to\pm\infty$, то $p(x)\to 0$. При зменшенні σ графік «стягається» до своєї осі симетрії $x=a$. Нормальний розподіл відіграє особливу роль у теорії ймовірностей і її застосувань. Це пов'язане з тим, що при виконанні певних умов сума великої кількості випадкових величин має «приблизно» нормальний розподіл.

Інтегральна функція цього розподілу виражається співвідношенням [4]

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t-a}{2\sigma^2}}dt\,.$$

Зміст кривої нормального розподілу полягає в наступному: значення випадкової величини з нормальним законом розподілу розташовуються по обидві сторони від генерального середнього *a*, і чим далі ці значення віддалені від *a*, тим вони менш імовірні. Крім того, чим менше σ^2, тим швидше гілки кривої наближаються до осі абсцис, тобто тим менш імовірні значення випадкової величини, що сильно ухиляються від *a* [5].

Незважаючи на те, що нормальний розподіл характеризує ряди з безперервними одиницями, при досить великій кількості спостережень воно добре описує й варіювання дискретних одиниць. До цього закону наближаються й інші типи розподілів. Прийнято вважати, що частотні одиниці мови (тексту) розподіляються за нормальним законом [6]. Навіть якщо на практиці отримані в ході лінгвостатистичних досліджень варіаційні ряди підкоряються якомусь іншому розподілу, вважається, що при низькій точності, пропонованої до лінгвостатистичним розрахунків, можна допустити, що розподіл є нормальним [8].

4. ЧИСЛОВІ ХАРАКТЕРИСТИКИ ВИПАДКОВОЇ ВЕЛИЧИНИ

4.1. 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 $x_1, x_2, \ldots, x_n$ (не обов'язково всі x різні між собою). Тоді говорять, що $(x_1, x_2, \ldots, x_n)$ 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, використаються 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 «вибіркове» опускають.

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

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

Щоб визначити характер розташування вибіркових значень щодо свого середнього арифметичного значення $\bar{x}$, обчислюється *вибіркова 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 розділить 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 *вибірковим середнім квадратичним відхиленням*.

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 *угрупування* вибіркових значень. Нехай є вибірка об'єму n $(x_1, x_2, \ldots, x_n)$. Можна вважати, що числа $x_1, x_2, \ldots, x_n$ розташовані в порядку зростання (цього завжди можна домогтися). Вибираємо деяке ціле число k, таке, що $10 \le k \le 20$, 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), …, $(x_1+(n\text{-}1)h, x_n)$. Кожне зі значень нашої випадкової вибірки $x_1, x_2, \ldots, x_n$ попадає в один із зазначених вище інтервалів. Так, якщо деяке вибіркове значення x_l задовольняє нерівності $x_1+lh \le x_l < x_1+(l+1)h$, то воно належить до інтервалу $(x_1+lh, x_1+(l+1)h)$. Позначимо через n_i кількість значень вибірки, що потрапили в інтервал $(x_1+(i\text{-}1)h, x_1+ih)$. Якщо деяке вибіркове значення потрапило на границю двох інтервалів, то воно належить до правого інтервалу. Числа n_i називаються *частотами відповідних інтервалів*. Тепер для визначення $\overline{x}$ 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:

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

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

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 вибіркового середнього й вибіркової дисперсії. Щоб нівелювати цю погрішність, можна внести виправлення на групування в $\overline{x}$ 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 вибіркова дисперсія добре наближає справжню дисперсію.

4.2. 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 $\overline{x}$, а σ^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 σ^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 $\overline{x}$ так само, як і вибіркова дисперсія S^2, є наближеннями до a и σ^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 σ^2 має наступну важливу властивість: з імовірністю β варіанти вилучені від a на відстань, не більше $d_\beta\sigma$, де d_β можна визначити з таблиці 4.1.

Таблиця 4.1

β	0.687	0.900	0.950	0.980	0.990	0.999
d_β	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 3σ. 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.

