ТІМС

1. Класифікація варіант
2. Статистичний матеріал
3. Табличне та графічне представлення статистичного матеріалу
    a. дискретний випадок
    b. неперервний випадок
4. Статистики центральної тенденції (медіана, мода)
5. Середнє арифметичне, його властивості
6. Статистики розсіяння
7. Інтерквантильні широти
8. Симетричні інтерквантильні широти
9. Моменти статистичного матеріалу
10. Статистичні форми
11. Майже вірогідна подія. Посилений закон великих чисел
12. Посилений закон великих чисел для функції розподілу
13. Схема статистичного доведення
14. Критерій Хі-квадрат. Умови застосованості
15. Метод максимуму правдоподібності
16. Статистичне оцінювання параметрів нормальної популяції
17. Оцінка невідомого математичного сподівання нормальної генеральної сукупності
18. Порівняння математичних сподівань двох нормальних популяцій
19. Інтервал довіря невідомого математичного сподівання
20. Гіпотезу про дисперсію нормального розподілу популяції
21. Інтервал довіря невідомої дисперсії нормального розподіл популяції
22. Порівняння дисперсій двох нормальних популяцій.
23. Критерій Колмогорова
24. Критерій Смирнова
25. Критерій знаків. Інтервал для прийняття рішень
26. Гіпотеза про медіану
27. Критерій Вілкінсона
28. Однофакторний варіансний аналіз
29. Двофакторний варіансний аналіз
30. Трифакторний варіансний аналіз
31. Варіансний аналіз за схемою латинського квадрата
32. Кореляційний аналіз (коваріація, кореляція, регресія)
33. Пряма регресія
34. Кореляція вищих порядків
35. Варіанси та стандарти вищих порядків

Класифікація варіант

Мінливі величини (варіанти) 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,....., і кількісні x, y, z,.....

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

Приклади кількісних варіант: число зерен у голівці 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.

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.

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

Табличне та графічне представлення статистичного матеріалу

a. дискретний випадок

Нехай серед спостережень $x_1, \ldots\ldots\ldots, x_n$ зустрічаються такі можливі значення одновимірної дискретної варіанти x, впорядковані за величиною:

$x_{(1)} < x_{(2)} < \ldots\ldots\ldots\ldots < x_{(k)}$

і нехай ці значення зустрічаються відповідно часто:

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

Число n_i називається частотою значення $x_{(i)} (i = 1,2,\ldots k).$ Тоді статистичний матеріал 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

(2)

$x_{(1)}$	$x_{(2)}$	$\ldots\ldots\ldots\ldots$	$x_{(k)}$	$\sum$
n_1	n_2	$\ldots\ldots\ldots\ldots$	n_k	n

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

Для графічного представлення частотної таблиці на вісь абсцис наносимо можливі значення дискретної мінливої величини та відкладемо в цих точках відповідні частоти. Отримаємо діаграму частот.

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

b. неперервний випадок

1. незгруповані дані

Якщо статистичний матеріал малий або середній, то спостереження над одновимірною неперервною варіантою впорядкуємо за величиною: від найменшого до найбільшого.

В силу обмеженої точності деякі спостереження можуть бути однакові, так упорядковані спостереження 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 над одновимірною неперервною мінливою величиною.

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

Емпірична функція розподілу або емпірична кумулята:

$$F_n(x)=\begin{cases}0, & x<x_{(1)}\\ \frac{m_i}{n}, & x_{(i)}\le x<x_{(i+1)}\quad (i=\overline{1,n-1})\\ 1, & x_{(n)}\le x,\end{cases}$$

2. згруповані дані

Якщо статистичний матеріал середній або великий, то знайдемо найменше та найбільше зі спостереження:

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

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

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

Інтервал розмаху ділимо досить довільним способом на *(r + 1)* однакові або неоднакові інтервали, де *r* натуральне , *r* = 1, 2, 3……

Центри одержаних інтервалів позначимо в зростаючому порядку через:

$z_1,\ldots,z_i,\ldots,z_{r+1}.$

Нехай на інтервалі з центром в т. z_i попадає n_i спостережень. Очевидно, що $n_1+n_2+\ldots\ldots n_{r+1}=n$.

Тоді статистичний матеріал представимо у вигляді таблиці з двох рядків:

1-й в зростаючому порядку - центри інтервалів

2 - відповідні частоти

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

Одержана таблиця – частотна.

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 полігон частот.

Якщо над інтервалом з центром в т. z_i поставити прямокутник висотою n_i, то одержимо гістограму частот.

Статистики центральної тенденції (медіана, мода)

Числові характеристики центральної тенденції (локації). До них відноситься:

1. медіана (Me)
2. мода (Mo)
3. середнє арифметичне ($\overline{x}$)

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

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

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

$$M_e = \left[x_{(k)}, x_{(k+1)}\right] \qquad M_e = \frac{x_{(k)} + x_{(k+1)}}{2}$$

Твердження. Лише медіана може мінімізувати суму абсолютних відхилень елементів статистичного матеріалу від сталої.
Доведення. Справді, позначимо через

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

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 = Me
.

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

Не виключено, що декілька значень статистичного матеріалу зустрічаються найчастіше та однаково часто, тоді всі вони модні.

Середнє арифметичне, його властивості.

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

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

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 $x_{(1)} \le \overline{x} \le x_{(n)}$.

Доведення. З означення:

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

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

2. Сума відхилень елементів статистичного матеріалу від середнього арифметичного рівна нулю. $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

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

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

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

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

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

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

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

$f'(a) = 2n > 0$ - точка мінімуму.

Статистики розсіяння

Числові характеристики розсіяння:
а) варіанса (s2)
б) стандарт (s)
в) розмах (p)
г) варіація (v)
д) інтерквантильність широт

Девіація — сума квадратів відхилень елементів статистичного матеріалу від середнього арифметичного (Dev)

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

Числом ступенів вільності — число n-1

$$d.A. = n-1 \qquad degreds\ of\ freedom$$

Варіансою — це девіація поділена число ступенів вільності.

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

Стандартом — арифметичний корінь з варіанси

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

Розмахом називається різниця між найбільшим і найменшим елементами статистичного матеріалу

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

Варіацією називається відношення стандарту 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 $\alpha\%$ елементів статистичного матеріалу (відповідного варіаційного ряду).

Статистичний матеріал $x_1, \ldots, x_n$ (1) має квантилі тільки порядків кратних $i \cdot \frac{100}{n}$, інші квантилі не існують; елемент $x_{(i)}$ є квантилем 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:

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

Симетричні інтерквантильні широти

Існують симетричні інтеквантильні широти

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

$$x_{(n+1-i)} - x_{(i)}.$$

Вони рівні

Квантилі порядку 25, 50, 75 називаються **квартилями**: першим Q1; другим Q2; третім Q3.
Різниця між третім і першим квартилем Q3 - Q1 називається *інтерквартильною широтою* (інтерквартильний розмах) (50 % елементів).

