

Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Теорія імовірностей та математична статистика.
Курс лекцій

*Рекомендовано Методичною радою
КПІ ім. Ігоря Сікорського*

Київ
2018

Теорія імовірностей та математична статистика. Курс лекцій. / Уклад.:
Т.А.Ліхоузова – К.: КПІ ім. Ігоря Сікорського, 2018. – 300 с.

Посібник призначений для студентів спеціальностей 121 «Інженерія програмного забезпечення» та 126 «Інформаційні системи та технології» всіх форм навчання.

Укладач

Ліхоузова Т.А., к.т.н., доцент

Відповідальний редактор

Ткач М.М., к.т.н., доцент

Рецензенти

Новицький В.В., д.ф.-м.н., проф.
провідний науковий співробітник
Інституту математики НАН України

Бридун Є.В., к.е.н., доцент
доцент кафедри фінансів
Національного університету
Києво-Могилянська академія

За редакцією укладачів

Зміст

Розділ 1. Теорія випадкових подій	10
1.1. Основні поняття теорії імовірностей. Операції над подіями	10
Предмет теорії імовірностей	10
Алгебра випадкових подій	13
Операції над подіями	16
Підсумки	20
1.2. Поняття імовірності. Імовірний простір	21
Класичне визначення імовірності	21
Геометричне визначення імовірності	23
Статистичне визначення імовірності	24
Основні властивості імовірності	26
Основні поняття та принципи комбінаторики	27
Правила суми і добутку	32
Підсумки	33
1.3. Основні теореми теорії імовірностей. Умовні імовірності	34
Додавання імовірностей несумісних подій	34
Імовірність появи хоча б однієї випадкової події	40
Імовірність суми сумісних випадкових подій	41
Надійність системи	43
Принцип практичної неможливості малоімовірних подій	45
Підсумки	46
1.4. Формули повної імовірності та Байєса	47

Формула повної імовірності.....	47
Формула Байєса.....	49
Підсумки	52
Розділ 2. Теорія випадкових величин.....	53
2.1. Випадкові величини	53
2.2. Закони розподілу імовірностей та числові характеристики дискретних випадкових величин	55
Способи задання закону розподілу ДВВ	55
Числові характеристики	57
Підсумки	67
2.3. Законі розподілу імовірностей та числові характеристики неперервних випадкових величин	68
Числові характеристики	73
Підсумки	77
2.4. Стандартні закони розподілу імовірностей дискретних випадкових величин	78
Повторні випробування.....	78
Біноміальний закон розподілу	82
Граничні теореми у схемі Бернуллі.....	83
Закон розподілу Пуассона.....	84
Локальна теорема Лапласа	86
Інтегральна теорема Лапласа	88
Геометричний розподіл	90
Гіпергеометричний розподіл	90

Поліноміальний розподіл	92
Послідовність випробувань із різними імовірностями	92
Підсумки	94
2.5. Стандартні закони розподілу імовірностей неперервних випадкових величин	96
Рівномірний розподіл	96
Показниковий розподіл	98
Нормальний розподіл	100
Правило трьох сигм	106
Підсумки	107
2.6. Закони великих чисел. Граничні теореми	108
Нерівність Чебишова	109
Граничні теореми	111
Теорема Бернуллі	111
Теорема Чебишова	115
Центральна гранична теорема (Ляпунова)	117
Підсумки	118
2.7. Функції випадкових величин та їх характеристики	120
Закон розподілу функції дискретного випадкового аргументу	120
Закон розподілу функції неперервного випадкового аргументу	122
Функція двох випадкових аргументів.....	125
Розподіл χ^2 («хі-квадрат»).....	127
Розподіл Стюдента.....	128

Розподіл Фішера-Снедекора	129
Підсумки	129
2.8. Багатомірні випадкові величини та їх закони розподілу імовірностей	131
Закон розподілу імовірностей дискретної двовимірної випадкової величини.....	132
Закон розподілу імовірностей неперервної двовимірної випадкової величини.....	135
Залежні та незалежні випадкові величини	137
Умовні закони розподілу складових системи випадкових величин	137
Числові характеристики багатомірних випадкових величин	138
Підсумки	142
Розділ 3. Математична статистика	144
3.1. Способи збору даних	144
Задачі, які розв'язує математична статистика	144
Генеральна та вибіркова сукупності	144
Джерела даних у статистиці. Способи відбору.....	147
Спотворення даних вибірок	150
Підсумки	152
3.2. Первинна обробка даних	153
Шкали для вимірювання ознак	153
Організація даних.....	154
Ряд розподілу.....	156
Розподіл частот.....	156

Підсумки	159
3.3. Описова статистика.....	160
Оцінка міри центральної тенденції	160
Оцінка міри варіативності.....	164
Графічні зображення статистичного розподілу	168
Гістограма частот та функція розподілу.....	169
Закон розподілу	172
Аналіз нормальності розподілу	176
Загальна схема аналізу даних	178
Підсумки	178
3.4. Точкові оцінки параметрів розподілу	180
Вимоги до статистичних оцінок	180
Поняття про точкові оцінки параметрів розподілу	182
Метод моментів для визначення точкових оцінок	185
Метод найбільшої подібності для визначення точкових оцінок.....	188
Підсумки	194
3.5. Інтервальні оцінки параметрів розподілу	195
Надійність оцінки.....	195
Загальна методика визначення інтервальних оцінок	196
Довірчий інтервал для оцінки математичного сподівання нормального розподілу при відомому σ	197
Знаходження об'єму вибірки, необхідного для розрахунку інтервальних оцінок із заданою надійністю	199

Довірчий інтервал для оцінки математичного сподівання нормального розподілу при невідомому σ	200
Довірчий інтервал для оцінки середньоквадратичного відхилення σ нормального розподілу	202
Підсумки	205
3.6. Статистичні гіпотези.....	206
Різновиди статистичних гіпотез	206
Похибки при перевірці гіпотез	208
Критерії узгодження	209
Критична область. Потужність критерію	213
Загальна схема при перевірці статистичних гіпотез	218
Проміжні підсумки.....	218
Перевірка гіпотез про вигляд закону розподілу	220
Перевірка гіпотез про параметри розподілу.....	230
Гіпотези про дисперсію	230
Гіпотези про середнє	241
Гіпотези про числові характеристики.....	254
Підсумки	257
3.7. Аналіз впливу факторів	259
Системи випадкових величин	259
Формулювання гіпотези	260
Залежні та незалежні вибірки	260
Вибір методу для аналізу впливу фактора при незалежних вибірках	262

t-критерій Стьюдента.....	262
U-критерій Манна-Уїтні.....	264
Дисперсійний аналіз (ANOVA)	266
Детальніше про дисперсійний аналіз.....	268
Критерій Краскала-Уоллеса.....	274
Критерій χ^2	275
Коефіцієнт кореляції.....	280
Вибір методу для аналізу впливу фактора при залежних вибірках.....	283
Підсумки	285
3.8. Елементи регресійного аналізу	286
Постановка задачі.....	287
Різновиди регресії	287
Порядок дій при регресійному аналізі	289
Визначення параметрів рівняння регресії. Одномірна лінійна регресія	295
Підсумки	300

Розділ 1. Теорія випадкових подій

Перші роботи, в яких виникли основні поняття теорії імовірностей, з'явилися у XV–XVI ст. як спроба побудови теорії азартних ігор і належать таким видатним ученим, як Б.Спіноза, Дж.Кардано, Галілео Галілей.

Наступний етап (кінець XVII — початок XVIII ст.) розвитку теорії імовірностей пов'язаний з роботами Б.Паскаля, П.Ферма (знаменита переписка), Х.Гюйгенса (1657р. перша книга з теорії імовірностей «О расчетах в азартной игре»), К.Гаусса, Я.Бернуллі та Н.Бернуллі (1713р. «Искусство предположения» перші теоретичні обґрунтування накопичених раніше фактів, класичне визначення імовірності), С.Пуассона, А.Муавра (ввів терміни незалежність подій, математичне сподівання, умовна імовірність), П.Лапласа, Т.Байєса.

В XIX ст. теорію імовірностей почали успішно застосовувати у страховій справі, артилерії, статистиці.

Лише наприкінці XIX ст. П.Л.Чебишов та його учні А.А.Марков та А.М.Ляпунов перетворили теорію імовірностей у математичну науку.

1.1. Основні поняття теорії імовірностей. Операції над подіями

Предмет теорії імовірностей

Розглянемо деякий дослід, у результаті якого може з'явитись або не з'явитись подія A .

Прикладами такого дослідження можуть бути:

- а) дослід — виготовлення певного виробу, подія A — стандартність цього виробу;
- б) дослід — кидання монети, подія B — випав герб;
- в) дослід — стрільба п'ятьма пострілами у мішень, подія C — вибито 30 очок;
- г) дослід — введення програми у комп'ютер, подія D — безпомилковий ввід.

Загальним для усіх дослідів є те, що кожен із них може реалізуватись у певних умовах скільки завгодно разів. Такі дослідни називають випробуваннями.

- **Випробуванням** називається експеримент, який можна проводити в однакових умовах (принаймні теоретично) будь-яке число разів.
- Найпростіший результат випробування називається **елементарною подією** або **наслідком**. При випробуванні обов'язково настає лише один наслідок.

Приклад. Стрілок робить постріл по мішені, що розділена на чотири області.

Постріл — це випробування.

Влучення в певну область мішені — подія.

Приклад. В урні знаходяться кольорові кулі. З урни навмання беруть одну кулю.

Виймання кулі — це випробування.

Поява кулі певного кольору — подія.

Події бувають достовірні, випадкові та неможливі.

- **Достовірною** називають таку подію, яка при розглянутих умовах обов'язково трапиться.
- **Неможливою** називають таку подію, яка при розглянутих умовах не може трапитись.
- **Випадковою** називають таку подію, яка при умовах, що розглядаються, може трапитися, а може й не трапитися.

Приклад.

Якщо в урні є лише білі кулі, то добування білої кулі з урни — достовірна подія, а добування з цієї урни кулі іншого кольору — неможлива подія.

Якщо кинути монету на площину, то поява герба буде випадковою подією, тому що замість герба може з'явитися надпис.

Випадкові події позначають великими літерами, наприклад

$A, B, C, D, X, Y, A_1, A_2, \dots, A_n$.

Кожна випадкова подія є наслідком багатьох випадкових або невідомих нам причин, які впливають на подію. Тому неможливо передбачити наслідок одиночного випробування.

Але якщо розглядати випадкову подію багато разів при однакових умовах, то можна виявити певну закономірність її появи або не появи. Таку закономірність називають **імовірною закономірністю масових однорідних випадкових подій**.

- ✓ У теорії імовірностей під **масовими однорідними випадковими подіями** розуміють такі події, які здійснюються багатократно при однакових умовах або багато однакових подій.

Наприклад, кинути одну монету 1000 разів або 1000 однакових монет кинути один раз в теорії імовірностей вважають однаковими подіями.

- ✓ Предметом теорії імовірностей є вивчення імовірнісних закономірностей масових однорідних випадкових подій.

Алгебра випадкових подій

- Події називають *несумісними*, якщо поява однієї з них виключає появу інших подій в одному і тому ж випробуванні.

Приклад. Серед однорідних деталей у ящику є стандартні та нестандартні. Навмання беруть із ящика одну деталь.

Події

A — взята стандартна деталь,

B — взята нестандартна деталь

несумісні тому, що одна деталь не може бути одночасно стандартною та нестандартною.

- Події називають *сумісними*, якщо поява однієї з них не виключає можливості появи інших.

Приклад. Два стрільця стріляють у мішень.

Події

A_1 — перший стрілок влучив у мішень,

A_2 — другий стрілок влучив у мішень

будуть сумісними випадковими подіями.

- Випадкові події A_1, A_2, \dots, A_n утворюють *повну групу подій*, якщо вони попарно несумісні і внаслідок випробування одна з них з'явиться обов'язково.

Приклад. Кидають шестигранний кубик.

Позначимо події так:

A_1 — випала грань 1;

A_2 — випала грань 2;

A_3 — випала грань 3;

A_4 — випала грань 4;

A_5 — випала грань 5;

A_6 — випала грань 6.

Події A_1, A_2, \dots, A_6 утворюють повну групу.

Приклад. Стрілок стріляє у мішень.

Події

A_1 — стрілок влучив у 1 коло мішені,

A_2 — стрілок влучив у 2 коло мішені.

Події A_1 та A_2 не утворюють повної групи. Але якщо позначити A_0 подію, що стрілок не влучив у мішень, тоді події A_0, A_1 та A_2 утворюють повну групу.

- Події називають **рівноможливими**, якщо немає причин стверджувати, що будь-яка з них можливіша за інші.

Приклад. Події — поява 1, 2, 3, 4, 5 або 6 очок при киданні шестигранного кубика — рівноможливі при умові, що центр його ваги не зміщений.

- Дві несумісні події, які утворюють повну групу, називають **протилежними**.

Подія, протилежна події A , позначається \bar{A} .

Приклад. Якщо позначити через A подію, що при стрільбі по мішені вибито 8 очок, то подія \bar{A} — при стрільбі по мішені вибито будь-яке інше число очок.

Тепер розглянемо важливе поняття ***простору елементарних наслідків***.

Нехай виконується деякий експеримент, який має елементи випадковості. Кожне випробування може мати різні наслідки.

Так, при киданні монети можуть бути два можливих наслідки: герб або надпис.

При киданні грального кубика можуть бути шість можливих наслідків.

У випробуванні «постріл у мішень» можна розглядати такі наслідки, як влучення у мішень, або кількість вибитих очок, або координати точки влучення.

Отже, що приймати за наслідок випробування, залежить від умови задачі.

- ***Елементарними наслідками*** називають такі події, які неможливо розділити на більш прості.
- Множину усіх можливих елементарних наслідків називають ***простором елементарних наслідків***.
- ✓ Простір елементарних наслідків може містити кінцеву, злічену або незлічену множину елементів.

У ролі елементарних наслідків можна розглядати точки n -вимірного простору, відрізок деякої лінії, точки поверхні S або об'єму V трьохвимірного простору, функцію однієї або багатьох змінних.

- ✓ У більшості випадків припускають, що елементарні наслідки рівноможливі.

Приклади.

а) При двократному киданні монети простір елементарних наслідків містить 4 точки

$$\{(G, G), (G, H), (H, G), (H, H)\},$$

де G — означає появу герба, H — появу надпису.

б) Нехай по мішені стріляють одиночними пострілами до першого влучення. Можливі такі елементарні події

w_1 {влучення при першому пострілі},

w_2 {влучення при другому пострілі},

w_3 {влучення при третьому пострілі} і т.д.

У цьому випадку простір елементарних наслідків може мати нескінченну кількість точок, які можна шляхом нумерації перелічити. Тому простір елементарних наслідків буде зліченим.

в) При виробництві кінескопів виникають неоднакові умови технологічного процесу, тому час роботи кінескопа відрізняється від його номінального значення, тобто буде випадковою подією.

Простір елементарних наслідків у цьому випадку буде нескінченною незліченою множиною, елементи якої неможливо пронумерувати.

Операції над подіями

Нехай A та B — випадкові події.

➤ **Об'єднанням (сумою) випадкових подій $A \cup B$** (або $A + B$) називають таку випадкову подію, яка полягає у появі подій A або B або A та B .

Якщо A та B — несумісні, то $A \cup B$ означає появу події A або події B .

Аналогічно визначають об'єднання (суму) більшої кількості випадкових подій.

➤ **Об'єднанням (сумою) випадкових подій** $A_1 \cup A_2 \cup \dots \cup A_n$ називають таку випадкову подію, яка полягає в появі хоча б однієї з цих подій.

Якщо події попарно несумісні, то їх сума полягає в тому, що повинна з'явитися подія A_1 або A_2 ... або A_n . Нескінченну суму випадкових подій позначають

$$\bigcup_{k=1}^{\infty} A_k$$

Приклад. Стрілок робить один постріл у мішень, поділену на три області. Позначимо

подія A_1 — влучення в першу область;

подія A_2 — влучення у другу область;

подія A_3 — влучення в третю область;

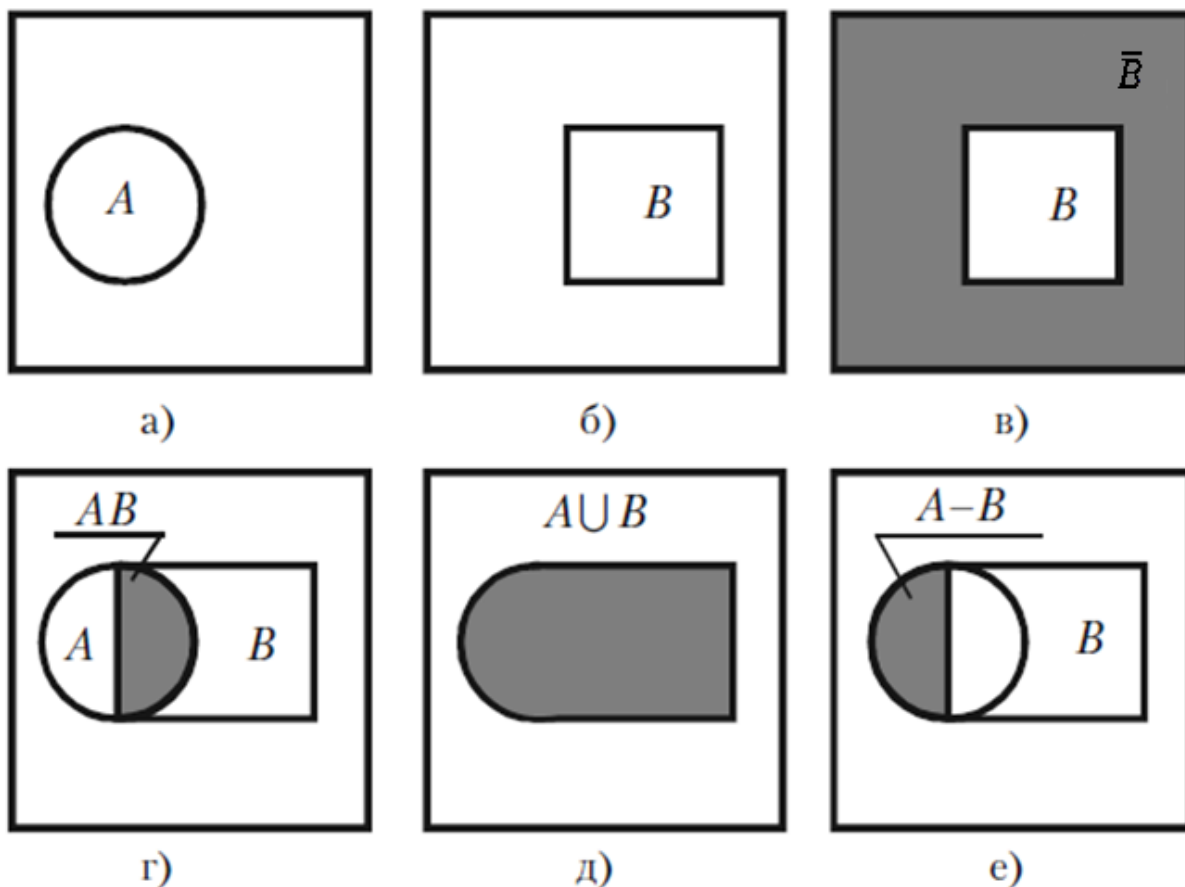
подія A_4 — немає влучення у мішень;

подія B — влучення в першу або другу області;

подія D — влучення хоча б в одну область мішені.

Тоді маємо $B = A_1 \cup A_2$; $D = A_1 \cup A_2 \cup A_3$.

В цьому прикладі події A_1, A_2, A_3 та A_4 — несумісні.



- **Різницею $B \setminus A$ (або $B - A$)** двох випадкових подій B, A називають усі наслідки, які полягають у тому, що подія B з'являється, а подія A не з'являється.
- **Добутком (перетином) $A \cdot B$ (або $A \cap B$)** випадкових подій A, B називають таку випадкову подію, яка полягає у появі подій A та B одночасно.

Якщо A та B — несумісні, то добуток $A \cdot B$ є множина, яка не має жодного елемента. Така множина називається порожньою і позначається \emptyset .

Таким чином, у разі несумісності подій A, B маємо

$$A \cdot B = A \cap B = \emptyset.$$

➤ **Добутком (перетином) скінченної кількості випадкових подій** A_1, A_2, \dots, A_n , називають таку випадкову подію, яка полягає у сумісній появі усіх цих подій одночасно.

Подія $\bigcap_{k=1}^n A_k$ означає, що розглядаються усі події A_k ($k=1, 2, \dots, n$) одночасно.

Приклад. Стрілець стріляє двічі по мішені. Описати простір елементарних наслідків. Записати подію, яка полягає в тому, що:

- а) стрілець влучив у мішень принаймні один раз (подія C);
- б) стрілець влучив рівно один раз (подія D);
- в) стрілець не влучив у мішень (подія F).

Розв'язок. Позначимо

подія A — влучення при першому пострілі,

подія B — влучення при другому пострілі.

Простір елементарних наслідків складається з чотирьох подій

$\{A B, A \bar{B}, \bar{A} B, \bar{A} \bar{B}\}$.

а) Якщо стрілець влучив у мішень принаймні один раз, то це означає, що він влучив або при першому пострілі $A \bar{B}$, або при другому пострілі $\bar{A} B$, або при обох $A B$.

$\{A B, A \bar{B}, \bar{A} B, \bar{A} \bar{B}\}$

Тобто, $C = A B \cup \bar{A} B \cup A \bar{B}$.

б) Рівно одне влучення може бути тільки тоді, коли стрілець при першому пострілі влучив, а при другому — ні, або при першому пострілі не влучив, а при другому — влучив.

$$\{A B, A \bar{B}, \bar{A} B, \bar{A} \bar{B}\}$$

Тому, $D = A \bar{B} \cup \bar{A} B.$

в) Якщо стрілець не влучив у мішень, то це означає, що він не влучив при обох пострілах,

$$\{A B, A \bar{B}, \bar{A} B, \bar{A} \bar{B}\}$$

Тобто, $F = \bar{A} \bar{B}.$

Підсумки

- ✓ Теорія імовірностей вивчає математичні моделі експериментів з випадковими результатами.
- ✓ Будь-який результат інтерпретується як випадкова подія, яка може відбутися або не відбутися при проведенні експерименту.
- ✓ Випадкові події можна порівнювати між собою за певною мірою можливості їх появи.

1.2. Поняття імовірності. Імовірний простір.

Для порівняння випадкових подій за степенем їх можливості треба кожну подію пов'язати з певним числом, яке повинно бути тим більше, чим більш можлива подія. Таке число P називають *імовірністю події*. Існує декілька означень імовірності.

- *Імовірність події* є чисельна міра степені об'єктивної можливості цієї події.

Це означення імовірності визначає філософську суть імовірності, але не вказує як знаходити імовірність будь-якої події.

Класичне визначення імовірності.

- (класичне) *Імовірність події* A дорівнює відношенню числа елементарних наслідків, які сприяють появі події A , до загального числа усіх несумісних та рівноможливих елементарних наслідків, що утворюють повну групу.

Імовірність події A позначають $P(A)$. За означенням

$$P(A) = m / n$$

де m — число елементарних наслідків, що сприяють події A ,

n — число усіх елементарних та рівноможливих наслідків.

Приклад. В урні 6 однакових за розміром куль: 2 червоні, 3 сині, 1 біла. Знайти імовірність появи червоної кулі, якщо беруть одну кулю з урни навмання.

Розв'язок.

Нехай подія A — навмання взята червона куля.

З урни можна взяти будь-яку кулю із шести, тому усіх можливих наслідків 6 ($n = 6$).

Для появи червоної кулі сприяти будуть лише 2 кулі, тому $m = 2$.

За формулою класичного визначення імовірності одержуємо

$$P(A) = 2/6 = 1/3.$$

Приклад. Кинута два гральних кубики. Знайти імовірність того, що сума очок, що випала, дорівнює 4.

Розв'язок. Нехай подія A — сума очок, що випала, дорівнює 4. Тоді протилежна подія — сума очок, що випала, не дорівнює 4. Події A сприяє один наслідок, $m = 1$; загальна кількість наслідків $n = 2$. За формулою класичного визначення імовірності одержуємо

$$P(A) = 1/2.$$

Чи правильний розв'язок?

Правильний розв'язок. Нехай подія A — сума очок, що випала, дорівнює 4. Тоді протилежна подія — сума очок, що випала, не дорівнює 4. Але ці події не рівноможливі. Розглянемо всі можливі елементарні наслідки даного експерименту: кожне число очок, що випало на першому кубіку, може поєднуватись з усіма числами другого кубіка, - отже загальна кількість несумісних рівноможливих наслідків $n = 6 \cdot 6 = 36$. Серед цих наслідків події A сприяє три: (1,3), (3,1), (2,2), $m = 3$. За формулою класичного визначення імовірності одержуємо

$$P(A) = 3/36 = 1/12.$$

- ✓ При розв'язанні багатьох задач знаходження чисел m та n має певні труднощі, запобігти яким допомагають принципи та формули комбінаторики.

- ✓ Класичне означення імовірності має місце лише тоді, коли m та n скінчені, усі елементарні наслідки рівноможливі (саме таке становище у більшості азартних ігор, що здійснюються без шахрайства).
- ✓ Якщо множина елементарних наслідків нескінченна або елементарні наслідки не рівноможливі, то формулою класичного визначення імовірності користуватись не можна. Замість нього користуються геометричним або статистичним визначенням імовірності.

Геометричне визначення імовірності.

Якщо множина усіх елементарних наслідків нескінченна і займає деяку область G , а події A сприяє лише частина $g \in G$, то обчислення імовірності події A виконують згідно геометричного означення:

- (геометричне) **Імовірність випадкової** події A дорівнює відношенню міри g до міри G

$$P(A) = m(g) / m(G)$$

Якщо область G — проміжок, поверхня, або просторове тіло, g — частина G , тоді мірою G та g буде довжина, площа або об'єм відповідної області. Якщо G та g проміжки часу, тоді їх мірою буде час.

- ✓ У загальному випадку міру області визначають аксіомами.

Приклад. Два туристичних пароплава повинні причалити до одного причалу. Час прибуття обох пароплавів рівноможливий на протязі доби. Визначити імовірність того, що одному з пароплавів доведеться чекати звільнення причалу, якщо час стоянки першого пароплава дорівнює одній годині, а другого — двом годинам.

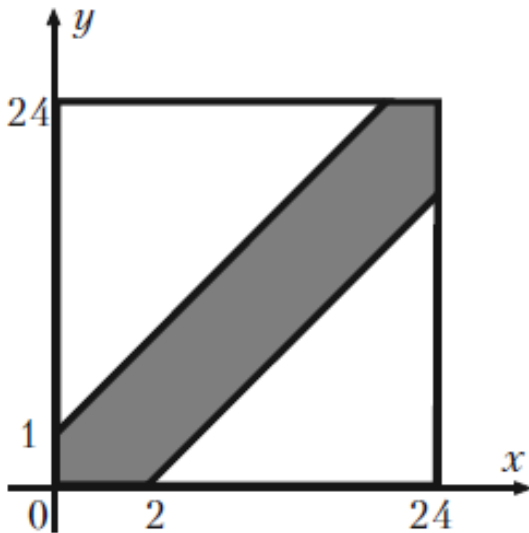
Розв'язок. Нехай X та Y — час прибуття пароплавів.

Можливі значення X та Y :

$$m(G) : 0 \leq X \leq 24 \text{ та } 0 \leq Y \leq 24.$$

Сприятливі значення

$$m(g) : Y - X \leq 1 \text{ та } X - Y \leq 2.$$



Відношення площі заштрихованої фігури $m(g)$ до площі квадрата, сторона якого дорівнює 24, згідно формули геометричного визначення імовірності дорівнює шуканій імовірності

$$P = \frac{24 \cdot 24 - 0.5 \cdot 23 \cdot 23 - 0.5 \cdot 22 \cdot 22}{24 \cdot 24} = 0.121$$

Статистичне визначення імовірності.

➤ **Відносною частотою або частотою події A** називають відношення числа випробувань, у яких подія A з'явилась, до числа фактично виконаних випробувань.

Відносну частоту події A позначають $W(A)$ або $P_n(A)$. Отже,

$$W(A) = P_n(A) = \frac{m}{n}$$

де m — кількість випробувань, у яких з'явилась подія A ,

n — кількість усіх випробувань.

Приклад. Відділ технічного контролю серед 100 виробів виявив 8 нестандартних. Чому дорівнює відносна частота появи нестандартних виробів?

Розв'язок. Позначимо через A таку подію, як поява нестандартного виробу. Тоді за означенням частоти події A одержимо

$$W(A) = \frac{8}{100} = 0.08$$

- ✓ Імовірність $P(A)$ події A обчислюється до випробування, а частоту $W(A)$ обчислюється після серії випробувань.

Частота має властивість стійкості: при великій кількості випробувань частота змінюється дуже мало, коливаючись біля деякого постійного числа — імовірності появи цієї події, тобто

- **Статистична імовірність** — це відносна частота (частота) або число, близьке до неї.

Для існування статистичної імовірності події A потрібно:

- можливість, хоча б принципова, виконати необмежену кількість випробувань, в кожному з яких подія A настає чи не настає;
- стійкість відносних частот появи A в різних серіях достатньо великої кількості випробувань.

Недоліком статистичного визначення є неоднозначність статистичної імовірності (наприклад, імовірність події можна прийняти не тільки 0.4, а і 0.41 або 0.39).

Основні властивості імовірності

що випливають з формули класичного означення імовірності і зберігаються при статистичному визначенні імовірності:

1. Якщо подія A достовірна, то її імовірність дорівнює одиниці, тобто $P(A) = 1$.
2. Якщо подія A неможлива, то її імовірність дорівнює нулеві, тобто $P(A) = 0$.
3. Якщо подія A випадкова, то її імовірність задовольняє співвідношення $0 < P(A) < 1$.

Отже, імовірність будь-якої події лежить в межах $0 \leq P(A) \leq 1$.

Дійсно, при розгляданих умовах достовірна подія обов'язково з'явиться, як наслідок, усі можливі елементарні наслідки сприяють події A , тобто $n = m$, одержимо

Якщо при умовах, що розглядаються, подія A неможлива, тоді серед усіх можливих наслідків немає тих, що сприяють події A , тобто $m = 0$, одержимо $P(A) = 0$.

Якщо подія A випадкова, то серед усіх n можливих наслідків є m наслідків, що сприяють події A , $0 < m < n$. Тому одержимо співвідношення $0 < P(A) < 1$.

- ✓ Остання властивість імовірності випадкових подій використовується для здійснення самоконтролю при розв'язанні багатьох задач теорії імовірностей.

Основні поняття та принципи комбінаторики

Часто для знаходження чисел m та n , що входять у класичне означення імовірності події, потрібно знати кількість різноманітних сполук, які можна одержати з n елементарних наслідків.

Класифікація та властивості таких сполук, а також формули для обчислення кількості різних сполук розроблені математиками і містяться у розділі «Комбінаторика» курсу алгебри.

Нехай задано скінченну непорожню множину

$A = \{a_1, a_2, \dots, a_n\}$ і виконано r таких кроків.

Крок 1. Із множини A вибирають якийсь елемент a_{i1} .

Крок 2. Із множини A чи з $A \setminus \{a_{i1}\}$ вибирають якийсь елемент a_{i2} .

...

Крок r . Якщо $a_{i1}, a_{i2}, \dots, a_{i, r-1}$ — елементи, які вибрані на перших $r - 1$ кроках ($r \geq 3$), то на цьому кроці вибирають якийсь елемент a_{ir} із множини A чи $A \setminus \{a_{i1}, a_{i2}, \dots, a_{i, r-1}\}$.

Елементи $a_{i1}, a_{i2}, \dots, a_{ir}$ утворюють **вибірку обсягом r** , або r -вибірку, із множини A .

- Вибірку називають **впорядкованою**, якщо задано порядок її елементів, а ні — то **невпорядкованою**.
- Впорядковані r -вибірки (a_1, a_2, \dots, a_r) з n -елементної множини називають **розміщеннями** з n елементів по r .
- Впорядковані n -вибірки (a_1, a_2, \dots, a_n) з n -елементної множини називаються **перестановками** n елементів.

- Невпорядковані r -вибірки з n -елементної множини — **сполученнями** з n елементів по r .

Розглянемо два способи вибору елементів.

Згідно з першим способом вибору на кожному кроці вибирають елемент з усієї множини A . Отже, один і той самий елемент із множини A може зустрітись у вибірці декілька разів. Такі вибірки називаються **вибірками з повтореннями**.

У разі застосування другого способу вибраний елемент вилучають із множини A . Це означає, що на кожному j -му кроці ($1 \leq j \leq k$) вибирають елемент із множини

$$A \setminus \bigcup_{k=1}^{j-1} \{a_{ik}\}$$

і вибірка не містить однакових елементів. Такі вибірки називають **вибірками без повторень**.

Перестановки і розміщення без повторень

Кількість всіх n -перестановок позначають через P_n .

При утворенні n -перестановок перший елемент може бути обраний n способами, оскільки існує можливість незалежного вибору з n елементів, другий — $(n - 1)$ різними способами, оскільки незалежний вибір здійснюється для решти $(n - 1)$ елемента, n -й елемент — 1 способом.

$$P_n = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 1 = n!$$

Приклад. Скільки п'ятизначних чисел можна записати, використовуючи п'ять різних цифр (крім нуля)?

Розв'язок. Вибірки, утворені з п'яти різних цифр - п'ятизначні числа, можуть відрізнятися лише порядком цифр, тому вони будуть перестановками з 5 елементів. Згідно формули їх кількість буде

$$P_5 = 5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$$

При утворенні розміщень перший елемент може бути обраний n способами, оскільки існує можливість незалежного вибору з n елементів, другий — $(n - 1)$ різними способами, оскільки незалежний вибір здійснюється для решти $(n - 1)$ елементів, r -й елемент — відповідно $(n-r+1)$ способами.

$$A_n^r = n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!}$$

- ✓ Два розміщення з n по r різні, якщо вони складаються з різних елементів або з однакових елементів, але розміщених у різному порядку.

Приклад. Студенти другого курсу згідно учбового плану вивчають 10 дисциплін. На один день можна планувати заняття з 4 дисциплін. Скількома способами можна скласти розклад занять на один день?

Розв'язок. Усі можливі розклади занять на один день — це вибірки з 10 по 4, які можуть відрізнятися дисциплінами або їх порядком, тобто ці вибірки — розміщення. Кількість таких розміщень згідно формули буде

$$A_{10}^4 = \frac{10!}{(10-4)!} = 10 \cdot 9 \cdot 8 \cdot 7 = 5040$$

Перестановки і розміщення з повтореннями

- *Розміщенням з повтореннями* з n елементів по r називається кортеж (вибірка з повторенням) довжини r з n елементів.

При утворенні розміщень з повтореннями перший елемент може бути обраний n способами, другий — теж n різними способами (повторюючи в одному з варіантів елемент, який знаходиться на першому місці), і так далі r разів.

$$\overline{A}_n^r = n \cdot n \cdot \dots \cdot n = n^r$$

Нехай є n елементів k різних типів, а число n_j ($j=1..k$) — кількість елементів j -то типу. Очевидно, що $n_1+n_2+\dots+n_k = n$. Перестановки з n елементів за такої умови називають **перестановками з повтореннями**.

Щоб знайти явний вираз для $P_n(n_1, n_2, \dots, n_k)$, візьмемо окрему перестановку та замінимо в ній усі однакові елементи різними. Тоді кількість різних перестановок, котрі можна отримати з узятій однієї перестановки, дорівнює $n_1! \cdot n_2! \cdot \dots \cdot n_k!$ Якщо зробити це для кожної перестановки, то одержимо $n!$ перестановок. Отже

$$P_n(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! n_2! \dots n_k!}$$

Сполучення без повторень

Сполученням з n елементів по k називається будь-яке k -розміщення з цих елементів, в якому враховується лише склад елементів і не враховується їх порядок.

Із неупорядкованої множини елементів $a_{i1}, a_{i2}, \dots, a_{i k}$ можна утворити $k!$ упорядкованих k -розміщень елементів. Тому кількість всіх k -сполучень з n елементів у $k!$ разів менше, ніж кількість всіх упорядкованих розміщень з n елементів по k :

$$C_n^k = \frac{A_n^k}{P_k} = \frac{n!}{(n-k)! k!}$$

Приклад. У футбольному чемпіонаті беруть участь 17 команд. За умовою, що 3 останні команди залишають вищу лігу, скільки варіантів такого завершення чемпіонату?

Оскільки нас не цікавить взаємний порядок команд, які посіли останні місця у лізі, застосуємо формулу підрахунку сполучень — кількості варіантів:

$$C_{17}^3 = \frac{17!}{(17-3)!3!} = \frac{17 \cdot 16 \cdot 15}{1 \cdot 2 \cdot 3} = 680$$

Сполучення з повтореннями

Маємо предмети n типів у такій кількості, що предметів кожного типу не менше, ніж k екземплярів. Скільки існує неупорядкованих k -вибірок предметів з можливими повтореннями?

$$\bar{C}_n^k = P(k, n-1) = \frac{(k+n-1)!}{(n-1)!k!} = C_{n+k-1}^k$$

Приклад. В магазині продаються тістечка чотирьох типів: бізе, наполеони, пісочні та еклери. Скільки існує комбінацій придбання 7 тістечок ($n=4$, $k=7$) ?

Часто доцільно використовувати такі властивості сполучень:

$$C_n^m = C_n^{n-m}$$

$$C_n^0 + C_n^1 + C_n^2 + \dots + C_n^n = 2^n$$

$$C_n^n = 1$$

$$C_n^1 = n$$

Приклад. У ящику 10 виробів, з яких 2 нестандартні. Навмання беруть 6 виробів. Яка імовірність того, що усі взяті вироби будуть стандартними?

Розв'язок. Позначимо подію A — взято 6 стандартних виробів. Згідно умови задачі, немає значення, в якому порядку беруть 6 виробів, тобто це

будуть сполучення. Тому кількість усіх можливих елементарних наслідків буде

$$n = C_{10}^6$$

Події A сприяють лише сполуки по 6 виробів з 8 стандартних у будь-якому порядку, тобто

$$m = C_8^6$$

Отже згідно класичному означенню імовірності події A маємо

$$P(A) = \frac{C_8^6}{C_{10}^6} = \frac{8!}{6! \cdot 2!} : \frac{10!}{6! \cdot 4!} = \frac{2}{15}$$

Правила суми і добутку

Приклад. Студент має вибрати тему курсової роботи зі списку, розміщеного на трьох аркушах. Аркуші містять відповідно 20, 15 і 17 тем.

З якої кількості можливих тем студент робить свій вибір? $20+15+17=52$

Приклад. З міста A у місто B вирушають 10 потягів, 5 літаків і 3 автобуси.

Скількома способами одній людині можна дістатися з A до B ? $10+5+3=18$

➤ **Правило суми.** Якщо множина A містить n елементів, а множина B містить m елементів і $A \cap B = \emptyset$, тоді множина $A \cup B$ містить $n + m$ елементів.

Приклад. Є 17 парубків і 21 дівчина. Скільки танцювальних пар вони можуть утворити?

Спочатку оберемо парубка — це можна зробити 17 способами, після цього кожний парубок обере собі партнершу (21 спосіб): $17 \cdot 21 = 357$ пар.

- **Правило добутку.** Якщо множина A містить n елементів, а множина B містить m елементів, то множина C усіх можливих пар (a_i, b_k) ($i=1, 2, \dots, n; k=1, 2, \dots, m$) містить $n \cdot m$ елементів.

Приклад. У кошику 4 яблука першого сорту та 5 яблук другого сорту. Навмання беруть 3 яблука. Знайти імовірність того, що будуть взяті 2 яблука першого сорту і 1 другого.

Розв'язок. Нехай подія A — навімання взяті 3 яблука: 2 яблука першого сорту і 1 другого.

Всього яблук 9, з них вибірок по 3 буде C_9^3 , тобто кількість усіх можливих наслідків $n=C_9^3$.

Події A будуть сприяти вибірки, утворені з трійок, елементами яких будуть 2 яблука першого сорту і 1 другого в довільному порядку. Згідно принципу добутку кількість таких пар буде $m = C_4^2 \cdot C_5^1$.

Використовуючи класичне визначення імовірності, одержимо шукану імовірність події A

$$P(A) = \frac{m}{n} = \frac{C_4^2 \cdot C_5^1}{C_9^3}$$

Підсумки

- ✓ Імовірність події є чисельна міра степені об'єктивної можливості цієї події.
- ✓ Є декілька визначень імовірності, всі вони мають свої обмеження на використання.
- ✓ Кількість наслідків випробування (сприятливих та загальну) зручно розраховувати, користуючись формулами комбінаторики.

1.3. Основні теореми теорії імовірностей. Умовні імовірності

Додавання імовірностей несумісних подій

- Теорема 1. Імовірність об'єднання двох випадкових несумісних подій дорівнює сумі їх імовірностей

$$P(A \cup B) = P(A) + P(B)$$

Доведення. Нехай число усіх можливих елементарних наслідків появи подій A та B дорівнює n ; m_1 та m_2 — числа наслідків, що сприяють подіям A та B відповідно. Тоді події $A \cup B$ будуть сприяти $m_1 + m_2$ наслідків. Отже, за класичним означенням імовірності, маємо

$$P(A \cup B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B)$$

тобто твердження теореми доведено.

- Теорема 2. Якщо випадкові події A_1, A_2, \dots, A_n попарно несумісні, то імовірність появи хоча б однієї з цих подій дорівнює сумі їх імовірностей

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Приклад. Імовірність влучення стрілкою у першу область мішені дорівнює 0.45, у другу область — 0.35, у третю — 0.15. Знайти імовірність того, що при одному пострілі стрілок влучить у першу або другу області мішені.

Позначимо:

подія A_1 — влучення у першу область мішені;

подія A_2 — влучення у другу область мішені.

$$P(A_1 \cup A_2) = ?$$

Розв'язок.

При одному пострілі події A_1 та A_2 несумісні. Тому імовірність влучення в першу або другу області мішені буде

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) = 0.45 + 0.35 = 0.8$$

➤ Теорема 3. Сума імовірностей повної групи випадкових подій дорівнює одиниці

$$P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

Доведення. Якщо випадкові події A_1, A_2, \dots, A_n утворюють повну групу, то вони попарно несумісні, а їх об'єднання буде достовірною подією. За Теоремою 2 маємо

$$P\left(\bigcup_{k=1}^n A_k\right) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Імовірність достовірної події дорівнює одиниці, тому

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = 1$$

Ліві частини двох останніх рівностей однакові, тому праві частини будуть рівними, тобто має місце перша рівність. Теорема доведена.

➤ Наслідок. Дві протилежні події A та \bar{A} утворюють повну групу, тому має місце рівність

$$P(A \cup \bar{A}) = 1$$

з якої одержуємо формулу знаходження імовірності протилежної події

$$P(\bar{A}) = 1 - P(A)$$

Приклад. Імовірність одержати повідомлення від певної особи на протязі доби дорівнює 0.25. Знайти імовірність того, що повідомлення на протязі доби від цієї особи не буде одержано.

Позначимо:

подія A — повідомлення від цієї особи на протязі доби буде одержано.

$$P(\bar{A}) = ?$$

Розв'язок.

За умовою задачі $P(A) = 0.25$.

Протилежна подія \bar{A} означає, що на протязі доби від цієї особи повідомлення не буде одержано.

$$P(A \cup \bar{A}) = 1$$

Одержимо

$$P(\bar{A}) = 1 - 0.25 = 0.75$$

- Випадкові події A та B називають **залежними**, якщо імовірність появи однієї з них залежить від появи або не появи другої події.
- Імовірність події B , обчислена при умові появи події A , називають **умовною імовірністю події B** і позначають $P(B/A)$ або $P_A(B)$
- Якщо імовірність появи однієї події не залежить від появи або не появи другої, то такі події називають **незалежними**.

Якщо події A та B незалежні, то умовна імовірність дорівнює безумовній імовірності, тобто

$$P_A(B) = P(B)$$

- Декілька подій називають **попарно незалежними**, якщо кожні дві з них незалежні.

Наприклад, події A , B , C попарно незалежні, якщо незалежні A і B , A і C , B і C .

- Декілька подій називають **незалежними в сукупності**, якщо незалежні кожні дві з них, а також незалежні кожна подія і всі можливі добутки решти подій.

Наприклад, події A , B , C незалежні в сукупності, якщо незалежні A і B , A і C , A і BC , B і C , B і AC , C і AB .

- ✓ Вимоги до незалежності в сукупності сильніші, ніж до незалежності попарно.

Приклад. В урні 10 куль: 3 білих і 7 чорних. Навмання беруть дві кулі. Нехай подія A — взята біла куля; подія B — взята чорна куля.

Якщо кулю, яку взяли першою, повертають до урни, то імовірність появи другої кулі не залежить від того, яка взята перша куля.

$$P_A(B) = 7/10$$

$$P_{\bar{A}}(B) = 7/10$$

Приклад. В урні 10 куль: 3 білих і 7 чорних. Навмання беруть дві кулі. Нехай подія A — взята біла куля; подія B — взята чорна куля.

Якщо перша куля не повертається до урни, то імовірність другої події залежить від результату першого випробування.

Якщо першою взяли білу кулю, то в урні залишилося 2 білих кулі та 7 чорних, тому

$$P_A(B) = 7/9$$

Якщо першою взяли чорну кулю, то в урні залишилося 3 білих кулі та 6 чорних куль, тому

$$P_A(B) = 6/9 = 2/3$$

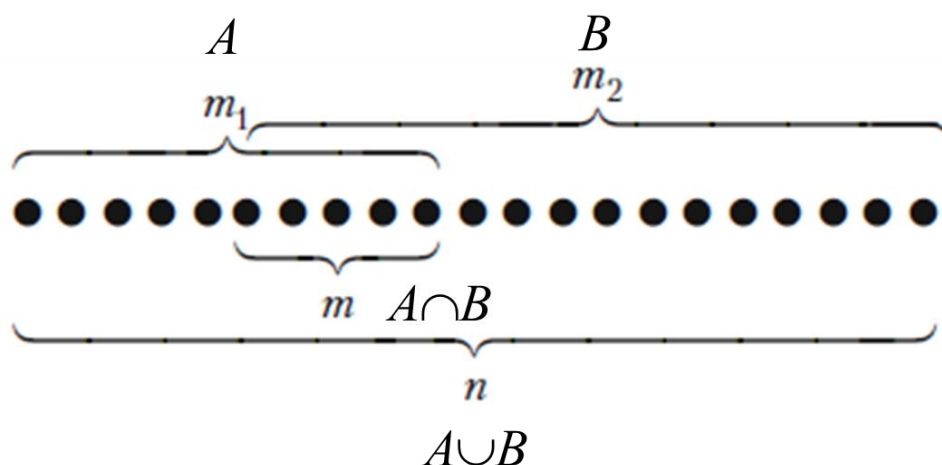
- **Теорема 4.** Імовірність сумісної появи двох випадкових подій A та B дорівнює добутку імовірностей однієї з цих подій та умовної імовірності другої події при умові, що перша подія з'явилась

$$P(A \cap B) = P(A) \cdot P_A(B) = P(B) \cdot P_B(A)$$

Співвідношення називають **формулою множення імовірностей залежних випадкових подій**.

Доведення.

Усі елементарні наслідки зобразимо у вигляді точок



Нехай появі події A сприяють m_1 наслідків, а появі події B — m_2 наслідків.

Усіх можливих наслідків n , а події $A \cap B$ будуть сприяти m наслідків. Так як

$$P(A \cap B) = \frac{m}{n}, \quad P(A) = \frac{m_1}{n}, \quad P_A(B) = \frac{m}{m_1}$$

то

$$P(A \cap B) = \frac{m}{n} = \frac{m_1}{n} \cdot \frac{m}{m_1} = P(A) \cdot P_A(B)$$

що і треба було довести.

- Наслідок. У випадку скінченної кількості залежних випадкових подій формула приймає вигляд

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P_{A_1}(A_2) \cdot P_{A_1 A_2}(A_3) \cdot \dots \cdot P_{A_1 A_2 \dots A_{n-1}}(A_n)$$

і називається **формулою множення імовірностей залежних подій**.

Імовірність кожної наступної події обчислюється за умови, що всі попередні події вже відбулися. Наприклад, для трьох подій

$$P(ABC) = P(A) \cdot P_A(B) \cdot P_{AB}(C)$$

- Наслідок. У випадку незалежних випадкових подій A та B формула приймає вигляд

$$P(A \cap B) = P(A) \cdot P(B)$$

і називається **формулою множення імовірностей незалежних подій**.

У випадку скінченної кількості незалежних в сукупності випадкових подій ця формула приймає вигляд

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_n)$$

Приклад. В урні 2 білих і 3 чорних кульки. Із урни виймають підряд дві кульки. Знайти імовірність того, що обидві кульки білі.

Позначимо через A – появу двох білих кульок.

$$P(A) = ?$$

Розв'язок.

Подія A являє собою добуток двох подій: $A = A_1 A_2$, де A_1 – поява білої кульки при першому вийманні, A_2 – поява білої кульки при другому вийманні.

Події A_1 і A_2 залежні, їх імовірності можна порахувати за класичним визначенням.

За теоремою множення імовірностей маємо:

$$P(A) = P(A_1 A_2) = P(A_1) \cdot P_{A_1}(A_2) = \frac{2}{5} \cdot \frac{1}{4} = 0.1$$

Імовірність появи хоча б однієї випадкової події

Нехай є n сумісних випадкових подій A_1, A_2, \dots, A_n . Позначимо через A подію, яка полягає в тому, що з'явиться хоча б одна з цих подій. Тоді подія \bar{A} полягає в тому, що події A_1, A_2, \dots, A_n не з'явилися

$$\bar{A} = \bar{A}_1 \cap \bar{A}_2 \cap \dots \cap \bar{A}_n = \bar{A}_1 \bar{A}_2 \dots \bar{A}_n$$

Події A та \bar{A} утворюють повну групу подій, тому

$$P(A) + P(\bar{A}) = 1 \Rightarrow P(A) = 1 - P(\bar{A})$$

звідси одержуємо

$$P(A) = 1 - P(\bar{A}_1 \bar{A}_2 \dots \bar{A}_n)$$

Приклад. Імовірність влучення у мішень першого стрілка дорівнює 0.7, другого стрілка — 0.8, а третього стрілка — 0.9. Знайти імовірність влучення у мішень хоча б одного стрілка.

Позначимо події:

A_i — у мішень влучив i -ий стрілок;

A — у мішень влучив хоча б один стрілок.

$$P(A) = P(A_1 + A_2 + A_3) = ?$$

Розв'язок.

За умовою задачі події A_1, A_2 та A_3 незалежні в сукупності, тому події \bar{A}_1, \bar{A}_2 та \bar{A}_3 також незалежні.

Імовірність події A можна шукати різними способами:

$$P(A) = P(A_1 + A_2 + A_3)$$

або

$$P(A) = 1 - P(\overline{A_1 + A_2 + A_3}) = 1 - P(\overline{A_1} \overline{A_2} \overline{A_3})$$

Згідно формули множення імовірностей незалежних подій маємо

$$P(A) = 1 - P(\overline{A_1} \overline{A_2} \overline{A_3}) = 1 - P(\overline{A_1}) \cdot P(\overline{A_2}) \cdot P(\overline{A_3})$$

Так як

$$P(\overline{A_1}) = 1 - 0.7 = 0.3;$$

$$P(\overline{A_2}) = 1 - 0.8 = 0.2;$$

$$P(\overline{A_3}) = 1 - 0.9 = 0.1$$

то підставивши значення одержимо

$$P(A) = 1 - 0.3 \cdot 0.2 \cdot 0.1 = 1 - 0.006 = 0.994$$

Імовірність суми сумісних випадкових подій

- Теорема 5. Якщо випадкові події A та B сумісні, то імовірність їх об'єднання дорівнює сумі їх імовірностей без імовірності їх сумісної появи, тобто

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Доведення. Згідно з умовою теореми події A та B сумісні, тому $A \cup B$ з'явиться, якщо з'явиться одна з трьох несумісних подій

$$A \cap \overline{B}, \quad \overline{A} \cap B, \quad A \cap B$$

Згідно з теоремою додавання імовірностей несумісних подій одержимо

$$P(A \cup B) = P(A \cap \overline{B}) + P(\overline{A} \cap B) + P(A \cap B)$$

Подія A з'явиться, якщо з'явиться одна з двох несумісних подій $A \cap \overline{B}$ або $A \cap B$.

Згідно з теоремою додавання імовірностей несумісних подій

$$\begin{aligned}P(A) &= P(A \cap \bar{B}) + P(A \cap B) \Rightarrow \\P(A \cap \bar{B}) &= P(A) - P(A \cap B)\end{aligned}$$

Аналогічно одержимо

$$\begin{aligned}P(B) &= P(\bar{A} \cap B) + P(A \cap B) \Rightarrow \\P(\bar{A} \cap B) &= P(B) - P(A \cap B)\end{aligned}$$

Підставимо

$$\begin{aligned}P(A \cup B) &= P(A \cap \bar{B}) + P(\bar{A} \cap B) + P(A \cap B) = \\&= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) = \\&= P(A) + P(B) - P(A \cap B)\end{aligned}$$

і одержимо рівність, яку треба було довести.

✓ Якщо події A та B незалежні, то формула приймає вигляд

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$$

✓ Для залежних випадкових подій

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P_A(B)$$

Приклад. У залежності від наявності сировини підприємство може виробити та відправити замовникам щодобово кількість певної продукції від 1 до 100. Яка імовірність того, що одержану кількість продукції можна розподілити без залишку

- а) трьома замовникам;
- б) чотирма замовникам;
- в) дванадцятьма замовникам;
- г) трьома або чотирма замовникам?

Розв'язок. Позначимо події:

A – одержана кількість виробів ділиться на 3 без залишку;

B – одержана кількість виробів ділиться на 4 без залишку.

Використовуючи класичне означення імовірності, знаходимо

$$a) P(A) = \frac{33}{100}; \quad б) P(B) = \frac{25}{100}; \quad в) P(AB) = \frac{8}{100}.$$

Події A та B – сумісні, тому за формулою додавання сумісних подій одержимо

$$г) P(A \cup B) = P(A) + P(B) - P(AB) = \frac{33}{100} + \frac{25}{100} - \frac{8}{100} = \frac{1}{2}$$

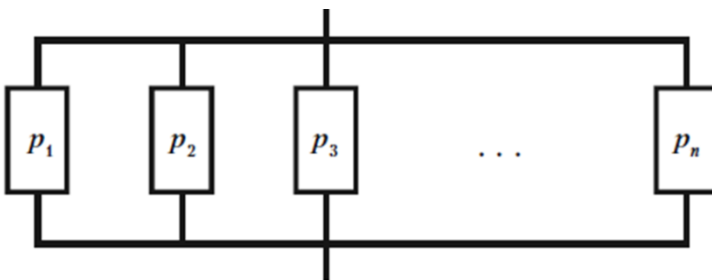
Надійність системи

➤ **Надійністю системи** називають імовірність її безвідмовної роботи в певний час t (гарантійний термін).

Системи складаються з елементів, поєднаних послідовно



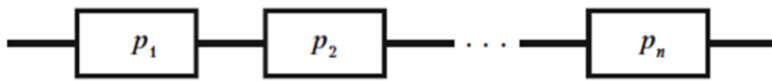
або паралельно



При обчисленні надійності систем необхідно виразити надійність системи через надійність елементів та блоків.

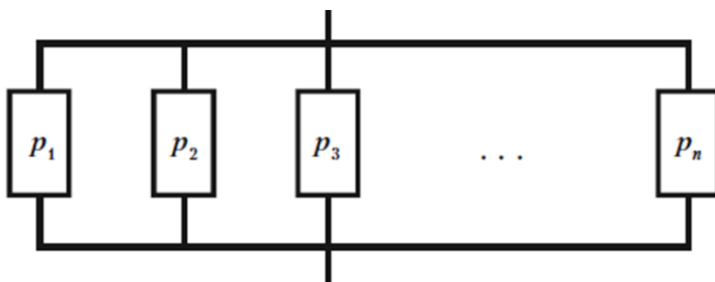
Надійність елементів вважається відомою, бо вона пов'язана з технологією їх виготовлення.