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 «чутливість», 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

$$\chi^2 = n\sum_{i=1}^{k}\frac{\left(\frac{n_i}{n}-p_i\right)^2}{p_i}, \tag{4.1}$$

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 χ^2 використається 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 χ^2 визначає критерій перевірки гіпотези H_0. Його називають *критерієм «хі-квадрат»*. Застосування критерію засноване на тім, що для обраного рівня значимості визначають табличне χ^2_{05} або χ^2_{01} й порівнюють зі значенням χ^2, обчисленим по формулі (4.1). Щоб знайти табличне χ^2, варто звернутися до таблиці розподілу χ^2 (див. Додаток, табл. 2) для значення $f=k-r-1$, де k - кількість інтервалів групування, використаних при обчисленні χ^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 χ^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 розбивка спостережень вибірки на k інтервалів і нехай $y_1, y_2, \ldots, y_{k+1}$ – границі всіх інтервалів.
2. Визначаються вибіркові $\overline{x}$ й S^2.
3. Обчислюються різниці $y_1 - \overline{x}$, $y_2 - \overline{x}$, …, $y_{k+1} - \overline{x}$ ...

4. Ці різниці нормуються, тобто обчислюються величини

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

5. По таблиці знаходимо величини (див. Додаток, табл.1) $\Phi(u_1)$, $\Phi(u_2)$, …, $\Phi(u_{k+1})$.
6. Складаємо різниці $p_1=|\Phi(u_1)-\Phi(u_2)|$, …, $p_k=|\Phi(u_k)-\Phi(u_{k+1})|$.
7. Обчислюємо $\tilde{n}_i=p_i n$.
8. Визначаємо значення χ^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 χ^2, що й формула (4.1), 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 противному 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 $\overline{x}$ й S^2).

Обчислене χ^2 порівнюємо з табличним χ^2_{05} (або χ^2_{01}). Якщо виявиться, що $\chi^2>\chi^2_{05}$, то гіпотеза H_0 відкидається з рівнем значимості 0.05. Рівень значимості α (звичайно α=0.05 або α=0.01) означає, що ймовірність одержання даної вибірки в умовах гіпотези H_0 менше α, тобто гіпотеза H_0 слабко погоджується з даною вибіркою, а тому повинна бути відкинута. Якщо ж $\chi^2<\chi^2_{05}$, 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 χ^2 задовольняє нерівності $\chi^2_{05}<\chi^2<\chi^2_{01}$, тому що дослідник (з обережності) може вибрати рівень значимості α=0.01, а тому не відкидає гіпотезу H_0, але наведена нерівність ставить під сумнів його висновок.

4.3. Критерій однорідності

Цінність критерію χ^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 аудірування 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. Висунуту гіпотезу можна перевірити за допомогою критерію χ^2. Процес обчислення χ^2 у цьому випадку розглянемо на наступному прикладі.

Приклад 4.1. Значення частоти основного тону голосного з головним наголосом в емоційних реченнях, що виражають подив, зазначені в таблиці 4.2.

Таблиця 4.2.

Інтервали (частота основного тону)	Частоти для положит. відтінків емоції (n_{1i})	Частоти для отрицат. відтінків емоції (n_{2i})	$n^{(j)}$	$\frac{n_1 n^{(j)}}{N}$	$\frac{n_2 n^{(j)}}{N}$
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 χ^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 χ^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 χ^2_{05}=9.488 і χ^2_{01}=13.30. Оскільки $\chi^2=1.61<9.488=\chi^2_{05}$, 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 $P_1, \ldots, P_k$ (у прикладі 4.5 якості характеризувалися інтервалами групування). Нехай є l вибірок відповідно об'ємів $n_1, \ldots, n_l$ (у прикладі 4.5 n=2). Позначимо через n_{ij} кількість елементів в j-й вибірці, що володіють якістю P_j. Очевидно, що $n_i=n_{i,1}+\ldots+n_{i,k}$. Позначимо ще $n^{(j)}=n_{1,j}+\ldots+n_{l,j}$; $N=n_1+\ldots+n_l$. Значення χ^2 обчислюється по формулі

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

Тут підсумовування ведеться по всіляких парах чисел (i, j) $i=1, \ldots, k$, $j=1, \ldots, k$. Це значення χ^2 порівнюється з табличним χ^2_{05} або χ^2_{01}, обчисленим для $f=(k-1)(n-1)$. Для випадку n=2 ми, мабуть, маємо випадок, уже розібраний у прикладі 4.5.