Квантилі порядку 12,5; 25,0;..., 87,5 називаються **октилями**: першим О1;другим О2; ...сьомим О7.
Різниця між сьомим і першим октилем О7 - О1 називається *інтероктильною широтою*. (75 % елементів)

Квантилі порядку 10; 20;..., 90 називаються **децилями**: першим Д1; другим Д2; ...дев'ятим Д9 .
Різниця між дев'ятим і першим децилем Д9 - Д1 називається *інтердецильною* широтою.(80 % елементів)

Квантилі порядку 01; 02;..., 99 називаються **центилями**: першим C01 ; другим C02; ...дев'яносто дев'ятим C99 .
Різниця між дев'яносто дев'ятим і першим центилем C99-C01 називається *інтерцентильною широтою.*(98 % елементів)

Квантилі порядку 00,1; 00,2;..., 99,9 називаються **мілілями**: першим M001; … дев'ятсот дев'яносто дев'ятим M999.
Різниця M999 - M001 називається *інтермілільною широтою.*(99,8 % елементів)

Моменти статистичного матеріалу

Означення. **Моментом** порядку *H* відносно сталої *a* називається вираз

$$\mu_H(a)=\frac{1}{n}\sum_{i=1}^{n}(x_i-a)^H \qquad (H=1,2,\ldots)$$

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

$$m_H=\frac{1}{n}\sum_{i=1}^{n}x_i^H \qquad (H=1,2,\ldots)$$

Покладемо, за означенням, $m_0=1$. .
Очевидно, що 1-ий початковий момент збігається із середнім арифметичним і позначається $m_1=\overline{x}$.

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

$$\mu_H=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^H \qquad (H=1,2,\ldots)$$

Очевидно, що 1-ий центральний момент дорівнює

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

2-й центральний момент запишеться у вигляді

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

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

Зауважимо, що практично найчастіше використовуються такі моменти:
1-й початковий $m_1=\overline{x}$

2-й центральний $\mu_2=s^2$

3-й центральний і 4-й центральний $\mu_3;\ \mu_4$.

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

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:

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

Статистичні форми

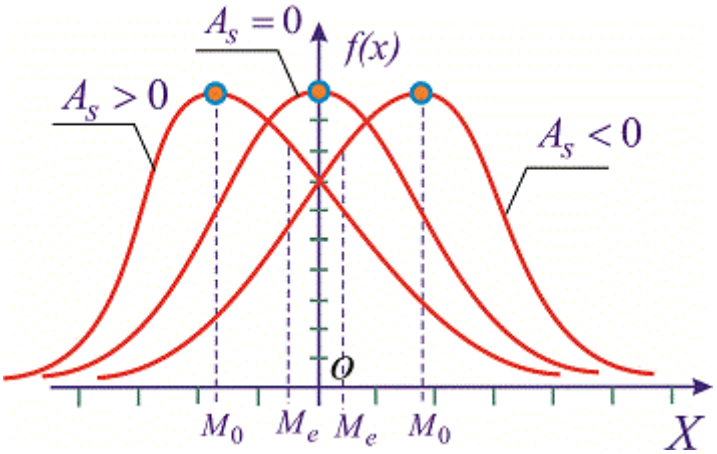
Для характеристики форми мінливості статистичного матеріалу, Фішер увів дві статистики:

1. асиметрію
2. ексцес.

Асиметрією або скошеністю статистичного матеріалу називається відношення 3-го центрального моменту до 2-го центрального моменту в степені півтора

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

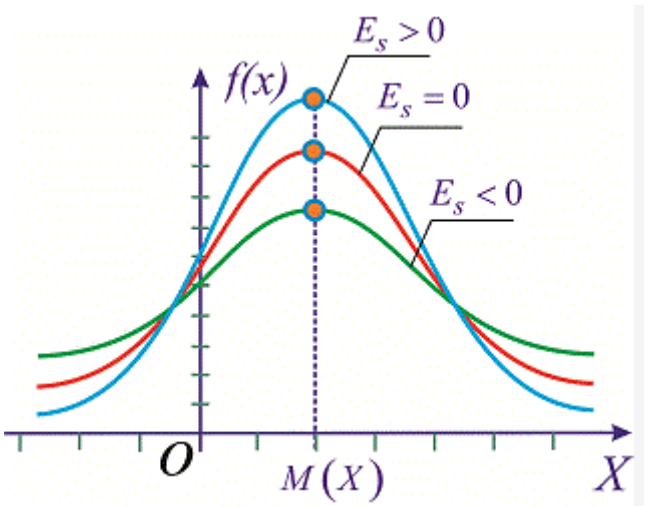
- Якщо As > 0, то статистичний матеріал скошений вправо (більшість елементів зліва)
- Якщо As < 0, то статистичний матеріал скошений вліво (більшість елементів справа)
- Якщо As = 0, то статистичний матеріал симетричний.



Ексцесом (крутістю, сплющеністю) статистичного матеріалу 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 Ek > 0, то статистичний матеріал – високовершинний.
При Ek < 0 , то статистичний матеріал – низьковершинний.
При Ek = 0, то статистичний матеріал – нормальновершинний.



Майже вірогідна подія. Посилений закон великих чисел

Коли подія А не еквівалентна вірогідній, але все одно її ймовірність P(A)=1, тоді подія А **майже вірогідна**.

Протилежна до майже вірогідної – це **майже неможлива подія**.

Можливості не попасти в точку (0,0) та попасти в неї при довільному виборі точки в координатні площині - це майже вірогідна та майже неймовірна події.

Посилений закон великих чисел

Теорема: Нехай μ - кількість появ події A з ймовірністю p в серії з n незалежних експериментів, а ε > 0 тоді

$$\lim_{n\to\infty} P\{|\frac{\mu}{n} - p| < \varepsilon\} = 1$$

.

У 1903 р. Барель узагальнив таку формулу до вигляду

$$P\{\frac{\mu}{n} \underset{n\to\infty}{\longrightarrow} p\} = 1$$

.

З цього випливає, що при великій кількості експериментів p наближено рівне m/n, де m – кількість сприятливих експериментів.

Посилений закон великих чисел для функції розподілу

Теорема 1. *Нехай $F(x)$ – теоретична функція розподілу випадкової змінної ξ, а $F_n(x)$ – емпірична функція розподілу результатів n незалежних спостережень над змінною ξ, проведених при незмінних умовах. Тоді напевно верхня грань модуля відхилень $F(x)$ від $F_n(x)$, при нескінченно зростаючій кількості експериментів, прямує до нуля для всіх значень аргументу, тобто*

$$P\left(\sup_{-\infty<x<+\infty} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{} 0\right) = 1. \qquad (6)$$

Схема статистичного доведення

В кожному статистичному доведенні є наступні кроки:

- Формулюється нульова гіпотеза H_0.
- Вибирається рівень значущості α (альфа)
- Обирається відповідно до нульової гіпотези статистика
- Знаходиться розподіл цієї статистики
- На основі знайденого розподілу визначаємо критичну область для статистики при рівні значущості α
- Знаходимо емпіричне значення статистики

- На основі отриманих даних робимо висновок про правомірність висунутої нульової гіпотези:

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

Статистичне доведення відрізняється від математичного доведення. Останнє базоване на логіці, тоді як статистичне може мати 4 ситуації:

- Гіпотеза вірна і в результаті статистичного доведення ми її приймаємо
- Гіпотеза хибна і в результаті статистичного доведення ми її відкидаємо
- Гіпотеза вірна і в результаті статистичного доведення ми її відкидаємо
- Гіпотеза хибна і в результаті статистичного доведення ми її приймаємо

В статистичному доведенні можливі 2 типи похибки: відкинення правдивої гіпотези (похибка 1-го типу) і прийняття хибної (похибка 2-го типу). Ймовірність відкинути істинну гіпотезу є рівнем значущості і його позначають α і він переважно вибирається 0.05, а рідше 0.1 або 0.01. Перше статистичне доведення в сучасному розумінні провів англійський математик Пірсон в 1900 р – році заснування математичної статистики.