Позначимо за p_k надійність k -того елемента, q_k – імовірність виходу з строю за час t k -того елемента, P – надійність блоку.



Такий блок буде працювати безвідмовно лише в тоді, коли усі елементи працюють безвідмовно. Згідно теореми множення імовірностей незалежних подій імовірність P безвідмовної роботи такого блоку буде

$$P = p_1 \cdot p_2 \cdot \dots \cdot p_n$$



Такий блок буде працювати безвідмовно, якщо хоч один елемент не вийде зі строю. Тому імовірність P безвідмовної роботи буде

$$P = 1 - q_1 \cdot q_2 \cdot \dots \cdot q_n$$

Будь-яку складну систему можна розглядати як послідовне або паралельне з'єднання блоків.

Приклад. Прилад складено з двох блоків, що з'єднані послідовно і працюють незалежно один від одного. Імовірність відмови блоків дорівнює 0.05 та 0.08. Знайти імовірність відмови приладу.

Розв'язок. Відмова приладу є подія протилежна до його безвідмовної роботи. Імовірності безвідмовної роботи блоків будуть

$$p_1 = 1 - 0.05 = 0.95; \quad p_2 = 1 - 0.08 = 0.92.$$

Імовірність безвідмовної роботи приладу буде

$$P(A) = P(A_1 A_2) = P(A_1) \cdot P(A_2) = 0.95 \cdot 0.92 = 0.874.$$

Тому імовірність відмови приладу буде

$$P(\bar{A}) = 1 - 0.874 = 0.126.$$

Принцип практичної неможливості малоїмовірних подій

При розв'язку багатьох практичних задач доводиться мати справу з подіями, імовірність яких дуже мала, тобто близька до нуля. Чи можна вважати, що малоїмовірна подія A в одиночному випробуванні не трапиться?

Відповідь залежить, насамперед, від умов задачі.

Приклад 1. Імовірність того, що парашут не розкриється при стрибку, дорівнює 0.01. Чи можна вважати, що парашут розкривається завжди?

Приклад 2. Імовірність того, що потяг запізниться на кілька хвилин, дорівнює 0.01. Чи можна вважати, що потяг завжди приходить вчасно?

На скільки малою повинна бути імовірність події, щоб можна було вважати, що в одиночному випробуванні ця подія не може трапитись?

Для задач, різних по суті, відповіді теж різні.

- Достатньо малу імовірність, при якій (в даній конкретній задачі) подію можна вважати практично неможливою, називають **рівнем значущості**.

На практиці зазвичай приймають рівні значущості в межах від 0.01 до 0.05.

Кажуть, що якщо подія A має дуже малу імовірність $p = \alpha$, то з рівнем значущості α можна вважати, що в одиночному випробуванні подія A не трапиться.

Підсумки

- ✓ На імовірність об'єднання випадкових подій впливає сумісність цих подій.
- ✓ Імовірність об'єднання двох протилежних подій дорівнює одиниці.
- ✓ На імовірність добутку випадкових подій впливає залежність цих подій.
- ✓ Безумовна та умовна імовірності характеризують подію по-різному.

1.4. Формули повної імовірності та Байєса

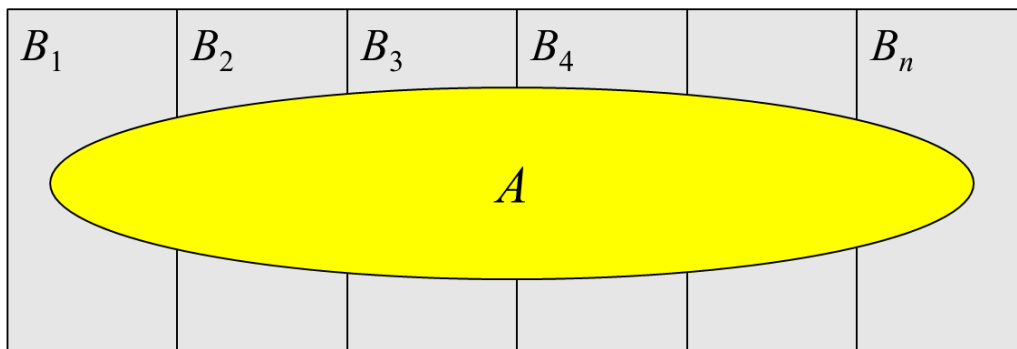
Формула повної імовірності

Нехай подія A може з'явитись лише сумісно з однією із несумісних між собою подій B_1, B_2, \dots, B_n , що утворюють повну групу.

Нехай відомі імовірності цих подій $P(B_k)$

та умовні імовірності події A $P_{B_k}(A)$

Потрібно знайти імовірність події A .



- Теорема 6. Якщо випадкова подія A може з'явитись лише сумісно з однією із несумісних між собою подій B_1, B_2, \dots, B_n , що утворюють повну групу, тоді імовірність події A обчислюється за формулою

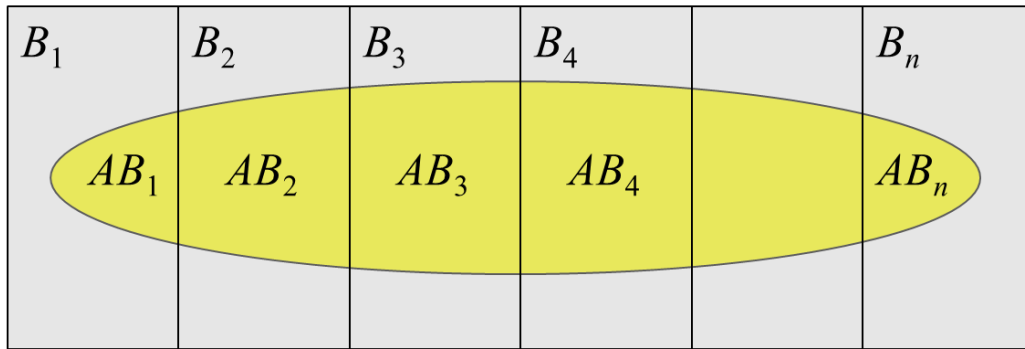
$$P(A) = \sum_{k=1}^n P(B_k)P_{B_k}(A) = \sum_{k=1}^n P(B_k)P(A/B_k)$$

- ✓ Формулу називають **формулою повної імовірності**.

Доведення. За умовою теореми поява події A означає появу однієї з подій AB_1, AB_2, \dots, AB_n , тобто

$$A = AB_1 \cup AB_2 \cup \dots \cup AB_n$$

Події B_1, B_2, \dots, B_n несумісні, тому й події AB_1, AB_2, \dots, AB_n також несумісні.



Згідно з теоремою додавання імовірностей несумісних подій маємо

$$P(A) = P(AB_1) + P(AB_2) + \dots + P(AB_n)$$

Події A та B_k – залежні, тому для обчислення $P(AB_k)$ можна використати теорему множення імовірностей залежних подій, тобто

$$P(AB_k) = P(B_k) \cdot P_{B_k}(A)$$

Підставимо це у попередню формулу і одержимо рівність, яку треба було довести.

Приклад. У першому ящику 20 деталей, з яких 15 стандартних. У другому ящику 10 деталей, з яких 9 стандартних. З другого ящика беруть навмання одну деталь і перекладають її до першого ящика. Знайти імовірність того, що взята після цього навмання деталь з першого ящика стандартна.

Позначимо події:

A – з першого ящика взято стандартну деталь;

B_1 – з другого ящика переклали до першого стандартну деталь;

B_2 – з другого ящика переклали до першого нестандартну деталь.

$$P(A) = ?$$

Розв'язок. Згідно з умовою задачі, з першого ящика можна взяти деталь лише після того, як здійсниться подія B_1 або подія B_2 . Події B_1 та B_2 несумісні, а подія A може з'явитись лише сумісно з однією із них. Тому

для знаходження імовірності події A можна використати формулу повної імовірності, яка у даному випадку прийме вигляд

$$P(A) = P(B_1)P_{B_1}(A) + P(B_2)P_{B_2}(A)$$

Знайдемо потрібні імовірності

$$P(B_1) = \frac{9}{10}; P_{B_1}(A) = \frac{16}{21}; P(B_2) = \frac{1}{10}; P_{B_2}(A) = \frac{15}{21}$$

Підставимо ці значення і одержимо

$$P(A) = \frac{9}{10} \cdot \frac{16}{21} + \frac{1}{10} \cdot \frac{15}{21} = \frac{144+15}{210} = \frac{53}{70}$$

Формула Байєса

В умовах Теорема 6 невідомо, з якою подією із несумісних подій B_1, B_2, \dots, B_n з'явиться подія A . Тому кожна з подій B_1, B_2, \dots, B_n можна вважати гіпотезою. Тоді $P(B_k)$ – імовірність k -тої гіпотези.

Якщо випробування проведено і в результаті його подія A з'явилась, то умовна імовірність $P_A(B_k)$ може не дорівнювати $P(B_k)$.

Порівняння імовірностей $P(B_k)$ та $P_A(B_k)$ дозволяє переоцінити імовірність гіпотези при умові, що подія A з'явилася.

Для одержання умовної імовірності використаємо теорему множення імовірностей залежних подій

$$P(AB_k) = P(B_k)P_{B_k}(A) = P(A)P_A(B_k) \Rightarrow$$
$$P_A(B_k) = \frac{P(B_k)P_{B_k}(A)}{P(A)}$$

Підставимо замість $P(A)$ її значення з формули повної імовірності. Одержимо

$$P_A(B_k) = \frac{P(B_k)P_{B_k}(A)}{\sum_{k=1}^n P(B_k)P_{B_k}(A)}$$

Цю формулу називають **формулою Байєса**. Вона дозволяє переоцінювати імовірності гіпотез.

Приклад. Деталі, виготовлені цехом заводу, потрапляють на перевірку їх стандартності до одного з двох контролерів. Імовірність того, що деталь потрапить до першого контролера, дорівнює 0.6, а до другого — 0.4. Імовірність того, що придатна деталь буде визнана стандартною першим контролером, дорівнює 0.94, а другим — 0.98.

Придатна деталь при перевірці визнана стандартною. Знайти імовірність того, що деталь перевіряв перший контролер.

Позначимо події:

A – придатна деталь признана стандартною;

B_1 – деталь перевіряв перший контролер;

B_2 – деталь перевіряв другий контролер.

$$P_A(B_1) = ?$$

Розв'язок. За умовою прикладу

$$P(B_1) = 0.6; P_{B_1}(A) = 0.94; P(B_2) = 0.4; P_{B_2}(A) = 0.98$$

За формулою Байєса при $k=1$ одержимо

$$P_A(B_1) = \frac{P(B_1)P_{B_1}(A)}{P(B_1)P_{B_1}(A) + P(B_2)P_{B_2}(A)} = \frac{0.6 \cdot 0.94}{0.6 \cdot 0.94 + 0.4 \cdot 0.98} = 0.59$$

✓ До появи події A імовірність $P(B_1)$ 0.6, а після появи події A імовірність перевірки деталі першим контролером $P_A(B_1)=0.59$ зменшилась.

Приклад. Імовірність знищення літака з одного пострілу для першої гармати дорівнює 0.2, а для другої гармати — 0.1. Кожна гармата робить по одному пострілу, причому було одне влучення у літак. Яка імовірність того, що влучила перша гармата?

Позначимо події:

A – знищення літака з одного пострілу першою гарматою;

B – знищення літака з одного пострілу другою гарматою;

C – одне влучення у літак.

Розв'язок.

В результаті випробування можливі такі наслідки:

$AB, A\bar{B}, \bar{A}B, \bar{A}\bar{B}$

Отже, маємо чотири гіпотези

$H_1 = AB, H_2 = A\bar{B}, H_3 = \bar{A}B, H_4 = \bar{A}\bar{B}$

які утворюють повну групу подій.

Імовірностями цих гіпотез будуть (події A і B незалежні)

$P(H_1) = 0.2 \cdot 0.1 = 0.02, P(H_2) = 0.2 \cdot 0.9 = 0.18,$

$P(H_3) = 0.8 \cdot 0.1 = 0.08, P(H_4) = 0.8 \cdot 0.9 = 0.72.$

Умовні імовірності події C будуть

$H_1 = AB, H_2 = A\bar{B}, H_3 = \bar{A}B, H_4 = \bar{A}\bar{B}$

$P_{H_1}(C) = 0, P_{H_2}(C) = 1, P_{H_3}(C) = 1, P_{H_4}(C) = 0.$

Тепер за формулою Байєса знаходимо шукану імовірність

$$P_C(H_2) = \frac{0.18 \cdot 1}{0.18 \cdot 1 + 0.08 \cdot 1} = 0.7$$

Підсумки

- ✓ Формула повної імовірності дозволяє обчислити імовірність події, що відбувається сумісно з повною групою подій-гіпотез.
- ✓ Формула Байєса дозволяє переоцінювати імовірності гіпотез.
- ✓ Імовірність гіпотези до та після випробування може відрізнитись.

Розділ 2. Теорія випадкових величин

2.1. Випадкові величини

При дослідженні багатьох проблем виникають такі випадкові події, наслідком яких є поява деякого числа, заздалегідь невідомого. Тому такі числові значення – випадкові.

Прикладом такої події є: кількість очок, що випадає при киданні грального кубика; кількість студентів, які прийдуть на лекцію; відстань, яку пролетить снаряд при пострілі з гармати, тощо.

- **Випадковою величиною** називають таку величину, яка в наслідок випробування може прийняти лише одне числове значення, заздалегідь невідоме і обумовлене випадковими причинами.

Випадкові величини доцільно позначати великими літерами X , Y , Z , а їх можливі значення – відповідними малими літерами з індексами.

Наприклад,

$$X : x_1, x_2, \dots, x_n; \quad Z : z_1, z_2, \dots, z_m.$$

Випадкові величини бувають дискретними та неперервними.

- **Дискретною випадковою величиною (ДВВ)** називають таку величину, яка може приймати відокремлені, ізольовані одне від одного числові значення (їх можна пронумерувати) з відповідними імовірностями.

Приклад. Кількість влучень у мішень при трьох пострілах буде $X : 0, 1, 2, 3$. Отже, X може приймати чотири ізольовані числові значення з різними імовірностями. Тому X – дискретна випадкова величина.

Приклад. Кількість викликів таксі Y на диспетчерському пункті також буде дискретною випадковою величиною, але при $t \rightarrow \infty$ значення Y також зростають, тобто їх кількість прямує до нескінченності $Y : 0, 1, 2, \dots, n, \dots$.

- **Неперервною випадковою величиною (НВВ)** називають величину, яка може приймати будь-яке числове значення з деякого скінченного або нескінченного інтервалу (a, b) . Кількість можливих значень такої величини є нескінченна.

Приклад. Величина похибки, яка може бути при вимірюванні відстані; час безвідмовної роботи приладу; зріст людини; розміри деталі, яку виготовляє станок-автомат.

Розглянемо випадкові величини: кількість очок, X та Y , що можуть з'явитись при киданні правильного грального кубика та неправильного (кривого) грального кубика. Їх можливі значення однакові:

$$X : 1, 2, 3, 4, 5, 6; \quad Y : 1, 2, 3, 4, 5, 6$$

Імовірність появи будь-якого значення x_k дорівнює $1/6$, однакова для усіх можливих значень X , а імовірності появи можливих значень Y будуть різними.

Отже, випадкові величини X та Y не рівні, бо при $x_k = y_k$ маємо

$$P(x_k) \neq P(y_k), \quad k=1, 2, 3, 4, 5, 6.$$

Для повної характеристики випадкової величини треба вказати не тільки усі її можливі значення, але й закон, за яким знаходять імовірності кожного значення

$$p_k = P(X = x_k) = f(x_k), \quad \text{або} \quad P(X) = f(x).$$

- **Законом розподілу випадкової величини** називають співвідношення, що встановлює зв'язок між можливими значеннями випадкової величини і відповідними їм імовірностями.

2.2. Закони розподілу імовірностей та числові характеристики дискретних випадкових величин

Способи задання закону розподілу ДВВ

Нехай випадкова дискретна величина X приймає значення x_1, x_2, \dots, x_n , з відповідними імовірностями p_1, p_2, \dots, p_n .

Задати закон розподілу такої випадкової величини – це задати рівність $p_k = P(X = x_k)$, яку можна розглядати як функцію. Тому закон розподілу X можна задати

- таблично
- графічно
- аналітично

Табличний спосіб задання ДВВ називають **рядом розподілу** і зображують у вигляді

X	x_1	x_2	\dots	x_n
$P(X)$	p_1	p_2	\dots	p_n

У першому рядку записані усі можливі значення X , а у другому рядку – відповідні імовірності, які мають властивість

$$\sum_{k=1}^n p_k = 1$$

Приклад. Умовами лотереї передбачено: один виграш – 100 гривень, два – 50 гривень, вісім – 10 гривень, дев’ятнадцять — 1 гривня. Знайти закон розподілу суми виграшу для власника одного лотерейного білету, якщо продано 1000 білетів.

Розв’язок. Будемо шукати закон розподілу суми виграшу X у вигляді ряду розподілу. Тоді

X	100	50	10	1	0
$P(X)$	0,001	0,002	0,008	0,019	0,97

де $p(0) = 1 - (0.001 + 0.002 + 0.008 + 0.019) = 1 - 0.03 = 0.97$

- ✓ Якщо випадкова дискретна величина може приймати нескінчену кількість значень, то її ряд розподілу (таблиця) буде мати нескінчену кількість елементів у кожному рядку, причому ряд

$$\sum_{k=1}^{\infty} p_k = 1$$

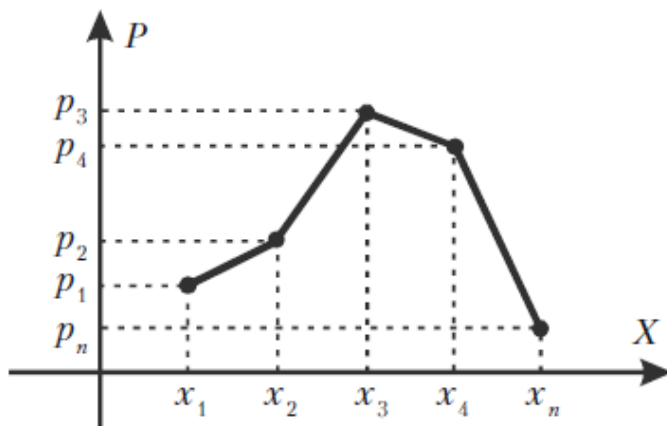
повинен бути збіжним, а його сума повинна дорівнювати одиниці.

Графічний спосіб задання ДВВ

Візьмемо прямокутну систему координат. На осі абсцис будемо відкладати можливі значення ДВВ, а на осі ординат – відповідні значення імовірності. Одержимо точки з координатами

$$(x_1, p_1), (x_2, p_2), \dots, (x_n, p_n).$$

Поєднавши ці точки прямими, одержимо графік у вигляді **многокутника розподілу випадкової дискретної величини**.



Аналітичний спосіб задання ДВВ базується на заданні певної функції, за якою можна знайти імовірність p відповідного значення x_k , тобто

$$p_k = f(x_k), \quad k = 1, 2, \dots, n$$

Приклад. Біноміальний розподіл

$$p_k = P(X = k) = C_n^k p^k (1-p)^{n-k}, \quad (k = 0, 1, 2, \dots, n)$$

Числові характеристики

Закони розподілу ДВВ повністю характеризують випадкові величини і дозволяють розв'язувати усі пов'язані з ними задачі.

Але в практичній діяльності не завжди вдається одержати закон розподілу, або закон надто складний для практичних розрахунків. Тому з'явилася потреба характеризувати ДВВ за допомогою числових характеристик, які характеризують особливості випадкових величин достатньо.

Математичне сподівання

- **Математичним сподіванням дискретної випадкової величини X** називають число, яке дорівнює сумі добутків усіх можливих значень X на відповідні їм імовірності.

Математичне сподівання ДВВ X характеризує середнє значення випадкової величини X із врахуванням імовірностей його можливих значень.

У практичній діяльності під математичним сподіванням розуміють центр розподілу випадкової величини.

Математичне сподівання ДВВ X позначають $M(X)$ або m_X

$$M(X) = \sum_{k=1}^n x_k p_k$$

Якщо X приймає нескінчену кількість значень, то

$$M(X) = \sum_{k=1}^{\infty} x_k p_k$$

Приклад. Знайти математичне сподівання випадкової величини X , якщо відомий її закон розподілу

X	3	5	2
p	0,1	0,6	0,3

Розв'язок. За визначенням математичного сподівання маємо

$$M(X) = 3 \cdot 0.1 + 5 \cdot 0.6 + 2 \cdot 0.3 = 3.9$$

Приклад. Знайти математичне сподівання випадкової величини X – числа появ події A в одному випробуванні, якщо імовірність події A дорівнює p .

Розв'язок. Випадкова величина X може прийняти тільки два значення:

$x_1=1$ (подія A з'явилася) з імовірністю p ,

$x_2=0$ (подія A не з'явилася) з імовірністю $q=1-p$.

За визначенням математичного сподівання маємо

$$M(X) = 1 \cdot p + 0 \cdot q = p$$

- ✓ Математичне сподівання числа появ події в одному випробуванні дорівнює імовірності цієї події.

Основні властивості математичного сподівання:

1. Математичне сподівання сталої величини дорівнює самій сталій

$$M(C) = C$$

2. Постійний множник можна виносити за знак математичного сподівання

$$M(CX) = C \cdot M(X)$$

Ці властивості випливають безпосередньо з визначення математичного сподівання.

3. Математичне сподівання добутку декількох взаємно незалежних дискретних випадкових величин дорівнює добутку їх математичних сподівань, тобто

$$M(X_1 \cdot X_2 \cdot \dots \cdot X_n) = M(X_1) \cdot M(X_2) \cdot \dots \cdot M(X_n)$$

Доведення. Нехай дві величини X та Y розподілені за законами

X	x_1	x_2
P	p_1	p_2

Y	y_1	y_2
G	g_1	g_2

для спрощення викладок взято лише по 2 можливих значення.

Тоді закон розподілу добутку $X \cdot Y$ буде

$X \cdot Y$	$x_1 y_1$	$x_2 y_1$	$x_1 y_2$	$x_2 y_2$
$P \cdot G$	$p_1 g_1$	$p_2 g_1$	$p_1 g_2$	$p_2 g_2$

За формулою одержимо математичне сподівання

$$\begin{aligned} M(X \cdot Y) &= y_1 g_1 (x_1 p_1 + x_2 p_2) + y_2 g_2 (x_1 p_1 + x_2 p_2) = \\ &= (x_1 p_1 + x_2 p_2) \cdot (y_1 g_1 + y_2 g_2) = M(X) \cdot M(Y) \end{aligned}$$

У випадку трьох випадкових величин маємо

$$\begin{aligned} M(X \cdot Y \cdot Z) &= M((X \cdot Y) \cdot Z) = \\ &= M(X \cdot Y) \cdot M(Z) = M(X) \cdot M(Y) \cdot M(Z) \end{aligned}$$

Методом математичної індукції тепер не важко завершити доведення.

$$M(X_1 \cdot X_2 \cdot \dots \cdot X_n) = M(X_1) \cdot M(X_2) \cdot \dots \cdot M(X_n)$$

4. Математичне сподівання суми випадкових величин дорівнює сумі їх математичних сподівань, тобто

$$M(X_1 + X_2 + \dots + X_n) = M(X_1) + M(X_2) + \dots + M(X_n)$$

Доведення аналогічне попередньому.

5. Математичне сподівання числа появ події A в n незалежних випробуваннях дорівнює добутку числа випробувань на імовірність появи події в кожному випробуванні, тобто

$$M(X) = np$$

Приклад. Незалежні випадкові величини X та Y розподілені так

X	5	2	4
P	0,6	0,1	0,3

Y	8	10
P	0,8	0,2

Знайти математичне сподівання випадкової величини $X \cdot Y$.

Розв'язок. Спочатку знайдемо математичні сподівання кожної з цих величин.

За формулою маємо

$$M(X) = 5 \cdot 0.6 + 2 \cdot 0.1 + 4 \cdot 0.3 = 4.4 ,$$

$$M(Y) = 8 \cdot 0.8 + 10 \cdot 0.2 = 8.4.$$

Випадкові величини X і Y незалежні, тому згідно властивості 3 математичного сподівання одержимо

$$M(X \cdot Y) = M(X) \cdot M(Y) = 4.4 \cdot 8.4 = 36.96$$

Приклад. Знайти математичне сподівання суми числа очок, які можуть з'явитися при киданні двох гральних кубиків.

Розв'язок.

Позначимо кількість очок, які можуть з'явитись на першому кубику X , а на другому – Y . Можливі значення цих величин 1, 2, 3, 4, 5, 6 однакові, імовірність кожного з цих значень дорівнює $1/6$. Тому

$$M(X) = M(Y) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

Згідно властивості 4 математичного сподівання, одержимо

$$M(X + Y) = M(X) + M(Y) = \frac{7}{2} + \frac{7}{2} = 7$$

Отже, математичне сподівання суми числа очок, що можуть з'явитись при киданні двох гральних кубиків, дорівнює 7.

Дисперсія

Математичне сподівання характеризує *центр розподілу* дискретної випадкової величини. Але цієї характеристики недостатньо, бо можливе значне відхилення значень випадкової величини від центру розподілу.

Наприклад, розглянемо дві випадкові величини X та Y :

X	—0,01	0,01	Y	—100	100
p	0,5	0,5	p	0,5	0,5

Їх математичні сподівання

$$M(X) = -0.01 \cdot 0.5 + 0.01 \cdot 0.5 = 0,$$

$$M(Y) = -100 \cdot 0.5 + 100 \cdot 0.5 = 0.$$

Нехай X – випадкова величина, $M(X)$ – її математичне сподівання.

Розглянемо в якості характеристики розсіювання випадкової величини різницю $X - M(X)$, яку називають **відхиленням**.

Нехай закон розподілу X відомий:

$$\begin{array}{ccccccc} X & x_1 & x_2 & \dots & x_n \\ p & p_1 & p_2 & \dots & p_n \end{array}$$

Тоді закон розподілу відхилення прийме вигляд

$$\begin{array}{ccccccc} X - M(X) & x_1 - M(X) & x_2 - M(X) & \dots & x_n - M(X) \\ p & p_1 & p_2 & \dots & p_n \end{array}$$

Математичне сподівання відхилення дорівнює нулю $M(X - M(X)) = 0$, тому часто таку випадкову величину ще називають **центрованою**.

$$\overset{\circ}{X} = X - M(X)$$

Для характеристики розсіювання можливих значень X відносно центру розподілу використовують дисперсію.

➤ **Дисперсією дискретної випадкової величини X** називають число, яке дорівнює математичному сподіванню квадрата відхилення ДВВ X від її математичного сподівання.

Дисперсію величини X позначають $D(X)$ або D_X .

$$D(X) = M((X - M(X))^2)$$

Основні властивості дисперсії:

1. Дисперсія будь-якої ДВВ X невід'ємна

$$D(X) \geq 0$$

Доведення.

Величина $(X - M(X))^2$ невід'ємна, тому згідно означення математичного сподівання та властивостей імовірностей $p_k, k = 1, 2, \dots, n$, DX також невід'ємна.

2. Дисперсія постійної величини C дорівнює нулеві

$$D(C) = 0$$

Доведення.

Якщо $X = C$, то $M(C) = C$, тому $C - M(C) = 0$.

3. Постійний множник C можна виносити за знак дисперсії, при цьому постійний множник треба піднести у квадрат

$$D(CX) = C^2 \cdot D(X)$$

Доведення.

$$CX - M(CX) = C(X - M(X))$$

тому

$$(CX - M(CX))^2 = C^2 (X - M(X))^2$$

Постійний множник C^2 можна виносити за знак математичного сподівання, тому з формули випливає потрібна рівність.

4. Дисперсія ДВВ X дорівнює різниці між математичним сподіванням квадрата випадкової величини X та квадрата її математичного сподівання

$$D(X) = M(X^2) - (M(X))^2$$

Доведення.

$$\begin{aligned} D(X) &= M((X - M(X))^2) = M(X^2 - 2 \cdot X \cdot M(X) + M^2(X)) = \\ &= M(X^2) - 2M^2(X) + M^2(X) = M(X^2) - M^2(X) \end{aligned}$$

5. Дисперсія алгебраїчної суми ДВВ X та Y дорівнює сумі їх дисперсій

$$D(X \pm Y) = D(X) + D(Y)$$

Доведення. Спочатку доведемо цю властивість для $X + Y$.

$$\begin{aligned} D(X + Y) &= M((X + Y)^2) - M^2(X + Y) = M(X^2 + 2XY + Y^2) - (M(X) + M(Y))^2 = \\ &= M(X^2) + 2M(X)M(Y) + M(Y^2) - M^2(X) - 2M(X)M(Y) - M^2(Y) = \\ &= (M(X^2) - M^2(X)) + (M(Y^2) - M^2(Y)) = D(X) + D(Y) \end{aligned}$$

Тепер розглянемо дисперсію різниці X та Y

$$D(X - Y) = D(X) + (-1)^2 \cdot D(Y) = D(X) + D(Y)$$

6. Дисперсія числа появ події A в n незалежних випробуваннях, в кожному з яких імовірність p появи події A постійна, дорівнює добутку числа випробувань на імовірності появи та неяви події в одному випробуванні

$$D(X) = npq$$

Приклад. Знайти дисперсію випадкової величини X , що задана законом

X	-5	0	4	5
P	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$

Розв'язок. Будемо шукати $D(X)$ з використанням формули

$$D(X) = M(X^2) - (M(X))^2$$

Математичним сподіванням X буде

$$M(X) = (-5) \cdot \frac{1}{8} + 0 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + 5 \cdot \frac{1}{8} = 1 \Rightarrow M^2(X) = 1$$

Щоб знайти математичне сподівання X^2 , тобто $M(X^2)$, запишемо закон розподілу X^2 у вигляді таблиці

X^2	25	0	16	25
P	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$

Усі значення X^2 отримані шляхом піднесення до квадрату відповідних значень X . Елементи другого рядка – імовірності цих значень – не змінюються.

$$M(X^2) = 25 \cdot \frac{1}{8} + 0 \cdot \frac{1}{2} + 16 \cdot \frac{1}{4} + 25 \cdot \frac{1}{8} = \frac{82}{8}$$

$$D(X) = \frac{82}{8} - 1 = \frac{74}{8} = 9.25$$

Середньоквадратичне відхилення дискретної випадкової величини

У більшості випадків випадкова величина X має розмірність, наприклад, метр, міліметр, грам, тому її дисперсія $D(X)$ буде вимірюватись у квадратних одиницях цієї розмірності.

У практичній діяльності доцільно знати величину розсіювання випадкової величини в розмірності цієї величини. Для цього використовують середньоквадратичне відхилення, яке дорівнює квадратному кореню з дисперсії і позначається

$$\sigma(X) = \sigma_X = \sqrt{D(X)}$$

Середньоквадратичне відхилення суми кінцевої кількості взаємно незалежних дискретних випадкових величин дорівнює квадратному кореню з суми дисперсій цих випадкових величин

$$\sigma(X_1 + X_2 + \dots + X_n) = \sqrt{D(X_1) + D(X_2) + \dots + D(X_n)} = \sqrt{\sigma^2(X_1) + \sigma^2(X_2) + \dots + \sigma^2(X_n)}$$

Поняття моментів розподілу

➤ *Початковим моментом порядку k* випадкової величини X називають математичне сподівання величини X^k і позначають

$$\nu_k = M(X^k), \quad k = 1, 2, \dots, n$$

➤ **Центральним моментом порядку k** випадкової величини X

називають математичне сподівання величини $(X - M(X))^k$ і

позначають

$$\mu_k = M\left((X - M(X))^k\right), \quad k = 1, 2, \dots, n$$

Відмітимо, що $\nu_1 = M(X)$, $\nu_2 = M(X^2)$, тому

$$D(X) = \nu_2 - \nu_1^2;$$

$$\mu_1 = M(X - M(X)) = 0$$

$$\mu_2 = M\left((X - M(X))^2\right) = D(X)$$

- ✓ Початкові та центральні моменти порядку $k \geq 2$ дозволяють краще враховувати вплив на математичне сподівання (центр розподілу випадкової величини X) тих можливих значень X , які сильно відрізняються від математичного сподівання та мають малу імовірність.

Приклад. Дискретна випадкова величина задана законом

X	1	2	5	100
p	0,6	0,2	0,19	0,01

Математичним сподіванням X буде

$$M(X) = 1 \cdot 0.6 + 2 \cdot 0.2 + 5 \cdot 0.19 + 100 \cdot 0.01 = 2.95$$

Законом розподілу X^2 буде

X^2	1	4	25	10000
p	0,6	0,2	0,19	0,01

Тому

$$M(X^2) = 1 \cdot 0.6 + 4 \cdot 0.2 + 25 \cdot 0.19 + 10000 \cdot 0.01 = 106.15$$

Отже, $M(X^2)$ значно більше $M(X)$, а це означає, що роль значення $X = 100$ зросла.

Підсумки

- ✓ Випадкові величини бувають дискретними та неперервними.
- ✓ Закон розподілу випадкової величини встановлює зв'язок між можливими значеннями випадкової величини і відповідними їм імовірностями.
- ✓ Числові характеристики характеризують особливості випадкових величин більш стисло порівняно з законом розподілу.

2.3. Законі розподілу імовірностей та числові характеристики неперервних випадкових величин

- **Неперервною випадковою величиною (НВВ)** називають величину, яка може приймати будь-яке числове значення з деякого скінченного або нескінченного інтервалу (a, b) . Кількість можливих значень такої величини є нескінченна.

Приклад. Величина похибки, яка може бути при вимірюванні відстані; час безвідмовної роботи приладу; зріст людини; розміри деталі, яку виготовляє станок-автомат.

- **Законом розподілу випадкової величини** називають співвідношення, що встановлює зв'язок між можливими значеннями випадкової величини і відповідними їм імовірностями.

У випадку неперервної випадкової величини неможливо скласти перелік всіх її значень.

Для можливості однотипного задання всіх різновидів випадкових величин використовують функцію розподілу.

- **Функцією розподілу (інтегральним законом розподілу)** називають імовірність того, що випадкова величина X прийме значення, менше x .

Функцію розподілу позначають $F(x)$. Таким чином,

$$F(x) = P(X < x)$$

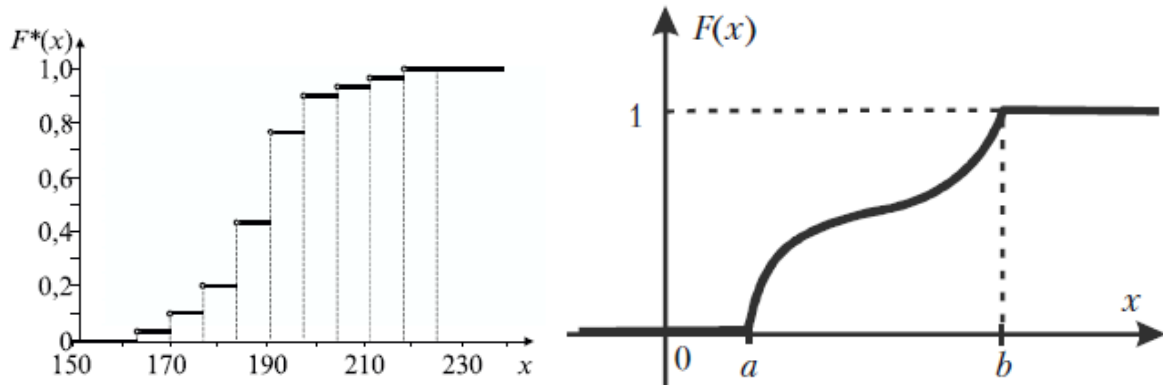
Функція розподілу для дискретної випадкової величини має вигляд

$$F(x) = P(X < x) = \sum_{x_i < x} p(x_i)$$

Властивості функції розподілу

1. $0 \leq F(x) \leq 1$;
2. $F(x)$ – зростаюча функція, тобто $F(x_1) > F(x_2)$, якщо $x_1 > x_2$;
3. $F(x) = 0$, при $x \leq a$, $F(x) = 1$, при $x \geq b$.

Графік функції розподілу $F(x)$ може мати вигляд



Якщо НВВ X може приймати будь-яке значення з $[a, b]$, то

$$P(a \leq X < b) = F(b) - F(a),$$

тобто імовірність прийняття величиною X значень з $[a, b]$ дорівнює приросту функції розподілу.

Цю формулу часто називають основною формулою теорії імовірностей.

- ✓ Неперервна випадкова величина X , що приймає значення у проміжку (a, b) , має незлічену кількість можливих значень, тому набуття X певних значень $X = a$ або $X = b$ буде майже неможливою подією. Це означає, що $P(X = a)$ та $P(X = b)$ будуть нескінченно малими величинами, які у практичних розрахунках можна не враховувати. Тому мають місце рівності

$$P(a \leq X < b) = P(a < X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

- **Щільністю розподілу імовірностей** (або **диференціальним законом розподілу**) **неперервної випадкової величини** називають похідну першого порядку від її функції розподілу

$$f(x)=F'(x)$$

тобто функція розподілу $F(x)$ є первісною для щільності розподілу $f(x)$.

Назва «щільність розподілу» випливає з рівності

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(X < x + \Delta x) - P(X < x)}{\Delta x}$$

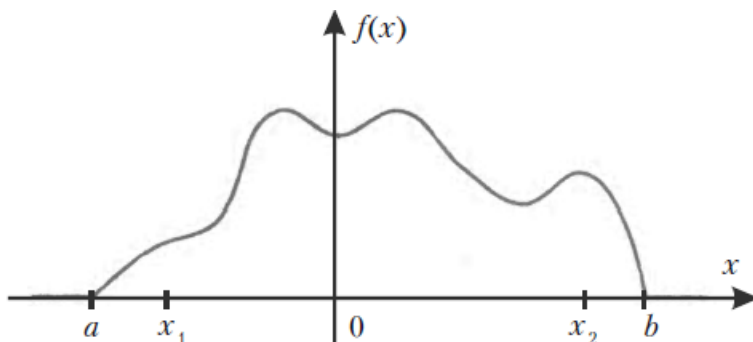
✓ Для опису дискретної випадкової величини щільність розподілу не використовують.

Властивості функції щільності розподілу:

Нехай $X \in (a, b)$

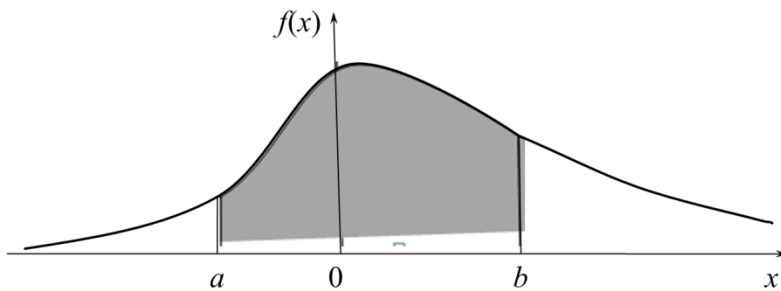
1. $f(x) \geq 0$, тому, що вона є похідною зростаючої функції $F(x)$;
2. $f(x) = 0$ при $x < a$ та $x \geq b$ тому, що є похідною $F(x) = 0$ при $x < a$ та $F(x) = 1$ при $x \geq b$.
3. $\int_{-\infty}^{\infty} f(x) dx = 1$ тому, що подія $\{-\infty < X < \infty\}$ - достовірна.

Графік щільності розподілу $f(x)$ називають **кривою розподілу**. Він може мати, наприклад, такий вигляд



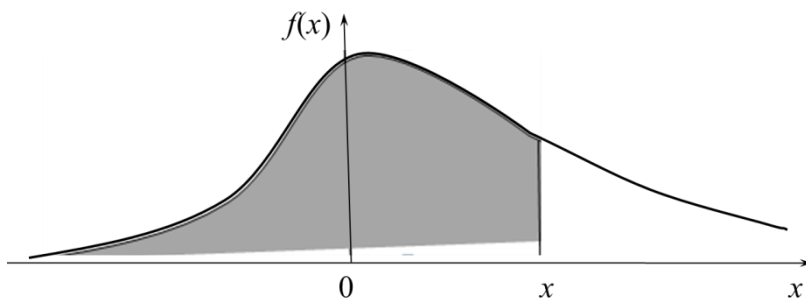
- Теорема 1. Імовірність того, що неперервна випадкова величина X прийме значення з інтервалу (a, b) , можна знайти за формулою

$$P(a < X < b) = \int_a^b f(x) dx$$



- Наслідок. Якщо щільність розподілу $f(x)$ (диференціальний закон розподілу) відома, то функцію розподілу $F(x)$ (інтегральний закон розподілу) можна знайти за формулою

$$F(x) = \int_{-\infty}^x f(x) dx$$



Приклад. Випадкова величина має щільність розподілу імовірностей

$$f(x) = \frac{a}{1+x^2}$$

Визначити параметр a та функцію розподілу.

Розв'язок.

Параметр a знайдено використовуючи властивість функції щільності розподілу

$$1 = \int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^{\infty} \frac{a}{1+x^2} dx = a \cdot \operatorname{arctg} x \Big|_{-\infty}^{\infty} =$$

$$= a \left[\lim_{b \rightarrow \infty} \operatorname{arctg} b - \lim_{c \rightarrow -\infty} \operatorname{arctg} c \right] = a \left[\frac{\pi}{2} - \left(-\frac{\pi}{2} \right) \right] = a \cdot \pi$$

Отже, одержали $1 = a \cdot \pi \Rightarrow a = \frac{1}{\pi}$

Функцію розподілу знайдемо

$$F(x) = \int_{-\infty}^x f(x) dx = \int_{-\infty}^x \frac{1}{\pi} \frac{1}{1+x^2} dx = \frac{1}{\pi} \cdot \operatorname{arctg} x \Big|_{-\infty}^x =$$

$$= \frac{1}{\pi} \left[\operatorname{arctg} x - \lim_{c \rightarrow -\infty} \operatorname{arctg} c \right] = \frac{1}{\pi} \operatorname{arctg} x - \frac{1}{\pi} \cdot \left(-\frac{\pi}{2} \right) \Rightarrow$$

$$F(x) = \frac{1}{\pi} \operatorname{arctg} x + \frac{1}{2}$$

Приклад. Випадкова величина X задана функцією розподілу $F(x) = x^2 - 4x + 4$.

Визначити область значень випадкової величини X та імовірність того, що $X \geq 2.3$.

Розв'язок.

Випадкова величина X задана функцією розподілу $F(x) = x^2 - 4x + 4$.

Згідно властивостям функції розподілу маємо

$0 \leq F(x) \leq 1$, тому повинні виконуватись умови

$$0 \leq x^2 - 4x + 4 \leq 1.$$

Якщо область значень випадкової величини $[a, b]$, то $F(a) = 0$ та $F(b) = 1$.

Підставимо в $F(x)$ замість x a та b , тоді одержимо

$$a^2 - 4a + 4 = 0 \Rightarrow (a - 2)^2 = 0 \Rightarrow a = 2$$

$$b^2 - 4b + 4 = 1 \Rightarrow b^2 - 4b + 3 = 0 \Rightarrow b_1 = 3, b_2 = 1$$

Але в проміжку $[a, b]$ $b > a$, тому $b = 3$. Отже, областю значень НВВ X буде $[2, 3]$.

Тепер знайдемо імовірність $P(X \geq 2.3)$. Подія $X < 2.3$ буде протилежною, тому

$$P(X \geq 2.3) = 1 - P(X < 2.3) = 1 - F(2.3)$$

З рівності $F(x) = x^2 - 4x + 4$ одержуємо

$$F(2.3) = (2.3)^2 - 4 \cdot 2.3 + 4 = 5.29 - 9.2 + 4 = 0.09$$

$$P(X \geq 2.3) = 1 - 0.09 = 0.91$$

Числові характеристики

У випадку неперервних випадкових величин (НВВ) математичне сподівання, дисперсія та середнє квадратичне відхилення мають такий же зміст та властивості, як і для дискретних випадкових величин, але обчислюють їх за іншими формулами.

Математичне сподівання

Нехай можливі значення неперервної випадкової величини X заповнюють відрізок $[a, b]$. Поділимо $[a, b]$ на n частин довжиною

$$\Delta x = \frac{b - a}{n}$$

У кожній частині візьмемо точку ξ_k , $k = 1, 2, \dots, n$.

Тоді щільність розподілу імовірностей $f(x)$ в точці ξ_k буде $f(\xi_k)$ – імовірність того, що X прийме значення ξ_k .

Одержимо розподіл НВВ X вигляду

X	ξ_1	ξ_2	\dots	ξ_n
P	$f(\xi_1)$	$f(\xi_2)$	\dots	$f(\xi_n)$

Сума $\sum_{k=1}^n \xi_k f(\xi_k)$ характеризує математичне сподівання X тим точніше, чим менше буде Δx . Ця сума буде дорівнювати математичному сподіванню $M(X)$ неперервної величини X , якщо перейти до границі при $\Delta x \rightarrow 0$. Згідно з означенням визначеного інтеграла маємо

$$M(X) = \lim_{\Delta x \rightarrow 0} \sum_{k=1}^n \xi_k f(\xi_k) = \int_a^b x f(x) dx$$

- Якщо неперервна випадкова величина X приймає значення на відрізку $[a, b]$ та має щільність розподілу імовірностей $f(x)$, то її **математичне сподівання** знаходиться за формулою

$$M(X) = \int_a^b x f(x) dx$$

- Якщо неперервна випадкова величина приймає значення на інтервалі $(-\infty, \infty)$ та має щільність розподілу імовірностей $f(x)$, то її математичне сподівання знаходиться за формулою

$$M(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- ✓ Якщо можливі значення X належать відрізку $[a, b]$, то центр розподілу $M(X)$ величини X знаходиться на цьому проміжку тому, що із нерівностей

$$\int_a^b a f(x) dx < \int_a^b x f(x) dx < \int_a^b b f(x) dx$$

та умови нормування

$$\int_a^b f(x) dx = 1$$

випливають співвідношення

$$a < M(X) = \int_a^b x f(x) dx < b$$

- ✓ Якщо щільність розподілу імовірностей $f(x)$ – парна функція, тобто $f(-x) = f(x)$, то центр розподілу X співпадає з початком $M(X) = 0$.
- ✓ Якщо графік функції $f(x)$ симетричний відносно прямої $x = a$, то $M(X) = a$.

Дисперсія

Як і у випадку дискретних випадкових величин, дисперсію неперервних випадкових величин X визначають так

$$D(X) = M((X - M(X))^2) = \int_{-\infty}^{\infty} (x - M(X))^2 \cdot f(x) dx$$

а обчислюють за формулою

$$D(X) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - M^2(X)$$

Якщо можливі значення X належать лише скінченному проміжку (a, b) , то формули приймають вигляд

$$D(X) = \int_a^b (x - M(X))^2 \cdot f(x) dx$$

$$D(X) = \int_a^b x^2 \cdot f(x) dx - M^2(X)$$

Середньоквадратичне відхилення неперервної випадкової величини визначають та обчислюють так

$$\sigma(X) = \sqrt{D(X)}$$

Приклад. Знайти числові характеристики випадкової величини X , яка задана функцією розподілу

$$F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^2}{25}, & 0 < x < 5 \\ 1, & x \geq 5 \end{cases}$$

Розв'язок. Випадкова величина X задана функцією розподілу

$$F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^2}{25}, & 0 < x < 5 \\ 1, & x \geq 5 \end{cases}$$

Спочатку знайдемо щільність розподілу імовірностей $f(x) = F'(x)$

$$f(x) = \begin{cases} \frac{2x}{25}, & 0 \leq x \leq 5 \\ 0, & x \notin [0, 5] \end{cases}$$

Математичне сподівання

$$M(X) = \int_0^5 x \frac{2x}{25} dx = \frac{2}{25} \cdot \frac{x^3}{3} \Big|_0^5 = \frac{10}{3}$$

Дисперсія

$$D(X) = \int_0^5 x^2 \cdot \frac{2x}{25} dx - \left(\frac{10}{3}\right)^2 = \frac{2}{25} \cdot \frac{x^4}{4} \Big|_0^5 - \frac{100}{9} = \frac{25}{18}$$

Середньоквадратичне відхилення

$$\sigma(X) = \sqrt{\frac{25}{18}} = \frac{5}{3\sqrt{2}} = \frac{5\sqrt{2}}{6} \approx 1.17$$

Підсумки

- ✓ Кількість можливих значень неперервної випадкової величини нескінчена.
- ✓ Для можливості однотипного задання всіх випадкових величин використовують функцію розподілу.
- ✓ Для задання неперервних випадкових величин можна використовувати функцію щільності розподілу.
- ✓ Числові характеристики неперервних випадкових величин такі ж, як і дискретних, але обчислюються з використанням функції щільності розподілу.

2.4. Стандартні закони розподілу імовірностей дискретних випадкових величин

Повторні випробування

У багатьох задачах теорії імовірностей, статистики та повсякденної практики треба досліджувати послідовність (серію) n випробувань.

Наприклад, випробування «кинуто 1000 однакових монет» можна розглядати як послідовність 1000 більш простих випробувань — «кинута одна монета». При киданні 1000 монет імовірність появи герба або надпису на одній монеті не залежить від того, що з'явиться на інших монетах. Тому можна казати, що у цьому випадку випробування повторюються 1000 разів незалежним чином.

- Якщо усі n випробувань проводити в однакових умовах і імовірність появи події A в усіх випробуваннях однакова та не залежить від появи або не появи A в інших випробуваннях, то таку послідовність незалежних випробувань називають *схемою Бернуллі*.

Нехай випадкова подія A може з'явитись у кожному випробуванні з імовірністю $P(A)=p$ або не з'явитись з імовірністю $q=P(\bar{A})=1-p$.

Поставимо задачу: знайти імовірність того, що при n випробуваннях подія A з'явиться m разів і не з'явиться $n - m$ разів. Шукану імовірність позначимо $P_n(m)$.

Спочатку розглянемо появу події A три рази в чотирьох випробуваннях. Можливі такі події

$AAAA, AA\bar{A}, A\bar{A}\bar{A}, \bar{A}\bar{A}\bar{A}$

тобто їх $4=C_4^3$.

Якщо подія A з'явилася 2 рази в 4 випробуваннях, то можливі такі події (їх $6=C_4^2$)

$AA\bar{A}\bar{A}$, $A\bar{A}A\bar{A}$, $A\bar{A}\bar{A}A$, $\bar{A}A\bar{A}\bar{A}$, $\bar{A}A\bar{A}A$, $\bar{A}\bar{A}AA$

У загальному випадку, коли подія A з'являється m разів у n випробуваннях, таких складних подій буде

$$C_n^m = \frac{n!}{m!(n-m)!}$$

Обчислимо імовірність однієї складної події, наприклад,

$$\underbrace{A \cdot A \cdots A}_m \cdot \underbrace{\bar{A} \cdot \bar{A} \cdots \bar{A}}_{n-m}$$

Імовірність сумісної появи n незалежних подій дорівнює добутку імовірностей цих подій згідно з теоремою множення імовірностей, тобто

$$P\left(\underbrace{(A \cdot A \cdots A)}_m \cdot \underbrace{(\bar{A} \cdot \bar{A} \cdots \bar{A})}_{n-m}\right) = P(\underbrace{A \cdot A \cdots A}_m) \cdot P(\underbrace{\bar{A} \cdot \bar{A} \cdots \bar{A}}_{n-m}) = P^m(A) \cdot P^{n-m}(\bar{A}) = p^m q^{n-m}$$

Кількість таких складних подій C_n^m і вони несумісні. Тому, згідно з теоремою додавання імовірностей несумісних подій, маємо

$$P_n(m) = C_n^m p^m q^{n-m}$$

- ✓ Формулу називають **формулою Бернуллі**. Вона дозволяє знаходити імовірність появи події A m разів при n випробуваннях, які утворюють схему Бернуллі.
- ✓ Імовірність появи події A в n випробуваннях схеми Бернуллі менше m разів знаходять за формулою
$$P_n(k < m) = P_n(0) + P_n(1) + \dots + P_n(m-1)$$
- ✓ Імовірність появи події A не менше m разів можна знайти за формулою

$$P_n(k \geq m) = P_n(m) + P_n(m+1) + \dots + P_n(n)$$

або за формулою

$$P_n(k \geq m) = 1 - \sum_{k=0}^{m-1} P_n(k)$$

- ✓ Імовірність появи події A хоча б один раз у n випробуваннях доцільно знаходити за формулою

$$P_n(1 \leq m \leq n) = 1 - q^n$$

Приклад. Прилад складено з 10 блоків, надійність кожного з них 0.8. Блоки можуть виходити з ладу незалежно один від одного. Знайти імовірність того, що

- а) відмовлять два блоки;
- б) відмовить хоча б один блок;
- в) відмовлять не менше двох блоків.

Розв'язок. Позначимо за подію A відмову блока. Тоді імовірність події A за умовою прикладу буде

$$P(A) = p = 1 - 0.8 = 0.2, \quad q = 0.8$$

Згідно з умовою задачі $n = 10$. Використовуючи формулу Бернуллі, одержимо

$$а) P_{10}(2) = C_{10}^2 p^2 q^8 = C_{10}^2 (0.2)^2 (0.8)^8 = 0.202$$

$$б) P_{10}(1 \leq m \leq 10) = 1 - P_{10}(0) = 1 - C_{10}^0 (0.2)^0 (0.8)^{10} = 0.8926$$

$$в) P_{10}(2 \leq m \leq 10) = 1 - (P_{10}(0) + P_{10}(1)) = 1 - (C_{10}^0 (0.2)^0 (0.8)^{10} + C_{10}^1 (0.2)^1 (0.8)^9) = 0.6244$$

- ✓ У багатьох випадках треба знаходити найбільш імовірне значення m_0 числа m появ події A . Це значення m визначається співвідношеннями

$$np - q \leq m_0 \leq np + p$$

або

$$(n+1)p - 1 \leq m_0 \leq (n+1)p$$

Число m_0 повинно бути цілим. Якщо $(n+1)p$ – ціле число, тоді найбільше значення імовірність має при двох числах

$$m_1 = (n+1)p - 1$$

$$m_2 = (n+1)p$$

Приклад. При новому технологічному процесі 80% усієї виготовленої продукції має найвищу якість. Знайти найбільш імовірне число виготовлених виробів найвищої якості серед 250 виготовлених виробів.

Розв'язок. Позначимо шукане число m_0 .

$$np - q \leq m_0 \leq np + p$$

За умовою прикладу $n=250$, $p=0.8$, $q=0.2$, отже

$$199.8 \leq m_0 \leq 200.8$$

Але m_0 повинно бути цілим числом, тому $m_0=200$.

- ✓ Якщо імовірність появи події A в кожному випробуванні дорівнює p , то кількість n випробувань, які необхідно здійснити, щоб з імовірністю P можна було стверджувати, що подія A з'явиться хоча б один раз:

$$P_n(m \geq 1) = 1 - q^n$$

$$q^n = 1 - P_n(m \geq 1)$$

$$n \geq \log_q(1 - P_n(m \geq 1)) = \frac{\ln(1 - P_n(m \geq 1))}{\ln(1 - p)}$$

Приклад. За одну годину автомат виготовляє 20 деталей. За скільки годин імовірність виготовлення хоча б однієї бракованої деталі буде не менше 0.952, якщо імовірність браку будь-якої деталі дорівнює 0.01?

Розв'язок.

Знайдемо спочатку таку кількість виготовлених деталей, щоб з імовірністю $P=0.952$ можна було стверджувати про наявність хоча б однієї бракованої деталі, якщо імовірність браку за умовою $p=0.01$

$$n \geq \frac{\ln(1-0.952)}{\ln(1-0.01)} \approx 300$$

Отже, за час $t=300/20=15$ (годин) автомат з імовірністю 0.952 виготовить хоча б одну браковану деталь.