4.4. 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 $\bar{x}_1, \ldots, \bar{x}_k$ 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 3σ. Тому, визначивши вибіркове середнє $\bar{x}$, можна з імовірністю 0.99 указати інтервал (його серединою буде $\bar{x}$), у якому втримується генеральне середнє $\hat{x}$. Але для цього треба знать середнє квадратичне σ нормального розподілу, якому підкоряється $\bar{x}$.

Як оцінку для σ 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 звузити. 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_β вибирається з таблиці 4.1 на підставі заданого значення ймовірності.

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

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

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 противному випадку до оцінок параметрів розподілу (a і σ), що даються цією вибіркою, варто ставитися з обережністю.

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 інтенограми значення частоти основного тону. Природно, що при вимірі частоти основного тону цілком може бути допущена якась помилка, величина якої визначається границями довірчого інтервалу X.

5. ЛІНГВІСТИЧНІ ГІПОТЕЗИ

5.1. Постановка й перевірка лінгвістичних гіпотез

Постановка лінгвістичної задачі визначається метою дослідження. Залежно від задачі планується й проводиться експеримент. Чітко поставлена лінгвістична задача є необхідною умовою ефективності статистичного аналізу.

Припустимо, досліджуються інтонаційні одиниці певної мови. Фрази, що представляють ту або іншу інтонаційну одиницю, відбираються дослідником у результаті лінгвістичного аналізу із залученням носіїв даної мови. Над отриманими групами фраз проводиться електроакустичний аналіз, у результаті якого кожна фраза подається фізичними характеристиками - частотою основного тону, інтенсивністю й тривалістю. Далі дослідник, спираючись на вже наявні результати експерименту, визначає окремі ознаки, гіпотетично істотні для даної інтонаційної одиниці. У підсумку будуть отримані інтонаційні ознаки, подані певним рядом значень тієї або іншої фізичної характеристики. До цих вихідних фізичних значень і застосовується ймовірносно-статистичний аналіз.

Звичайно виходять із припущення, що ряд значень фізичної величини являє собою розподіл деякої випадкової величини. З досліджень по лінгвістиці відомо, що мовним явищам властиво статичне коливання, потенційність. Однак границі варіювання даних характеристик не довільні, існують «області розсіювання» кожної лінгвістичної одиниці. Виходячи із властивості потенційності мовної одиниці й існування певних границь варіювання її, можна зробити висновок про ймовірносно-статистичний характер прояву даної одиниці. Розподіл значень фізичних характеристик розпізнавальних ознак двох лінгвістичних одиниць, незважаючи на деяке перетинання їхніх інтервалів, має досить чітке протиставлення, що виражається в статистичних параметрах. Це підтверджується даними багатьох експериментально-лінгвістичних досліджень. Отже, кожна лінгвістична одиниця характеризується певним імовірнісним розподілом (або розподілами), представленими у своїх фізичних характеристиках.

На початковому етапі статистичного аналізу йде ***опис*** у статистичних термінах того різноманіття значень, у яких представлений кожний з гіпотетично істотних лінгвістичних ознак. Без такої обробки фізичних даних не можна виділити із усього різноманіття величин, що змінюються безупинно, самі характерні прояви, типові для досліджуваного явища. Цей етап припускає обчислення середнього арифметичного значення й дисперсії, трьохсигмених границь для середньоарифметичного, побудову розподілу (гістограми) варіант і т.д.