Критерій Хі-квадрат. Умови застосованості

Нехай дана вибірка з генеральної сукупності і нам варто перевірити гіпотезу про вид розподілу, тобто про приналежність розподілу вибірки деякому параметричному

$$X = (x_1, x_2, \dots, x_n)$$

$$H: F(x)$$

сімейству.

Поділимо генеральну сукупність довільним чином на r+1 частин. За міру відхилення теоретичного розподілу від вибірки Пірсон прийняв величину

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

Також він довів, що для вибірок великого обсягу статистика має розподіл, який задається густиною

$$P_{\chi^2}(n)(x) = \begin{cases} 0; x < 0 \\ \frac{1}{\Gamma(\frac{r}{2}) 2^{\frac{r}{2}}} * x^{\frac{r}{2}-1} e^{-\frac{x}{2}}; x \geq 0 \end{cases}$$

Якщо $\chi^2_{\text{емпіричне}} > \chi^2_{\text{критичне}}$, то гіпотезу відхиляємо

Зауважимо, що критерій Пірсона можна застосовувати при умовах:

- n > 4
- в кожному класі не менше 10 спостережень, в іншому разі класи варто об’єднати
- 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 = r-s

Метод максимуму правдоподібності

Задача. Нехай x1, x2, . . . , xn - ряд незалежних спостережень проведених в однакових умовах над статичною змінною ξ, що має функцію розподілу $\mathcal{F}$, залежну від s невідомих параметрів α1, α2, . . . , αs

$$\xi : \mathcal{F}(x; (\alpha_1, \ldots, \alpha_s)) \qquad (1)$$

Завдання: оцінити невідомі параметри на основі вибірки.

Англійський статистик Р.Фішер у 1912 р. запропонував наступний метод оцінки невідомих параметрів. Якщо статистична змінна ξ абсолютно неперервна і має густину *P(x, α1, . . . , αs)*, то розглядатимемо таку функцію:

$$L(x_1, x_2, \ldots, x_n; \alpha_1, \alpha_2, \ldots \alpha_s) = \prod_{i=1}^{n} P(x_i, \alpha_1, \ldots, \alpha_s) \qquad (2)$$

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

$$L(x_1, x_2, \ldots, x_n; \alpha_1, \ldots, \alpha_s) = \prod_{i=1}^{n} P(x_i; \alpha_1, \ldots, \alpha_s) \qquad (3)$$

В обох випадках функцію *L = L(x1, .., xn; α1, . . . , αs)* називають **функцією правдоподібності**. Невідомі параметри оцінюємо з необхідної умови максумуму функції правдоподібності.
Очевидно, що необхідною умовою максимуму функцій багатьох змінних є зникнення 1-х частинних похідних від функції правдоподібності

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

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

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

тобто при ln *L* - точка досягає максимуму

Остання система рівнянь називається **системою рівнянь правдоподібності.**
Кожне розв’язання системи рівнянь правдоподібності, що залежить від вибіркових значень x1, ...xn називається **оцінкою максимальної правдоподібності** для α1, . . . , αs.
Таким чином, метод максимуму правдоподібності полягає в тому, що за оцінку невідомих параметрів приймаємо такі розв’язки системи (3), відносно α1, . . . , αk.

Зауваження: дуже частно 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 теоретичних і емнір-х моментів, скільки невідомих параметрів.

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

Статистичне оцінювання параметрів нормальної популяції

Означення 1. Оцінка θ називається *незміщеною оцінкою* параметра θ, якщо $E(\bar{\theta}) = \theta$, тобто $E(\bar{\theta} - \theta) = 0$. ■

Незміщеність оцінки означає, що похибка від заміни невідомого параметра θ на $\bar{\theta}$ не має систематичного характеру.

Означення 2. Незміщена оцінка θ параметра θ називається *ефективною, найкращою незміщеною оцінкою*, якщо

$$D(\theta) = \inf_{\bar{\theta}: E\bar{\theta} = \theta} D(\bar{\theta}). \quad ■ \quad (2)$$

Означення 3. Оцінка $\bar{\theta}_n$ параметра θ називається *спроможною, вагомою,* якщо:

$$\lim_{n \to \infty} P(|\bar{\theta}_n - \theta| < \varepsilon) = 1, \quad (3)$$

для будь-якого $\varepsilon > 0$, тобто, коли вони збігаються по ймовірності до θ при $n \to \infty$. ■

Означення 4. Послідовність оцінок $\bar{\theta}_1, \bar{\theta}_2, \ldots$ невідомого параметра θ називається *асимптотично незміщеною* послідовністю оцінок цього параметра, якщо

$$\lim_{n \to \infty} E(\bar{\theta}_n) = \theta, \text{ тобто } \lim_{n \to \infty} E(\overline{\theta_n} - \theta) = 0. \quad ■ \quad (4)$$

Оцінка невідомого математичного сподівання нормальної генеральної сукупності

Нехай x1, x2 …. xN вибірка з нормальної популяції ξ

Потрібно перевірити що математиче сподівання рівне заданому $E\xi = a$.

Емпіричне $t = \frac{\bar{x} - a}{s}\sqrt{n}$.

Критичне

Порівняння математичних сподівань двох нормальних популяцій

Нехай x1, x2 …. xN вибірка з нормальної популяції ξ
а y1, y2, …. yM вибірка з нормальної популяції n
Потрібно перевірити гіпотезу, що вибірки мають однакове математичне сподівання

$$H_0 : E\xi = E\eta .$$

Тоді емпіричне значення:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}}}\sqrt{\frac{mn}{m+n}}.$$

А число ступенів вільності: $d.f. = m + n - 2.$

Емпіричне значення шукаємо з таблички при вибраному альфа і за числом ступенів вільності.

Інтервал довіря невідомого математичного сподівання

При визначення критичної області для гіпотези на підставі означення рівня значущості одержуємо співвідношення

$$P\left\{\left|\frac{a-\overline{x}}{s}\sqrt{n}\right| < t_{\alpha/2}\right\} = 1-\alpha .$$

Звідси

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

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

$$\left(\overline{x} - \frac{s}{\sqrt{n}}t_{\alpha/2}; \overline{x} + \frac{s}{\sqrt{n}}t_{\alpha/2}\right) \qquad (5)$$

накриває невідоме сподівання a нормальної популяції. Цей інтервал називається *інтервалом довір'я, довірчим інтервалом.*

Гіпотезу про дисперсію нормального розподілу популяції

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

Нехай x1, x2 … xN ряд незалежних над незалежною нормальною статистичною змінною ξ.

Потрібно перевірити гіпотезу про те, що дисперсія нормальної популяції з

якої взята вибірка рівна заданій, тобто $H_0 : D\xi = \sigma^2$.

Для гіпотези будуть сприятливі ті випадки, коли значення статистики близьке до 1. Тому критична область складається із двох частин: дуже малих і дуже великих значень статистики .

