Елементи математичної статистики

Часто потрібно вивчати явище, точний перебіг яких не можливо передбачити і які виступають не поодиноко, а масово. Такі явища називаються **масовими випадковими**.
Означення. Масові випадкові явища залежні від часу називаються **випадковими**, або **стохастичними процесами.**
Різні прояви стохастичного процесу називаються **мінливими величинами** або **варіантами**.
Оскільки варіанту пізнаємо в результаті спостереження то її називаємо ще **статистичною змінною**.
Три терміни: мінлива величина, варіанта і статистична змінна – значення еквівалентні.

1. Класифікація варіант

Мінливі величини діляться на:
1) якісні та кількісні
2) дискретні та неперервні
3) одновимірні, двовимірні і т.д.
Такий поділ вказує на три підходи до однієї і тієї ж варіанти.

Наприклад мінлива величина може бути кількісною, неперервною та тривимірною, або якісною, дискретною і одновимірною і т.д. якісні варіанти будемо позначати $A, B, C, \ldots\ldots$, і кількісні $x, y, z, \ldots\ldots$

Приклади якісних варіант(або мінливих величин): яскравість зірок, ступінь захмареності в даній місцевості, колір очей, стать.

Приклади кількісних варіант: число зерен у головці маку, число бракованих виробів за зміну на якомусь виробництві масової продукції.

Приклади дискретних варіант: число пелюсток на квітці бузку, число осіб у сім'ї, число букв у слові, число телефонних викликів за одиницю часу.

Приклади неперервних варіант: ріст допризивників, тривалість телефонної розмови.

Приклади n – вимірних варіант: вік, ріст і вага людини, тривимірна мінлива величина; число і вага зерен одного складного колоса озимої пшениці – двовимірна величина.

Вибір мінливої величини при дослідженні випадкового явища є справою зручності. Те саме випадкове явище можна досліджувати за допомогою якісної або кількісної мінливої величини.

Приклад. 1.Землетрус може бути

перехід якісної в кількісну варіанту

1) непомітний (1бал)
2) дуже слабкий (2бали)
3) слабкий (3 бали)
4) помірний (4 бали)
5) досить сильний (5 балів)
6) сильний (6 балів)
7) дуже сильний (7 балів)
8) руйнівний (8 балів)
9) спустошувальний (9 балів)
10) знищувальний (10 балів)
11) катастрофа (11 балів)
12) сильна катастрофа (12 балів)

Нумерація вказує на кількісну варіанту (на бали).

Приклад.

2. У стародавньому Єгипті для вимірювання рівня виловів риби у річці Ніл побудували колодязі. В різні роки зареєстровано max виловів в ліктях (міра 3 лікті = 1 метру). За багато століть одержано такі значення max виловів.

12 ліктів – голод
13 ліктів – достаток
14 ліктів – радість
15 ліктів – спокій
16 ліктів – багатство
17 ліктів – неспокій
18 ліктів – тривога
19 ліктів – руйнування , голод
20 ліктів – катастрофа, епідемія

2. Статистичний матеріал

Випадкові явища, стохастичні процеси, мінливі величини пізнаємо спостереженнями, тобто у результаті відповідно поставлених експериментів. **Означення.** Кількість спостережень називається **обсягом** (розміром, об'ємом, довжиною, тривалістю) спостережень.

Означення. Сукупність спостережень називається **статистичним матеріалом**.

Означення. Кожне окреме спостереження називається **елементом** статистичного матеріалу.

Якщо обсяг статистичного матеріалу в межах від 2 до кільканадцяти, то статистичний матеріал називається малим статистичним матеріалом; від кільканадцяти до кількадесяти – то статистичний матеріал середній; в межах від кількадесяти до кількасот – великий.

Якщо число спостережень рівне багатьом сотням, тисячам і т.д. то статистичний матеріал дуже великий, колосальний, гігантський і т.д. Такий поділ не є строгий.

В залежності від обсягу статистичного матеріалу існують різні математичні методи його обробки, а також використовуються обчислювальна техніка різної потужності від рахівниць і арифмометрів до найсучасніших ЕОМ.

Статистичний матеріал може бути дуже непроглядний. Про такий первісний статистичний матеріал кажуть, що він сирий. Відносно сирого статистичного матеріалу виникає таке 1-е питання: Як компактно представити статистичний матеріал. Компактно значить зручно для аналізу і прогнозу.

Статистичний матеріал можна представити словесно, таблично, графічно та аналітично.

3. Табличне та графічне представлення статистичного матеріалу

В дальнішому обмежимося дослідженням одновимірної кількісної мінливої величини.

Нехай (1) $x_1,.....,x_n$ – результат n спостережень (коротко x_j – результат j – го спостереження) над одновимірною кількісною мінливою величиною. Тоді послідовність (1) представляє собою статистичний матеріал обсягом n спостережень.

1) Дискретний випадок

Нехай серед спостережень (1) зустрічаються такі можливі значення одновимірної дискретної варіанти x, впорядковані за величиною:

$$x_{(1)} < x_{(2)} < ..... < x_{(k)}$$

і нехай ці значення зустрічаються відповідно часто:

$$n_1, n_2, ..., n_k, \quad (n_1 + n_2 + ... + n_k = n)$$

Число n_i називається частотою значення $x_{(i)} (i = 1,2,...k)$. Тоді статистичний матеріал (1) зручно записати в формі таблички з двома рядками у першому рядку виписуємо в зростаючому порядку можливі значення варіанти, а в другому – відповідні їм частоти. Дістанемо частотну таблицю

$x_{(1)}$ $x_{(2)}$ ..... $x_{(k)}$	$\sum$

n_1 $n_2 \ldots\ldots\ldots\ldots n_k$	n

(2)

Частотна таблиця (2) називається ще **статистичним розподілом дискретної варіанти** x.

Для графічного представлення частотної таблиці на вісь абцис наносимо можливі значення дискретної мінливої величини та відкладемо в цих точках відповідні частоти $n_i (i = 1, 2, ...)$. Отримаємо **діаграму частот**.

Якщо з'єднати відрізками сусідні пункти $(x_{(i)}, n_i)$, то дістанемо **полігон** частот.

2) **Неперервний випадок.**

а) **не згруповані дані**.

Якщо статистичний матеріал малий або середній, то спостереження (1) над одновимірною неперервною варіантою впорядкуємо за величиною: від найменшого до найбільшого. Нехай $x_{(1)}$ буде найменше зі спостережень (1) і т.д., в кінці $x_{(n)}$ буде найбільше зі спостережень (1)

$$x_{(1)} = \min \quad (x_1, \ldots\ldots\ldots, x_n)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_{(n)} = \max \quad (x_1, \ldots\ldots\ldots, x_n)$$

В силу обмеженої точності деякі спостереження можуть бути одинакові. так упорядковані спостереження (1) записуємо у формі ряду:

$$x_{(1)} \leq x_{(2)} \leq \ldots\ldots\ldots\ldots \leq x_{(n)} \qquad (3)$$

Ряд (3) називається **варіаційним рядом** для спостережень (1) над одновимірною неперервною мінливою величиною.

Приклад 1. Періодична система Менделєєва, з огляду на атомну вагу елементів, утворює варіаційний ряд.

Для графічного представлення варіаційного ряду наносимо на вісь абцис елементи варіаційного ряду $x_{(i)} (i = 1, 2, ..., n)$ та пов'яжемо з кожною точкою $x_{(i)}$ масу $\frac{1}{n}$.

Нарисуємо східчасту лінію зі стрибками вверх у пунктах $x_{(i)}$ на $\frac{1}{n}$. Від $-\infty$ до $x_{(1)}$ маємо лінію на рівні нуль. У точці $x_{(1)}$ маємо стрибок на $\frac{1}{n}$ і відрізок на висоті $\frac{1}{n}$ до точки $x_{(2)}$ у точці $x_{(n)}$ останній стрибок на $\frac{1}{n}$ і лінія на висоті 1буде продовжуватися до безмежності. Якщо б зустрілося два одинакові $x_{(i)}$, або більше, то в цій точці був би стрибок на $\frac{2}{n}$, або відповідно більше.

Одержане графічне представлення варіаційного ряду називаються **емпіричною функцією розподілу** або **емпіричною кумулятою.**

Таким чином, емпірична функція розподілу

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x_{(k)} \leq x < x_{(k+1)} \quad (k = 1, ..., .n - 1) \\ 1, & x_{(n)} \leq x \end{cases}$$

Для кожного x емпірична функція розподілу $F_n(x)$ є випадковою змінною з розподілом

$$P\left\{F_n(x) = \frac{k}{n}\right\} = C_n^k \left[F_n(x)\right]^k \cdot \left[1 - F_n(x)\right]^{n-k}$$

б) **Згруповані дані**. Якщо статистичний матеріал середній або великий, то знайдемо найменше та найбільше зі спостереження

$$x_{(1)} = \min \quad (x_1, ..., x_n), \quad x_{(n)} = \max \quad (x_1, ..., x_n)$$

Означення. Різниці між найбільшим і найменшим елементами статистичного матеріалу називається **розмахом статистичного матеріалу**

$$\rho = x_{(n)} - x_{(1)} \qquad 2^r < n \leq 2^{r+1}$$

Інтервал розмаху ділимо досить довільним способом на $(r+1)$ однакові або неоднакові інтервали, де r – натуральне , $r = 1,\ 2,.....$

Центри одержаних інтервалів позначимо в зростаючому порядку через $z_1,....., z_i,.....z_{r+1.}$

Нехай на інтервалі з центром в точці z_i попадає n_i спостережень.

Очевидно, що $n_1 + \ n_2 + ... \ n_{r+1} \ = n$

Тоді статистичний матеріал представимо у вигляді таблиці з двох рядків:

1-й в зростаючому порядку- центри інтервалів

2-й - відповідні частоти

$z_1 \quad z_2 ....z_i ....z_{r+1}$	$\sum$
$n_1 n_2 .... .n_i ....n_{r+1}$	n

У цій таблиці замість кожного з n_i індивідуальних значень статистичного матеріалу (1), що попадають у інтервал з центром в т. z_i, розглядається n_i – кратно повторений центр i – го інтервалу z_i.

Одержана таблиця – частотна. При такому представлені дальша математична обробка статистичного матеріалу значно спрощується.

Для графічного представлення одержаної частотної таблиці наносимо на абцису центри інтервалів. В точці z_i ставимо ординату n_i. Одержимо **графік частот**

Якщо з'єднати верхушки сусідніх вершин графіка частот відрізками, то одержимо многокутник частот або **<u>полігон частот</u>** .

Якщо над інтервалом з центром в т. z_i поставити прямокутник висотою n_i, то одержимо **<u>гістограму частот.</u>**

Ми розглянули лише такі графічні представлення, які нагадують функцію розподілу або густину.

СТАТИСТИКИ

Нехай $x_1, \ldots, x_n$ (1) ряд незалежних спостережень проведених в однакових умовах над одновимірною кількісною мінливою величиною.

Табличне та графічне представлення статистичного матеріалу все ж містить немало інформації (елементів).

Тому друге питання, що виникає відносно статистичного матеріалу таке: як охарактеризувати статистичний матеріал одним або кількома числами.

Вже давно зауважено, що статистичний матеріал взагалі групується в одному або кількох місцях, в околі одного або кількох значень. Причому в околі цих значень він більш або менш розсіяний, а також форма розсіяння може бути досить різна. Тому числові характеристики поділяються на три групи:

1. Числові характеристики **<u>центральної тенденції</u> (локації)**. До них відноситься:
   - а) медіана (M_e)
   - б) мода (M_o)
   - в) середнє арифметичне ($\bar{x}$)
2. Числові характеристики **розсіяння**. До них відноситься:
   - а) варіанса (s^2)
   - б) стандарт (s)
   - в) розмах (ρ)
   - г) варація (v)
   - д) інтерквантильність широт
3. Числові характеристики **форми**: До них відноситься:
   - а) асиметрія (γ_1) *(Ас)*
   - б) ексцес (γ_2) *(Ек)*

Кожна з перерахованих числових характеристик є деякою функцією від елементів статистичного матеріалу.

<u>Означення.</u> Функція від елементів статистичного матеріалу називається **<u>статистикою.</u>**

Таким чином ми розглянемо три групи статистик.

1. статистики центральної тенденції

2. статистики розсіяння

3. статистики форми

Статистики центральної тенденції

Медіана.

Означення. Медіаною називають цей елемент статистичного матеріалу, який ділить відповідний варіаційний ряд (3) на дві рівні за обсягом частини. Медіану позначаємо M_e.

Якщо обсяг статистичного матеріалу непарний, то медіана визначається однозначно. Наприклад, якщо варіаційний ряд статистичного матеріалу буде ($n = 2k+1$)

$\underbrace{x_{(1)} \le x_{(2)} \le \cdots \le x_{(k)}}_{k} \le x_{(k+1)} \le \underbrace{\cdots \le x_{(2k+1)}}_{k}$, то $M_e = x_{(k+1)}$

Якщо обсяг статистичного матеріалу парний, то медіаною може бути інтервал. Наприклад, якщо варіаційний ряд статистичного матеріалу буде ($n = 2k$)

$\underbrace{x_{(1)} \le x_{(2)} \le \cdots \le}_{k-1} \; x_{(k)} \le x_{(k+1)} \le \underbrace{\cdots \le x_{(2k)}}_{k-2}$

To $M_e = \left[x_{(k)}, x_{(k+1)}\right]$ $\quad M_e = \dfrac{x_{(k)} + x_{(k+1)}}{2}$

Твердження. Лише медіана може мінімізувати суму абсолютних відхилень елементів статистичного матеріалу від сталої.

Доведення. Справді, позначимо через

$$f(a) = \sum_{i=1}^{n} |x_i - a| = \sum_{x_i > a} (x_i - a) + \sum_{x_i < a} (a - x_i)$$

Ця функція має похідну рівну нулю (що є необхідною умовою екстремуму)

$$f'(a) = \sum_{x_i > a} (-1) + \sum_{x_i < a} 1 = 0$$

тільки тоді, коли число елементів статистичного матеріалу більших від a рівне числу елементів статистичного матеріалу менших від a, тобто, коли $a = M_e$.

Мода.

Означення. Модою називають цей елемент статистичного матеріалу, який найчастіше зустрічається. Моду позначаємо M_o.

Не виключено, що декілька значень статистичного матеріалу зустрічаються найчастіше та однаково часто, тоді всі вони модні. Мода типове значення статистичного матеріалу. Мода широко використовується в демографії. У демографії, при багатовершинних розподілах, краще вказати моди, ніж середнє арифметичне.

5	4	3	2	1	Σ	$M_o = 4$ (найчастіше зустрічається)
7	12	5	0	1		

Середнє арифметичне.

Означення. Середнім арифметичним називається сума всіх елементів статистичного матеріалу, поділена на обсяг статистичного матеріалу, позначається $\bar{x}$:

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

Відмітимо, що середнє арифметичне може не зустрічатися серед елементів статистичного матеріалу.

Властивості:

1. Середнє арифметичне не менше від найменшого елемента і не більше від найбільшого елемента статистичного матеріалу $x_{(1)} \le \bar{x} \le x_{(n)}$

Доведення. З означення:

$$\bar{x} = = \frac{1}{n}\sum_{i=1}^{n} x_i \ge \frac{1}{n}\sum_{i=1}^{n} x_{(1)} = x_{(1)}$$

$$\bar{x} = = \frac{1}{n}\sum_{i=1}^{n} x_i \le \frac{1}{n}\sum_{i=1}^{n} x_{(n)} = x_{(n)}$$

2. Сума відхилень елементів статистичного матеріалу від середнього арифметичного рівна нулю.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Доведення. $\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\bar{x}=n\cdot\frac{1}{n}\sum_{i=1}^{n}x_i-n\bar{x}=n\bar{x}-n\bar{x}=0$.

3. Середнє арифметичне мінімізує суму квадратів відхилень елементів статистичного матеріалу від сталої.

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

Доведення. Позначимо через $f(a)$ функцію, яка дорівнює

$$f(a)=\sum_{i=1}^{n}(x_i-a)^2$$

$$f'(a)=-2\sum_{i=1}^{n}(x_i-a)=-2(n\bar{x}-na)=2n(a-\bar{x})=0$$

$a=\bar{x}$ - точка підозріла на екстремум

$f''(a)=2n>0$ - точка мінімуму.

Означення. Сума квадратів відхилень елементів статистичного матеріалу від її середнього називається **девіацією**.

Девіація виражається у квадратних одиницях.

Статистики розсіяння

Варіанса.

Означення. **Варіансою** називається девіація (сума квадратів відхилень елементів статистичного матеріалу від середнього арифметичного) поділена на обсяг статистичного матеріалу без одного і позначається s^2.

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

Число $n-1$ називається **числом ступенів вільності** статистичного матеріалу з очікуваним середнім і позначають $d.f.$ ($d.f.=n-1$).

Таким чином **варіанса** – це девіація поділена число ступенів вільності даної вибірки.

Стандарт.

Означення. Стандартом (флуктуацією, середнім квадратичним відхиленням) називається арифметичний корінь з варіанси і позначається

$$s=+\sqrt{s^2}$$

Стандарт має ту саму розмірність як і статистична змінна.

Розмах.

Означення. Розмахом називається різниця між найбільшим і найменшим елементами статистичного матеріалу і позначається ρ.

$$\rho=\max(x_1,\cdots,x_n)=\min(x_1,\cdots,x_n)=x_{(n)}-x_{(1)}$$

Тобто розмах – це різниця між крайніми елементами варіаційного ряду.

Для додатніх мінливих величин за міру розсіяння часто можна прийняти варіацію.

Варіацією вибірки називається відношення стандарту цієї вибірки до середнього арифметичного

$$v=\frac{s}{\bar{x}}.$$

Інтерквантильні широти.

Означення. Квантилем порядку α, якщо він існує, називається цей елемент статистичного матеріалу (відповідного варіаційного ряду), до якого включно маємо α% елементів статистичного матеріалу (відповідного варіаційного ряду).

Статистичний матеріал (1) має квантилі тільки порядків кратних $\cdot\frac{100}{n}$, інші квантилі не існують; елемент $x_{(i)}$ є кантилем порядку $i\cdot\frac{100}{n}$ $(i=1,\cdots,n)$.

Означення. При $\alpha<\beta$, різницю між квантилем порядку β і квантилем порядку α називають **інтерквантильною широтою порядку** $\beta-\alpha$.

Для статистичного матеріалу (1) існують тільки інтерквантильні широти наступних порядків

$$q_{ij}=(j-i)\cdot\frac{100}{n},\ j>i\ \ (i=1,2,\cdots n-1;\ j=2,3,\cdots n)$$

Практично важливими є **симетричні інтерквантильні широти**.

Для статистичного матеріалу (1) існують симетричні інтерквантильні широти тільки таких порядків

$q_{i,n+1-i} = (n+1-2i)\cdot\frac{100}{n}$; $\left(i = 1,2,\cdots,\left[\frac{n+1}{2}\right]-1\right)$, вони рівні $x_{(n+1-i)} - x_{(i)}$.

Квантилі порядку 25, 50, 75 називаються **квартилями:** першим Q_1; другим Q_2; третім Q_3. Різниця між третім і першим квартилем $Q_3 - Q_1$ називається **інтерквартильною широтою** (інтерквартильний розмах).

Очевидно, що

$$Q_1 = x_{\left(\frac{n}{4}\right)},\ Q_3 = x_{\left(\frac{3n}{4}\right)}.$$

Від Q_1 виключно до Q_3 включно розташовано 50% центральних елементів статистичного матеріалу.

Квантилі порядку 12,5; 25,0;…, 87,5 називаються **октилями**: першим O_1; другим O_2; …сьомим O_7. Різниця між сьомим і першим октилем $O_7 - O_1$ називається **інтероктильною широтою**. Очевидно, що

$$O_1 = x_{\left(\frac{n}{8}\right)},\ \ldots,\ O_7 = x_{\left(\frac{7n}{8}\right)}$$

Від O_1 виключно до O_7 включно розташовано 75% центральних елементів статистичного матеріалу.

Квантилі порядку 10; 20;…, 90 називаються **децилями**: першим $Д_1$; другим $Д_2$; …дев'ятим $Д_9$. Різниця між дев'ятим і першим децилем $Д_9 - Д_1$ називається **інтердецильною широтою**. Очевидно, що

$$Д_1 = x_{\left(\frac{n}{10}\right)},\ \ldots, Д_9 = x_{\left(\frac{9n}{10}\right)}$$

Від $Д_1$ виключно до $Д_9$ включно розташовано 80% центральних елементів статистичного матеріалу.

Квантилі порядку 01; 02;…, 99 називаються **центилями**: першим C_{01}; другим C_{02}; …дев'яносто дев'ятим C_{99}. Різниця між дев'яносто дев'ятим і першим центилем $C_{99} - C_{01}$ називається **інтерцентильною широтою**.