Наступний етап застосування статистичного аналізу є саме перевіркою певних лінгвістичних гіпотез. Це може бути вирішення питання істотності тієї або іншої ознаки для лінгвістичної одиниці в результаті застосування критеріїв розходження (критерію Стьюдента, Ван дер Вардена).

Перебираючи в такий спосіб ознаки, установлюється набір розпізнавальних ознак для кожної лінгвістичної одиниці (тут виходять із припущення, що розходження на лінгвістичному рівні обов'язково знаходить своє відбиття у фізичних характеристиках). Ймовірносно-статистичний аналіз дозволяє також розглянути взаємодію різних ознак усередині лінгвістичної одиниці. При цьому рекомендується застосовувати кореляційний аналіз або аналіз кореляційних відносин (якщо розподіл варіант відрізняється від нормального). Визначити ступінь залежності тих або інших ознак може допомогти регресійний аналіз. Якщо лінгвіст припускає дію на розподіл варіант якогось другорядного фактору, то перевірити цю гіпотезу можна за допомогою дисперсійного аналізу [5].

5.2. 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 5.1. Для двох груп дикторів (група **А** – чоловіки, група **В** – 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 5.1

Група А

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 В

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 **А**, а $\bar{y}$ - середнє значення для групи **В**.

Розходження вибіркових середніх (310 і 236) ще не доводить, що групи **А** і **В** істотно різні з точки зору впливу на значення сумарної енергії. Висувається гіпотеза H_0: розходження вибіркових середніх $\bar{x}$ і $\bar{y}$ є несуттєвим, випадковим, тобто справжні середні значення сумарної енергії для груп **А** і **В** однакові. Для перевірки цієї гіпотези обчислюємо $s_{\bar{x}-\bar{y}}$ по формулі:

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

Де n_1 і n_2 – об'єми відповідно вибірок у групах **А** і **В**, 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 **А** і **В**. Тому

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

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

Тому

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

Тепер обчислюємо

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

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 А і В на сумарну енергію помилково.

У розглянутому прикладі вибірки для груп **А** і **В** 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

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

Критерій Стьюдента в лінгвістиці часто використається 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 (попарно). 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 розходження 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 включать 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 вибіркових 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 використається 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.

5.3. 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 5.2. 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 5.2

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. Складемо 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 Ψ (Додаток, табл. 4) знаходимо $\Psi(\frac{r}{n+1})$, а потім складаємо суму $X=\sum_r \Psi(\frac{r}{n+1})$. У розглянутому прикладі

$$X=\Psi(\frac{3}{16})+\Psi(\frac{4}{16})+\Psi(\frac{7}{16})+\Psi(\frac{8}{16})+\Psi(\frac{11}{16})+\Psi(\frac{12}{16})+\Psi(\frac{14}{16})+\Psi(\frac{15}{16})=$$

$$=(-0.89)+(-0.67)+(-0.16)+(0.00)+(0.49)+(0.67)+(1.15)+(1.53)=(-1.72)+(3.84)=2.12.$$

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 противному 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 $|n_1 - n_2| \leq 5$).

6. ВИВЧЕННЯ ЗАЛЕЖНОСТІ ЛІНГВІСТИЧНИХ ОЗНАК

6.1. 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 багатопричинним зв'язком. Зведення складної системи відносин до більш простих її видів, виділення тих зв'язків, які є основними в досліджуваній функції язикового 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 *тіснотою* кореляційного зв'язку розуміється сила впливу досліджуваних ознак однієї на одну. По <u>тісноті</u> кореляція може бути *слабкою, середньою* й *сильною*. *Форма* кореляційного зв'язку показує, як у середньому змінюються значення однієї ознаки при зміні іншої. За <u>формою</u> 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.

<u>Приклад 6.1</u>. Нехай є дві ознаки, X і Y, що характеризують одне й теж 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*.