Біноміальний закон розподілу

Нехай проводиться n незалежних випробувань, в кожному з яких випадкова подія A може з'явитись з імовірністю $P(A)=p$ або не з'явитись з імовірністю $q=P(\bar{A})=1-p$ (ці значення сталі для всіх випробувань).

Розглянемо в якості ДВВ X число появ події A у цих випробуваннях.

Поставимо задачу: знайти закон розподілу величини X .

Для знаходження закону розподілу величини X потрібно визначити її можливі значення та їх імовірності.

Очевидно, що подія A в n випробуваннях може або не з'явитися, або з'явитися 1 раз, або 2 рази, ..., або n разів. Таким чином, можливі значення X :

$$X : x_1 = 0, x_2 = 1, x_3 = 2, \dots, x_{n+1} = n$$

Оскільки при проведенні випробувань використовується схема Бернуллі, то імовірності кожного із значень величини X можна знайти за формулою Бернуллі.

$$p_m = P_n(m) = C_n^m p^m q^{n-m}, \quad (m = 0, 1, 2, \dots, n)$$

Біноміальний закон у вигляді таблиці

$$\begin{array}{ccccccc} X & n & n-1 & \dots & k & \dots & 0 \\ P & p^n & np^{n-1}q & \dots & C_n^k p^k q^{n-k} & \dots & q^n \end{array}$$

Граничні теореми у схемі Бернуллі

Знаходження імовірностей $P_n(m)$ та $P_n(m_1 \leq m \leq m_2)$ за формулою Бернуллі ускладнюється при досить великих значеннях n та при малих p або q . У таких випадках часто можна використовувати замість формули Бернуллі наближені асимптотичні формули.

➤ Теорема Пуассона

Якщо $n \rightarrow \infty$ і $p \rightarrow 0$ так, що $np \rightarrow \lambda$ і $0 < \lambda < \infty$

то

$$P_n(m) = C_n^m p^m q^{n-m} \rightarrow \frac{\lambda^m}{m!} e^{-\lambda}$$

для будь-якого цілого $m = 0, 1, 2, \dots$

Імовірність появи події A m разів у n випробуваннях схеми Бернуллі можна знаходити за наближеною формулою Пуассона

$$P_n(m) = \frac{\lambda^m}{m!} e^{-\lambda}$$

де $\lambda = np \leq 10$

✓ Формулу доцільно застосовувати при великих n та малих p .

Приклад. Підручник надруковано тиражем 100000 екземплярів. Імовірність невірного брошурування підручника дорівнює 0.0001. Знайти імовірність того, що тираж має 5 бракованих підручників.

Розв'язок.

Брошурування кожного підручника можна розглядати як випробування. Випробування незалежні і мають однакову імовірність невірного

брошурування, тому задача вкладається у схему Бернуллі. Згідно з умовою задачі $n = 100000$ досить велике; $p = 0.0001$ мала; $m = 5$.

Застосовуючи формулу Пуассона, одержимо

$$P_{100000}(5) = \frac{10^5}{5!} e^{-10} = 0.0375$$

Закон розподілу Пуассона

Якщо у схемі незалежних повторних випробувань n досить велике, а p або $1 - p$ прямує до нуля, то біноміальний розподіл апроксимується розподілом Пуассона, параметр якого $a = np$, причому при $p \leq 0.1$ або $p \geq 0.9$ ця апроксимація дає добрі результати незалежно від величини n .

ДВВ X приймає злічену множину значень ($m = 0, 1, 2, \dots$) з імовірностями

$$P(X = m) = \frac{a^m}{m!} e^{-a} \quad (a > 0)$$

Проста течія подій

- **Течією подій** називають послідовність таких подій, які з'являються у випадкові моменти часу.

Прикладами простої течії подій можуть бути:

- поява викликів на АТС, на пункти швидкої медичної допомоги,
- прибуття літаків до аеропорту або клієнтів у підприємство побутового обслуговування,
- серія відмов елементів або блоків приладів та інше.

- Течія подій називається **пуассонівською**, якщо вона:

1. **Стационарна**, тобто залежить від кількості k появ події та часу t і не залежить від моменту свого початку.

2. Має *властивість відсутності післядії*, тобто імовірність появи події не залежить від появи або не появи події раніше та не впливає на найближче майбутнє.
3. *Ординарна*, тобто імовірність появи більше однієї події в малий проміжок часу є величина нескінченно мала у порівнянні з імовірністю появи події один раз у цей проміжок часу.

- Середнє число λ появ події A в одиницю часу називають ***інтенсивністю течії***.
- Теорема. Якщо течія подій пуассонівська, то імовірність появи події A k разів за час t можна знайти за формулою

$$P_t(k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

де λ – інтенсивність течії.

- ✓ Цю формулу іноді звать ***математичною моделлю простої течії подій***.

Приклад. Середня кількість замовлень, що поступають до комбінату побутового обслуговування кожну годину, дорівнює 3. Знайти імовірність того, що за дві години поступлять

- а) 5 замовлень;
- б) менше 5 замовлень;
- в) не менше 5 замовлень.

Розв'язок.

Маємо просту течію подій з інтенсивністю $\lambda = 3$. За формулою одержуємо

$$a) P_2(5) = \frac{(3 \cdot 2)^5}{5!} e^{-3 \cdot 2} = \frac{6^5}{5!} e^{-6}$$

$$б) P_2(k < 5) = P_0(0) + P_0(1) + P_0(2) + P_0(3) + P_0(4) = 115e^{-6}$$

$$в) P_2(k \geq 5) = 1 - P_2(k < 5) = 1 - 115e^{-6}$$

Локальна теорема Лапласа

➤ *Локальною функцією Лапласа* називають функцію вигляду

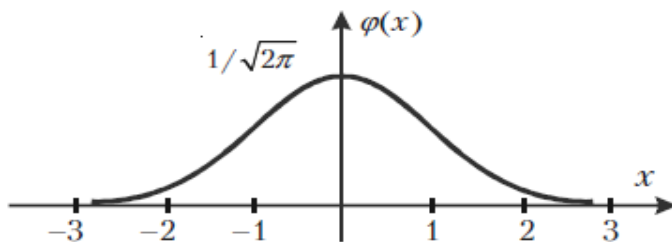
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Ця функція часто використовується, тому її значення для різних x наведені в підручниках та посібниках із теорії імовірностей. Вона табульована для додатних x .

Основні властивості локальної функції Лапласа:

1. Функція Лапласа x парна, тобто $\varphi(-x) = \varphi(x)$.
2. Функція $\varphi(x)$ визначена для усіх $x \in (-\infty, \infty)$.
3. $\varphi(x) \rightarrow 0$, коли $x \rightarrow \pm\infty$.
4. $\varphi_{\max} = \varphi(0) = \frac{1}{\sqrt{2\pi}}$

Графік локальної функції Лапласа має вигляд



➤ Теорема (локальна теорема Муавра-Лапласа). Якщо у схемі Бернуллі кількість випробувань n достатньо велика, а імовірність p появи події A в усіх випробуваннях однакова, то імовірність появи події A m разів може бути знайдена за наближеною формулою

$$P_n(m) = \frac{1}{\sqrt{npq}} \varphi(x_m)$$

де

$$x_m = \frac{m - np}{\sqrt{npq}}$$

Формулу доцільно використовувати при $n > 100$ та $npq > 10$.

Приклад. Гральний кубик кидають 800 разів. Яка імовірність того, що кількість очок, кратна трьом, з'явиться 267 разів.

Розв'язок.

У даному випадку n та m досить великі, щоб користуватися формулою біноміального закону розподілу, але імовірність появи події в одному випробуванні занадто велика, щоб можна було використати формулу Пуассона. Тому для знаходження $P_{800}(267)$ можна використати формулу Муавра-Лапласа. Маємо

$$P(A) = p = \frac{2}{6}, \quad q = 1 - \frac{1}{3} = \frac{2}{3}$$

$$x_{267} = \frac{m - np}{\sqrt{npq}} = \frac{267 - 800 \cdot \frac{1}{3}}{\frac{40}{3}} = 0.025$$

$$P_{800}(267) = \frac{3}{40} \varphi(0.025) = \frac{3}{40} \cdot 0.3988 = 0.03$$

Приклад. Імовірність влучення у мішень при одному пострілі 0.75. Знайти імовірність того, що при 10 пострілах стрілок влучить у мішень 8 разів.

Розв'язок.

У даному випадку $n=10$, $m=8$, $p=0.75$, $q=0.25$. Використаємо формулу Муавра-Лапласа для знаходження $P_{10}(8)$. Маємо

$$x_8 = \frac{m - np}{\sqrt{npq}} = \frac{8 - 10 \cdot 0.75}{\sqrt{10 \cdot 0.75 \cdot 0.25}} = 0.36$$

$$P_{10}(8) = \frac{1}{\sqrt{10 \cdot 0.75 \cdot 0.25}} \varphi(0.36) = 0.7301 \cdot 0.3739 = 0.273$$

Формула Бернуллі дає інший результат $P_{10}(8) = 0.282$. Значна похибка пояснюється недостатньо великим значенням n .

Інтегральна теорема Лапласа

➤ *Інтегральною функцією Лапласа* називають функцію вигляду

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

Інтегральна функція Лапласа $\Phi(x)$ табульована для $x \in [0, 5]$.

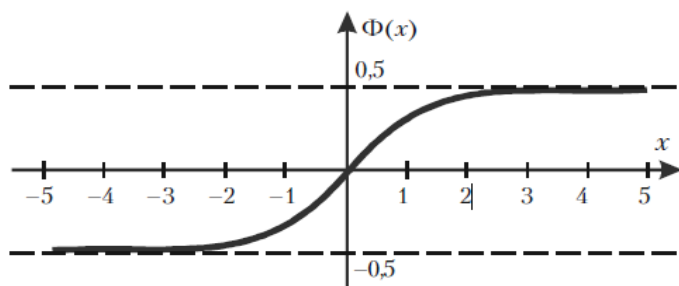
Між локальною функцією $\varphi(x)$ та інтегральною функцією $\Phi(x)$ існує простий зв'язок

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt$$

Основні властивості інтегральної функції Лапласа:

1. Інтегральна функція Лапласа є непарною функцією, тобто $\Phi(-x) = -\Phi(x)$.
2. $\Phi(0) = 0$.
3. $\Phi(x) = 0.5$ для $x \geq 5$.

Графік інтегральної функції Лапласа має вигляд



- Теорема (Інтегральна теорема Муавра-Лапласа). Якщо у схемі Бернуллі в кожному із n незалежних випробувань подія A може з'явитися з постійною імовірністю p , тоді імовірність появи події A не менш m_1 та не більш m_2 разів може бути знайдена за формулою
- $$P_n(m_1 \leq m \leq m_2) = \Phi(x_2) - \Phi(x_1)$$

де

$$x_2 = \frac{m_2 - np}{\sqrt{npq}}, \quad x_1 = \frac{m_1 - np}{\sqrt{npq}}$$

Приклад. Гральний кубик кидають 800 разів. Яка імовірність того, що кількість очок, кратна трьом, з'явиться не менше 260 та не більше 274 разів?

$$P_{800}(260 \leq m \leq 274) = ?$$

Розв'язок.

Для знаходження імовірності використаємо інтегральну теорему Муавра-Лапласа. Маємо

$$x_2 = \frac{274 - 800 \cdot \frac{1}{3}}{\frac{40}{3}} = 0.55, \quad x_1 = \frac{260 - 800 \cdot \frac{1}{3}}{\frac{40}{3}} = -0.5$$

$$\begin{aligned} P_{800}(260 \leq m \leq 274) &= \Phi(0.55) - \Phi(-0.5) = \\ &= \Phi(0.55) + \Phi(0.5) = 0.2088 + 0.1915 = 0.4003 \end{aligned}$$

Значення інтегральної функції Лапласа взято з таблиці і використана властивість непарності $\Phi(-0.5) = -\Phi(0.5)$ функції $\Phi(x)$.

Геометричний розподіл

Нехай проводяться незалежні випробування, в кожному з яких випадкова подія A може з'явитись з імовірністю $P(A)=p$ або не з'явитись з імовірністю $q=P(\bar{A})=1-p$ (ці значення стали для всіх випробувань).

Випробування припиняються, як тільки з'явиться подія A . Таким чином, якщо подія A з'явилася в m випробуванні, то в попередніх $m-1$ випробуваннях вона не з'являлась.

Розглянемо в якості ДВВ X число випробувань, які потрібно провести до появи події A у цих випробуваннях.

Можливі значення X : $x_1=1, x_2=2, \dots$

Цей розподіл має вигляд

$$P(X = m) = pq^{m-1}$$

де $p = P(A)$ – імовірність появи події A в кожному випробуванні, $q = 1 - p$, X – кількість випробувань до появи події A в серії незалежних повторних випробувань.

Ряд імовірностей цього розподілу буде нескінченно спадною геометричною прогресією із знаменником q , сума якої дорівнює одиниці.

Геометричний розподіл застосовують у різноманітних задачах статистичного контролю якості виробів, в теорії надійності, тощо.

Гіпергеометричний розподіл

Нехай маємо партію з N виробів, з яких k – стандартні ($k < N$).

Проводиться випробування, в якому з партії відбирається n виробів (кожен виріб може бути вибраний з однаковою імовірністю), причому взятий виріб не повертається у партію перед вибором наступного (тому формула Бернуллі не може бути застосована).

Розглянемо в якості ДВВ X число m стандартних виробів серед n відібраних.

Можливі значення X : $0, 1, 2, \dots, \min(k, n)$.

Цей розподіл має вигляд

$$P(X = m) = \frac{C_k^m \cdot C_{N-k}^{n-m}}{C_N^n}, \quad m = 0, 1, 2, \dots, n$$

Він вказує імовірність появи m елементів з певною властивістю серед n елементів, взятих із сукупності N елементів, яка містить k елементів саме такої властивості.

Цей розподіл використовують у багатьох задачах статистичного контролю якості.

- ✓ Якщо об'єм вибірки n малий у порівнянні з об'ємом N сукупності, тобто

$$\frac{n}{N} \leq 0.1; \quad \frac{n}{k} \leq 0.1$$

то імовірності у гіпергеометричному розподілі будуть близькими до відповідних імовірностей біноміального розподілу з $p = \frac{k}{N}$

У статистиці це означає, що розрахунки імовірностей для неповторної вибірки будуть мало відрізнятись від розрахунків імовірностей для повторної вибірки.

Поліноміальний розподіл

Цей розподіл має вигляд

$$P_n(X_1 = m_1; X_2 = m_2; \dots; X_s = m_s) = \\ = \frac{n!}{m_1! m_2! \dots m_s!} \cdot p_1^{m_1} \cdot p_2^{m_2} \cdot \dots \cdot p_s^{m_s}$$

Він застосовується тоді, коли внаслідок кожного із здійснених повторних незалежних випробувань може з'явитися s різних подій A_i з імовірністю p_i , причому

$$\sum_{i=1}^s p_i = 1$$

Послідовність випробувань із різними імовірностями

У схемі Бернуллі імовірність появи події A в усіх випробуваннях однакова.

Але у практичній діяльності іноді зустрічаються і такі випадки, коли у n незалежних випробуваннях імовірності появи події A різні, наприклад, вони дорівнюють p_1, p_2, \dots, p_n .

У цьому випадку не можна обчислювати за формулою Бернуллі імовірність появи події A m разів у n випробуваннях, а треба використовувати твірну функцію.

- У випадках, коли у n незалежних випробуваннях імовірності появи події A різні, імовірність появи події A m разів у n випробуваннях визначається через твірну функцію

$$\varphi_n(z) = \prod_{k=1}^n (p_k z + q_k)$$

Шукана імовірність $P_n(m)$ дорівнює коефіцієнту, що стоїть при z^m твірної функції.

Приклад. Імовірність відмови кожного з 4 приладів у 4 незалежних випробуваннях різні і дорівнюють

$$p_1 = 0.1, \quad p_2 = 0.2, \quad p_3 = 0.3, \quad p_4 = 0.4.$$

Знайти імовірність того, що внаслідок випробувань

- а) не відмовить жоден прилад;
- б) відмовлять один, два, три, чотири прилади;
- в) відмовить хоча б один прилад;
- г) відмовлять не менше двох приладів.

Розв'язок. Імовірності відмови приладів у випробуваннях різні, тому застосовуємо твірну функцію, яка у даному випадку матиме вигляд

$$\varphi_n(z) = (0.9 + 0.1z)(0.8 + 0.2z)(0.7 + 0.3z)(0.6 + 0.4z)$$

Розкриємо дужки та зведемо подібні члени. Тоді матимемо

$$\varphi_n(z) = 0.3024 + 0.4404z + 0.2144z^2 + 0.0404z^3 + 0.0024z^4$$

Звідси одержуємо відповіді на питання прикладу

- а) $P_4(0) = 0.3024$
- б) $P_4(1) = 0.4404$; $P_4(2) = 0.2144$; $P_4(3) = 0.0404$; $P_4(4) = 0.0024$
- в) $P_4(1 \leq m \leq 4) = 1 - P_4(0) = 0.6976$
- г) $P_4(m \geq 2) = 1 - (P_4(0) + P_4(1)) = 1 - (0.3024 + 0.4404) = 0.2372$

Приклад. Працівник обслуговує три станка, що працюють незалежно один від одного. Імовірність того, що на протязі години перший станок не вимагатиме уваги працівника, дорівнює 0.9, а для другого та третього станків — 0.8 та 0.85, відповідно. Якою є імовірність того, що на протязі години

- а) жоден станок не потребуватиме уваги працівника;

- б) усі три станки потребують уваги працівника;
в) хоча б один станок потребує уваги працівника?

Розв'язок. Цей приклад можна розв'язати з використанням теорем множення та додавання імовірностей. Розв'яжемо тепер цей приклад з використанням твірної функції, яка у даному випадку прийме вигляд

$$\begin{aligned}\varphi_n(z) &= \prod_{k=1}^3 (q_k + p_k z) = \\ &= (0.1 + 0.9z)(0.2 + 0.8z)(0.15 + 0.85z) = \\ &= 0.003 + 0.056z + 0.0329z^2 + 0.612z^3\end{aligned}$$

Отже, коефіцієнт при z^k ($k = 0, 1, 2, 3$) дорівнює імовірності того, що на протязі години уваги працівника не потребують k станків.

Одержуємо відповіді на питання прикладу:

а) імовірність того, що усі три станка не потребують уваги працівника, дорівнює коефіцієнту при z^3 , тобто

$$P_3(3) = 0.612$$

$$б) P_3(0) = 0.003$$

$$в) P_3(1 \leq m \leq 3) = 1 - P_3(0) = 1 - 0.003 = 0.997$$

Підсумки

- ✓ Схема Бернуллі – усі n випробувань проводяться в однакових умовах і імовірність появи події A в усіх випробуваннях однакова та не залежить від появи або не появи A в інших випробуваннях.
- ✓ Граничні теореми у схемі Бернуллі дозволяють спростити обчислення імовірностей при великій кількості випробувань.
- ✓ Закони розподілу, якими можна описувати ДВВ при проведенні випробувань по схемі Бернуллі:
 - біноміальний – число появ події у серії випробувань;

- Пуассона – число появ події у серії випробувань, що є простою течією подій;
- геометричний – число випробувань, які потрібно провести до появи події.
- ✓ Закони розподілу, якими можна описувати ДВВ при проведенні випробувань по схемі Бернуллі, але без повернення елементів у початкову сукупність:
 - гіпергеометричний – число m шуканих виробів серед n відібраних;
 - поліноміальний – внаслідок кожного із здійснених повторних незалежних випробувань може з'явитися s різних подій.
- ✓ У випадках, коли у n незалежних випробуваннях імовірності появи події різні, імовірність появи події m разів у n випробуваннях визначається через твірну функцію.

2.5. Стандартні закони розподілу імовірностей неперервних випадкових величин

Рівномірний розподіл

- Величина X розподілена *рівномірно на проміжку* (a, b) , якщо усі її можливі значення належать цьому проміжку і її щільність розподілу імовірностей на цьому проміжку постійна, тобто

$$f(x) = \begin{cases} C = \frac{1}{b-a}, & \text{при } x \in (a, b) \\ 0, & \text{при } x \notin (a, b) \end{cases}$$

Величина сталої $C = \frac{1}{b-a}$ визначається умовою нормування

$$P(a < X < b) = C(b-a) = 1$$

Приклад. Шкала вимірювального приладу градуйована в певних одиницях. Похибку при округленні відліку до найближчої цілої поділки можна розглядати як випадкову величину X , яка може набувати зі сталою щільністю розподілу імовірності значення між двома сусідніми поділками. Таким чином, X має рівномірний розподіл.

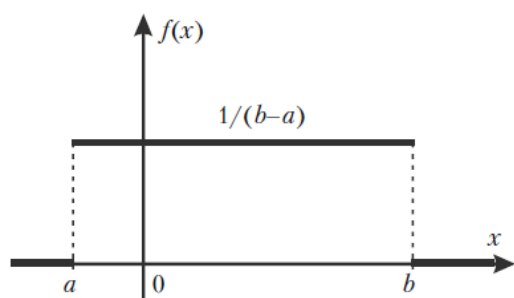
Якщо X рівномірно розподілена на проміжку (a, b) , то імовірність належності X будь-якому інтервалу $(x_1, x_2) \in (a, b)$ пропорційна довжині цього інтервалу

$$P(x_1 < X < x_2) = C(x_2 - x_1) = \frac{x_2 - x_1}{b - a}$$

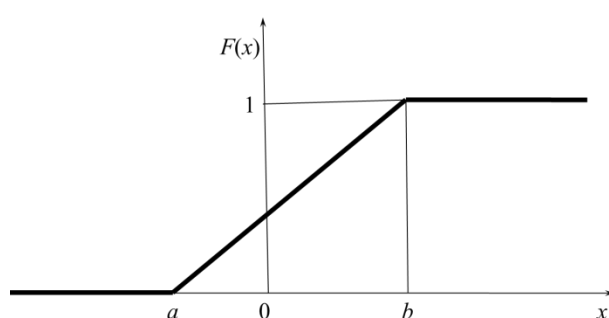
Іншими словами, імовірність влучення X в інтервал (x_1, x_2) дорівнює відношенню довжини цього інтервалу до довжини усього проміжку (a, b) .

Цей розподіл задовольняють, наприклад, похибки округлення різноманітних розрахунків.

Графік щільності рівномірного розподілу НВВ X



Графік функції рівномірного розподілу НВВ X



Числовими характеристиками НВВ X , що розподілена за рівномірним законом, будуть:

математичне сподівання

$$\begin{aligned} M(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^a x f(x) dx + \int_a^b x f(x) dx + \int_b^{\infty} x f(x) dx = \\ &= \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2} \end{aligned}$$

дисперсія

$$D(X) = \int_{-\infty}^{\infty} (x - M(X))^2 f(x) dx = \int_a^b \frac{(x - M(X))^2}{b-a} dx = \frac{\left(x - \frac{b+a}{2}\right)^3}{3(b-a)} \Big|_a^b = \frac{(b-a)^2}{12}$$

середньоквадратичне відхилення

$$\sigma(X) = \frac{(b-a)\sqrt{3}}{6}$$

Показниковий розподіл

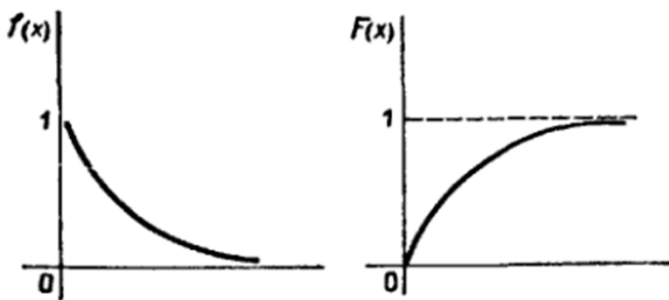
- Випадкову величину X називають **розподіленою за показниковим законом**, якщо її щільність розподілу імовірностей має вигляд

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{при } x \geq 0 \\ 0, & \text{при } x < 0 \end{cases}$$

де $\lambda > 0$ – параметр.

Показниковому розподілу задовольняють: час телефонної розмови, час ремонту техніки, час безвідмовної роботи комп'ютера.

Графік функції показникового розподілу



Числовими характеристиками показникового розподілу будуть

$$M(X) = \frac{1}{\lambda}; \quad D(X) = \frac{1}{\lambda^2}; \quad \sigma(X) = \frac{1}{\lambda}$$

Отже, якщо НВВ X розподілена за показниковим законом, то вона має однакові математичне сподівання та середнє квадратичне відхилення.

Приклад. Знайти числові характеристики випадкової величини, розподіленої за законом

$$f(x) = \begin{cases} 4e^{-4x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Розв'язок. У даному випадку випадкова величина X розподілена за показниковим законом із параметром $\lambda = 4$. Згідно з формулами маємо

$$M(X) = \sigma(X) = \frac{1}{4} = 0.25; \quad D(X) = \frac{1}{4^2} = 0.0625$$

Якщо випадкова величина X розподілена за показниковим законом, то її функція розподілу (інтегральний закон розподілу) має вигляд

$$F(x) = 1 - e^{-\lambda x}$$

Тому основна формула теорії імовірностей набуде вигляду

$$P(a < X < b) = e^{-\lambda a} - e^{-\lambda b}$$

Приклад. Величина X розподілена за законом

$$f(x) = \begin{cases} 4e^{-4x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Знайти імовірність того, що X потрапить в інтервал $(0.4; 1)$.

Розв'язок.

Випадкова величина X розподілена за показниковим законом із параметром $\lambda = 4$. Згідно з формулою маємо

$$P(0.4 < X < 1) = e^{-0.4 \cdot 4} - e^{-1 \cdot 4} = e^{-1.6} - e^{-4}$$

Часто тривалість часу безвідмовної роботи елемента має показниковий розподіл, функція розподілу якого

$$F(x) = 1 - e^{-\lambda x}$$

➤ **Функцією надійності** $R(t)$ називають функцію, що визначає імовірність безвідмовної роботи елемента за час тривалістю t :

$$R(t) = 1 - F(t) = e^{-\lambda t}$$

де λ - інтенсивність відмов.

- ✓ Імовірність безвідмовної роботи елемента на інтервалі часу тривалістю t не залежить від часу попередньої роботи до початку даного інтервалу, а залежить тільки від тривалості часу t (при заданій інтенсивності відмов).

Приклад. Час безвідмовної роботи елемента розподілений за законом

$$f(t) = 0.02e^{-0.02t}, \quad t \geq 0$$

Знайти імовірність того, що елемент працюватиме без відмов протягом 100 годин.

Розв'язок. За умовою стала інтенсивність відмов $\lambda = 0.02$. Згідно з формулою маємо

$$R(100) = e^{-0.02 \cdot 100} = e^{-2} = 0.13534$$

Шукана імовірність безвідмовної роботи елемента протягом 100 годин приблизно дорівнює 0.14.

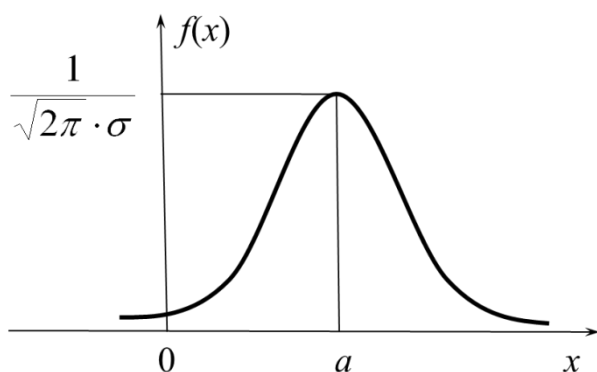
Нормальний розподіл

- Випадкову величину X називають **розподіленою нормально**, якщо її щільність розподілу імовірностей має вигляд

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}$$

де a та σ – параметри розподілу.

Графік функції $f(x)$ називають **нормальною кривою** або **кривою Гаусса**.



Властивості нормальної кривої:

1. Функція визначена на всій осі x .
2. При всіх значеннях x функція приймає додатні значення.
3. При необмеженому зростанні x значення функції прямує до нуля, тобто вісь Ox є горизонтальною асимптотою графіка.
4. Різниця $x-a$ є у аналітичному виразі в квадраті, отже графік функції симетричний відносно прямої $x=a$.
5. Перша похідна функції

$$f'(x) = -\frac{x-a}{\sigma^3 \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}$$

$$y' = 0 \text{ при } x = a$$

$$y' > 0 \text{ при } x < a$$

$$y' < 0 \text{ при } x > a$$

Отже, при $x=a$ функція має екстремум, рівний

6. Друга похідна функції

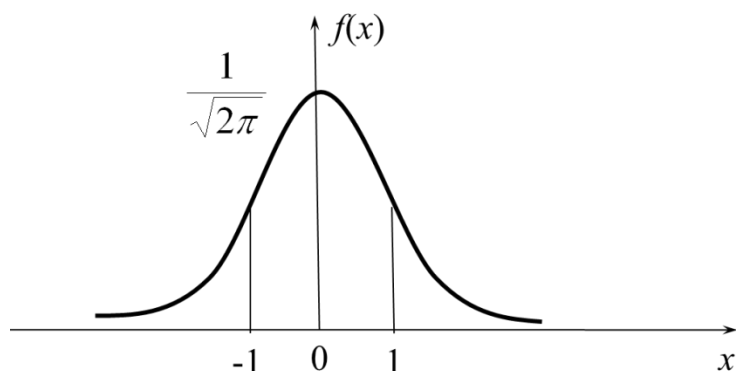
$$f''(x) = -\frac{x-a}{\sigma^3 \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}} \left(1 - \frac{(x-a)^2}{\sigma^2} \right)$$

При $x = a + \sigma$ та $x = a - \sigma$ похідна $y'' = 0$ і при переході через ці значення міняє знак. Отже, ці точки є точками перегину.

При $a = 0$ та $\sigma = 1$ нормальну криву називають **нормованою**, її рівняння буде

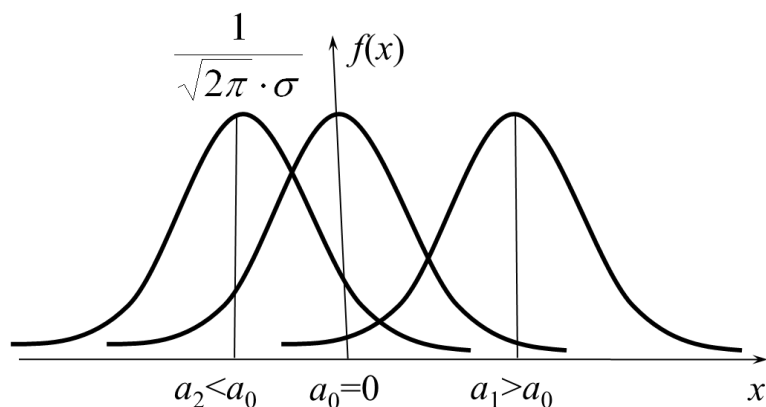
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

Тобто це є функція Лапласа, яка табульована.

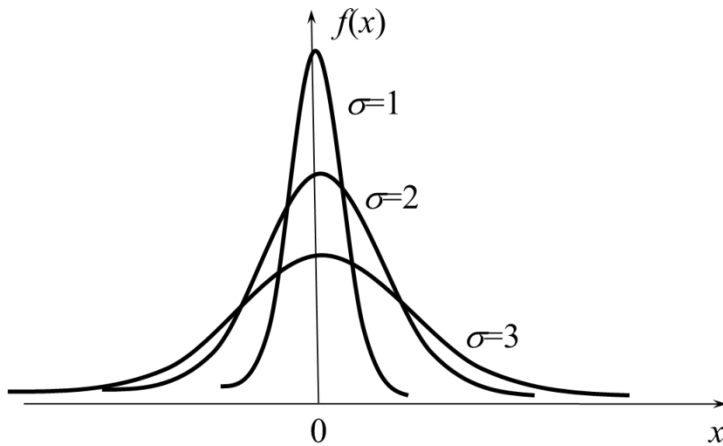


Вплив параметрів нормального розподілу на форму нормальної кривої:

Зміна величини параметра a не міняє форму кривої, а призводить до її зсуву вздовж осі Ox : вправо, якщо a збільшити, вліво – якщо a зменшити.



Зміна величини параметра σ міняє форму кривої: збільшення σ призводить до зменшення максимальної ординати кривої і вона стає більш пологою; при зменшенні σ крива розтягується вздовж Oy .



Числові характеристики нормально розподіленої НВВ X

$$M(X) = a; \quad D(X) = \sigma^2; \quad \sigma(X) = \sigma$$

Отже, математичне сподівання нормального розподілу дорівнює параметру a цього розподілу, а середнє квадратичне відхилення дорівнює параметру σ .

Якщо випадкова величина X розподілена за нормальним законом з параметрами a та σ , то випадкова величина

$$Z = \frac{X - a}{\sigma}$$

буде розподілена за нормованим нормальним законом і $M(Z) = 0$; $\sigma(Z) = 1$

Інтегральний закон розподілу нормальної НВВ

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(z-a)^2}{2\sigma^2}} dz$$

а для нормованої нормальної НВВ

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

Імовірність влучення в інтервал (c, d) нормально розподіленої випадкової величини X знаходять за формулою

$$P(c < X < d) = \Phi\left(\frac{d-a}{\sigma}\right) - \Phi\left(\frac{c-a}{\sigma}\right)$$

де функція Лапласа $\Phi(x)$ має вигляд

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

Приклад. Випадкова величина X розподілена за нормальним законом, її математичне сподівання дорівнює 30, а середньоквадратичне відхилення — 10. Знайти імовірність того, що X матиме значення з інтервалу $(10, 50)$.

$$P(10 < X < 50) = ?$$

Розв'язок.

Згідно умови $a = 30$, $\sigma = 10$, тому за формулою одержимо

$$\begin{aligned} P(10 < X < 50) &= \Phi\left(\frac{50-30}{10}\right) - \Phi\left(\frac{10-30}{10}\right) = \\ &= 2\Phi(2) = 2 \cdot 0.4772 = 0.9544 \end{aligned}$$

Тут використано властивості непарності інтегральної функції Лапласа $\Phi(-x) = -\Phi(x)$ та значення $\Phi(2)$ з таблиці значень цієї функції.

Приклад. Зріст студентів розподілено за нормальним законом.

Математичне сподівання зросту студентів дорівнює 175 см, а середньоквадратичне відхилення — 6 см. Визначити імовірність того, що хоча б один із п'яти викликаних навмання студентів буде мати зріст від 170 до 180 см.

Розв'язок.

Зріст студента X — випадкова величина, яка за умовою задачі розподілена нормально з математичним сподіванням $M(X) = 175$ см. та середньоквадратичним відхиленням $\sigma = 6$ см.

Позначимо події:

A_i — зріст i -го студента належить проміжку $(170, 180)$;

$A = A_1 + A_2 + A_3 + A_4 + A_5$ — із 5 викликаних студентів зріст хоча б одного належить проміжку $(170, 180)$;

\bar{A} — зріст усіх 5 викликаних студентів не належать проміжку $(170, 180)$.

Величина X розподілена нормально, тому імовірність того, що зріст одного викликаного студента належить проміжку $(170, 180)$

$$P(A_i) = P(170 < X < 180) = \Phi\left(\frac{180-175}{6}\right) - \Phi\left(\frac{170-175}{6}\right) = 2\Phi\left(\frac{5}{6}\right) = 2 \cdot 0.2967 = 0.5934$$

Імовірність того, що зріст одного викликаного студента не належить проміжку $(170, 180)$ буде

$$P(\bar{A}_i) = 1 - P(170 < X < 180) = 1 - 0.5934 = 0.4066$$

Застосовуючи теорему множення імовірностей незалежних подій, знайдемо імовірність події \bar{A}

$$P(\bar{A}) = P(\bar{A}_1 \bar{A}_2 \bar{A}_3 \bar{A}_4 \bar{A}_5) = (0.4066)^5 = 0.0111$$

Отже, імовірність шуканої події A буде

$$P(A) = 1 - P(\bar{A}) = 1 - 0.0111 = 0.9889 \approx 0.99$$

Обчислення імовірності заданого відхилення

Часто потрібно обчислити імовірність того, що відхилення нормально розподіленої випадкової величини за абсолютним значенням менше заданого числа δ , тобто потрібно знайти

$$P(|X - a| < \delta) = P(-\delta < X - a < \delta) = P(a - \delta < X < a + \delta) = \\ = \Phi\left(\frac{(a + \delta) - a}{\sigma}\right) - \Phi\left(\frac{(a - \delta) - a}{\sigma}\right) = \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(-\frac{\delta}{\sigma}\right) = 2\Phi\left(\frac{\delta}{\sigma}\right)$$

Приклад. Випадкова величина X розподілена за нормальним законом.

Математичне сподівання дорівнює 20, а середньоквадратичне відхилення 10. Визначити імовірність того, що відхилення випадкової величини від математичного сподівання за абсолютним значенням буде менше 3.

$$P(|X - a| < \delta) = P(|X - 20| < 3) = ?$$

Розв'язок.

За умовою задачі $a = 20$, $\sigma = 10$, $\delta = 3$.

$$P(|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right)$$

$$P(|X - 20| < 3) = 2\Phi\left(\frac{3}{10}\right) = 0.2358$$

Правило трьох сигм

Якщо випадкова величина X розподілена нормально, то

$$P(|X - a| > 3\sigma) \rightarrow 0$$

тобто імовірність того, що абсолютна величина відхилення X від її математичного сподівання більше 3σ прямує до 0, а це означає, що $|X - a| < 3\sigma$ – практично достовірна подія.

У практиці це правило використовують так:

- Якщо закон розподілу випадкової величини X невідомий, але $|X - a| < 3\sigma$ тоді можна припустити, що X розподілена нормально.

Підсумки

- ✓ Для опису неперервних випадкових величин найчастіше використовують такі закони розподілу:
 - рівномірний;
 - показниковий;
 - нормальний.
- ✓ Якщо закон розподілу випадкової величини X невідомий, але $|X - a| < 3\sigma$ тоді можна припустити, що X розподілена нормально.

2.6. Закони великих чисел. Граничні теореми

Як вже відомо, не можна заздалегідь впевнено передбачити, яке з можливих значень прийме випадкова величина у результаті випробування; це залежить від багатьох випадкових причин, врахувати які неможливо. Здавалося б, оскільки про кожну випадкову величину ми маємо в цьому сенсі вельми скромні відомості, то навряд чи можна встановити закономірності поведінки і суми достатньо великого числа випадкових величин. Насправді це не так. Виявляється, що при певних порівняно широких умовах сумарна поведінка достатньо великого числа випадкових величин майже втрачає випадковий характер і стає закономірною.

- ✓ Граничні теореми, які встановлюють відповідність між теоретичними та дослідними характеристиками випадкових подій, об'єднують загальною назвою – **закон великих чисел**.
- ✓ Граничні теореми, що встановлюють граничні закони розподілу випадкових величин, об'єднують загальною назвою – **центральна гранична теорема**.

Необхідність граничних теорем обумовлена потребою розв'язання, наприклад, таких задач:

1. Коли сума багатьох випадкових величин мало відрізняється від постійної величини, тобто майже перестає бути випадковою величиною і тому її поведінка може прогнозуватись із значною імовірністю?
2. При яких умовах можна із значною імовірністю прогнозувати число появ деякої випадкової події при великій кількості незалежних випробувань?
3. При яких обмеженнях сума багатьох випадкових величин буде розподілена за нормальним законом?

При доведенні різних граничних теорем, а також при розв'язанні різних задач важливу роль грає нерівність Чебишова, яка справедлива для дискретних та неперервних величин.

Нерівність Чебишова

Перша форма нерівності Чебишова

Для довільної випадкової величини X , яка приймає невід'ємні значення та має скінчене математичне сподівання

$$P(X \geq 1) \leq M(X)$$

Якщо X – дискретна випадкова величина, то

$$P(X \geq 1) = \sum_i p(x_i) \leq \sum_i x_i p(x_i) = M(X)$$

Якщо X – неперервна випадкова величина, $f(x)$ – щільність її імовірностей, то

$$P(X \geq 1) = \int_1^{\infty} f(x) dx \leq \int_1^{\infty} x f(x) dx \leq \int_0^{\infty} x f(x) dx = M(X)$$

Наслідок. Якщо X приймає лише невід'ємні значення, $M(X) < \infty$, $\alpha > 0$, то

$$P(X \geq \alpha) \leq \frac{M(X)}{\alpha}$$

Дійсно,

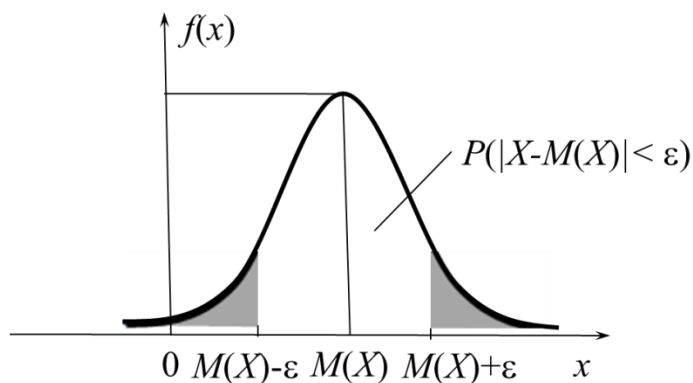
$$P(X \geq \alpha) = P\left(\frac{X}{\alpha} \geq 1\right) \leq M\left(\frac{X}{\alpha}\right) = \frac{M(X)}{\alpha}$$

Цю нерівність ще називають *нерівністю Маркова*.

Друга форма нерівності Чебишова

Якщо випадкова величина X має скінчені математичне сподівання та дисперсію, то для довільного $\varepsilon > 0$ має місце нерівність

$$P(|X - M(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}$$



Доведення. Спочатку розглянемо протилежну подію $|X - M(X)| \geq \varepsilon$

Легко бачити, що ця подія еквівалентна події $(X - M(X))^2 \geq \varepsilon^2$ тому до неї можна застосувати першу форму нерівності Чебишова

$$\begin{aligned} P(|X - M(X)| \geq \varepsilon) &= P\left(\frac{1}{\varepsilon^2} |X - M(X)|^2 \geq 1\right) \leq \\ &\leq \frac{M((X - M(X))^2)}{\varepsilon^2} = \frac{D(X)}{\varepsilon^2} \end{aligned}$$

Тепер імовірність протилежної події $|X - M(X)| < \varepsilon$ задовольняє нерівність, що треба було довести.

✓ Нерівність Чебишова дає оцінку імовірності

$$P(|X - M(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}$$

а не її точне значення. Його треба шукати за інтегральною теоремою Лапласа.

$$\begin{aligned} P(|X - MX| < \varepsilon) &= P(-\varepsilon < X - MX < \varepsilon) = P(MX - \varepsilon < X < MX + \varepsilon) = \\ &= \Phi\left(\frac{(MX + \varepsilon) - MX}{\sigma_X}\right) - \Phi\left(\frac{(MX - \varepsilon) - MX}{\sigma_X}\right) = 2\Phi\left(\frac{\varepsilon}{\sigma_X}\right) \end{aligned}$$

Приклад. Дисперсія випадкової величини X дорівнює 0.002. Оцініть імовірність того, що випадкова величина X відрізняється від її математичного сподівання $M(X)$ більше ніж на 0.1?

$$P(|X - M(X)| > 0.1) = ?$$

Розв'язок. За нерівністю Чебишова маємо

$$P(|X - M(X)| > 0.1) \leq \frac{D(X)}{0.1^2} = \frac{0.002}{0.01} = 0.2$$

Граничні теореми

Нерівності Чебишова дозволяють довести граничну теорему Бернуллі та інші важливі граничні теореми про стійкість середніх.

Теорема Бернуллі

- Якщо в кожному із n незалежних повторних випробувань імовірність p появи події A стала, то як завгодно близька до одиниці імовірність того, що відхилення відносної частоти m/n (m – число появ події A в n випробуваннях) від імовірності p буде як завгодно малим, якщо кількість випробувань досить велика. Тобто

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1, \quad \varepsilon > 0$$

Доведення. Відносну частоту m/n можна розглядати як невід'ємну випадкову величину X . Знайдемо її математичне сподівання

$$M(X) = M\left(\frac{m}{n}\right) = \frac{1}{n} M(m) = \frac{1}{n} \cdot np = p$$

Отже, необхідно оцінити імовірність відхилення випадкової величини X від її математичного сподівання. Для цього знайдемо дисперсію цієї випадкової величини

$$D(X) = D\left(\frac{m}{n}\right) = \frac{1}{n^2} D(m) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$$

За нерівністю Чебишова одержимо

$$P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right) = 1 - \frac{p(1-p)}{n \cdot \varepsilon^2}$$

Звідси граничним переходом при $n \rightarrow \infty$ одержуємо

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1, \quad \varepsilon > 0$$

що й треба було довести.

➤ Наслідок. Рівність $\left|\frac{m}{n} - p\right| = 0$ може відрізнятися від практично достовірної події $\left|\frac{m}{n} - p\right| < a, a > 0$ на нескінченно малу величину. Це

означає, що $\frac{m}{n} \rightarrow p$, тобто відносна частота $W(A) = \frac{m}{n}$ події A відрізняється від імовірності p події A на нескінченно малу величину, яку практично можна не враховувати.

✓ Формулу Бернуллі можна записати з використанням інтегральної функції Лапласа $\Phi(x)$ у вигляді

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right) = 2\Phi\left(\varepsilon \sqrt{\frac{n}{pq}}\right)$$

Звідси одержуємо формулу, яка дозволяє розв'язувати багато задач:

$$P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right) \approx 2\Phi\left(\varepsilon \sqrt{\frac{n}{pq}}\right)$$

Приклад. Імовірність того, що деталь нестандартна, $p=0.1$. Знайти імовірність того, що серед випадково відібраних 400 деталей відносна

частота появи нестандартних деталей відхилиться від імовірності за абсолютною величиною не більш ніж на 0.03.

$$P\left(\left|m/400 - 0.1\right| \leq 0.03\right) = ?$$

Розв'язок. За умовою $n=400$, $p=0.1$, $q=0.9$, $\varepsilon=0.03$.

$$P\left(\left|m/400 - 0.1\right| \leq 0.03\right) = 2 \cdot \Phi\left(0.03\sqrt{400/(0.1 \cdot 0.9)}\right) = 2 \cdot \Phi(2)$$

За таблицею $\Phi(2)=0.4772$, отже

$$P\left(\left|m/400 - 0.1\right| \leq 0.03\right) = 0.9544$$

Якщо взяти достатньо велику кількість проб деталей по 400 в кожній, то приблизно в 95.44% цих проб відхилення відносної частоти появи від імовірності $p=0.1$ за абсолютною величиною не перевищить 0.03.

Приклад. Імовірність появи події в кожному із 625 незалежних випробувань дорівнює 0.8. Знайти імовірність того, що відносна частота появи події відхиляється від імовірності за абсолютною величиною не більше ніж на 0.04.

Розв'язок. За умовою прикладу $n = 625$, $p = 0.8$, $q = 1-0.8=0.2$, $\varepsilon = 0.04$.

Треба знайти

$$P\left(\left|\frac{m}{625} - 0.8\right| \leq 0.04\right) = ?$$

За теоремою Бернуллі маємо

$$P\left(\left|\frac{m}{625} - 0.8\right| \leq 0.04\right) \approx 2\Phi\left(0.04\sqrt{\frac{625}{0.8 \cdot 0.2}}\right) = 2\Phi(2.5)$$

З таблиці значень функції Лапласа $\Phi(x)$ знаходимо $\Phi(2.5) = 0.4938$. Отже,

$$2\Phi(2.5) = 2 \cdot 0.4938 = 0.9876 \Rightarrow$$

$$P\left(\left|\frac{m}{625} - 0.8\right| \leq 0.04\right) = 0.9876$$

Приклад. Імовірність появи події в кожному із незалежних випробувань дорівнює 0.5. Знайти число випробувань n , при якому з імовірністю 0.7698 можна чекати, що відносна частота появи події відхиляється від її імовірності за абсолютною величиною не більше ніж на 0.02.

Розв'язок. За умовою $p = 0.5$, $q = 0.5$, $\varepsilon = 0.02$.

$$P\left(\left|\frac{m}{n} - 0.5\right| \leq 0.02\right) = 0.7698$$

Треба знайти n .

Застосуємо теорему Бернуллі.

$$2\Phi\left(0.02\sqrt{\frac{n}{0.5 \cdot 0.5}}\right) = 0.7698 \Rightarrow \Phi(0.04\sqrt{n}) = 0.3849$$

Із таблиці значень інтегральної функції Лапласа знайдемо

$$0.3849 = \Phi(1.2) \Rightarrow 0.04\sqrt{n} = 1.2 \Rightarrow \sqrt{n} = 30 \Rightarrow n = 900$$

Отже, шукана кількість випробувань $n = 900$.

Приклад. Відділ технічного контролю перевіряє стандартність 900 виробів. Імовірність того, що виріб стандартний, дорівнює 0.9. Знайти з імовірністю 0.9544 межі інтервалу, що містить число m стандартних виробів серед перевірених.

Розв'язок. За умовою $n = 900$, $p = 0.9$, $q = 0.1$,

$$P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right) = P\left(\left|\frac{m}{900} - 0.9\right| \leq \varepsilon\right) = 0.9544$$

Треба знайти ε та m .

За теоремою Бернуллі

$$P\left(\left|\frac{m}{900} - 0.9\right| \leq \varepsilon\right) = 0.9544 = 2\Phi\left(\varepsilon \sqrt{\frac{900}{0.9 \cdot 0.1}}\right)$$

Із таблиці значень інтегральної функції Лапласа знайдемо

$$2\Phi\left(\varepsilon \sqrt{\frac{900}{0.9 \cdot 0.1}}\right) = 0.9544 \Rightarrow \Phi(100 \cdot \varepsilon) = 0.4772$$

$$0.4772 = \Phi(2) \Rightarrow 100 \cdot \varepsilon = 2 \Rightarrow \varepsilon = 0.02$$

Отже, з імовірністю 0.9544 відхилення частоти кількості стандартних виробів від імовірності 0.9 задовольняє нерівність

$$\left|\frac{m}{900} - 0.9\right| \leq 0.02 \Rightarrow 0.88 \leq \frac{m}{900} \leq 0.92$$

З останніх співвідношень випливає, що шукане число m стандартних виробів серед 900 перевірених з імовірністю 0.9544 належить інтервалу $792 \leq m \leq 828$.

Теорема Чебишова

➤ Нехай X_1, X_2, \dots, X_n – послідовність попарно незалежних випадкових величин, які задовольняють умовам

$$1) M(X_i) = a_i \quad 2) D(X_i) \leq c \text{ для усіх } i = 1, 2, \dots, n.$$

Тоді

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{\sum_{i=1}^n X_i}{n} - \frac{\sum_{i=1}^n a_i}{n}\right| < \varepsilon\right) = 1$$

Доведення. Знайдемо математичне сподівання та дисперсію середньої випадкових величин, тобто

$$\frac{X_1 + X_2 + \dots + X_n}{n}$$

$$M\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n} M\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n M(X_i) = \frac{1}{n} \sum_{i=1}^n a_i;$$

$$D\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2} D\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \frac{c \cdot n}{n^2} = \frac{c}{n}$$

Застосуємо для випадкової величини $\frac{1}{n} \sum_{i=1}^n X_i$ нерівність Чебишова

$$P\left(\left|\frac{\sum_{i=1}^n X_i}{n} - \frac{\sum_{i=1}^n a_i}{n}\right| < \varepsilon\right) = 1 - \frac{c}{n \cdot \varepsilon^2}$$

Границя цієї імовірності при $n \rightarrow \infty$ дорівнює одиниці, тобто теорему доведено.

Зміст теореми Чебишова

Середнє арифметичне досить великого числа незалежних випадкових величин (дисперсії яких рівномірно обмежені) втрачає характер випадкової величини.

Пояснюється це тим, що відхилення кожної з величин від своїх математичних сподівань можуть бути як позитивними, так і негативними, а в середньому арифметичному вони взаємно погашаються.

Таким чином, не можна впевнено передбачити, яке можливе значення прийме кожна з випадкових величин, але можна передбачати, яке значення прийме їх середнє арифметичне.

Наслідок з теореми Чебишова

Якщо X_1, X_2, \dots, X_n – послідовність попарно незалежних випадкових величин, що мають однакове математичне сподівання та обмежену дисперсію, тобто які задовольняють умовам

$$1) M(X_i) = a \quad 2) D(X_i) \leq c$$

тоді, яким би малим не було число $\varepsilon > 0$, імовірність нерівності

$$\left| \frac{X_1 + X_2 + \dots + X_n}{n} - a \right| < \varepsilon$$

буде як завгодно близька до одиниці, якщо кількість випадкових величин досить велика.

Приклад. Скільки доданків треба взяти у теоремі Чебишова, щоб з надійністю 96% і точністю до 0.01 виконувалась наближена рівність

$$\frac{1}{n} \sum_{i=1}^n X_i \approx \frac{1}{n} \sum_{i=1}^n M(X_i)$$

Розв'язок. В цьому прикладі $\varepsilon = 0.01$. Щоб одержати надійність 96% згідно формули достатньо підібрати таке n , яке задовольняє нерівність

$$\frac{c}{\varepsilon^2 n} \leq 0.04 \Rightarrow n \geq \frac{c}{0.04 \cdot 0.0001} = 250000c$$

- ✓ Приклад показує, що навіть у випадку не дуже великих точності та надійності, треба брати значну кількість доданків (n – досить велике число). Це означає, що оцінки, одержані з використанням нерівності, – завищені. Більш точні оцінки можна одержати за допомогою теореми Ляпунова.

Центральна гранична теорема (Ляпунова)

- Якщо випадкова величина Y являє собою суму великої кількості взаємно незалежних випадкових величин X_1, X_2, \dots, X_n , вплив кожної

з яких на всю суму досить малий, то сума Y буде розподілена за законом, близьким до нормального.

Нехай задана послідовність незалежних випадкових величин $X_1, X_2, \dots, X_n, \dots$, які задовольняють умовам $M(X_i)=0, D(X_i)=b_i^2$ для усіх $i=1, 2, \dots, n, \dots$

Побудуємо суму випадкових величин

$$Y_n = \sum_{i=1}^n X_i$$

Позначимо сумарну дисперсію

$$B_n^2 = \sum_{i=1}^n b_i^2$$

Якщо виконується умова рівномірної малості величин, що утворюють суму

$$\frac{1}{B_n^{2+\delta}} \cdot \sum_{i=1}^n M(X_i)^{2+\delta} \rightarrow 0 \quad \text{при } \delta > 0, n \rightarrow \infty$$

то сума Y_n буде розподіленою нормально з математичним сподіванням $M(Y_n)=0$ та дисперсією

$$D(Y_n) = B_n^2$$

Наслідок. При $n \geq 30$ розподіл суми однаково розподілених випадкових величин мало відрізняється від нормального розподілу.

Підсумки

- ✓ Граничні теореми теорії імовірностей встановлюють відповідність між теоретичними та дослідними характеристиками випадкових величин або випадкових подій при великій кількості випробувань, а також описують граничні закони розподілу.

- ✓ Нерівність Чебишова дає оцінку імовірності

$$P(|X - M(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}$$

а не її точне значення. Його треба шукати за інтегральною теоремою Лапласа.

- ✓ Середнє арифметичне досить великого числа незалежних випадкових величин втрачає характер випадкової величини.
- ✓ При $n \geq 30$ розподіл суми однаково розподілених випадкових величин мало відрізняється від нормального розподілу.

2.7. Функції випадкових величин та їх характеристики

У багатьох випадках треба розглядати дві випадкові величини X та Y . Так, наприклад, при аналізі діяльності підприємства треба враховувати кількість усіх працюючих X та кількість зроблених виробів Y . З різних причин кількість працюючих та зроблених виробів кожного дня можуть бути різними, тобто X та Y будуть випадковими величинами.