Очевидно, що

$$C_{01} = x_{\left(\frac{n}{100}\right)},\ \ldots,\ C_{99} = x_{\left(\frac{99n}{100}\right)}$$

Від C_{01} виключно до C_{99} включно розташовано 98% центральних елементів статистичного матеріалу.

Квантилі порядку 00,1; 00,2;…, 99,9 називаються **мілілями**: першим M_{001}; … дев'ятсот дев'яносто дев'ятим M_{999}. Різниця $M_{999} - M_{001}$ називається **інтермілільною широтою**.

Очевидно, що

$$M_{001} = x_{\left(\frac{n}{1000}\right)},\ \ldots,\ M_{999} = x_{\left(\frac{999n}{1000}\right)}$$

Від M_{001} виключно до M_{999} включно розташовано 99,8 % центральних елементів статистичного матеріалу.

Перший центиль – однопроцентний квантиль, перший міліль – однопроцентний квантиль.

Квартиль, октиль, дециль, центиль, міліль використовується тоді, коли об'єм статистичного матеріалу буде кратний відповідно 4, 8, 10, 100, 1000.

Моменти статистичного матеріалу

Означення. Моментом порядку k відносно сталої a називається вираз

$$\mu_k(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^k \quad (k = 1,2,\ldots)$$

При $a = 0$ момент називається **початковим** і позначається через

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad (k = 0,1,2,\ldots)$$

Покладемо, за означенням, $m_0 = 1$.

Очевидно, що 1-ий початковий момент збігається із середнім арифметичним і позначається

$$m_1 = \bar{x}$$

При $a = \bar{x}$ момент називається **центральним** і позначається через

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \quad (k = 1,2,\ldots)$$

Очевидно, що 1-ий центральний момент дорівнює

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

2-й центральний момент запишеться у вигляді

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}s^2.$$

Очевидно, що при великих n другий центральний момент практично дорівнює варіансі.

Зауважимо, що практично найчастіше використовуються такі моменти:

1-й початковий $m_1 = \bar{x}$

2-й центральний $\mu_2 = s^2$

3-й центральний і 4-й центральний μ_3; μ_4.

Очевидно, що центральний момент порядку k можна виразити через початкові моменти до порядку k.

Наприклад. 2-ий центральний момент дорівнює

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\bar{x} + \bar{x}^2 = m_2 - m_1^2.$$

Статистики форми.

Для характеристики форми мінливості статистичного матеріалу (1), Фішер увів дві статистики: **1) асиметрію і 2) ексцес.**

Означення. **Асиметрією** або скошеністю статистичного матеріалу (1) називається відношення 3-го центрального моменту до 2-го центрального моменту в степені півтора

$$A_s = \gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

Якщо $A_s > 0$ $(\gamma_1 > 0)$, то статистичний матеріал скошений вправо і має додатну асиметрію.

Якщо $A_s < 0$ $(\gamma_1 < 0)$, то статистичний матеріал скошений вліво і має від'ємну асиметрію.

Якщо $A_s = 0$ $(\gamma_1 = 0)$, то статистичний матеріал симетричний.

Асиметрія – безрозмірна величина (видно з формули).

Означення. **Ексцесом** (крутістю, сплющеністю) статистичного матеріалу (1) називається відношення 4-го центрального моменту до 2-го центрального моменту в квадраті мінус три

$$E_k = \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3.$$

При $E_k > 0$ $(\gamma_2 > 0)$, то статистичний матеріал – високовершинний.

При $E_k < 0$ $(\gamma_2 < 0)$, то статистичний матеріал – низьковершинний.

При $E_k = 0$ $(\gamma_2 = 0)$, то статистичний матеріал – нормальновершинний.

Неважко перевірити, що для нормально розподіленої випадкової змінної відношення 4-го центрального моменту до 2-го в квадраті дорівнює 3.

Моменти випадкової змінної або: момент порядку k, $(k = 1,\cdots,n)$ випадкової змінної ξ відносно сталої a називається сподівання k-го степеня відхилення цієї змінної від константи a і позначають $M_a^{(k)}(\xi)$

$$M_a^{(k)}(\xi) = M(\xi - a)^k.$$

Отже момент існує, якщо існує сподівання змінної $(\xi - a)^k$.

При $a = 0$ момент називається **початковим** і позначається

$$m_k(\xi) = M(\xi)^k.$$

Очевидно, що 1-ий початковий момент є математичним сподіванням

$$m_1(\xi) = M(\xi).$$

При $a = M(\xi)$ момент називається центральним і позначається

$$\mu_k(\xi) = M(\xi - M(\xi))^k .$$

Перший центральний момент будь-якої випадкової змінної дорівнює 0:

$$\mu_1(\xi) = M(\xi - M(\xi)) = M(\xi) - M(M(\xi)) = 0 .$$

Другий центральний момент будь-якої випадкової змінної є дисперсією цієї змінної, тобто

$$\mu_2(\xi) = M(\xi - M(\xi))^2 = D(\xi).$$

Майже вірогідна подія

Означення. Масова випадкова подія A не еквівалентна вірогідній ($A \neq u$) і така що її ймовірність дорівнює одиниці ($P(A)=1$), називається **майже вірогідною**. Подія протилежна до майже вірогідної називається **майже неможливою**. Очевидно, що майже неможлива подія не еквівалентна вірогідній. ($B \neq v \rightarrow P(B)=0$).

Приклад. На кулю нанесена сітка географічних координат. Кулю кидаємо навмання на горизонтальну площину. Дотик точкою поза екватором майже вірогідна подія, дотик полюс – майже неможлива подія. Це очевидно на підставі геометричної ймовірності.

Посилений закон великих чисел

В 1713 році була опублікована теорема Я. Бернуллі – закон великих чисел.

Теорема. Нехай μ – кількість появ події A з імовірністю p в серії з n незалежних спроб, а $\varepsilon > 0$ тоді: $\lim\limits_{n\to\infty} P\left\{\left|\frac{\mu}{n} - p\right| < \varepsilon\right\} = 1.$

У 1909 р. французький математик Еміль Борель (1871-1956) провів узагальнення теореми Я. Бернуллі.

Теорема. Нехай μ – кількість появ події A з імовірністю p в серії з n незалежних спроб. Тоді майже відносна частота збігається до ймовірності появи події в одній спробі, тобто

$$P\left\{\frac{\mu}{n} \underset{n\to\infty}{\to} p\right\} \to 1.$$

Теорема Е.Бореля посилює теорему Я.Бернуллі і називається **посиленим законом великих чисел.**

Практичний висновок з теореми Е.Бореля такий самий як з теореми Я.Бернуллі: при великих n *apriori* невідома ймовірність появи події в одній спробі наближено дорівнює відносній частоті: $p \approx \frac{m}{n}$, де m – кількість сприятливих спроб.

Посилений закон великих чисел для функції розподілу

Статистичну змінну пізнаємо за допомогою спостережень над нею, пізнаємо в результаті спроб. Нехай спроби будуть взаємно незалежні та проведені в незмінних умовах, а результати спостережень над одномірною статистичною змінною нехай будуть такі:

(1) $x_1, x_2, \ldots, x_n$.

Утворимо варіаційний ряд $x_{(1)} \le x_{(2)} \le \ldots x_{(n)}$ для статистичного матеріалу (1) і розглянемо функцію $F_n(x) = \frac{k}{n}$, яка в точці x ($<-\infty<x<\infty$) дорівнює відносній частоті тих елементів варіаційного ряду, що не більші від x. Тобто

(2) $F_n(x) = \frac{k}{n}, \quad x_{(k)} \le x \prec x_{(k+1)}, \quad k = 1,2,\ldots,n-1.$

Функція $F_n(x)$ називається **емпіричною функцією розподілу**. При кожному x ордината емпіричної функції розподілу $F_n(x)$ є випадковою змінною з $n+1$ можливими

значеннями: *0, 1/n, 2/n, …,(n-1)/n, 1*. Припустимо, що спостереження (1) проведено над випадковою змінною ξ з функцією розподілу $F(x)$. Функція $F(x)$ тоді називається **теоретичною функцією розподілу**. У 1933 р. В.Г. Глівенко (1897-1940) довів, що майже вірогідна емпірична функція розподілу збігається до теоретичної функції розподілу $P\left\{F_n(x) \underset{n\to\infty}{\to} F(x)\right\} \to 1$.

Практичний висновок. Якщо теоретична функція розподілу відома, то емпірична функція розподілу вказує на погодженість теорії з експериментом. Якщо ж теоретична функція розподілу не відома, то емпірична функція розподілу дає уявлення про її можливий вигляд.

Обидва підходи використовуються практично. У першому випадку змінна ξ називається **випадковою змінною**, а в другому – **статистичною змінною**.

Таким чином, терміни випадкова і статистична змінні вказують на два аспекти тієї ж самої мінливої величини. Теорема Глівенка вказує на те, що емпірична функція розподілу несе інформацію про теоретичну функцію розподілу, або що те саме, що статистичний матеріал несе інформацію про теоретичну функцію розподілу. Щоб це зручніше висловити впровадимо такі два поняття: **генеральна сукупність і вибірка**.

Означення. Сукупність всеможливих значень випадкової змінної називають **генеральною сукупністю або популяцією**.

Означення. Ряд незалежних спостережень над випадковою змінною називають **вибіркою з генеральної сукупності**.

Таким чином теорема Глівенка доводить, що вибірка несе інформацію про генеральну сукупність.

Основна задача математичної статистики полягає в тому, щоб виявити інформацію, яку несе вибірка про генеральну сукупність.

Означення. Всяке твердження про генеральну сукупність на основі вибірки називаємо **гіпотезою**.

Означення. Міркування, на основі яких приходимо до висновків про гіпотезу називаємо **статистичним доведенням**.

Кожне статистичне доведення в основному проводиться за такою схемою.

Схема статистичного доведення

В кожному статистичному доведенні є наступні кроки:

- Формулюємо гіпотеза H.
- Вибираємо рівень значущості α.
- Вибираємо відповідно гіпотезі статистика S_t.
- Знаходимо розподіл цієї статистики.
- На основі знайденого розподілу визначаємо критичну область для статистики.
- Знаходимо емпіричне значення статистики.
- Приймаємо вирішення про гіпотезу.

Якщо емпіричне значення статистики попадає в критичну область для гіпотези, то гіпотезу відкидаємо. Якщо ж емпіричне значення статистики не попадає в критичну для гіпотези, то гіпотезу приймаємо і кажемо що вона не суперечить експериментальним даним.

Статистичні гіпотези відносно генеральної сукупності можуть бути дуже різноманітними. Наприклад: про розподіл, про параметр розподілу, про рівність математичних сподівань або дисперсій тощо.

Статистичне доведення істотно відрізняються від математичного доведення. Математичне доведення базоване на логіці, тоді як статистичне доведення може мати такі чотири ситуації:

- Гіпотеза істинна і в результаті статистичного доведення ми її приймаємо.
- Гіпотеза хибна і в результаті статистичного доведення ми її відкидаємо.
- Гіпотеза істинна, але в результаті статистичного доведення ми її відкидаємо.
- Гіпотеза хибна, але в результаті статистичного доведення ми її приймаємо.

Таким чином при статистичному доведенні можливо допустити одну з двох похибок: або відкинути істинну гіпотезу (похибка 1-го типу), або прийняти хибну гіпотезу (похибка 2-го типу). Умовно допустити похибку 1-го, тобто ймовірність відкинути істинну похибку називаємо **рівнем значущості даного критерія** і позначаємо через α. В наукових дослідженнях α переважно вибирається 0.05, а рідше 0.1 або 0.01. Перше статистичне доведення в сучасному розумінні провів англійський математик К. Пірсон у 1900 р. – році заснування математичної статистики.

При кожному статистичному доведенні центральною точкою є розподіл статистики на основі якої приймаємо рішення про гіпотезу. Тому цей розподіл за означенням є **критерієм**.

Критерій Хі-квадрат (χ^2)

Нехай $x_1,\ldots,x_n$ – вибірка з генеральної сукупності ξ. Потрібно перевірити гіпотезу H, про те, що функція розподілу генеральної сукупності є $F(x)$. $H: F(x)$

Поділимо генеральну сукупність довільним чином на $r+1$ частин, які позначимо $S_1,\ldots,S_i,\ldots,S_{r+1}$. Нехай в область S_i попадає m_i елементів вибірки. Очевидно, що $m_1+m_2+\ldots+m_{r+1}=n$.

На основі гіпотетичної функції розподілу $F(x)$ знаходимо, що імовірність того, що спостережувана змінна буде в комірці S_i $P\{\xi\in S_i\}=p_i$.

Очевидно, що сума таких ймовірностей буде рівна 1.

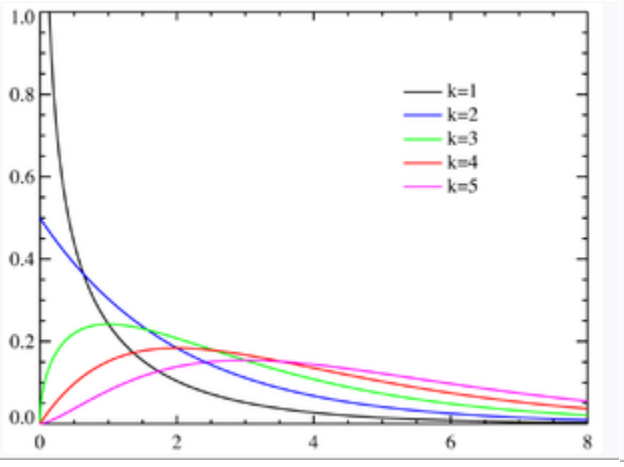
За міру відхилення теоретичного розподілу від вибірки К. Пірсон прийняв величину:

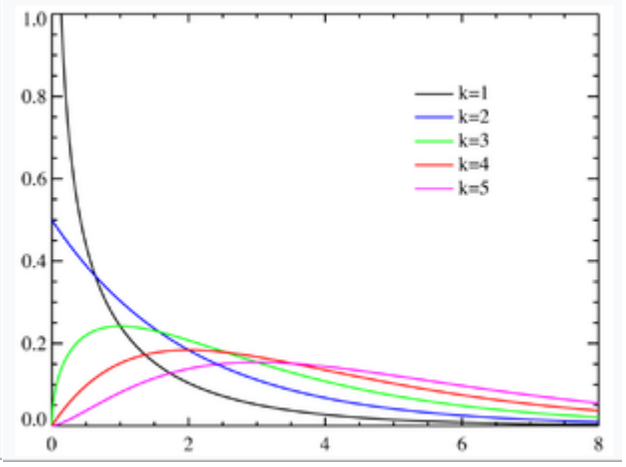
$$\chi^2(r,n,F)=\sum_{i=1}^{r+1}\frac{(m_i-np_i)^2}{np_i}=\sum_{i=1}^{r+1}\frac{m_i^2}{np_i}-n$$

і довів, що для $n\to\infty$ статистика χ^2 має розподіл, який задається густиною:

$$p_r(x)=\begin{cases}0, & x<0\\ \dfrac{1}{\Gamma(r/2)2^{r/2}}x^{(r/2)-1}e^{-x/2}, & x>0\end{cases}.$$

Розподіл з густиною $p_r(x)$ називають розподілом χ^2 з r ступенями свободи. Нижче на двох рисунках зображені густини $p_r(x)$ та відповідно хі-квадрат розподіли при $r=k$ від 1 до 5.





На основі густини $p_r(x)$ визначаємо критичну область для гіпотези при заданому рівні значущості α. Для різних α і різних ступенів свободи (d .f), критерій χ^2 табульований. Якщо χ^2 емпіричне $>\chi^2$ критичне, то гіпотезу H відкидають.

Умови застосовності критерію хі-квадрат:

- Першою необхідною умовою застосовності критерію хі-квадрат є те, щоб вибірка була велика ($n \geq 5$).
- Друга необхідна умова, щоб у кожному класі поділу генеральної сукупності було не менше ніж 10 спостережень ($m_i \geq 10$). Якщо в якомусь класі менше ніж 10 спостережень, то його об'єднують із сусіднім класом в один. Інколи треба об'єднати кілька класів в один.
- Третя необхідна умова застосовності критерію хі-квадрат: якщо на основі вибірки оцінюється s невідомих параметрів, то число ступенів вільності зменшують на s . Якщо число класів після об'єднання є $r+1$, а число параметрів s, то число ступенів вільності дорівнює $r-s=d.f. \geq 1$.

Метод максимуму правдоподібності

Математична статистика – це наука про те, як

1. планувати експеримент;
2. збирати статистичний матеріал;
3. аналізувати статистичний матеріал математичними методами;
4. робити прогнози.

Надалі будемо розкривати пункт 3, причому будемо розглядати статистичні методи для одновимірної кількісної мінливої величини.

Нехай $x_1,...,x_n$ – ряд незалежних спостережень проведених в однакових умовах над статистичною змінною ξ, що має функцію розподілу $F(x)$ залежну від s невідомих параметрів $\alpha_1,...,\alpha_s$

$$\xi : F(x;\alpha_1,...,\alpha_s) \tag{1}$$

Задача: оцінити невідомі параметри на основі вибірки.

Англійський статистик Р. Фішер у 1912р. запропонував наступний метод оцінки невідомих параметрів. Якщо статистична змінна ξ абсолютно неперервна і має густину $p(x,\alpha_1,...,\alpha_s)$, то розглядають таку фунцію:

$$L(x_1,...,x_n;\alpha_1,...,\alpha_s) = \prod_{i=1}^{n} p(x_i;\alpha_1,...,\alpha_s). \tag{2}$$

Якщо статична змінна ξ дискретна і приймає значення з ймовірністю $P_j(\alpha_1,...,\alpha_s) = P\{\xi = j\}$, (j=0,1,…), то розглядаємо функцію

$$L(x_1,...,x_n;\alpha_1,...,\alpha_s)=\prod_{i=1}^{n}p_{x_i}(\alpha_1,...,\alpha_s). \qquad (3)$$

В обидвох випадках функцію $L=L(x_1,...,x_n;\alpha_1,...,\alpha_s)$ називають функцією **правдоподібності**. Невідомі параметри оцінюємо з необхідної умови максумуму функції правдоподібності.

Очевидно, що необхідною умовою максимуму функції багатьох змінних є рівність нулю частинних похідних від функції правдоподібності

$$\frac{\partial L}{\partial \alpha_k}=0, \quad (k=1,...,s).$$

Обидві сторони останьої рівності помножимо на $\frac{1}{L}$, одержимо:

$$\frac{\partial \ln L}{\partial \alpha_k}=0, \quad (k=1,...,s). \qquad (4)$$

Остання система рівнянь називається **системою рівнянь правдоподібності**.

Кожне розв'язання системи рівнянь правдоподібності, що залежить від вибіркових значень $x_1,...,x_n$ називається **оцінкою максимальної правдоподібності** для $\alpha_1,...,\alpha_s$. Ми не будемо розглядати умов існування розв'язку системи рівнянь правдоподібності, їх єдиності, а в конкретних випадках на ці питання дамо відповіді по змозі і до кінця.

Таким чином, метод максимуму правдоподібності полягає в тому, що за оцінку невідомих параметрів приймаємо такі розв'язки системи (4), відносно $\alpha_1,...,\alpha_k$, при яких функції (2), (3) досягають найбільшого значення.

Приклад 1. Методом максимуму правдоподібності оцінити на основі заданої вибірки параметри нормального розподілу (сподівання та диспер- сію)

Нехай $x_1,...,x_n$ вибірка з нормальної популяції, що має густину

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

Для оцінки невідомих параметрів a та σ^2 розглянемо функцію правдоподібності:

$$L=\frac{1}{(2\pi)^{(n/2)}\left(\sigma^2\right)^{(n/2)}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2}.$$

Звідси

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2.$$

Запишемо систему рівнянь правдоподібності:

$$\begin{cases}\dfrac{\partial \ln L}{\partial a}\equiv\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-a)^2=0\\ \dfrac{\partial \ln L}{\partial \sigma^2}\equiv-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-a)^2=0\end{cases}$$

Дана система має єдиний розв'язок:

$$a=\frac{1}{n}\sum_{i=1}^{n}x_i\equiv\bar{x}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2\Big|_{a=\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2 \equiv \frac{n-1}{n}s^2.$$

Розв'язок єдиний, а чи він надає максимум.
Оскільки,

$$\left|\begin{matrix}\dfrac{\partial^2 \ln L}{\partial a^2} & \dfrac{\partial^2 \ln L}{\partial a\,\partial\sigma^2}\\ \dfrac{\partial^2 \ln L}{\partial a\,\partial\sigma^2} & \dfrac{\partial^2 \ln L}{(\partial\sigma^2)^2}\end{matrix}\right|_{\substack{a=\bar{x}\\ \sigma^2=\frac{n-1}{n}s^2}} = \left|\begin{matrix}-\dfrac{n^2}{(n-1)s^2} & 0\\ 0 & -\dfrac{n^3}{2(n-1)^2 s^4}\end{matrix}\right| > 0 \neq 0$$

і

$$\left.\frac{\partial^2 \ln L}{\partial a^2}\right|_{\substack{a=\bar{x}\\ \sigma^2=\frac{n-1}{(n-1)s^2}}} = -\frac{n^2}{(n-1)s^2} < 0$$

то при $a=\bar{x}$ і $\sigma^2 = \dfrac{n-1}{n}s^2$ функція правдоподібності нормального розподілу має максимум.