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

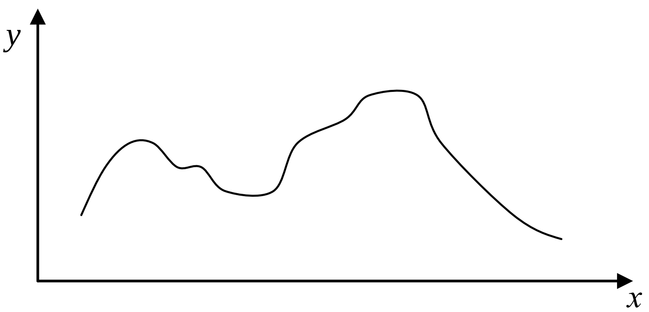
Повна зворотна кореляція

x	4	6	9	10	12
y	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 уважається *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. 6.1 – 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.

У лінгвістичних дослідженнях кореляційний аналіз застосовується досить широко. Він служить для встановлення подібності й зв'язку між одиницями мови й мовлення, якщо відомо частоту зустрічальності цих одиниць. Зокрема, використання кореляційного аналізу для встановлення парадигматичних зв'язків між словами засновано на припущенні, що чим більшою мірою збігається лексична сполучуваність двох слів, тим більшою мірою проявляється семантична подібність між цими словами. Досвід лінгвістичних досліджень показує, що при використанні кореляційного аналізу для встановлення парадигматичних відносин у лексиці необхідно розрізняти два основних процедурних варіанти [6]:

- ***формальна процедура***, коли аналіз здійснюється на основі частоти сполучуваності досліджуваних одиниць без якої-небудь додаткової семантичної обробки й угруповання останніх;

- ***напівформальна процедура***, коли аналіз проводиться на основі спільної зустрічальності (сполучуваності) досліджуваної одиниці із семантичними підкласами слів, угруповання й вичленовування яких здійснюється дослідником неформальним способом.

Експериментальним шляхом установлене, що застосування чисто формальної процедури підвищує ступінь об'єктивності одержуваних результатів, але знижує ступінь їхньої повноти й відповідності інтуїтивним уявленням про зв'язки слів, відбитих у тлумачних і синонімічних словниках. Застосування напівформальної процедури знижує ступінь об'єктивності (тому що дослідник інтуїтивно робить якісне угруповання контекстуальних партнерів по подібності їхньої семантики) одержуваних результатів, але більш повно відбиває сполучуванісні властивості всіх лексико-семантичних варіантів, а разом з тим і парадигматичні зв'язки аналізованих слів. Ще одна особливість використання кореляційного аналізу полягає в тому, що часто вихідна таблиця, на основі якої здійснюється кореляційний аналіз, залежно від цілей і задач дослідження використається двічі: при знаходженні коефіцієнтів кореляції таблиця розвертається таким чином, щоб рядки й стовпці в ній мінялися місцями. Наприклад, були досліджені парадигматичні відносини між прикметниками зі значенням «міцний» в англійській мові на основі їхньої сполучуваності з підкласами іменників. У результаті були отримані дані про парадигматичні зв'язки цих прикметників і на цій основі побудоване «семантичне поле». Однак, строго говорячи, дослідження констатують у таких ситуаціях не семантичне, а лексичне поле, тому що встановлюється зв'язок між словами, а не між значеннєвими одиницями. Розгорнувши вихідну таблицю, де була зафіксована частота зустрічальності прикметників з підкласами іменників («значеннєвими одиницями», вираженими й обмеженими певним набором слів), були знайдені величини коефіцієнтів кореляції між частотами семантичних підкласів і на цій основі побудоване семантичне поле, що покриває лексемами зі значенням «міцний» в англійській мові. Потім був здійснений наступний крок: було проведене порівняння семантичного поля «міцний» в англійській мові з відповідним семантичним полем у французькій мові [14, 15]. В інших аналогічних випадках можна досліджувати зв'язки між іншими одиницями мови на фонетичному, морфологічному або синтаксичному рівнях.

6.2. 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

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

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

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 $x_1, x_2, \ldots, x_n$ (відповідно, $y_1, y_2, \ldots, y_n$) 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.