- Якщо вказано закон, за яким кожному можливому значенню випадкової величини X відповідає певне значення випадкової величини Y , то кажуть, що Y **функція** X і позначають $Y = \varphi(X)$.

Закон розподілу функції дискретного випадкового аргументу

Нехай $Y = \varphi(X)$, аргумент X – дискретна випадкова величина з можливими значеннями x_1, x_2, \dots, x_n , імовірності яких дорівнюють p_1, p_2, \dots, p_n відповідно, тобто X задана законом

X	x_1	x_2	\dots	x_n
$P(X)$	p_1	p_2	\dots	p_n

У цьому випадку Y також дискретна випадкова величина з можливими значеннями

$$y_1 = \varphi(x_1), y_2 = \varphi(x_2), \dots, y_n = \varphi(x_n).$$

Із події «величина X прийняла значення x_k » впливає подія «величина Y прийняла значення $\varphi(x_k)$ », тому імовірності можливих значень Y також дорівнюють p_1, p_2, \dots, p_n .

Це означає, що закон розподілу Y буде мати вигляд

Y	$\varphi(x_1)$	$\varphi(x_2)$	\dots	$\varphi(x_n)$
$P(y)$	p_1	p_2	\dots	p_n

- ✓ У випадку, коли різним значенням X відповідають однакові значення Y , то слід додати імовірності повторюваних значень Y .

Числові характеристики

Математичне сподівання

$$M(Y) = \sum_{k=1}^n \varphi(x_k) p_k$$

Дисперсія

$$D(Y) = M(Y^2) - M^2(Y) = \sum_{k=1}^n [\varphi(x_k)]^2 p_k - M^2(Y)$$

Середньоквадратичне відхилення

$$\sigma(Y) = \sqrt{D(Y)}$$

Початкові моменти розподілу

$$\nu_k = \sum_{i=1}^n (\varphi(x_i))^k p_i$$

Центральні моменти розподілу

$$\mu_k = \sum_{i=1}^n (\varphi(x_i) - M(Y))^k p_i$$

Приклад. Дискретна випадкова величина задана законом розподілу

X	1	3	5
P	0,2	0,5	0,3

Знайти математичне сподівання функції $Y = X^2 + 1$.

Розв'язок. Можливими значеннями Y будуть

$$y_1 = 1^2 + 1 = 2; \quad y_2 = 3^2 + 1 = 10; \quad y_3 = 5^2 + 1 = 26.$$

За формулою знаходимо математичне сподівання Y

$$M(Y) = M(X^2 + 1) = 2 \cdot 0.2 + 10 \cdot 0.5 + 26 \cdot 0.3 = 13.2.$$

Закон розподілу функції неперервного випадкового аргументу

Нехай X — неперервна випадкова величина, закон розподілу якої заданий у вигляді щільності розподілу імовірностей $f(x)$; випадкова величина $Y = \varphi(X)$.

Якщо φ — диференційовна функція, монотонна на усьому проміжку можливих значень X , то щільність розподілу $Y = \varphi(X)$ визначають так

$$g(y) = f(\psi(y)) \cdot |\psi'(y)|$$

де ψ — функція, обернена по відношенню до функції φ , ψ' — похідна першого порядку.

Якщо φ — не монотонна функція в області визначення аргументу X , то обернена функція неоднозначна і щільність розподілу $g(y)$ визначається як сума доданків, кількість яких дорівнює кількості значень оберненої функції, тобто

$$g(y) = \sum_{i=1}^k f(\psi_i(y)) \cdot |\psi'_i(y)|$$

де $\psi_i(y)$ — обернені функції при заданому y .

Приклад. Випадкова величина X розподілена за нормальним законом з нульовим математичним сподіванням. Знайти закон розподілу функції $Y = X^3$.

Розв'язок. Згідно означенню нормального розподілу неперервної випадкової величини X та умови прикладу диференціальний закон розподілу X має вигляд

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2\sigma^2}}$$

Функція $Y=X^3$ диференційовна, $Y'=3X^2 > 0$ тому вона зростає для усіх x $(-\infty, \infty)$. Отже, можна застосувати формулу для знаходження диференціального закону розподілу $g(y)$ випадкової величини Y .

У даному випадку $Y = X^3 \Rightarrow X = \sqrt[3]{Y}$, тобто $\psi(y) = \sqrt[3]{y} = y^{\frac{1}{3}}$.

Тому щільність розподілу $g(y)$ прийме вигляд

$$g(y) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{\left(y^{\frac{1}{3}}\right)^2}{2\sigma^2}} \cdot \left| \left(y^{\frac{1}{3}}\right)' \right| = \frac{e^{-\frac{y^{\frac{2}{3}}}{2\sigma^2}}}{\sigma\sqrt{2\pi} \cdot y^{\frac{2}{3}}}$$

Для знаходження математичного сподівання від $Y=\varphi(X)$ можна спочатку знайти $g(y)$ – диференціальний закон розподілу величини Y , а потім використати формулу

$$M(Y) = \int_{-\infty}^{\infty} y \cdot g(y) dy$$

Більш доцільно знаходити математичне сподівання функції неперервного випадкового аргументу $\varphi(X)$ безпосередньо за формулою

$$M(Y) = M(\varphi(X)) = \int_{-\infty}^{\infty} \varphi(x) \cdot f(x) dx$$

де $f(x)$ – щільність розподілу імовірностей величини X .

Якщо величина X може приймати значення лише в проміжку $[a, b]$, то

$$M(Y) = M(\varphi(X)) = \int_a^b \varphi(x) \cdot f(x) dx$$

Приклад. Неперервна випадкова величина X задана диференціальним законом розподілу

$$f(x) = \begin{cases} \sin x, & \text{при } x \in \left(0, \frac{\pi}{2}\right) \\ 0, & \text{при } x \notin \left(0, \frac{\pi}{2}\right) \end{cases}$$

Знайти математичне сподівання функції $Y = X^2$.

Розв'язок. У даному випадку $\varphi(X) = X^2$; $x \in (0, \pi/2)$, тому за формулою

$M(Y) = M(\varphi(X))$ одержимо

$$M(Y) = M(X^2) = \int_0^{\frac{\pi}{2}} x^2 \sin x \, dx$$

Інтегруючи частинами два рази, одержимо потрібне математичне сподівання

$$\int_0^{\frac{\pi}{2}} x^2 \sin x \, dx = -x^2 \cos x \Big|_0^{\frac{\pi}{2}} + 2 \int_0^{\frac{\pi}{2}} x \cos x \, dx = 2 \left(x \sin x \Big|_0^{\frac{\pi}{2}} - \int_0^{\frac{\pi}{2}} \sin x \, dx \right) = \pi + 2 \cos x \Big|_0^{\frac{\pi}{2}} = \pi - 2$$

Отже, одержали $M(Y) = M(X^2) = \pi - 2$

Дисперсію функції Y неперервного випадкового аргументу X визначають звичайним чином

$D(Y) = M(Y^2) - M^2(Y)$, а обчислюють за формулою

$$D(Y) = D(\varphi(X)) = \int_{-\infty}^{\infty} \varphi^2(x) f(x) \, dx - \left(\int_{-\infty}^{\infty} \varphi(x) f(x) \, dx \right)^2$$

У випадку, коли X змінюється лише в проміжку $[a, b]$, дисперсію функції $Y = \varphi(X)$ знаходять за формулою

$$D(Y) = D(\varphi(X)) = \int_a^b \varphi^2(x) f(x) \, dx - \left(\int_a^b \varphi(x) f(x) \, dx \right)^2$$

Функція двох випадкових аргументів

Якщо кожній парі можливих значень випадкових величин X та Y відповідає одне можливе значення випадкової величини Z , то Z називають функцією двох випадкових аргументів

$$Z = \varphi(X, Y)$$

На практиці найчастіше зустрічається задача знайти розподіл функції $Z = X + Y$ по відомих розподілам доданків.

Приклад. X – похибка показників вимірювального приладу (розподілена нормально, параметри закону розподілу відомі), Y – похибка округлення показників до найближчої поділки шкали (розподілена рівномірно, параметри закону розподілу відомі). Знайти закон розподілу сумарної похибки $Z = X + Y$.

Нехай X та Y – дискретні незалежні випадкові величини.

Щоб знайти закон розподілу сумарної похибки $Z = X + Y$ потрібно знайти всі можливі значення Z та їх імовірності.

Приклад. Дискретні незалежні випадкові величини задані розподілами. Знайти $Z = X + Y$.

X	1	2	Y	3	4
p	0,4	0,6	p	0,2	0,8

Розв'язок. Можливі значення Z – це суми кожного значення X з усіма можливими значеннями Y :

$$z_1 = 1 + 3 = 4; \quad z_2 = 1 + 4 = 5; \quad z_3 = 2 + 3 = 5; \quad z_4 = 2 + 4 = 6.$$

Аргументи X та Y незалежні, тому події $X=1$ та $Y=3$ теж незалежні, імовірність їх сумісної появи (тобто події $Z=1+3=4$) можна визначити за теоремою множення незалежних подій:

$$P(Z = z_1 = 1+3=4) = 0.4 \cdot 0.2 = 0.08;$$

Аналогічно

$$P(Z = z_2 = 1+4=5) = 0.4 \cdot 0.8 = 0.32;$$

$$P(Z = z_3 = 2+3=5) = 0.6 \cdot 0.2 = 0.12;$$

$$P(Z = z_4 = 2+4=6) = 0.6 \cdot 0.8 = 0.48;$$

Події z_2 та z_3 несумісні, тому їх імовірності додаємо

$$P(Z=5 = z_2+z_3) = 0.32+0.12=0.44$$

Z	4	5	6
p	0,08	0,44	0,48

$$0.08+0.44+0.48=1$$

Нехай X та Y – неперервні випадкові величини.

Доведено: якщо X та Y незалежні, то щільність розподілу $g(z)$ суми $Z = X+Y$ (за умови, що щільність хоча б одного з аргументів задана на інтервалі $(-\infty, \infty)$ однією формулою) може бути знайдена з рівності

$$g(z) = \int_{-\infty}^{\infty} f_1(x)f_2(z-x)dx = \int_{-\infty}^{\infty} f_1(z-y)f_2(y)dy$$

де $f_1(x)$, $f_2(y)$ – щільності розподілу аргументів.

- Щільність розподілу суми незалежних випадкових величин називають **композицією**.
- Закон розподілу називають **стійким**, якщо композиція таких законів є тим самим законом (що, можливо, відрізняється параметрами).

Приклад. Нормальний закон має властивість стійкості: композиція нормальних законів також має нормальний розподіл (математичне

сподівання та дисперсія цієї композиції дорівнюють відповідно суммам математичних сподівань та дисперсій доданків).

Приклад. Незалежні величини X та Y задані щільностями розподілів

$$f(x) = \frac{1}{3} e^{-\frac{x}{3}} ; f(y) = \frac{1}{4} e^{-\frac{y}{4}} \quad (0 \leq x, y < \infty)$$

Знайти щільність розподілу $g(z)$ суми $Z = X + Y$.

Розв'язок. Щільність розподілу $g(z)$ суми $Z = X + Y$

$$g(z) = \int_{-\infty}^{\infty} f_1(x) f_2(z-x) dx = \int_{-\infty}^{\infty} \left[\frac{1}{3} e^{-\frac{x}{3}} \right] \cdot \left[\frac{1}{4} e^{-\frac{(z-x)}{4}} \right] dx = \frac{1}{12} e^{-\frac{z}{4}} \int_0^z e^{-\frac{x}{12}} dx = e^{-\frac{z}{4}} \left(1 - e^{-\frac{z}{12}} \right)$$

Перевірка: $\int_{-\infty}^{\infty} g(z) dz = 1$

Розподіл χ^2 («хі-квадрат»)

Нехай X_i ($i = 1, 2, \dots, n$) – нормальні, нормовані незалежні величини, тобто їх математичне сподівання дорівнює нулю, середнє квадратичне відхилення дорівнює одиниці і кожна з них розподілена за нормальним законом. Тоді сума квадратів цих величин

$$\chi^2 = \sum_{i=1}^n X_i^2$$

розподілена по закону χ^2 з $k = n$ степенями свободи.

Якщо величини X_i зв'язані одним лінійним співвідношенням, наприклад,

$$\sum_{i=1}^n X_i = n \bar{X}$$

то число степенів свободи буде $k = n - 1$.

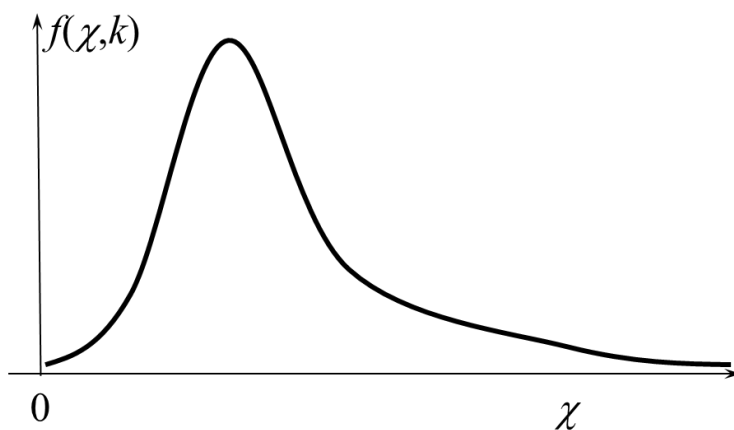
Щільність розподілу χ^2 має вигляд

$$f(x) = \begin{cases} 0, & \text{при } x \leq 0 \\ \frac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot e^{-\frac{x}{2}} \cdot x^{\frac{k}{2}-1}, & \text{при } x > 0 \end{cases}$$

де $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$ – гама-функція, $\Gamma(n+1) = n!$

- ✓ Розподіл χ^2 визначається одним параметром – числом степенів свободи k .
- ✓ Коли k зростає, розподіл χ^2 дуже повільно прямує до нормального розподілу.

Закон розподілу статистики χ^2 – несиметрична функція.



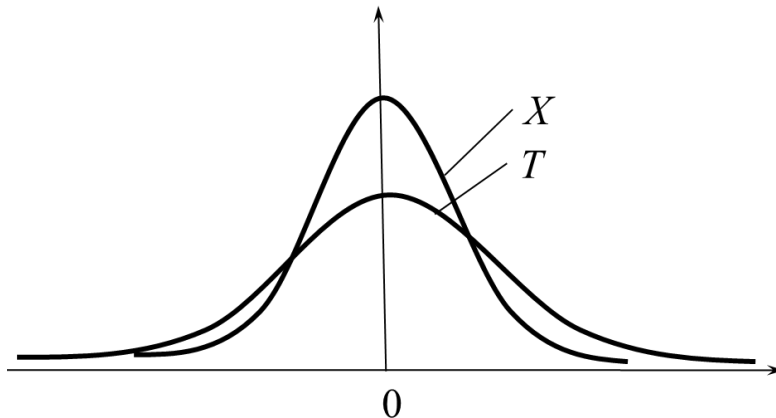
Розподіл Стюдента

Нехай X – нормальна нормована випадкова величина, а Y – незалежна від X величина, яка розподілена за законом χ^2 з k степенями свободи. Тоді величина

$$T = \frac{X}{\sqrt{Y/k}}$$

має розподіл, який називають ***t-розподілом*** або ***розподілом Стьюдента*** з k степенями свободи.

При зростанні k розподіл Стьюдента швидко наближається до нормального розподілу.



Розподіл Фішера-Снедекора

Нехай X та Y – незалежні випадкові величини, розподілені за законом χ^2 з k_1 та k_2 степенями свободи відповідно. Тоді величина

$$F = \frac{X/k_1}{Y/k_2}$$

має розподіл, який називають ***F-розподілом*** або ***розподілом Фішера-Снедекора*** з k_1 та k_2 степенями свободи.

Підсумки

- ✓ Якщо вказано закон, за яким кожному можливому значенню випадкової величини X відповідає певне значення випадкової величини Y , то кажуть, що Y функція X і позначають $Y = \varphi(X)$.
- ✓ На практиці найчастіше зустрічається задача знайти розподіл функції $Z = X + Y$. Щільність розподілу суми незалежних випадкових величин називають композицією.

✓ Закон розподілу називають стійким, якщо композиція таких законів є тим самим законом.

✓ На практиці найчастіше зустрічаються такі функції випадкових величин:

- розподіл χ^2 $\chi^2 = \sum_{i=1}^n X_i^2$

- розподіл Стюдента $T = \frac{X}{\sqrt{Y/k}}$

- розподіл Фішера-Снедекора $F = \frac{X/k_1}{Y/k_2}$

2.8. Багатомірні випадкові величини та їх закони розподілу імовірностей

Розглянуті раніше випадкові величини X , які при кожному випробуванні визначались одним можливим числовим значенням. Тому таку випадкову величину X називають *одновимірною*.

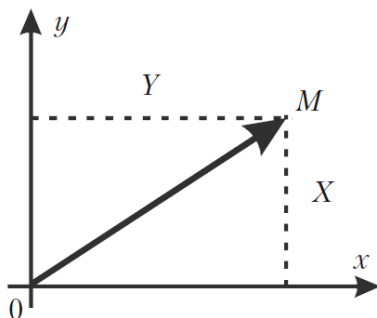
Якщо можливі значення випадкової величини визначаються у кожному випробуванні 2, 3, ..., n числами, то такі величини називають відповідно *дво-, три-, ..., n -вимірними* відповідно.

Двовимірну випадкову величину будемо позначати (X, Y) , X та Y при цьому будуть компонентами. Величини X та Y , що розглядаються одночасно, утворюють систему двох випадкових величин.

Аналогічно можна розглядати систему n випадкових величин.

- Сукупність n випадкових величин (X_1, X_2, \dots, X_n) , що розглядаються одночасно, називають *системою випадкових величин*.

Систему n випадкових величин (X_1, X_2, \dots, X_n) можна розглядати як випадкову точку в n -вимірному просторі з координатами (X_1, X_2, \dots, X_n) або як випадковий вектор, направлений з початку системи координат у точку $M(X_1, X_2, \dots, X_n)$. При $n = 2$ маємо систему двох випадкових величин (X, Y) .



Закон розподілу імовірностей дискретної двовимірної випадкової величини

➤ *Законом розподілу дискретної двовимірної випадкової величини*

називають перелік можливих значень цієї величини (x_i, y_k) та їх імовірностей $p(x_i, y_k)$, $i = 1, 2, \dots, n$; $k = 1, 2, \dots, m$.

Закон розподілу двовимірної дискретної випадкової величини найчастіше задають у вигляді таблиці

Y	X					
	x_1	x_2	\dots	x_i	\dots	x_n
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$	\dots	$p(x_i, y_1)$	\dots	$p(x_n, y_1)$
y_2	$p(x_1, y_2)$	$p(x_2, y_2)$	\dots	$p(x_i, y_2)$	\dots	$p(x_n, y_2)$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
y_k	$p(x_1, y_k)$	$p(x_2, y_k)$	\dots	$p(x_i, y_k)$	\dots	$p(x_n, y_k)$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
y_m	$p(x_1, y_m)$	$p(x_2, y_m)$	\dots	$p(x_i, y_m)$	\dots	$p(x_n, y_m)$

Події $(X=x_i, Y=y_k)$, $i = 1, 2, \dots, n$; $k = 1, 2, \dots, m$, утворюють повну групу, тому сума імовірностей таблиці дорівнює одиниці, тобто

$$\sum_{i=1}^n \sum_{k=1}^m p(x_i, y_k) = 1$$

Закон розподілу двовимірної випадкової величини дозволяє отримати закони розподілу кожної компоненти.

Події $(x_i, y_1), (x_i, y_2), \dots, (x_i, y_m)$, несумісні, тому імовірність $P(x_i)$ того, що X прийме значення x_i за теоремою додавання імовірностей буде

$$P(x_i) = P(x_i, y_1) + P(x_i, y_2) + \dots + P(x_i, y_m)$$

тобто дорівнює сумі імовірностей, що розташовані в i -тому стовпчику таблиці розподілу.

Аналогічно, додаванням імовірностей k -ого рядка таблиці, одержимо імовірність

$$P(y_k) = P(x_1, y_k) + P(x_2, y_k) + \dots + P(x_n, y_k)$$

Приклад. Знайти закони розподілу компонент двовимірної випадкової величини, закон розподілу якої заданий таблицею

Y	X		
	x_1	x_2	x_3
y_1	0,1	0,3	0,2
y_2	0,06	0,18	0,16

Розв'язок. Закони розподілу X та Y будуть мати вигляд

X	x_1	x_2	x_3
P	0,16	0,48	0,36

Y	y_1	y_2
P	0,6	0,4

Імовірності відповідних значень X та Y знаходимо так

$$p(x_1) = 0.1 + 0.06 = 0.16;$$

$$p(x_2) = 0.3 + 0.18 = 0.48;$$

$$p(x_3) = 0.2 + 0.16 = 0.36;$$

$$p(y_1) = 0.1 + 0.3 + 0.2 = 0.6;$$

$$p(y_2) = 0.06 + 0.18 + 0.16 = 0.4$$

Контроль:

$$0.16 + 0.48 + 0.36 = 1;$$

$$0.6 + 0.4 = 1.$$

- **Функцією розподілу (інтегральним законом розподілу) двовимірної випадкової величини (X, Y)** називають функцію двох змінних $F(x, y)$, яка визначає для кожної пари чисел (X, Y) імовірність виконання нерівностей $X < x; Y < y$, тобто
- $$F(x, y) = P(X < x; Y < y)$$

Аналогічно визначають функцію розподілу системи n випадкових величин

$$F(x_1, x_2, \dots, x_n) = P(X_1 < x_1, X_2 < x_2, \dots, X_n < x_n)$$

Властивості функції розподілу двовимірної випадкової величини:

1) $0 \leq F(x, y) \leq 1$;

2) $F(x, y)$ не спадна функція за кожним аргументом, тобто

$$F(x_2, y) \geq F(x_1, y), \text{ якщо } x_2 \geq x_1;$$

$$F(x, y_2) \geq F(x, y_1), \text{ якщо } y_2 \geq y_1.$$

3) Мають місце граничні співвідношення

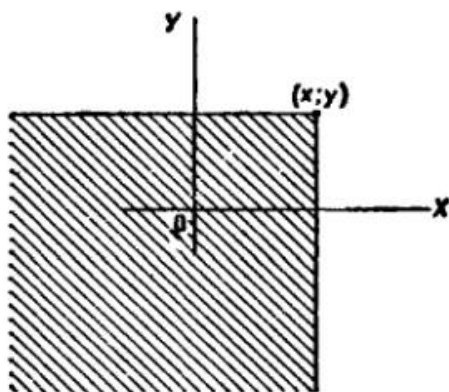
$$F(-\infty, y) = 0; F(x, -\infty) = 0; F(-\infty, -\infty) = 0; F(\infty, \infty) = 1.$$

4) Імовірність влучення випадкової точки до прямокутника $\{x_1 \leq X \leq x_2; y_1 \leq Y \leq y_2\}$

можна знайти за формулою

$$P(x_1 < X < x_2, y_1 < Y < y_2) = \{F(x_2, y_2) - F(x_1, y_2)\} - \{F(x_2, y_1) - F(x_1, y_1)\}$$

Геометричний зміст функції розподілу $F(x, y)$ – це імовірність того, що випадкова точка $M(X, Y)$ потрапить у нескінчений прямокутник з вершиною в точці (X, Y) і розміщений нижче та лівіше цієї вершини.



Приклад. Знайти імовірність влучення випадкової точки (X, Y) у прямокутник, обмежений лініями

$$x_1 = \frac{\pi}{6}; \quad x_2 = \frac{\pi}{2}; \quad y_1 = \frac{\pi}{3}; \quad y_2 = \frac{\pi}{4};$$

якщо задана функція розподілу вигляду

$$F(x, y) = \sin(x) \cdot \sin(y); \quad 0 \leq x \leq \frac{\pi}{2}; \quad 0 \leq y \leq \frac{\pi}{2};$$

Розв'язок. Згідно з формулою одержуємо

$$\begin{aligned} P\left(\frac{\pi}{6} < X < \frac{\pi}{2}; \frac{\pi}{3} < Y < \frac{\pi}{4}\right) &= \left[\sin \frac{\pi}{2} \cdot \sin \frac{\pi}{3} - \sin \frac{\pi}{6} \cdot \sin \frac{\pi}{3}\right] - \\ &- \left[\sin \frac{\pi}{2} \cdot \sin \frac{\pi}{4} - \sin \frac{\pi}{6} \cdot \sin \frac{\pi}{4}\right] = \left[\frac{\sqrt{3}}{2} - \frac{1}{2} \cdot \frac{\sqrt{3}}{2}\right] - \left[\frac{\sqrt{2}}{2} - \frac{1}{2} \cdot \frac{\sqrt{2}}{2}\right] = \\ &= \frac{\sqrt{3} - \sqrt{2}}{4} \approx 0.08 \end{aligned}$$

Закон розподілу імовірностей неперервної двовимірної випадкової величини

➤ ***n -вимірною щільністю імовірностей (диференціальним законом розподілу)*** неперервної випадкової величини називають мішану частинну похідну від інтегральної функції розподілу

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n F(x_1, x_2, \dots, x_n)}{\partial x_1 \cdot \partial x_2 \cdot \dots \cdot \partial x_n}$$

Таким чином, якщо функція розподілу n -вимірної випадкової величини відома, то за цією формулою можна знайти диференціальний закон розподілу цієї випадкової величини.

Двовимірну випадкову величину можна задавати функцією розподілу $F(x, y) = P(X < x; Y < y)$ або щільністю розподілу.

➤ **Двовимірною щільністю імовірностей $f(x, y)$ (диференціальним законом розподілу)** двовимірної випадкової величини (X, Y) називають мішану частинну похідну другого порядку від інтегральної функції розподілу

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

Якщо відома щільність розподілу імовірностей $f(x, y)$ двовимірної випадкової величини, то її функцію розподілу знаходять за формулою

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(x, y) dx dy$$

тобто з використанням невластного двократного інтегралу.

Імовірність влучення випадкової точки (X, Y) в довільну область D знаходять за формулою

$$P((X, Y) \in D) = \iint_D f(x, y) dx dy$$

Функція щільності розподілу імовірностей $f(x, y)$ задовольняє властивостям:

1) $f(x, y) \geq 0$, тобто вона не від'ємна;

$$2) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

Закон розподілу неперервної двовимірної випадкової величини дозволяє отримати закони розподілу кожної компоненти.

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(x, y) dx dy$$

$$F_1(x) = F(x, \infty) = \int_{-\infty}^x \int_{-\infty}^{\infty} f(x, y) dx dy$$

$$f_1(x) = \frac{dF_1(x)}{dx} = \int_{-\infty}^{\infty} f(x, y) dy$$

$$f_2(y) = \frac{dF_2(y)}{dy} = \int_{-\infty}^{\infty} f(x, y) dx$$

Залежні та незалежні випадкові величини

- Дві випадкові величини **незалежні**, якщо закон розподілу однієї з них не залежить від того, які можливі значення прийняла друга величина.
- Випадкові величини **залежні**, якщо закон розподілу однієї величини залежить від того, які значення прийняла друга величина.
- Теорема. Щоб випадкові величини X та Y були незалежні, необхідно і достатньо, щоб функція розподілу системи (X, Y) дорівнювала добутку функцій розподілу кожної з них

$$F(x, y) = F_1(x) \cdot F_2(y)$$
- Наслідок. Щоб неперервні випадкові величини X та Y були незалежними, необхідно і достатньо, щоб функція щільності розподілу системи (X, Y) дорівнювала добутку функцій щільності розподілу складових

$$f(x, y) = f_1(x) \cdot f_2(y)$$

Умовні закони розподілу складових системи випадкових величин

Коли події A та B залежні, то умовна імовірність події B відрізняється від її безумовної імовірності і при цьому

$$P_A(B) = \frac{P(AB)}{P(A)}$$

Аналогічно і для випадкових величин.

- **Умовним розподілом** складової Y двовимірної випадкової величини при $X=x_i$ називають сукупність умовних імовірностей $p(y_1/x_i), p(y_2/x_i), \dots, p(y_m/x_i)$, обчислених при припущенні, що подія $X=x_i$ (i має одне і теж значення при всіх значеннях Y) вже настала.

$$p\left(\frac{y_j}{x_i}\right) = \frac{p(x_i, y_j)}{p(x_i)}, \quad (j=1, 2, \dots, m)$$

Сума імовірностей умовного розподілу дорівнює 1.

$$\sum_{j=1}^m p\left(\frac{y_j}{x_i}\right) = 1$$

Нехай (X, Y) – неперервна випадкова величина.

- **Умовною щільністю розподілу** складових Y при заданому значенні $X=x$ називають відношення щільності сукупного розподілу $f(y, x)$ системи (X, Y) до щільності розподілу складової X .

$$\varphi\left(\frac{y}{x}\right) = \frac{f(x, y)}{f_1(x)} \quad \int_{-\infty}^{\infty} \varphi\left(\frac{y}{x}\right) dy = 1$$

Щільність розподілу складової X можна знайти так:

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

Числові характеристики багатомірних випадкових величин

Математичне сподівання двовимірної випадкової величини (X, Y) характеризує координати центру розподілу випадкової величини. Ці координати у випадку неперервних величин знаходять за формулами

$$m_x = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy; \quad m_y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy$$

Дисперсії D_X та D_Y характеризують розсіювання випадкової точки (X, Y) вздовж координатних осей Ox та Oy , відповідно. Їх знаходять за формулами

$$D_X = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 \cdot f(x, y) dx dy - m_X^2;$$

$$D_Y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y^2 \cdot f(x, y) dx dy - m_Y^2$$

середньоквадратичні відхилення

$$\sigma_X = \sqrt{D_X}; \quad \sigma_Y = \sqrt{D_Y}$$

- **Умовним математичним сподіванням** випадкової величини Y при $X=x$ (x певне можливе значення X) називають добуток можливих значень Y на їх умовні імовірності.

Для дискретних величин:

$$M(Y/X = x) = \sum_{j=1}^m y_j p(y_j/x)$$

Для неперервних величин:

$$M(Y/X = x) = \int_{-\infty}^{\infty} y \cdot \varphi(y/x) dy$$

- Умовне математичне сподівання $M(Y/x)$ є функцією від x , яку називають **функцією регресії** Y на X .

Для опису двовимірної випадкової величини крім математичного сподівання, дисперсії та середньоквадратичних відхилень використовують також інші характеристики, а саме — *кореляційний момент* (або *коваріація*).

- **Кореляційним моментом** випадкових величин X та Y називають математичне сподівання добутку відхилень цих величин

$$\text{cov}(X, Y) = K_{XY} = M((X - m_X)(Y - m_Y))$$

Для обчислення кореляційного моменту дискретних величин X та Y використовують формулу

$$K_{XY} = \sum_{i=1}^n \sum_{j=1}^m (x_i - M(X))(y_j - M(Y))p(x_i, y_j)$$

Для неперервних величин X та Y

$$K_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_X)(y - m_Y)f(x, y)dx dy$$

- ✓ Враховуючи, що відхилення – це центровані випадкові величини, кореляційний момент можна визначити як математичне сподівання добутку центрованих випадкових величин

$$K_{XY} = M(\overset{\circ}{X} \overset{\circ}{Y})$$

- ✓ Кореляційний момент можна записати у вигляді

$$K_{XY} = M(XY) - M(X) \cdot M(Y)$$

Кореляційний момент використовується для характеристики зв'язку між величинами X та Y .

- Теорема. Кореляційний момент двох незалежних випадкових величин X та Y дорівнює нулю.
- Теорема. Абсолютна величина кореляційного моменту двох випадкових величин X та Y не перевищує середнього геометричного їх дисперсій.

$$|K_{XY}| \leq \sqrt{D_X D_Y} = \sigma_X \sigma_Y$$

- ✓ З визначення кореляційного моменту слідує, що він має розмірність, що дорівнює добутку розмірностей величин X та Y . Отже, величина кореляційного моменту залежить від одиниць виміру випадкових величин, що утруднює порівняння.

Для усунення цього недоліку використовують іншу числову характеристику – коефіцієнт кореляції.

- **Коефіцієнт кореляції** є кількісна характеристика залежності випадкових величин X та Y (безрозмірна величина)

$$r_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y}$$

Коефіцієнт кореляції для дискретних величин X та Y

$$r_{XY} = \sum_{i=1}^n \sum_{j=1}^m \left(\frac{x_i - M(X)}{\sigma_X} \right) \left(\frac{y_j - M(Y)}{\sigma_Y} \right) p(x_i, y_j)$$

для неперервних величин X та Y

$$r_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\frac{x_i - m_X}{\sigma_X} \right) \left(\frac{y_j - m_Y}{\sigma_Y} \right) f(x_i, y_j) dx dy$$

Властивості коефіцієнта кореляції:

- 1) $|r_{XY}| \leq 1$;
- 2) якщо X та Y незалежні, то $r_{XY} = 0$;
- 3) якщо між X та Y є лінійна залежність $Y = aX + b$, де a та b – постійні, то $|r_{XY}| = 1$.

- ✓ Дві корельовані величини обов'язково залежні. Але дві залежні випадкові величини можуть бути корельованими або некорельованими, тобто їх коефіцієнт кореляції може як дорівнювати, так і не дорівнювати нулеві.

- ✓ Із незалежності двох величин випливає їх некорельованість, але із некорельованості ще не випливає незалежність цих величин. У випадку нормального розподілу величин із некорельованості випадкових величин випливає їх незалежність.
- ✓ Дві випадкові величини можуть бути пов'язані або функціональною залежністю, або статистичною (кореляційною) залежністю, або бути незалежними.
- ✓ Строга *функціональна* залежність зустрічається рідко, бо обидві або одна випадкова величина може знаходитися під впливом випадкових факторів. Серед цих факторів можуть бути і такі, що впливають на обидві випадкові величини. У цьому випадку виникає *статистична* залежність.
- ✓ Якщо статистична залежність проявляється в тому, що при зміні однієї з величин змінюється середнє значення іншої, то таку статистичну залежність називають *кореляційною*.

Підсумки

- ✓ Багатомірні випадкові величини задають або інтегральним законом розподілу (функцією розподілу), або диференціальним законом розподілу (багатомірною імовірністю для ДВВ і функцією щільності для НВВ).
- ✓ Закон розподілу багатовимірної випадкової величини дозволяє отримати закони розподілу кожної компоненти.
- ✓ Компоненти багатовимірної випадкової величини можуть бути залежними або незалежними між собою.
- ✓ Для характеристики зв'язку між величинами використовують коваріацію або коефіцієнт кореляції.

- ✓ Умовне математичне сподівання $M(Y/x)$ є функцією регресії Y на X .

Розділ 3. Математична статистика

3.1. Способи збору даних

Задачі, які розв'язує математична статистика

Мета кожного наукового дослідження – виявлення закономірностей спостережуваних явищ та використання цих закономірностей у повсякденній практичній діяльності. Для встановлення цих закономірностей проводять спеціальні дослідження та спостерігають одиничні явища. Далі роблять узагальнений висновок у вигляді закону.

У тих випадках, коли явище знаходиться під дією багатьох факторів і неможливо виявити вплив усіх цих факторів, застосовують інший метод вивчення – *статистичний*, тобто систематизацію та обробку статистичних даних однорідних дослідів.

Основні задачі, які розв'язує математична статистика:

- 1) вказати способи збору та групування (якщо даних дуже багато) статистичних відомостей;
- 2) визначити закон розподілу випадкової величини або системи випадкових величин за статистичними даними;
- 3) визначити невідомі параметри розподілу;
- 4) перевірити правдоподібність припущень про закон розподілу випадкової величини, про значення параметра, який оцінюють, або про форму зв'язку між випадковими величинами.

Генеральна та вибіркова сукупності

Нехай потрібно вивчити сукупність об'єктів відносно деякої ознаки, яка характеризує ці об'єкти. Кожен об'єкт спостереження в загальному

випадку має декілька ознак. Розглядаючи лише одну ознаку кожного об'єкта, припускаємо, що інші ознаки рівноправні, або що **множина об'єктів однорідна**. Такі множини однорідних об'єктів називають **статистичною сукупністю**.

Розрізняють декілька різновидів статистичних сукупностей.

- **Генеральною** називають сукупність об'єктів, відносно яких проводиться дослідження та з яких можна зробити вибірку.
- **Вибірковою сукупністю (вибіркою)** називають сукупність випадково взятих об'єктів генеральної сукупності.
- **Об'ємом сукупності** (вибіркової або генеральної) називають кількість об'єктів цієї сукупності.

Наприклад, якщо з 5000 виробів для дослідження взято 50, тоді об'єм генеральної сукупності $N = 5000$, а об'єм вибірки $n = 50$.

Перевірку сукупності об'єктів можна провести двома способами:

- 1) провести перевірку (контроль) усіх об'єктів;
 - 2) перевірити лише певну частину об'єктів.
- **Переписами** називають обстеження, що мають своєю метою дослідження кожного елементу сукупності (генеральної сукупності), що вивчається.

Зразками перепису є перепис населення у країні, звіт про виробничі показники усіх підприємств у галузі (наприклад, шахт вугільної промисловості).

Якщо об'єктів дуже багато або перевірка пов'язана з руйнуванням об'єкту (наприклад, випробування деталі на міцність), тоді перший спосіб перевірки не доцільний.

Якщо дослідити усі об'єкти неможливо, тоді відбирають із усієї сукупності обмежене число об'єктів і перевіряють лише їх. Перевага вивчення вибірки порівняно з переписом: малі затрати коштів, обладнання та часу.

- ✓ Вибірку можна ефективно використовувати для вивчення відповідної ознаки генеральної сукупності лише тоді, коли дані вибірки вірно відображають цю ознаку, тобто вибірка повинна бути ***репрезентативною***.

Згідно із законом великих чисел теорії імовірностей можна стверджувати, що вибірка буде репрезентативною лише тоді, коли її здійснюють випадково.

Одні й ті ж самі дані можна розглядати або як генеральну сукупність, або як вибірку, в залежності від мети їх збору та аналізу. Наприклад, оцінки за екзамен для всіх учнів класу – генеральна сукупність, якщо хочемо описати розподіл оцінок в цьому класі, але ці оцінки можна розглядати як вибірку, якщо мета аналізу - на підставі цих оцінок зробити висновок про оцінки інших учнів в школі.

Аналіз генеральної сукупності передбачає, що набір даних містить всі об'єкти дослідження, отже можна напряму робити висновки про характеристики цієї групи. В протилежність цьому при аналізі вибірки працюємо тільки з частиною генеральної сукупності, і будь-які висновки, які робляться про цю велику групу на підставі вибірки, є імовірнісними, а не абсолютними.

Різниця принципова, і для її фіксації використовують різні терміни та позначення при роботі з різними статистичними сукупностями. Так, числа, що характеризують генеральну сукупність, називають параметрами і

позначають грецькими літерами, а числа, що характеризують вибірку, називають статистиками і позначають латинськими літерами.

Джерела даних у статистиці. Способи відбору

Джерелами даних можуть бути:

- вибіркові обстеження
- спеціально поставлені експерименти
- дані, що є результатом повсякденної (рутинної) роботи

Всі джерела даних можна поділити на первинні і вторинні.

- **Первинні дані** збираються спеціально для статистичного дослідження. Для цих даних є відомості про методи збирання, точність даних і т.д.

Більш цінними даними у статистиці є первинні дані, але їх не завжди можливо отримати, тому часто використовуються вторинні дані.

- **Вторинними даними** є дані, що використовуються у статистиці, але спочатку збиралися для інших цілей.

Рутинні записи про діяльність фірм, офіційні статистичні звіти є вторинними даними.

В залежності від обсягу та структури статистичної сукупності можна використовувати різні способи формування вибірки. В більшості випадків для математичної статистики найбільш зручним засобом здійснення випадкового відбору є простий випадковий відбір.

- **Простим випадковим** є такий **відбір** з генеральної сукупності, при якому кожний об'єкт, що відбирається, має однакову імовірність потрапити до вибірки.

- ✓ Вибірка, що здійснена за допомогою простого випадкового відбору, називається **простою випадковою вибіркою**.

За допомогою простого випадкового відбору можна сформувати повторні та безповторні вибірки.

- **Повторною** називають вибірку, при якій відібраний об'єкт повертається до генеральної сукупності перед відбором іншого об'єкта.
- Вибірку називають **безповторною**, якщо взятий об'єкт до генеральної сукупності не повертається.

Найчастіше використовують безповторні вибірки.

Для здійснення простої випадкової вибірки необхідна наявність **основи вибірки**, тобто такого представлення генеральної сукупності, при якому її елементи були хоча б перераховані.

- ✓ Основа вибірки повинна повністю відображати ознаку генеральної сукупності, що вивчається - порушення цієї вимоги може зробити вибірку не репрезентативною.

Приклад.

а) Генеральна сукупність – всі покупці крамниці.

Мета дослідження – виявити товари, що мають найбільший попит серед різних категорій покупців.

Основою вибірки можуть бути робочі списки покупців, що веде крамниця.

б) Генеральна сукупність – всі жителі міста, що мають телефон.

Мета дослідження – виявити категорії абонентів, відсоток міжміських переговорів яких більше 30%.

Основою вибірки може бути довідкова телефонна книга, не може – список телефонів батьків першокласників.

За допомогою розшарованого випадкового відбору генеральна сукупність розділяється на частини за певним принципом, використовуючи:

- – механічний відбір;
- – серійний відбір;
- – типовий відбір.

➤ **Механічним** називають відбір, при якому генеральна сукупність механічно поділяється на стільки частин, скільки має бути об'єктів у вибірці. З кожної частини випадковим чином відбирають один об'єкт.

Наприклад, якщо потрібно перевірити 25% усіх виготовлених станком-автоматом виробів, то відбирають кожен четвертий виріб.

✓ Щоб механічний відбір був репрезентативним, треба враховувати специфіку технологічного процесу.

➤ **Серійним** називають відбір, при якому об'єкти із генеральної сукупності відбирають не по одному, а серіями, які і досліджують.

✓ Серійний відбір використовують тоді, коли ознака, яку досліджують, мало змінюється в різних серіях.

➤ **Типовим** називають відбір, при якому об'єкти відбирають не з усієї генеральної сукупності, а лише з її типових частин.

Наприклад, якщо вироби виготовлені на різних станках, то відбір проводять лише з виробів кожного станка окремо.

✓ Типовий відбір доцільно використовувати тоді, коли типові частини суттєво відрізняються.

В економічних та соціальних дослідженнях іноді використовують **комбінований** відбір.

Наприклад, спочатку поділяють генеральну сукупність на серії однакового об'єму, випадковим чином відбирають декілька серій і, нарешті, з кожної серії випадковим чином беруть окремі об'єкти.

Загальні рекомендації щодо обсягу вибірки:

- Найбільший обсяг вибірки потрібен при перевірці гіпотез про вигляд розподілу 200-2000 елементів.
- Якщо потрібно порівняти дві вибірки, то їх загальний обсяг має бути не менше 50 елементів та обсяг кожної з них приблизно однаковий.
- Якщо вивчається зв'язок між якимись ознаками (перевірка гіпотез про параметри розподілу або аналіз впливу факторів), то обсяг вибірки не менше 30-40 елементів.
- Чим більша варіативність ознаки, що вивчається, тим більшим має бути обсяг вибірки.

Спотворення даних вибірок

Більшість досліджень проводиться на вибірках об'єктів з генеральної сукупності. Щоб дослідник міг спокійно використовувати отриману вибірку для характеристики всієї генеральної сукупності, вона має добре характеризувати цю генеральну сукупність (тобто бути репрезентативною). Якщо ж вибірка зсунута (містить спотворення типу систематичних помилок), то висновки, зроблені на її основі, не можна розповсюджувати на всю генеральну сукупність. Можна назвати кілька причин спотворення вибірки.

Неповне охоплення - виникає, якщо деякі об'єкти не мають шансів бути включеними у вибірку. Наприклад, телефонні опитування з використанням

номерів з довідника залишають поза межами потенційних респондентів, що не мають телефонів або змінили номер після публікації довідника. Це може спричинити спотворення даних, якщо виключені з дослідження люди систематично виділяються по досліджуваним ознакам. (Зокрема, люди, що мешкають в будинках без телефону, зазвичай бідніші тих, хто має телефон, а люди, в яких є тільки мобільний телефон, зазвичай молодші тих, в кого є ще й домашній. Якщо рівень прибутків або вік пов'язані з досліджуваною ознакою, виключення таких людей з вибірки призведе до зсуву результатів дослідження.)

Нерівномірність вибору – виникає, якщо деякі об'єкти мають більше шансів бути включеними у вибірку. *Вплив волонтерів* – відображає той факт, що люди, які добровільно погоджуються на участь в дослідженні, як правило, не типові для генеральної сукупності. (При телефонному опитуванні вдень скоріш за все відповіді будуть отримані від пенсіонерів та домогосподарок.)

Неотримані дані:

- *Відсутність відповіді* – люди, що відмовляються від участі в дослідженнях, скоріш за все відрізняються від тих, хто погодився.
- *Втрата об'єктів при тривалому дослідженні* – це типове явище, яке призводить до спотворення результатів у випадку, коли об'єкти випадають не випадково, а по причинах, пов'язаних з предметом дослідження. (При дослідженні ефективності нової методики лікування в першу чергу від участі в експерименті відмовляться ті, кому не стає краще вже через кілька днів.)

Спотворення відгуку (інформаційне спотворення)

- *Вплив опитувача* – може виникнути, якщо опитувач знає мету дослідження та якимось чином проявляє своє відношення до тих чи інших відповідей.
- *Можливість виявлення* – певні характеристики можуть бути виявленими або зафіксованими у різних груп об'єктів по різному. (контроль допінгу в спорті набагато жорсткіше проводиться у вищій лізі, але це не означає, що в інших змаганнях допінгом користуються менше.)
- *Соціальна привабливість* – викликана прагненням людей представити себе у вигідному світлі. (Це часто спонукає людей давати такі відповіді, які, на їхню думку, сподобаються опитувачу.)

Підсумки

- Мета кожного наукового дослідження – виявлення закономірностей спостережуваних явищ та використання цих закономірностей у повсякденній практичній діяльності.
- Дослідження проводиться на статистичній сукупності об'єктів – генеральній або вибірковій.
- В залежності від обсягу та структури статистичної сукупності можна використовувати різні способи формування вибірки.
- Щоб дослідник міг спокійно використовувати отримані для вибірки результати для характеристики всієї генеральної сукупності, вибірка має бути репрезентативною.

3.2. Первинна обробка даних

Загальна схема процесу дослідження включає три кроки:

- 1) Визначити об'єкт спостереження
 - мета аналізу
 - ознаки, що цікавлять
 - шкали для вимірювання ознак
- 2) Сформувати вибірку
 - результати спостережень
 - дані для аналізу
 - ряд розподілу
 - частотні таблиці
- 3) Провести аналіз
 - описативний аналіз
 - перевірка гіпотез

Шкали для вимірювання ознак

Кожен об'єкт спостереження в загальному випадку має декілька ознак. Всі ознаки можна поділити на:

- якісні
- кількісні

Приклад. При дослідженні партії деталей якісною ознакою може бути стандартність або нестандартність кожної деталі, а кількісною ознакою – розмір деталі.

Кількісні ознаки бувають *неперервними* та *дискретними*.

Можливий перехід від більш детальної шкали до простішої.

Якщо порядкова шкала має 4 і менше значень, то для більшості досліджень краще трактувати її як номінальну.

Приклад.

значення ознаки	шкала			
	відносна °K	інтервальна °C	порядкова	номінальна
	323	> +40	сильна спека	тепло
	300	+26 .. +40	спека	
	289	+16 .. +25	тепло	
	274	+1 .. +15	прохолодно	холодно
	253	-20 .. 0	холодно	
	150	< -20	дуже холодно	
	0			

Організація даних

Дані у статистиці надходять до дослідника у вигляді неорганізованої маси, незалежно від того, чи є вони вибірковими даними, чи даними з генеральної сукупності. Тому перед початком аналізу їх необхідно певним чином підготувати.

Якість аналізу суттєво залежить від якості даних: «Сміття на вході – сміття (хибні переконання) на виході».

Загальні моменти по керуванню даними.

Ієрархія – для ефективного керування даними у великих проектах необхідно визначити ієрархію людей, уповноважених приймати рішення при виникненні проблем.

Кодифікатор – спосіб збору та організації інформації про проект, містить:

- інформацію про проект та методи збору даних;
- методи введення даних в комп'ютер;

- рішення, прийняті відносно даних;
- процедури кодування.

При наявності кодифікатора ті, хто приєднується до проекту або аналізує дані значно пізніше, ніж їх збирали, знатимуть, що це за дані та як їх інтерпретувати.

Файл даних – існує багато способів збереження даних в електронному вигляді, але найбільш розповсюджений формат – це прямокутний файл даних.

При визначенні структури зберігання даних в електронному вигляді є сенс обирати її з огляду на способи її подальшої обробки. Інколи при зборі даних зручніше записати їх спочатку в одному форматі (Excel), а потім перенести в інший (SPSS, CSV).

В призначених для статистичного аналізу даних прийнято, щоб рядки відповідали об'єктам, а стовпчики – ознакам.

Перевірка файлу даних виконується для прийняття рішення про можливість використання даних для аналізу.

Дано: файл, супровідна документація, тип аналізу. Перевірити треба:

- Чи достатньо спостережень?
- Чи всі потрібні ознаки присутні?
- Чи є повторювані спостереження? Чим викликані?
- Чи вірно конвертовані значення, назви, підписи?
- Чи в розумних межах лежать значення ознак?
- Скільки значень пропущено і чи є в цьому якась закономірність? Що з ними робити?

Ряд розподілу

У математичній статистиці замість слова «дані» вживається термін «варіанти».

- **Варіанта** – це числова характеристика ознаки, для дослідження якої робиться вибірка.

Нехай із генеральної сукупності взята вибірка об'єктів $\{x_1, x_2, \dots, x_n\}$ об'єму n , для вивчення ознаки X . Тобто, значення x_1, x_2, \dots, x_n є варіанти ознаки X .

Першим кроком обробки є впорядкування варіант.

- Варіанти, що записані до таблиці у зростаючому (спадяючому) порядку, називають **варіаційним рядом** (дискретним варіаційним рядом).

Після впорядкування можна отримати більше інформації про ознаку, наприклад, про межі зміни.

- Різниця максимального та мінімального значень варіанти $R = x_{\max} - x_{\min}$ називається **розмахом варіант**.

До визначення розмаху варіант доцільно виключити з вибірки аномальні спостереження (наприклад, з вибірки зарплат по підприємству видалити зарплату власника).

Розподіл частот

Нехай у вибірці обсягу n з варіантами x_1, x_2, \dots, x_m ознака X прийняла значення x_1 - n_1 раз, значення x_2 - n_2 раз, ..., значення x_m - n_m раз.

- Додатне число, що вказує, скільки раз та чи інша варіанта зустрічається в таблиці даних, називається **частотою**.
- Ряд n_1, n_2, \dots, n_m називається **рядом частот**.

Сума усіх частот повинна дорівнювати об'єму вибірки.

$$\sum_{i=1}^m n_i = n$$

Статистичний розподіл вибірки встановлює зв'язок між рядом варіант, що зростає або спадає, і відповідними частотами. Він може бути представлений таблицею, де n — об'єм вибірки, $n = n_1 + n_2 + \dots + n_m$.

x_i	x_1	x_2	\dots	x_m
n_i	n_1	n_2	\dots	n_m

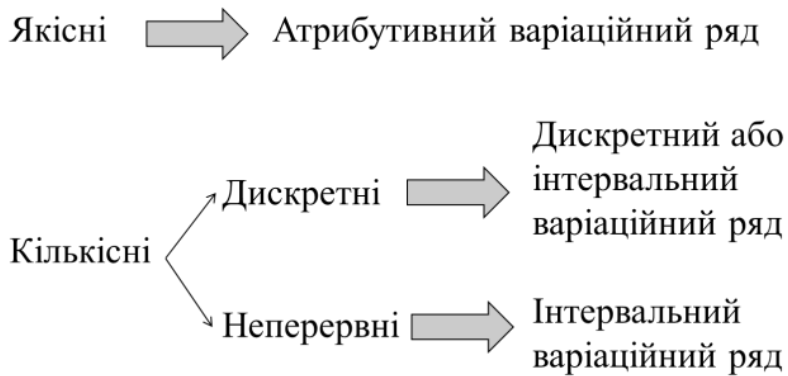
Нерідко замість значень частот використовуються відносні частоти. Нехай існує m частот n_1, n_2, \dots, n_m .

- Відношення частоти n_i варіанти x_i до об'єму вибірки n називається **відносною частотою** або **частотою**, причому, сума усіх відносних частот дорівнює 1.

Залежність між упорядкованим рядом варіант і відповідними їм відносними частотами або частотами також називається **статистичним розподілом вибірки**.

Вторинні дані, що використовуються у статистиці і спочатку збиралися для інших цілей, надходять до дослідника у вигляді неорганізованої досить великої маси. Тому перед початком аналізу їх доцільно спочатку групувати.

Види статистичних ознак, по яким можна виконувати групування даних:



У випадку використання інтервального ряду спочатку задаються кількістю інтервалів.

➤ Кожний інтервал називається **класом інтервалів** або **класом**.

Рекомендації щодо вибору:

- число класів k зручніше приймати непарним;
- при загальному числі замірів $n \geq 100$ доцільно брати $9 \leq k \leq 15$;
- при загальному числі замірів $n \leq 100$ можна вважати $k = 7$.

Вибір кількості інтервалів групування можна робити з використанням формули Стерджесса

$$k = 1 + 3.322 \cdot \lg n$$

Її використання не рекомендується, якщо:

- вибірка не досить велика;
- розподіл вибірки не схожий на нормальний;
- в результаті отримано порожні класи.

Інколи неможливо, або небажано вибирати ширину класів інтервалів однаковою. Нерівна ширина класів бажана, наприклад, коли значення частоти одного чи декількох класів набагато більше (менше) значень частот інших інтервалів. Як правило, ширина інтервалів зростає (або

спадає) і може містити інтервали відкритого типу «більше ніж...», «менше ніж...».

Загальна схема побудови згрупованого розподілу частот

1. Визначити найбільше x_{\max} та найменше x_{\min} значення варіанти x_i і визначити розмах варіант $R = x_{\max} - x_{\min}$.
2. Задатися певним числом класів k .
3. Визначити ширину класу $h = R/k$. Для спрощення розрахунків, отримане значення h округлити у будь-яку сторону.
4. Встановити границі класів і підрахувати кількість варіант у кожному класі. При підрахунку числа варіант значення x_i , що знаходиться на границі класів, слід відносити завжди до одного й того ж класу, наприклад, до наступного класу, де це число зустрілось вперше. Воно, таким чином, стає нижньою границею класу.
5. Визначити частоту для кожного класу і записати ряд розподілу.

Підсумки

- Якість аналізу суттєво залежить від якості даних.
- Збирати дані потрібно орієнтуючись на специфіку об'єкту та мету дослідження.
- Перед формуванням вибірки необхідно проконтролювати якість даних спостережень.
- При великих обсягах даних є сенс формувати згруповані вибірки.

3.3. Описова статистика

У більшості випадків під терміном «статистичний аналіз» мають на увазі статистичну перевірку гіпотез (вивідну статистику). Однак є ще один різновид статистики – описовий, який використовує методи статистики та графічні підходи для представлення інформації про дані, що вивчаються. Майже всі, хто працює з обробкою даних, використовують обидва різновиди статистики, і часто обчислення описових статистик – це попередній етап перед перевіркою гіпотез. Особливо широко практикують аналіз графічного представлення даних та розрахунок найпростіших описових статистик, щоб краще відчувати дані, що аналізуються.