Таким чином максимум правдоподібності сподівання нормального розподілу оцінюється середнім арифметичним $a=\bar{x}$, а дисперсія – другим центральним моментом, або варіансою наближення, $\sigma^2 = \mu_2 = \dfrac{n-1}{n}s^2$.

Приклад 2. Методом максимуму правдоподібності оцінити на основі вибірки параметр експонентного розподілу.

Нехай $x_1,...,x_n$ - вибірка з експонентної популяції, що має густину

$$p(x,\lambda) = \lambda e^{-\lambda x},\ x>0.$$

1-ий крок. Записуємо функцію правдоподібності: замість x ставимо $x_1,...,x_n$ і перемножаємо:

$$L = \lambda^n e^{-\lambda\sum_{i=1}^{n}x_i}.$$

2-ий крок. Логарифмуємо функцію L

$$\ln L = n\ln\lambda - \lambda\sum_{i=1}^{n}x_i.$$

Записуємо рівняння правдоподібності

$$\frac{\partial \ln L}{\partial\lambda} \equiv \frac{n}{\lambda} - \sum_{i=1}^{n}x_i = 0,$$

який має єдиний розв'язок $\lambda = \dfrac{1}{\bar{x}}$, а чи надає максимуму функції L, то треба взяти другу похідну:. $\left.\dfrac{\partial^2 \ln L}{\partial\lambda^2}\right|_{\lambda=\frac{1}{\bar{x}}} = -n\bar{x}^2 < 0$. Оскільки друга похідна від'ємна то маємо максимум.

Таким чином методом максимуму правдоподібності параметр експоненціального розподілу оцінюється оберненою величиною середнього арифметичного.

Приклад 3. ММП оцінити параметр розподілу Пуассона на основі вибірки (тобто знайти оцінку для параметра λ).

Нехай $x_1,...,x_n$ - вибірка з генеральної сукупності пуассонівської розподіленої змінної

$$\mathrm{P}\left(\xi = j\right) = e^{-\lambda}\frac{\lambda^j}{j!},\quad \left(j=0,1,2,...;\quad \lambda>0\right).$$

Функція правдоподібності у нашому випадку приймає вигляд

$$L = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n}x_i}}{x_1!x_2!...x_n!}.$$

Звідси

$$\ln L = -n\lambda + \left(\sum_{i=1}^{n}x_i\right)\ln\lambda - \sum_{i=1}^{n}\ln(x_i!).$$

На підставі (4) записуємо рівняння правдоподібності

$$\frac{\partial \ln L}{\partial \lambda} = -n + (\sum_{i=1}^{n}x_i)\frac{1}{\lambda} = 0.$$

Це рівняння має єдиний розв'язок при $\lambda = \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{x}$,

оскільки $\left.\frac{\partial^2 \ln L}{\partial \lambda^2}\right|_{\lambda=\overline{x}} = -\frac{n}{\overline{x}} < 0$, то при $\lambda = \overline{x}$ функція правдоподібності пуассонівського розподілу має максимум.

Таким чином методом максимуму правдоподібності параметр розподілу пуассона на основі вибірки оцінюється середнім арифметичним.

Зауваження. Часто система рівнянь правдоподібності трансцендентна і навіть за допомогою сучасних ЕОМ буває нелегко її розв'язати. В таких випадках інколи вдається оцінити невідомі параметри методом моментів. Метод моментів полягає в тому, що ми прирівнюємо між собою стільки початкових теоретичних і емпіричних моментів, скільки невідомих параметрів.

Приклад 4. Оцінити на основі вибірки методом максимуму правдоподібності параметр розподілу λ із заданою густиною

$$p(x;\lambda) = \frac{1}{\Gamma(\lambda)}e^{-x}x^{\lambda-1}, \quad (x>0, \lambda>0). \qquad (A)$$

Нехай $x_1,...,x_n$ – вибірка із популяцій, що має густину $p(x;\lambda)$. Функція правдоподібності у нашому випадку буде

$$L = \frac{1}{\Gamma^n(\lambda)}\left(\prod_{i=1}^{n}x_i\right)^{\lambda-1}e^{-\sum_{i=1}^{n}x_i}.$$

Звідси

$$\ln L = -n\ln\Gamma(\lambda) - \sum_{i=1}^{n}x_i + (\lambda-1)\sum_{i=1}^{n}\ln x_i = 0$$

Рівняння правдоподібності має вигляд

$$\frac{\partial \ln L}{\partial \lambda} \equiv -n\frac{\Gamma'(\lambda)}{\Gamma(\lambda)} + \sum_{i=1}^{n}\ln x_i = 0.$$

Логарифмічна похідна гама $\Gamma(\lambda)$ - функції

$$\psi(\lambda) = \frac{\Gamma'(\lambda)}{\Gamma(\lambda)},$$

при $\lambda > 0$, монотонно зростає від $-\infty$ до $+\infty$. Тому рівняння правдоподібності

$$\psi(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\ln x_i$$

має єдиний розв'язок

$$\lambda = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right)$$

де ψ^{-1} – функція обернена до ψ; $\psi\left[\psi^{-1}(x)\right] = x$. Оскільки

$$\frac{\partial^2 \ln L}{\partial \lambda^2} = -n\psi'(\lambda) > 0$$

і похідна всюди невід'ємна, то при $\lambda = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right)$

функція правдоподібності має максимум. Але вираз досить складний для практичного застосування. Тому постараємось оцінити невідомий параметр λ методом моментів.

Перший початковий момент розподілу (А) збігається із сподіванням і рівний

$$m_1 = E\xi = \int_0^\infty x \frac{1}{(\lambda)e^{\lambda-1}dx} = \frac{1}{(\lambda)}\int_0^\infty xe^{\lambda-1}dx = \lambda$$

Відомо, що перший емпіричний початковий момент збігається із середнім арифметичним $\frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$. Таким чином оцінкою невідомого параметра λ, згідно з методом моментів є вибіркове середнє

$$\lambda = \bar{x}$$

Ця оцінка значно простіша від отриманої.

Слід підкреслити, що оцінка параметрів одержані ММП мають взагалі більше властивостей, ніж оцінки одержані методом моментів.

Гіпотези про параметри нормальних розподілів

1. Гіпотеза про сподівання

Нехай $x_1, x_2, \cdots, x_n$ (1) – ряд незалежних спостережень проведених в однакових умовах над незалежною нормальною статистичною змінною ξ. Потрібно перевірити нульову гіпотезу H_0 про те, що математичне сподівання генеральної сукупності , з якої отримана ця вибірка, дорівнює параметру a.

$H_0 : E\xi = a$, де a – деяка стала.

На основі вибірки (1) обчислюємо середнє та варіансу

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2.$$

Розглянемо статистику:

$$t = \frac{\bar{x} - a}{s}\sqrt{n}. \qquad (2)$$

Англійський статистик В.Госсет (Gosset) у 1908 році в роботі під псевдонімом Стьюдент довів, що статистика t має розподіл із густиною

$$p(x, n-1) = \frac{1}{\sqrt{(n-1)\pi}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} (1 + \frac{x^2}{n-1})^{-\frac{n}{2}}, \quad (-\infty < x < +\infty, \quad n = 2,3,...).$$

Густину $p(x, n-1)$ називають **густиною розподілу Стьюдента** з $(n-1)$ ступенями вільності: $d.f. = n-1$.

При $n = 2$ отримаємо густину Коші

$$p(x) = \frac{1}{\pi}(\frac{1}{1+x^2}).$$

На підставі густини розподілу $p(x, n-1)$ та статистики Стьюдента (2) будується критерій істотності для відхилення вибіркового середнього значення $\bar{x}$ від гіпотетичного значення a.

Зазначимо, що при $n \to \infty$, густина $p(x, n-1)$ збігається до нормально розподіленої густини.

На основі розподілу Стьюдента визначаємо критичну область для заданого рівня значущості α та числа ступенів вільності $d.f. = n-1$. Вона складається з двох частин, симетричних відносно ординат, кожна з яких має площу по $\alpha/2$. Статистику (2) називають **статистикою Стьюдента**. Для статистики Стьюдента складені таблиці критичних значень для різних ступенів вільності (додаток 7).

Сформулюємо алгоритм перевірки нульової гіпотези за допомогою критерію Стьюдента:

1. Вибираємо рівень значущості α.
2. Обчислюємо за формулою (2) емпіричне значення статистики Стьюдента, визначивши попередньо за формулою (1) середнє вибіркове та варіансу.
3. При $\alpha/2$ та кількості ступенів вільності $d.f. = n-1$ з додатка 7 знаходимо критичне значення цієї статистики.
4. Якщо $|t_{емп}| < t_{кр}$, то нульову гіпотезу про те, що $E\xi = a$, приймаємо. У протилежному випадку вважаємо, що вона суперечить експериментальним даним і може бути відхилена.

У класичній роботі про t - розподіл Стьюдент наводить такі дані про додаткові години сну в 10 пацієнтів, викликаними снодійними препаратами A та B.

1,1	A(x)	B(y)	Різниця z=x-y
1	1,9	0,7	1,2
2	0,8	-1,6	2,4
3	1,1	-0,2	1,3
4	0,1	-1,2	1,3
5	-0,1	-0,1	0,0
6	4,4	3,4	1,0
7	5,5	3,7	1,8
8	1,6	0,8	0,8
9	4,6	0,0	4,6
10	3,4	2,0	1,4

Перевірити, чи істотна різниця між дією двох снотворних засобів A та B. Якщо припустити, що різниця між додатковими годинами сну викликана дією засобів A та B, тоді $z = x - y$ буде нормально розподіленою вибіркою.

Формулюємо нульову гіпотезу: H_0 – немає істотної різниці між дією двох снотворних засобів.

Доведення проведемо за допомогою критерію Стьюдента.
Розглянемо різницю $z = x - y$. На основі експериментальних даних знаходимо середню різницю $\bar{z} = 1,58$. Та стандарт

$$s_z = \sqrt{\frac{1}{10-1}\sum_{i=1}^{10}(z_i - \bar{z})^2} = 1,111.$$

Відношення Стьюдента дорівнює

$$|t|_{емп} = \frac{1,58-0}{1,111}\sqrt{10} = 4,49.$$

Критичне значення знаходимо з таблиці для $\alpha = 0,05$ та числа ступенів вільності $d.f. = 9$ $t_{кр} = 2,26$. Оскільки $t_{кр} < |t|_{емп}$ то гіпотезу про еквівалентність снодійних засобів відкидаємо, тобто існує істотна різниця між дією двох снотворних засобів.

2. Інтервал довіри для невідомого сподівання

На основі області прийому гіпотези про сподівання нормальної величини можемо записати співвідношення, тобто при визначенні критичної області для критерію Стьюдента одержуємо таке співвідношення

$$P\{|t| < t_{\alpha/2}\} = 1 - \alpha .$$

Умова не зміниться, якщо її замінити на еквівалент

$$P\{|\frac{a - \bar{x}}{s}\sqrt{n}| < t_{\alpha/2}\} = 1 - \alpha ,$$

$$P\{-t_{\alpha/2} < \frac{a - \bar{x}}{s}\sqrt{n} < t_{\alpha/2}\} = 1 - \alpha ,$$

$$P\{\bar{x} - \frac{s}{\sqrt{n}}t_{\alpha/2} < a < \bar{x} + \frac{s}{\sqrt{n}}t_{\alpha/2}\} = 1 - \alpha .$$

Отже з імовірністю $1 - \alpha$ випадковий інтервал:

$$[\bar{x} - \frac{s}{\sqrt{n}}t_{\alpha/2};\ \bar{x} + \frac{s}{\sqrt{n}}t_{\alpha/2}] \qquad (2)$$

накриває невідоме сподівання a нормально розподіленої генеральної сукупності. Цей інтервал називають **інтервалом довіри,** або **довірчим інтервалом** для сподівання a, при рівні значущості α.

клад. Проведене 31 спостереження над нормально розподіленою статистичною змінною ξ, на основі яких одержали середнє вибіркове $\bar{x} = 38,61$, та варіансу $s^2 = 33,43$.

Знайти 90% інтервал довіри для невідомого сподівання генеральної сукупності.

У цьому випадку $n = 31; \alpha = 0,1; 1 - \alpha = 0,9; d.f. = 30$. З таблиці, при таких вихідних даних, беремо $t_{\frac{\alpha}{2}} = t_{\frac{1}{2}} = t_{кр} = 1,7$ (додаток 7)

Шуканим інтервалом згідно формули (2) є

$[\bar{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}] = [56,85;\ 60,57]$, який з ймовірністю 90 % містить невідоме сподівання цієї генеральної сукупності.

3. Порівняння сподівань двох нормальних розподілів

Нехай $x_1, x_2, \cdots, x_{n_1}$ – ряд незалежних спостережень над незалежною нормальною статистичною змінною ξ, а $y_1, y_2, \cdots, y_{n_2}$ – ряд незалежних спостережень над незалежною нормальною статистичною змінною η. Перевірити гіпотезу про те, що математичне сподівання популяції, з яких одержані ці вибірки мають однакове математичне сподівання, тобто: $H_0 : E\xi = E\eta$.

На основі вибірок знаходимо середні

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \quad \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$$

та варіанси

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2, \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2 .$$

Виберемо рівень значущості α та розглянемо статистику:

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}\sqrt{\frac{n_1n_2}{n_1+n_2}}.$$

У 1925 році Р.Фішер довів, що статистика t має розподіл Стьюдента з кількістю ступенями вільності $d.f.=n_1+n_2-2$. Наступна частина аналогічна попередньо розглянутому випадку.

Приклад. В одному класі з 20 дітей випадково вибирають 10, яким щоденно почали видавати апельсиновий сік. Решта 10 щоденно отримували молоко. Через деякий час зафіксоване таке збільшення ваги дітей (у фунтах):

I група: 4 2,5 3,5 4 1,5 1 3,5 3 2,5 3,5

II група:1,5 3,5 2,5 3 2,5 2 2 2,5 1,5 3

Середній приріст ваги в першій групі дорівнює 2,9 а в другій групі – 2,4. Чи істотна ця різниця?

Формулюємо нульову гіпотезу H_0 : нема істотної різниці між збільшенням ваги у 2-х групах. Використаємо критерій Стьюдента.

$$t_{емп}=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}\sqrt{\frac{n_1n_2}{n_1+n_2}}=\frac{2,9-2,4}{\sqrt{\frac{13,3}{18}}}\sqrt{\frac{100}{20}}=1,3$$

При $\alpha=0,05$ і $d.f.=18$: $t_{кр}=2,10$.

Гіпотезу приймаємо, бо $t_{емп}<t_{кр}$

потеза про дисперсію нормальної популяції

Дисперсія характеризує точність машин і приладів, точність технологічного процесу, похибку показань вимірювального приладу і таке інше. Тому важливо приймати рішення про гіпотези відносно дисперсії.

Нехай $x_1,x_2,\cdots,x_n-$ ряд незалежних спостережень над нормально розподіленою статистичною змінною ξ.

Потрібно перевірити гіпотезу про те, що дисперсія нормальної популяції з якої взята вибірка дорівнює σ^2, тобто, $H_0: D\xi=\sigma^2$.

За даними вибірки обчислимо її середнє та варіансу:

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

Виберемо рівень значущості α та розглянемо статистику $\frac{s^2}{\sigma^2}$. Якщо гіпотеза вірна, то ця статистика має розподіл $\frac{\chi^2}{d.f.}$, де $d.f.=n-1-$ число ступенів вільності. На основі розподілу статистики $\frac{\chi^2}{d.f.}$ визначаємо критичну область для гіпотези. Очевидно, що для гіпотези будуть сприятливі ті випадки, коли значення статистики $\frac{s^2}{\sigma^2}$ близьке до 1. Тому критична область для статистики складається із двох частин: дуже малих і дуже великих значень змінної $\frac{\chi^2}{d.f.}$. При рівні значущості α нижнім критичним значенням є $\frac{\chi^2_{\alpha/2}}{d.f.}$, а верхнім $-\frac{\chi^2_{1-\alpha/2}}{d.f.}$.

Якщо емпіричне значення статистики $(\frac{\chi^2}{d.f.})_{емп} = \frac{s^2}{\sigma^2}$ знаходиться між $\frac{\chi^2_{\alpha/2}}{d.f.}$ і $\frac{\chi^2_{1-\alpha/2}}{d.f.}$, то гіпотезу про те, що дисперсія дорівнює σ^2 приймаємо, а в інших випадках відхиляємо.

клад. Стандартні екзамени проходили кілька років із сподіванням a=70 і дисперсією 9. Для деякої групи 25 студентів за результатами їх успішності отримали середнє ення сумарної їх успішності $\bar{x} = 71$ і варіансу $s^2 = 12$. Чи є підстави сумніватися, що ерсія сумарної успішності всіх студентів буде дорівнювати 9? $D\xi = 9$.

одимо $(\frac{\chi^2}{d.f.})_{емп} = \frac{12}{9} = 1,33$.

При $\alpha = 0,05$ і $d.f. = 24$ $(\frac{\chi^2}{d.f.})_{в.кр} = 1,64 > (\frac{\chi^2}{d.f.})_{емп}$.

Підстав сумніватися немає. Гіпотезу приймаємо.

тервал довіри для невідомої дисперсії нормальної популяції

Розподіл статистики $\frac{\chi^2}{d.f.}$ дає можливість при заданому рівні значущості α на підставі вибірки вказати з імовірністю 1-α інтервал довіри для невідомої дисперсії σ^2.

При вибраному числу значущості α і числу степенів вільності $d.f.$ для яких виконується співвідношення

$$P\{\frac{\chi^2_{\alpha/2}}{d.f.} < \frac{s^2}{\sigma^2} < \frac{\chi^2_{1-\alpha/2}}{d.f.}\} = 1-\alpha,$$

отримаємо інтервал довіри для дисперсії нормальної популяції

$$P\{\frac{s^2}{\frac{\chi^2_{1-\alpha/2}}{d.f.}} < \sigma^2 < \frac{s^2}{\frac{\chi^2_{\alpha/2}}{d.f.}}\} = 1-\alpha.$$

Отже з імовірністю 1-α випадковий інтервал $[\frac{s^2}{\frac{\chi^2_{1-\alpha/2}}{d.f.}}; \frac{s^2}{\frac{\chi^2_{\alpha/2}}{d.f.}}]$ накриває невідоме значення дисперсії генеральної сукупності. Цей інтервал називають $100\cdot(1-\alpha)\%$ **інтервалом довіри**, або **довірчим інтервалом** для дисперсії σ^2 при рівні значущості α.

Приклад. Проведене 31 спостереження над нормально розподіленою статистичною змінною ξ. Одержали $\bar{x} = 58,61$, $s^2 = 33,43$.
Знайти 90 % інтервал довіри до дисперсії генеральної сукупності.

За умовою задачі $\alpha = 0,1$ і $d.f. = 30$. Тому з таблиці [] маємо $\frac{\chi^2_{0,05}}{d.f.} = 0,616$,

$\frac{\chi^2_{0,95}}{d.f.} = 1,46$.

Шуканий інтервал $[22,91;\ 54,22]$ є 90 %-им інтервалом довіри для невідомої дисперсії генеральної сукупності.

Порівняння дисперсій двох нормальних розподілів

Нехай $x_1, x_2, \cdots, x_{n_1}$.– вибірка з нормально розподіденої статистичної змінної ξ_1 , а $y_1, y_2, \cdots, y_{n_2}$ – незалежна від неї вибірка з нормально розподіленої статистичної змінної ξ_2.