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 ± 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.

6.3. 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 використають 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 ґрат 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 $e_{y/x}$:

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

6.4. Парціальна кореляція

У більшості випадків при лінгвістичних дослідженнях характер кореляції між двома ознаками повністю або частково залежить від того, що дві досліджуваних 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:

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

6.5. Критерії кореляційного аналізу

6.5.1. *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.

6.5.2. *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 використається 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.

6.6. Регресійний аналіз

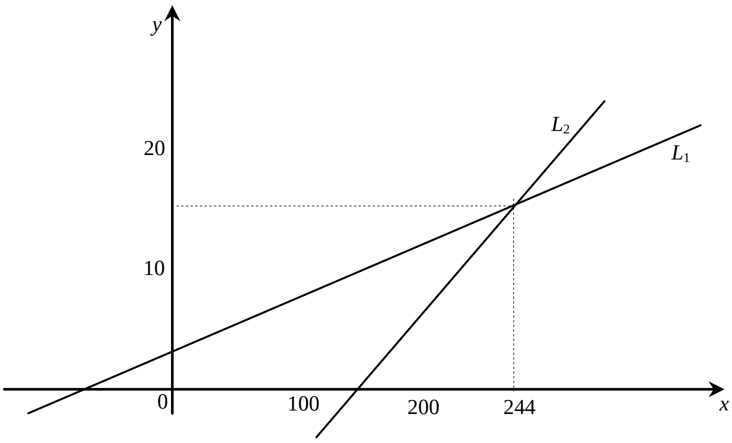
З кореляційного аналізу відомо, як перевірити наявність залежності між ознаками й установити тісноту цього зв'язку. Кореляційні відносини дають нам ступінь однобічного зв'язку ознак, але не вказують, наскільки змінюється одна ознака зі зміною іншої. Часто буває необхідно довідатися, як поводиться ознака 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 ±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 6.2. Тоді

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

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).



Мал. 6.3 – 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 6.2 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 90^0, то практично лінійний зв'язок між ознаками відсутній. У цьому випадку доцільно побудувати емпіричну криву регресії Y по X (і аналогічно X по Y). Процес побудови емпіричної лінії регресії виглядає в такий спосіб. Нехай (x_1, y_1), …, (x_n, y_n) – n пар отриманих значень ознак X і Y, і нехай X_1, …, X_k – всі різні значення X серед чисел x_1, …, x_n (деякі з x_i можуть збігатися, так що $k \le n$). Якщо значення X_1 зустрілося серед чисел x_1, …, x_n рівно n_1 раз, то йому відповідає n_1 значень Y серед чисел y_1, …, y_n. Позначимо через $\overline{Y}_1$ середнє значення цих значень Y. Одержуємо пари $(X_1, \overline{Y}_1)$. Аналогічно знаходять пари $(X_2, \overline{Y}_2)$, …, $(X_k, \overline{Y}_k)$. Пари точок $(X_1, \overline{Y}_1)$, …, $(X_k, \overline{Y}_k)$ зображуємо на малюнку, а потім побудовані точки з'єднуємо прямими відрізками. Отримана ламана лінія і є емпіричною лінією регресії Y по X.

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.

Додаток 1

Значення функції нормального розподілу Φ(*u*)

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

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

Додаток 2

Критичні значення χ^2

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

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

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	∞	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 $t \le t_{05}$ й відкидається при $t > t_{01}$.

Додаток 4

Значення функції Ψ

p→ ↓	0	1	2	3	4	5	6	7	8	9
0.00	- ∞	- 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→ ↓	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	∞	∞	∞	∞	∞	∞
7	∞	∞	∞	∞	∞	∞
8	2.40	∞	2.30	∞	∞	∞
9	2.48	∞	2.20	∞	∞	∞
10	2.60	3.20	2.49	3.10	2.30	∞
11	2.72	3.40	2.58	3.40	2.40	∞
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 firmy 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.