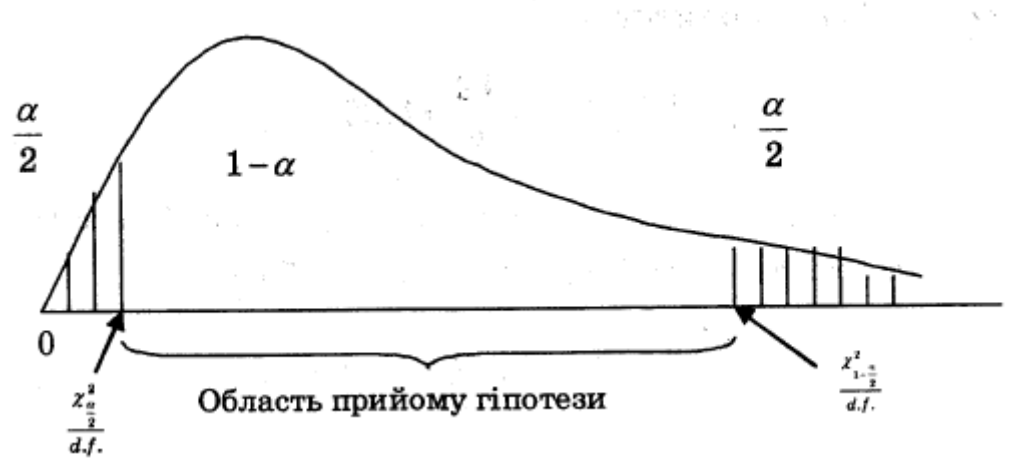
При рівні значущості α нижнім критичним
значенням є

$$\frac{\chi^2_{\frac{\alpha}{2}}}{d.f.} = \left(\frac{\chi^2}{\nu}\right)$$

, а верхнім

$$\frac{\chi^2_{1-\frac{\alpha}{2}}}{d.f.} = \left(\frac{\chi^2}{\nu}\right)$$

Емпіричне має бути в межах критичного, аби прийняти гіпотезу.



Емпіричне значення рахуємо як $\frac{s^2}{\sigma^2}$.

Інтервал довіря невідомої дисперсії нормального розподіл популяції

При визначенні критичної області статистики $\frac{s^2}{\sigma^2}$ використовуючи означення рівня значущості отримуємо співвідношення

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

Звідси безпосередньо знаходимо

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

Таким чином, з імовірністю $1-\alpha$ інтервал $\left(\frac{s^2}{\chi^2_{\frac{\alpha}{2}}/d.f.}; \frac{s^2}{\chi^2_{1-\frac{\alpha}{2}}/d.f.}\right)$ накриває невідоме значення дисперсії генеральної сукупності. Цей інтервал називається $100(1-\alpha)\%$ *інтервалом довіри, довірчим інтервалом* для дисперсії.

В ІНТЕРВАЛІ ЗНАЧЕННЯ МІСЦЯМИ ПОМІНЯТИ!

Порівняння дисперсій двох нормальних популяцій.

Нехай x1, x2 …. xN буде ряд незалежних спостережень над статистичною змінною ξ,
а y1, y2, …. yM - над змінною n.
Потрібно перевірити гіпотезу про те, що дисперсії двох нормальних популяцій , з яких взято вибірки x і y збігаються.

$$H_0 : D\xi = D\eta.$$

Розглянемо статистику

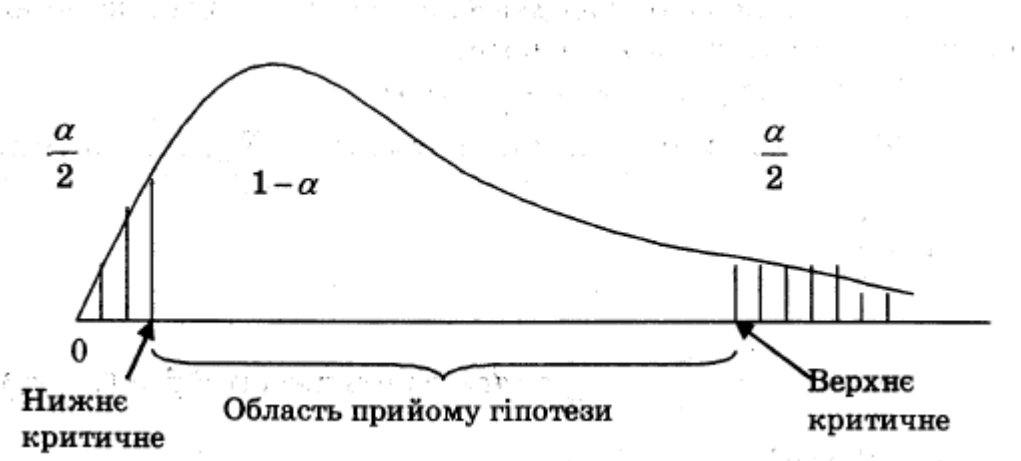
$$F = \frac{s_1^2}{s_2^2}.$$

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

$$\textbf{Нижнє критичне: } F_{\text{ниж.кр.}} = F_{\frac{\alpha}{2}}(\nu_1, \nu_2) = \frac{1}{F_{1-\frac{\alpha}{2}}(\nu_2, \nu_1)}.$$

$$\textbf{Верхнє критичне: } F_{\text{в.кр.}} = F_{1-\frac{\alpha}{2}}(\nu_1, \nu_2).$$

Гіпотезу приймаємо якщо емпіричне значення в межах критичного



Якщо в чисельнику більша з варіант то використовуємо лише верхню критичну область

Критерій Колмогорова

Нехай x1, x2 …. xN — вибірка з абсолютно неперервної популяції ξ
Потрібно перевірити гіпотезу про те що популяція керується неперервною функцією розподілу

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

Будуємо статистичний ряд
Знаходимо значення емпіричної функції розподілу.
Знаходимо значення гіпотетичної функції розподілу.
Знаходимо модуль різниці значень цих функцій у точках безмежно близьких зліва до точок цього ряду

варіаційний ряд

функція розподілу

твоя функція

-0 означає, що значення за індексом -1

x_i	$F_{25}(x_i)$	$F(x_i)$	$\lvert F_{25}(x_i)-F(x_i)\rvert$	$\lvert F_{25}(x_i-0)-F(x_i)\rvert$
-1.54	0.04	0.0618	0.0218	0.0618
-1.53	0.08	0.0630	0.0170	0.0230
-1.17	0.12	0.1210	0.0010	0.0410
-1.02	0.16	0.1539	0.0061	0.0339
-0.86	0.20	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
-0.30	0.32	0.3821	0.0621	0.1021
-0.26	0.36	0.3974	0.0374	0.0774
-0.17	0.40	0.4325	0.0325	0.0725
-0.06	0.44	0.4761	0.0361	0.0761

Знаходимо супремум останньої колонки

Емпіричне значення: $K = \sqrt{n} * D$

Критичне

α	0.10	0.05	0.01
$K_{крит}$	1.23	1.36	1.63

$$K(x)=\begin{cases}0, & x<0\\ \sum\limits_{k=-\infty}^{+\infty}(-1)^k e^{-2x^2k^2}, & x\geq 0\end{cases}.$$

x	$K(x)$
0.40	0.003
0.50	0.036
0.60	1.136
0.70	0.289
0.80	0.456
0.90	0.607
1.0	0.730
1.10	0.822
1.20	0.888
1.23	0.903
1.30	0.932
1.36	0.950
1.40	0.96
1.50	0.978
1.60	0.988
1.63	0.980
1.95	0.999
2.23	0.9999

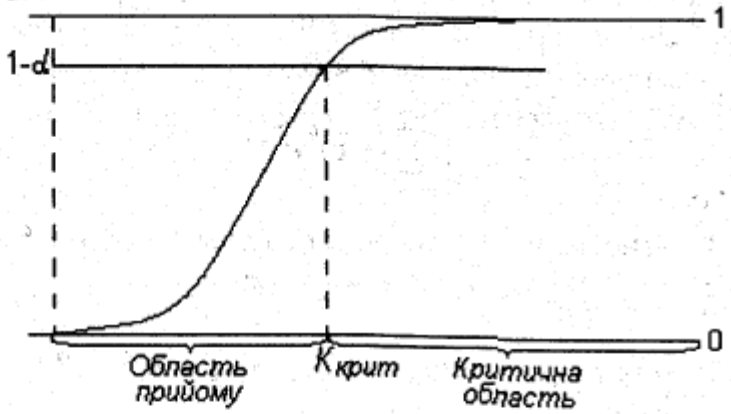


Рис. 2.