Потрібно перевірити гіпотезу про те, що дисперсії генеральних сукупностей, з яких взяті ці вибірки, однакові: $H_0: D\xi_1 = D\xi_2$

На основі даних вибірок знаходимо їх вибіркові середні та варіанси відповідно:

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \quad \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i, \quad s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2, \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2.$$

Виберемо рівень значущості α та розглянемо статистику

$$F_{емп} = \frac{s_1^2}{s_2^2}. \qquad (*)$$

Р. Фішер (англійський статистик) у 1926 р. довів, що у випадку істинності висунутої нульової гіпотези статистика $F_{емп}$ має розподіл із густиною розподілу:

$$p(x; m, n) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2}) \cdot \Gamma(\frac{n}{2})} (\frac{m}{n})^{\frac{m}{2}} x^{\frac{m}{2}-1} (1 + \frac{m}{n} x)^{-\frac{m+n}{2}}, \quad (**)$$

де $m = n_1 - 1, n = n_2 - 1.$

Вираз (*) називають дисперсійним відношенням Фішера, а (**) – щільністю Фішера з (m, n) ступенями вільності:

$m = n_1 - 1$- число ступенів вільності чисельника

$n = n_2 - 1$ - число ступенів вільності знаменника дисперсійного відношення (*). Зауважимо, що при $n \to \infty$ статистика $F_{емп}$ збігається до статистики $\frac{\chi^2}{d.f.}$. п.4 ($\frac{S^2}{\sigma^2}$).

На основі статистики $F_{емп}$ визначаємо критичну область для гіпотези. Очевидно, що для гіпотези сприятливими будуть ті випадки коли $F_{емп}$ близьке до 1. Тому критична область для гіпотези складається з двох частин: із дуже малих і дуже великих значень статистики $F_{емп}$.

Довірча і критична область критерію Фішера є наступні: критична зліва на величину α/2, довірча посередині на величину 1-α і критична справа на величину α/2.

Тут α – заданий рівень значущості і, якщо $F_{емп} < F_{н.кр.} = F_{\frac{\alpha}{2}}$, або $F_{емп} > F_{в.кр.} = F_{1-\frac{\alpha}{2}}$, то гіпотезу відкидаємо. В протилежному випадку кажемо, що вона не суперечить експериментальним даним.

Зауваження 1. $F_{в.кр.}$ – табульована для різних пар чисел ступенів вільності $d.f. = (m, n)$ і рівнів значущості α. Нижнє критичне значення дістаємо як обернену величину верхнього критичного значення статистики $F_{емп}$ при зміненій черговості ступенів вільності, тобто

$$F_{\frac{\alpha}{2}}(m, n) = \frac{1}{F_{1-\frac{\alpha}{2}}(n, m)}.$$

Якщо в чисельнику дисперсійного відношення $F_{емп}$ ставити більшу варіансу, то це відношення буде більше одиниці, і тоді достатньо розглядати лише верхні критичні

значення. У цьому випадку, якщо емпіричне значення статистики $F_{емп}$ попаде у верхню критичну область ($F_{емп} > F_{в.кр.} = F_{1-\frac{\alpha}{2}}$), то гіпотезу відхиляємо як хибне твердження.

Приклад. На основі вибірок з двох нормально розподілених генеральних сукупностей обсягів $n_1 = 15, n_2 = 12$ обчислено $S_1^2 = 17$, $S_2^2 = 29$. Чи можна вважати при десяти процентному рівні значущості, що дисперсії обох популяцій однакові?

У цьому випадку $\alpha = 0.1$. Згідно означення дисперсійного відношення Фішера $F_{емп} = \frac{29}{17} = 1,7$. У цьому відношенні у чисельнику маємо більшу за величиною варіансу. Тому з таблиці (додаток 8), при $1-\frac{\alpha}{2} = 0,95$ та кількості ступенів вільності $d.f. = (11,14)$, знаходимо лише верхнє критичне значення статистики Фішера: $F_{в.кр.} = 2,57$. Оскільки $F_{емп} < F_{в.кр.}$, то немає підстав для сумнівів, що дисперсії генеральних сукупностей рівні між собою.

Приклади застосування критерію

При вивченні міцності 2-х типів бетону одержано

$n_1 = 8\ \bar{x}_1 = 4492\ s_1^2 = 16610$

$n_2 = 8\ \bar{x}_2 = 4150\ s_2^2 = 29050$

Гіпотеза Н1: міцність бетону не залежить від типу.

Шукаємо $t_{емп} = \frac{4492 - 4150}{\sqrt{\frac{7(16610 + 29030)}{8+8-2}}} \sqrt{\frac{8*8}{8+8}} = 4,52$

Вибираємо α=0,05 d.f.=14 |-> tкр=2,14

Гіпотезу відкидаємо, бо $|t_{емп}| > t_{кр}$

Гіпотеза Н2: Ті, що виготовляють бетон 2-х типів однаково кваліфіковані.

Шукаємо $F_{емп} = \frac{29050}{16610} = 1,75$

Вибираємо α=0,05 d.f.=14 |-> Fкр=3,79

Гіпотезу приймаємо, бо $F_{емп} < F_{кр.в}$

На певному підприємстві розробили 2 методи виготовлення означеного виробу.
Витрати сировини при роботі 2-ма методами є наступні:
I: 2.0, 2.7, 2.5, 2.9, 2.3, 2.6
II: 2.5, 3.2, 3.5, 3.5, 3.5
Що можна про це сказати?
Гіпотеза Н1: методи однаково матеріалоємкі.
Визначаємо величини:

$n_1 = 6\ \bar{x}_1 = 2,5\ s_1^2 = 0,1 \qquad n_2 = 5\ \bar{x}_2 = 3,24\ s_2^2 = 0,195$

Шукаємо $t_{емп} = \frac{2,5 - 3,24}{\sqrt{\frac{5*0,1 + 4*0,195}{6+5-2}}} \sqrt{\frac{6*5}{6+5}} = -3,24$

Вибираємо α=0,05 d.f.=9 |-> tкр=2,26

Гіпотезу приймаємо, бо $|t_{емп}| > t_{кр}$

Гіпотеза Н2: Ті, що виготовляють продукцію 2-х видів однаково кваліфіковані.

Шукаємо $F_{емп} = \frac{0,195}{0,1} = 1,95$

Вибираємо α=0,05 d.f.=(4,5) |-> Fкр=5,13

Гіпотезу приймаємо, бо $F_{емп} < F_{кр.в}$

Одновибірковий критерій погодженості(Крит.Колмогорова)

Нехай $x_1, x_2, \cdots, x_{n_1}$ буде ряд незалежних спостережень над неперервною статистичою змінною ξ. Потрібно перевірити Н про те, що популяція, з якої взята вибірка має функцію розподілу F(x), де F(x) – вповні означена. Н: F(x)

На основі вибірки знаходимо емпіричну функцію розподілу $F_n = \frac{числох_i \le x}{n}$

Одновибірковий критерій погодженості

Розглянемо статистику $D_n = \sup\left|F_n(x) - F(x)\right|$

А. М. Колмогоров (1933) довів, що статистика типу $K_n = \sqrt{n} D_n$ має розподіл незалежний від неперервної гіпотетичної функції розподілу F(x) і при $n \to \infty$ збігається до розподілу

$$K(x) = \begin{cases} 0, x < 0 \\ \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2} \end{cases}$$

Ця функція K(x) табульована в Таблиці 7 при різних рівнях значущості α.

Приклад. Дано вибірку з n=25 незалежних спостережень xi над незалежною змінною ξ. Потрібно перевірити Н про те, що вибірка взята з нормально (розподіленої) популяції з середнім α=5 і стандартним відхиленням (стандартом) σ=10.

-1,5 4,8 2,0 8,7 4,3 -10,4 -10,3 2,4 9,4 7,1 7,7 7,8 15,0 11,0 17,5 13,8 19,3 4,4 -3,1 -5,2 -3,6 12,7 3,3 21,8 -6,7

Якщо гіпотеза вірна, то лінійно перетворена за формулою $y_i = \frac{x_i - 5}{10}$ переведе нашу вибірку у вибірку з нормальною популяцією із сподіванням 0 і стандартом 1

Запишемо лінійно перетворену вибірку:

-0,65 -0,02 -0,80 0,37 -0,01 -1,54 -1,53 -0,26 0,44 0,21 0,27 0,28 1,00 0,60 1,25 0,88 1,43 -0,26 -0,81 -1,02 -0,80 0,77 -0,17 1,68 -1,17

Запишемо для останніх даних варіаційний ряд y(i), значення емпіричного та нормального розподілів $F_{25}(y_{(i)}), F(y_{(i)})$ у пунктах варіаційного ряду, а також абсолютні різниці в тих пунктах і зліва в них між обома розподілами $\left|F_{25}(y_{(i)}) - F(y_{(i)})\right|$ і $\left|F_{25}(y_{(i)} - 0) - F(y_{(i)})\right|$

Результати запишемо далі.

$y_{(i)}$	$F_{25}(y_{(i)})$	$F(y_{(i)})$	$F_{25}(y_{(i)}) - F(y_{(i)})$	$F_{25}(y_{(i)} - 0) - F(y_{(i)})$
-1,54	0,04	0,0618	0,0218	0,0618
-1,53	0,08	0,0630	0,0171	0,0230
-1,17	0,12	0,1210	0,0010	0,0410

-1,02	0,16	0,1539	0,0061	0,0339
-0,86	0,2	0,1949	0,0051	0,0349
-0,81	0,24	0,2090	0,0310	0,0090
-0,65	0,28	0,2578	0,0222	0,0178
-0,3	0,32	0,3821	0,0621	0,1021 max
-0,26	0,36	0,3974	0,0374	0,0774
-0,17	0,4	0,4325	0,0325	0,0725
-0,06	0,44	0,4761	0,0361	0,0761
-0,02	0,48	0,4920	0,0120	0,0520
-0,01	0,52	0,4960	0,0240	0,0160
0,21	0,56	0,5832	0,0232	0,0632
0,27	0,6	0,6064	0,0064	0,0464
0,28	0,64	0,6103	0,2944	0,0103
0,37	0,68	0,6443	0,357	0,0043
0,44	0,72	0,6700	0,0500	0,0100
0,60	0,76	0,7257	0,0343	0,0057
0,77	0,80	0,7294	0,0206	0,0194
0,88	0,84	0,8106	0,0294	0,0106
1,00	0,88	0,8413	0,0387	0,0013
1,25	0,92	0,8944	0,0256	0,0144
1,43	0,96	0,9236	0,0304	0,0036
1,68	1,00	0,9535	0,0465	0,0065

З цього видно, що максимальне відхилення між емпіричним та гіпотетичним розподілами є 0,1021, тобто D25= 0,1021 $K_{25} = \sqrt{25} * 0{,}1021 = 0{,}5105$

Воно значно менше від критичного значення при рівні значущості α=0,05, тому гіпотезу приймаємо. Ми довели гіпотезу для уі, але вона – лінійне перетворення. Тому ми довели нашу початкову гіпотезу.

Двовибірковий критерій погодженості (Крит.Смірнова)

Нехай $x_1, x_2, \cdots, x_{n_1}$ буде ряд незалежних спостережень над неперервною статистичною змінною ξ1, а $y_1, y_2, \cdots, y_{n_1}$ - над неперервною статистичною змінною ξ2. Потрібно перевірити,гіпотезу про те, що популяції з двох взятих вибірок є однаково розподілені:

$$H : F(x) \equiv G(x)$$

На основі вибірки знаходимо емпіричну форму розподілу

$$F_m(x) = \frac{число\ x_i \le x}{m} \quad G_n(y) = \frac{число\ y_i \le y}{n}$$

Розглянемо статистику

$$D_{\min} = \sup(F_m(x) - G_n(y))$$

М.В Смірнов (1933) довів, що статистика $C_{mn} = \sqrt{\frac{mn}{m+n}} D_{mn}$ має розподіл незалежний від неперервних гіпотетичних розподілів F і G, який при $m, n \to \infty$ збігається з розподілом Колмогорова K(x)

Критерій Смірнова стосується випадку великих вибірок. Для малих m,n (2<= m,n <=50) існують окремі таблиці розподілу Смірнова

Приклад: Дано 2 незалежні вибірки незалежних спостережень над абсолютно неперервною випадковою змінною.

(x): 1,9 3,2 0,4 1,2 1,1 1,0 1,5 2,7 0,6 2,0

(y): 0,6 3,0 2,3 2,1 1,2 2,6 1,9 1,0 2,4 2,7

H: Обидві множини спостережень взято у одинаково розподілених абсолютно неперервних популяцій $F \equiv G$

Запишемо для обох рядів спостережень спільний варіаційний ряд значень обох емпіричних функцій розподілу в точках спільного варіаційного ряду та абсолютну різницю в цих точках між розподілами.

$x_{(i)}$	$y_{(i)}$	F_{10}	G_{10}	$\lvert F_{10} - G_{10} \rvert$
0,4		0,1	0	0,1
0,6	0,6	0,2	0,1	0,1
1,0	1,0	0,3	0,2	0,1
1,1		0,4	0,2	0,2
1,2	1,2	0,5	0,3	0,2
1,5		0,6	0,3	0,3
1,9	1,9	0,7	0,4	0,3
2,0		0,8	0,4	0,4 max
	2,1	0,8	0,5	0,3
	2,2	0,8	0,6	0,2
	2,4	0,8	0,7	0,1
	2,6	0,8	0,8	0

2,7	2,7	0,9	0,9	0,1
	3,0	0,9	1	0,1
3,2		1	1	0

$D_{10,10} = 0,4$

Для рівня значущості α=0,05, Sкр=1,36

$$S_{10,10} = \sqrt{\frac{10*10}{10+10}}0,4 = 0,894 < 1,36$$

Гіпотезу приймаємо

Приклад: Дано 2 незалежні вибірки незалежних спостережень над абсолютно неперервною випадковою змінною.

(x): 0,4 -0,5 1,7 0,0 -1,1 1,2 -0,3 -0,9 -0,4 0,5

(y): -0,9 1,1 1,5 0,4 0,8 -0,5 0,6 0,9 -0,3 1,2

H: Обидві множини спостережень взято у однаково розподіленої абсолютно неперервної генеральної сукупності

$x_{(i)}$	$y_{(i)}$	F_{10}	G_{10}	$\lvert F_{10} - G_{10} \rvert$
-1,1		0,1	0	0,1
-0,9	-0,9	0,2	0,1	0,1
-0,5	-0,5	0,3	0,2	0,1
-0,4		0,4	0,2	0,2
-0,3	-0,3	0,5	0,3	0,2
0,0		0,6	0,3	0,3
0,4	0,4	0,7	0,4	0,3
0,5		0,8	0,4	0,4 max
	0,6	0,8	0,5	0,3
	0,8	0,8	0,6	0,2
	0,9	0,8	0,7	0,1
	1,1	0,8	0,8	0
1,2	1,2	0,9	0,9	0
	1,5	0,9	1	0,1
1,7		1	1	0

$D_{10,10} = 0,4$

Для рівня значущості α=0,05, Sкр=1,36

$$S_{10,10} = \sqrt{\frac{10*10}{10+10}}0,4 = 0,894 < 1,36$$

Гіпотезу приймаємо

Одновибірковий критерій погодженості Колмогорова

Нехай $x_1, x_2, \cdots, x_n$ – вибірка з генеральної сукупності ξ. Потрібно перевірити гіпотезу про те, що випадкова змінна ξ керується неперервною теоретичною функцією розподілу $F(x)$:

$$H_0 : P\{\xi \le x\} = F(x).$$

Методика перевірки цієї нульової гіпотези грунтується на висновку теореми Глівенка, згідно якої емпірична функція розподілу $F_n(x)$ збігається до теоретичної функції розподілу $F(x)$:

$$P\{F_n(x) \underset{n\to\infty}{\to} F(x)\} \to 1, \quad -\infty < x < +\infty,$$

де емпірична функція розподілу має вигляд: $F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{m_k}{n}, & x_{(1)} \le x < x_{(k+1)}, \quad k = (\overline{1, n-1}) \\ 1, x \ge x_n, \end{cases}$

, (1)

де m_k – кількість елементів $x_{(k)}$ варіаційного ряду даної вибірки, які не більші за x.

Розглянемо статистику:

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F(x)| \qquad (2)$$

А. М. Колмогоров (1933р.) довів, що статистика типу $K_n = \sqrt{n}D_n$ має розподіл незалежний від неперервної гіпотетичної функції розподілу $F(x)$, який при $n \to \infty$ збігається до розподілу:

$$K(x) = \begin{cases} 0, & x < 0, \\ \sum_{k=-\infty}^{+\infty} (-1)^k e^{2x^2} k^2, & x \ge 0, \end{cases} \qquad (3)$$

Статистику K_n називають **статистикою Колмогорова**.

На основі функції розподілу $K(x)$, при заданому рівні значущості α, визначаємо критичну область для гіпотези H_0 (рис.13.2.3 нарисувати графік!!!!). Якщо при цьому емпіричне значення статистики Колмогорова, обчисленої за формулою (2), більше критичного її значення, взятого з таблиці (додаток 12), то гіпотезу H_0 відкидаємо, як хибне твердження.

Для практичних цілей складено таблицю значень функції $K(x)$ (Додаток 11).

З означення критичної області статистики на основі (3), наприклад, для окремих рівнів значущості α маємо таку таблицю статистики Колмогорова, які найчастіше використовують (табл.13.2.1, нарисувати)

Оскільки висновок теореми Колмогорова асимптотичний, то обсяг вибірки повинен бути великим $(n > 40)$. Якщо обсяг вибірки в межах $(2 \le n \le 40)$, то для кожного n існує свій розподіл Колмогорова, на основі якого знаходимо критичні значення.

Сформулюємо на основі критерію Колмогорова наступний алгоритм:

1. Вибираємо рівень значущості α.
2. Будуємо варіаційний ряд для заданої вибірки.
3. Знаходимо значення гіпотетичної функції розподілу в точках варіаційного ряду та модулі різниць значень емпіричної функції у точках безмежно близьких зліва до точок цього ряду та гіпотетичної функції розподілу в точках варіаційного ряду.
4. Знаходимо емпіричне значення $(K_n)_{емп}$ статистики Колмогорова як максимальне значення модулів таких різниць.
5. При вибраному рівні значущості α, використовуючи таблиці (додаток 11-12), або формулу (3), знаходимо критичне значення $(K_n)_{кр}$ статистики Колмогорова.
6. Якщо $(K_n)_{емп} < (K)_{кр}$, то гіпотезу приймаємо, в протилежному випадку – відкидаємо, як хибну.

Приклад. Дано вибірку з n=25 незалежних спостережень над незалежною змінною ξ. Потрібно перевірити Н про те, що вибірка взята з нормально (розподіленої) популяції з середнім α=5 і стандартним відхиленням (стандартом) σ=10.

-1,5 4,8 2,0 8,7 4,3 -10,4 -10,3 2,4 9,4 7,1 7,7 7,8 15,0 11,0 17,5 13,8 19,3 4,4 -3,1 -5,2 -3,6 12,7 3,3 21,8 -6,7

Якщо гіпотеза вірна, то лінійно перетворена за формулою $y_i = \frac{x_i - 5}{10}$ переведе нашу вибірку у вибірку з нормальною популяцією із сподіванням 0 і стандартом 1

Запишемо лінійно перетворену вибірку:

-0,65 -0,02 -0,80 0,37 -0,01 -1,54 -1,53 -0,26 0,44 0,21 0,27 0,28 1,00 0,60 1,25 0,88 1,43 -0,26 -0,81 -1,02 -0,80 0,77 -0,17 1,68 -1,17

Запишемо для останніх даних варіаційний ряд y(i), значення емпіричного та нормального розподілів $F_{25}(y_{(i)}), F(y_{(i)})$ у пунктах варіаційного ряду, а також абсолютні різниці в тих пунктах і зліва в них між обома розподілами $\left|F_{25}(y_{(i)}) - F(y_{(i)})\right|$ і $\left|F_{25}(y_{(i)} - 0) - F(y_{(i)})\right|$

Результати запишемо далі.

$y_{(i)}$	$F_{25}(y_{(i)})$	$F(y_{(i)})$	$F_{25}(y_{(i)}) - F(y_{(i)})$	$F_{25}(y_{(i)} - 0) - F(y_{(i)})$
-1,54	0,04	0,0618	0,0218	0,0618
-1,53	0,08	0,0630	0,0171	0,0230
-1,17	0,12	0,1210	0,0010	0,0410
-1,02	0,16	0,1539	0,0061	0,0339
-0,86	0,2	0,1949	0,0051	0,0349
-0,81	0,24	0,2090	0,0310	0,0090
-0,65	0,28	0,2578	0,0222	0,0178
-0,3	0,32	0,3821	0,0621	0,1021 max
-0,26	0,36	0,3974	0,0374	0,0774
-0,17	0,4	0,4325	0,0325	0,0725
-0,06	0,44	0,4761	0,0361	0,0761

-0,02	0,48	0,4920	0,0120	0,0520
-0,01	0,52	0,4960	0,0240	0,0160
0,21	0,56	0,5832	0,0232	0,0632
0,27	0,6	0,6064	0,0064	0,0464
0,28	0,64	0,6103	0,2944	0,0103
0,37	0,68	0,6443	0,357	0,0043
0,44	0,72	0,6700	0,0500	0,0100
0,60	0,76	0,7257	0,0343	0,0057
0,77	0,80	0,7294	0,0206	0,0194
0,88	0,84	0,8106	0,0294	0,0106
1,00	0,88	0,8413	0,0387	0,0013
1,25	0,92	0,8944	0,0256	0,0144
1,43	0,96	0,9236	0,0304	0,0036
1,68	1,00	0,9535	0,0465	0,0065

З цього видно, що максимальне відхилення між емпіричним та гіпотетичним розподілами є 0,1021, тобто D25= 0,1021 $K_{25} = \sqrt{25} * 0{,}1021 = 0{,}5105$

Воно значно менше від критичного значення при рівні значущості α=0,05, тому гіпотезу приймаємо. Ми довели гіпотезу для y_i, але вона – лінійне перетворення. Тому ми довели нашу початкову гіпотезу.