Описова статистика (дескриптивний аналіз вибірки) включає в себе:

- оцінку міри центральної тенденції
- оцінку міри варіативності
- графічні зображення статистичного розподілу
- аналіз нормальності розподілу

Оцінка міри центральної тенденції

Міри центральної тенденції, також відомі як міри розташування, є одними з перших статистик, які розраховуються для ознак з щойно отриманих даних. Головна мета їх розрахунку – дати уявлення про типові значення досліджуваної ознаки. Три міри, що найчастіше використовуються – мода, медіана, середнє.

➤ **Модой** називають значення варіанти, яка має найбільшу частоту або для інтервального ряду

$$Mo = x_0 + h \cdot \frac{n_s - n_{s-1}}{(n_s - n_{s-1}) + (n_s - n_{s+1})}$$

де x_0 – початок модального інтервалу, h – крок інтервального ряду, n_s – частота модального інтервалу, n_{s-1} – частота інтервалу, що передує модальному, n_{s+1} – частота наступного за модальним інтервалу.

- **Медіаною** називається значення середнього елемента варіаційного ряду при непарній кількості елементів і середньоарифметичне двох середніх елементів при їхній парній кількості, або для інтервального ряду

$$Me = x_0 + h \cdot \frac{n/2 - T_{s-1}}{n_s}$$

де x_0 – початок медіанного інтервалу, тобто інтервалу, в якому утримується середній елемент, h – крок інтервального ряду, n_s – частота медіанного інтервалу, T_{s-1} – сума частот інтервалів, які передують медіанному.

Медіану згрупованого розподілу частот можна знайти як значення або точку на горизонтальній осі гістограми розподілу частот, в якій перпендикулярна лінія, що проходить через неї, поділяє площу гістограми на дві рівні частини.

Крім структурних середніх - моди та медіани, які не залежать від значень варіант, що розташовані на краях розподілу, у статистиці застосовуються ще різноманітні середні.

- **Генеральною середньою** називають середнє арифметичне варіант ознаки, розраховане по генеральній сукупності

$$\bar{x}_G = \frac{1}{N} \sum_{i=1}^N x_i = \mu$$

де x_i ($i=1, 2, \dots, N$) – значення ознаки, N – об'єм генеральної сукупності.

- **Простою середньоарифметичною вибірки** називають суму варіант вибірки, поділену на їх кількість

$$\bar{x}_a = \frac{1}{m} \sum_{i=1}^m x_i$$

де x_i ($i=1, 2, \dots, m$) – варіанти вибірки.

➤ **Вибірковою середньою** або **зваженою середньоарифметичною** називають середню арифметичну варіант вибірки з врахуванням їх частот

$$\bar{x} = \frac{1}{n} \sum_{i=1}^m n_i x_i$$

де n – об'єм вибірки, m – число різних варіант, n_1, \dots, n_m – частоти варіант ($n = n_1 + \dots + n_m$), x_i – значення i -тої варіанти.

Вибіркова середня є аналогом математичного сподівання і використовується дуже часто.

✓ Вибіркова середня може приймати різні числові значення при різних вибірках однакового об'єму.

Основні властивості вибіркової середньої:

1. При множенні усіх варіант вибірки на однаковий множник вибіркова середня також множиться на цей множник

$$\frac{1}{n} \sum_{i=1}^m n_i (c x_i) = \frac{c}{n} \sum_{i=1}^m n_i x_i = c \bar{x}$$

2. Якщо додати (відняти) до всіх варіант вибірки однакове число, то вибіркова середня зростає (зменшується) на це число

$$\frac{1}{n} \sum_{i=1}^m n_i (c + x_i) = \frac{c}{n} \sum_{i=1}^m n_i + \frac{1}{n} \sum_{i=1}^m n_i x_i = c + \bar{x}$$

➤ **Степеневою середньою вибірки** називають таку середню, яку знаходять за формулою

$$\bar{x}_C = \left(\sum_{i=1}^m \frac{n_i}{n} x_i^\alpha \right)^{\frac{1}{\alpha}}$$

При $\alpha = 1$ одержимо вибірккову середню.

$$\bar{x}_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^m n_i x_i$$

При $\alpha = 2$ одержимо *середньоквадратичну вибірки*

$$\bar{x}_2 = \sqrt{\sum_{i=1}^m \frac{n_i}{n} x_i^2}$$

При $\alpha = -1$ одержуємо *середню гармонічну*

$$\bar{x}_{-1} = \bar{x}_{zap} = \frac{n}{\sum_{i=1}^m \frac{n_i}{x_i}}$$

Середню гармонічну застосовують у тому випадку, коли шуканий показник є величина, що обернена середньому значенню ознаки.

При $\alpha = 0$ вираз буде невизначеним. Застосовуючи логарифмування та правило Лопітала для розкриття невизначеності, одержимо *середню геометричну*

$$\bar{x}_0 = \bar{x}_z = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_m^{n_m}}$$

Ця середня обчислюється лише при умові, що усі варіанти додатні $x_i > 0$, $i = 1, 2, \dots, m$.

Середня геометрична застосовується у статистиці для визначення темпу зростання при дослідженні змін ознаки з часом.

Групова та загальна середні.

Припустимо, що всі значення кількісної ознаки X розбито на декілька груп (тобто побудовано інтервальний варіаційний ряд). Якщо розглядати кожну групу як самостійну сукупність, можна знайти її середню арифметичну.

- **Груповою середньою** називають середнє арифметичне варіант ознаки, розраховане по групі.
- **Загальною середньою** \bar{x} називають середнє арифметичне варіант ознаки, розраховане по генеральній сукупності.

Загальна середня дорівнює середньому арифметичному групових середніх, зважених по обсягам груп.

Середнє – це інтуїтивно зрозуміла міра центральної тенденції, але середнє в цій якості слід використовувати не для будь-яких даних, бо воно чутливе до екстремальних значень (викидів), а також може призвести до невірних висновків у випадку асиметричного розподілу даних.

Отже, обрання тієї чи іншої міри центральної тенденції для характеристики ознаки пов'язано з якісним аналізом розподілу цієї ознаки:

- можна брати будь-яку, якщо розподіл близький до симетричного та унімодальний (схожий на нормальний); найчастіше обирають середнє вибіркове;
- краще брати медіану, якщо розподіл мультимодальний;
- краще брати моду, якщо розподіл має суттєво виражену асиметрію;
- краще брати моду, якщо шкала виміру ознаки порядкова або номінальна.

Оцінка міри варіативності

До мір варіативності відносять:

- розмах варіант
- міжквартильний розмах

- дисперсія
- середньоквадратичне відхилення

Розташуємо всі елементи сукупності в порядку зростання (тобто проранжуємо їх).

- **Квантиль** – значення, яке випадкова величина не перевищує із заданою імовірністю.

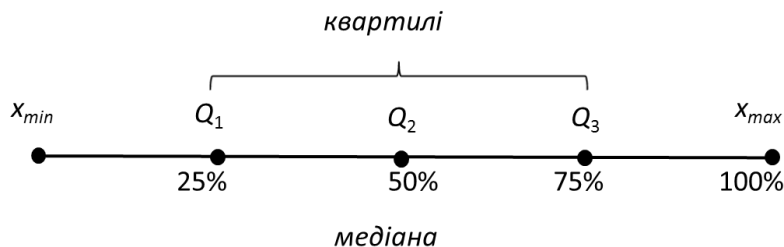
Використовуються квартилі, децилі та персентилі.

- Різниця максимального та мінімального значень варіанти називається **розмахом варіант**.

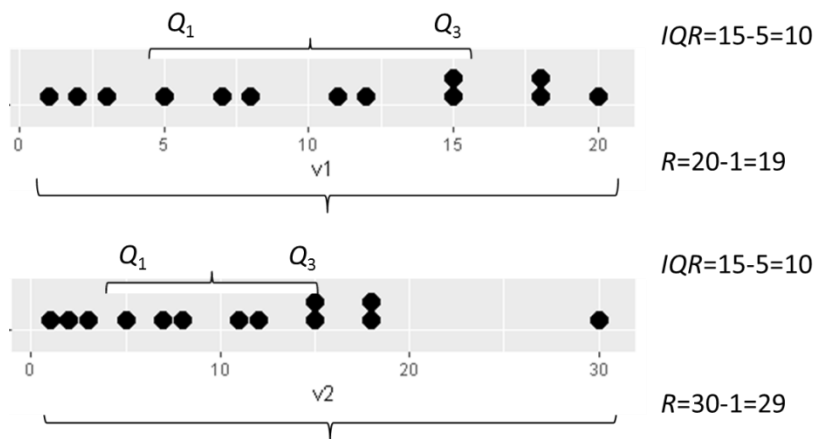
$$R = x_{\max} - x_{\min}$$

- Різниця 75% та 25% рівнями значень варіанти називається **міжквартильним розмахом**.

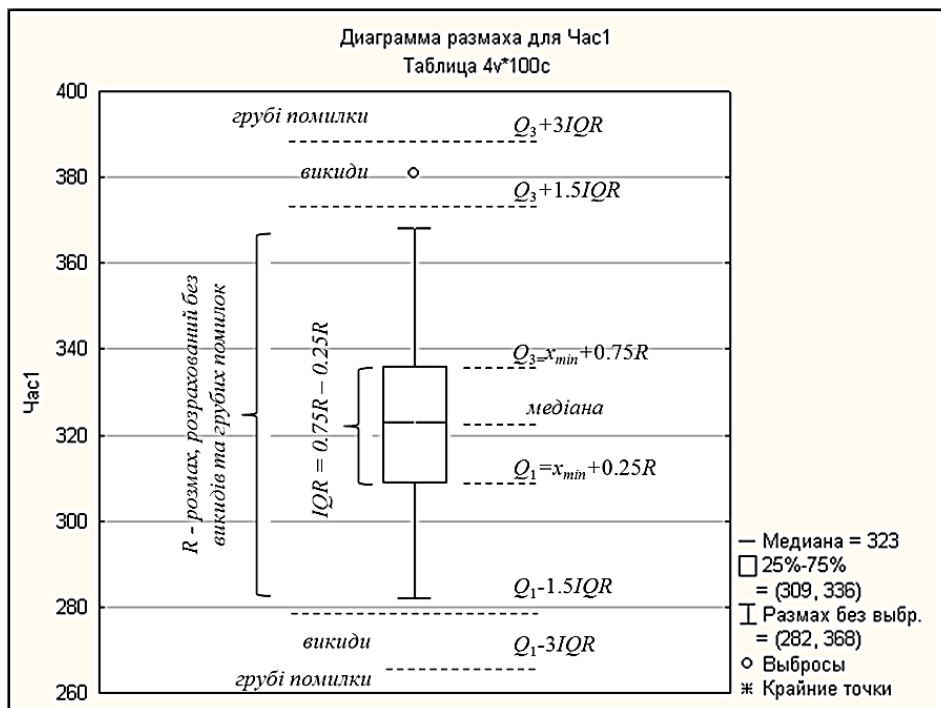
$$IQR = Q_3 - Q_1$$



Міжквартильний розмах більш стійкий до викидів.



Для ілюстрації цих мір варіативності зручно використовувати діаграму розмаху.



На цій діаграмі значення, що сильно відрізняються від основної маси даних, позначені як викиди та грубі помилки. Виявлення та аналіз причин появи таких значень – важливий етап попереднього аналізу даних, тому що наявність навіть одного такого значення може сильно спотворити значення багатьох статистик (наприклад, середнього вибіркового).

Найбільш часто в якості мір варіативності для кількісних ознак використовують дисперсію та стандартне відхилення.

Формули їх розрахунку залежать від того, для якої статистичної сукупності міри розраховуються.

- **Генеральною дисперсією D** називають середню квадратів відхилення варіант від генеральної середньої з врахуванням відповідних частот

$$D = \frac{1}{n} \sum_{i=1}^m n_i (x_i - \mu)^2$$

Обчислення дисперсії спрощується, якщо її знаходити за формулою

$$D = \frac{1}{n} \sum_{i=1}^m n_i x_i^2 - \left(\frac{1}{n} \sum_{i=1}^m n_i x_i \right)^2 = \frac{1}{n} \sum_{i=1}^m n_i x_i^2 - \mu^2$$

- **Вибірковою дисперсією** називають середню квадратів відхилення варіант від вибіркової середньої з врахуванням відповідних частот

$$D_B = s^2 = \frac{1}{n} \sum_{i=1}^m n_i (x_i - \bar{x})^2$$

Обчислення вибіркової дисперсії спрощується, якщо її знаходити за формулою

$$D_B = \frac{1}{n} \sum_{i=1}^m n_i x_i^2 - \left(\frac{1}{n} \sum_{i=1}^m n_i x_i \right)^2 = \frac{1}{n} \sum_{i=1}^m n_i x_i^2 - (\bar{x})^2$$

- **Генеральним середньоквадратичним відхиленням** (стандартним відхиленням) називають квадратний корінь із генеральної дисперсії

$$\sigma = \sqrt{D}$$

➤ **Вибірковим середньоквадратичним відхиленням** (стандартним відхиленням) називають квадратний корінь із вибіркової дисперсії

$$s = \sqrt{D_B}$$

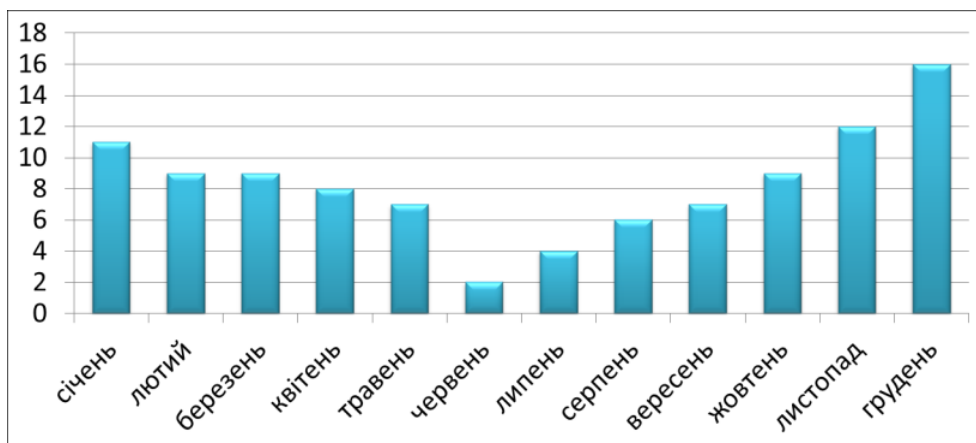
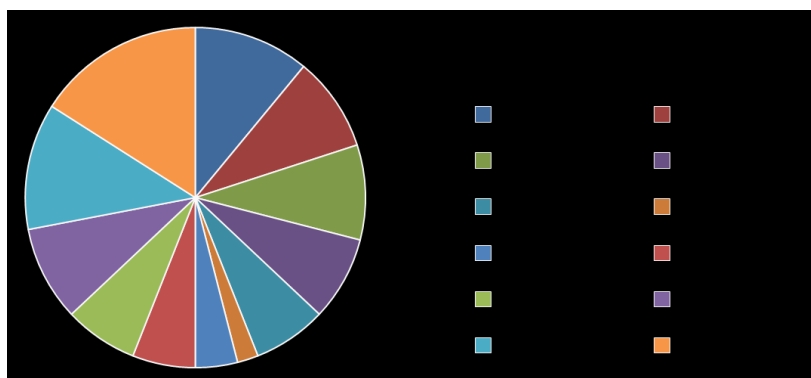
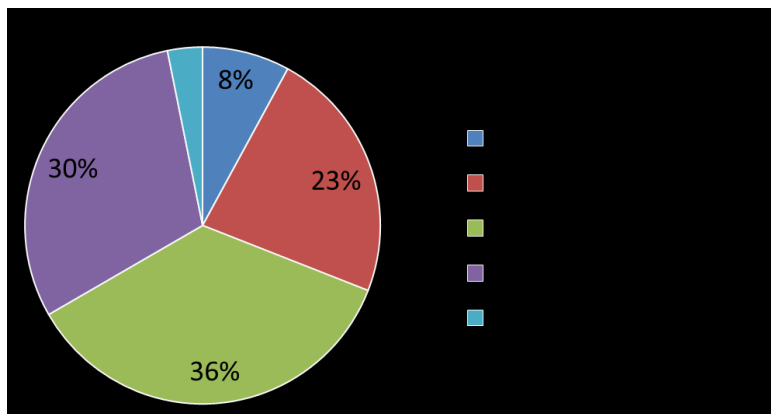
Не всі міри однаково доступні при використанні різних шкал вимірювання

шкала		номінальна	порядкова	інтервальна	відносна
міри центральної тенденції	мода	V	V	V	V
	медіана		V	V	V
	середнє			V	V
міри варіативності	розмах варіант		V	V	V
	дисперсія			V	V
	СКВ			V	V

Графічні зображення статистичного розподілу

Загальне правило при виборі типу графічного зображення – використовуйте найпростіший, який зрозуміло представить дані. Можна орієнтуватися також на тип ознаки та шкалу, в якій вона виміряна.

Тип ознаки	Шкала	Тип графіки		
якісна	номінальна	кругова діаграма	стовпчикова діаграма	лінійний графік
	порядкова			
кількісна	інтервальна			
	відносна			



Гістограма частот та функція розподілу

Для зображення розподілів згрупованих та незгрупованих даних використовують різні графіки.

Незгруповані дані.

Якщо в результаті вибірки одержано статистичний розподіл ознаки X , яку треба дослідити, то маємо

перелік варіант ознаки	x_1, x_2, \dots, x_m
та відповідних їм частот	n_1, n_2, \dots, n_m
або відносних частот	W_1, W_2, \dots, W_m

Значення варіант та частот або відносних частот можна розглядати як координати точок.

➤ **Полігоном частот** називають ламану, відрізки якої з'єднують точки $(x_1, n_1), (x_2, n_2), \dots, (x_m, n_m)$

➤ **Полігоном відносних частот** (частостей) називають ламану, відрізки якої проходять через точки $(x_1, W_1), (x_2, W_2), \dots, (x_m, W_m)$

Полігони частот та частостей є аналогами щільності імовірності. Площа фігури, обмежена полігоном щільності відносних частот приблизно дорівнює площі контуру діаграми, що в свою чергу дорівнює 1. Отже, полігон щільності відносних частот є приблизним зображенням функції щільності імовірності генеральної сукупності.

Якщо збільшувати об'єм вибірки n , то полігон щільності відносних частот буде більш точно зображати функцію щільності імовірності генеральної сукупності.

Згруповані дані.

Для згрупованого розподілу частоти, відносної частоти, щільності частоти і щільності відносної частоти можуть бути побудовані спеціальні діаграми, складені з прямокутників ступінчасті фігури, що називаються **гістограмами**.

Для побудови гістограми на горизонтальну вісь наносяться класи інтервалів. На кожному класі будується прямокутник, висота якого рівна

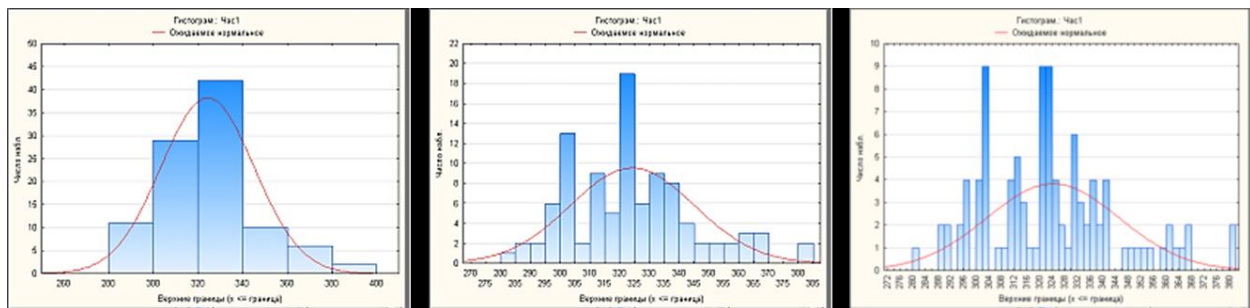
значенню частоти (або відносної частоти, або щільності частоти, або щільності відносної частоти) на цьому інтервалі.

- **Гістограмою частот** називають фігуру, яка складається з прямокутників, основами яких є інтервали варіант довжиною $h = x_k - x_{k-1}$, а висоти дорівнюють n_k/h (щільність частоти).
- **Гістограмою відносних частот** називають фігуру, яка складається з прямокутників, основами яких є інтервали варіант, а висоти дорівнюють відношенню W_k/h (щільність відносної частоти).

Площа гістограми частот дорівнює об'єму вибірки, а площа гістограми відносних частот – одиниці.

Від способу обрання ширини інтервалів h залежить виразність гістограми:

- якщо h занадто мале, то гістограма містить багато випадкового;
- якщо h занадто велике, то в гістограмі зникають індивідуальні особливості вибірки.



В деяких задачах замість полігону (гістограми) частот (відносних частот), які є статистичними аналогами функції щільності розподілу, зручніше користуватися накопиченими частотами, які є статистичними аналогами функції розподілу.

- **Емпіричною функцією розподілу** (або **функцією розподілу вибірки**) називають функцію $F^*(x)$, яка визначає для кожного значення x відносну частоту події $X < x$.

$$F^*(x) = \frac{n_x}{n}$$

де n_x – кількість варіант, які менше від x , n – об'єм вибірки.

- ✓ Інтегральну функцію розподілу $F(x)$ генеральної сукупності у математичній статистиці називають **теоретичною функцією розподілу**. Вона відрізняється від емпіричної функції розподілу $F^*(x)$ тим, що визначає імовірність події $X < x$, а не частість цієї події.

З теореми Бернуллі випливає, що частість події $X < x$ прямує до імовірності цієї події. Тому $F(x)$ та $F^*(x)$ мало відрізняються одна від одної.

Доцільно використовувати $F^*(x)$ для наближеного представлення функції розподілу $F(x)$ генеральної сукупності.

Властивості емпіричної функції розподілу $F^*(x)$

1. $0 \leq F^*(x) \leq 1$.
2. $F^*(x)$ – зростаюча функція.
3. $F^*(x) = \begin{cases} 0, & x \leq x_1 \\ 1, & x > x_m \end{cases}$

де x_1 – найменша варіанта, x_m – найбільша варіанта.

Закон розподілу

Теоретичні розподіли імовірностей корисні для статистичних висновків, оскільки їх властивості та характеристики визначені. Якщо реальний розподіл досліджуваного набору даних близький до теоретичного, більшість обчислень для цих даних можна зробити з використанням припущень, що спираються на властивості теоретичного розподілу. Крім того, завдяки центральній граничній теоремі при певних умовах можна вважати, що вибіркові середні розподілені нормально, навіть якщо

значення генеральної сукупності, з якої взяті ці вибірки, мають інший розподіл.

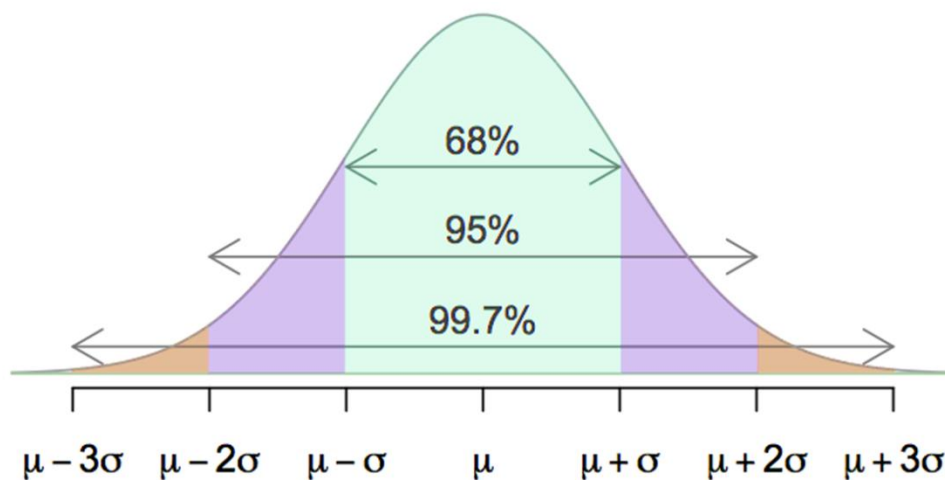
Нормальний розподіл найчастіше використовується у статистиці.

Для всіх нормальних розподілів незалежно від їх середнього значення та середньоквадратичного відхилення характерні загальні властивості:

- симетричність
- унімодальність (одне найбільше значення)
- неперервність
- рівність середнього, медіани та моди.

Оскільки всі нормальні розподіли мають однакову загальну форму, можна сформулювати певні тези про те, як розподілені дані при будь-якому нормальному розподілі:

- близько 68% даних знаходяться в інтервалі \pm одне стандартне відхилення від середнього;
- близько 95% даних знаходяться в інтервалі \pm два стандартних відхилення від середнього;
- близько 99% даних знаходяться в інтервалі \pm три стандартних відхилення від середнього.



Для того, щоб порівняти два нормальних розподіли (або один, про який відомо, що він нормальний, і інший, у якого «підозрюємо» цю властивість), необхідно привести їх до одного масштабу – стандартного нормального розподілу (Z-розподілу) з нульовим середнім та одиничним середньоквадратичним відхиленням.

$$Z = \frac{X - \mu}{\sigma}$$

Цей перехід називають нормалізацією.

На практиці статистичні висновки настільки часто спираються на припущення про те, як розподілені дані, що в статистиці прийнято перетворювати дані так, щоб їх було легко порівняти з відомими типами розподілів.

Перший крок в перетворенні даних – розглянути уважно набір даних та вирішити, яке перетворення підійде в даному випадку та чи потрібне воно взагалі. Для аналізу даних з цією метою можна побудувати гістограму частот та накласти на неї криву нормального розподілу – це дозволить візуально оцінити розподіл даних, а також виявити нехарактерні та екстремальні значення. Розуміння загальної форми розподілу також допомагає вирішити, який тип перетворення можна спробувати. Інший підхід полягає у обчисленні однієї із статистик, розроблених для перевірки відповідності даних певному розподілу. Найчастіше з цією метою використовуються статистики Колмогорова-Смирнова або Андерсона-Дарлінга.

Зсунутий вліво розподіл даних може бути наближений до нормального за допомогою обчислення квадратного кореня або логарифмування.

Якщо розподіл зсунутий вправо, та можна спочатку дзеркально відобразити дані (віднявши кожне значення від максимального плюс одиниця), а потім з них взяти квадратний корінь або прологарифмувати.

Перетворення даних – це не гарантоване вирішення проблеми з розподілом, тому після нього обов’язково потрібно знову перевірити дані на нормальність. Крім того, перетворення міняє одиницю вимірювання даних, тому інтерпретація будь-якої статистики, порашованої по новим значенням, повинна враховувати цю особливість.

Нормалізовані дані також дають можливість оцінити імовірність появи тих чи інших значень у вибірці.

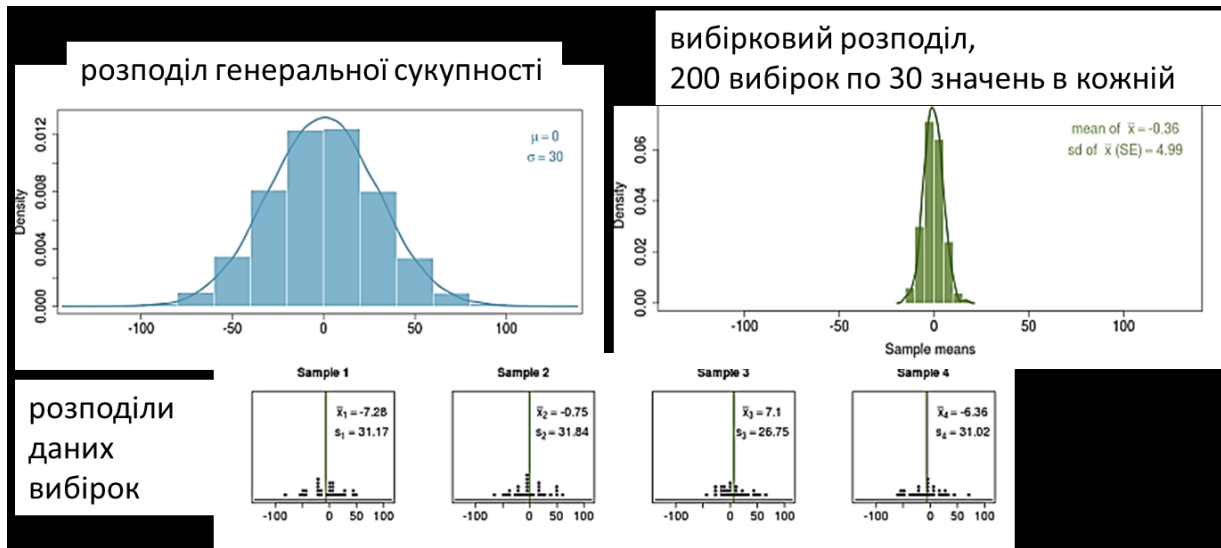
При аналізі даних важливо розрізняти статистичні розподіли:

1. розподіл генеральної сукупності
2. розподіл даних вибірки
3. вибірковий розподіл (розподіл вибіркових середніх).

Розподіл генеральної сукупності може бути абсолютно довільним, чим більшу вибірку з генеральної сукупності беремо – тим більше розподіл даних вибірки буде на нього схожий. Вибірковий розподіл (розподіл вибіркових середніх) завжди нормальний, якщо обсяги вибірок досить великі (більше 30) – за центральною граничною теоремою.

статистики	розподіл генеральної сукупності	розподіл даних вибірки	вибірковий розподіл
міра центральної тенденції	μ	\bar{x}	$\mu_{\bar{x}}$
міра варіативності	σ СКВ	s вибіркове	$se_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ стандартне

		СКВ	відхилення
--	--	-----	------------



Аналіз нормальності розподілу

Методи аналізу вигляду закону розподілу поділяють на:

- непрямі
- графічні
- розрахункові

Непрямі методи аналізу не дають повної картини даних, лише можуть застерегти, що вигляд закону розподілу не схожий на нормальний.

До непрямих методів аналізу розподілу відносять оцінку коефіцієнта асиметрії та ексцесу.

➤ **Коефіцієнтом асиметрії** називають міру

$$Sk = \frac{n \cdot \mu_3}{(n-1) \cdot (n-2) \cdot \sigma^3} \quad \mu_3 = \sum_{i=1}^n (x_i - \bar{x})^3$$

де μ_3 - центральний момент третього порядку.

➤ **Ексцесом** називають міру

$$K = \frac{n \cdot (n+1) \cdot \mu_4 - 3 \cdot (\mu_2)^2 \cdot (n-1)}{(n-1) \cdot (n-2) \cdot (n-3) \cdot \sigma^4}$$

де μ_2 та μ_4 - центральні моменти другого та четвертого порядку. $K=0$ відповідає нормальному розподілу.

➤ **Стандартна помилка коефіцієнта асиметрії**

$$S_{Sk} = \sqrt{\frac{6 \cdot n \cdot (n-1)}{(n-2) \cdot (n+1) \cdot (n+3)}}$$

➤ **Стандартна помилка ексцесу**

$$S_K = \sqrt{\frac{4 \cdot (n^2 - 1) \cdot (S_{Sk})^2}{(n-3) \cdot (n+5)}}$$

Розподіл вважається нормальним, якщо абсолютні величини показників асиметрії та ексцесу менші за їх стандартні помилки в 3 або більше разів.

Графічні методи аналізу дозволяють візуально оцінити вигляд кривої розподілу та її відхилення від нормальної кривої. Для цього можна скористатися:

- гістограмою частот
- нормально-імовірністним графіком
- діаграмою розмаху (boxplot)

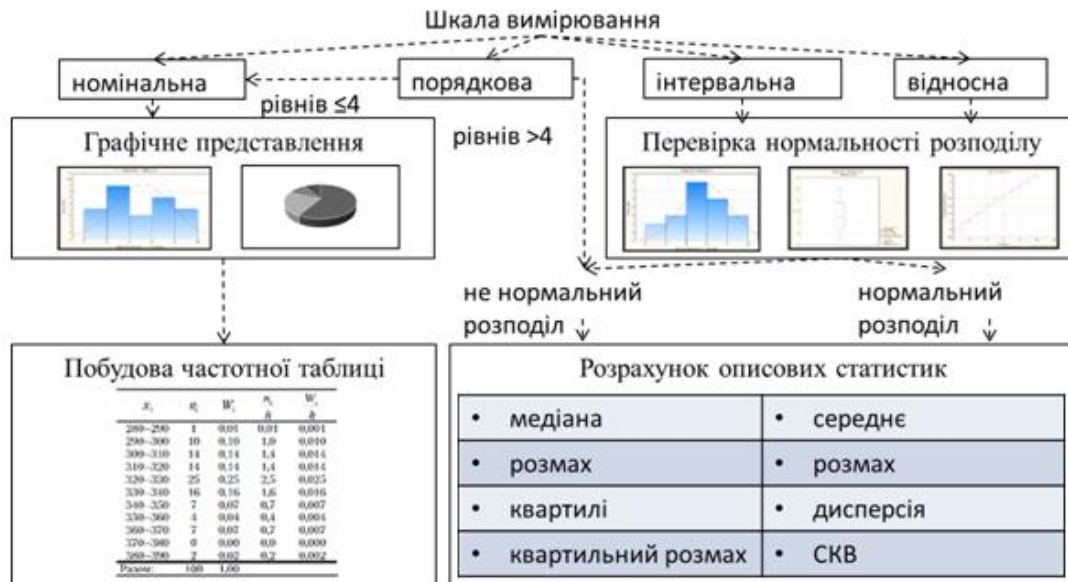
На гістограмі (полігоні) частот звертають увагу на форму кривої, що описує дані. На нормально-імовірністному графіку звертають увагу на кількість і місце варіант, що лежать далеко від нормальних даних. На діаграмі розмаху (boxplot) звертають увагу на симетричність, на кількість варіант, що потрапили в області викидів та грубих помилок.

Розрахункові методи аналізу нормальності розподілу – це перевірка гіпотез про вигляд розподілу.

Для перевірки на нормальність розподілу найчастіше користуються:

- критерієм Колмогорова-Смирнова (для вибірок обсягом >60);
- критерієм Шапіро-Уїлка (для вибірок обсягом <60).

Загальна схема аналізу даних



Підсумки

- Описова статистика – це попередній етап обробки даних перед перевіркою гіпотез.
- Міри центральної тенденції дають уявлення про типове значення досліджуваної ознаки. Найчастіше використовуються мода, медіана, середнє. Обрання тієї чи іншої міри пов'язане з якісним аналізом розподілу ознаки.
- Міри варіативності характеризують мінливість досліджуваної ознаки. Найчастіше використовуються міжквартильний розмах, дисперсія, середньоквадратичне відхилення.
- Не всі міри однаково доступні при використанні різних шкал вимірювання.

- Графічне зображення розподілу дає загальне уявлення про досліджувані дані. При виборі типу графічного зображення беремо найпростіший, який зрозуміло представить дані.
- Для того, щоб порівняти два розподіли, необхідно привести їх до одного масштабу.

3.4. Точкові оцінки параметрів розподілу

Вимоги до статистичних оцінок

Для того, щоб скласти враження про деяку ознаку X генеральної сукупності, використовуючи результати вибірки, часто достатньо знати наближені значення її міри центральної тенденції та міри варіативності.

Іноді з деяких міркувань вдається встановити закон розподілу X . Тоді треба вміти оцінювати параметри цього закону розподілу.

Наприклад, відомо, що випадкова величина X розподілена за нормальним законом, її щільність імовірностей має вигляд

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}$$

Необхідно оцінити (знайти наближені значення) параметрів a , який дорівнює $M(X)$, та σ , який дорівнює $\sigma(X)$. Ці параметри повністю визначають нормальний розподіл X .

Дослідник має у своєму розпорядженні лише дані вибірки, одержані в результаті спостережень. Саме через ці дані і треба виразити потрібний параметр випадкової величини X генеральної сукупності.

➤ **Статистичною оцінкою θ^* невідомого параметра θ випадкової величини X генеральної сукупності** (теоретичного розподілу X) називають функцію від випадкових величин (результатів вибірки), що спостерігаються

$$\theta^* = \theta^*(X_1, \dots, X_n)$$

Щоб статистичні оцінки давали найкращі наближення параметрів, вони повинні задовольняти певним вимогам.

Нехай θ^* є статистична оцінка невідомого параметра θ теоретичного розподілу.

Припустимо, що за вибіркою об'єму n знайдена оцінка θ_1^* . При інших вибірках того ж об'єму одержимо деякі інші оцінки $\theta_2^*, \theta_3^*, \dots, \theta_m^*$. Саме оцінку θ^* можна розглядати як випадкову величину, а числа $\theta_1^*, \theta_2^*, \dots, \theta_m^*$ як її можливі значення.

Якщо числа θ_k^* ($k=1, 2, \dots, m$) будуть більші значення параметра θ , тоді оцінка θ^* дає наближене значення θ з надлишком. У цьому випадку математичне сподівання випадкової величини θ^* буде більше θ , $M(\theta^*) > \theta$.

Якщо θ^* дає оцінку параметра θ з недостачею, тоді $M(\theta^*) < \theta$.

Таким чином, використання статистичної оцінки, математичне сподівання якої не дорівнює параметру θ , приводить до систематичних (одного знака) похибок.

Вимога $M(\theta^*) = \theta$ застерігає від систематичних похибок.

➤ Статистичну оцінку θ^* параметра θ називають *незсунutoю*, якщо

$$M(\theta^*) = \theta.$$

Вимога про незсунутість оцінки θ^* є недостатньою тому, що можливі значення θ^* можуть бути сильно розсіяні навколо свого середнього значення, а отже дисперсія $D(\theta^*)$ може бути великою. Тоді знайдена за даними однієї вибірки оцінка, наприклад, θ_k^* може набагато відрізнятись від середнього значення θ^* , отже і від параметра θ .

Якщо $D(\theta^*)$ буде малою, тоді можливість допустити велику помилку буде виключена.

Отже, до статистичної оцінки виникає вимога про її ефективність.

- **Ефективною** називають таку статистичну оцінку θ^* , яка при заданому об'ємі вибірки n має найменшу можливу дисперсію.

При розгляді вибірки великого об'єму ($n \rightarrow \infty$) до статистичних оцінок пред'являють вимогу їх обґрунтованості.

- **Обґрунтованою** називають статистичну оцінку, яка при $n \rightarrow \infty$ прямує за імовірністю до оцінюваного параметра.

Наприклад, якщо дисперсія незсунутої оцінки при $n \rightarrow \infty$ прямує до нуля, то оцінка буде і обґрунтованою.

Поняття про точкові оцінки параметрів розподілу

- **Точковими оцінками параметрів розподілу генеральної сукупності** називають такі оцінки, які визначаються одним числом.

Наприклад, вибіркова середня \bar{x} , вибіркова дисперсія D_B – точкові оцінки відповідних числових характеристик генеральної сукупності.

Точкові оцінки параметрів розподілу є випадковими величинами, їх можна вважати первинними результатами обробки вибірки оскільки невідомо, з якою точністю кожна з них оцінює відповідну числову характеристику генеральної сукупності.

Якщо об'єм вибірки досить великий, то точкові оцінки задовольняють практичні потреби точності.

Оцінка математичного сподівання $M(X)$ по вибірковій середній \bar{x}

Нехай з генеральної сукупності в результаті незалежних спостережень за кількісною ознакою X отримано вибірку обсягом n із значеннями ознаки x_1, \dots, x_n (вважаємо, що всі вони різні). Нехай потрібно оцінити математичне сподівання ознаки X за даними вибірки.

Спробуємо в якості оцінки взяти вибіркову середню

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Перевіримо, чи є ця оцінка незсунутою.

Будемо розглядати \bar{x} як випадкову величину \bar{X} і x_1, \dots, x_n як незалежні, однаково розподілені випадкові величини X_1, \dots, X_n .

Математичне сподівання середнього арифметичного однаково розподілених величин дорівнює математичному сподіванню кожної з них.

$$M(\bar{X}) = M\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n M(X_i) = M(X)$$

Отже, оцінка незсунута.

Дослідимо оцінку на ефективність.

$$D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \frac{1}{n} D(X)$$

Отже, оцінка ефективна.

$$\lim_{n \rightarrow \infty} D(\bar{X}) = \lim_{n \rightarrow \infty} \left(\frac{1}{n} D(X)\right) = 0$$

Отже, оцінка обґрунтована.

Оцінка дисперсії $D(X)$ по вибірковій дисперсії D_B

Нехай з генеральної сукупності в результаті незалежних спостережень за кількісною ознакою X отримано повторну вибірку обсягом n із значеннями ознаки x_1, \dots, x_n (вважаємо, що всі вони різні).

Потрібно оцінити дисперсію ознаки X за даними вибірки.

Спробуємо в якості оцінки взяти вибіркову дисперсію

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

Можна довести, що вибіркова дисперсія дає занижені значення для дисперсії $D(X)$ генеральної сукупності, тобто вона буде зсунутою оцінкою $D(X)$. При цьому математичне сподівання D_B буде

$$M(D_B) = \frac{n-1}{n} D(X)$$

Тому вибіркoву дисперсію доцільно виправити таким чином, щоб вона стала незсунутою оцінкою.

➤ **Виправлену вибіркoву дисперсію** позначають

$$s^2 = \frac{n-1}{n} D_B = \frac{1}{n-1} \sum_{i=1}^m n_i (x_i - \bar{x})^2$$

➤ Тоді **виправленим середньоквадратичним відхиленням** вибірки буде

$$s = \sqrt{s^2}$$

При досить великих об'ємах вибірки вибіркoва дисперсія D_B та виправлена вибіркoва дисперсія s^2 різняться дуже мало. Тому в практичних задачах виправлену дисперсію s^2 та виправлене середньоквадратичне відхилення вибірки s використовують лише при об'ємі вибірки $n < 30$.

Для оцінки інших параметрів розподілу зручніше не підбирати підходящі статистики (з подальшою їх перевіркою на незсунутість, ефективність та обґрунтованість), а скористатися одним з наступних методів.

Для першого з них визначимо аналогічно початковому та центральному моментам розподілу із теорії імовірностей деякі числові характеристики вибірки.

➤ **Моментом порядку k** називають середнє значення k -го степеня різниці $x_i - C$.

При $C = 0$ одержимо початковий момент порядку k вибірки

$$v_k^\bullet = \frac{1}{n} \sum_{i=1}^n x_i^k$$

При $C = \bar{x}$ одержимо центральний момент порядку k вибірки

$$\mu_k^\bullet = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

Метод моментів для визначення точкових оцінок

Метод запропоновано К.Пірсоном.

Метод заснований на тому, що початкові та центральні емпіричні моменти є обґрунтованими оцінками відповідних початкових та центральних теоретичних моментів того ж порядку.

Перевага методу – простота.

Нехай випадкова величина задана щільністю розподілу $f(x, \theta_1, \dots, \theta_r)$, де $\theta_1, \dots, \theta_r$ - невідомі параметри розподілу. Потрібно знайти точкові оцінки $\theta_1^*, \dots, \theta_r^*$ цих параметрів, для чого необхідно сформулювати r незалежних умов.

Суть методу моментів полягає в тому, що такі умови отримують прирівнюванням довільних r теоретичних моментів (початкових або центральних) відповідним їм емпіричним моментам, які визначаються за вибіркою.

$$\begin{aligned} v_k(\theta_1, \dots, \theta_r) &= v_k^\bullet, k = \overline{1, m} \\ \mu_s(\theta_1, \dots, \theta_r) &= \mu_s^\bullet, s = \overline{m+1, r} \end{aligned}$$

Вимоги до рівнянь системи:

1. Рівняння повинні бути інформативними (не можна використовувати

$$v_0, \mu_0, \mu_1);$$

2. Рівняння повинні бути незалежними (якщо використано ν_1, ν_2 , то не можна використовувати μ_2 , бо $\mu_2 = \nu_2 - \nu_1^2$).

Можна показати, що отримані оцінки будуть більш ефективними, якщо для формування системи рівнянь використовувати моменти більш низьких порядків.

Оцінка одного параметра

Нехай задано вигляд щільності розподілу $f(x, \theta)$, що визначається одним невідомим параметром θ . Потрібно знайти точкову оцінку цього параметра.

Для оцінки одного параметра достатньо мати одне рівняння відносно цього параметра. Тому прирівнюємо, наприклад, початковий теоретичний момент першого порядку $\nu_1 = M(X)$ до початкового емпіричного моменту першого порядку $\nu_1^* = \bar{x}$:

$$\begin{aligned} \nu_1 = M(X) &= \int_{-\infty}^{\infty} x \cdot f(x, \theta) dx = \varphi(\theta) = \\ &= \nu_1^* = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \end{aligned}$$

Математичне сподівання є функція від θ , тому вираз можна розглядати як рівняння з одним невідомим, розв'язавши яке відносно θ знайдемо його точкову оцінку θ^* , яка є функцією від вибіркової середньої, а отже і від варіант вибірки:

$$\theta^* = \psi(x_1, x_2, \dots, x_n)$$

Приклад. Знайти методом моментів за вибіркою x_1, \dots, x_n точкову оцінку невідомого параметра λ показникового розподілу, щільність розподілу якого $f(x) = \lambda e^{-\lambda x}$ ($x \geq 0$)

Розв'язок. Прирівнюємо теоретичний та емпіричний початкові моменти першого порядку

$$M(X) = \bar{x}$$

Прийнявши до уваги, що математичне сподівання показникового розподілу дорівнює $1/\lambda$, маємо

$$1/\lambda = \bar{x}$$

звідки шукана точкова оцінка параметра $\lambda^* = 1/\bar{x}$.

Оцінка двох параметрів

Нехай задано вигляд щільності розподілу $f(x, \theta_1, \theta_2)$, що визначається двома невідомими параметрами θ_1, θ_2 . Потрібно знайти точкові оцінки цих параметрів.

Для оцінки достатньо мати два рівняння відносно цих параметрів. Тому прирівнюємо початковий теоретичний момент першого порядку до початкового емпіричного моменту першого порядку та центральний теоретичний момент другого порядку до центрального емпіричного моменту другого порядку

$$\begin{cases} M(X) = \bar{x} \\ D(X) = D_B \end{cases}$$

Математичне сподівання та дисперсія є функціями від θ_1, θ_2 , тому систему можна розглядати як систему рівнянь з двома невідомими, розв'язавши яку відносно θ_1, θ_2 знайдемо їх точкові оцінки θ_1^*, θ_2^* , які є функцією від вибіркової середньої та вибіркової дисперсії, а отже і від варіант вибірки:

$$\theta_1^* = \psi_1(x_1, x_2, \dots, x_n)$$

$$\theta_2^* = \psi_2(x_1, x_2, \dots, x_n)$$

Приклад. Знайти методом моментів за вибіркою x_1, \dots, x_n точкові оцінки невідомих параметрів a та σ нормального розподілу, щільність якого

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)}$$

Розв'язок. Прирівнюємо теоретичний та емпіричний початкові моменти першого порядку і центральні моменти другого порядку

$$M(X) = \bar{x}, \quad D(X) = D$$

Прийнявши до уваги, що математичне сподівання нормального розподілу дорівнює a , а дисперсія σ^2 , маємо

$$a^* = \bar{x}, \quad \sigma^* = \sqrt{D_B}$$

Метод найбільшої подібності для визначення точкових оцінок

Метод запропоновано Р.Фішером.

Метод найбільшої правдоподібності має ряд переваг:

- оцінки найбільшої правдоподібності, взагалі кажучи, обґрунтовані (але вони можуть бути зсунутими), розподілені асимптотично нормально (при великих значеннях n приблизно нормальні) і мають найменшу дисперсію в порівнянні з іншими асимптотично нормальними оцінками;
- якщо для оцінюваного параметра θ існує ефективна оцінка θ^* , то рівняння правдоподібності має єдине рішення θ^* ;
- цей метод найповніше використовує дані вибірки про оцінюваний параметр, тому він особливо корисний у разі малих вибірок.

Недолік методу полягає в тому, що він часто вимагає складних обчислень.

Метод найбільшої подібності для дискретної випадкової величини

Нехай X – дискретна випадкова величина, яка в результаті n випробувань набувала значень x_1, \dots, x_n . Припустимо, що вигляд закону розподілу випадкової величини задано, але невідомо параметр θ , яким визначається цей закон. Потрібно знайти його точкову оцінку.

Позначимо через $p(x_i, \theta)$ імовірність того, що в результаті випробування випадкова величина X прийме значення x_i ($i=1, 2, \dots, n$).

➤ **Функцією правдоподібності дискретної випадкової величини X** називають функцію аргументу θ , де x_1, \dots, x_n – варіанти вибірки.

В якості точкової оцінки параметра θ приймають таке його значення $\theta^* = \theta^*$ (x_1, \dots, x_n), при якому функція правдоподібності досягає максимуму.

➤ Оцінку θ^* називають **оцінкою найбільшої правдоподібності**.

Зауваження:

✓ Функція правдоподібності – це функція від аргументу θ , оцінка найбільшої правдоподібності – це функція від незалежних аргументів x_1, \dots, x_n .

Оцінка найбільшої правдоподібності не завжди співпадає з оцінкою, знайденою методом моментів.

Функції L та $\ln L$ досягають максимуму при одному й тому ж значенні θ , тому замість пошуку максимуму функції L шукають (що зручніше) максимум функції $\ln L$.

➤ **Логарифмічною функцією правдоподібності** випадкової величини X називають функцію $\ln L$.

Точку максимуму функції правдоподібності можна знайти так:

1. Перейти від функції L до $\ln L$.

2. Знайти першу похідну $\frac{d \ln L}{d\theta}$
3. Прирівняти похідну до нуля та знайти критичну точку – корінь отриманого рівняння (його називають рівнянням правдоподібності)
4. Знайти другу похідну $\frac{d^2 \ln L}{d\theta^2}$. Якщо друга похідна при $\theta = \theta^*$ від'ємна, то θ^* – точка максимуму.
5. Знайдену точку максимуму θ^* прийняти в якості оцінки найбільшої правдоподібності параметра θ .

Приклад. Знайти методом найбільшої правдоподібності оцінку параметра λ розподілу Пуассона

$$P_m(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

де m – число проведених випробувань, x_i – число появ події в i -тому ($i=1,2,\dots,n$) досліді (дослід складається з m випробувань).

Розв'язок. Побудуємо функцію правдоподібності, врахувавши, що $\theta = \lambda$

$$\begin{aligned} L &= p(x_1, \lambda) \cdot p(x_2, \lambda) \cdot \dots \cdot p(x_n, \lambda) = \\ &= \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \cdot \dots \cdot \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} = \\ &= \frac{\lambda^{\sum x_i} \cdot e^{-n\lambda}}{x_1! \cdot x_2! \cdot \dots \cdot x_n!} \end{aligned}$$

Знайдемо логарифмічну функцію

$$\ln L = (\sum x_i) \ln \lambda - n\lambda - \ln(x_1! \cdot x_2! \cdot \dots \cdot x_n!)$$

Знайдемо першу похідну

$$\frac{d \ln L}{d\lambda} = \frac{\sum x_i}{\lambda} - n$$

Запишемо рівняння правдоподібності

$$0 = \frac{\sum x_i}{\lambda} - n$$

та знайдемо критичну точку

$$\lambda = \frac{\sum x_i}{n} = \bar{x}$$

Перевіримо, чи це максимум

$$\frac{d^2 \ln L}{d\lambda^2} = -\frac{\sum x_i}{\lambda^2}; \quad -\frac{\sum x_i}{(\bar{x})^2} < 0$$

Отже, оцінка найбільшої правдоподібності параметра λ розподілу Пуассона

$$\lambda^* = \bar{x}$$

Метод найбільшої подібності для неперервної випадкової величини

Нехай X – неперервна випадкова величина, яка в результаті n випробувань набувала значень x_1, \dots, x_n . Вигляд функції щільності розподілу $f(x)$ випадкової величини задано, але невідомо параметр θ , яким визначається ця функція.

Потрібно знайти його точкову оцінку.

➤ **Функцією правдоподібності неперервної випадкової величини X** називають функцію аргументу θ , де x_1, \dots, x_n – варіанти вибірки.

$$L(x_1, x_2, \dots, x_n, \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \dots \cdot f(x_n, \theta)$$

Оцінку найбільшої правдоподібності невідомого параметра розподілу неперервної випадкової величини шукають так само, як і у випадку дискретної величини.

Приклад. Знайти методом найбільшої правдоподібності оцінку параметра λ показникового розподілу

$$f(x) = \lambda e^{-\lambda x}, \quad (0 < x < \infty)$$

якщо в результаті n проведених випробувань випадкова величина X набула значень x_1, \dots, x_n .

Розв'язок. Побудуємо функцію правдоподібності, врахувавши, що $\theta = \lambda$

$$\begin{aligned} L &= f(x_1, \lambda) \cdot f(x_2, \lambda) \cdot \dots \cdot f(x_n, \lambda) = \\ &= \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdot \dots \cdot \lambda e^{-\lambda x_n} = \lambda^n e^{-\lambda \sum x_i} \end{aligned}$$

$$\ln L = n \ln \lambda - \lambda \sum x_i$$

Знайдемо першу похідну

$$\frac{d \ln L}{d \lambda} = \frac{n}{\lambda} - \sum x_i$$

Запишемо рівняння правдоподібності

$$0 = \frac{n}{\lambda} - \sum x_i$$

та знайдемо критичну точку

$$\lambda = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$$

Перевіримо, чи це максимум

$$\frac{d^2 \ln L}{d \lambda^2} = -\frac{n}{\lambda^2}; \quad -\frac{n}{(\bar{x})^2} < 0$$

Отже, оцінка найбільшої правдоподібності параметра λ показникового розподілу

$$\lambda^* = 1/\bar{x}$$

Якщо щільність розподілу неперервної випадкової величини X визначається двома невідомими параметрами θ_1 та θ_2 , то функція правдоподібності є функцією двох незалежних аргументів θ_1, θ_2

$$L(x_1, x_2, \dots, x_n, \theta_1, \theta_2) = f(x_1, \theta_1, \theta_2) \cdot f(x_2, \theta_1, \theta_2) \cdot \dots \cdot f(x_n, \theta_1, \theta_2)$$

де x_1, \dots, x_n – спостережувані значення X .

Далі знаходять логарифмічну функцію правдоподібності і для пошуку її максимуму складають та розв'язують систему

$$\begin{cases} \frac{\partial \ln L}{\partial \theta_1} = 0 \\ \frac{\partial \ln L}{\partial \theta_2} = 0 \end{cases}$$

Приклад. Знайти методом найбільшої правдоподібності точкові оцінки невідомих параметрів a та σ нормального розподілу

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)}$$

якщо в результаті n проведених випробувань випадкова величина X набула значень x_1, \dots, x_n .

Розв'язок. Побудуємо функцію правдоподібності, врахувавши, що $\theta_1 = a$, $\theta_2 = \sigma$

$$\begin{aligned} L &= f(x_1, a, \sigma) \cdot f(x_2, a, \sigma) \cdot \dots \cdot f(x_n, a, \sigma) = \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_1-a)^2/2\sigma^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_2-a)^2/2\sigma^2} \dots \times \\ &\quad \times \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_n-a)^2/2\sigma^2} = \\ &= \frac{1}{\sigma^n (\sqrt{2\pi})^n} e^{-\sum (x_i-a)^2/2\sigma^2} \end{aligned}$$

Побудуємо логарифмічну функцію правдоподібності

$$\ln L = -n \ln \sigma + \ln \frac{1}{(\sqrt{2\pi})^n} - \frac{\sum (x_i - a)^2}{2\sigma^2}$$

Знайдемо частинні похідні по a та σ

$$\frac{\partial \ln L}{\partial a} = \frac{\sum x_i - na}{\sigma^2}; \quad \frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum (x_i - a)^2}{\sigma^3}$$

Прирівнявши частинні похідні до нуля та розв'язавши отриману систему рівнянь відносно a та σ^2 , отримаємо

$$a^* = \frac{\sum x_i}{n} = \bar{x}; \quad \sigma^* = \sqrt{\left(\sum (x_i - \bar{x})^2\right)/n}$$

Підсумки

- Точкова оцінка - це оцінка, визначена одним числом.
- Щоб точкова оцінка була коректною, вона має бути незсунутою, ефективною і обґрунтованою.
- Є кілька методів визначення точкової оцінки, у кожного з яких свої переваги та недоліки.