Критерій Смирнова

Нехай x1, x2 …. xN — вибірка з абсолютно неперервної популяції ξ,
а y1, y2, …. yM - з абсолютно неперервної популяції n.

Потрібно перевірити гіпотезу про те, що обидві генеральні сукупності однаково розподілені, тобто ξ і n стохастично еквівалентні

Застосовують при n, m >= 40

Вибираємо рівень значущості альфа.
Знаходимо значення функції розподілу обох вибірок .
Розглядаємо статистику

$$D_{mn} = \sup_{-\infty<x<+\infty} |F_m(x) - G_n(x)|.$$

Дві вибірки сортуємо
Якщо елементи однакові то в один рядок
Якщо елемент відсутній то значення емпіричної функції розподілу не міняємо

$x_{(i)}$	$y_{(i)}$	F_{10}	G_{10}	$\lvert F_{10} - G_{10} \rvert$
-1.1		0.1	0.0	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
	0.6	0.8	0.5	0.3

$$S = \sqrt{\frac{m * n}{m + n}} * D$$

Емпіричне:
Критичне значення беремо з Критерію Колмагорова.

Критерій знаків. Інтервал для прийняття рішень

Нахай є пари спостережень (x1, y1), (x2, y2) Перевіримо те що розподіли спостережень співпадають.

Однакові пари відкидаємо
Шукаємо різницю пар і рахуємо кількість додатніх знаків
Якщо n>= 16 і альфа = 0,05

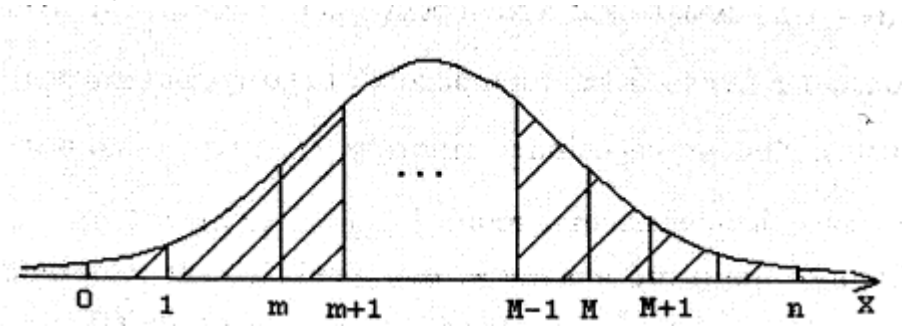
$$m = \frac{n}{2} - 0.98 * \sqrt{n}$$

$$M = \frac{n}{2} + 0.98 * \sqrt{n}$$

Інтервал прийняття рішень є розвязком нерівності

$$\sum_{s=0}^{m-1} \frac{C_n^s}{2^n} \le \frac{\alpha}{2}$$

$$\sum_{s=M+1}^{n} \frac{C_n^s}{2^n} \le \frac{\alpha}{2},$$



Гіпотеза про медіану

Нехай x1, x2 …. xN — вибірка з абсолютно неперервної популяції ξ.

Перевірити гіпотезу про те що медіана рівна заданому числу $H_0: Me = a.$

Для цього можна використати критерій знаків.
Будуємо пари (x1, a), (x2, a)... для Xi != a
Якщо Xi == a, то пару відкидаємо, а N зменшуємо

Маємо спостереження, від кожного віднімаємо медіану, рахуємо додатні знаки і перевіряємо чи попадає в межі.

Якщо n>= 16 і альфа = 0,05

$$m = \frac{n}{2} - 0.98 * \sqrt{n}$$

$$M = \frac{n}{2} + 0.98 * \sqrt{n}$$

Критерій Вілкінсона

Нехай $x_1, \dots, x_i, \dots, x_n$ - вибірка абсолютної неперервної популяції, що має неперервну ф-ію розподілу $\mathcal{F}(x)$, а $y_1, \dots, y_n$ - вибірка з абсолютно неперервних популяцій з ф-ією розподілу $\mathcal{G}(x)$.

Потрібно перевірити гіпотезу про те, що: популяції, з яких взято вибірки однаково розподілені: $\mathcal{F}(x) \equiv \mathcal{G}(x)$. (Порівняння за критерієм Смирнова)

2 вибірки з нормальної популяції, не обов'язково однакових розмірів
Сортуємо в один ряд
Рахуємо кількість У перед Х для кожного Х, і напаки

x x y x y x y x x x y y x x y y, то

$$W(y/x) = 1+2+3+3+3+5+5 = 22;$$

$$W(x/y) = 2+3+4+7+7+9+9 = 41;$$

для перевірки: 22+41 = m*n

Якщо альфа = 0,05 і m>=4, n>=4, m+n >=20, то межі такі:

$$E = \frac{m * n}{2}$$

$$\partial = \sqrt{\frac{m * n}{12} * (m + n + 1)}$$

$$m = E - 1.96 * \partial$$

$$M = E + 1.96 * \partial$$

Однофакторний варіансний аналіз

Варіансний аналіз (В.А.) – це метод статистичного дослідження впливу різних факторів на яке не-будь явище, який базується на порівнянні варіанс.

Він застосовується тоді, коли вибірки можна згрупувати.

Основна задача варіансного аналізу полягає в тому, щоб дослідити мінливості, викликані різними факторами.

Нехай дано m груп (класів, рівнів) незалежних спостережень над деякою одновимірною кількістю мінливою величиною.

$$m = 2,3,.....$$

Позначимо через x_{ij} - $j-$е спостережень в $i-$й групі, а через n_i - обсяг $i-$ї групи. Тоді всі m груп спостережень можна записати в такій таблиці:

$$x_{11}.....x_{1j}.....x_{1n_{11}}$$
$$.....$$
$$x_{i1}.....x_{ij}.....x_{in_{ii}}$$
$$.....$$
$$x_{m1}.....x_{mj}.....x_{mn_m}$$

Позначимо через N обсяг всіх спостережень

$$N = n_1 + ... + n_i + ... + n_m$$

$x_{i\bullet}$ - середнє спостереження в $i-$й групі

$$x_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_{ij}} x_{ij} \quad (i = 1, m)$$

$x_{\bullet\bullet}$ – загальне середнє всіх спостережень

$$x_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij}$$

$$S^2 = \frac{m-1}{N-1} S_1^2 + \frac{N-m}{N-1} S_2^2$$

Мінливість	Девіація	$d.f.$	Варіанса
між групами	$\sum_{i=1}^{m} n_i (x_{i0} - x_{\bullet\bullet})^2$	$m-1$	S_1^2
у групах	$\sum_{i-1}^{m} \sum_{j=1}^{n_i} (x_{ij} - x_{i\bullet})^2$	$N-m$	S_2^2
повна	$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - x_{\bullet\bullet})^2$	$N-1$	-

Для перевірки гіпотези можна використовувати критерій Фішера:

$$F = \frac{S_1^2}{S_2^2}$$

Якщо при заданому рівні значущості α і даних ступенях вільності $d.f. = (m-1, N-m)$ $F_{emn} > F_{kp}$, то гіпотезу однорідності відкидаємо. В протилежному випадку – гіпотеза не суперечить емпіричним даним.