Двовибірковий критерій погодженості Смірнова

Нехай $x_1, x_2, \cdots, x_{n_1}$ – вибірка з випадкової змінної ξ_1, а незалежна від неї вибірка $y_1, y_2, \cdots, y_{n_2}$ – з випадкової змінної ξ_2. Потрібно перевірити, гіпотезу про те, що популяції з двох взятих вибірок є однаково розподілені:

$$H_0 : F(x) \equiv G(x).$$

На основі вибірки знаходимо емпіричні функції розподілів цих випадкових змінних

$$F_{n_1}(x) = \begin{cases} 0, & x < x_{(1)}, \\ \dfrac{(n_1)_i}{n_1}, & x_{(i)} \le x < x_{(i+1)}, \quad (i = \overline{1,\ n_1 - 1}), \\ 1, & x \ge x_{n_1} \end{cases} \qquad (1)$$

де $(n_1)_i$ – кількість елементів x_j варіаційного ряду першої вибірки, які не більші за x та:

$$G_{n_2}(x) = \begin{cases} 0, & x < y_{n_2}, \\ \dfrac{(n_2)_i}{n_2}, & y_i \le x < y_{(i+1)}, \ i = \overline{(1, n_2 - 1)}, \\ 1, & x \ge y_{n_2} \end{cases} \qquad (2)$$

де $(n_2)_i$ – кількість елементів y_k варіаційного ряду другої вибірки, які не більші за x.
Розглянемо статистику:

$$D_{n_1 n_2} = \sup_{-\infty<x<+\infty} \left| F_{n_1}(x) - G_{n_2}(x) \right|. \qquad (3)$$

М. В. Смірнов (1933р.) довів, що статистика

$$S_{n_1 n_2} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1 n_2} \qquad (4)$$

має розподіл незалежний від неперервних гіпотетичних розподілів F і G, який при $n_1, n_2 \to \infty$ збігається з розподілом Колмогорова $K(x)$.
Критерій Смірнова стосується випадку великих вибірок $(n_1, n_2 \geq 40)$
Для малих n_1, n_2 $(2 \leq n_1, n_2 \leq 40)$ існують окремі таблиці розподілу Смірнова.
Сформулюємо алгоритм критерію Смірнова:
1. Вибираємо рівень значущості α.
2. Будуємо спільний варіаційний ряд даних двох вибірок.
3. Згідно формул (1)–(4) знаходимо емпіричне значення $(S_{n_1 n_2})_{кр} = (K)_{кр}$ статистики Смірнова.
4. При вибраному рівні значущості α, використовуючи таблицю додатка 11 або з таблиці додатка 12 критичних значень статистики Колмогорова, знаходимо критичне значення $(S_{n_1 n_2})_{кр} = (K)_{кр}$ статистики Смірнова.
5. Якщо $(S_{n_1 n_2})_{емп} < (S_{n_1 n_2})_{кр}$, то висунуту гіпотезу приймаємо, в протилежному випадку – відкидаємо, як хибну.

Приклад: Дано 2 незалежні вибірки незалежних спостережень над абсолютно неперервною випадковою змінною.
(x): 1,9 3,2 0,4 1,2 1,1 1,0 1,5 2,7 0,6 2,0
(y): 0,6 3,0 2,3 2,1 1,2 2,6 1,9 1,0 2,4 2,7
H: Обидві множини спостережень взято у одинаково розподілених абсолютно неперервних популяцій $F \equiv G$
Запишемо для обох рядів спостережень спільний варіаційний ряд значень обох емпіричних функцій розподілу в точках спільного варіаційного ряду та абсолютну різницю в цих точках між розподілами.

$x_{(i)}$	$y_{(i)}$	F_{10}	G_{10}	$\lvert F_{10} - G_{10} \rvert$
0,4		0,1	0	0,1
0,6	0,6	0,2	0,1	0,1
1,0	1,0	0,3	0,2	0,1
1,1		0,4	0,2	0,2
1,2	1,2	0,5	0,3	0,2
1,5		0,6	0,3	0,3
1,9	1,9	0,7	0,4	0,3
2,0		0,8	0,4	0,4 max
	2,1	0,8	0,5	0,3

	2,2	0,8	0,6	0,2
	2,4	0,8	0,7	0,1
	2,6	0,8	0,8	0
2,7	2,7	0,9	0,9	0,1
	3,0	0,9	1	0,1
3,2		1	1	0

$D_{10,10} = 0,4$

Для рівня значущості α=0,05, Sкр=1,36

$$S_{10,10} = \sqrt{\frac{10*10}{10+10}}0,4 = 0,894 < 1,36$$

Гіпотезу приймаємо

Приклад: Дано 2 незалежні вибірки незалежних спостережень над абсолютно неперервною випадковою змінною.

(x): 0,4 -0,5 1,7 0,0 -1,1 1,2 -0,3 -0,9 -0,4 0,5

(y): -0,9 1,1 1,5 0,4 0,8 -0,5 0,6 0,9 -0,3 1,2

H: Обидві множини спостережень взято у однаково розподіленої абсолютно неперервної генеральної сукупності

$x_{(i)}$	$y_{(i)}$	F_{10}	G_{10}	$\lvert F_{10} - G_{10} \rvert$
-1,1		0,1	0	0,1
-0,9	-0,9	0,2	0,1	0,1
-0,5	-0,5	0,3	0,2	0,1
-0,4		0,4	0,2	0,2
-0,3	-0,3	0,5	0,3	0,2
0,0		0,6	0,3	0,3
0,4	0,4	0,7	0,4	0,3
0,5		0,8	0,4	0,4 max
	0,6	0,8	0,5	0,3
	0,8	0,8	0,6	0,2
	0,9	0,8	0,7	0,1
	1,1	0,8	0,8	0

1,2	1,2	0,9	0,9	0
	1,5	0,9	1	0,1
1,7		1	1	0

$D_{10,10} = 0,4$

Для рівня значущості α=0,05, Sкр=1,36

$$S_{10,10} = \sqrt{\frac{10*10}{10+10}}0,4 = 0,894 < 1,36$$

Гіпотезу приймаємо

Критерій порівняння експерименту

Часто потрібно порівняти між собою два різні виробничі методи або два методи оброки.

Означення. Ряд спостережень відносно одного способу обробки назвемо **контрольним**, а відносно іншого способу обробки **рядом спостережень обробки**.

Для вияснення того, що ряд спостережень обробки істотно відрізняється від контрольного ряду спостережень формулюємо нульову гіпотезу (H_o: немає істотної різниці між обома рядами спостережень). Сформульовану гіпотезу перевіряємо за допомогою відповідного критерію.

1. Критерій знаків

Нехай $(x_1, y_1),...,(x_i, y_i),...,(x_n, y_n)$ - ряд незалежних пар незалежних у кожній парі спостережень над деякою абсолютною неперервною статистичною змінною з неперервними функціями розподілу $F_i(x)$ та $G_i(x)$, $(i = \overline{1,n})$.

Задача. Перевірити гіпотезу про те, що розподіли у кожній парі спостережень однакові, тобто

$$H_0: F_i(x) = G_i(x), (i = \overline{1, n}).$$

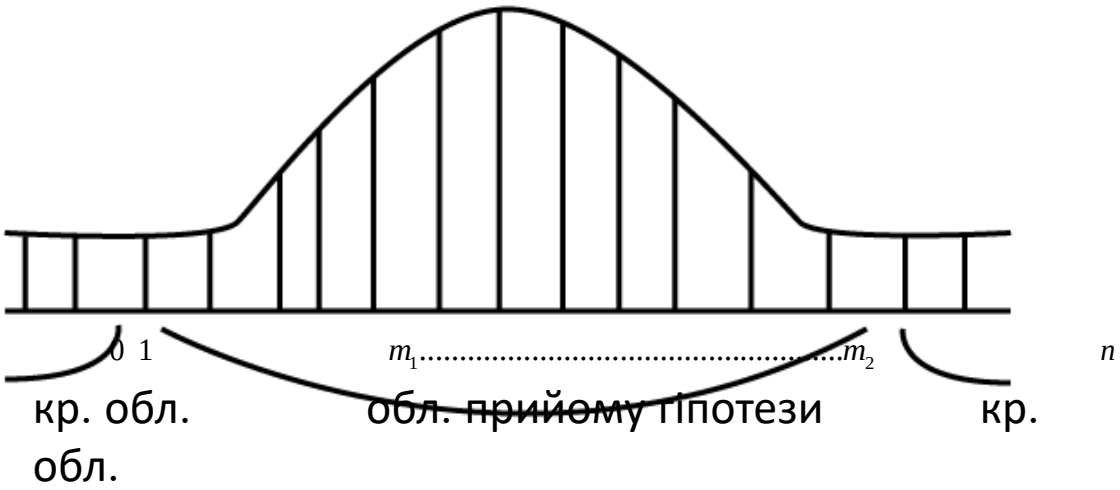
Будемо вважати, що у всіх парах спостережень $x_i \neq y_i$. Тоді, якщо гіпотеза істинна, то різниці $x_i - y_i$ повинні.однаково часто бути додатними та від'ємними,.тобто.ймовірність того, що $\mathrm{P}\{x_i = y_i\} = 0,\ \mathrm{P}\{x_i - y_i > 0\} = \mathrm{P}\{x_i - y_i < 0\} = \frac{1}{2},\quad (i = \overline{1, n})$.

За статистику приймаємо число додатних різниць $x_i - y_i$, яке позначимо через $k(+)$. Очевидно, що статистика $k(+)$ є випадкова змінна і вона біномно розподілена, причому:

$$P\{k(+) = s\} = \frac{C_n^s}{2^n}, (s = 0,1,...,n),\ p = \frac{1}{2}, \text{ то і } q = \frac{1}{2}.$$

На основі цього розподілу визначаємо критичну область гіпотези при заданому рівні значущості α. Оскільки статистика $k(+)$ дискретна, то область прийому гіпотези (m_1, m_2) визначаємо системою нерівностей, де m_1 - найбільше, а m_2 - найменше значення, що задовольняють нерівність

$$\sum_{s=0}^{m_1-1}\frac{C_n^s}{2^n} \le \frac{a}{2},\quad \sum_{s=m_2+1}^{n}\frac{C_n^s}{2^n} \le \frac{a}{2}. \ (*)$$



Наприклад при $n=6$, і $a=0,05$ отримуємо $(m_1,m_2)=(1,5)$. Для різних рівнів значущості та різного числа спостережень табульовано межі області прийому гіпотези.

Наприклад при $a=0,05$ область прийому для гіпотези для різних n визначаються числами m_1 і m_2.

n	m_1	m_2
5	0	5
6	1	5
7	1	6
8	1	7
9	2	8
10	2	8
11	2	9
12	3	9
13	3	10
14	3	11
15	4	11
16	4	12
17	5	12
18	5	13
19	5	14
20	6	14
21	6	15
36	12	24
49	18	31

100	40	60

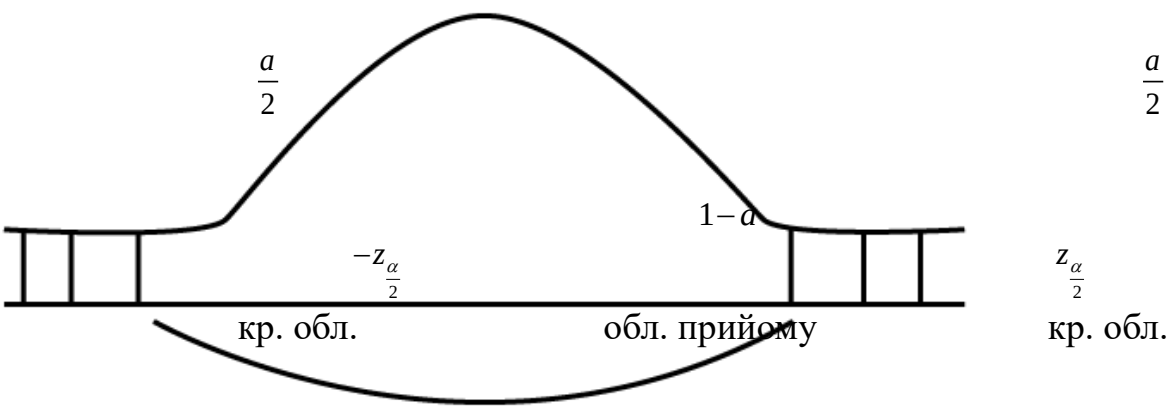
Оскільки біномний розподіл у симетричному випадку досить швидко збігається до нормального розподілу, то при $n \geq 16$ суми (*) досить добре наближаються за допомогою інтегральної теореми Муавра-Лапласа.

Наприклад : при $a = 0,5$ критична область визначається із співвідношення

$$P\left\{ \left| \frac{\mu - \frac{n}{2}}{\frac{\sqrt{n}}{2}} \right| < z_{\frac{\alpha}{2}} \right\} = 1 - a \quad \text{, звідси}$$

$$P\{\frac{n}{2} - 0,98\sqrt{n} < \mu < \frac{n}{2} + 0,98\sqrt{n}\} = 0,95,$$

$z_{\frac{\alpha}{2}}$ - табульована, наприклад при $a = 0,05$, $z_{\frac{\alpha}{2}} = 1,96$.



Отже область прийому гіпотез визначається числами

$m_1 = [\frac{n}{2} - 0,98\sqrt{n}]$ - ціла частина числа,

$m_2 = \{\frac{n}{2} + 0,98\sqrt{n}\}$ - найменше число з доповненням до цілого числа.

Наприклад. при $n = 16$, тоді

$m_1 = [8 - 0,98 * 4] = [4,08] = 4$

$m_2 = \{8 + 0,98 * 4\} = \{8 + 3,92\} = \{11,92\} = 12$

$(m_1, m_2) = (4,12).$

Зауваження. У результаті обмеженої точності вимірювань може трапитись, що деякі пари мають одинакові елементи. Тоді такі пари пропускаємо. Число спостережень при цьому зменшується на число пропущених пар.

Якщо $k(+)$ емпіричне менше ніж m_1 або більше m_2, то гіпотезу відкидаємо.

Приклад. Дано 21 пару незалежних у кожній парі спостережень над абсолютно неперервною статистичною змінною.

(2,4; 1,9) +	(1,4; 1,9) -	(1,4; 1,9) -
(1,2; 0,5) +	(0,8; 1,1) -	(0,9; 2,1) -
(2,2; 1,2) +	(2,0; 3,2) -	(1,1; 3,0) -
(1,4; 1,3) -	(2,7; 2,6) +	(1,2; 2,1) -
(2,3; 2,6) -	(1,9; 2,7) -	(0,3; 2,3) -
(3,3; 1,2) +	(0,6; 2,4) -	(3,1; 2,9) +
(3,4; 2,3) +	(1,5; 0,9) +	(1,9; 0,6) +

Перевірити гіпотезу про те, що елементи кожної пари однаково неперервно розподілені $F_i = G_i,\ (i = \overline{1,21})$.

Для доведення гіпотези використовується перевірка знаків. Число додатних різниць дорівнює 9: $k(+) = 9$. З таблиці Т-9 відчитуємо межі області прийому гіпотези при рівні значущості $a = 0,05$ і обсязі спостереження $n = 21$. Дістаємо $(m_1, m_2) = (6,15)$. Оскільки число додатних знаків різниць попадає в область прийому гіпотези, то гіпотеза не суперечить статистичним даним. Нормальне наближення дає ті самі межі.

2. Гіпотеза про медіану

Нехай $x_1,...,x_i,...,x_n$ - ряд незалежних спостережень над абсолютною неперервною статистичною змінною. Потрібно перевірити гіпотезу про те, що медіана популяції, з якої взято вибірку дорівнює $a : H_0 : M_e = a$. Якщо гіпотеза істинна, то $\mathrm{P}\{x_i = M_e\} = 0,\quad \mathrm{P}\{x_i - M_e > 0\} = \mathrm{P}\{x_i - M_e < 0\} = \frac{1}{2}$.

Отже, за критерій можна прийняти статистику $k(+)$ - число додатних різниць $(x_i - M_e)$, яка є біномно розподіленою випадковою змінною з параметром n та

$p = \frac{1}{2}$, тому визначення меж області прийому гіпотези H_0, та висновок про її істинність здійснюється як у пункті **1**.

3. Критерій інверсій (Wilcoxon, 1945)

Нехай $x_1,...,x_i,...,x_n$ - вибірка з абсолютної неперервної популяції, що має неперервну функцію розподілу $F(x)$, а $y_1,...,y_m$ - вибірка з абсолютно неперервної популяції з функцією розподілу $G(x)$.

Задача. Перевірити гіпотезу про те, що популяції з яких взято вибірки однаково розподілені. Нехай нульова гіпотеза H_0, полягає в тому, що $F(x) = G(x)$, тобто, що ці генеральні сукупності є стохастично еквівалентними випадковими змінними. (Порівняння за критерієм Смирнова)

Якщо гіпотеза істинна, то в середньому на кожному відрізку із заданою пропорцією *x*-ів ми повинні отримати приблизно таку ж пропорцію *y*- ів. Якщо на якомусь відрізку із заданою пропорцією *x*-ів пропорція *y*-ів дуже мала, або дуже велика, то цей факт буде свідчити проти гіпотези. Тому, за статистику приймаємо число інверсій *x*-ів відносно *y*-ів у спільному варіаційному ряді, яке позначимо через $W(x/y)$. (Кількість інверсій елементів першої вибірки відносно елементів другої вибірки у спільному варіаційному ряді).

Аналогічно, $W(y/x)$ - число інверсій *y*-ів відносно *x*-ів.

Число інверсій для даного x_i визначається як число тих y_k, що задовольняють нерівність $y_k < x_i$; число інверсій для даного y_i дорівнює числу тих x_k, що задовольняють нерівність $x_k < y_i$.

$$W(y/x) = \sum_{i=1}^{n} (\text{число } y_k < x_i),\ k = \overline{1,m},$$

$$W(x/y) = \sum_{i=1}^{m} (\text{число } x_k < y_i),\ k = \overline{1,n}.$$

Наприклад, якщо порядок розміщення елементів у спільному варіаційному ряді елементів першої та другої вибірки є таким

$$x_9 < x_5 < y_4 < x_1 < y_7 < x_8 < y_5 < x_4 < x_6 < x_7 < y_6 < y_1 < x_3 < x_2 < y_2 < y_3;\ x_i < y_k,$$

то число інверсій *y*-ів відносно *x*-ів (число тих *y*-ів, що стоять перед *x*-ми) дорівнює

$$W(y/x) = 1+2+3+3+3+3+5+5 = 22,\ y \le x.$$

Аналогічно, число інверсій *x*-ів відносно *y*-ів

$$W(x/y) = 2+3+4+7+7+9+9 = 41,\ x \le y.$$

Очевидно, що статистика $W(y/x)$ і $W(x/y)$ може приймати цілочисельні значення від 0 (коли всі елементи першої вибірки розміщені перед усіма елементами другої вибірки) до mn (коли всі елементи другої вибірки стоять перед всіма елементами першої вибірки), тобто $0 \le W(y/x) \le mn$ і $0 \le W(x/y) \le mn$. Неважко перевірити, що

$$W(y/x) + W(x/y) = mn.$$

Це служить контролем правильності обчислення: $22+41 = 9*7 = 63$.

Статистики $W(y/x)$ і $W(x/y)$ виступають симетрично, тому досить обмежитись однією з них, яку позначимо через *W*. Статистику *W* ввів у 1945 р. Вілкоксон (*Wilcoxon*) і тому критерій інверсій часто називають критерієм Вілкоксона.

На основі цього розподілу визначаємо критичні області для гіпотези, подібно ж як у випадку критерію знаків. Зазначимо, що при $m \ge 4$ і $n \ge 4$, але так, що $m+n \ge 20$ розподіл статистики Вілкоксона досить добре наближається нормальним розподілом з параметрами a та σ^2 (тобто зі математичним сподіванням та дисперсією), відповідно

$$a = E(W) = \frac{mn}{2}, \qquad \sigma^2 = D(W) = \frac{mn}{12}(m+n+1).$$

Тому у таких випадках, при рівні значущості $\alpha = 0,05$, з високим ступенем точності можна вважати, що область прийому гіпотези розташована в інтервалі $(a - 1,96\sigma\ ;\ a + 1,96\sigma)$.

