Розділ 4. Перевірка наявності зв'язку між змінними

У статистичному аналізі зазвичай розрізняють наступні види зв'язків між факторами (змінними):

– функціональні;

– стохастичні;

– статистичні.

Функціональний зв'язок – зв'язок між змінними, при якому кожному значенню однієї величини відповідає суворо певне значення іншої, тобто $Y=F(x_1,x_2,\ldots,x_n)$. Дослідженням таких зв'язків статистика не займається.

Стохастичний зв'язок відповідає ситуації, коли зміна значення однієї змінної призводить до зміни закону розподілу іншої. Для дискретного випадку це означає, що кожному значенню однієї змінної відповідає набір значень іншої, причому кожне значення має свою ймовірність реалізації. Тут такі методи не описуються. В практичних дослідженнях найбільш відомим видом таких зв'язків є Марківські ланцюги.

Статистичний зв'язок означає, що значення однієї змінної змінюється в середньому залежно від того, яких значень набуває інша.

Дуже часто розглядається як функціональна залежність з випадковою похибкою, тобто $Y=F(x_1,x_2,\ldots,x_n)+\varepsilon$, (4.1)

де $F(x_1,x_2,\ldots,x_n)$ – функція, що описує залежність Y від сукупності незалежних змінних $x_1,x_2,\ldots,x_n$, а ε – деяка випадкова похибка. Відомо, що сума константи і випадкової величини є випадковою величиною. В зв'язку з цим значення Y, що розраховані за вказаною формулою, будуть в наслідок додавання випадкової величини ε також випадковими величинами.

В даному розділі розглядуються методи, що призначені для аналізу (але не опису) статистичних зв'язків.

4.1. Вибір методу перевірки наявності зв'язку

Ці методи призначені для перевірки гіпотез про наявність зв'язків між змінними. Вибір методу залежить від шкал вимірювання, в яких вимірюються аналізовані змінні та від їх кількості (див. таблицю. 4.1).

Кореляційний аналіз (параметричний та непараметричний) використовують в тих випадках, коли змінні вимірюються в шкалах відношень, інтервалів або порядку.

Дисперсійний аналіз використовують, якщо залежна змінна вимірюється в шкалах відношень, інтервалів або порядку, а впливаючі змінні мають нечислову природу (шкала найменувань).

Аналіз таблиць зв'язаності використовується, коли впливаючі змінні мають нечислову природу (шкала найменувань), а залежна змінна відображає кількість спостережень (% або частку від загальної кількості), для яких ознака присутня або відсутня.

Таблиця 4.1. Вибір методу аналізу зв'язків між змінними.

Загальна кількість змінних	**Шкали виміру**		**Закон розподілу**	**Метод**
	впливаючих змінних	**залежної змінної**		
Дві	Інтервалів або відношень		Нормальний	Парам. кореляція Пірсона
			Відрізняється від нормального	Непарам. кореляція Спірмена
	Хоч би одна шкала порядку		–	Непарам. кореляція Спірмена або Кендалла
Три і більше	Порядку		–	Конкордація
Дві і більше	Найменувань	Інтервалів або відношень	Нормальний для залежної змінної	Парам. дисперсійний аналіз (критерій Фішера)
			Порядку для залежної змінної	Непарам.дисперсійний аналіз (критерій Зігеля і Тьюкі)
			Відрізняється від нормального для залежної змінної	Непарам. дисперсійний аналіз (критерій Зігеля і Тьюкі)
Три і більше		Порядку	–	Багатовимірний непарам. диспер. аналіз Фрідмана
		Інтервалів або відношень	Відрізняється від нормального для залежної змінної	Багатовимірний непарам. дисперсійний аналіз Фрідмана
			Нормальний для залежної змінної	Багатовимірний параметр. дисперсійний аналіз
Дві	Обидві шкали найменувань, одна виду «є/немає», інша має тільки два значення		–	Чотирьохліткові таблиці зв'язаності (сполучень)
Дві	Обидві шкали найменування, одна виду «є/немає», інша має *K* значень		–	Таблиці зв'язаності (сполучень) виду 2**K*
	Обидві найменування, одна має *K* значень рівнів, інша – *L*		–	Таблиці зв'язаності (сполучень) виду *K***L*

4.2. Кореляційний аналіз

Кореляційний зв'язок є окремим випадком статистичного зв'язку $M(Y/X=x)=\hat{y}(x)$, тобто математичне очікування змінної Y, за умови, що випадкова величина X набуває значення x.