Двофакторний варіансний аналіз

Нехай дані про деяку мінливу величину поділяються на m груп за ознакою A і n груп за ознакою B. Одержимо mn класифікаційних підгруп. Припустимо, що для кожної підгрупи проводиться лише одне спостереження.

Позначимо через x_{ij} - спостереження в $i-$й групі за ознакою A, та в $j-$й групі за ознакою B. Тоді всі mn спостережень можна записати в наступній таблиці

	1 ..... j ..... n
	x_{11} ..... x_{1j} ..... x_{1n}
	x_{i1} ..... x_{ij} ..... x_{in}
(*)	x_{m1} ..... x_{mj} ..... x_{mn}

Позначимо через $x_{i\bullet}$ – середнє $i-$ої групи за ознакою A ($i-$рядка)

$$x_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad (i = 1, m)$$

через $x_{\bullet j}$ - середнє $j-$ої групи за ознакою B

$$x_{\bullet j} = \frac{1}{m}\sum_{i=1}^{m} x_{ij}, \quad (j = \overline{1}, n)$$

через $x_{\bullet\bullet}$ – загальне середнє всіх спостережень

$$x_{\bullet\bullet} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$$

Кожна з цих дивіацій має своє число $d.f.$ (ступенів вільності)

$mn-1 \quad m-1 \quad n-1 \quad mn-(m+n-1) = (m-1)(n-1)$

$$S_A^2 = \frac{1}{m-1} n \sum_{i=1}^{m} (x_{i\bullet} - x_{\bullet\bullet})^2$$

$$S_B^2 = \frac{1}{n-1} \cdot m \sum_{j=1}^{n} (x_{\bullet j} - x_{\bullet\bullet})^2$$

$$S_r^2 = \frac{1}{(m-1)\quad(n-1)} \sum_{i=1}^{m} \quad \sum_{j=1}^{n} (x_{ij} - x_{i\bullet} - x_{\bullet j} + x_{\bullet\bullet})^2$$

Мінливість	Девіація	$d.f.$	Варіанса
між групами A	$n\sum_{i=1}^{m} (x_{i\bullet} - x_{\bullet\bullet})^2$	$m-1$	S_A^2
між групами B	$m\sum_{j=1}^{n} (x_{\bullet j} - x_{\bullet\bullet})^2$	$n-1$	S_B^2
Залишкова	$\sum_{i=1}^{m}\sum_{j=1}^{n} (x_{ij} - x_{i\bullet} - x_{j\bullet} + x_{\bullet\bullet})^2$	$(m-1)(n-1)$	S_r^2
Повна	$\sum_{i=1}^{m}\sum_{j=1}^{n} (x_{ij} - x_{\bullet\bullet})^2$	$mn-1$	-

Доводемо за фішером $F_A = \frac{\hat{S}_A^2}{S_r^2}$

Трифакторний варіансний аналіз

Нехай дані про деяку мінливу величину класифікуються за трьома ознаками:

- на m груп за ознакою A,
- на n груп за ознакою B
- на l груп за ознакою C.

Дістаємо mnl класифікаційних підгруп.

Припустимо, що в кожній групі є тільки одне спостереження.

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

Всі mnl спостережень можна розмістити в l таблиць 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:

x_(ij•) ;x_(i•k) ;x_(•jk) ; x_(i••) ;x_(•j•);x_(••k) ;x_(•••) (*)

де наприклад,

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

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

$$x_{\cdots} = \frac{1}{mnl}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l} x_{ijk}$$

$$d.f. = m-1, \qquad d.f. = n-1, \qquad d.f. = l-1;$$
$$d.f. = mn - m - n + 1 = (m-1)(n-1),$$
$$d.f. = ml - m - l + 1 = (m-1)(l-1),$$
$$d.f. = nl - n - l + 1 = (n-1)(l-1);$$
$$d.f. = mnl - mn - ml - nl + m + n + l - 1 = (m-1)(n-1)(l-1).$$

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

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

$$S_{ABC}^2 = S_r^2 = \frac{1}{(m-1)(n-1)(l-1)}\cdot$$

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

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

Аналогічно записують інші варіанси: $\hat{S}_B^2, \hat{S}_C^2, \hat{S}_{AC}^2, \hat{S}_{BC}^2$.

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 розглянемо такі статистики:

$$F_A = \frac{S_A^2}{S_r^2}, \qquad F_{AB} = \frac{S_{AB}^2}{S_r^2} \qquad (***)$$

Мінливість	Девіація	d.f.	Варіанса
між групами A	$nl\sum_{i=1}^{m}(x_{i\cdot\cdot}-x_{\cdot\cdot\cdot})^2$	$m-1$	S_A^2
між групами B	$ml\sum_{j=1}^{n}(x_{\cdot j\cdot}-x_{\cdot\cdot\cdot})^2$	$n-1$	S_B^2
між групами C	$mn\sum_{k=1}^{l}(x_{\cdot\cdot k}-x_{\cdot\cdot\cdot})^2$	$l-1$	S_C^2
взаємодія AB	$l\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij\cdot}-x_{i\cdot\cdot}-x_{\cdot j\cdot}+x_{\cdot\cdot\cdot})^2$	$(m-1)(n-1)$	S_{AB}^2
взаємодія AC	$n\sum_{i=1}^{m}\sum_{k=1}^{l}(x_{i\cdot k}-x_{i\cdot\cdot}-x_{\cdot\cdot k}+x_{\cdot\cdot\cdot})^2$	$(m-1)(l-1)$	S_{AC}^2
взаємодія BC	$m\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{\cdot jk}-x_{\cdot j\cdot}-x_{\cdot\cdot k}+x_{\cdot\cdot\cdot})^2$	$(n-1)(l-1)$	S_{BC}^2
залишкове	$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{ijk}-x_{ij\cdot}-x_{i\cdot k}-x_{\cdot jk}+x_{i\cdot\cdot}+x_{\cdot j\cdot}+x_{\cdot\cdot k}-x_{\cdot\cdot\cdot})^2$	$(m-1)(n-1)\cdot\cdot(l-1)$	S_r^2
Повне	$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}(x_{ijk}-x_{\cdot\cdot\cdot})^2$	$mnl-1$	

Варіансний аналіз за схемою латинського квадрата

Латинським квадратом порядку m називається таке розміщення 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.