Критичні значення статистики *W* для малих m і n табульовано при різних рівнях значущості α. Якщо емпіричне значення статистики Вілкоксона попадає в область прийому гіпотези, то кажемо, що гіпотеза не суперечить експериментальним даним.

Приклад. Дано 2 незалежні вибірки незалежних спостережень над двома неперервними популяціями: одна вибірка обсягом $m = 10$, а друга - $n = 15$, відповідно

$n = 15$

$(Y)\ y_1$	2,2	y_9	2,7
y_2	4,0	y_{10}	1,6
y_3	1,4	y_{11}	3,4

y_4 2,9 y_{12} 2,5

y_5 2,3 y_{13} 0,6(1)

y_6 1,9 y_{14} 2,6

y_7 1,4 y_{15} 0,9(3)

y_8 2,9

$m = 10$

$(X)x_1$ 2,3

x_2 1,7

x_3 2,4

x_4 2,7

x_5 0,8(2)

x_6 1,2(5)

x_7 1,5(7)

x_8 1,7

x_9 0,9(4)

x_{10} 2,6

Перевірити H про те, що: популяції, з яких взято вибірки однаково розподілені: $F(x) \equiv G(x)$. Позначимо елементи першої вибірки через $x_1, ..., x_{10}$, а другої – $y_1, ..., y_{15}$. Тоді, спільний варіаційний ряд буде таким:

$y_{13}x_5y_{15}x_9x_6y_3x_7yxxyyxyyxyyyxxyyyy$

Як видно число інверсій y -ів відносно x -ів рівне.

$$W(y/x) = 1+2+2+3+4+4+6+8+11+11 = 52$$

Оскільки $a = \frac{10*15}{2} = 75$

$$\sigma^2 = \frac{10*15}{12}(10+15+1) = 325 \quad \sigma = \sqrt{325} = 18,027,$$

То область прийому гіпотези при п’ятипроцентному рівні значущості буде

$$(a - 1,96\sigma; a + 1,96\sigma) = (39,66; 110,34)$$

Отже, гіпотеза про однаковий неперервний розподіл обох популяцій, з яких взято вибірки не суперечить вибірковим даним H прийнято.

Зауваження. Якщо при застосуванні критерію знаків зустрічаються пари з однаковими елементами, то такі при пропускають, а об'єм вибірки зменшується на число пропущених пар.

Трифакторний варіансний аналіз

Нехай дано вибірку $x_1, x_2, ..., x_n$ незалежних спостережень над деякою одновимірною нормально розподіленою мінливою величиною, елементи якої можна згрупувати за рівнями трьох виділених факторів A, B та C: на m груп за ознакою A, на n груп за ознакою B і на l груп за ознакою C. Отримаємо mnl класифікаційних груп. Припустимо, що в кожній групі є тільки одне спостереження. Позначимо через x_{ijk} – спостереження. в i-тій групі за ознакою A, в j-тій групі за ознакою B та в k-тій групі за ознакою C. Всі mnl спостережень можна розмістити в l таблицях вигляду двофакторного варіансного аналізу (mn). У кожній з l – таблиць третій індекс k сталий, $(k = 1, 2, ..., l)$. Перший індекс – індекс довготи, другий – широти, третій – глибини . Введемо середні відповідних класифікаційних підгруп:

$$x_{ij.};\ x_{i.k}\ ;\ x_{.jk}\ ;\ x_{i..}\ ;\ x_{.j.}\ ;\ x_{..k}\ ;\ x_{...} \qquad (*)$$

де наприклад:

$$x_{ij.} = \frac{1}{l}\sum_{k=1}^{l} x_{ijk}, (i = \overline{1,m},\ j = \overline{1,n}), \qquad x_{i..} = \frac{1}{nl}\sum_{j=1}^{n}\sum_{k=1}^{l} x_{ijk}, (i = \overline{1,m}), \qquad x_{...} = \frac{1}{mnl}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l} x_{ijk}.$$

Спостереження x_{ijk} та середні $(*)$ пов'язані алгебраїчною тотожністю

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{ijk} - x_{...})^2 = nl\sum_{i=1}^{m}(x_{i..} - x_{...})^2 + ml\sum_{j=1}^{n}(x_{.j.} - x_{...})^2 +$$

$$+mn\sum_{k=1}^{l}(x_{..k} - x_{...})^2 + l\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij.} - x_{i..} - x_{.j.} + x_{...})^2 +$$

$$+n\sum_{i=1}^{m}\sum_{k=1}^{l}(x_{i.k} - x_{i..} - x_{..k} + x_{...})^2 + m\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{.jk} - x_{.j.} - x_{..k} + x_{...})^2 + \qquad (**)$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{ijk} - x_{ij.} - x_{i.k} - x_{.jk} + x_{i..} + x_{.j.} + x_{..k} - x_{...})^2.$$

Тотожність $(**)$ випливає з тотожності

$$(x_{ijk} - x_{...}) = (x_{i..} - x_{...}) + (x_{.j.} - x_{...}) + (x_{..k} - x_{...}) + (x_{ij.} - x_{i..} - x_{.j.} + x_{...}) + (x_{i.k} - x_{i..} - x_{..k} +$$

$$+x_{...}) + (x_{.jk} - x_{.j.} - x_{..k} + x_{...}) + (x_{ijk} - x_{ij.} - x_{i.k} - x_{.jk} + x_{i..} + x_{.j.} + x_{..k} - x_{...})$$

і того, що сума відхилень вибіркових даних від їх середнього арифметичного у кожній відповідній групі дорівнює нулю. Співвідношення $(**)$ вказує на те, що повна девіація (ліва частина рівності $(**)$) розкладається на сім девіацій (справа рівності $(**)$): перші три характеризують відповідно мінливість між групами ознак A, B, C, наступні три оцінюють взаємодію добутків AB, AC, BC (інтеракції) відповідно, остання виражає залишкову мінливість, яку можна вважати взаємодією другого порядку, ABC. Тотальну мінливість назвемо мінливістю нульового порядку,$[\ ]$; мінливість між групами – мінливістю першого порядку $[A, B, C]$; мінливість взаємодії – мінливістю другого порядку $[AB, AC, BC]$; залишкова мінливість – мінливістю третього порядку $[ABC]$. У тотожності $(**)$ сума зліва має $mnl - 1$ ступенів вільності. Для семи сум справа маємо відповідно таке число ступенів вільності:

$$d.f. = m - 1, \qquad d.f. = n - 1, \quad d.f. = l - 1;$$

$$d.f. = mn - m - n + 1 = (m-1)(n-1);$$

$$d.f. = ml - m - l + 1 = (m-1)(l-1);$$

$$d.f. = nl - n - l + 1 = (n-1)(l-1);$$

$$d.f. = mnl - mn - ml + m + n + l + 1 = (m-1)(n-1)(l-1).$$

Число ступенів вільності девіацій, що виступають у тотожності (**) утворюють наступну тотожність:

$$mnl-1=(m-1)+(n-1)+(l-1)+(m-1)(n-1)+(m-1)(l-1)+$$
$$+(n-1)(l-1)+(m-1)(n-1)(l-1).$$

Зліва даної тотожності маємо число ступенів вільності нульового порядку; справа – три сукупності ступенів вільності першого, другого і третього порядків відповідно.

Якщо поділити тотожність (**) на $mnl-1$, отримаємо, що повна варіанса є лінійною комбінацією варіанс між групами, інтеракцій (взаємодії) та залишкової варіанси:

$$S^2=\frac{m-1}{mnl-1}S_A^2+\frac{n-1}{mnl-1}S_B^2+\frac{l-1}{mnl-1}S_C^2+\frac{(m-1)(n-1)}{mnl-1}S_{AB}^2+$$
$$+\frac{(m-1)(l-1)}{mnl-1}S_{AC}^2+\frac{(n-1)(l-1)}{mnl-1}S_{BC}^2+\frac{(m-1)(n-1)(l-1)}{mnl-1}S_{ABC}^2.$$

Тут наприклад,

$$S_A^2=nl\frac{1}{m-1}\sum_{i=1}^{m}(x_{i..}-x_{...})^2\ ,$$

$$S_{AB}^2=l\frac{1}{(m-1)(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij.}-x_{i..}-x_{.j.}+x_{...})^2,$$

$$S_{ABC}^2=S_r^2=\frac{1}{(m-1)(n-1)(l-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{ijk}-x_{ij.}-x_{i.k}-x_{.jk}+x_{i..}+x_{.j.}+x_{..k}-x_{...})^2\ .$$

Аналогічно записують інші варіанси: S_B^2 , S_C^2 , S_{AC}^2 , S_{BC}^2. Перелічені варіанси можна використати при доведенні гіпотези однорідності за допомогою критерію Фішера, для чого порівняємо варіанси між групами або варіанси взаємодій із залишковою. Отже, при доведенні відповідних гіпотез розглядаємо відповідні статистики. Зокрема, для аналізу можливого впливу факторів *A, B,C* на генеральну сукупність використовуємо статистики:

$$F_A=\frac{S_A^2}{S_r^2},\quad F_B=\frac{S_B^2}{S_r^2},\quad F_C=\frac{S_C^2}{S_r^2},\quad (***)$$

відповідно для аналізу сумісного впливу факторів *AB, AC, BC* використовуємо такі статистики

$$F_{AB}=\frac{S_{AB}^2}{S_r^2},\qquad F_{AC}=\frac{S_{AC}^2}{S_r^2},\quad F_{BC}=\frac{S_{BC}^2}{S_r^2}.\qquad (****)$$

Якщо гіпотеза про однорідність даних в нормальній популяції – вірна, то статистики (***) та (*****) мають розподіл Фішера відповідно до чисел ступеня вільності:

$$d.f.=((m-1,(m-1)(n-1)(l-1)),\qquad d.f.=((n-1),(m-1)(n-1)(l-1)),$$
$$d.f.=((l-1),(m-1)(n-1)(l-1))$$

для виразу (***) та

$$d.f.=((m-1)(n-1),(m-1)(n-1)(l-1)),\quad d.f.=((m-1)(n-1),(m-1)(n-1)(l-1)),$$
$$d.f.=((n-1)(l-1),(m-1)(n-1)(l-1)).$$

для виразу (****) відповідно.

Обчислення при статистичному доведенні однорідності зручно розмістити в таблиці трифакторного варіансного аналізу при одному спостереженні в кожній групі:

сть	Девіація	*d.f.*	Варіанса
лами *A*	$nl\sum_{i=1}^{m}(x_{i..}-x_{...})^2$	$m-1$	S_A^2
лами *B*	$ml\sum_{j=1}^{n}(x_{.j.}-x_{...})^2$	$n-1$	S_B^2

лами C	$mn\sum_{k=1}^{l}(x_{..k}-x_{...})^2$	$l-1$	S_C^2
ія AB	$l\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij.}-x_{i..}-x_{.j.}+x_{...})^2$	$(m-1)(n-1)$	S_{AB}^2
ія AC	$n\sum_{i=1}^{m}\sum_{k=1}^{l}(x_{i.k}-x_{i..}-x_{..k}+x_{...})^2$	$(m-1)(l-1)$	S_{AC}^2
ія BC	$m\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{.jk}-x_{.j.}-x_{..k}+x_{...})^2$	$(n-1)(l-1)$	S_{BC}^2
ове	$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{ijk}-x_{ij.}-x_{i.k}-x_{.jk}+x_{i..}+x_{.j.}+x_{..k}-x_{...})^2$	$(m-1)(n-1)(l-1)$	S_r^2
	$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{ijk}-x_{...})^2$	$mnl-1$	

Приклад. Перевірити дію 2 добрив на 5 сортів якогось збіжжя. Проводимо експеримент у $4-x$ дослідних центрах. Дістанемо 40 даних про врожаї, які класифікуємо в таблиці $4\times5\times2$ (чотири центри, п'ять сортів збіжжя і два типи добрив). Нехай врожаї, виміряні від якоїсь зручної вихідної вартості та виражені в означених одиницях, будуть такі, як у таблиці 1, де T_1 і T_2 - два типи добрив.

Таблиця 1

Дослідницькі центри	Сорти збіжжя											
	1		2		3		4		5		Разом	
	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
1	-6	-4	-4	-2	-10	-7	-3	-5	1	0	-22	-18
2	-2	-1	-5	-1	-3	-4	-4	-1	-2	1	-16	-6
3	3	2	-2	3	-4	0	4	1	3	3	4	9
4	3	6	3	2	6	3	-1	5	6	8	17	24
Разом	-2	3	-8	2	-11	-8	-4	0	8	12	-17	3

З умови задачі випливає, що дослідницькі центри – це фактор A, сорти збіжжя – фактор B, добрива – фактор C, генеральна сукупність – врожайність збіжжя.
Тому $m=4,\ n=5,\quad l=2;\quad x_{...}=-0,2$. Одно- і двоіндексні середні розмістимо в таблиці 2.

Таблиця 2

ення ного ксу	$x_{i..}$	$x_{.j.}$	$x_{..k}$	$x_{3j.}$	$x_{2j.}$	$x_{3j.}$	$x_{4j.}$	$x_{i.1}$	$x_{i.2}$	$x_{.j1}$	$x_{.j1}$
	-4,0	0,125	-0,85	-5,0	-1,5	2,5	4,5	-4,4	-3,6	-0,50	0,75
	-2,2	-0,75	0,45	-3,0	-3,0	0,5	2,5	-3,2	-1,2	-2,00	0,5
	1,3	-2,375		-8,5	-3,5	-2,0	4,5	0,8	1,8	-2,75	-2,00
	4,1	-0,5 2,5		-4,0 0,5	-2,5 - 0,5	2,5 3,0	2,0 7,0	3,4	4,8	-1,00 2,00	0,00 3,00

Девіації, ступені вільності та варіанси для різних мінливостей розмістимо в таблиці 3.

Таблиця 3

Мінливість	Девіація	$d.f.$	Варіанса
Між центрами А	381,8	3	130,60
Між сортами В	100,15	4	25,04
Між добривами Т	16,90	1	16,9
Інтеракція(взаємодія) АВ	62,45	12	5,20
Інтеракція(взаємодія) АТ	2,10	3	0,70
Інтеракція(взаємодія) ВТ	3,85	4	0,96
Залишкове	61,15	12	5,10
Повна	638,40	39	-

За допомогою варіансного аналізу перевірити гіпотезу про можливий вплив вказаних факторів на врожайність збіжжя. На основі даних таблиці 3 обчислимо емпіричні значення статистик Фішера для ітеракцій із залишковою варіансою при рівні значущості $\alpha = 0,05$:

$$(F_{AB})_{емп} = \frac{S_{AB}^2}{S_r^2} = \frac{5,20}{5,10} = 1,02; \quad (F_{AC})_{емп} = \frac{S_r^2}{S_{AC}^2} = \frac{5,10}{0,70} = 7,29;$$

$$(F_{BC})_{емп} = \frac{S_r^2}{S_{BC}^2} = \frac{5,10}{0,36} = 5,32.$$

При рівні значущості $\alpha = 0,05$ та кількостях ступенів вільності (12,12), 12,3), (12,4), відповідно, наприклад з таблиці (додаток 8) маємо:

$(F_{AB})_{кр} = 2,69;$ $(F_{AC})_{кр} = 8,74;$ $(F_{BC})_{кр} = 5,91.$

Усі три емпіричні значення статистик інтеракцій менші відповідних критичних значень цих статистик. Отже, всі ці фактори діють незалежно. Дослідимо чи впливають фактори *A, B, C* на генеральну сукупність. Для цього вобчислимо емпіричні значення статистики Фішера для цих факторів:

$$(F_A)_{емп} = \frac{130,60}{5,10} = 25,4; \quad (F_B)_{емп} = \frac{25,04}{5,10} = 4,93; \quad (F_C)_{емп} = \frac{16,30}{5,10} = 3,34.$$

При кількості ступенів вільності (3,12), (4,12) , (1,12), відповідно, з таблиць (додаток 8) маємо:

якщо $\alpha = 0,05$, то

$(F_A)_{кр} = 3,49;$ $(F_B)_{кр} = 3,26;$ $(F_C)_{кр} = 4,75;$

якщо $\alpha = 0,01$, то

$(F_A)_{кр} = 5,95;$ $(F_B)_{кр} = 5,41;$ $(F_C)_{кр} = 9,43.$

Отже різниці між центрами істотні; між сортами – неістотні при рівні значущості 1%, і істотні при рівні значущості 5%, між добривами неістотні. Отже, мінливість врожаїв залежить від мінливості між центрами (і можливо між сортами), чого не можна сказати без деяких досліджень про залежність її від мінливості між добривами.

Латинський квадрат

Означення. Латинським квадратом порядку m називають таке розміщення m різних елементів, кожен з яких повторений m разів у m рядках і m стовпчиках квадрату, при якому кожний елемент зустрічається точно один раз у кожному рядку і кожному стовпчику.

Приклад трьох латинських квадратів 5 порядку

1	2	3	4	5						A	B	C	D	E
2	3	4	5	1						B	C	D	E	A
3	4	5	1	2						C	D	E	A	B
4	5	1	2	3	α	β	γ	δ	ε	D	E	A	B	C
5	1	2	3	4	β	γ	δ	ε	α	E	A	B	C	D
					γ	δ	ε	α	β					
					δ	ε	α	β	γ					
					ε	α	β	γ	δ					

Латинський квадрат називають **стандартним**, якщо у першому рядку і в першому стовпчику символи виступають у загальноприйнятому порядку.

Два стандартні квадрати називають **спряженими**, якщо рядки одного є стовпчиками другого.

Латинський квадрат називають **симетричним**, якщо від зміни рядків на стовпчики він не зміниться. Перший з квадратів стандартний.

Два латинські квадрати називаються **ортогональними** якщо при накладанні кожний символ одного квадрата зустрічається з кожним символом другого квадрата точно один раз.

Другий і третій квадрати **ортогональні.**

Два ортогональні латинські квадрати один з яких заданий латинськими літерами, а другий грецькими, називається греко- латинським квадратом. Греко-латинський квадрат можна записати одним квадратом

$A\alpha$	$B\beta$	$C\gamma$	$D\delta$	$E\xi$
$C\delta$	$D\xi$	$E\alpha$	$A\beta$	$B\gamma$
$E\beta$	$A\gamma$	$B\delta$	$C\xi$	$D\alpha$
$B\xi$	$C\alpha$	$D\beta$	$E\gamma$	$A\delta$
$D\gamma$	$E\delta$	$A\xi$	$B\alpha$	$C\beta$

Число латинських квадратів дуже швидко зростає з порядком m . Це видно з наступної таблиці

Порядок **латинського квадрату** m	Число ортогональних латинських квадратів	Число стандартних латинських квадратів S	Число латинських квадратів з одного стандартного $m!(m-1)!$	число всіх латинських квадратів $Sm!(m-1)!$
2	0	1	2	2
3	2	1	12	12
4	3	4	144	576
5	4	56	2 880	161 280
6	0	94 08	86 400	812 850 200

7	6	16 942 080	3 628 800	61 479 419 904 000

Випадковий експеримент за планом латинського квадрату

Нехай деяка мінлива величина поділяється на m груп за кожною з трьох ознак: A, B, C. Отримаємо m^3 класифікаційний підгруп.

Припустимо, що проводиться по-одному спостереженні в m^2 класифікаційних підгрупах. Ці спостереження проводимо за планом навмання вибраному латинського квадрату порядку m.

Позначимо через x_{ijk} спостереження в i-ій групі за ознакою A, в j-ій групі за ознакою B, і в k-ій групі за ознакою C.