4.2.1. Параметрична кореляція (коефіцієнт кореляції Пірсона)

Передумови. Всі спостереження взаємно незалежні. Спостереження мають нормальний закон розподілу.

Опис методу.

Значення коефіцієнта кореляції обчислюється за формулою:

$$r_{xy} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2 \sum_{i=1}^{N}(Y_i - \overline{Y})^2}} \qquad (4.2)$$

Коефіцієнт кореляції показує тісноту лінійного зв'язку між двома вибірками випадкових величин. Його значення змінюється від –1, що відповідає зворотному зв'язку, до +1, що є відповідним прямо пропорційному зв'язку (значення *0* означає відсутність зв'язку).

Залежність коефіцієнта кореляції

Оскільки маємо справу з випадковими величинами, однієї величини коефіцієнта парної кореляції для висновку про статистичне значимість зв'язку недостатньо.

Необхідно перевірити, чи значущо він відрізняється від нуля. Це можна зробити за допомогою критерію Стьюдента. Фактично перевіряється гіпотеза про рівність коефіцієнта кореляції нулю. Для цього розраховується критеріальне значення за формулою:

$$t_{розр} = \frac{r\sqrt{(N-2)}}{\sqrt{(1-r^2)}}, \qquad (4.3)$$

де N – кількість спостережень, за якими розраховується коефіцієнт кореляції; r – значення коефіцієнта кореляції.

Якщо розрахункове значення $t_{розр}$ більше табличного коефіцієнта Стьюдента, взятого з N-2 степенями свободи – $t_{N-2,p}$, то нульова гіпотеза відкидається. Це означає, що коефіцієнт кореляції значимо відрізняється від нуля (з вибраним рівнем значимості).

Напівширина довірчого інтервалу Δ для коефіцієнта кореляції r визначається за формулою:

$$\Delta = \frac{t_{N-2,p}\{1-r^2)}{\sqrt{N}}. \qquad (4.4)$$

Примітка. При використанні кореляційного аналізу треба пам'ятати, що коефіцієнт кореляції показує тісноту тільки лінійного зв'язку. Тому у разі, коли залежності більш складні, ніж лінійні, коефіцієнт кореляції показуватиме відсутність зв'язку. Тому для визначення складних залежностей

між змінними використовуються інші статистичні методи, найчастіше і ефективно – регресійний аналіз.

4.2.2. Часткова кореляція

Для того, щоб вплив кореляційного зв'язку між змінними очистити від можливого впливу третьої, введено поняття часткової кореляції. По ній коефіцієнт між двома змінними X і Z визначається за формулою:

$$r = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}, \qquad (4.5)$$

де r_{12}, r_{13}, r_{23} – коефіцієнти парної кореляції між змінними X і Y, X і Z, Y і Z відповідно. При використанні часткового коефіцієнта кореляції необхідно пам'ятати:

– взаємовпливаючих змінних може бути не три, а скільки завгодно;

– можна не знати про всіх взаємовпливаючих змінних;

– деякі автори стверджують, що для коректного використання часткового коефіцієнта кореляції необхідна наявність багатовимірного нормального закону розподілу, проте перевірити виконання цієї передумови практично нереально.

4.2.3. Рангова кореляція (РК)

РК є аналогом парної кореляції для тих випадків, коли величини, наявність зв'язку між якими потрібно перевірити, представлені не в шкалі відношень, а в якій-небудь іншій. Найчастіше така ситуація виникає, якщо ми маємо справу з суб'єктивними оцінками об'єктивних явищ, які не можна виміряти, тобто з експертними оцінками. Крім того, РК використовується в тих випадках, коли закон розподілу змінних, що вивчаються, не є нормальним.

Коефіцієнти кореляції називають ранговими, оскільки перед обчисленнями значення змінних перетворюють на ранги. Для цього наявні значення змінних розташовують в ранжируваний ряд (впорядкований за величиною). Значення в початковому стані можуть бути таким ранжируваним рядом. Потім кожному значенню надається ранг від 1 до N, де N – кількість аналізованих об'єктів. В тому випадку, якщо декілька елементів мають один і той же ранг, то кожному з них надається середнє від займаних ними місць.