Нехай деяка мінлива величина поділяється на **m** груп за кожною з 3 ознак: A, B, C. Отримуємо m^3 класифікаційний підгруп. Припустимо, що проводиться по-одному спостережені в m^2 класифікаційний підгрупах. Ці спостереження проводимо за планом навмання вибраному латинського квадрату порядку m. Позначимо через χ_{ijk} спостереження в i-тій групі за ознакою A, j-ій групі за ознакою B, k-ій за ознакою C.
Ці m^2 спостережень розташуємо в m рядках і m стовпчиках, нашого 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
χ_{113}	χ_{121}	χ_{134}	χ_{142}			χ_{113}	χ_{121}	χ_{134}	χ_{142}		
χ_{214}	χ_{222}	χ_{233}	χ_{241}			χ_{211}	χ_{224}	χ_{232}	χ_{243}		
χ_{312}	χ_{323}	χ_{331}	χ_{344}			χ_{312}	χ_{323}	χ_{331}	χ_{344}		
χ_{411}	χ_{424}	χ_{432}	χ_{443}			χ_{414}	χ_{422}	χ_{433}	χ_{441}		

Упорядковуємо за алфавітом

H L M S

1 2 3 4

Позначимо через $\chi_{i..}$ середнє і рядка $\chi_{i..} = \frac{1}{m}\sum_{j=1}^{m}\chi_{ijk}, i = \overline{1,m}$

Через $\chi_{.j.}$ середнє j стовпчика: $\chi_{.j.} = \frac{1}{m}\sum_{i=1}^{m}\chi_{ijk}, j = \overline{1,m}$

Через $\chi_{..k}$ середнє k рівня ознаки C: $\chi_{..k} = \frac{1}{m}\sum_{k=1}^{m}\chi_{ijk}, k = \overline{1,m}$

Наприклад, $\chi_{..1} = \frac{\chi_{121}+\chi_{241}+\chi_{331}+\chi_{411}}{4}$

Позначимо через $\chi_{...}$ середнє всіх спостережень

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

Повна мінливість всіх спостережень виражається рівністю

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

, яку можна представити суму таких чотирьох девіацій

$$=\mathrm{m}\sum_{i=1}^{m}(\mathcal{X}_{i..}-\mathcal{X}_{...})^{2}+\mathrm{m}\sum_{j=1}^{m}(\mathcal{X}_{.j.}-\mathcal{X}_{...})^{2}+\mathrm{m}$$
$$\sum_{k=1}^{m}(\mathcal{X}_{..k}-\mathcal{X}_{...})^{2}+\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathcal{X}_{ijk}-\mathcal{X}_{...}-\mathcal{X}_{.j.}-\mathcal{X}_{..k}+2\mathcal{X}_{...})^{2} \quad (1)$$

Тотожність (1) очевидна на основі тотожності

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

Таким чином тотожність (1) вказує на те, що повна девіація розкладається на 4 девіації: девіація між груп ознаки А, девіація між груп ознаки В, девіація між груп ознаки С і залишкова девіація.

Кожна, з яких має своє число вільності. Повна має m^2-1, кожна між групами має m-1 ступінь вільності(для кожної групи)

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

Звідси виходить, що порядок латинського квадрату m може бути найменше 3 (m≥3)

Очевидно, що між ступенями вільності має місце тотожність
m^2-1 = (m-1) + (m-1) + (m-1) + ((m^2-(3m-2)) = (m-1)(m-2))

Припустимо, що спостереження X_{ijk} однорідні і взяті з нормальної генеральної сукупності, тоді варіанси

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

$$S_B^2 = \frac{m}{(m-1)} \sum_{j=1}^{m} (X_{.j.} - X_{...})^2,$$

....

є незміщеними і незалежними оцінками дисперсій нормальної генеральної сукупності. Звідси виходить, що для перевірки гіпотез однорідності можна використати критерій Фішера F.

Наприклад, статистика F
Обчислення при варіансному аналізі латинського квадрату, записуємо у вигляді такої таблиці

Мінливість	Девіація	d. f.	Варіанса
Між гр. A	$m\sum_{i=1}^{m}(X_{i..}-X_{...})^2$	m-1	S_A^2
Між гр. B	$m\sum_{j=1}^{m}(X_{.j.}-X_{...})^2$	m-1	S_B^2
Між гр. C	$m\sum_{k=1}^{m}(X_{..k}-X_{...})^2$	m-1	S_C^2
Залишкова	$\sum_{k=1}^{m}\sum_{i=1}^{m}(X_{ijk}-X_{i..}-X_{.j.}-X_{..k}+2X_{...})^2$	(m-1)(m-2)	S_3^2
Повна	$\sum_{j=1}^{m}\sum_{j=1}^{m}(X_{ijk}-X_{...})^2$	m^2-1	

Зазначимо, що варіансний аналіз за планом є часто випадком неповного трифакторного варіансного аналізу, коли число рівнів кожної ознаки є однаковою.

Кореляційний аналіз (коваріація, кореляція, регресія)

Коваріація

Розглянемо довільний двовимірний випадковий вектор з компонентами ξ, η, для яких відомі їх сподівання і дисперсії E(ξ), E(η), D(ξ), D(η): і нехай α = const деяка стала, α > 0.

Знайти зв'язок між компонентам ξ та η ?

Коваріацією між випадковими змінними ξ та η називається сподівання добутку відхилень цих змінних від своїх сподівань

$$cov(\xi,\eta) = E[(\xi - E\xi)(\eta - E\eta)]$$

коваріація симетрична відносно змінних

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

Якщо ξ та η – незалежні, то

$$cov(\xi,\eta) = 0$$

Доведення

$$E[(\xi - E\xi)(\eta - E\eta)] = E(\xi - E\xi)E(\eta - E\eta) = (E\xi - E\xi)(E\eta - E\eta) = 0$$

Оскільки коваріація між незалежними випадковими змінними = 0, то для залежних випадкових змінних коваріацію можна прийняти за міру залежності.

Кореляцією між випадковими змінними ξ та η називають відношення коваріції між змінними до стандартів цих змінних.

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

$$\rho(\xi,\eta) = \rho(\eta,\xi)$$

Якщо ξ та η незалежні випадкові змінні, то ρ(ξ, η) = 0.

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

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

Не симетрчина

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

Якщо ξ і η – незалежні випадкові змінні і кореляція між ними дорівнює нулю, то

$$R(\xi/\eta) = R(\eta/\xi) = \alpha$$

Пряма регресія

6. **Пряма регресія.** Нехай на двовимірним впорядкованим вектором $(\xi; \eta)$ проведено в однакових умовах n незалежних спостережень : $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$.

Знайдено середні (незрозуміло) компонентів вектора :

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

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,\ S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{2}$$

Аналогічно до рівностей (1), (2), (9), (10) знайдемо вибіркову коверіацію між компонентами

$$C_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

Вибіркову кореляцію між компонентами

$$r_{12} = \frac{c_{12}}{s_1 s_2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

Вибіркову регресію другої компоненти відносно першої

$$b_{21} = r_{12}\frac{s_2}{s_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

Вибіркову регресію першої компоненти відносно другої

$$b_{12} = r_{12}\frac{s_1}{s_2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

Випадкова варіанта випадкової змінної $\xi = -a\xi + \eta$ дорівнює

$$S^2_{\xi=-a\xi+\eta} = a^2 S_1^2 - 2aC_{12} + S_2^2 = a^2 S_1^2 - 2ar_{12}S_1\ S_2\ + S_2^2$$

і буде найменшою при $a = r_{12}\frac{s_2}{s_1}$. Це мінімальне вибіркове (незрозуміло) дорівнює

$$S^2_{\xi=-a\xi+\eta} = (1 - r_{12}^2)S_2^2 \tag{7}$$

Отже, завжди $-1 \le r_{12} \le 1$. Якщо $r_{12} = \pm 1$, то з огляду на нерівність всі $y_i - ax_i$ набувають однакових значень $\boldsymbol{b}$, тобто всі спостереженні точки з координатами $(x_i; y_i)$, $(i = \overline{1,n})$ лежать на прямій $y = ax + b$. Якщо ж точки $(x_i; y_i)$, $(i = \overline{1,n})$ не лежать на одній прямій, то через пункт $(\bar{x}, \bar{y})$. Рівномірно пряму з кутовим коефіцієнтом, визначеним за формулою (5).