Ці m^2 спостережень розташуємо в m рядках і m стовпчиках, навмання вибраного латинського квадрату. Рядки характеризують групи ознаки A, стовпчики-групи ознаки B, а символи латинського квадрату характеризують групи ознаки C. Наприклад, якщо випадковий експеримент проводиться за схемою латинського квадрату, то спостереження записуємо так:

M	H	S	L	M	H	S	L
S	L	M	H	H	S	L	M
L	M	H	S	L	M	H	S
H	S	L	M	S	L	M	H
x_{113}	x_{121}	x_{134}	x_{142}	x_{113}	x_{121}	x_{134}	x_{142}
x_{214}	x_{222}	x_{233}	x_{241}	x_{211}	x_{224}	x_{232}	x_{243}
x_{312}	x_{323}	x_{331}	x_{344}	x_{312}	x_{323}	x_{331}	x_{344}
x_{411}	x_{424}	x_{432}	x_{443}	x_{414}	x_{422}	x_{433}	x_{441}

Упорядковуємо за алфавітом

H	L	M	S
1	2	3	4

Позначимо через $x_{i..}$ середнє арифметичне i-го рядка:

$$x_{i..} = \frac{1}{m}\sum_{j=1}^{m} x_{ijk}, \quad i = \overline{1,m},$$

через $x_{.j.}$ середнє арифметичне j-го стовпчика утвореної матриці:

$$x_{.j.} = \frac{1}{m}\sum_{i=1}^{m} x_{ijk,}, \quad j = \overline{1,m},$$

середнє арифметичне тих елементів вибірки, які мають k-й прояв ознаки C:

наприклад $x_{..1} = \dfrac{x_{121} + x_{241} + x_{331} + x_{411}}{4}$

Через $x_{...}$ позначимо середнє арифметичне всіх спостережень

$$x_{...} = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ijk}.$$

Повна мінливість всіх спостережень виражається рівністю

$$\sum_{i=1}^{m}\sum_{j=1}^{m}(x_{ijk} - x_{...})^2 =$$

(яку можна представити як суму таких чотирьох девіацій)

$$= m\sum_{i=1}^{m}(x_{i..} - x_{...})^2 + m\sum_{j=1}^{m}(x_{.j.} - x_{...})^2 + m\sum_{k=1}^{m}(x_{..k} - x_{...})^2 +$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{m}(x_{i..} - x_{i..} - x_{.j.} - x_{..k} + 2x_{...})^2 . \qquad (1)$$

Доведення тотожності (1) грунтується на тому, що

$$\left(x_{ijk} - x_{...}\right)^2 = [\left(x_{ijk} - x_{i..} - x_{.j.} - x_{..k} + 2x_{...}\right) + \left(x_{i..} - x_{...}\right) + \left(x_{.j.} - x_{...}\right) + \left(x_{..k} - x_{...}\right)]^2 .$$

та тому, що сума відхилень вибіркових значень у групах від середнього арифметичного відповідної групи дорівнює нулю. Зауважимо, що для обчислення $x_{..k}$ потрібно вибрати з утвореної таблиці всі ті значення, які мають однакове значення k третього індекса (мають k-ий прояв ознаки C), додати їх і розділити на m.

Таким чином тотожність (1) вказує на те, що повна девіація розкладається на чотири девіації: девіації між групами ознак A, B, C і залишкової девіації. Аналіз квадратичних форм у тотожності (1) показує, що кількість ступенів вільності в ній така: лівої частини $d.f. = m^2 - 1$, кожного з трьох перших членів правої частини $d.f. = m - 1$, останього – $d.f. = m^2 - 3m + 2 = (m-1)(m-2)$.

Отже, ступені вільності доданків тотожності (1) самі утворюють тотожність:

$$m^2 - 1 = 3(m-1) + (m-1)(m-2)$$

Розділивши тотожність (1) на $m^2 - 1$ отримаємо вираз повної варіанси у вигляді лінійної комбінації варіанс між групами ознак A, B, C, відповідно, та залишкової варіанси:

$$S^2 = \alpha_1 S_A^2 + \alpha_2 S_B^2 + \alpha_3 S_C^2 + \alpha_4 S_r^2 .$$

Тут

$$\alpha_1 = \alpha_2 = \alpha_3 = \frac{m-1}{m^2-1}, \qquad \alpha_4 = \frac{(m-1)(m-2)}{(m^2-1)};$$

$$S^2 = \frac{1}{m^2-1}\sum_{i=1}^{m}\sum_{j=1}^{m}(x_{ijk} - x_{...})^2, \quad S_A^2 = m\frac{1}{m-1}\sum_{i=1}^{m}(x_{i..} - x_{...})^2 ,$$

$$S_B^2 = m\frac{1}{m-1}\sum_{j=1}^{m}(x_{.j.} - x_{...})^2 , \quad S_C^2 = m\frac{1}{m-1}\sum_{k=1}^{m}(x_{..k} - x_{...})^2 ,$$

$$S_r^2 = \frac{1}{(m-1)(m-2)}\sum_{i=1}^{m}\sum_{j=1}^{m}(x_{ijk} - x_{i..} - x_{.j.} - x_{..k} + 2x_{...})^2 .$$

Чотири останні варіанси є незміщеними і незалежними оцінками дисперсій нормальної генеральної сукупності. Тому для перевірки гіпотези однорідності про вплив факторів на генеральну сукупність можна застосувати відповідні статистики Фішера:

$$F_A = \frac{S_A^2}{S_r^2} , \qquad F_B = \frac{S_B^2}{S_r^2} , \qquad F_C = \frac{S_C^2}{S_r^2} .$$

Якщо гіпотеза про однорідність даних у класифікаційних групах істинна, тобто віповідний фактор не впливає на генеральну сукупність, то відповідна з цих статистик має розподіл Фішера з кількістю ступенів вільності $d.f. = ((m-1), (m-1)(m-2))$. При обчисленнях проміжні результати зручно розмістити у таблиці.

Мінливість	Девіація	d. f.	Варіанса

Між групами ознаки A	$m\sum_{i=1}^{m}\left(x_{i..}-x_{...}\right)^2$	$m-1$	S_A^2
Між групами ознаки B	$m\sum_{j=1}^{m}\left(x_{.j.}-x_{...}\right)^2$	$m-1$	S_B^2
Між групами ознаки C	$m\sum_{k=1}^{m}\left(x_{..k}-x_{...}\right)^2$	$m-1$	S_C^2
Залишкова	$\sum_{i=1}^{m}\sum_{j=1}^{m}(x_{ijk}-x_{...}-x_{.j.}-x_{..k}+2x_{...})^2$	$(m-1)(m-2)$	S_r^2
Повна	$\sum_{i=1}^{m}\sum_{j=1}^{m}\left(x_{ijk}-x_{...}\right)^2$	m^2-1	-

Зазначимо, що варіансний аналіз за планом латинського квадрату є неповним трифакторним варіансним аналізом. Він вимагає в *m* разів менше спостережень, ніж у повному трифакторному варіансному аналізі і значно менше обчислень. Однак, при цьому всі три фактори повинні мати однакове число в *m* рівнів, та й при наявності взаємодій між факторами він може бути помилкови

Приклад:

Для оцінки впливу добрив на урожай ячменю статистична лабораторія Ротамстедської дослідної станції в Англії(1929 рік, керівник Рональд Фішер). провела експеримент методом латинського квадрата. При цьому п'ять сортів добрив C, M, N, S, U (фактор T) були внесені на 25 ділянок землі за такою схемою:

N	U	M	S	C
S	C	N	M	U
M	S	U	C	N
C	N	S	U	M
U	M	C	N	S

У результаті цього експерименту отримали такі врожаї ячменю (в чвертях фунта на 1/40 акра) [22].

1 акр =40 арів=4046,86м2

Варіансним аналізом за схемою латинського квадратуа, вибраного в цьому експерименті, перевірити гіпотезу про вплив добрив на урожай ячменю.

Занесемо результати всіх проміжних обчислень у типову таблицю 2.

Таблиця 2

Мінливість	**Девіація**	**d.f.**	**Варіанса**
Між рядками (А)	5028,8	4	1257,20
Між стовпчиками (В)	5954,8	4	1488,70
Між добривами (Т)	2452,0	12	613,00
Залишкова	1306,4		108,87
Повна	14742,0	24	-

Обчислимо емпіричне значення статистики Фішера фактора (T):

$$(F)_{емп} = \frac{S_T^2}{S_r^2} = \frac{613,00}{108,87} = 5,84.$$

Виберемо рівень значущості $\alpha = 0,10$. Тоді при кількості ступенів вільності $d.f. = (4,12)$ з таблиці (додаток 8) маємо $(F)_{кр} = 3,25$.

Оскільки $F_{емп.} > F_{кр.}$ то гіпотезу про те, що фактор T на генеральну сукупність не впливає, потрібно відхилити, тобто наведені результати експериментів дають підставу стверджувати при $\alpha = 0,05$, що ці добрива суттєво впливають на урожайність ячменю.

Кореляційний аналіз

1. Коваріація

Розглянемо довільний двовимірний випадковий вектор з компонентами ξ, η, для яких відомі їх сподівання і дисперсії: $E(\xi)$, $E(\eta)$, $D(\xi)$, $D(\eta)$: і нехай a деяка стала, $a > 0$.

Задача. Знайти зв'язок між компонентам ξ та η двовимірного випадкового вектора?

З цією метою утворимо з координат цього вектора випадкову змінну

$$\zeta = -a\xi + \eta.$$

Знайдемо її дисперсію: за означенням дисперсії та властивостями сподівання маємо:

$$D(\zeta) = D(-a\xi + \eta) = E[(-a\xi + \eta) - E(-a\xi + \eta)]^2 =$$
$$= E[-a(\xi - E\xi) + (\eta - E\eta)]^2 =$$
$$= a^2 E(\xi - E\xi)^2 - 2aE\left[(\xi - E\xi)(\eta - E\eta)\right] + E\left(\eta - E\eta\right)^2 =$$
$$= a^2 D\xi - 2aE\left[\left(\xi - E\xi\right)\left(\eta - E\eta\right)\right] + D\eta.$$

Отже,

$$D\zeta = a^2 D\xi - 2aE[(\xi - E\xi)(\eta - E\eta)] + D\eta. \qquad (1)$$

З другого доданку (1) отримуємо наступне означення.

Означення 1. Коваріацією між випадковими змінними ξ та η називають сподівання добутку відхилень цих змінних від своїх сподівань (або другому змішаному центральному моменту цих змінних), тобто

$$cov(\xi, \eta) = E\left[\left(\xi - E\xi\right)\left(\eta - E\eta\right)\right]. \qquad (2)$$

З цього означення безпосередньо випливає, що коваріація є симетричною функцією за змінними ξ та η, тобто

$$cov(\xi,\eta)=cov(\eta,\xi).$$

Якщо ξ та η – незалежні випадкові змінні , то $cov(\xi,\eta)=0$.

Справді, для незалежних випадкових змінних маємо
$cov(\xi,\eta)=E\left[(\xi-E\xi)(\eta-E\eta)\right]=E(\xi-E\xi)E(\eta-E\eta)=(E\xi-E\xi)(E\eta-E\eta)=0.$

Оскільки коваріація між незалежними випадковими змінними дорівнює нулю, то для залежних випадкових змінних коваріацію можна прийняти за міру залежності.

2. Кореляція

Коваріація між компонентами випадкового вектора вимірюється тими ж одиницями, що і добуток компонент. Але іноді треба мати абстрактну міру залежності між випадковими змінними. Вони одержуються діленням коваріації на добуток стандартів цих змінних. Одержану міру називають **кореляцією** і позначають через $\rho(\xi,\eta)$.

Означення. 2. Кореляцією між випадковими змінними ξ та η називають відношення коваріації між змінними ξ,η до стандартів цих змінних

$$\rho(\xi,\eta)=\frac{cov(\xi,\eta)}{\sqrt{D\xi}\sqrt{D\eta}}.$$

З цього означення випливає, що кореляція є симетричною функцією, тобто $\rho(\xi,\eta)=\rho(\eta,\xi)$.

Якщо ξ та η незалежні випадкові змінні, то $\rho(\xi,\eta)=0$.

Це випливає з того, що для незалежних випадкових змінних, коваріація дорівнює нулю, та з означення кореляції.

У термінах коваріації дисперсія (1) випадкової змінної ζ запишеться у вигляді:

$$D\zeta=a^2D\xi-2a\cdot cov(\xi,\eta)+D\eta,$$

а в термінах кореляції – таким чином:

$$D\zeta=a^2D\xi-2a\cdot\rho(\xi,\eta)\sqrt{D\xi}\sqrt{D\eta}+D\eta. \qquad (3)$$

3. Регресія

Знайдемо таке a, при якому випадкова змінна ζ має найменшу дисперсію. Для цього дисперсію (3) розглядатимемо як функцію параметра a.

$$f(a)=D\zeta=a^2D\xi-2a\cdot\rho(\xi,\eta)\sqrt{D\xi}\sqrt{D\eta}+D\eta. \qquad (4)$$

Обчислимо перші дві похідні за параметром a. З виразу (4), отримаємо

$$f'(a) = \left(D\zeta\right)'_a = 2aD\xi - 2\rho\left(\xi,\eta\right)\sqrt{D\xi}\sqrt{D\eta}, \quad f''(a) = 2D\xi > 0.$$

Оскільки друга похідна додатна, то при

$$a = \rho\left(\xi,\eta\right)\frac{\sqrt{D\eta}}{\sqrt{D\xi}}$$

з (4), отримаємо, що дисперсія випадкової змінної ζ набуває мінімального значення

$$(D\zeta)_{\min} = \left[1-\rho^2\left(\xi,\eta\right)\right]D\eta. \qquad (5)$$

Значення a, що мінімізує дисперсію випадкової змінної ζ називають **регресією** η на ξ.

Означення 3. Регресією випадкової змінної η на ξ називають добуток кореляції між цими змінними на відношення стандарту випадкової змінної η до стандарту випадкової змінної ξ

$$R\left(\eta/\xi\right) = \rho\left(\xi,\eta\right)\frac{\sqrt{D\eta}}{\sqrt{D\xi}}. \qquad (6)$$

З означення рівності (6) випливає, що регресія не є симетричною відносно випадкових змінних. Справді, регресія випадкової змінної ξ на η визначається як

$$R\left(\xi/\eta\right) = \rho\left(\xi,\eta\right)\frac{\sqrt{D\xi}}{\sqrt{D\eta}}. \qquad (7)$$

Якщо кореляція між його координатами дорівнює нулю ($\rho(\xi,\eta) = 0$), або стандарти координат випадкового вектора (ξ,η) рівні між собою, то регресія буде симетричною

$$R\left(\xi/\eta\right) = R\left(\eta/\xi\right).$$

Якщо ξ і η – незалежні випадкові змінні і кореляція між ними дорівнює нулю, то $R\left(\xi/\eta\right) = R\left(\eta/\xi\right) = 0$.

Із рівностей (6), (7) випливає, що знаки кореляції та регресії однакові, а добуток регресій дорівнює квадрату кореляції.

$$sgnR\left(\xi/\eta\right) = sgnR\left(\eta/\xi\right) = sgn\rho\left(\xi,\eta\right)$$

$$R\left(\xi/\eta\right)R\left(\eta/\xi\right) = p^2\left(\xi,\eta\right).$$

Розглянемо іншу випадкову змінну:

$$\vartheta = -c\eta + \xi,\ (c > 0,\ c = const).$$

Її дисперсія є такою:

$$D\vartheta = c^2 D\eta - 2c\cdot\mathrm{cov}(\eta,\xi) + D\xi,$$

а тому своє мінімальне значення

$$(D\vartheta)_{\min} = \left[1-\rho^2\left(\eta,\xi\right)\right]D\xi \qquad (8)$$

вона набуває при:

$$c = R(\xi / \eta) = \rho(\eta, \xi)\frac{\sqrt{D\xi}}{\sqrt{D\eta}},$$

де $R(\xi / \eta)$ – регресія першої координати випадкового вектора (ξ, η) не другу його координату.

Із невід'ємності дисперсії та з виразів (5), (8) випливає, що

$$1 - \rho^2(\xi, \eta) \geq 0, \quad 1 - \rho^2(\eta, \xi) \geq 0,$$

тобто $-1 \leq \rho(\xi, \eta) = \rho(\eta, \xi) \leq 1$.

1. Лінійно-залежні та некорельовані випадкові змінні

Нехай $\rho(\xi, \eta) = \pm 1$. З (5), (8) випливає, що тоді $D\zeta = 0$ та $D\vartheta = 0$. Це означає, що в цьому випадку з ймовірністю рівною одиниці випадкові змінні ζ та ϑ є константами. Нехай вони є відповідно b та d. Тоді $-a\xi + \eta = b, \quad -c\eta + \xi = d$, тобто

$$\eta = a\xi + b, \qquad (9)$$

$$\xi = c\eta + d, \qquad (10)$$

де $a = R(\eta / \xi), c = R(\xi / \eta)$. Отже, тоді η є лінійною функцією змінної ξ і навпаки. Очевидно, що в загальному випадку це різні функції.

Означення 3. Компоненти випадкового вектора (ξ, η) лінійно-залежні, якщо вони пов'язані співвідношеннями (9), або (10)

Означення 4. Компоненти випадкового вектора (ξ, η) називають некорельованими, якщо кореляція між ними дорівнює нулю.

Примітка. Можна навести приклад, у якому некорельовані випадкові змінні не обовязково є незалежними.

Нехай $0 < \rho(\xi, \eta) < 1$. Покажемо, що у співвідношенні

$$\eta = R(\eta / \xi)\xi + \zeta \qquad (11)$$

випадкові змінні ξ та ζ некорельовані, тобто $\rho(\xi, \zeta) = 0$ або $\operatorname{cov}(\xi, \zeta) = 0$. Справді за означенням (2) маємо

$$\operatorname{cov}(\xi, \zeta) = E\left\{[\xi - E\xi][-\rho(\xi, \eta)\frac{\sqrt{D\eta}}{\sqrt{D\xi}}(\xi - E\xi) + (\eta - E\eta)]\right\} =$$

$$= -\rho(\xi, \eta)\sqrt{D\xi}\sqrt{D\eta} + \operatorname{cov}(\xi, \eta) = 0.$$

Таким чином вираз (11) вказує на розклад випадкової змінної η на суму двох некорельованих випадкових змінних $R(\eta / \xi)\xi$ та ζ.

Із некорельованості випадкових змінних ξ та ζ у формулі (11) випливає, що дисперсія суми двох випадкових змінних дорівнює сумі дисперсій доданків

$$D\eta = \rho^2(\xi,\eta)D\eta + D\zeta\ .$$

Звідси випливає, що

$$D\eta = \rho^2(\xi,\eta)D\eta + [1-\rho^2(\xi,\eta)]D\eta\ . \qquad (12)$$

Тотожність (12) вказує на розклад дисперсії випадкової змінної η на частини, відповідні некорельованим доданкам у формулі (11). Отже, квадрат кореляції $\rho(\xi,\eta)$
між випадковими змінними ξ та η у формулі (12) вказує на те, яка частина дисперсії випадкової змінної η припадає на доданок, пропорціональний до випадкової змінної ξ у формулі (11).
Аналогічно можна показати, що у співвідношенні

$$\xi = R(\xi/\eta)\eta + \vartheta$$

випадкові змінні η та ϑ є некорельованими.
Зауваження. Властивості кореляції між двома випадковими змінними ξ та η нагадують властивості косинуса кута між двома векторами $\bar{a}$ та $\bar{b}$

Встановимо відповідність між випадковими змінними та векторами. Для цього звернемо увагу на таку аналогію між кореляцією $\rho(\xi,\eta)$ та величинами $\cos(\bar{a},\bar{b})$, де $\bar{a},\bar{b}$ - деякі вектори:

	$\rho(\xi,\eta)$	$\cos(\bar{a},\bar{b})$
1	Кореляція симетрична за змінними ξ та η: $\rho(\xi,\eta)=\rho(\eta,\xi)$	Косинус-парна функція $\cos(\bar{a},\bar{b})=\cos(\bar{b},\bar{a})$
2	Кореляція між незалежними випадковими змінними ξ та η дорівнює нулю: $\rho(\xi,\eta)=0$	Косинус кута між ортогональними векторами дорівнює нулю: $\cos(\bar{a},\bar{b})=0,$ *якщо* $\bar{a}\perp\bar{b}$
3	Кореляція набуває значення від -1 до +1: $\lvert\rho(\xi,\eta)\rvert\le 1$	Косинус кута між двома векторами набуває значення від -1 до +1: $\lvert\cos(\bar{a},\bar{b})\rvert\le 1$
4	$\rho(\xi,\eta)=1,$ якщо ξ та η лінійно залежні	$\cos(\bar{a},\bar{b})=1,$ *якщо* $\bar{a}\uparrow\uparrow\bar{b}$
5	$\rho(\xi,\eta)=-1,$ якщо ξ та η лінійно залежні	$\cos(\bar{a},\bar{b})=-1,$ якщо $\bar{a}\uparrow\downarrow\bar{b}$

Наявність описаної аналогії між кореляцією і косинусом дає можливість будь - які дві випадкові змінні схематично зобразити на площині двома векторами, що мають довжини рівні відповідним стандартам цих змінних і виходять з однієї точки під кутом, косинус якого дорівнює кореляції між цими

змінними. Тоді співвідношення між випадковими змінними можна виразити геометрично та всі дослідження здійснювати відповідними методами геометрії.