Припущення :

– всі спостереження взаємно незалежні;

– всі значення спостережень вибрані з однієї і тієї ж двовимірної генеральної сукупності, тобто X і Y однаково розподілені.

Існує декілька різних способів обчислення коефіцієнтів РК. Найчастіше використовують коефіцієнт кореляції Спірмена (r, інколи позначається r_s) і коефіцієнт Кендалла (τ).

Коефіцієнт РК Спірмена

Коефіцієнт Спірмена обчислюється за формулою:

$$r_s = \rho(A,B) = 1 - \frac{6\sum_{i=1}^{n}(R_{1i} - R_{2i})^2}{n^3 - n}, \quad (4.6)$$

де R_{1i} і R_{2i} – ранги *i*-го об'єкту для кожної з порівнюваних змінних. Значення r_s не залежить від способу впорядкування рангів.

Цей коефіцієнт є повним аналогом коефіцієнта парної кореляції. Після перетворення його можна представити в такому вигляді:

$$\rho(A,B) = 1 - \frac{\sum_{i=1}^{n}(R_{1i} - \frac{n+1}{2})(R_{2i} - \frac{n+1}{2})}{\frac{1}{12}(n^3 - n)}. \quad (4.7)$$

За наявності співпадаючих значень (зв'язок) знаменник зменшується на величину:

$$\frac{\sum_{i=1}^{L1}(T_{1i}^3 - T_{1i}) + \sum_{j=1}^{L2}(T_{2j}^3 - T_{2j})}{2}, \quad (4.8)$$

де L_1 і L_2 – кількість зв'язок в T_{1i} і T_{2j}; T_{1i}, T_{2j} – розміри зв'язок (кількість елементів в них).

Для перевірки значущості коефіцієнта РК Спірмена при n>9 можна скористатися критерієм Стьюдента (як для звичайного коефіцієнта парної кореляції). Перевірка значущості для загального випадку виконується за допомогою спеціальних таблиць.

Коефіцієнт кореляції Кендалла

Коефіцієнт кореляції Кендалла обчислюється за формулою:

$$\tau = 1 - \frac{4}{n(n-1)}Q, \quad (4.9)$$

де n – кількість спостережень; Q – кількість неузгоджених пар (X_j, Y_j) і (X_i, Y_i) для всіх комбінацій i і j. Пари називаються неузгодженими, якщо для них виконується наступна умова:

$$sign(X_j - Y_j)\,sign(X_i - Y_i) = -1, \quad (4.10)$$

де *sign* – означає знак. Ця функція набуває значення +1 для додатного числа і –1 для від'ємного. Іншими словами, наведена умова означає, що збільшення X приводить до зменшення Y, і навпаки.

Зауваження до коефіцієнтів Кендалла і Спірмена

1. Розраховані для одних і тих же даних значення коефіцієнтів Кендалла і Спірмена не збігаються, окрім крайніх значень (–1, 0, 1).

2. Асимптотично ці коефіцієнти збігаються.

Конкордація

У тих випадках, коли необхідне порівняння не двох змінних, а більшої кількості (наприклад, при з'ясуванні узгодженості думок групи експертів), використовується коефіцієнт конкордації, що запропонований Кендаллом:

$$W = \frac{12}{m^2(n^3-n)} \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \left(R_{ij} - \frac{n+1}{2} \right) \right)^2, \qquad (4.11)$$

де n – кількість аналізованих об'єктів;

m – кількість експертів;

R_{ij} – ранг j-го об'єкту, який наданий йому i-тим експертом.

Треба звернути увагу на різницю в значеннях коефіцієнта конкордації від коефіцієнта кореляції. Якщо думки експертів повністю протилежні, коефіцієнт конкордації дорівнює нулю (W=0), але коефіцієнт кореляції в цьому випадку буде дорівнювати –1.

За наявності зв'язок (однакових значень) формула набуває наступного вигляду:

$$W = \frac{12}{\frac{1}{12} m^2(n^3-n) - m\sum_{j=1}^{m} T_j} \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \left(R_{ij} - \frac{n+1}{2} \right) \right)^2, \qquad (4.12)$$

де $T_j = \frac{\sum_{i=1}^{Li} (n_i^2 - n_i)}{12}$. При цьому L_i – кількість зв'язок, n_i – кількість елементів в i-й зв'язці для j-го експерта.