Ця пряма

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

Називають емпіричною прямою регресії другої компоненти відносно першої. Покажемо, що сума квадратів відхилень точок у напрямі y від прямої є найменшою.

Дійсно, сума квадратів відхилень точок у напрямі від довільної прямої $y = ax + b$, дорівнює

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

Оскільки система рівнянь $f'_a = 0$, $f'_b = 0$ має єдиний розв'язок $a = b_{21}$, $b = \bar{y} - b_{21}$, причому

$$f''_{a^2} = 2\sum_{i=1}^{n} x_i > 0,\ f'_{a^2}f'_{b^2} - (f''_{ab})^2 = \varphi n\Big\{\sum_{i=1}^{n} x_i^2 - n\bar{x}\Big\} > 0,$$

то $f(a,b)$ набуває найменшого значення для прямої $y = b_{21}x + \bar{y} - b_{21}\bar{x}$, яка і являє собою емпіричну пряму регресію.

Таким же шляхом можна показати, що емпірична пряма регресії першої компоненти відносно другої ()

Мінімізує суму квадратів відхилень точок у напрямі від довільної прямої ().

Зауваження 1. Суми, що виступають у формулах (20) – (25), можна обчислювати різними способами, які себе взаємно контролюють.

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

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

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

Зауваження 2. Кореляція у математичній статистиці, вказує на зв'язок між явищами, який виникає тоді, коли одне з них належить до сукупності причин, що (незрозуміло) інші явище або коли є щільні причини, що впливають на ці явища.

Кореляція вищих порядків

Нехай випадкові змінні $\mathfrak{z}_1, \ldots, \mathfrak{z}_n$ будуть найзагальнішими компонентами n-вимірного випадкового вектора $(\mathfrak{z}_1, \ldots, \mathfrak{z}_n)$, $(n = 3, 4, \ldots)$, що описує деякий випадковий процес. Щоб знайти кореляцію між двома з цих змінних при сталих значеннях (незрозуміло), позначимо вибіркову кореляцію змінних між $\mathfrak{z}_1$ та $\mathfrak{z}_2$ при сталих значеннях $\mathfrak{z}_3, \ldots, \mathfrak{z}_m$ через $r_{12[3,4,\ldots,m}$ (незрозуміло), відповідні сталим значенням змінних, назвемо німими, а число німих індексів – порядком частинної кореляції .

Отже, $r_{12[3,4,..,m}$ – частинна вибіркова кореляція (m-2)-го порядку порядку між ξ_1 та ξ_2. Визначимо кореляцію (m-2)-го порядку через три кореляції (m-3)-го за рекурентною формулою :

$$r_{12[3,4,..,m} = \frac{r_{12[3,4,..,(m-1)} - r_{1m[3,4,..,(m-1)} r_{2m[3,4,..,(m-1)}}{\sqrt{[1-r^2{}_{1m[3,4,..,(m-1)}][1-r^2{}_{2m[3,4,..,(m-1)}]}} \qquad (*)$$

Аналогічно можна записати рекуренті формули дія інших пар випадкових змінних. Наприклад,

$$r_{1m[3,4,..,(m-1)} = \frac{r_{1m[3,4,..,(m-1)} - r_{1(m-1)[3,4,..,(m-2)} r_{m(m-1)[3,4,..,(m-2)}}{\sqrt{[1-r^2{}_{1(m-1)[3,4,..,(m-2)}][1-r^2{}_{m(m-1)[3,4,..,(m-2)}]}}$$

Якщо випадковий процес описується трьома випадковими змінними (), () та (), то частинна кореляція 1-го порядку залишеться через три кореляції нульового порядку. Наприклад,

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

Відзначимо, що перестановка індексів двох нефіскованих змінних у формулі (*) не змінює вартість кореляції, а перестановка індексів фіксованих змінних змінює її. Наприклад,

$r_{12[34} = r_{12[34} \neq r_{12[43}$.

Варіанси та стандарти вищих порядків

Позначимо вибіркову варіансу для $\mathfrak{z}_1$ при сталих $\mathfrak{z}_2, \ldots, \mathfrak{z}_m$ через $S^2_{1[23\ldots m}$. Тут *23…m* – німі індекси. Назвемо цю частинну варіансу **варіансою** (m-1)-го порядку. Варіансу (m-1)-го порядку визначимо рекурентно – через варіансу (m-2)-го та кореляцію (m-2)-го порядку за формулою

$$S^2_{1[23\ldots m} = S^2_{1[23\ldots(m-1)}[1-r^2_{1m[23\ldots(m-1)}] \quad (1)$$

Звідси

$$S^2_{1[23\ldots(m-1)} = S^2_{1[23\ldots(m-2)}[1-r^2_{1(m-1)[23\ldots(m-2)}]$$

…

$$S^2_{1[23} = S^2_{1[2}[1-r^2_{13[2}]$$

$$S^2_{1[2} = S^2_1[1-r^2_{12}]$$

Отже,

$$S^2_{1[23\ldots m} = S^2_1[1-r^2_{12}]\,[1-r^2_{13[2}]\ldots[1-r^2_{1m[23\ldots(m-1)}] \quad (2)$$

Таким чином, варіанса (m-1)-го виражається через варіансу нульового порядку та кореляції від нульових до (m-2)-го порядку.

Аналогічно одержуємо варіанси вищих порядків для інших випадкових змінних без (незрозуміло) на решту змінних (коли інші змінні фіксовані).

Наприклад,

$$S^2_{3[12} = S^2_1[1-r^2_{13}]\,[1-r^2_{23[1}]$$

Стандарти

Арифметичний квадратний корінь із варіанси відповідного порядку називають стандартом того же порядку. Наприклад, S_1 – стандарт (m-1)-го порядку, () – стандарт нульового порядку.

Запишемо співвідношення (2) у вигляді

$$S^2_{1[23\ldots m} = S^2_1[1\text{-}R^2_{1(23\ldots m)}], \qquad (3)$$

Де $R_{1(23\ldots m)}$ – багатократна вибіркова кореляція (m-1)-го порядку між $\mathfrak{z}_1$ і ($\mathfrak{z}_2$, …, $\mathfrak{z}_m$). Із співвідношення (2), (3) випливає, що (m-1)-кратна кореляція виражається через частинні кореляції від нульового до (m-2)-го порядку співвідношеннями.

1- $R_{1(23\ldots m)} = [1 - r^2_{12}]\,[1 - r^2_{13[2]}] \ldots [1 - r^2_{1m[23\ldots(m-1)}]$.

Аналогічно представляємо багатократну кореляцію кожної іншої змінної відносно решти змінних. Очевидно, що $R_{1(2)} = r_{12}$.