3.5. Інтервальні оцінки параметрів розподілу

Надійність оцінки

Якщо об'єм вибірки малий, то точкові оцінки можуть давати значні похибки, тому питання точності оцінок у цьому випадку дуже важливе.

- **Інтервальною** називають оцінку, яка визначається двома числами — кінцями інтервалу.

Інтервальні оцінки дозволяють встановити точність та надійність оцінок.

Нехай знайдена за даними вибірки статистична оцінка θ^* буде оцінкою невідомого параметра θ .

Ясно, що оцінка θ^* тим точніше визначає параметр θ , чим менше абсолютна величина різниці $\theta - \theta^*$. Іншими словами, якщо $\delta > 0$ і $|\theta - \theta^*| < \delta$, тоді меншому δ відповідає більш точна оцінка. Тому число δ характеризує точність оцінки.

Але статистичні методи не дозволяють категорично стверджувати, що оцінка θ^* задовольняє нерівність $|\theta - \theta^*| < \delta$. Таке твердження можна зробити лише з імовірністю $P_{\text{дов}} = \gamma$.

- **Надійністю (довірчою імовірністю) оцінки параметра θ за θ^*** називають імовірність

$$\gamma = P(|\theta - \theta^*| < \delta)$$

з якою виконується нерівність $|\theta - \theta^*| < \delta$.

Найчастіше число $P_{\text{дов}} = \gamma$ задається наперед і, залежно від обставин, воно дорівнює 0.95, або 0.99, або 0.999.

Формулу $\gamma = P(|\theta - \theta^*| < \delta)$ можна записати у вигляді

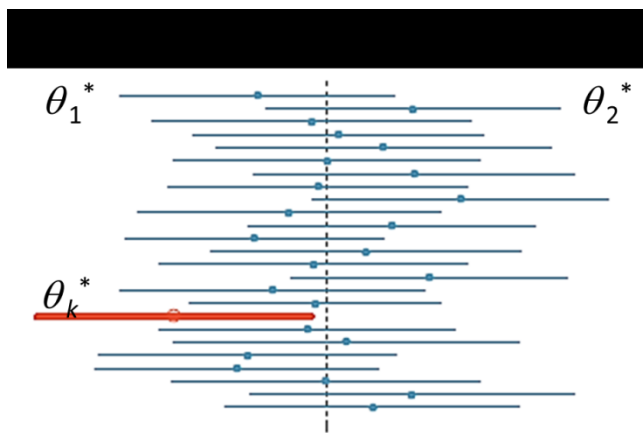
$$P(\theta^* - \delta < \theta < \theta^* + \delta) = \gamma$$

З цієї рівності випливає, що інтервал $(\theta^* - \delta, \theta^* + \delta)$ містить невідомий параметр θ генеральної сукупності.

➤ **Інтервал** $(\theta^* - \delta, \theta^* + \delta)$ називають **довірчим**, якщо він покриває невідомий параметр θ із заданою надійністю $P_{\text{дов}} = \gamma$.

Кінці довірчого інтервалу є випадковими величинами.

Якщо $P_{\text{дов}} = 0.95$, то в 95% випадків інтервал $(\theta^* - \delta, \theta^* + \delta)$ покриває невідомий параметр θ .



Загальна методика визначення інтервальних оцінок

1. Необхідно вибрати підходящу статистику Y для побудови довірчого інтервалу.

Y – підходяща статистика, якщо:

- містить параметр θ ;
- містить оцінку параметра θ^* ;
- якщо містить інші параметри β , то вони відомі;
- має відомий ЗРІ (бажано табличний)

$$Y = Y(\theta^*, \theta, \beta)$$

2. За вибраним (заданим) $P_{\text{дов}}$ та законом розподілу статистики Y визначити довірчий інтервал (y_1, y_2) , в який потрапить статистика Y з імовірністю $P_{\text{дов}}$ при виконанні експерименту.
3. За виразом $Y=Y(\theta^*, \theta, \beta)$ знайти в загальному вигляді інтервал (θ_1^*, θ_2^*) , в який потрапить θ^* при виконанні експерименту.
4. Виконати експеримент, за вибіркою $\{x\}_n$ визначити емпіричне значення $y_{\text{ем}}$ статистики Y та підставити його в п.3.



Довірчий інтервал для оцінки математичного сподівання нормального розподілу при відомому σ

Нехай кількісна ознака X генеральної сукупності розподілена за нормальним законом, середньоквадратичне відхилення σ відомо.

Треба знайти довірчий інтервал, що покриває математичне сподівання $\mu = a$ генеральної сукупності із заданою надійністю $P_{\text{дов}} = \gamma$.

В якості підходящої статистики для побудови довірчого інтервалу можна використати Z -статистику

$$Z = Z(\bar{X}, a, \sigma_{\bar{X}}) = \frac{\bar{X} - a}{\sigma_{\bar{X}}} = \frac{(\bar{X} - a)\sqrt{n}}{\sigma_X}$$

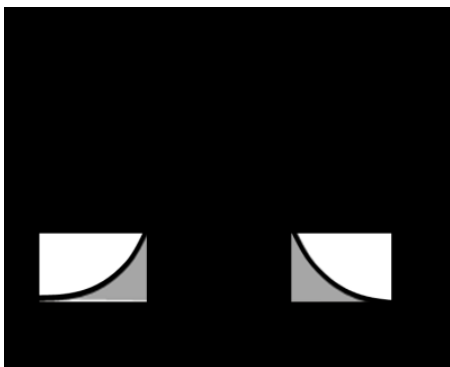
Z має нормальний розподіл з параметрами $f(z, 0, 1)$, тобто це таблична функція Лапласа.

За заданим $P_{\text{дов}}$ та законом розподілу статистики Z визначаємо довірчий інтервал (z_1, z_2) , в який потрапить статистика Z з імовірністю $P_{\text{дов}}$ при виконанні експерименту.

$$z_1 < Z < z_2$$

$$z_1 < \frac{(\bar{X} - a)\sqrt{n}}{\sigma_X} < z_2$$

Закон розподілу статистики Z - симетрична функція, тому довірчий інтервал буде найвужчим при $z_1 = -z_2$.



Виконуємо експеримент, за вибіркою $\{x\}_n$ визначаємо емпіричне значення $\bar{X} = \bar{x}$ та підставимо його в нерівність.

$$\bar{x} - \frac{z_{\gamma/2}\sigma_X}{\sqrt{n}} < a < \bar{x} + \frac{z_{\gamma/2}\sigma_X}{\sqrt{n}}$$

Отриманий результат відповідає здоровому глузду:

- при збільшенні $P_{\text{дов}}$ довірчий інтервал розширюється, отже, збільшення надійності оцінки зменшує її точність.
- при збільшенні обсягу інформації ($n \uparrow$) та поліпшенні її якості ($\sigma_X \downarrow$) довірчий інтервал звужується.

Приклад. Випадкова величина розподілена за нормальним законом з параметром $\sigma = 3$. Зроблена вибірка об'єму $n = 36$. З надійністю $\gamma = 0.95$ знайти довірчий інтервал невідомого параметра a цього розподілу.

Розв'язок. В даному випадку підходить Z-статистика. Довірчий інтервал буде найвужчим при $z_1 = -z_2$, тому з рівності

$$\Phi(z) = \gamma/2 \Rightarrow \Phi(z) = 0.475$$

З таблиці інтегральної функції Лапласа Φ знайдемо число $z = 1.96$. Тоді точність оцінки буде

$$\frac{z \cdot \sigma}{\sqrt{n}} = \frac{3 \cdot 1.96}{\sqrt{36}} = 0.98$$

Отже, довірчий інтервал буде $(\bar{x} - 0.98; \bar{x} + 0.98)$

Наприклад, якщо $\bar{x} = 4.1$, то з надійністю 95% інтервал $(3.12; 5.08)$ покриває параметр a .

Знаходження об'єму вибірки, необхідного для розрахунку інтервальних оцінок із заданою надійністю

Нехай ознака X генеральної сукупності розподілена за нормальним законом з відомим параметром σ . Треба знайти об'єм вибірки n , який із заданою точністю δ та надійністю γ дозволить знайти оцінку параметра a .

Оскільки при таких вихідних умовах для оцінки довірчого інтервалу використовується Z-статистика, то і обсяг вибірки слід шукати з її допомогою. Для надійності γ , використовуючи $\Phi(z) = \gamma/2$ (бо довірчий інтервал буде найвужчим при $z_1 = -z_2$) та таблицю значень інтегральної функції Лапласа, знайдемо відповідне число z .

$$\text{Із формули } z = \frac{\delta \cdot \sqrt{n}}{\sigma} \text{ одержуємо } n = \frac{z^2 \sigma^2}{\delta^2}$$

Тепер z , δ та σ відомі, тому можна знайти потрібний об'єм вибірки.

Приклад. Випадкова величина розподілена за нормальним законом з параметром σ . Знайти мінімальний об'єм n вибірки, щоб з надійністю γ та точністю δ виконувалась рівність $\bar{x} = a$, якщо $\sigma = 0.5$; $\gamma = 0.95$; $\delta = 0.1$.

Розв'язок. Для $\gamma = 0.95$ маємо

$$\Phi(z) = 0.457 \Rightarrow z = 1.96$$

Використовуючи знайдене z та задані σ , δ , одержуємо

$$n = \left(\frac{1.96 \cdot 0.5}{0.1} \right)^2 = (9.8)^2 = 96.04$$

Отже, мінімальний об'єм вибірки $n = 96$.

Довірчий інтервал для оцінки математичного сподівання нормального розподілу при невідомому σ

Нехай кількісна ознака X генеральної сукупності розподілена за нормальним законом, середньоквадратичне відхилення σ невідомо.

Треба знайти довірчий інтервал, що покриває математичне сподівання a генеральної сукупності із заданою надійністю γ .

В якості підходящої статистики для побудови довірчого інтервалу можна використати випадкову величину T , що має розподіл Стюдента з $k = n-1$ степенями свободи

$$T = \frac{\bar{X} - a}{S / \sqrt{n}}$$

\bar{X} – вибіркова середня, n – обсяг вибірки, S – виправлене середньоквадратичне відхилення. Це таблична функція, що має щільність розподілу

$$T(t, n) = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)} \cdot \Gamma((n-1)/2)} \left[1 + \frac{t^2}{n-1} \right]^{-n/2}, \text{ де } \Gamma(x) = \int_0^{\infty} y^{x-1} \cdot e^{-y} dy - \text{гама-функція.}$$

За заданим $P_{\text{дов}}$ та законом розподілу статистики T визначаємо довірчий інтервал (t_1, t_2) , в який потрапить статистика T з імовірністю $P_{\text{дов}}$ при виконанні експерименту. Закон розподілу статистики T - симетрична функція, тому довірчий інтервал буде найвужчим при $t_1 = t_2$.

$$P\left(\frac{|\bar{X} - a|\sqrt{n}}{s} < t_\gamma\right) = 2 \int_0^{t_\gamma} T(t, n) dt = P_{\text{дов}} = \gamma$$

Знаходимо в загальному вигляді інтервал

$$\bar{X} - \frac{t_\gamma S}{\sqrt{n}} < a < \bar{X} + \frac{t_\gamma S}{\sqrt{n}}$$

Виконуємо експеримент, за вибіркою $\{x\}_n$ визначаємо емпіричне значення

$\bar{X} = \bar{x}$ та $S = s$ підставимо його

$$\bar{x} - \frac{t_\gamma s}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma s}{\sqrt{n}}$$

З граничних співвідношень слідує, що при зростанні обсягу вибірки n розподіл Стюдента прямує до нормального. Тому практично при $n > 30$ можна замість розподілу Стюдента користуватись нормальним розподілом.

Приклад. Випадкова величина X розподілена за нормальним законом. За вибіркою об'єму $n = 16$, знайдено вибіркиму середню $\bar{x} = 20.2$ та виправлене середньоквадратичне відхилення $s = 0.8$. Оцінити невідоме математичне сподівання за допомогою довірчого інтервалу з надійністю $\gamma = 0.95$.

Розв'язок. По $\gamma = 0.95$ та $n=16$ з таблиці розподілу Стюдента знайдемо число $t = 2.13$. Тоді границі довірчого інтервалу будуть

$$20.2 - 2.13 \cdot 0.8 / \sqrt{16} = 19.774$$

$$20.2 + 2.13 \cdot 0.8 / \sqrt{16} = 20.626$$

Отже, з надійністю 95% невідомий параметр μ знаходиться в довірчому інтервалі (19.774; 20.626).

Довірчий інтервал для оцінки середньоквадратичного відхилення σ нормального розподілу

Нехай кількісна ознака X генеральної сукупності розподілена за нормальним законом.

Треба знайти довірчий інтервал, що покриває середньоквадратичне відхилення σ генеральної сукупності із заданою надійністю $P_{\text{дов}} = \gamma$.

Оскільки треба оцінити невідоме середньоквадратичне відхилення σ генеральної сукупності по виправленому вибірковому середньоквадратичному відхиленню s , потрібно щоб виконувалося співвідношення

$$P(|S - \sigma| < \delta) = \gamma \quad \text{або} \quad P(S - \delta < \sigma < S + \delta) = \gamma$$

В якості підходящої статистики для побудови довірчого інтервалу беремо χ^2 -статистику

$$\chi^2 = \sum_{i=1}^n Z_i^2$$

де Z_i ($i = 1, 2, \dots, n$) – нормальні, нормовані незалежні величини, тобто їх математичне сподівання дорівнює нулю, середнє квадратичне відхилення дорівнює одиниці і кожна з них розподілена за нормальним законом.

Ця статистика розподілена по закону χ^2 з $k = n-1$ степенями свободи.

В загальному випадку маємо справу не з нормованими величинами, тому запишемо χ^2 -статистику у вигляді

$$\chi^2 = \sum_{i=1}^n \frac{(X_i - MX)^2}{\sigma^2}$$

Вибіркову χ^2 – статистику отримаємо наступним чином. Виправлена вибіркова дисперсія

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - MX)^2$$

Домножимо на $(n-1)/\sigma^2$

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - MX}{\sigma} \right)^2 = \sum_{i=1}^n Z_i^2 = \chi_n^2$$

За заданим $P_{\text{дов}}$ та законом розподілу статистики χ^2 визначаємо довірчий інтервал $\chi_1 < \chi < \chi_2$, в який потрапить статистика χ^2 з імовірністю $P_{\text{дов}}$ при виконанні експерименту. Закон розподілу статистики χ^2 - несиметрична функція.



Знаходимо в загальному вигляді інтервал

$$\sqrt{\frac{S^2(n-1)}{\chi_2^2}} < \sigma < \sqrt{\frac{S^2(n-1)}{\chi_1^2}}$$

Щоб можна було користуватися більш зручною таблицею (значення q , а не χ^2), перетворимо нерівність у рівносильну їй

$$S - \delta < \sigma < S + \delta \Rightarrow S(1 - \delta/S) < \sigma < S(1 + \delta/S)$$

поклавши $q = \delta/S$ отримаємо $S(1 - q) < \sigma < S(1 + q)$

Імовірність того, що нерівність

$$\frac{\sqrt{n-1}}{1+q} < \chi < \frac{\sqrt{n-1}}{1-q} \text{ а отже і рівносильна їй } S(1-q) < \sigma < S(1+q)$$

виконується, дорівнює $\int_{\sqrt{n-1}/(1+q)}^{\sqrt{n-1}/(1-q)} R(\chi, n) d\chi = \gamma$

З цього рівняння можна за заданими n та γ знайти q .

Практично для знаходження q користуються таблицею.

Виконуємо експеримент, за вибіркою $\{x\}_n$ визначаємо емпіричне значення $S = s$ та підставимо його в нерівність

$$s(1 - q) < \sigma < s(1 + q) \text{ або } \sqrt{\frac{s^2(n-1)}{\chi_2^2}} < \sigma < \sqrt{\frac{s^2(n-1)}{\chi_1^2}}$$

Приклад. Випадкова величина X розподілена за нормальним законом. За вибіркою об'єму $n = 25$, знайдено виправлене середньоквадратичне відхилення $s = 0.8$. Оцінити невідоме середньоквадратичне відхилення σ за допомогою довірчого інтервалу з надійністю $\gamma = 0.95$.

Розв'язок. По $\gamma = 0.95$ та $n=25$ з таблиці розподілу χ^2 знайдемо число $q = 0.32$. Тоді границі довірчого інтервалу будуть

$$0.8(1 - 0.32) < \sigma < 0.8(1 + 0.32)$$

Отже, з надійністю 95% невідомий параметр σ знаходиться в довірчому інтервалі (0.544; 1.056).

Приклад. По 15 рівноточним вимірам знайдено виправлене середньоквадратичне відхилення $s = 0.12$. Знайти точність вимірювань з надійністю $\gamma = 0.99$.

Розв'язок. Точність вимірювань характеризується середньоквадратичним відхиленням σ випадкових помилок, тому задача зводиться до пошуку довірчого інтервалу, що покриває σ з заданою надійністю.

По $\gamma = 0.99$ та $n=15$ з таблиці розподілу χ^2 знайдемо число $q = 0.73$. Тоді границі довірчого інтервалу будуть

$$0.12(1 - 0.73) < \sigma < 0.12(1 + 0.73)$$

$$0.03 < \sigma < 0.21$$

Підсумки

- Якщо об'єм вибірки малий, то точкові оцінки можуть давати значні похибки, тому виникає потреба знати наскільки цій оцінці можна вірити. Інтервальні оцінки дозволяють встановити точність та надійність оцінок.
- При збільшенні $P_{\text{дов}}$ довірчий інтервал розширюється, отже, збільшення надійності оцінки зменшує її точність.
- При збільшенні обсягу інформації ($n \uparrow$) та поліпшенні її якості ($\sigma_x \downarrow$) довірчий інтервал звужується.
- В залежності від обсягу апіорної інформації про випадкову величину для розрахунку довірчого інтервалу використовують різні статистики.

3.6. Статистичні гіпотези

Різновиди статистичних гіпотез

Вивідна статистика – це методологія, що дозволяє охарактеризувати генеральну сукупність або сформулювати міркування про неї спираючись на інформацію, отриману з вибірки, що зроблена з цієї генеральної сукупності. Більша частина практичної діяльності в області статистики пов'язана саме з статистичним виведенням.

- **Статистичними** називають гіпотези про генеральну сукупність, які висуваються та перевіряються на основі даних вибірки.

Наприклад, статистичними будуть гіпотези:

- генеральна сукупність розподілена за нормальним законом;
- дисперсії двох сукупностей, розподілених за законом Пуассона, рівні між собою.

У багатьох задачах необхідно знати закон розподілу генеральної сукупності. Якщо закон розподілу невідомий, але є міркування для припущення щодо його певного вигляду A , наприклад, розподіл рівномірний, показниковий або нормальний, тоді висувають гіпотезу: *генеральна сукупність розподілена за законом A* . У цій гіпотезі йде мова про вигляд невідомого розподілу.

Іноді закон розподілу генеральної сукупності відомий, але його параметри (числові характеристики) невідомі. Якщо є міркування припустити, що невідомий параметр θ дорівнює певному значенню θ_0 то висувають гіпотезу: $\theta = \theta_0$. Ця гіпотеза вказує припущену величину параметра відомого розподілу.

Можливі також інші гіпотези: про рівність параметрів двох різних розподілів, про незалежність вибірок та багато інших.

За своїм прикладним змістом статистичні гіпотези можна поділити на декілька основних типів:

- про однорідність вибірок (тобто належності їх одній і тій же генеральній сукупності);
- про закон розподілу;
- про рівність числових характеристик генеральних сукупностей;
- про числові значення параметрів;
- про стохастичну незалежність елементів вибірки.

Разом з основною гіпотезою завжди можна розглядати протилежну їй гіпотезу. Якщо основна гіпотеза була відхилена, тоді має місце протилежна гіпотеза. Отже, ці гіпотези доцільно відрізняти.

➤ **Основною (нульовою)** називають припущену гіпотезу і позначають H_0 .

➤ **Альтернативною (конкурентною)** називають гіпотезу, що суперечить основній, її позначають H_1 .

Наприклад, якщо $H_0: M(X) = 6$, то $H_1: M(X) \neq 6$.

Гіпотези можуть містити тільки одне або більше одного припущення.

➤ Гіпотезу називають **простою**, якщо вона містить лише одне припущення.

Наприклад, якщо λ – параметр показникового розподілу $f(x) = \lambda e^{-\lambda x}$, то гіпотеза $H_0: \lambda = 5$ буде проста.

➤ Гіпотезу називають **складною**, якщо вона складається із скінченної або нескінченної кількості простих гіпотез.

Основну гіпотезу H_0 краще формулювати як просту – це полегшить її перевірку.

Похибки при перевірці гіпотез

Довільна запропонована статистична гіпотеза може бути правильною або неправильною, тому виникає необхідність її перевірки.

- Перевірка гіпотези здійснюється за даними вибірки, тобто статистичними методами. Тому перевірку гіпотези за даними вибірки називають **статистичною**.

При перевірці гіпотези можна зробити хибний висновок. При цьому можуть бути похибки першого та другого роду.

- Якщо за висновком буде відкинута правильна гіпотеза, то кажуть, що це **похибка першого роду**.
- Якщо за висновком буде прийнята неправильна гіпотеза, то кажуть, що це **похибка другого роду**.

Наслідки цих похибок можуть бути різними.

- ✓ При контролі якості продукції імовірність визнати нестандартними стандартні вироби називають **ризиком виробника**, а імовірність визнати придатними браковані вироби називають **ризиком споживача**.
- Імовірність здійснити похибку першого роду називають **рівнем значущості α** .

Найчастіше рівень значущості приймають рівним 0.05 або 0.01.

Якщо прийнято рівень значущості рівним 0.05, то це означає, що в п'яти випадках із 100 ризикуємо одержати похибку першого роду (відкинути правильну гіпотезу).

Критерії узгодження

Перевірку статистичної гіпотези можна здійснити лише з використанням даних вибірки. Для цього слід обрати деяку випадкову статистичну характеристику (вибіркову функцію), точний або наближений розподіл якої відомий, і за допомогою цієї характеристики перевірити основну гіпотезу.

- *Статистичним критерієм узгодження перевірки гіпотези* (або просто критерієм) називають випадкову величину K , розподіл якої (точний або наближений) відомий і яка застосовується для перевірки основної гіпотези.

В означенні не враховується вид розподілу статистичної характеристики.

Якщо статистична характеристика розподілена нормально, то критерій позначають не літерою K , а літерами U або Z .

У випадку розподілу статистичної характеристики за законом Стюдента її позначають T , а у випадку закону «хі-квадрат» – χ^2 .

Якщо статистична характеристика розподілена за законом Фішера-Снедекора, то її позначають F .

Наприклад, для перевірки гіпотези про рівність дисперсії двох нормальних генеральних сукупностей за статистичну характеристику K вибирають відношення виправлених вибірових дисперсій

$$F = \frac{S_1^2}{S_2^2}$$

В різних дослідках дисперсія буде приймати різні, наперед невідомі значення, тому ця величина випадкова. Вона розподілена за законом Фішера-Снедекора. Існує багато інших критеріїв узгодження.

- **Спостереженим значенням критерію узгодження** називають значення відповідного критерію, обчислене за даними вибірки.

Наприклад, якщо за даними вибірок із двох нормальних генеральних сукупностей знайдено виправлені вибіркові дисперсії $s_1^2 = 18$ та $s_2^2 = 6$, тоді спостереженим значенням критерію узгодження буде

$$F_{cn} = \frac{18}{6} = 3$$

Різновиди статистичних критеріїв (за способом перевірки):

- **Параметричні критерії** - група статистичних критеріїв, які включають розрахунок параметрів імовірнісного розподілу ознаки (середнього і дисперсії)
- **Непараметричні критерії** - група статистичних критеріїв, які не включають в розрахунок параметрів імовірнісного розподілу і засновані на оперуванні частотами або рангами.

Непараметричні критерії найбільш прийнятні, коли обсяг вибірок малий. Якщо даних багато (наприклад, $n > 100$), то не має сенсу використовувати непараметричні статистики. Причина в тому, що коли вибірки стають дуже великими, то вибіркові середні підкоряються нормальному закону, навіть якщо вихідна змінна не є нормальною або виміряна з похибкою. Таким чином, параметричні методи, які є більш чутливими (мають велику статистичну потужність), завжди підходять для великих вибірок.

Проте, якщо вибірка мала, ці критерії слід використовувати тільки при наявності впевненості, що змінна дійсно має нормальний розподіл. Однак немає способу перевірити це припущення на малій вибірці. Використання критеріїв, заснованих на припущенні нормальності, крім того, обмежене шкалою вимірювань. Такі статистичні методи, як t-критерій, регресія і т. д.

припускають, що вихідні дані безперервні. Однак є ситуації, коли дані, скоріше, просто ранжовані (виміряні в порядкової шкалою), ніж виміряні точно. Для аналізу малих вибірок і для даних, виміряних в бідних шкалах, застосовують непараметричні методи. По суті, для кожного параметричного критерію є, принаймні, одна непараметрична альтернатива.

Всі критерії можна віднести до однієї з наступних груп:

- критерії відмінності між групами (незалежні вибірки);
- критерії відмінності між групами (залежні вибірки);
- критерії залежності між змінними.

Відмінності між незалежними групами. Зазвичай, коли є дві вибірки, які ви хочете порівняти щодо середнього значення деякої досліджуваної змінної, ви використовуєте t-критерій для незалежних вибірок. Непараметричними альтернативами цьому критерію є: U критерій Манна-Уїтні, двовибірковий критерій Колмогорова-Смирнова і критерій серій Валда-Волфовіца. Якщо ви маєте кілька груп, то можете використовувати дисперсійний аналіз. Його непараметричними аналогами є: ранговий дисперсійний аналіз Краскела-Уолліса і медіанний тест.

Відмінності між залежними групами. Якщо ви хочете порівняти дві змінні, що відносяться до однієї і тієї ж вибірки (наприклад, математичні успіхи студентів на початку і в кінці семестру), то зазвичай використовується t-критерій для залежних вибірок. Альтернативними непараметричними тестами є: критерій знаків і критерій Вілкоксона для парних порівнянь. Якщо розглянуті змінні по природі своїй категоріальні або є категоризованими (тобто представлені у вигляді частот, що потрапили в певні категорії), то відповідним буде критерій хі-квадрат Мак-Немара. Якщо розглядається більше двох змінних, що відносяться до

однієї і тієї ж вибірки, то зазвичай використовується дисперсійний аналіз (ANOVA) з повторними вимірами. Альтернативним непараметричним методом є ранговий дисперсійний аналіз Фрідмана або Q критерій Кохрена (останній застосовується, наприклад, якщо змінна виміряна в номінальній шкалі). Q критерій Кохрена використовується також для оцінки змін частот.

Залежності між змінними. Для того щоб оцінити залежність між двома змінними, зазвичай обчислюють коефіцієнт кореляції Пірсона. Непараметричними аналогами є коефіцієнти рангової кореляції Спірмена R, статистика Кендала і коефіцієнт Гамма. Якщо є більше двох змінних, то використовують коефіцієнт конкордації Кендала. Наприклад, він застосовується для оцінки узгодженості думок незалежних експертів (суддів), наприклад, балів, виставлених одному і тому ж учаснику конкурсу.

Різновиди статистичних критеріїв:

- *Критерій узгодження.* Перевірка на узгодженість має на увазі, що випадкова величина, що досліджується, підкорюється закону, що розглядається. Критерій узгодження можна також сприймати, як критерій значущості.
- *Критерій значущості.* Перевірка за значущістю припускає перевірку гіпотези про числові значення параметрів відомого закону розподілу.
- *Критерій однорідності.* При перевірці на однорідність випадкові величини досліджуються на факт взаємної відповідності їх законів розподілу (чи підкорюються ці величини одному і тому ж закону). Використовуються у факторному аналізі для визначення наявності залежностей.

Критична область. Потужність критерію

Після обрання певного критерію узгодження, множину усіх його можливих значень поділяють на дві підмножини, що не перетинаються: одна з них містить значення критерію, при яких основна гіпотеза відхиляється, а друга – при яких вона приймається.

- **Критичною областю** називають сукупність значень критерію, при яких основна гіпотеза відхиляється.
- **Областю прийняття гіпотези (областю допустимих значень)** називають множину значень критерію, при яких немає підстав відхилити основну гіпотезу.

Критерій узгодження K – одновимірна випадкова величина, усі її можливі значення належать деякому інтервалу. Тому критична область та область прийняття гіпотези також будуть інтервалами, а це означає, що існують точки, які ці інтервали відокремлюють.

- **Критичними точками (межами) критерію K** називають точки $k_{кр}$, які відокремлюють критичну область від області прийняття гіпотези.

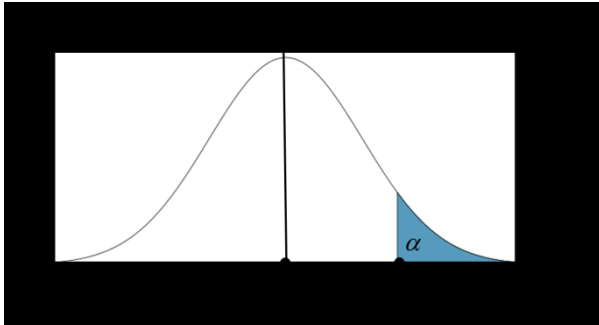
Для кожного критерію узгодження є відповідні таблиці критичних точок.

Щоб знайти критичну область, треба знайти критичну точку $k_{кр}$. Для цього задають достатньо малу імовірність – *рівень значущості* α , а потім шукають критичну точку з врахуванням вимоги $P(K \text{ належить критичній області}) = \alpha$

Розрізняють однобічну (правобічну та лівобічну) та двобічну критичні області.

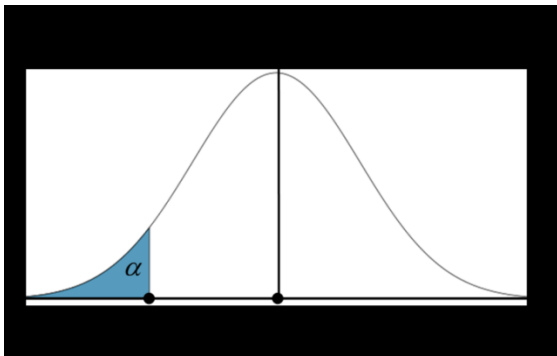
- **Правобічною** називають критичну область, що визначається нерівністю $K > k_{кр}$, де $k_{кр}$ – додатне число.

$$P(K > k_{kp}) = \alpha$$



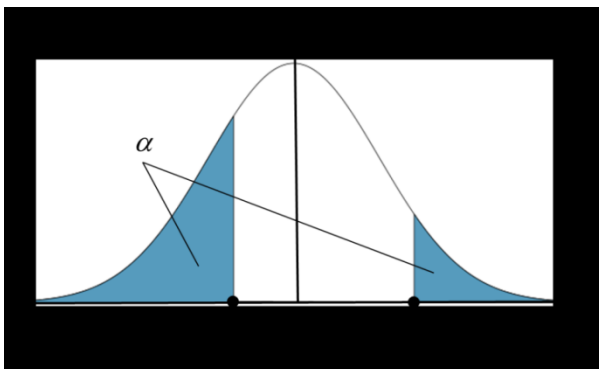
➤ **Лівобічною** називають критичну область, що визначається нерівністю $K < k_{kp}$, де k_{kp} – від’ємне число.

$$P(K < k_{kp}) = \alpha$$



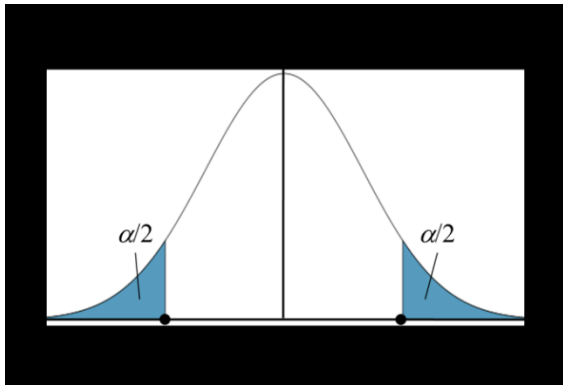
➤ **Двосторонньою** називають критичну область, що визначається нерівностями $K < k_1$, $K > k_2$ де $k_2 > k_1$.

$$P(K < k_1) + P(K > k_2) = \alpha$$



У випадку, коли критичні точки симетричні відносно 0, двостороння критична область визначається нерівністю

$$P(|K| > k_{kp}) = \alpha$$



Коли критична точка вже знайдена, за даними вибірок обчислюють спостережене значення критерію і, якщо виявиться, що $K_{\text{сп}} > k_{kp}$ то нульову гіпотезу відкидають (у випадку правосторонньої критичної області); якщо ж $K_{\text{сп}} < k_{kp}$ то немає підстав, щоб відкинути нульову гіпотезу.

- ✓ Спостережене значення критерію може виявитися більшим за k_{kp} не тому, що нульова гіпотеза помилкова, а з інших причин (малий об'єм вибірки, недоліки методики експерименту та ін.). В цьому випадку, відкинувши правильну нульову гіпотезу, здійснюють помилку першого роду. Вірогідність цієї помилки дорівнює рівню значущості α .
- ✓ Нехай нульова гіпотеза прийнята; помилково думати, що тим самим вона доведена - бо один приклад, що підтверджує справедливість деякого загального твердження, ще не доводить його. Правильніше говорити «дані спостережень узгоджуються з нульовою гіпотезою і, отже, не дають підстав її відхилити».

На практиці для більшої впевненості прийняття гіпотези її перевіряють іншими способами або повторюють експеримент, збільшивши об'єм вибірки.

- ✓ Відкидають гіпотезу категоричніше, ніж приймають. Дійсно, відомо, що досить привести один приклад, що суперечить деякому загальному твердженню, щоб це твердження відхилити. Якщо виявилось, що спостережене значення критерію належить критичній області, то цей факт і служить прикладом, що суперечить нульовій гіпотезі, що дозволяє її відхилити.

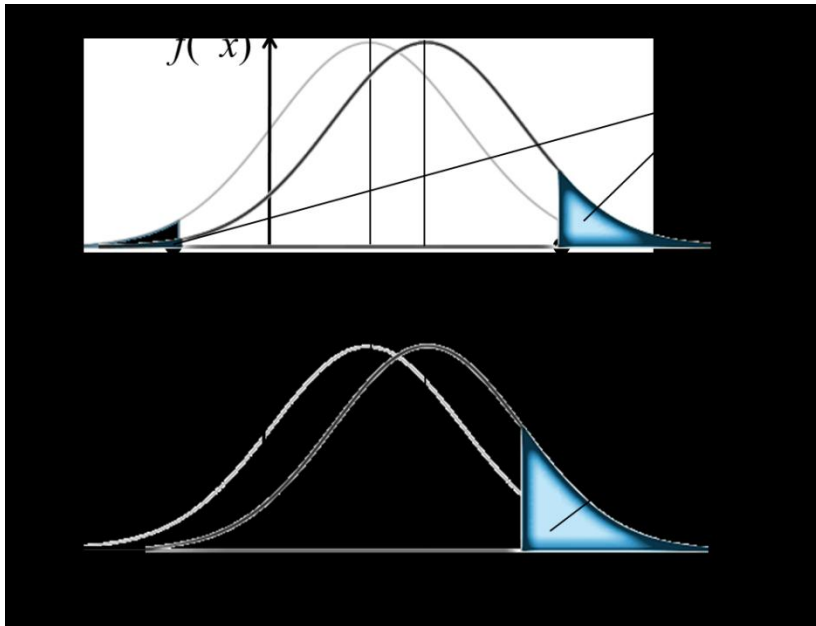
При знаходженні критичної області доцільно враховувати потужність критерію.

- **Потужністю критерію** називають імовірність належності критерію критичній області при умові, що правильна альтернативна гіпотеза.

Іншими словами, потужність критерію є імовірність того, що основна гіпотеза буде відхилена, якщо альтернативна гіпотеза правильна.

Приклад. Перевіряється гіпотеза $H_0: M(X) = 5$, в якості критерію вибрано $K = \bar{X}$, критична область двостороння. Нехай правильною є гіпотеза $H_1: M(X) = 7$.

Тоді з імовірністю P_d буде відхилена основна гіпотеза, якщо альтернативна гіпотеза правильна.



Тепер візьмемо правосторонню критичну область. Тоді з імовірністю $P_{\pi} > P_d$ буде відхилена основна гіпотеза, якщо альтернативна гіпотеза правильна.

Отже, в даному прикладі при правосторонній критичній області імовірність відхилити неправильну гіпотезу більше.

Якщо рівень значущості α вже обрано, то критичну область доцільно будувати (вибирати) так, щоб потужність критерію була максимальною – таким чином забезпечується мінімальна імовірність похибки другого роду. Для цього треба врахувати наявність та вигляд альтернативної гіпотези:

- якщо альтернативна гіпотеза приписує критерію K більші значення, ніж основна гіпотеза, то краще брати правосторонню критичну область;
- якщо альтернативна гіпотеза приписує критерію K менші значення, ніж основна гіпотеза, то краще брати лівосторонню критичну область;
- якщо альтернативна гіпотеза відсутня, то краще брати двосторонню критичну область.

Єдиний спосіб одночасного зменшення імовірностей похибок першого та другого роду – це збільшення об'єму вибірки.

Загальна схема при перевірці статистичних гіпотез

- 1) сформулювати гіпотезу H_0 ;
- 2) обрати статистичну характеристику – критерій узгодження для перевірки гіпотези;
- 3) задати допустиму імовірність похибки першого роду – рівень значущості α ;
- 4) визначити гіпотезу H_1 , альтернативну до гіпотези H_0 ;
- 5) обрати вигляд критичної області з урахуванням наявної альтернативної гіпотези;
- 6) знайти за відповідною таблицею критичну область (критичну точку) для обраної статистичної характеристики;
- 7) за вибіркою обчислити емпіричне значення критерію та порівняти його з критичним;
- 6*) за вибіркою обчислити емпіричне значення критерію та визначити його імовірність (рівень значущості);
- 7*) порівняти обчислений рівень значущості з заданим α ;
- 8) зробити висновок про прийняття чи відхилення гіпотези H_0 .

Проміжні підсумки

- Статистична гіпотеза – це твердження про генеральну сукупність, яке висувається та перевіряється на основі даних вибірки.
- Разом з основною гіпотезою завжди можна розглядати протилежну їй альтернативну гіпотезу. Вони формуються на основі даних вибірки та апіорних даних про генеральну сукупність – чим їх більше, тим обґрунтованіші висновки можна зробити.

- Для перевірки гіпотез користуємось критерієм – міркою, орієнтуючись на яку можемо робити висновки про генеральну сукупність.
- При перевірці статистичної гіпотези за даними випадкової вибірки можна зробити хибний висновок. При цьому можуть бути похибки I та II роду. Імовірність похибки I роду регулюється рівнем значущості, II роду – виглядом критичної області, обидві – обсягом вибірки.
- Відкидають гіпотезу категоричніше, ніж приймають.

Перевірка гіпотез про вигляд закону розподілу

Якщо закон розподілу невідомий, але є міркування для припущення щодо його певного вигляду A , наприклад, розподіл рівномірний, показниковий або нормальний, тоді висувають гіпотезу: *генеральна сукупність розподілена за законом A* .

Вибір критерію узгодження для перевірки гіпотез про вигляд розподілу залежить від того, як саме сформульована гіпотеза, від шкали вимірювання випадкової величини та від обсягу вибірки. Критерії узгодження, що найчастіше використовуються при перевірці гіпотез про вигляд розподілу: для перевірки гіпотези про нормальний розподіл генеральної сукупності:

- критерій Шапіро-Уїлка;
- критерій Пірсона (χ^2);

для перевірки гіпотези про довільний розподіл генеральної сукупності:

- критерій Колмогорова;
- критерій Мізеса (ω^2);

для перевірки гіпотези про однаковий розподіл двох генеральних сукупностей:

- критерій інверсії (Вілкоксона).

Критерій узгодження Шапіро-Уїлка

Критерій узгодження Шапіро-Уїлка використовують для перевірки гіпотези про нормальний розподіл генеральної сукупності. Він базується на відношенні оптимальної лінійної незсунутої оцінки дисперсії до її звичайної оцінки методом найбільшої правдоподібності.

Статистика критерію має вигляд

$$W = \frac{1}{s^2} \left(\sum_{i=1}^n a_{n-i+1} (x_{n-i+1} - x_i) \right)^2$$

Коефіцієнти a та критичні значення $W_{кр}$ беруться з таблиць.

Критерій надійний при $8 \leq n \leq 50$ (існує модифікований критерій Шапіро-Франчича, який можна застосовувати при n до 2000). Критерій є найбільш ефективним, оскільки він має найбільшу потужність порівняно з іншими критеріями перевірки на нормальність.

Критерій узгодження Пірсона (χ^2)

Критерій узгодження Пірсона ефективно використовують для перевірки гіпотези про нормальний розподіл генеральної сукупності у випадку великих вибірок.

Нехай вибірка об'єму n має такий розподіл

варіанти	x_i	x_1	x_2	...	x_m
емпіричні частоти	n_i	n_1	n_2	...	n_m

Нехай зроблено припущення про нормальний вигляд розподілу і за цим припущенням обчислені теоретичні частоти n'_i

варіанти	x_i	x_1	x_2	...	x_m
теоретичні частоти	n'_i	n'_1	n'_2	...	n'_m

Потрібно з рівнем значущості α перевірити основну гіпотезу H_0 : *генеральна сукупність розподілена нормально.*

Критерієм перевірки цієї гіпотези беруть випадкову величину χ^2 , яка у різних випробуваннях приймає різні, наперед невідомі значення

$$\chi^2 = \sum (n_i - n'_i)^2 / n'_i$$

Очевидно, що чим менше відрізняються емпіричні та теоретичні частоти, тим менше значення приймає критерій, тому обираємо правосторонню критичну область.

Доведено, що при $n \rightarrow \infty$ закон розподілу цієї випадкової величини незалежно від того, за яким законом розподілена генеральна сукупність, прямує до закону розподілу χ^2 з $k=m-1-r$ степенями свободи, де m – число груп (варіант, інтервалів), r – кількість параметрів розподілу, визначених за даними вибірки.

Для розподілу генеральної сукупності за нормальним законом степінь свободи буде $k = m - 3$, де m – кількість варіант вибірки або інтервалів варіант.

Критичне значення цієї випадкової величини залежить від рівня значущості α та степенів свободи її розподілу k

$$P(\chi^2 > \chi_{kp}^2(\alpha, k)) = \alpha$$

Ці критичні значення табульовані для різних α та k .

Порядок дій при перевірці гіпотези

Щоб при заданому рівні значущості α перевірити основну гіпотезу H_0 : генеральна сукупність розподілена нормально, треба:

- 1) обчислити теоретичні частоти n'_i для варіант вибірки;
- 2) обчислити спостережене значення критерію

$$\chi_{cn}^2 = \sum (n_i - n'_i)^2 / n'_i$$

- 3) знайти степінь свободи;
- 4) знайти з таблиці критичну точку χ_{kp}^2 , яка відповідає заданому рівню значущості α та степені свободи k ;
- 5) порівняти χ_{cn}^2 та χ_{kp}^2 зробити висновок:

якщо $\chi_{cn}^2 < \chi_{kp}^2$, то немає підстав відхилити гіпотезу H_0 ;

якщо $\chi_{cn}^2 > \chi_{kp}^2$, то гіпотезу H_0 треба відхилити.

Критерій узгодження Пірсона має певні недоліки:

- вибір кількості інтервалів суттєво впливає на результат;
- при вузьких інтервалах математичне сподівання кількості елементів, що потрапляють в інтервал, має велику дисперсію;
- при малому обсязі вибірки розподіл випадкової величини погано описується нормальним законом розподілу.

Рекомендації по використанню критерію:

- ✓ Обсяг вибірки повинен бути достатньо великий в усякому разі, не менше 50. Кожна група повинна містити не менше 5-8 варіант; групи з невеликою кількістю варіант слід об'єднувати в одну, підсумовуючи частоти.
- ✓ Оскільки можливі помилки першого і другого роду, особливо якщо узгодження теоретичних і емпіричних частот «дуже хороше», слід проявляти обережність. Наприклад, можна повторити дослід, збільшивши число спостережень, скористатися іншими критеріями, побудувати графік розподілу.
- ✓ Для спрощення обчислень формулу для χ^2_{cn} перетворюють до вигляду

$$\chi^2_{cn} = \left(\sum n_i^2 / n_i' \right) - n$$

Методика знаходження теоретичних частот нормального розподілу

Згідно з статистичним визначенням імовірності

$$p_i = \frac{n_i'}{n} \Rightarrow n_i' = p_i \cdot n, \quad i = 1, 2, \dots, m$$

Отже, для знаходження теоретичних частот n_i' треба знайти імовірність

$$p_i = P(X = x_i) \quad \text{або} \quad p_i = P(x_i < X < x_{i+1}).$$

Імовірність $p_i = P(X=x_i)$ для дискретного варіаційного ряду можна знайти, використовуючи локальну функцію Лапласа $\varphi(x)$ та дані вибірки за формулою

$$p_i = P(X = x_i) = \frac{h}{s} \varphi(u_i), \text{ де } h = x_{i+1} - x_i; u_i = \frac{x_i - \bar{x}}{s} \quad \varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

варіанти x_i рівновіддалені (у випадку різної ширини інтервалів формула трохи ускладнюється).

Імовірність $p_i = P(x_i < X < x_{i+1})$ для інтервального варіаційного ряду можна знайти, використовуючи інтегральну функцію Лапласа $\Phi(x)$ за формулою

$$p_i = P(x_i < X < x_{i+1}) = \Phi\left(\frac{x_{i+1} - \bar{x}}{s}\right) - \Phi\left(\frac{x_i - \bar{x}}{s}\right)$$

Приклад. При рівні значущості перевірити $\alpha = 0.05$ гіпотезу про нормальний розподіл генеральної сукупності, якщо відомі емпіричні частоти. Вважаємо, що середньоквадратичне відхилення відоме і дорівнює 3.

варіанта	частота n_i
8	2
9	3
10	5
11	4
12	2
13	1
14	2
15	1

Розв'язок. У даному випадку теоретичні частоти n'_k невідомі, для їх розрахунку використаємо інтегральну функцію Лапласа.

$$p_i = P(x_i < X \leq x_{i+1}) = \Phi\left(\frac{x_{i+1} - \bar{x}}{s}\right) - \Phi\left(\frac{x_i - \bar{x}}{s}\right)$$

варіанта	частота	теоретична імовірність	теоретична частота
8	2	0.1193	2.4
9	3	0.1293	2.6
10	5	0.1293	2.6
11	4	0.1193	2.4
12	2	0.0927	1.9
13	1	0.0669	1.3
14	2	0.0443	0.9
15	1	0.0475	1.0

Кількість варіант вибірки $m = 8$, тому степінь свободи $k = 8 - 3 = 5$. З таблиці критичних точок розподілу $\chi^2(\alpha, k)$ для $\alpha = 0.05$ та $k = 5$ знаходимо $\chi^2_{кр} = 11.07$.

Спостережене значення критерію

$$\chi^2_{cn} = \left(\sum n_i^2 / n'_i\right) - n = 10.0$$

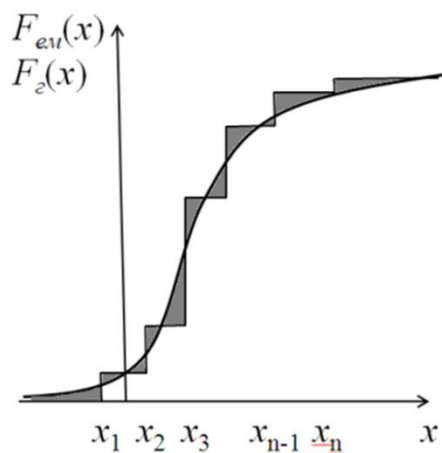
Оскільки спостережене значення критерію менше критичного, то немає підстав відхилити гіпотезу про нормальний розподіл.

Для подолання недоліків критерія Пірсона розроблено критерії Колмогорова та Мізеса.

Критерій узгодження Колмогорова

Як критерій перевірки нульової гіпотези про вигляд розподілу приймемо статистику

$$D = |F_{em}(x) - F_c(x)|$$



Критерій Колмогорова має вигляд

$$\lambda(n) = \sqrt{n} \cdot \max(D(n))$$

Величина λ має табличний розподіл. Критична область правостороння. Для розрахунку критичної точки задайте рівень значущості α та обсяг вибірки n .

Приклад. При рівні значущості перевірити $\alpha = 0.05$ гіпотезу про нормальний розподіл генеральної сукупності, якщо відомі емпіричні частоти.

варіанта	частота n_i
8	2
9	3
10	5
11	4
12	2
13	1
14	2
15	1

Розв'язок. У даному випадку теоретичні та емпіричні функції розподілу треба розрахувати, емпіричні – за відносними частотами по вибірці, теоретичні - використавши інтегральну функцію Лапласа.

варіанта	частота	відносна частота	F_{em}	F_2
8	2	0.10	0.10	0.25
9	3	0.15	0.25	0.37
10	5	0.25	0.50	0.50
11	4	0.20	0.70	0.63
12	2	0.10	0.80	0.75
13	1	0.05	0.85	0.84
14	2	0.10	0.95	0.91
15	1	0.05	1.00	0.95

Максимальна різниця значень теоретичної та емпіричної функцій розподілу дорівнює 0.15.

$$\lambda(n) = \sqrt{n} \cdot \max(D(n)) = \sqrt{20} \cdot 0.15 = 0.67$$

Критичне значення критерію Колмогорова для правосторонньої критичної області при рівні значущості $\alpha = 0.05$ та обсягу вибірки $n=20$

$$\lambda_{kp}(20) = 1.315$$

Оскільки спостережене значення критерію менше критичного, то немає підстав відхилити гіпотезу про нормальний розподіл.

Рекомендації по використанню критерію:

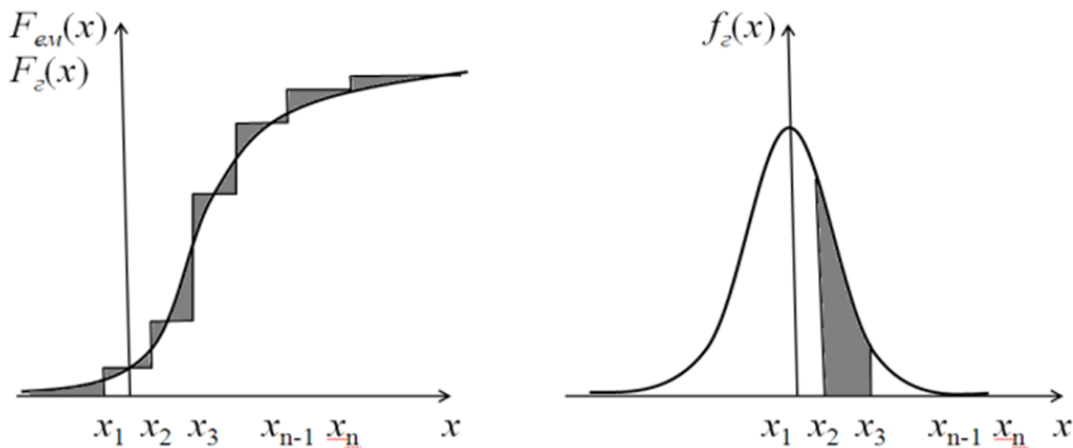
- ✓ Критерій Колмогорова можна застосовувати у випадку, коли параметри теоретичного закону розподілу визначаються не по даним вибірки.
 - ✓ Критерій також можна застосовувати для статистичної перевірки гіпотези про належність двох вибірок одній генеральній сукупності.
- В цьому випадку критерій має вигляд

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max(F_1(x) - F_2(x))$$

Критерій узгодження Мізеса (ω^2)

Як критерій перевірки нульової гіпотези про вигляд розподілу використовується випадкова величина

$$\omega^2 = \int_{-\infty}^{\infty} (F_{em}(x) - F_z(x))^2 f_z(x) dx$$



Величина $m=(n-\omega^2)$ має табличний розподіл. Критична область правостороння. Застосовувати критерій ω^2 можна для $n>40$.

Критерій однорідності Вілкоксона (інверсії)

Нехай маємо дві вибірки $\{x\}_n$, $\{y\}_m$ двох випадкових величин X та Y . Потрібно з рівнем значущості α перевірити основну гіпотезу H_0 :

обидві вибірки належать одній генеральній сукупності,

або дві генеральні сукупності мають однаковий розподіл.

Запишемо елементи вибірок $\{x\}_n$, $\{y\}_m$ в одну загальну вибірку в порядку зростання значень

$$\{x_1, y_1, y_2, y_3, x_2, y_4, \dots\}$$

Підрахуємо кількість інверсій, пов'язаних з цією вибіркою

$$u = \sum_{i=1}^n u_{y_i}$$

u_{y_i} – кількість елементів вибірки $\{x\}_n$, що знаходяться перед елементом y_i .

Ця випадкова величина використовується як критерій перевірки нульової гіпотези. При $n \geq m \geq 4$ та $n+m \geq 50$ вона досить добре описується нормальним законом розподілу з параметрами

$$M(u) = \frac{m \cdot n}{2}; \quad D(u) = \frac{m \cdot n}{12} \cdot (m + n + 1)$$

Критична область правостороння.

Перевірка гіпотез про параметри розподілу

Для перевірки гіпотез про числові значення параметрів відомого закону розподілу використовують критерії значущості.

Гіпотези про дисперсію

Найчастіше виникає потреба перевірити такі гіпотези:

- Гіпотеза про рівність дисперсій двох нормальних генеральних сукупностей. Критерій Фішера
- Порівняння кількох дисперсій нормальних генеральних сукупностей за вибірками різного обсягу. Критерій Бартлетта
- Порівняння кількох дисперсій нормальних генеральних сукупностей за вибірками однакового обсягу. Критерій Кочрена
- Порівняння виправленої вибіркової дисперсії з гіпотетичною генеральною дисперсією нормальної сукупності. Критерій Пірсона

Перевірка гіпотези про рівність дисперсій двох нормальних генеральних сукупностей (критерій Фішера)

На практиці задача порівняння дисперсій виникає, коли потрібно порівняти точність приладів, інструментів, методів вимірювання тощо. Очевидно, що кращим є той прилад, інструмент або метод, що забезпечує найменше розсіювання результатів вимірювань, тобто має найменшу дисперсію.

Нехай дві генеральні сукупності розподілені нормально. Із сукупностей зробили вибірки об'єму n_1 та n_2 і знайшли виправлені дисперсії s_1^2 та s_2^2 , відповідно.

Потрібно по виправленим дисперсіям при заданому рівні значущості перевірити гіпотезу H_0 : дисперсії генеральних сукупностей рівні

$$H_0 : D(X_1) = D(X_2)$$

Враховуючи, що виправлені дисперсії є незсунутими оцінками генеральних дисперсій, тобто $M(s_1^2) = D(X_1)$, $M(s_2^2) = D(X_2)$, нульову гіпотезу можна записати так:

$$H_0 : M(s_1^2) = M(s_2^2)$$

Таким чином, потрібно перевірити, що математичні сподівання виправлених вибірових дисперсій рівні між собою. Таке завдання ставиться тому, що зазвичай виправлені дисперсії виявляються різними. Виникає питання: значуще (істотно) або незначуще розрізняються виправлені дисперсії?

Якщо виявиться, що нульова гіпотеза справедлива, тобто генеральні дисперсії однакові, то відмінність виправлених дисперсій незначуща і пояснюється випадковими причинами, зокрема випадковим відбором об'єктів вибірки. Наприклад, якщо відмінність виправлених вибірових дисперсій результатів вимірювань, виконаних двома приладами, виявилася незначущою, то прилади мають однакову точність.