Значущість коефіцієнта конкордації при малій кількості експертів перевірити складно. Для малих значень існують неповні таблиці, наприклад таблиці, фрагмент якої наведено в таблиці 4.3.

Таблиця 4.3. Критичні значення коефіцієнта конкордації.

n=3; m=10		n=5; m=3	
50	0.092	56	0.096
62	0.046	62	0.056
104	0.0034	78	0.053
126	0.0008	86	0.0009

Для отримання критичного значення коефіцієнта конкордації необхідно взяти з таблиці значення S і підставити у формулу:

$$12\frac{S}{m^2(n^3-n)}.$$

Якщо ж кількість експертів більше 7, то можливе порівняння значення вираження $n(m-1)W$ з табличним значенням, що розподілене по χ^2 з n–1 степенями свободи.

4.3. Дисперсійний аналіз (ДА)

У задачах, які вирішуються з використанням ДА, відгук має числову природу. присутній об'єкт числової природи. На цей відгук діють декілька змінних, що мають номінальну природу. Тут розглядається модель

$$(X_i - \bar{\bar{X}}) = \alpha_i - \beta_i + \varepsilon_i, \qquad (4.13)$$

тобто розсіювання дорівнює виміру, що залежить від одного фактора α_i, плюс розсіювання, залежне від другого фактора β_i, плюс випадкова помилка ε_i. Тоді спільне розсіювання складається з декількох компонент:

$$S^2 = S_a^2 + S_b^2 + S_c^2.$$

Виділивши відповідні компоненти, за допомогою критерію Фішера можна визначити їх значущість.

4.3.1. Параметричний ДА

Однофакторна задача.

Для найбільш простого випадку таблиця початкових даних має вигляд, що представлено в таблиці 4.4.

Таблиця 4.4. Загальний вигляд початкових даних для однофакторного дисперсійного аналізу.

Номер елементів совокупностей	1	2	.	j	.	n
Номер совокупностей						
1	x_{11}	x_{12}	.	x_{1j}	.	x_{1n}
2	x_{21}	x_{22}	.	x_{2j}	.	x_{2n}
.	.	.	.	.	.	.
i	x_{i1}	x_{i2}	.	x_{ij}	.	x_{in}
.	.	.	.	.	.	.
m	x_{m1}	x_{m2}	.	x_{mj}	.	x_{mn}

Необхідно виявнити, чи відрізняються характеристики кожної з m груп в n групах. Сенс ДА полягає в тому, щоб порівняти дисперсію, що обумовлена випадковими причинами з дисперсією, викликаною наявністю деякого фактора. Якщо вони значущо відрізняються, то фактор має статистично значущий вплив на досліджувану змінну.

Відмінність вважається значущущим, якщо розрахункове значення критерію Фішера (відношення міжгрупової дисперсії до внутрішньогрупової)

буде більше табличного, взятого із заданим рівнем значущості і степенями свободи *(m–1)* і *m(n–1)*.

Міжгрупова дисперсія розраховується за формулою:

$$S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{\bar{X}}_i - \bar{\bar{X}})^2 , \qquad (4.14)$$

внутрішньогрупова

$$S_2^2 = \frac{1}{m(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{\bar{X}}_i)^2 . \qquad (4.15)$$

Тут $\bar{X}$ – загальне середнє, а $\bar{\bar{X}}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$.

Для ДА в англійській мові прийнято скорочення ANOVA (ANalys Of VAriantes, що означає «дисперсійний аналіз»), яке використовується і в деяких російськомовних джерелах.

Для однофакторного ДА результати розрахунків прийнято представляти у вигляді таблиці 4.5.

Таблиця 4.5. Результати розрахунків однофакторного дисперсійного аналізу.

Компоненти дисперсії	Сума квадратів	Кількість степенів свободи	Середній квадрат (дисперсія)
Міжгрупова (впливаючий фактор)	$\sum_{i=1}^{m}(\bar{X}_i - \bar{X})^2$	m-1	$S_1^2 = \sum_{i=1}^{m}(\bar{X}_i - \bar{X})^2/(m-1)$
Внутрішньогрупова (випадковий вплив)	$\sum_{j=1}^{n}\sum_{i=1}^{m}(X_{ij} - \bar{X})^2$	$m(n$-1)	$S_2^2 = \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{(X_{ij} - \bar{X})^2}{m(n-1)}$
Загальна	$\sum_{j=1}^{n}\sum_{i=1}^{m}(X_{ij} - \bar{X})^2$	mn-1	$S^2 = \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{(X_{ij} - \bar{X})^2}{mn-1}$

Однофакторна задача з неоднковою кількістю випробувань

Бувають задачі, коли кількість дослідів для різних значень рівня факторів відрізняється. Це може пов’язано, якщо, наприклад, з тим, що частина лабораторних тварин під час експерименту загинула або в них не було необхідної реакції і тому подібне.