Прямі регресії

Нехай над двовимірним впорядкованим випадковим вектором $(\xi;\eta)$.проведено в однакових умовах n пар незалежних спостережень: $(x_1, y_1),\ldots, (x_n, y_n)$.
Випадкові змінні ξ та η нам невідомі, тому на основі лише цих спостережень потрібно зробити висновок про кореляцію між ними.
Обчислимо середні арифметичні вибіркових значень координат цього вектора:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

та їх варіанси

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\ ,\quad s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}\ )^2\,.$$

Варіанси s_1^2, s_2^2 є незміщеними оцінками дисперсій випадкових змінних ξ та η, відповідно. Тому, за аналогією з рівностями (2), (3), (6) див. Кореляційний аналіз, **вибіркова коваріація між координатами вектора** (ξ,η) визначається так:

$$c_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad (1)$$

вибіркова кореляція –

$$r_{12} = \frac{c_{12}}{s_1 s_2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - y)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (2)$$

вибіркова регресія другої координати на першу

$$b_{21} = r_{12}\frac{s_2}{s_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ , \qquad (3)$$

вибіркова регресія першої координати на другу

$$b_{12} = r_{12}\frac{s_1}{s_2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\,. \qquad (4)$$

Вибірковим аналогом дисперсії випадкової змінної

$$\zeta = -a\xi + \eta,\quad (a > 0, a = const)$$

є варіанса:

$$s_\zeta^2 = a^2 s_1^2 - 2\,a\,c_{12}\,s_1\,s_2 + s_2^2,$$

яка при

$$a = b_{12} = r_{12}\frac{s_2}{s_1}$$

досягає мінімуму

$$(s_\zeta^2)_{\min} = (1 - r_{12}^2)\,s_2. \qquad (5)$$

З (5) випливає, що $-1 \le r_{12} \le 1$. Зокрема, якщо $r_{12} = \pm 1$, то $(s_\zeta^2)_{\min} = 0$. Отже, всі $-ax_i + y_i$ приймають однакове значення $b = const$, тобто всі спостережувані точки $(x_1, y_1), \cdots, (x_n, y_n)$ лежать на одній прямій $y = a\,x + b$.

Зауваження 1. Якщо $r_{12} = \pm 1$, то звідси не випливає, що і в генеральній сукупності коефіцієнт кореляції $\rho(\xi, \eta) = \pm 1$, тобто, що і в ній ξ та η пов'язані функціональним лінійним зв'язком. Дійсно, вибіркову кореляцію і всі висновки ми отримуємо на основі конкретного статистичного розподілу представленого парами точок. Якщо здійснимо ряд спостережень навіть того ж обсягу над тими ж випадковими змінними ξ та η, то отримаємо, взагалі кажучи, інший статистичний розподіл, для якого $r_{12} \ne \pm 1$. ■

Якщо $-1 < r_{12} < 1$, то проведемо через точку $(\bar{x}, \bar{y})$ пряму $y = ax + b$, яка мінімізує суму квадратів відхилень вибіркових значень вектора (ξ, η) в напрямку осі *OY*. Для цього потрібно мінімізувати функцію

$$f(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2. \qquad (6)$$

Необхідна умова екстремуму функції (6) еквівалентна системі рівнянь:

$$\begin{cases} f'_a = 0 \\ f'_b = 0 \end{cases}, \text{ тобто } \begin{cases} \sum_{i=1}^{n} 2(y_i - ax_i - b)(-x_i) = 0, \\ \sum_{i=1}^{n} 2(y_i - ax_i - b)(-1) = 0. \end{cases}$$

Звідси остаточно маємо систему:

$$\begin{cases} \sum_{i=1}^{n} (y_i - ax_i - b)x_i = 0, \\ \sum_{i=1}^{n} (y_i - ax_i - b) = 0. \end{cases} \qquad (7)$$

Друге рівняння цієї системи еквівалентне рівнянню $n\bar{y} - an\bar{x} - nb = 0$, тому $b = \bar{y} - a\bar{x}$. Підставимо цей вираз у перше рівняння системи (7). Тоді послідовно отримаємо:

$$\sum_{i=1}^{n} ((y_i - \bar{y}) - a(x_i - \bar{x}))x_i = 0,\ \sum_{i=1}^{n} (y_i - \bar{y})x_i - a\sum_{i=1}^{n} (x_i - \bar{x})x_i = 0,$$

$$a = \frac{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \bar{y})x_i}{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})x_i} = \frac{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})^2} = b_{21}.$$

Справді,

$$\sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y})x_i - \sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y})\bar{x} =$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y})x_i - \bar{x}\sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y}) = \sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y})x_i,$$

$$\sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})y_i - \sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})\bar{y} =$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})y_i,$$

$$\sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y})x_i = \sum_{\mathrm{i}=1}^{\mathrm{n}}x_i y_i - n\bar{x}\bar{y}.$$

Аналогічно доводимо, що

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

$$\sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y})^2 = \sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y})y_i.$$

Отже

$$\begin{cases} a = b_{21}, \\ b = \bar{y} - b_{21}\bar{x}, \end{cases}$$

- розв’язок системи рівнянь (7).

Безпосередньо переконуємося, що

$$f''_{a^2} = 2\sum_{i=1}^{n}x_i^2 > 0,\ f''_{a^2}f''_{b^2} - (f'_a\ f'_b)^2 = 4n\,(\sum_{i=1}^{n}x_i^2 - n\,\bar{x}^{-2}) > 0.$$

Отже точка з координатами (a,b) є точкою мінімуму функції (6), тобто сума відхилень точок вибірки (x_i, y_i), $(\mathrm{i}=1,...,\mathrm{n})$ від прямої

$$y - \bar{y} = b_{21}(x - \bar{x}) \qquad (8)$$

в напрямку осі OY є найменшою. Пряма (8) називається **прямою регресії другої компонента вектора** (ξ,η) **на першу**.

Аналогічно можна показати, що пряма

$$x - \bar{x} = b_{12}(y - \bar{y}) \qquad (9)$$

мінімізує суму квадратів відхилень від точок даної вибірки, але в напрямку осі OX.

Проведемо через точку $(\bar{x},\bar{y})$ пряму $x = c\ y + d$, яка мінімізує суму квадратів відхилень вибіркових значень вектора (ξ,η) в напрямку осі *OX*. Для цього потрібно мінімізувати функцію

$$g(a,b) = \sum_{i=1}^{n}(x_i - c\ y_i - d)^2 . \qquad (6')$$

Пряма (9) називається **прямою регресії першої компонента вектора** (ξ,η) **на другу**.

Зауваження 2. Якщо вибіркова кореляція $r_{12} = \pm 1$, то обидві прямі регресії першої компонента вектора (ξ,η) на другу та другої координати на першу співпадають і навпаки, - якщо ці прямі співпадають, то вибіркова кореляція $r_{12} = \pm 1$.

Справді, обидва рівняння прямих регресії

$$x - \bar{x} = r_{12}\frac{s_1}{s_2}(y - \bar{y}) \text{ та } y - \bar{y} = r_{12}\frac{s_2}{s_1}(x - \bar{x})$$

при $r_{12} = \pm 1$ еквівалентні одному й тому ж рівнянню

$$\frac{x - \bar{x}}{s_1} = \pm\frac{y - \bar{y}}{s_2}.$$

Якщо ж обидва ці рівняння прямих регресії співпадають, то

$$r_{12}\frac{s_2}{s_1} = \frac{1}{r_{12}}\frac{s_2}{s_1}.$$

Отже тоді $r^2 = 1$, тобто $r_{12} = \pm 1$.

Багатовимірний регресійний аналіз.

Кореляції вищих порядків

У попередніх викладках ми розглядали кореляцію між компонентами двовимірного вектора. Очевидно, такі ж питання виникають, коли розглядаємо випадкові вектори з кількістю координат більшою двох.

Нехай випадковий вектор має координати $(\xi_1,\xi_2,...,\xi_n)$. Хочемо вивчити кореляцію, наприклад, між ξ_1 та ξ_2. Для цього будемо вважати координати $\xi_3,\xi_4,...,\xi_n$ зафіксованими, тобто такими, які набули певних значень з відповідних їх областей визначення і далі не змінюються. Однак, проблема полягає в тому, що кореляція між виділеними для аналізу змінними ξ_1 та ξ_2 може бути обумовлена тим, що обидві вони цілком, чи частково залежать, наприклад, від третьої змінної $\xi_i, i = \overline{3,n}$, в той час як при кожному фіксованому значенні цієї змінної $\xi_i, i = \overline{3,n}$ змінні ξ_1 та ξ_2 – стохастично незалежні. Тому спочатку розглянемо кореляцію між ξ_1 та ξ_2 тривимірного випадкового вектора (ξ_1,ξ_2,ξ_3).

Вплив третьої координати на перші дві спробуємо нейтралізувати замінивши змінні ξ_1 та ξ_2 такими величинами $\tilde{\xi}_1 = \xi_1 - \lambda\,\xi_3$, $\tilde{\xi}_2 = \xi_2 - \mu\,\xi_3$, коефіцієнти кореляції яких з ξ_3 рівні нулю. Якщо $E(\xi_1) = E(\xi_2) = E(\xi_3) = 0$, то коефіцієнт кореляції

$$\rho(\xi_1, \xi_2) = \frac{E(\xi_1\,\xi_2)}{\sqrt{D(\xi_1)}\sqrt{D(\xi_2)}}. \qquad (1)$$

Отже множники λ, μ потрібно вибирати так, щоб

$$E(\hat{\xi}_1\,\xi_3) = E(\xi_1\xi_3 - \lambda\,\xi_3^2) = 0,\ E(\hat{\xi}_2\,\xi_3) = E(\xi_2\xi_3 - \mu\,\xi_3^2) = 0.$$

Звідси

$$\lambda = \frac{E(\xi_1\,\xi_3)}{E(\xi_3^2)} = \frac{\rho(\xi_1, \xi_3)\sqrt{D(\xi_1)}\sqrt{D(\xi_3)}}{D(\xi_3)} = \rho(\xi_1, \xi_3)\frac{\sqrt{D(\xi_1)}}{\sqrt{D(\xi_3)}}, \qquad (2)$$

$$\mu = \frac{E(\xi_2\,\xi_3)}{E(\xi_3^2)} = \frac{\rho(\xi_2, \xi_3)\sqrt{D(\xi_2)}\sqrt{D(\xi_3)}}{D(\xi_3)} = \rho(\xi_2, \xi_3)\frac{\sqrt{D(\xi_2)}}{\sqrt{D(\xi_3)}}. \qquad (3)$$

Коефіцієнт кореляції між ξ_1 та ξ_2 без впливу змінної ξ_3 дорівнює

$$\rho((\xi_1, \xi_2)/\xi_3) = \frac{E((\xi_1 - \lambda\,\xi_3)(\xi_2 - \mu\,\xi_3))}{\sqrt{D(\xi_1 - \lambda\,\xi_3)}\sqrt{D(\xi_2 - \mu\,\xi_3)}}. \qquad (4)$$

Розкривши дужки та використовуючи відповідні властивості математичного сподівання, формулу (1), переконуємося, що

$$E((\xi_1 - \lambda\,\xi_3)(\xi_2 - \mu\,\xi_3)) =$$
$$= \rho(\xi_1, \xi_2)\sqrt{D(\xi_1)}\sqrt{D(\xi_2)} - \lambda\,\rho(\xi_2, \xi_3)\sqrt{D(\xi_2)}\sqrt{D(\xi_3)} -$$
$$- \mu\,\rho(\xi_1, \xi_3)\sqrt{D(\xi_1)}\sqrt{D(\xi_3)} + \lambda\,\mu\,D(\xi_3).$$

Підставивши в цю рівність формули (2), (3), одержимо:

$$E((\xi_1 - \lambda\,\xi_3)(\xi_2 - \mu\,\xi_3)) =$$
$$= (\rho(\xi_1, \xi_2) - \rho(\xi_1, \xi_3)\rho(\xi_2, \xi_3))\sqrt{D(\xi_1)}\sqrt{D(\xi_2)}. \qquad (5)$$

Аналогічно доводимо, що

$$D(\xi_1 - \lambda\,\xi_3) = E(\xi_1 - \lambda\,\xi_3)^2 = (1 - \rho^2(\xi_1, \xi_3))D(\xi_1), \qquad (6)$$

$$D(\xi_2 - \mu\,\xi_3) = E(\xi_2 - \mu\,\xi_3)^2 = (1 - \rho^2(\xi_2, \xi_3))D(\xi_2). \qquad (7)$$

Підставивши (5), (6), (7) в (4), остаточно маємо:

$$\rho((\xi_1, \xi_2)/\xi_3) = \frac{\rho(\xi_1, \xi_2) - \rho(\xi_1, \xi_3)\rho(\xi_2, \xi_3)}{\sqrt{1 - \rho^2(\xi_1, \xi_3)}\sqrt{1 - \rho^2(\xi_2, \xi_3)}}. \qquad (8)$$

Використовуючи замість випадкових змінних ξ_1, ξ_2, ξ_3 їх вибіркові значення відповідно x_i, y_i, z_i, цілком так само, як отримано формулу (8), отримуємо співвідношення

$$r_{12[3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}, \qquad (9)$$

де r_{12}, r_{13}, r_{23} вибіркові кореляції, відповідно, між першою і другою координатою, між першою та третьою, між другою та третьою координатою випадкового вектора (ξ_1, ξ_2, ξ_3), $r_{12[3}$ – умовна кореляція між першою та другою координатою цього вектора при зафіксованій третій координаті.

Формулу (9) узагальнюємо індукцією за розмірністю простору випадкових змінних наступним чином. Індекси, які відповідають постійним значенням змінних, будемо називати ***німими***, а кількість німих індексів – ***порядком частинної кореляції***. Німі індекси записуватимемо після відкриваючої квадратної дужки. Тоді, наприклад, $r_{12[34...m}$ – ***частинна вибіркова кореляція*** $(m-2)$ – го порядку між координатами ξ_1 та ξ_2. Визначимо рекурентно частинну вибіркову кореляцію $(m-2)$ – го порядку через три частинні вибіркові кореляції $(m-3)$ – го порядку за формулою:

$$r_{12[34..m} = \frac{r_{12[34...(m-1)} - r_{1m[34...(m-1)}\, r_{2m[34..(m-1)}}{\sqrt{1-r^2_{1m[34...(m-1)}}\,\sqrt{1-r^2_{2m[34...(m-1)}}}. \qquad (10)$$

Приклад. Частинна кореляція між першою і другою координатою тривимірного вектора, згідно (10), як і згідно формули (9), рівна:

$$r_{12[3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r^2_{13}}\,\sqrt{1-r^2_{23}}},$$

а кореляція між тими ж координатами чотиривимірного вектора:–

$$r_{12[34} = \frac{r_{12[3} - r_{14[3}\, r_{24[3}}{\sqrt{1-r^2_{14[3}}\,\sqrt{1-r^2_{24[3}}}. \quad \blacksquare$$

З означення кореляцій вищого порядку випливає , що ***активні,*** незафіксовані індекси можна переставляти місцями, а зафіксовані, німі, ***пасивні*** переставляти не можна – відносно них кореляція несиметрична.

Отже маємо цілий набір частинних кореляцій, які визначаються різними перестановками індексів в них. Щоб обчислити частинні кореляції $(m-2)$ порядку необхідно мати всі такі частинні кореляції нульового порядку:

$$r_{12}, r_{13}, ..., r_{1m},$$
$$r_{23}, ..., r_{2m},$$
$$r_{m-1,m}.$$

Варіанси вищих порядків

Варіанси вищих порядків також будуємо за аналогією до формування кореляцій вищих порядків, тобто фіксуванням всіх неактивних координат випадкового вектора. Отже така варіанса має один активний індекс, а інші $(m-1)$ – пасивні, німі. Цю ***вибіркову частинну варіансу*** будемо називати ***варіансою*** $(m-1)$ ***– го порядку***. Варіансу $(m-1)$ –го порядку визначаємо рекурентно через варіансу $(m-2)$ – го порядку та кореляцію $(m-2)$ – го порядку за формулою:

$$s^2_{1[23...m} = s^2_{1[23...(m-1)}(1-r^2_{1m[23...(m-1)}). \qquad (11)$$

Приклад. Частинна варіанса першої координати двовимірного вектори при фіксованій другій координаті, згідно (11), дорівнює:

$$s^2_{1[2} = s^2_1(1-r^2_{12}),$$

а частинна варіанса першої координати тривимірного вектори при фіксованих двох інших:

$$s^2_{1[23} = s^2_{1[2}(1-r^2_{13[2}). \quad \blacksquare$$

Послідовно застосовуючи формулу (11), можна виразити варіансу $(m-2)$ – го порядку через варіансу першого порядку:

$$s^2_{1[23\ldots m} = s^2_1 \underbrace{(1-r^2_{12})(1-r^2_{13[2})\ldots(1-r^2_{1m[2\ldots(m-1)})}_{1-R^2_{1m(23\ldots(m-1)}} = s^2_1(1-R_{1m(23\ldots(m-1)}) \quad . \quad (12)$$

Стандарт m-ого порядку визначають звичайним чином:

$$s_{1[23\ldots m} = \sqrt{s^2_{1[23\ldots m}} \quad . \quad (13)$$

Регресії вищих порядків

Регресії вищих порядків також визначаємо за індукцією. Позначимо вибіркову регресію між ξ_1 та ξ_2 при постійних значеннях всіх інших координат $\xi_3, \xi_4, \ldots, \xi_m$ через $b_{12[3,\ldots m}$. Тоді ***частинну регресію*** $(m-2)$ ***- порядку*** визначають через кореляцію ($m-2$) – го порядку та два стандарти $(m-2)$ – го порядку за формулою:

$$b_{12[3,\ldots,m} = r_{12[3,\ldots,m} \frac{s_{1[3,4,\ldots,m}}{s_{2[3,4,\ldots,m}} . \quad (14)$$

Зокрема, при $m=3$ одержуємо регресію першого порядку

$$b_{12[3} = r_{12[3} \frac{s_{1[3}}{s_{2[3}},$$

а при m =2 – відому вже раніше регресію нульового порядку

$$b_{12} = r_{12} \frac{s_1}{s_2} .$$

Лінійне рівняння регресії однієї змінної на інші

За аналогією до двовимірного випадку, можна говорити і про лінійне рівняння регресії однієї координати m вимірного випадкового вектора ($m>2$) відносно всіх інших координати цього ж вектора. Очевидно, що вивід таких рівнянь буде подібним до двовимірному випадку з урахуванням того, що тепер потрібно аналізувати частинні регресії вищих порядків.

Нехай в результаті N спостережень над випадковим вектором $(\xi_1, \cdots, \xi_m)$ одержано k різних значень $(x_{1j}, \cdots, x_{mj})$, $(j = \overline{1,k})$ його, причому значення $(x_{1j}, \cdots, x_{mj})$ цей вектор набував n_j разів. Очевидно, що $N = n_1 + \cdots + n_k$.Обчислимо середні арифметичні

$$\bar{x}_i = \frac{1}{N}\sum_{j=1}^{k} n_j x_{ij}$$

координат вектора $(\xi_1, \cdots, \xi_m)$. Будемо шукати, наприклад, гіперплощину

$$x_1 - \bar{x}_1 = a_{12}(x_2 - \bar{x}_2) + a_{13}(x_3 - \bar{x}_3) + \cdots + a_{1m}(x_m - \bar{x}_m), \qquad (15)$$

яка проходить через точку $(\bar{x}_1, \cdots, \bar{x}_m)$ і найкращим чином, в сенсі принципу найменших квадратів (див. § 5), узгоджується з експериментальними даними $(x_{1j}, \cdots, x_{mj})$, $(j = \overline{1,k})$. Отже задача полягає у відшуканні таких коефіцієнтів $a_{1i}, i = 2, 3, \cdots, m$, при яких сума

$$S = \sum_{j=1}^{k} n_j (x_{1j} - \bar{x}_1 - \sum_{i=2}^{m} a_{1i}(x_{ij} - \bar{x}_i))^2$$

набуває найменшого значення. З необхідної умови мінімуму цієї суми маємо систему рівнянь:

$$-\frac{1}{2}\frac{\partial S}{\partial a_{1k}} = \sum_{j=1}^{k} n_j (x_{kj} - \bar{x}_k)(x_{1j} - \bar{x}_1 - \sum_{i=2}^{m} a_{1i}(x_{ij} - \bar{x}_i)) = 0, \quad (k = \overline{2,m}), \qquad (16)$$

розв'язавши яку з (15) отримаємо лінійне рівняння регресії першої компоненти m - мірного випадкового вектора на всі інші координати:

$$x_1 - \bar{x}_1 = b_{12[3,4,\ldots,m}(x_2 - \bar{x}_2) + b_{13[2,4,\ldots,m}(x_3 - \bar{x}_3) + \ldots + b_{1m[2,3,\ldots(m-1)}(x_m - \bar{x}_m). \qquad (17)$$

При $m = 2$ це рівняння перетворюється у відоме рівняння прямої регресії першої координати двовимірного випадкового вектора на другу координату.

Емпіричне рівняння регресії m - го порядку аналітично зображає випадковий процес, що описується випадковими змінними $\xi_1, \ldots, \xi_m$ і слугує для подання великого за обсягом статистичного матеріалу компактно аналітично.