Якщо нульова гіпотеза відхилена, тобто генеральні дисперсії неоднакові, то відмінність виправлених дисперсій значуща і не може бути пояснена випадковими причинами, а є наслідком того, що самі генеральні дисперсії різні. Наприклад, якщо відмінність виправлених вибірових дисперсій результатів вимірювань, проведених двома приладами, опинилася значущою, то точність приладів різна.

Як критерій перевірки нульової гіпотези про рівність генеральних дисперсій приймемо відношення більшої виправленої дисперсії до меншої, тобто випадкову величину

$$F = \frac{S_o^2}{S_m^2}$$

Величина F за умови справедливості нульової гіпотези має розподіл Фішера-Снедекора із степенями свободи $k_1=n_1-1$ та $k_2=n_2-1$, де n_1 – обсяг вибірки, за якою обчислена більша виправлена дисперсія, n_2 – обсяг вибірки, за якою обчислена менша виправлена дисперсія.

Критична область будується в залежності від вигляду альтернативної гіпотези.

Випадок перший. $H_0: D(X_1) = D(X_2)$, $H_1: D(X_1) > D(X_2)$.

Вибираємо правосторонню критичну область (в цьому випадку потужність критерію буде більшою), виходячи з вимоги, що імовірність влучення критерію F в цю область при справедливості нульової гіпотези буде дорівнювати прийнятому рівню значущості

$$P(F > F_{kp}(\alpha; k_1; k_2)) = \alpha$$

Критичну точку $F_{кр}$ знаходимо по таблиці критичних точок розподілу Фішера-Снедекора.

Розраховуємо спостережене значення критерію як відношення більшої виправленої дисперсії до меншої

$$F_{cn} = \frac{s_{\sigma}^2}{s_m^2}$$

Якщо $F_{cn} < F_{кр}$, то немає підстав відхилити нульову гіпотезу. Якщо $F_{cn} > F_{кр}$, то нульову гіпотезу відхиляємо.

Випадок другий. $H_0: D(X_1) = D(X_2)$, $H_1: D(X_1) \neq D(X_2)$.

Вибираємо двосторонню критичну область, виходячи з вимоги, що імовірність влучення критерію F в цю область при справедливості нульової гіпотези буде дорівнювати прийнятому рівню значущості

$$P(F < F_1) = \alpha / 2, \quad P(F > F_2) = \alpha / 2$$

Праву критичну точку $F_2 = F_{кр}(\alpha/2; k_1; k_2)$ знаходимо безпосередньо з таблиці критичних точок розподілу Фішера-Снедекора за рівнем значущості $\alpha/2$ та степенями свободи k_1 і k_2 .

Лівих критичних точок таблиця не містить, тому знайти F_1 по таблиці неможливо. В загальному випадку при двосторонній альтернативній гіпотезі в цьому немає потреби, достатньо знати лише F_2 (доведення подивитись самостійно).

Приклад. По двом незалежним вибіркам обсягом $n_X = 12$ та $n_Y = 15$ зробленим з нормальних генеральних сукупностей X та Y знайдено виправлені вибіркові дисперсії $s^2_X = 11.41$ та $s^2_Y = 6.52$. При рівні значущості 0.05 перевірити нульову гіпотезу $H_0: D(X) = D(Y)$ про рівність генеральних дисперсій при альтернативній гіпотезі $H_1: D(X) > D(Y)$

Розв'язок. Знайдемо відношення більшої виправленої дисперсії до меншої

$$F_{ср} = 11.41 / 6.52 = 1.75$$

Альтернативна гіпотеза має вигляд $H_1: D(X) > D(Y)$, тому критична область – правостороння. По таблиці розподілу Фішера для рівня значущості $\alpha = 0.05$ та степенів свободи $k_1 = 12 - 1 = 11$ та $k_2 = 15 - 1 = 14$ знаходимо критичну точку $F_{кр}(\alpha; k_1; k_2) = 2.56$.

Оскільки $F_{ср} < F_{кр}$ - немає підстав відкинути нульову гіпотезу про рівність генеральних дисперсій.

Порівняння кількох дисперсій нормальних генеральних сукупностей за вибірками різного обсягу (критерій Бартлетта)

Нехай генеральні сукупності X_1, X_2, \dots, X_l розподілені нормально. З цих сукупностей зроблені незалежні вибірки різних обсягів n_1, n_2, \dots, n_l (деякі

обсяги можуть бути однаковими; якщо всі вибірки мають однаковий обсяг, то краще користуватися критерієм Кохрена).

По вибірках знайдені виправлені вибіркові дисперсії $s_1^2, s_2^2, \dots, s_l^2$.

Потрібно по виправлених вибіркових дисперсіях при заданому рівні значущості α перевірити нульову гіпотезу: генеральні дисперсії даних сукупностей рівні між собою.

$$H_0 : D(X_1) = D(X_2) = \dots = D(X_l)$$

Таку гіпотезу про рівність кількох дисперсій називають ще *гіпотезою про однорідність дисперсій*.

Числом степенів свободи дисперсії s_i^2 називають число $k_i = n_i - 1$, тобто число, на одиницю менше обсягу вибірки, по якій обчислено дисперсію.

Позначимо $\overline{s^2}$ середню арифметичну виправлених дисперсій, зважену по числам степенів свободи

$$\overline{s^2} = \frac{\sum_{i=1}^l k_i s_i^2}{k}, \quad k = \sum_{i=1}^l k_i$$

В якості критерію значущості для перевірки гіпотези про однорідність дисперсій використовують критерій Бартлетта – випадкову величину

$$B = V/C, \text{ де}$$

$$V = 2.303 \left[k \lg \overline{s^2} - \sum_{i=1}^l k_i \lg s_i^2 \right]$$

$$C = 1 + \frac{l}{3(l-1)} \left[\sum_{i=1}^l \frac{1}{k_i} - \frac{1}{k} \right]$$

Бартлетт встановив, що випадкова величина B при умові справедливості нульової гіпотези розподілена приблизно як χ^2 з $l-1$ степенями свободи, якщо всі $k_i > 2$, тобто об'єм кожної з вибірок повинен бути не менше 4.

Критична область правостороння

$$P(B > \chi_{кр}^2(\alpha; l-1)) = \alpha$$

Критичну точку $\chi_{кр}^2(\alpha, l-1)$ знаходять по таблиці за рівнем значущості α і числом степенів свободи $k=l-1$, і тоді правостороння критична область визначається нерівністю $B > \chi_{кр}^2$, а область прийняття гіпотези - нерівністю $B < \chi_{кр}^2$.

- ✓ Не слід поспішати обчислювати сталу C . Спочатку треба знайти V і порівняти з $\chi_{кр}^2$; якщо виявиться, що $V < \chi_{кр}^2$, то тим паче $B=(V/C) < \chi_{кр}^2$ (оскільки $C > 1$) і, отже, C обчислювати не потрібно. Якщо ж $V > \chi_{кр}^2$, то треба обчислити C і потім порівняти B з $\chi_{кр}^2$.
- ✓ Критерій Бартлетта вельми чутливий до відхилень розподілів від нормального, тому до висновків, отриманих по цьому критерію, треба відноситися з обережністю.

Порівняння кількох дисперсій нормальних генеральних сукупностей за вибірками однакового обсягу (критерій Кохрена)

Нехай генеральні сукупності X_1, X_2, \dots, X_l розподілені нормально. З них зроблено l незалежних вибірок однакового обсягу n і знайдено виправлені вибіркові дисперсії $s_1^2, s_2^2, \dots, s_l^2$, всі з однаковим числом степенів свободи $k=n-1$.

Потрібно по виправлених вибіркових дисперсіях при заданому рівні значущості α перевірити нульову гіпотезу: генеральні дисперсії даних сукупностей рівні між собою.

$$H_0 : D(X_1) = D(X_2) = \dots = D(X_l)$$

У випадку вибірок однакового обсягу можна було б по критерію Фішера-Снедекора порівняти найбільшу і найменшу дисперсії; якщо опиниться, що відмінність між ними незначуща, то незначуща і відмінність між рештою дисперсій. Недолік цього методу полягає в тому, що інформація, яку містить решта дисперсій, окрім найменшої і найбільшої, не враховується.

Можна також застосувати критерій Бартлетта. Проте відомо лише наближений розподіл цього критерію, тому краще використовувати критерій Кохрена, розподіл якого знайдено точно.

Критерій Кохрена - відношення максимальної виправленої дисперсії до суми всіх виправлених дисперсій:

$$G = \frac{S_{\max}^2}{S_1^2 + S_2^2 + \dots + S_l^2}$$

Розподіл цієї випадкової величини залежить тільки від числа степенів свободи $k = n - 1$ і кількості вибірок l .

Критична область правостороння.

$$P(G > G_{kp}(\alpha; k; l)) = \alpha$$

Критичну точку $G_{kp}(\alpha, k, l)$ знаходять по таблиці, і тоді правостороння критична область визначається нерівністю $G > G_{kp}$, а область прийняття нульової гіпотези - нерівністю $G < G_{kp}$.

Порівняння виправленої вибіркової дисперсії з гіпотетичною генеральною дисперсією нормальної сукупності (критерій Пірсона)

На практиці така задача виникає, коли потрібно перевірити точність приладів, інструментів, методів вимірювання та стійкість технологічних процесів. Наприклад, якщо допустима характеристика розсіювання

розміру деталі дорівнює σ_0^2 , а знайдена за вибіркою виявиться значно більше σ_0^2 , то станок потребує наладки.

Нехай генеральна сукупність розподілена нормально, при цьому генеральна дисперсія хоча і невідома, але можна припустити, що вона дорівнює певному гіпотетичному значенню σ_0^2 . На практиці σ_0^2 встановлюється на підставі попереднього досвіду або теоретично.

Із генеральної сукупності зробили вибірку об'єму n і знайшли виправлену дисперсію S^2 з $k=n-1$ степенями свободи.

Потрібно по виправленій дисперсії при заданому рівні значущості перевірити гіпотезу H_0 : генеральна дисперсія сукупності дорівнює гіпотетичному значенню σ_0^2

$$H_0 : D(X) = \sigma_0^2$$

Враховуючи, що виправлена дисперсія S^2 є незсунутою оцінкою генеральної дисперсії, нульову гіпотезу можна записати так:

$$H_0 : M(S^2) = \sigma_0^2$$

Таким чином, потрібно перевірити, що математичне сподівання виправленої вибіркової дисперсії дорівнює гіпотетичному значенню генеральної дисперсії.

Як критерій перевірки нульової гіпотези приймемо відношення виправленої дисперсії до гіпотетичного значення генеральної дисперсії, тобто випадкову величину

$$\frac{(n-1)S^2}{\sigma_0^2}$$

Величина має розподіл χ^2 із $k=n-1$ степенями свободи.

Критична область будується в залежності від вигляду альтернативної гіпотези.

Випадок перший. $H_0: \sigma^2 = \sigma_0^2, H_1: \sigma^2 > \sigma_0^2$.

Вибираємо правосторонню критичну область (в цьому випадку потужність критерію буде більшою), виходячи з вимоги, що імовірність влучення критерію χ^2 в цю область при справедливості нульової гіпотези буде дорівнювати прийнятому рівню значущості

$$P(\chi^2 > \chi_{кр}^2(\alpha; k)) = \alpha$$

Критичну точку $\chi_{кр}^2$ знаходимо по таблиці критичних точок розподілу χ^2 . Тоді правостороння критична область визначається нерівністю $\chi^2 > \chi_{кр}^2$, а область прийняття нульової гіпотези – нерівністю $\chi^2 < \chi_{кр}^2$.

Розраховуємо спостережене значення критерію

$$\chi_{сп}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Якщо $\chi_{сп}^2 < \chi_{кр}^2$, то немає підстав відхилити нульову гіпотезу. Якщо $\chi_{сп}^2 > \chi_{кр}^2$, то нульову гіпотезу відхиляємо.

Випадок другий. $H_0: \sigma^2 = \sigma_0^2, H_1: \sigma^2 \neq \sigma_0^2$.

Вибираємо двосторонню критичну область.

$$P(\chi^2 < \chi_{лів.кр}^2(\alpha/2; k)) = \alpha/2$$

$$P(\chi^2 > \chi_{прав.кр}^2(\alpha/2; k)) = \alpha/2$$

Праву критичну точку $\chi_{прав.кр}^2$ знаходимо по таблиці критичних точок розподілу χ^2 за рівнем значущості $\alpha/2$ та степенем свободи k .

Ліву критичну точку знаходимо з умови

$$P(\chi^2 < \chi_{\text{лів.кр}}^2(\alpha/2; k)) + P(\chi^2 > \chi_{\text{прав.кр}}^2(\alpha/2; k)) = 1$$

$$P(\chi^2 < \chi_{\text{лів.кр}}^2(\alpha/2; k)) = 1 - P(\chi^2 > \chi_{\text{лів.кр}}^2(\alpha/2; k)) = 1 - \alpha/2$$

по таблиці критичних точок розподілу χ^2 за рівнем значущості $(1-\alpha/2)$ та степенем свободи k .

Розраховуємо спостережене значення критерію

$$\chi_{\text{сп}}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Якщо $\chi_{\text{лів.кр}}^2 < \chi_{\text{сп}}^2 < \chi_{\text{прав.кр}}^2$, то немає підстав відхилити нульову гіпотезу.

Якщо $\chi_{\text{лів.кр}}^2 > \chi_{\text{сп}}^2$ або $\chi_{\text{сп}}^2 > \chi_{\text{прав.кр}}^2$, то нульову гіпотезу відхиляємо.

Випадок третій. $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 < \sigma_0^2$.

Вибираємо лівосторонню критичну область.

$$P(\chi^2 < \chi_{\text{кр}}^2(1-\alpha; k)) = 1 - \alpha$$

Критичну точку $\chi_{\text{кр}}^2$ знаходимо по таблиці критичних точок розподілу χ^2 за рівнем значущості $(1-\alpha)$ та степенем свободи k .

Тоді лівостороння критична область визначається нерівністю $\chi_{\text{кр}}^2 < \chi^2$, а область прийняття нульової гіпотези – нерівністю $\chi^2 > \chi_{\text{кр}}^2$.

Розраховуємо спостережене значення критерію

$$\chi_{\text{сп}}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Якщо $\chi_{\text{кр}}^2 < \chi_{\text{сп}}^2$, то немає підстав відхилити нульову гіпотезу. Якщо $\chi_{\text{кр}}^2 > \chi_{\text{сп}}^2$, то нульову гіпотезу відхиляємо.

- ✓ У випадку, коли знайдено вибіркву дисперсію $D_{\text{в}}$, в якості критерію приймають випадкову величину

$$\chi^2 = \frac{nD_B}{\sigma_0^2}$$

яка має розподіл χ^2 із $k=n-1$ степенями свободи, або переходять до

$$S^2 = \frac{n}{n-1} D_B$$

Приклад. З нормальної генеральної сукупності X зроблено вибірку обсягом $n=13$ та знайдено по ній виправлену вибірккову дисперсію $s^2=10.3$. При рівні значущості 0.02 перевірити нульову гіпотезу $H_0: \sigma^2 = 12$ при альтернативній гіпотезі $H_1: \sigma^2 \neq 12$

Розв'язок. Знайдемо спостережене значення критерію

$$\chi_{\text{сп}}^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(13-1)10.3}{12} = 10.3$$

Альтернативна гіпотеза має вигляд $\sigma^2 \neq 12$, тому критична область – двостороння. По таблиці розподілу χ^2 знаходимо критичні точки:

$$\text{ліву } \chi_{\text{лів.кр}}^2(1-\alpha/2; k) = \chi_{\text{лів.кр}}^2(0.99; 12) = 3.57$$

$$\text{та праву } \chi_{\text{прав.кр}}^2(\alpha/2; k) = \chi_{\text{прав.кр}}^2(0.01; 12) = 26.2.$$

Оскільки $\chi_{\text{лів.кр}}^2 < \chi_{\text{сп}}^2 < \chi_{\text{прав.кр}}^2$ ($3.57 < 10.3 < 26.2$) - немає підстав відкинути нульову гіпотезу.

Гіпотези про середнє

Найчастіше виникає потреба перевірити такі гіпотези:

- про рівність математичних сподівань нормальних генеральних сукупностей, дисперсії яких відомі (незалежні вибірки);
- про рівність двох середніх довільно розподілених генеральних сукупностей (великі незалежні вибірки);
- про рівність двох середніх нормальних генеральних сукупностей, дисперсії яких невідомі та однакові (малі незалежні вибірки);
- про рівність вибіркової середньої генеральній середній нормальної сукупності, дисперсія якої відома;
- про рівність вибіркової середньої генеральній середній нормальної сукупності, дисперсія якої невідома;
- про рівність двох середніх нормальних генеральних сукупностей з невідомими дисперсіями (залежні вибірки).

Перевірка гіпотези про рівність математичних сподівань нормальних генеральних сукупностей, дисперсії яких відомі (незалежні вибірки)

Нехай дві генеральні сукупності розподілені нормально та їх дисперсії відомі (наприклад, з попередніх дослідів або обчислені теоретично).

Із сукупностей зробили вибірки об'єму n_1 та n_2 і знайшли вибіркові середні \bar{x}_1 та \bar{x}_2 .

Потрібно перевірити гіпотезу: математичні сподівання цих сукупностей рівні.

$$H_0 : M(X_1) = M(X_2)$$

Враховуючи, що вибіркові середні є незсунутими оцінками генеральних середніх

$$M(X_1) = M(\bar{X}_1), \quad M(X_2) = M(\bar{X}_2)$$

нульову гіпотезу можна записати так:

$$H_0 : M(\bar{X}_1) = M(\bar{X}_2)$$

Таким чином, потрібно перевірити, що математичні сподівання вибірових середніх рівні між собою.

Таке завдання ставиться тому, що зазвичай вибірові середні виявляються різними. Виникає питання: значуще (істотно) або незначуще розрізняються вибірові середні?

Як критерій перевірки нульової гіпотези про рівність вибірових середніх приймемо випадкову величину

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma(\bar{X}_1 - \bar{X}_2)} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{D(X_1)}{n_1} + \frac{D(X_2)}{n_2}}}$$

Величина Z має нормальний нормований розподіл, оскільки є лінійною комбінацією нормально розподілених величин \bar{X}_1 та \bar{X}_2 ; Z – нормована величина, бо $M(Z)=0$; при справедливості нульової гіпотези $\sigma(Z)=1$, бо вибірки незалежні.

Критична область будується в залежності від вигляду альтернативної гіпотези.

Випадок перший. $H_0: M(X_1) = M(X_2), \quad H_1: M(X_1) \neq M(X_2).$

Вибираємо двосторонню критичну область, виходячи з вимоги, що імовірність влучення критерію Z в цю область при справедливості нульової гіпотези буде дорівнювати прийнятому рівню значущості

$$P(Z < z_{\text{лів.кр}}) = \alpha / 2, \quad P(Z > z_{\text{прав.кр}}) = \alpha / 2$$

Оскільки Z – нормована нормальна величина, її розподіл симетричний відносно нуля. Потужність критерію буде найбільшою, якщо критичні точки теж симетричні відносно нуля.

$$P(Z < -z_{кр}) = \alpha / 2, \quad P(Z > z_{кр}) = \alpha / 2$$

Праву критичну точку $z_{кр}$ знаходимо, користуючись інтегральною функцією Лапласа $\Phi(z)$. Відомо, що функція Лапласа визначає імовірність влучення нормованої нормальної випадкової величини в інтервал $(0; z)$:

$$P(0 < Z < z) = \Phi(z)$$

Оскільки розподіл Z симетричний відносно нуля, то імовірність влучення Z в інтервал $(0; \infty)$ дорівнює 0.5. Якщо розбити цей інтервал точкою $z_{кр}$ на дві частини, то

$$P(0 < Z < z_{кр}) + P(Z > z_{кр}) = 1/2$$

$$\Phi(z_{кр}) + \alpha/2 = 1/2 \quad \Rightarrow \quad \Phi(z_{кр}) = (1 - \alpha)/2$$

Знаходимо $z_{кр}$ по таблиці розподілу $\Phi(z)$ за рівнем значущості $(1 - \alpha)/2$.

Розраховуємо спостережене значення критерію

$$Z_{cn} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{D(X_1)}{n_1} + \frac{D(X_2)}{n_2}}}$$

Якщо $-z_{кр} < Z_{cn} < z_{кр}$, то немає підстав відхилити нульову гіпотезу. Якщо $Z_{cn} < -z_{кр}$ або $z_{кр} < Z_{cn}$, то нульову гіпотезу відхиляємо.

Випадок другий. $H_0: M(X_1) = M(X_2), \quad H_1: M(X_1) > M(X_2)$.

На практиці такий випадок трапляється, наприклад, при вдосконаленні технологічного процесу, і можна припустити, що це призведе до збільшення випуску продукції.

Вибираємо правосторонню критичну область

$$P(Z > z_{кр}) = \alpha$$

Критичну точку $z_{кр}$ знаходимо, користуючись інтегральною функцією Лапласа $\Phi(z)$.

$$P(0 < Z < z_{кр}) + P(Z > z_{кр}) = 1/2$$

$$\Phi(z_{кр}) + \alpha = 1/2 \Rightarrow \Phi(z_{кр}) = (1 - 2\alpha)/2$$

Знаходимо $z_{кр}$ по таблиці розподілу $\Phi(z)$ за рівнем значущості $(1 - 2\alpha)/2$.

Розраховуємо спостережене значення критерію

$$Z_{cn} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{D(X_1)}{n_1} + \frac{D(X_2)}{n_2}}}$$

Якщо $Z_{cn} < z_{кр}$, то немає підстав відхилити нульову гіпотезу. Якщо $z_{кр} < Z_{cn}$, то нульову гіпотезу відхиляємо.

Випадок третій. $H_0: M(X_1) = M(X_2)$, $H_1: M(X_1) < M(X_2)$.

Вибираємо лівосторонню критичну область

$$P(Z < z'_{кр}) = \alpha$$

Критичну точку знаходимо, користуючись властивістю $z'_{кр} = -z_{кр}$.

$$P(0 < Z < z_{кр}) + P(Z > z_{кр}) = 1/2$$

$$\Phi(z_{кр}) + \alpha = 1/2 \Rightarrow \Phi(z_{кр}) = (1 - 2\alpha)/2$$

Знаходимо $z_{кр}$ по таблиці розподілу $\Phi(z)$ за рівнем значущості $(1 - 2\alpha)/2$.

Розраховуємо спостережене значення критерію

$$Z_{cn} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{D(X_1)}{n_1} + \frac{D(X_2)}{n_2}}}$$

Якщо $-z_{кр} < Z_{сн}$, то немає підстав відхилити нульову гіпотезу. Якщо $Z_{сн} < -z_{кр}$, то нульову гіпотезу відхиляємо.

Приклад. По двом незалежним вибіркам обсягом $n_1=10$ та $n_2=10$ зробленим з нормальних генеральних сукупностей X та Y знайдено вибіркові середні $\bar{x}=14.3$ та $\bar{y}=12.2$. Генеральні дисперсії відомі $D(X)=22$ та $D(Y)=18$. При рівні значущості 0.05 перевірити нульову гіпотезу $H_0: M(X)=M(Y)$ про рівність математичних сподівань при альтернативній гіпотезі $H_1: M(X) > M(Y)$

Розв'язок. Знайдемо спостережене значення критерію

$$Z_{сн} = \frac{14.3 - 12.2}{\sqrt{22/10 + 18/10}} = 1.05$$

Альтернативна гіпотеза має вигляд $H_1: M(X) > M(Y)$, тому критична область – правостороння. По таблиці розподілу Лапласа $\Phi(z)$ за рівнем значущості $(1-2\alpha)/2$ знаходимо критичну точку $z_{кр} = 1.64$.

Оскільки $z_{сн} < z_{кр}$ - немає підстав відкинути нульову гіпотезу.

Порівняння двох середніх довільно розподілених генеральних сукупностей (великі незалежні вибірки)

У попередньому випадку передбачалося, що генеральні сукупності розподілені нормально, а їх дисперсії відомі. При цих припущеннях у разі справедливості нульової гіпотези про рівність середніх і незалежних вибірках критерій Z розподілений точно нормально з параметрами 0 і 1.

Якщо хоч би одна з приведених вимог не виконується, метод порівняння середніх, описаний вище, непридатний.

Проте якщо незалежні вибірки мають великий об'єм (не менше 30 кожна), то вибіркові середні розподілені приблизно нормально, а вибіркові

дисперсії є достатньо точними оцінками генеральних дисперсій і в цьому сенсі їх можна вважати відомими приблизно. В результаті критерій

$$Z' = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{D_B(X_1)}{n_1} + \frac{D_B(X_2)}{n_2}}}$$

має приблизно нормальний розподіл з параметрами $M(Z')=0$ (за умови справедливості нульової гіпотези) та $\sigma(Z')=1$ (якщо вибірки незалежні).

Отже, якщо:

- 1) генеральні сукупності розподілені нормально, а дисперсії їх невідомі,
 - 2) генеральні сукупності не розподілені нормально і дисперсії їх невідомі, причому вибірки мають великий об'єм і незалежні,
- можна порівнювати середні так, як описано в попередньому випадку, замінивши точний критерій Z наближеним критерієм Z' .

✓ Оскільки критерій, що використовується – наближений, до висновків, отриманих за цим критерієм, слід відноситися обережно.

Приклад. По двом незалежним вибіркам обсягом $n_1=100$ та $n_2=120$ зробленим з нормальних генеральних сукупностей X та Y знайдено вибіркові середні $\bar{x}=32.4$ та $\bar{y}=30.1$ та вибіркові дисперсії $D_B(X)=15.0$ та $D_B(Y)=25.2$. При рівні значущості 0.05 перевірити нульову гіпотезу $H_0: M(X)=M(Y)$ про рівність математичних сподівань при альтернативній гіпотезі $H_1: M(X)>M(Y)$

Розв'язок. Знайдемо спостережене значення критерію

$$Z'_{cn} = \frac{32.4 - 30.1}{\sqrt{15/100 + 25.2/120}} = 3.83$$

Альтернативна гіпотеза має вигляд $H_1: M(X)>M(Y)$, тому критична область – правостороння.

$$\Phi(z_{кр}) = (1 - 2\alpha)/2 = (1 - 2 \cdot 0.05)/2 = 0.45$$

По таблиці розподілу Лапласа $\Phi(z)$ знаходимо критичну точку $z_{кр} = 1.64$.

Оскільки $z_{сн} > z_{кр}$ - нульову гіпотезу відкидаємо, тобто вибіркові середні відрізняються суттєво.

Порівняння двох середніх нормальних генеральних сукупностей, дисперсії яких невідомі та однакові (малі незалежні вибірки)

Нехай дві генеральні сукупності розподілені нормально та їх дисперсії невідомі (наприклад, по вибіркам малого обсягу неможливо отримати хороші оцінки генеральних дисперсій).

Якщо припустити, що невідомі генеральні дисперсії однакові, то можна побудувати критерій Стюдента для порівняння середніх.

Якщо немає підстав вважати дисперсії однаковими, то до того, як порівнювати середні, слід, користуючись критерієм Фішера-Снедекора, попередньо перевірити гіпотезу про рівність генеральних дисперсій.

Як критерій перевірки нульової гіпотези про рівність вибірових середніх можна викоистати випадкову величину

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

Величина T має розподіл Стюдента при справедливості нульової гіпотези з $k = n_1 + n_2 - 2$ степенями свободи.

Критична область будується в залежності від вигляду альтернативної гіпотези.

Випадок перший. $M(X_1) = M(X_2)$, $H_1: M(X_1) \neq M(X_2)$.

Вибираємо двосторонню критичну область, виходячи з вимоги, що імовірність влучення критерію T в цю область при справедливості нульової гіпотези буде дорівнювати прийнятому рівню значущості α .

$$P(T < t_{\text{лів.кр}}) = \alpha / 2, \quad P(T > t_{\text{прав.кр}}) = \alpha / 2$$

Потужність критерію буде найбільшою, якщо критичні точки симетричні відносно нуля (T - симетрична)

$$P(T < -t_{\text{кр}}) = \alpha / 2, \quad P(T > t_{\text{кр}}) = \alpha / 2$$

Критичні точки $t_{\text{кр}}$ знаходимо по таблиці критичних точок розподілу Стюдента за рівнем значущості α та степенем свободи $k = n_1 + n_2 - 2$.

Критична область двостороння: $T < -t_{\text{кр}}$ або $t_{\text{кр}} < T$.

Область прийняття нульової гіпотези: $-t_{\text{кр}} < T < t_{\text{кр}}$.

Випадок другий. $H_0: M(X_1) = M(X_2)$, $H_1: M(X_1) > M(X_2)$.

Вибираємо правосторонню критичну область

$$P(T > t_{\text{прав.кр}}) = \alpha$$

Критичну точку $t_{\text{кр}}$ знаходимо по таблиці критичних точок розподілу Стюдента за рівнем значущості α та степенем свободи $k = n_1 + n_2 - 2$.

Критична область: $t_{\text{кр}} < T$.

Область прийняття нульової гіпотези: $T < t_{\text{кр}}$.

Випадок третій. $H_0: M(X_1) = M(X_2)$, $H_1: M(X_1) < M(X_2)$.

Вибираємо лівосторонню критичну область

$$P(T < t_{\text{лів.кр}}) = \alpha$$

В силу симетрії розподілу Стюдента відносно нуля $t_{\text{лів.кр}} = -t_{\text{прав.кр}}$. Тому спочатку знаходимо допоміжну критичну точку $t_{\text{прав.кр}}$ як у попередньому випадку, а потім покладаємо $t_{\text{лів.кр}} = -t_{\text{прав.кр}}$

Критична область: $T < t_{\text{кр.}}$.

Область прийняття нульової гіпотези: $t_{\text{кр}} < T$

Перевірка гіпотези про рівність вибіркової середньої генеральній середній нормальної сукупності, дисперсія якої відома

Нехай генеральна сукупність розподілена нормально, причому генеральна середня $M(X) = a$ хоч і невідома, але є підстави вважати, що вона дорівнює гіпотетичному значенню a_0 .

Наприклад, розглядається сукупність розмірів деталей, що виготовлені станком-автоматом, тоді можна припустити, що генеральна середня a цих розмірів дорівнює проектному розміру a_0 .

Дисперсія генеральної сукупності відома, наприклад, з попередніх дослідів або обчислена теоретично.

Із сукупності зробили вибірку об'єму n та знайшли вибірку середню \bar{x} , генеральна дисперсія σ^2 відома.

Потрібно по вибірковій середній при заданому рівні значущості перевірити гіпотезу про рівність генеральної середньої a гіпотетичному значенню a_0

$$H_0 : M(X) = a_0$$

Враховуючи, що вибірка середня є незсунутою оцінкою генеральної середньої

$$M(\bar{X}) = a$$

нульову гіпотезу можна записати так:

$$H_0 : M(\bar{X}) = a_0$$

Як критерій перевірки нульової гіпотези про рівність вибірових середніх приймемо випадкову величину

$$U = \frac{\bar{X} - a_0}{\sigma(\bar{X})} = \frac{(\bar{X} - a_0)\sqrt{n}}{\sigma}$$

Величина U має нормальний нормований розподіл, $M(U)=0$, $\sigma(U)=1$ при справедливості нульової гіпотези.

Критична область будується в залежності від вигляду альтернативної гіпотези.

Перевірка гіпотези про рівність вибіркової середньої генеральній середній нормальної сукупності, дисперсія якої невідома

Якщо дисперсія генеральної сукупності невідома, наприклад, у випадку малих вибірок, в якості критерію перевірки нульової гіпотези про рівність вибірових середніх приймемо випадкову величину

$$T = (\bar{X} - a_0)\sqrt{n}/S$$

Величина T має розподіл Стюдента при $k=n-1$ степенях свободи.

Критична область будується в залежності від вигляду альтернативної гіпотези.

Порівняння двох середніх нормальних генеральних сукупностей з невідомими дисперсіями (залежні вибірки)

У попередніх випадках розглядались незалежні вибірки. Тепер розглянемо вибірки однакового об'єму, варіанти яких попарно залежні. Наприклад, якщо x_i ($i=1,2,\dots,n$) – результати вимірювань деталей першим приладом, а y_i

– результати вимірювань цих же деталей, зроблені в тому ж порядку другим приладом, то x_i і y_i попарно залежні і у цьому сенсі самі вибірки залежні. Оскільки, як правило, $x_i \neq y_i$, то виникає необхідність встановити значуще або незначуще розрізняються пари цих чисел.

Аналогічне завдання ставиться при порівнянні двох методів дослідження, здійснених однією лабораторією або якщо дослідження проведене одним і тим же методом двома різними лабораторіями.

Нехай дві генеральні сукупності розподілені нормально, їх дисперсії невідомі. Із сукупностей зробили вибірки однакового об'єму n і знайшли вибіркові середні \bar{x} та \bar{y} .

Потрібно перевірити гіпотезу

$$H_0 : M(X) = M(Y)$$

про рівність математичних сподівань сукупностей з невідомими дисперсіями при альтернативній гіпотезі

$$H_1 : M(X) \neq M(Y)$$

по двом залежним вибіркам однакового обсягу.

Зведемо цю задачу порівняння двох середніх до задачі порівняння однієї вибіркової середньої з гіпотетичним значенням генеральної середньої, вирішеною раніше. З цією метою введемо в розгляд випадкові величини - різниці $D_i = X_i - Y_i$ і їх середню

$$\bar{D} = \frac{\sum D_i}{n} = \frac{\sum (X_i - Y_i)}{n} = \frac{\sum X_i}{n} - \frac{\sum Y_i}{n} = \bar{X} - \bar{Y}$$

Якщо нульова гіпотеза справедлива, то

$$M(\bar{X}) - M(\bar{Y}) = M(\bar{D}) = 0$$

Розраховуємо спостережене значення критерію

$$T_{cn} = \frac{\bar{d} \cdot \sqrt{n}}{s_d} = \frac{\bar{d} \cdot \sqrt{n}}{\sqrt{\frac{\sum d_i^2 - (\sum d_i)^2 / n}{n-1}}}$$

Критичну точку знаходимо по таблиці розподілу Стюдента. Якщо $-t_{кр} < T_{cn} < t_{кр}$, то немає підстав відхилити нульову гіпотезу. Якщо $T_{cn} < -t_{кр}$ або $t_{кр} < T_{cn}$, то нульову гіпотезу відхиляємо.

Приклад. Двома приладами в одному й тому ж порядку виміряно 5 деталей та отримано наступні результати:

$$\begin{array}{ccccc} x_1=6 & x_2=7 & x_3=8 & x_4=5 & x_5=7 \\ y_1=7 & y_2=6 & y_3=8 & y_4=7 & y_5=8 \end{array}$$

При рівні значущості 0.05 з'ясувати, суттєво чи несуттєво відрізняються результати вимірювань

Розв'язок. Віднімаючи від першого рядка другий отримаємо

$$d_1=-1 \quad d_2=1 \quad d_3=0 \quad d_4=-2 \quad d_5=-1$$

Вибіркова середня

$$\bar{d} = \sum d_i / n = (-1+1+0-2-1)/5 = -0.6$$

Виправлене середньоквадратичне відхилення

$$\begin{aligned} \sum d_i^2 &= 1+1+4+1=7 & \sum d_i &= 3 \\ s_d &= \sqrt{\frac{\sum d_i^2 - (\sum d_i)^2 / n}{n-1}} = \sqrt{\frac{7-9/5}{5-1}} = \sqrt{1.3} \end{aligned}$$

Знайдемо спостережене значення критерію

$$T_{cn} = \bar{d} \sqrt{n} / s_d = -0.6 \sqrt{5} / \sqrt{1.3} = -1.18$$

По таблиці розподілу Стюдента за рівнем значущості 0.05 та числу степенів свободи $k=n-1=5-1=4$ знаходимо критичну точку $t_{кр}=2.78$.

Оскільки $|T_{cn}| < t_{kp}$ – немає причин відкинути нульову гіпотезу, тобто результати вимірювань відрізняються несуттєво.

Гіпотези про числові характеристики

Найчастіше виникає потреба:

- порівняння спостереженої відносної частоти з гіпотетичною імовірністю появи події;
- порівняння двох імовірностей біноміальних розподілів;
- перевірка гіпотези про значущість вибіркового коефіцієнта кореляції.

Порівняння спостереженої відносної частоти з гіпотетичною імовірністю появи події

Нехай по досить великій кількості n незалежних випробувань, у кожному з яких імовірність p появи події постійна, але невідома, знайдена відносна частота m/n . Хай є підстави припускати, що невідома імовірність рівна гіпотетичному значенню p_0 .

Потрібно при заданому рівні значущості α перевірити нульову гіпотезу: невідома імовірність p рівна гіпотетичній імовірності p_0 .

Оскільки імовірність оцінюється по відносній частоті, дане завдання можна сформулювати і так: потрібно встановити, значуще або незначуще відрізняються спостережувана відносна частота і гіпотетична імовірність.

Доведено (теорема Лапласа), що при досить великих значеннях n відносна частота має приблизно нормальний розподіл з математичним сподіванням p і середнім квадратичним відхиленням $\sqrt{pq/n}$.

Нормуючи відносну частоту, отримаємо

$$U = \frac{M/n - p}{\sqrt{pq/n}} = \frac{(M/n - p)\sqrt{n}}{\sqrt{pq}} \quad \begin{array}{l} M(U) = 0, \\ \sigma(U) = 1 \end{array}$$

При справедливості нульової гіпотези, тобто $p = p_0$

$$U = \frac{(M/n - p_0)\sqrt{n}}{\sqrt{p_0 q_0}}$$

Спостережене значення критерію

$$U_{cn} = \frac{(m/n - p_0)\sqrt{n}}{\sqrt{p_0 q_0}}$$

Критична область будується в залежності від вигляду альтернативної гіпотези. Подальша перевірка стандартна.

Порівняння двох імовірностей біноміальних розподілів

Нехай в двох генеральних сукупностях проводяться незалежні випробування; в результаті кожного випробування подія A може з'явитися або не з'явитися.

Позначимо невідому імовірність появи події A в першій сукупності через p_1 , а в другій - через p_2 .

Припустимо, що в першій сукупності проведено n_1 випробувань, причому подія A спостерігалася m_1 раз. Отже, відносна частота появи події в першій сукупності

$$w_1(A) = m_1/n_1$$

Припустимо, що в другій сукупності проведено n_2 випробувань, причому подія A спостерігалася m_2 раз. Отже, відносна частота появи події в другій сукупності

$$w_2(A) = m_2/n_2$$

Приймемо відносні частоти, що спостерігалися, в якості оцінок невідомої імовірності появи події A : $p_1 \sim w_1$, $p_2 \sim w_2$.

Потрібно при заданому рівні значущості α перевірити нульову гіпотезу: імовірність p_1 і p_2 рівні між собою.

Оскільки імовірність оцінюється по відносним частотам, дане завдання можна сформулювати і так: потрібно встановити, значуще або незначуще розрізняються відносні частоти w_1 і w_2 .

Як критерій перевірки нульової гіпотези візьмемо випадкову величину

$$U = \frac{M_1 / n_1 - M_2 / n_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}$$

яка при справедливості нульової гіпотези розподілена приблизно нормально з параметрами $M(U)=0$, $\sigma(U)=1$.

У формулі імовірність p невідома, тому замінимо її оцінкою найбільшої правдоподібності

$$p^* = \frac{m_1 + m_2}{n_1 + n_2}$$

крім того, замінимо випадкові величини M_1 і M_2 їх значеннями m_1 і m_2 , отриманими у випробуваннях. В результаті отримаємо робочу формулу для обчислення спостереженого значення критерію:

$$U_{\text{сп}} = \frac{m_1 / n_1 - m_2 / n_2}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2} \left(1 - \frac{m_1 + m_2}{n_1 + n_2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Критична область будується в залежності від вигляду альтернативної гіпотези.

Перевірка гіпотези про значущість вибіркового коефіцієнта кореляції

Нехай двовимірна генеральна сукупність (X, Y) розподілена нормально. З цієї сукупності зроблена вибірка обсягу n і по ній знайдений вибірковий коефіцієнт кореляції $r_{\text{в}}$, який виявився відмінним від нуля. Оскільки

вибірка відібрана випадково, то ще не можна вважати, що коефіцієнт кореляції генеральної сукупності r_T також відмінний від нуля. Тобто виникає необхідність при заданому рівні значущості α перевірити нульову гіпотезу $H_0: r_T = 0$ про рівність нулю генерального коефіцієнта кореляції при конкуруючій гіпотезі $H_1: r_T \neq 0$.

Якщо нульова гіпотеза відкидається, то це означає, що вибірковий коефіцієнт кореляції значуще відрізняється від нуля, а X і Y корельовані, тобто зв'язані лінійною залежністю.

Якщо нульова гіпотеза буде прийнята, то вибірковий коефіцієнт кореляції незначущий, а X і Y некорельовані - не зв'язані лінійною залежністю.

Як критерій перевірки нульової гіпотези приймемо випадкову величину

$$T = \frac{r_B \sqrt{n-2}}{\sqrt{1-r_B^2}}$$

Величина T при справедливості нульової гіпотези має розподіл Стюдента з $k = n - 2$ степенями свободи. Оскільки конкуруюча гіпотеза має вигляд $r_T \neq 0$, критична область - двостороння.

Підсумки

- Вибір критерію залежить від гіпотези, що перевіряється, та наявної апіорної інформації.
- Потрібно акуратно вибирати статистику, яка буде оцінюватись за критерієм – вона повинна мати такий же закон розподілу імовірностей, як і критерій.
- При різних альтернативних гіпотезах критичні значення визначаються по-різному. Це варто мати на увазі як при користуванні таблицями, так і при використанні спеціальних статистичних пакетів.

3.7. Аналіз впливу факторів

Системи випадкових величин

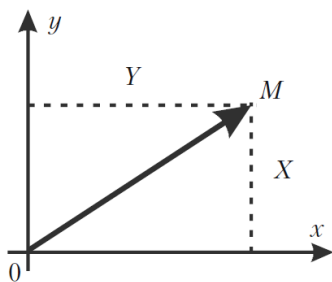
Розглянуті раніше випадкові величини при кожному випробуванні визначались одним можливим числовим значенням. Такі випадкові величини називають *одновимірними*.

Якщо можливі значення випадкової величини визначаються у кожному випробуванні 2, 3, ..., n числами, то такі величини називають дво-, три-, ..., n -вимірними відповідно.

Двовимірну випадкову величину будемо позначати (X, Y) , X та Y при цьому будуть компонентами. Величини X та Y , що розглядаються одночасно, утворюють систему двох випадкових величин.

- Сукупність n випадкових величин (X_1, X_2, \dots, X_n) , що розглядаються одночасно, називають *системою випадкових величин*.

Систему n випадкових величин (X_1, X_2, \dots, X_n) можна розглядати як випадкову точку в n -вимірному просторі з координатами (X_1, X_2, \dots, X_n) або як випадковий вектор, направлений з початку системи координат у точку $M(X_1, X_2, \dots, X_n)$. При $n = 2$ маємо систему двох випадкових величин (X, Y) .



При дослідженнях досить часто потрібно з'ясувати, чи взаємодіють між собою компоненти системи випадкових величин, і якщо це так, то наскільки сильний зв'язок між ними.

Формулювання гіпотези

Розглянемо систему випадкових величин. Нехай нам цікаво, від чого залежить поведінка однієї з її компонент. Цю компоненту будемо називати відгуком. Решта компонент – це фактори, які можуть впливати (або не впливати) на відгук.

Вибір фактора та відгука виконується на підставі:

- природи досліджуваної проблеми;
- інтуїції спеціаліста;
- досвіду аналогічних досліджень.

Як правило, гіпотеза H_0 формулюється як проста:

H_0 : фактор не впливає на відгук.

В цьому випадку, незалежно від того, яким критерієм для перевірки гіпотези будемо користуватись, можна звести інтерпретацію результатів до оцінки отриманого рівня значущості критерію. Якщо гіпотеза H_0 відхиляється, тобто вплив фактора суттєвий, обчислений рівень значущості буде менше заданого (як правило, 0.05).

При виконанні аналізу неважливо, яка із змінних є фактором, а яка відгуком – визначальними є шкали, в яких виміряні та представлені змінні. Тому на етапі інтерпретації результатів необхідно керуватися знанням досліджуваної проблеми та здоровим глуздом.

Результати досліджень рекомендується представляти не тільки в аналітичній, а і в графічній формі, яка добре їх ілюструє.

Залежні та незалежні вибірки

При розгляді системи випадкових величин можуть траплятися випадки, коли апріорно відомо про певний зв'язок між її компонентами. Наприклад,

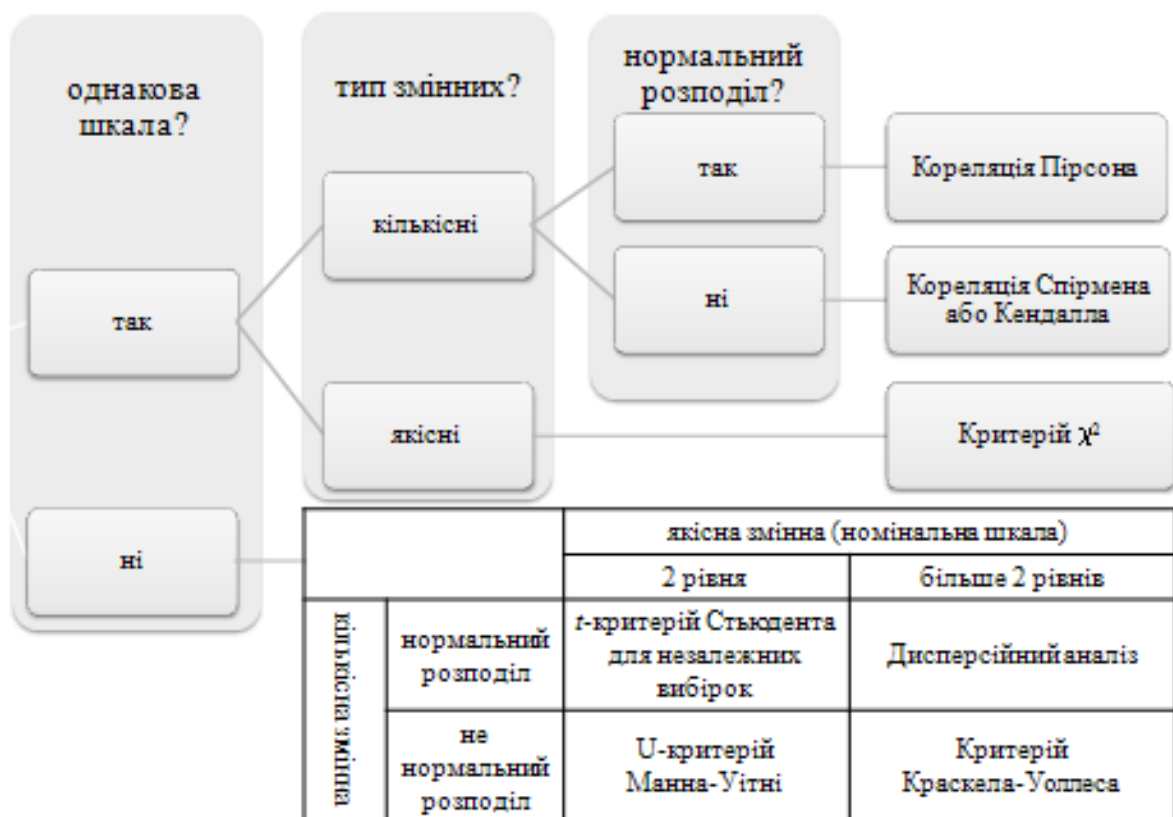
при тестуванні ліків міряємо рівень цукру в крові пацієнтів до та після їх прийому. В результаті формується дві вибірки, де для кожного пацієнта є відповідні виміри.

➤ *Залежними* називають вибірки, в яких кожній варіанті однієї вибірки відповідає варіанта в іншій вибірці.

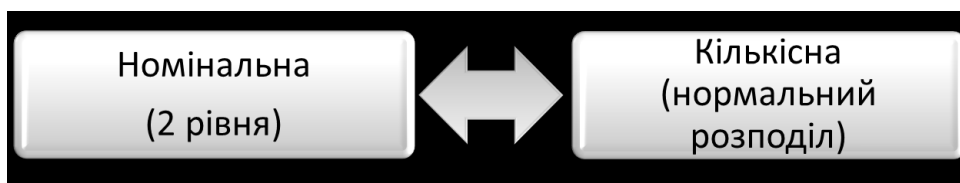
Методи аналізу таких даних дещо відрізняються від роботи з результатами незалежних випробувань, бо потрібно враховувати додаткову інформацію.

Вибір методу для аналізу впливу фактора при незалежних вибірках

Вибір методу для аналізу фактора залежить від того, чи в однакових шкалах виміряно змінні та типу змінних. У випадку, коли аналізуються кількісні змінні (фактор чи відгук, чи обидва), до вибору критерію узгодження необхідно виконати їх перевірку на нормальність розподілу. Від цього суттєво залежить можливість використання багатьох критеріїв узгодження.



t-критерій Стюдента



Хочемо перевірити: чи впливає якісний 2-рівневий фактор на кількісний відгук (або навпаки).

H_0 : якісний 2-рівневий фактор не впливає на кількісний відгук.

Математичний зміст критерію: перевірка рівності середніх в двох підгрупах кількісної змінної (підгрупи відповідають рівням якісної змінної).

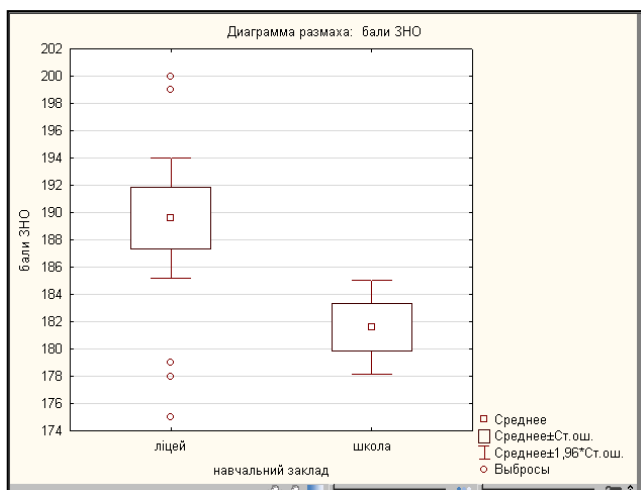
Вимоги до даних:

- Якісна змінна вимірювалась номінальною шкалою з двома рівнями;
- Дані всередині порівнюваних підгруп (тобто підгруп кількісної змінної, що відповідають рівням якісної змінної) розподілені нормально;
- Кожному рівню якісної змінної відповідає не менше 30 значень кількісної змінної (для можливості перевірки підгруп на нормальність).

Загальний алгоритм:

1. Перевірка підгруп кількісної шкали на нормальність;
2. Розрахунок значення t -критерія Стьюдента та оцінка його рівня значущості (якщо рівень значущості менше 0.05, то фактор суттєво впливає на кількісний відгук – нульова гіпотеза відкинута);
3. Порівняння середніх значень кількісної змінної по підгрупам.

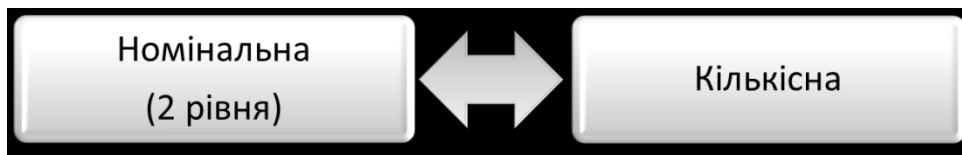
Приклад. Чи суттєво відрізняються бали ЗНО випускників середніх шкіл та ліцеїв?



Т-критерий; Группир.: навчальний заклад (Tab					
Группа 1:ліцей					
Группа 2:школа					
Переменная	Среднее ліцей	Среднее школа	t-знач.	сс	p
бали ЗНО	189,6000	181,6000	2,802006	38	0,007948

Різниця між балами ЗНО випусників середніх шкіл та ліцеїв суттєва. Бали ЗНО випусників ліцеїв в середньому на 8 вищі, ніж у випусників шкіл.

U-критерій Манна-Уїтні



Хочемо перевірити: чи впливає якісний 2-рівневий фактор на кількісний відгук (або навпаки).

H_0 : якісний 2-рівневий фактор не впливає на кількісний відгук.

Математичний зміст критерію: перевірка рівності медіан в двох підгрупах кількісної змінної (підгрупи відповідають рівням якісної змінної).

Вимоги до даних:

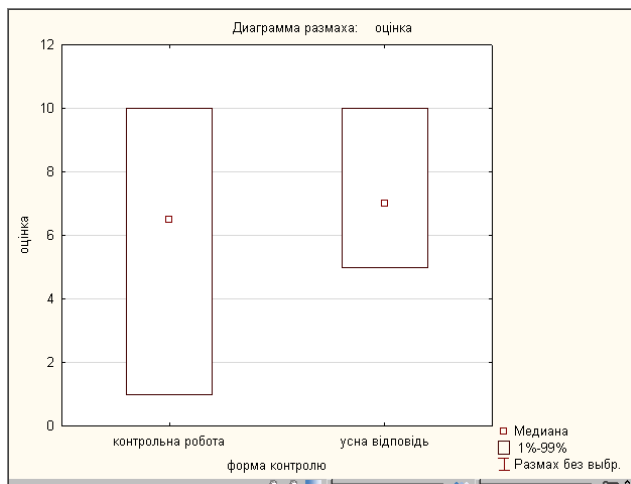
- Якісна змінна вимірювалась номінальною шкалою з двома рівнями;

- Дані всередині порівнюваних підгруп (тобто підгруп кількісної змінної, що відповідають рівням якісної змінної) можуть мати довільний розподіл.

Загальний алгоритм:

1. Перевірка підгруп кількісної шкали на нормальність;
2. Розрахунок значення U-критерія Манна-Уїтні оцінка його рівня значущості (якщо рівень значущості менше 0.05, то фактор суттєво впливає на кількісний відгук – нульова гіпотеза відкинута);
3. Порівняння медіан кількісної змінної по підгрупах.

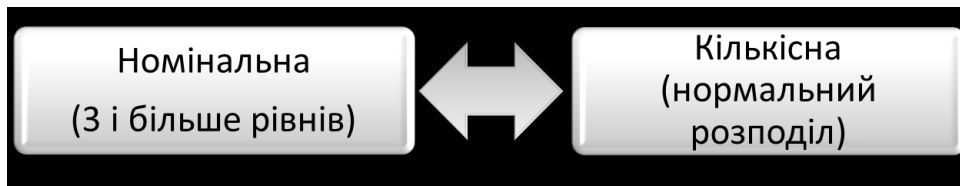
Приклад. Чи суттєво відрізняються оцінки студентів при різних формах контролю знань?



U критерий Манна-Уитни (Таблица данных1)					
По перем. форма контролю					
Отмеченные критерии значимы на уровне $p < 0,05000$					
Перем.	Сум. ранг контрольна робота	Сум. ранг усна відповідь	U	Z	p-уров.
оцінка	361,0000	459,0000	151,0000	-1,31193	0,189545

Різниця між оцінками за контрольні та усні відповіді несуттєва.

Дисперсійний аналіз (ANOVA)



Хочемо перевірити: чи впливає якісний багаторівневий фактор на кількісний відгук (або навпаки).

H_0 : якісний фактор не впливає на кількісний відгук.

Математичний зміст критерію: перевірка рівності середніх в кількох підгрупах кількісної змінної (підгрупи відповідають рівням якісної змінної).