Тоді загальна дисперсія визначається за формулою:

$$SS_{заг} = \sum_{j=1}^{m}\sum_{i=1}^{n_j} X_{ij}^2 - \frac{\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j} X_{ij}\right)^2}{n}, \qquad (4.16)$$

де X_{ij} – значення відповідного спостереження; n_j – кількість спостережень для j-го рівня фактора; $n = \sum_{j=1}^{m} n_j$ – загальна кількість спостережень.

Міжгрупова (викликана впливом факторів) сума квадратів визначається за формулою:

$$SS_{фактор} = \sum_{j=1}^{m} \frac{\left(\sum_{j=1}^{n_j} X_{ij}\right)^2}{n_j} - \frac{\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j} X_{ij}\right)^2}{n}. \qquad (4.17)$$

Залишкова сума знаходиться як різниця між загальною і факторною:

$$SS_{залишк} = SS_{заг} - SS_{факт}.$$

Потім визначається залишкова (внутрішньогрупова) $S^2_{заишкл} = \frac{SS_{залишк}}{(n-m)}$ і факторна $S^2_{фактор} = \frac{SS_{фактор}}{(n-m)}$ дисперсії, а також розрахункове значення критерію Фішера $F = \frac{S^2_{фактор}}{S^2_{залишк}}$.

Двофакторна задача з однаковою кількістю спостережень

У такій задачі є два фактори, що вимірювані в шкалі найменувань, які впливають на загальний відгук.

Загальний вигляд представлення даних для такої задачі наведено в таблиці 4.10.

Тут $X_{111}, X_{112}, X_{mnk}$ – спостережені значення досліджуваної змінної;

$\bar{\bar{X}}_{ij*} = \frac{1}{k}\sum_{l=1}^{k} X_{ijl}$ – середнє значення в комірці;

$\bar{\bar{X}}_{i**} = \frac{1}{n}\sum_{j=1}^{n} X_{ij*}$ – середнє значення в рядку;

$\bar{\bar{X}}_{*j*} = \frac{1}{m}\sum_{i=1}^{m} \bar{\bar{X}}_{ij*}$ – середнє значення по стовпцю;

$\bar{\bar{X}}_{*j*} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} \bar{X}_{ij*}$ – загальне середнє.

Таблиця 4.10. Загальний вигляд представлення початкових даних для двофакторного дисперсійного аналізу.

Фактор A	Фактор B				
	B_1	B_2	...	B_n	$\overline{X}_{1**}$
A_1	$\overline{X}_{11*} X_{111}$, $X_{112},...,X_{11k}$	$\overline{X}_{12*} X_{121}$, $X_{122},...,X_{12k}$	...	$\overline{X}_{1n*} X_{1n1}$, $X_{1n2},...,X_{1nk}$	$\overline{X}_{1**}$
A_2	$\overline{X}_{21*} X_{211}$, $X_{212},...,X_{21k}$	$\overline{X}_{22*} X_{221}$, $X_{222},...,X_{22k}$	...	$\overline{X}_{2n*} X_{2n1}$, $X_{2n2},...,X_{2nk}$	$\overline{X}_{2**}$
...	...	...	$\overline{X}_{ij*} X_{ij1}$, $X_{ij2},...,X_{ijk}$	...	...
A_m	$\overline{X}_{m1*} X_{m11}$, $X_{m12},...,X_{m1k}$	$\overline{X}_{m2*} X_{m21}$, $X_{m22},...,X_{m2k}$	...	$\overline{X}_{mn*} X_{mn1}$, $X_{mn2},...,X_{mnk}$	$\overline{X}_{m**}$
$\overline{X}_{*j*}$	$\overline{X}_{*1*}$	$\overline{X}_{*2*}$	...	$\overline{X}_{*n*}$	$\overline{X}$

Результати розрахунків зазвичай представляються в вигляді таблиці (табл. 4.11).