Вимоги до даних:

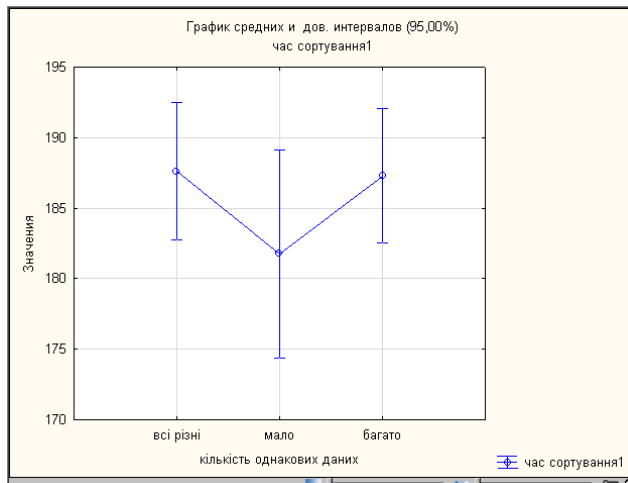
- Якісна змінна вимірювалась номінальною шкалою з 3 або більше рівнями;
- Дані всередині порівнюваних підгруп (тобто підгруп кількісної змінної, що відповідають рівням якісної змінної) розподілені нормально;
- Кожному рівню якісної змінної відповідає не менше 30 значень кількісної змінної (для можливості перевірки підгруп на нормальність) – вимога не строга, але виконання бажане.

Загальний алгоритм:

1. Перевірка підгруп кількісної шкали на нормальність, якщо хоча б в одній підгрупі дані розподілені не нормально – критерій не підходить;
2. Розрахунок значення F -критерія Фішера та оцінка його рівня значущості (якщо рівень значущості менше 0.05, то є рівні фактору, що суттєво впливають на кількісний відгук);

3. Порівняння обсягів підгруп та виконання парних тестів HSD (Tukey honest significant difference, з корекцією, якщо обсяги підгруп різні);
4. Порівняння середніх значень кількісної змінної по підгрупах, що суттєво відрізняються.

Приклад. Чи впливає кількість однакових даних на час сортування вибором?



Переменная	Дисперсионный анализ (Таблица данных1)							
	Отмечены эффекты, значимые на уров. $p < ,05000$							
	Сум. квадрат эффект	Ст. св. эффект	Ср. квадрат эффект	Сум. квадрат ошибки	Ст. св. ошибки	Ср. квадрат ошибки	F	p
час сортування1	283,3582	2	141,6791	3454,242	37	93,35789	1,517591	0,232572

Різниця між часом сортування повністю різних та частково однакових даних несуттєва.

Детальніше про дисперсійний аналіз

На практиці дисперсійний аналіз застосовують, щоб встановити, чи істотно впливає певний якісний фактор F , який має p рівнів F_1, F_2, \dots, F_p на кількісну ознаку X , що вивчається.

Нехай генеральні сукупності X_1, X_2, \dots, X_p розподілені нормально і мають однакову, хоч і невідому, дисперсію; математичні сподівання також невідомі, але можуть бути різними. Потрібно при заданому рівні значущості по вибірковим середнім перевірити нульову гіпотезу

$$H_0: M(X_1) = M(X_2) = \dots = M(X_p)$$

про рівність всіх математичних сподівань. Іншими словами, потрібно встановити, значуще чи незначуще розрізняються вибіркові середні.

Здавалося б для порівняння декількох середніх ($p > 2$) можна порівняти їх попарно. Проте із зростанням числа середніх зростає і найбільша відмінність між ними: середнє нової вибірки може опинитися більше найбільшого або менше найменшого з середніх, отриманих до нового випробування. З цієї причини для порівняння декількох середніх користуються іншим методом, який заснований на порівнянні дисперсій і тому названий **дисперсійним аналізом** (в основному розвинений англійським статистиком Р. Фішером).

Основна ідея дисперсійного аналізу полягає в порівнянні «дисперсії фактора», що породжується його дією, і «залишкової дисперсії», обумовленої випадковими причинами. Якщо відмінність між цими дисперсіями значима, то фактор має істотний вплив на X ; в цьому випадку середні спостережуваних значень на кожному рівні (групові середні) відрізняються також значуще.

Якщо вже встановлено, що фактор істотно впливає на X , а потрібно з'ясувати, який з рівнів має найбільший вплив, то додатково виконують попарне порівняння середніх.

Наприклад, якщо потрібно з'ясувати, що найбільше впливає на швидкість сортування бульбашковим методом, то в якості фактору F можна розглянути обсяг масиву, попередню впорядкованість даних, кількість різних значень даних тощо.

Рівнями фактору будуть

для обсягу:

$F_1 - 100$,

$F_2 - 1000$,

$F_3 - 3000$;

для впорядкованості:

F_1 – невлпорядкований,

F_2 – частково впорядкований,

F_3 – майже впорядкований.

Нехай на кількісну нормально розподілену ознаку X впливає фактор F , що має p постійних рівнів. Вважаємо, що кількість спостережень (випробувань) на кожному рівні однакова і дорівнює q .

Нехай спостережувалось $n=pq$ значень x_{ij} ознаки X , де i – номер випробування ($i=1,2,\dots,q$), j – номер рівня фактора ($j=1,2,\dots,p$).

Номер випробування	Рівні фактора F_j			
	F_1	F_2	...	F_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
...	
q	x_{q1}	x_{q2}	...	x_{qp}
Групова середня	$\bar{x}_{гр1}$	$\bar{x}_{гр2}$...	$\bar{x}_{грp}$

Позначимо загальну суму квадратів відхилень спостережених значень від загальної середньої \bar{x}

$$S_{\text{загальна}} = \sum_{j=1}^p \sum_{i=1}^q (x_{ij} - \bar{x})^2$$

факторну суму квадратів відхилень спостережених значень від загальної середньої, яка характеризує розсіювання між групами

$$S_{\text{факт}} = q \sum_{j=1}^p (\bar{x}_{Грj} - \bar{x})^2$$

залишкову суму квадратів відхилень спостережених значень від своєї групової середньої, яка характеризує розсіювання всередині груп

$$S_{\text{зал}} = \sum_{j=1}^p \sum_{i=1}^q (x_{ij} - \bar{x}_{Грj})^2 = S_{\text{загальна}} - S_{\text{факт}}$$

Переконаємося, що $S_{\text{факт}}$ характеризує дію фактора F . Припустимо, що фактор має істотний вплив на X . Тоді група спостережуваних значень ознаки на одному певному рівні відрізняється від груп спостережень на інших рівнях. Отже, розрізняються і групові середні, причому вони тим більше розсіяні навколо загальної середньої, чим більшою виявиться дія фактора. Звідси випливає, що для оцінки дії фактора доцільно скласти суму квадратів відхилень групових середніх від загальної середньої (відхилення зводять в квадрат щоб виключити погашення додатних і від'ємних відхилень). Помноживши цю суму на q , отримаємо $S_{\text{факт}}$. Отже, $S_{\text{факт}}$ характеризує дію фактора.

Переконаємося, що $S_{\text{зал}}$ характеризує вплив випадкових причин. Здавалося б, спостереження однієї групи не повинні розрізнятися. Проте, оскільки на X , окрім фактора F , впливають і випадкові причини, спостереження однієї і тієї ж групи різні і, значить, розсіяні навколо своєї групової середньої.

Звідси витікає, що для оцінки впливу випадкових причин доцільно скласти суму квадратів відхилень спостережуваних значень кожної групи від своєї групової середньої. Отже, $S_{\text{зал}}$ характеризує дію випадкових причин.

Переконаємося, що $S_{\text{загальна}}$ характеризує вплив і фактора і випадкових причин. Розглядатимемо всі спостереження як єдину сукупність. Спостережувані значення ознаки різні внаслідок дії фактора і випадкових причин. Для оцінки цієї дії доцільно скласти суму квадратів відхилень спостережуваних значень від загальної середньої. Отже, $S_{\text{загальна}}$ характеризує вплив фактора і випадкових причин.

Поділивши суми квадратів відхилень на відповідне число степенів свободи, отримаємо загальну, факторну та залишкову дисперсії

$$s_{\text{загальна}}^2 = \frac{S_{\text{загальна}}}{pq-1}$$

$$s_{\text{факт}}^2 = \frac{S_{\text{факт}}}{p-1}$$

$$s_{\text{зал}}^2 = \frac{S_{\text{зал}}}{p(q-1)}$$

де p – число рівнів фактора, q – число спостережень на кожному рівні, $(pq-1)$ – число степенів свободи загальної дисперсії, $(p-1)$ – число степенів свободи факторної дисперсії, $p(q-1)$ – число степенів свободи залишкової дисперсії.

Якщо гіпотеза про рівність середніх справедлива, то всі ці дисперсії є незсунутими оцінками генеральної дисперсії.

Для того, щоб перевірити нульову гіпотезу про рівність групових середніх нормальних сукупностей з однаковими дисперсіями, достатньо перевірити

по критерію Фішера гіпотезу про рівність факторної та залишкової дисперсії.

$$F = \frac{s_{\text{факт}}^2}{s_{\text{зал}}^2}$$

Критичну область знаходять з урахуванням умови

$$P(F > f_{\alpha}) = \alpha$$

де f_{α} – критичне (і табульоване) значення розподілу Фішера з $(p-1)$ та $p(q-1)$ степенями свободи.

Зауваження:

- ✓ Якщо факторна дисперсія виявиться менше залишкової, то вже звідси слідує справедливості гіпотези про рівність групових середніх і, значить, немає потреби вдаватися до критерію F .
- ✓ Якщо немає впевненості в справедливості припущення про рівність дисперсій p сукупностей, що розглядаються, то це припущення слід перевірити заздалегідь, наприклад по критерію Кохрена.

Приклад. Є дані про час виконання програми (в секундах) при трьох тестових запусках на різних комп'ютерах лабораторії

ПК1	ПК2	ПК3	ПК4	ПК5
10.2	10.8	10.7	13	12
11.5	9.8	11.5	13.2	11.5
12	12.1	12	11.5	11.8

Припускаючи нормальний закон розподілу часу виконання для кожного ПК та рівність дисперсій, перевірити гіпотезу $H_0: a_1 = a_2 = a_3 = a_4 = a_5$ при рівні значущості $\alpha = 0.05$, тобто час виконання програми не залежить від обраного комп'ютера.

Розв'язок. Умови прикладу дозволяють застосувати до розв'язання задачі критерій дисперсійного аналізу. У цьому випадку маємо: число рівнів фактора $p = 5$; число спостережень на кожному рівні $q = 3$.

За формулами знаходимо групові та загальну середні

$$x_1 = 11.2; x_2 = 10.8; x_3 = 11.4; x_4 = 12.6; x_5 = 11.8; \bar{x} = 11.6$$

Зробимо обчислення сум

$$S_{\text{факт.}} = \sum_{i=1}^5 (x_{pi} - \bar{x})^2 q = 0.48 + 1.92 + 0.12 + 3 + 0.12 = 5.64$$

$$S_{\text{залишк.}} = \sum_{i=1}^5 \sum_{j=1}^3 (x_{ij} - \bar{x})^2 = 12.34$$

Тепер знайдемо значення статистичної характеристики

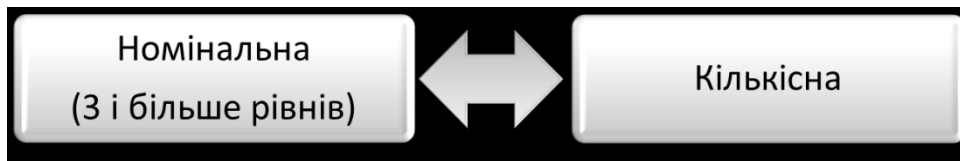
$$F = \frac{\frac{1}{4} \cdot 5.64}{\frac{1}{10} \cdot 12.34} = 1.14$$

Із таблиці критичних значень розподілу Фішера за степенями свободи факторної та залишкової дисперсій ($p-1=5-1=4$ та $p(q-1)=5(3-1)=10$) і рівнем значущості $\alpha = 0.05$ знаходимо

$$F_{kp} = f_{0.05} = 3.48$$

Одержали, що $F_{cn}=1.14 < F_{kp}=3.48$, тому немає підстав відхилити гіпотезу H_0 .

Критерій Краскела-Уоллеса



Хочемо перевірити: чи впливає якісний багаторівневий фактор на кількісний відгук (або навпаки).

H_0 : якісний фактор не впливає на кількісний відгук.

Математичний зміст критерію: перевірка рівності медіан в кількох підгрупах кількісної змінної (підгрупи відповідають рівням якісної змінної).

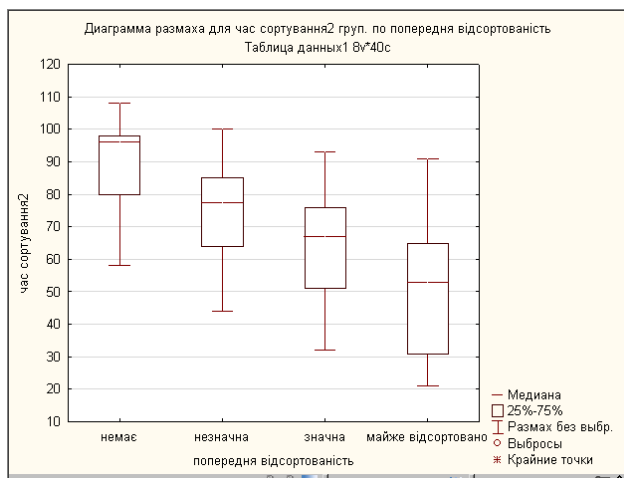
Вимоги до даних:

- Якісна змінна вимірювалась номінальною шкалою з 3 або більше рівнями;
- Дані всередині порівнюваних підгруп (тобто підгруп кількісної змінної, що відповідають рівням якісної змінної) розподілені довільно.

Загальний алгоритм:

1. Перевірка підгруп кількісної шкали на нормальність, щоб дізнатись про можливість використання ANOVA;
2. Розрахунок значення критерію та оцінка його рівня значущості (якщо рівень значущості менше 0.05, то є рівні фактору, що суттєво впливають на кількісний відгук – нульова гіпотеза відкинута);
3. Виконання парних тестів;
4. Порівняння медіан кількісної змінної по підгрупам, що суттєво відрізняються.

Приклад. Чи впливає ступінь попередньої відсортованості даних на час сортування вибором?

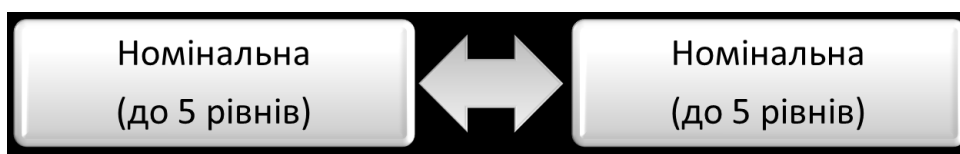


Зависим.: час сортування2	Ранговый ДА Краскела-Уоллиса; час сортування2 (Таблиця даних1) переменная: попередня відсортованість Кр.Краскела-Уоллиса: $H(3, N=40) = 14,91297$ $p = ,0019$					
	Код	Допуст N	Сумма Рангов	Среднее Ранг		
	немає	1	10	307,0000	30,70000	
	незначна	2	10	227,5000	22,75000	
	значна	3	10	172,0000	17,20000	
	майже відсортовано	4	10	113,5000	11,35000	

Зависим.: час сортування2	р знач. (2-сторонние) для множеств. сравнений; час сортування2 (Таблиця даних1) переменная: попередня відсортованість Кр.Краскела-Уоллиса: $H(3, N=40) = 14,91297$ $p = ,0019$				
	немає R:30,700	незначна R:22,750	значна R:17,200	майже відсортовано R:11,350	
	немає	0,770131	0,058906	0,001288	
	незначна	0,770131	1,000000	0,175317	
	значна	0,058906	1,000000	1,000000	
	майже відсортовано	0,001288	0,175317	1,000000	

Різниця між часом сортування випадкових та частково відсортованих даних суттєва.

Критерій χ^2



Хочемо перевірити: чи впливає якісний багаторівневий фактор на якісний багаторівневий відгук (або навпаки).

H_0 : якісний фактор не впливає на якісний відгук.

Математичний зміст критерію зводиться до виконання наступних кроків:

1. побудова таблиці, що містить фактичні частоти;
2. знаходження теоретичних значень частот;
3. перевірка близькості фактичних та теоретичних частот.

Вимоги до даних:

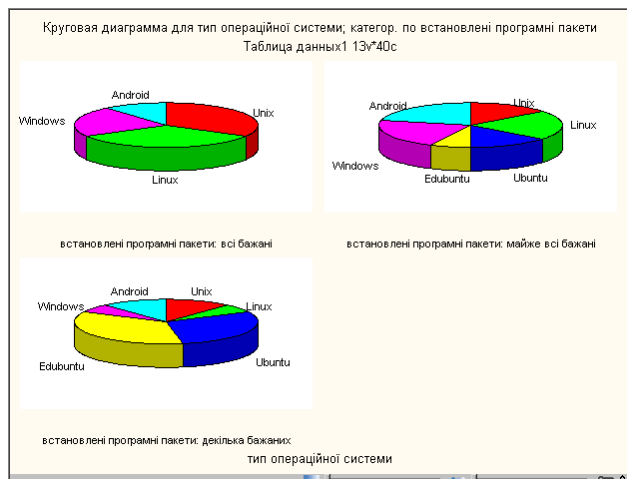
- Якісні змінні вимірювались номінальною шкалою не більше ніж з 5 рівнями (інакше знайдемо те, чого немає);
- Не більше 20% теоретичних частот мають значення менше 5.

Загальний алгоритм:

1. Побудова кругових діаграм для візуальної оцінки;
2. Аналіз можливості застосування критерію (не більше 20% теоретичних частот менше 5);
3. Розрахунок значення χ^2 - критерію для всіх підгруп в цілому та оцінка його рівня значущості (якщо рівень значущості менше 0.05, то фактор суттєво впливає на відгук – нульова гіпотеза відкинута);
4. Розрахунок значення χ^2 - критерію для всіх пар підгруп.

Приклад. Чи впливає тип операційної системи на можливість використання типових програмних пакетів?

Якісний фактор – тип операційної системи (6 рівнів), якісний відгук – можливість встановлення програмних пакетів (3 рівня).



Итоговая таблица частот (Таблица данных1)
Частоты выделенных ячеек > 4
(Маргинальные суммы не отмечены)

тип операційної системи	встановлені програмні пакети всі бажані	встановлені програмні пакети майже всі бажані	встановлені програмні пакети декілька бажаних	Всего по стр.
Unix	19	12	12	43
Linux	15	15	10	40
Ubuntu	5	12	28	45
Edubuntu	6	11	29	46
Windows	12	12	12	36
Android	10	11	9	30
Всего	67	73	100	240

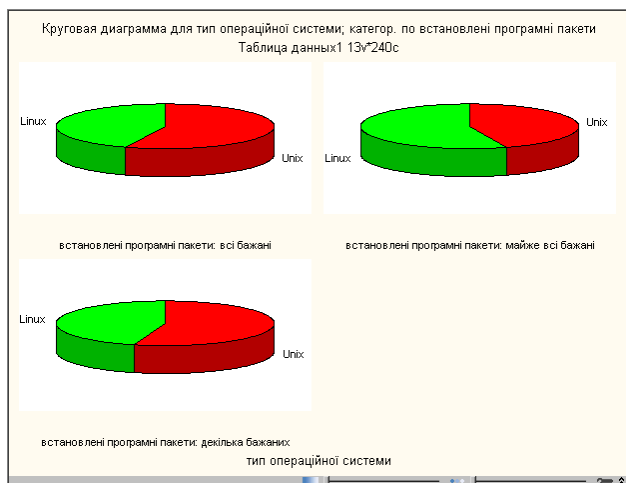
Итог. таблица: Ожидаемые частоты (Таблица данных1)
Частоты выделенных ячеек > 4
хи-квадрат Пирсона: 32,1756, cc=10, **0,00004**

тип операційної системи	встановлені програмні пакети всі бажані	встановлені програмні пакети майже всі бажані	встановлені програмні пакети декілька бажаних	Всего по стр.
Unix	12,00417	13,07917	17,9167	43,0000
Linux	11,16667	12,16667	16,6667	40,0000
Ubuntu	12,56250	13,68750	18,7500	45,0000
Edubuntu	12,84167	13,99167	19,1667	46,0000
Windows	10,05000	10,95000	15,0000	36,0000
Android	8,37500	9,12500	12,5000	30,0000
Всего	67,00000	73,00000	100,0000	240,0000

В таблиці немає значень теоретичних частот, що менше 5, отже можна продовжити аналіз. Якби такі частоти були виявлені, потрібно було б зменшити кількість рівнів фактору (укрупнити групи).

Тип встановленої операційної системи суттєво впливає на можливість використання типових програмних пакетів.

Найбільше можливостей надає операційна система Unix, найменше Edubuntu.



Итог. таблица: Ожидаемые частоты (Таблица данных1)
Частоты выделенных ячеек > 4
хи-квадрат Пирсона: ,878454, cc=2,07,044335

тип операційної системи	встановлені програмні пакети всі бажані	встановлені програмні пакети майже всі бажані	встановлені програмні пакети декілька бажаних	Всего по стр.
Unix	17,61446	13,98795	11,39759	43,00000
Linux	16,38554	13,01205	10,60241	40,00000
Всего	34,00000	27,00000	22,00000	83,00000

Різниця між Unix і Linux несуттєва.

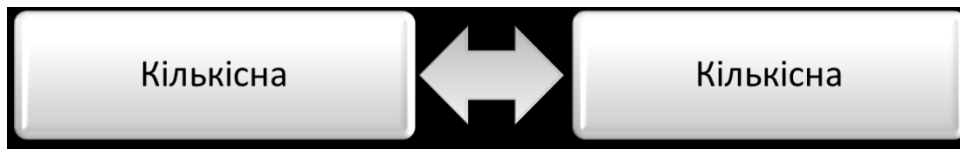


Итог. таблица: Ожидаемые частоты (Таблица данных1)
Частоты выделенных ячеек > 4
хи-квадрат Пирсона: 8,38588, cc=2, $p=0,015102$

тип операційної системи	встановлені програмні пакети всі бажані	встановлені програмні пакети майже всі бажані	встановлені програмні пакети декілька бажаних	Всього по стр.
Ubuntu	9,44444	13,33333	22,22222	45,00000
Windows	7,55556	10,66667	17,77778	36,00000
Всього	17,00000	24,00000	40,00000	81,00000

Різниця між Ubuntu і Windows суттєва.

Коефіцієнт кореляції



Хочемо перевірити: чи впливає кількісний фактор на кількісний відгук.

H_0 : фактор не впливає на відгук.

Математичний зміст критерію: оцінка величини коефіцієнта кореляції – міри лінійного зв'язку між фактором та відгуком.

Значення коефіцієнта кореляції може лежати в межах $-1 \leq r \leq 1$. Тракується коефіцієнт наступним чином. Тип зв'язку між змінними залежить від знаку коефіцієнта кореляції:

- $r > 0$ - зв'язок прямий;
- $r < 0$ - зв'язок обернений.

Сила зв'язку між змінними залежить від величини коефіцієнта кореляції:

- $0.75 \leq r$ – зв'язок сильний;
- $0.25 \leq r \leq 0.75$ – зв'язок помірний;
- $r < 0.25$ – зв'язок слабкий;
- r близький до 0 та рівень значущості більше 0.05 – відсутність зв'язку.

Вибір методу кореляційного дослідження залежить від нормальності розподілу даних та типів шкали, в яких виміряні змінні.



Коефіцієнт кореляції Пірсона між двома змінними дорівнює коваріації двох змінних, або сумі добутків відхилень, поділеній на добуток їх стандартних відхилень.

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}}$$

Коефіцієнт кореляції Спірмена визначається як коефіцієнт кореляції Пірсона між ранжованими змінними. Для вибірки обсягу n множини X_i, Y_i перетворюються в ряди x_i, y_i та обчислюється наступним чином.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Однаковим значенням (ранг зв'язків або величина дублікатів) присвоюється ранг, що дорівнює середньому числу їхніх позицій в порядку зростання величини. У наведеній нижче таблиці зверніть увагу, що ранг значень x_i при однаковій величині змінної X_i є однаковими:

Зміна X_i	Позиція в порядку зростання	Ранг x_i
0.8	1	1
1.2	2	$(2+3)/2=2.5$
1.2	3	$(2+3)/2=2.5$
2.3	4	4
18	5	5

Коефіцієнт кореляції Кендала є мірою рангової кореляції, тобто подібності упорядкування даних, коли вони упорядкованні за своєю величиною.

$$\tau = \frac{s_1 - s_2}{\frac{1}{2}n(n-1)}$$

де s_1 - кількість узгоджених пар, s_2 - кількість неузгоджених пар.

Нехай $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ - набір спільних спостережень випадкових величин X і Y відповідно, так що всі значення (x_k) і (y_k) не є однаковими для будь якого $k=1..n$.

Будь-яка пара спостережень (x_i, y_i) і (x_j, y_j) називається узгодженою, якщо узгоджені ряди для обох елементів: тобто, якщо $x_i > x_j$ та $y_i > y_j$ або якщо $x_i < x_j$ та $y_i < y_j$. Вони називаються неузгодженими (або дисонуючими), якщо $x_i > x_j$ та $y_i < y_j$ або якщо $x_i < x_j$ та $y_i > y_j$.

Якщо $x_i = x_j$ або $y_i = y_j$, то пара не є ні узгодженою ні неузгодженою.

Вимоги до даних.

На результати дослідження сильно впливають:

- викиди та грубі помилки (їх видно на діаграмі розмаху);
- неоднорідність даних (кілька ізольованих груп даних, їх добре видно на діаграмі розсіювання);

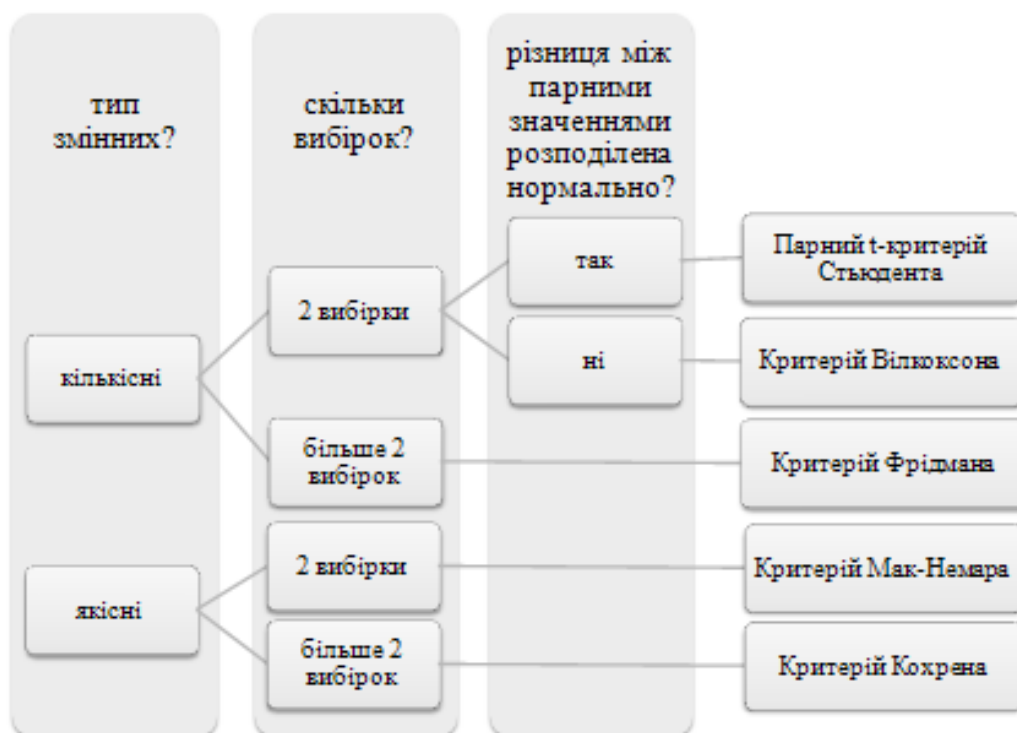
- нелінійна залежність.
- ✓ Обов'язково проводити попередній аналіз вибірки.

Загальний алгоритм:

1. Перевірка фактору та відгуку на нормальність;
2. Вибір методу дослідження;
3. Побудова діаграми розсіювання з лінією регресії;
4. Розрахунок коефіцієнта кореляції і його рівня значущості та їх оцінка.

Вибір методу для аналізу впливу фактора при залежних вибірках

Для обробки таких даних вибірки об'єднують в одну, а номер вибірки вказують у вигляді значення фактора (додаткова якісна змінна). При такому підході дані будуть розглядатися без фіксації номеру випробування та усереднено по всій серії випробувань.



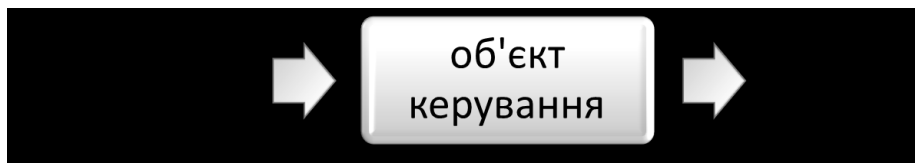
Підсумки

- При дослідженнях досить часто потрібно з'ясувати, чи взаємодіють між собою компоненти системи випадкових величин, і якщо це так, то наскільки сильний зв'язок між ними. Це можна зробити, використовуючи методи аналізу впливу факторів.
- Вибір фактора та відгука виконується на підставі:
 - природи досліджуваної проблеми;
 - інтуїції спеціаліста;
 - досвіду аналогічних досліджень.
- При виконанні аналізу неважливо, яка із змінних є фактором, а яка відгуком – визначальними є шкали, в яких виміряні та представлені змінні.
- Якщо і фактор, і відгук є кількісними змінними, то можна не тільки зафіксувати наявність зв'язку між ними, але й оцінити його напрям і силу.
- Методи аналізу даних залежних вибірок відрізняються від роботи з результатами незалежних випробувань, бо потрібно враховувати додаткову інформацію.
- Результати досліджень рекомендується представляти не тільки в аналітичній, а і в графічній формі, яка добре їх ілюструє.

3.8. Елементи регресійного аналізу

Для розв'язку різних задач керування, прогнозування та інших необхідно мати математичні моделі відповідних об'єктів або процесів. Для їх побудови можна використати різні підходи:

- фізичний
- статистичний
- комбінований



Фізичний підхід складається в побудові моделей об'єктів на основі їх фізичної структури і значень їх фізичних параметрів.

Нехай маємо RC ланцюг, що описується рівнянням

$$\underbrace{R \cdot C}_K \cdot \frac{dy}{dt} + y = u$$

В даному випадку структура моделі задається порядком диференційного рівняння, а параметр K визначається значеннями R і C . Модель буде точною, якщо точно знаємо параметри R і C (опір та ємність), які на практиці ідеально знати неможливо.

Статистичний підхід складається в визначенні моделі об'єкту на основі обробки інформації, що міститься в деяких реалізаціях $\{u(t), y(t)\}_{\Delta t}$ вхідної та вихідної величин на деякому інтервалі Δt , або в деякій вибірці.

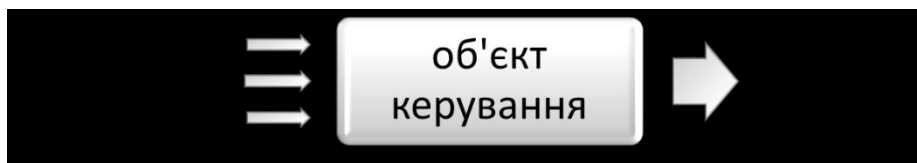
$$\{u(t_i), y(t_i)\}_n, \quad i = \overline{1, n}$$

Комбінований підхід складається в об'єднанні перших двох. При цьому фізичний підхід використовується для вибору структури моделі, а статистичний – для визначення її параметрів.

Слід відмітити, що використана інформація $\{u(t), y(t)\}_{\Delta t}$ або $\{u(t_i), y(t_i)\}_n$ містить випадкові завади, у зв'язку з чим неможливо знайти точно структуру та параметри об'єктів дослідження, а можна знайти лише їх оцінки. Для цього потрібні відповідні методи.

Постановка задачі

- *Задача регресії* складається у визначенні структури та параметрів об'єктів дослідження на основі інформації, що міститься у вибірках вхідних та вихідних величин.



Задача, яку будемо вирішувати: побудувати модель, що описує реакцію об'єкту керування у вигляді відгуку на вплив факторів.

Проміжні задачі:

1. які з факторів впливають на відгук, які ні;
2. ранжування факторів за ступенем впливу на відгук;
3. прогнозування значень відгуку.

Різновиди регресії

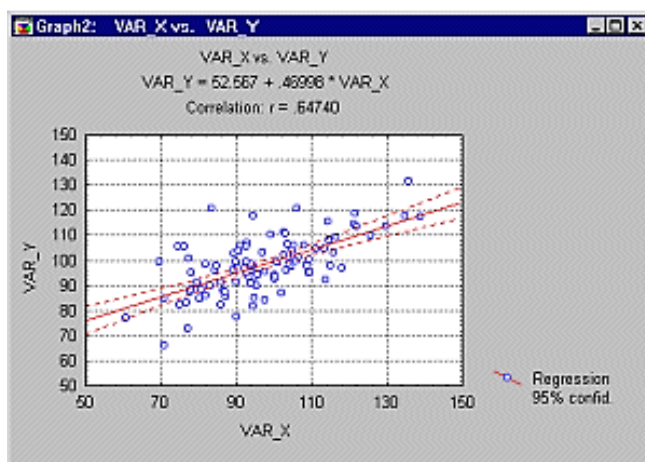
Існують різноманітні види регресійного аналізу - одномірний і багатомірний, лінійний і нелінійний, параметричний і непараметричний.

Для проведення лінійного регресійного аналізу залежна змінна повинна мати інтервальну (або порядкову) шкалу. Бінарна логістична регресія

виявляє залежність дихотомічної змінної від якоїсь іншої змінної, що вимірюється за будь-якою шкалою. Якщо залежна змінна є категоріальною, але має більше двох категорій, то тут відповідним методом буде поліноміальна логістична регресія. Порядкову регресію можна використовувати, коли залежні змінні відносяться до порядкової шкали. І, звичайно ж, можна аналізувати і нелінійні зв'язки між змінними, які вимірюються за інтервальною чи відносною шкалою. Для цього призначений метод нелінійної регресії.

Метою процедур лінійної регресії є підгонка прямої лінії до деякого набору точок так, щоб мінімізувати квадрати відхилень цієї лінії від спостережуваних точок.

У найпростішому випадку, коли є одна залежна і одна незалежна змінна, це можна побачити на діаграмі розсіювання.



Зазвичай, ступінь залежності двох або більше факторів (незалежних змінних або змінних X) з відгуком (залежною змінною Y) виражається за допомогою коефіцієнта множинної кореляції R . Для інтерпретації напрямку зв'язку між змінними дивляться на знаки (плюс або мінус) регресійних коефіцієнтів чи b -коефіцієнтів.

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_k \cdot X_k$$

Регресійні коефіцієнти представляють незалежні внески кожного фактору в передбачення відгуку. Іншими словами, змінна X_1 , наприклад, корелює зі змінною Y після врахування впливу всіх інших незалежних змінних. Цей тип кореляції згадується також під назвою часткової кореляції.

Вимоги до даних: обсяг вибірки $\geq 10 \cdot$ кількість факторів.

Обов'язково необхідно проводити попередній аналіз вибірки, оскільки на результати дослідження сильно впливають:

- викиди та грубі помилки;
- неоднорідність даних;
- нелінійна залежність.

Можна скористатися методами перетворення даних (розглядалися в описовій статистиці) для того, щоб зробити залежність більш схожою на лінійну. В цьому випадку моделювання буде простішим, але ускладнюється розуміння і трактовка моделі.

Порядок дій при регресійному аналізі

Розглянемо загальний алгоритм на прикладі лінійного багатомірного регресійного аналізу.

Нехай на об'єкт дослідження впливають декілька кількісних та/або якісних факторів. Потрібно спрогнозувати кількісний відгук.

Вирішуються задачі:

- які з факторів впливають на відгук, які ні;
- ранжування факторів за ступенем впливу на відгук;
- прогнозування значень відгуку.

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_k \cdot X_k$$

Загальний алгоритм

1. Вибір факторів та відгуку.

Виконується на базі уявлень про природу досліджуваної проблеми, інтуїції дослідника, досвіду схожих досліджень.

2. Пошук мультиколінеарних факторів.

Мультиколінеарними називають фактори, між якими є досить сильний кореляційний зв'язок (0.7 або більше). Такі фактори ускладнюють ранжування інших факторів за ступенем їх впливу на відгук. Для їх виявлення необхідно побудувати кореляційну матрицю. Рекомендовано один з двох мультиколінеарних факторів видаляти, який саме – вибираємо або керуючись здоровим глуздом, або обчислюємо на наступному кроці.

3. Дослідження відносної важливості мультиколінеарних факторів.

Для кожного фактора визначається його стандартизований коефіцієнт регресії (β). Як правило, видаляється той з мультиколінеарних факторів, стандартизований коефіцієнт регресії якого менший. Далі потрібно повернутися на попередній крок (тобто побудувати оновлену кореляційну матрицю).

4. Аналіз залишків (попередній аналіз моделі).

Полягає в оцінці різниці фактичних значень відгуку та значень, обчислених за побудованим рівнянням регресії:

- залишки мають нормальний розподіл;
- залишки не залежать від прогнозованих (обчислених) за рівнянням регресії значень відгука.

Лінія регресії виражає найкраще передбачення залежної змінної (Y) з незалежних змінних (X). Однак, зазвичай є істотний розкид спостережуваних точок щодо підібраної прямої. Відхилення окремої точки від лінії регресії (від передбаченого значення) називається залишком.

Аналіз залишків можна провести по гістограмі частот, нормально-імовірністному графіку або діаграмі розсіювання.

5. Аналіз регресійного рівняння та видалення факторів, що не впливають на відгук

Фактори, які мають рівень значущості $p > 0.05$, можуть бути виключені з аналізу. Всі видалення потрібно робити послідовно, кожен раз перебудовуючи модель.

6. Оцінка прийнятності моделі в цілому

Розраховується рівень значущості по таблиці для дисперсійного аналізу. Якщо $p < 0.05$ – модель можна використовувати для опису впливу факторів на відгук.

7. Аналіз коефіцієнта детермінації

Чим менше розмах значень залишків поблизу лінії регресії по відношенню до загального розмаху значень, тим, очевидно, краще прогноз. Наприклад, якщо зв'язок між змінними X і Y відсутній, то відношення залишкової мінливості змінної Y до вихідної дисперсії дорівнює 1. Якщо X і Y жорстко пов'язані, то залишкова мінливість відсутня, і відношення дисперсій дорівнюватиме 0. У більшості випадків відношення буде лежати десь між цими екстремальними значеннями, тобто між 0 і 1.

➤ 1 мінус відношення залишкової мінливості змінної Y до вихідної дисперсії називається R -квадратом або **коефіцієнтом детермінації**.

R^2 показує долю мінливості відгуку, що може бути пояснена одночасним впливом всіх включених в модель факторів. В ідеалі бажано мати пояснення якщо не для всієї, то хоча б для більшої частини вихідної мінливості.


- ✓ Значення R^2 є індикатором ступеня підгонки моделі до даних (значення R^2 близьке до 1 показує, що модель пояснює майже всю мінливість відповідних змінних).
- ✓ Якщо $R^2 < 0.3$ – поганий вибір факторів.
- ✓ Якщо факторів у моделі більше 10, коефіцієнт детермінації необхідно корегувати.

8. Побудова прогнозу

Слід зауважити, що прогноз має максимальну точність при значеннях факторів, близьких до середніх і точність прогнозу погіршується по мірі віддалення від середнього значення. Звідси витікає, що по оцінкам регресійних залежностей неможна зробити прогноз задовільної точності при значному віддаленні x від \bar{x} .

Приклад.

1. Вибір факторів та відгуку.



	1 оценка	2 посещено лекций	3 посещено практик	4 сдано работ
1	100	18	18	10
2	99	16	18	10
3	93	8	18	9
4	95	15	18	10
5	96	13	18	10
6	98	17	18	10
7	87	14	18	9

2. Пошук мультиколінеарних факторів.

Знайдено пару мультиколінеарних факторів.

Переменная	Корреляции (Таблица данных1)			
	посещено лекций	посещено практик	сдано работ	оценка
посещено лекций	1,000000	0,773961	0,546741	0,784837
посещено практик	0,773961	1,000000	0,415116	0,876671
сдано работ	0,546741	0,415116	1,000000	0,614073
оценка	0,784837	0,876671	0,614073	1,000000

3. Дослідження відносної важливості мультиколінеарних факторів.

Визначаємо стандартизовані коефіцієнти регресії (beta) факторів та видаляємо той фактор, у якого стандартизований коефіцієнт регресії менший.

Итоги регрессии для зависимой переменной: оценка (Таблица данных1)						
R= ,92111384 R2= ,84845071 Скоррект. R2= ,84416158 F(3,106)=197,81 p<0,0000 Станд. ошибка оценки: 5,7335						
N=110	БЕТА	Ст.Ош. БЕТА	В	Ст.Ош. В	t(106)	p-знач.
Св.член			10,23651	3,874546	2,64199	0,009490
посещено лекций	0,112457	0,064887	0,49427	0,285191	1,73312	0,085983
посещено практик	0,676887	0,059719	2,41034	0,212653	11,33459	0,000000
сдано работ	0,271601	0,045164	3,40408	0,566057	6,01367	0,000000

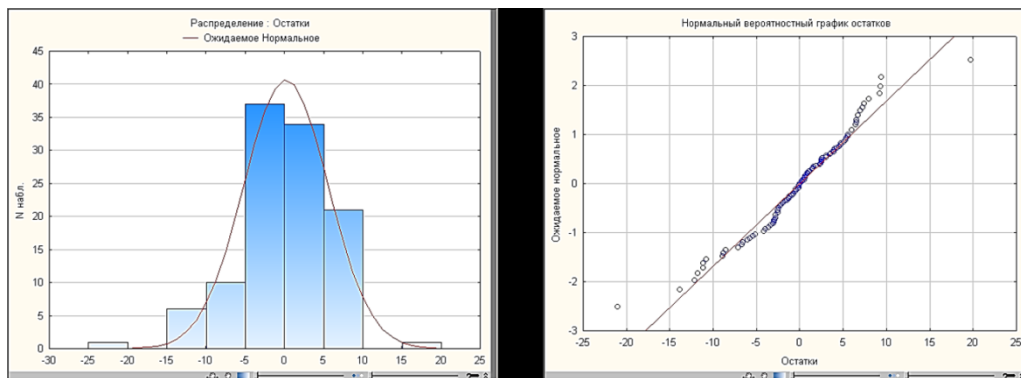
Повторно будуємо кореляційну таблицю.

Переменная	Корреляции (Таблица данных1)		
	посещено практик	сдано работ	оценка
посещено практик	1,000000	0,415116	0,876671
сдано работ	0,415116	1,000000	0,614073
оценка	0,876671	0,614073	1,000000

Мультиколінеарних факторів більше немає. Визначаємо коефіцієнти регресії.

Итоги регрессии для зависимой переменной: оценка (Таблица данных1)						
R= ,91877979 R2= ,84415630 Скоррект. R2= ,84124334 F(2,107)=289,79 p<0,0000 Станд. ошибка оценки: 5,7870						
N=110	БЕТА	Ст.Ош. БЕТА	В	Ст.Ош. В	t(107)	p-знач.
Св.член			8,367824	3,756180	2,22775	0,027990
посещено практик	0,751209	0,041949	2,674990	0,149377	17,90765	0,000000
сдано работ	0,302234	0,041949	3,788017	0,525764	7,20479	0,000000

4. Аналіз залишків (попередній аналіз моделі).



Можна вважати розподіленими нормально.

5. Аналіз регресійного рівняння та видалення факторів, що не впливають на відгук

Итоги регрессии для зависимой переменной: оценка (Таблица)						
R= ,91877979 R2= ,84415630 Скоррект. R2= ,84124334 F(2,107)=289,79 p<0,0000 Станд. ошибка оценки: 5,7870						
N=110	БЕТА	Ст.Ош. БЕТА	В	Ст.Ош. В	t(107)	p-знач.
Св.член			8,367824	3,756180	2,22775	0,027990
посещено практик	0,751209	0,041949	2,674990	0,149377	17,90765	0,000000
сдано работ	0,302234	0,041949	3,788017	0,525764	7,20479	0,000000

Обидва фактори важливі.

6. Оцінка прийнятності моделі в цілому

Дисперсионный анализ; ЗП: оценка (Таблица)					
Эффект	Сумма квадр.	сс	Средн. квадр.	F	p-знач.
Регресс.	19409,76	2	9704,881	289,7927	0,000000
Остатки	3583,33	107	33,489		
Итого	22993,09				

$p < 0.05$ – модель можна використовувати для опису впливу факторів на відгук.

7. Аналіз коефіцієнта детермінації

Итоги регрессии для зависимой переменной: оценка (Таблица)						
R= ,91877979 R2= ,84415630 Скоррект. R2= ,84124334 F(2,107)=289,79 p<0,0000 Станд. ошибка оценки: 5,7870						
N=110	БЕТА	Ст.Ош. БЕТА	В	Ст.Ош. В	t(107)	p-знач.
Св.член			8,367824	3,756180	2,22775	0,027990
посещено практик	0,751209	0,041949	2,674990	0,149377	17,90765	0,000000
сдано работ	0,302234	0,041949	3,788017	0,525764	7,20479	0,000000

Значення R^2 досить велике і показує, що модель пояснює 84% мінливості оцінки.

8. Побудова прогнозу

Задаємо значення факторів: 18 відвіданих практик і 9 зданих робіт; 5 відвіданих практик і 5 зданих робіт.

Переменная	Предск.значения для (Таблица д перемен.: оценка			Переменная	Предск.значения для (Таблица д перемен.: оценка		
	В-Веса	Значение	В-Веса * знач.		В-Веса	Значение	В-Веса * знач.
посещено практик	2,674990	18,00000	48,14982	посещено практик	2,674990	5,000000	13,37495
сдано работ	3,788017	9,00000	34,09215	сдано работ	3,788017	5,000000	18,94008
Св. член			8,36782	Св. член			8,36782
Предсказанные			90,60980	Предсказанные			40,68286
-95,0%ИС			88,70998	-95,0%ИС			37,64141
+95,0%ИС			92,50961	+95,0%ИС			43,72430

Визначення параметрів рівняння регресії. Одномірна лінійна регресія

Розглянемо двовимірну випадкову величину (X, Y) , де X і Y - залежні випадкові величини. Представимо одну з величин як функцію іншої. Обмежимося наближеним уявленням (точне наближення, взагалі кажучи, неможливо) величини Y у вигляді лінійної функції величини X :

$$Y \cong g(X) = \alpha + \beta \cdot X$$

де α і β - параметри, що підлягають визначенню. Це можна зробити різними способами, найбільш часто використовують метод найменших квадратів.

Функцію $g(X) = \alpha + \beta \cdot X$ називають *найкращим наближенням* Y в сенсі методу найменших квадратів, якщо математичне сподівання $M(Y - g(X))^2$ приймає найменше можливе значення.

➤ Функцію $g(x)$ називають *середньоквадратичною регресією* Y на X .

З курсу теорії імовірностей нагадаємо теорему про вигляд

Теорема. Лінійна середньоквадратична регресія Y на X має вигляд

$$g(X) = m_y + r \frac{\sigma_y}{\sigma_x} (X - m_x)$$

де

$$m_y = M(Y), \quad m_x = M(X),$$

$$\sigma_y = \sqrt{D(Y)}, \quad \sigma_x = \sqrt{D(X)}$$

$$r = \frac{K_{XY}}{\sigma_x \sigma_y} \text{ - коефіцієнт кореляції величин } Y \text{ та } X.$$

Рівняння можна записати в іншому вигляді

$$g(X) = \underbrace{\left(m_y - r \frac{\sigma_y}{\sigma_x} \cdot m_x \right)}_{\alpha} + \underbrace{r \frac{\sigma_y}{\sigma_x}}_{\beta} \cdot X$$

Коефіцієнт $\beta = r \frac{\sigma_y}{\sigma_x}$ називають **коефіцієнтом регресії Y на X** .

Визначення параметрів вибіркового рівняння прямої лінії середньоквадратичної регресії по незгрупованим даним

Вивчається двовимірна випадкова величина (X, Y) , де X і Y - залежні випадкові величини. В результаті n незалежних випробувань отримано n пар чисел $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Необхідно за даними спостережень знайти рівняння прямої лінії середньоквадратичної регресії у вигляді

$$Y = b + \rho_{YX} \cdot X$$

Коефіцієнт ρ_{YX} називають **вибірковим коефіцієнтом регресії Y на X** .

Задача зводиться до вибору параметрів b та ρ_{YX} так, щоб точки (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) , побудовані за даними спостережень на площині xOy лежали якомога ближче до прямої

$$Y = b + \rho_{YX} \cdot X$$

Назвемо відхиленням різницю $Y_i - y_i$, $i=1, n$; де Y_i – ордината, що відповідає x_i , обчислена за рівнянням, y_i – спостережена ордината, що відповідає x_i .

Підберемо параметри b та ρ_{YX} так, щоб сума квадратів відхилень була мінімальною, тобто мінімізуємо функцію

$$F(b, \rho_{YX}) = \sum_{i=1}^n (Y_i - y_i)^2 = \sum_{i=1}^n (\rho_{YX} \cdot x_i + b - y_i)^2$$

Для пошуку мінімуму прирівнюємо частинні похідні до нуля

$$\begin{cases} \frac{\partial F}{\partial \rho} = 2 \sum_{i=1}^n (\rho_{YX} \cdot x_i + b - y_i) \cdot x_i = 0 \\ \frac{\partial F}{\partial b} = 2 \sum_{i=1}^n (\rho_{YX} \cdot x_i + b - y_i) = 0 \end{cases}$$

отримали систему рівнянь

$$\begin{cases} \rho_{YX} \sum x^2 + b \sum x = \sum xy \\ \rho_{YX} \sum x + bn = \sum y \end{cases}$$

Розв'язавши систему рівнянь, отримаємо

$$\begin{cases} \rho_{YX} = \frac{n \sum xy - \sum x \cdot \sum y}{n \sum x^2 - (\sum x)^2} \\ b = \frac{\sum x^2 \cdot \sum y - \sum x \cdot \sum xy}{n \sum x^2 - (\sum x)^2} \end{cases}$$

Визначення параметрів вибіркового рівняння прямої лінії середньоквадратичної регресії при нормальному розподілі похибки

Вивчається двовимірна випадкова величина (X, Y) , де X і Y - залежні випадкові величини. В результаті n незалежних випробувань отримано n пар чисел $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Представимо рівняння прямої лінії середньоквадратичної регресії у вигляді

$$Y = b + \rho_{YX} \cdot X + \delta$$

δ - випадкова величина (наприклад, похибка вимірювання або завада), розподілена за нормальним законом з $M(\delta) = 0, D(\delta) = \sigma_Y^2$.

Задача зводиться до вибору параметрів b та ρ_{YX} так, щоб точки $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, побудовані за даними спостережень на площині xOy лежали якомога ближче до прямої

$$Y = b + \rho_{YX} \cdot X$$

В цьому випадку відхилення буде дорівнювати випадковій величині $\delta_i = Y_i - y_i$, закон розподілу якої відомий.

Для визначення параметрів b та ρ_{YX} можна скористатися методом найбільшої подібності, тобто мінімізуємо функцію

$$\begin{aligned} L(\delta_1, \dots, \delta_n, b, \rho_{YX}) &= \prod_{i=1}^n \frac{1}{\sigma_Y \sqrt{2\pi}} e^{-\delta_i^2 / 2\sigma_Y^2} = \\ &= C \cdot e^{-\frac{1}{2\sigma_Y^2} \sum \delta_i^2} = C \cdot e^{d \sum \delta_i^2} = C \cdot e^{d \sum (y_i - b - \rho_{YX} x_i)^2} \end{aligned}$$

C та d - деякі константи, причому d від'ємна.

Позначимо

$$Q = \sum_{i=1}^n (y_i - Y_i)^2 = \sum_{i=1}^n (y_i - b - \rho_{YX} x_i)^2$$

Тоді

$$\{b, \rho_{YX}\} = \arg \left\{ \max_{b, \rho_{YX}} L \right\} = \arg \left\{ \min_{b, \rho_{YX}} Q \right\} =$$

$$= \arg \left\{ \frac{\partial Q}{\partial b} = 0, \frac{\partial Q}{\partial \rho_{YX}} = 0 \right\}$$

Розв'язавши систему рівнянь, отримаємо

$$\begin{cases} \rho_{YX} = \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\ b = \bar{y} - \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \cdot \bar{x} \end{cases}$$

При $\bar{x} = 0$

$$\begin{cases} \rho_{YX} = \frac{\sum y_i x_i}{\sum x_i^2} \\ b = \bar{y} \end{cases}$$

Визначення параметрів вибіркового рівняння прямої лінії середньоквадратичної регресії по згрупованим даним ($n > 50$)

При великій кількості спостережень одне й теж саме значення x може зустрітися n_x раз, одне й теж саме значення y може зустрітися n_y раз, одна й таж пара (x, y) може зустрітися n_{xy} раз. Тому дані спостережень групують у вигляді кореляційної таблиці

Y	X				n_y
	10	20	30	40	
0,4	5	—	7	14	26
0,6	—	2	6	4	12
0,8	3	19	—	—	22
n_x	8	21	13	18	$n = 60$

Перепишемо систему рівнянь для визначення параметрів рівняння регресії так, щоб вони відображали дані кореляційної таблиці.

$$\begin{cases} \rho_{YX} \bar{x}^2 + b n \bar{x} = \sum n_{xy} xy \\ \rho_{YX} \bar{x} + b = \bar{y} \end{cases}$$

Вибірковий коефіцієнт кореляції

$$r_B = \rho_{YX} \cdot \frac{\sigma_X}{\sigma_Y} = \frac{\sum n_{xy} xy - n \bar{x} \bar{y}}{n \sigma_X \sigma_Y}, \quad \rho_{YX} = r_B \frac{\sigma_Y}{\sigma_X}$$

де σ_X , σ_Y – вибіркові середньоквадратичні відхилення.

В результаті отримаємо рівняння регресії у вигляді

$$Y - \bar{y} = r_B \frac{\sigma_Y}{\sigma_X} (X - \bar{x})$$

Інтервали довіри для оцінок параметрів рівняння регресії b та ρ_{YX} та прогнозованого значення g , можна побудувати на основі z –статистик при відомій дисперсії, або t –статистик при невідомій дисперсії.

Ширина інтервалу довіри для оцінки b буде збільшуватися при збільшенні рівня довіри та по мірі віддалення від \bar{x} .

Підсумки

- Регресія - один із варіантів моделювання структури та параметрів об'єктів дослідження на основі інформації, що міститься у вибірках.
- Вид регресійного аналізу, який можна застосувати в рамках конкретної задачі, визначається шкалою, в якій вимірювана залежна змінна (відгук), і виглядом зв'язку між змінними.
- Якість регресійної моделі сильно залежить від обсягу вхідних даних, їх якості та правильності вибору типу моделі (шкала вимірювання, форма зв'язку).

- Алгоритм побудови регресійної моделі має певні відмінності для різних типів моделей.
- Корисна регресійна модель може бути побудована з використанням відносно невеликої кількості незалежних змінних.

Література

1. Барковський В.В., Барковська Н.В., Лопатін О.К. Теорія імовірностей та математична статистика. 5-те видання. / Київ: Центр учбової літератури, 2010. – 424 с
2. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и её инженерные приложения / 1988р.
3. Гмурман В.Н. Теория вероятностей и математическая статистика / М.: Высшая школа, 2002.
4. Жлуктенко В.І., Наконечний С.І. Теорія ймовірностей і математична статистика / 1997р.
5. Свешников С.В. Сборник задач по теории вероятностей, математической статистики и случайным функциям.
6. Хом'юк І.В., Хом'юк В.В., Краєвський В.О. Теорія імовірностей та математична статистика. Навчальний посібник. / Вінниця: ВНТУ, 2009. – 189 с.