Таблиця 4.11. Форма представлення результатів двофакторного дисперсійного аналізу.

Компоненти дисперсії	Суми квадратів	Кількість степенів свободи	Дисперсії
S^2_A-між середніми по рядках (по фактору *A*)	$nk\sum_{i=1}^{m}\left(\overline{X}_{i**}-\overline{X}\right)^2$	*n-1*	$\frac{nk\sum_{i=1}^{m}\left(\overline{X}_{i**}-\overline{X}\right)^2}{(n-1)}$
S^2_B-між середніми по стовпцях (по фактору *B*)	$mk\sum_{i=1}^{m}\left(\overline{X}_{*j*}-\overline{X}\right)^2$	*m-1*	$\frac{mk\sum_{i=1}^{m}\left(\overline{X}_{*j*}-\overline{X}\right)^2}{(m-1)}$
Взаємодія S^2_{AB}	$k\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\overline{X}_{ij*}-\overline{X}_{i**}-\overline{X}_{*j*}+\overline{X}\right)^2$	*(m-1)* *(n-1)*	$\frac{k\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\overline{X}_{ij*}-\overline{X}_{i**}-\overline{X}_{*j*}+\overline{X}\right)^2}{(m-1)(n-1)}$
Залишкова $S^2_{зал}$	$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k}\left(X_{ijk}-\overline{X}_{ij*}\right)^2$	*nm(k-1)*	$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k}\left(X_{ijk}-\overline{X}_{ij*}\right)^2}{\text{nm(k -1)}}$
Повна S^2	$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k}\left(X_{ijk}-\overline{X}\right)^2$	*nmk-1*	$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k}\left(X_{ijk}-\overline{X}\right)^2}{\text{(nmk -1)}}$

Тут S^2_A – характеризує вплив фактора A; S^2_B – характеризує вплив фактора B; S^2_{AB} – спільний вплив обох факторів; $S^2_{зал}$ – вплив випадкових факторів, які неможливо віднести ні до A, ні до B.

Необхідно перевірити значущість відмінності дисперсій S^2_A і $S^2_{зал}$, S^2_B і $S^2_{зал}$, S^2_{AB} і $S^2_{зал}$. Для цього знаходять розрахункові значення F-критерия Фішера $F_A = \frac{S_A^2}{S_{зал}^2}$, $F_B = \frac{S_B^2}{S_{зал}^2}$.

Якщо виконуються умови $F_A > F_{\alpha,n-1,nm(k-1)}$, $F_B > F_{\alpha,m-1,nm(k-1)}$, то вплив факторів A і B є значимим.

4.3.2. Непараметричний дисперсійний аналіз Фрідмана

Призначення – застосовується в випадку, коли закон розподілу не є нормальним.

Нульова гіпотеза – середні значення всіх вибірок однакові.

Передумови – всі випадкові величини взаємно незалежні; дані кожної вибірки розподілені по одному закону розподілу. Закон розподілу однієї вибірки, може відрізнятися від закону розподілу іншої.

Опис методу

Початкові дані представляються у вигляді таблиці 4.17.

Таблиця 4.17. Загальний вигляд початкових даних для непараметричного дисперсійного аналізу.

Номери елементів сукупностей / Номери сукупностей	1	2	...	j	...	n
1	X_{11}	X_{12}		X_{1j}		X_{1n}
2	X_{21}	X_{22}		X_{2j}		X_{2n}
...	...	...		...		...
i	X_{j1}	X_{j2}		X_{ij}		X_{in}
...	...	...		...		...
m	X_{m1}	X_{m2}		X_{mj}		X_{mn}

Для цього в кожному стовпці значення X замінюють їх рангами (замість значень змінних ставиться їх номер в ряду, впорядкований за їх зростанням). Потім розраховується значення критерію

$$\chi^2 = \frac{12\sum_{j=1}^{n}\left(\sum_{i=1}^{m} R_{ij}\right)^2}{mn(n+1)} - 3m(n+1), \qquad (4.18)$$

де R_{ij} – відповідні значення рангів.

Якщо розрахункове значення χ^2 буде більше критичного, взятого з заданим рівнем значущості α і (n-1) степенями свободи, гіпотеза про відмінність між вибірками (сукупностями) приймається.

При розрахунках можна перевірити правильність простановки рангів і розрахунків, знаючи, що має місце співвідношення:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} R_{ij} = \frac{nm(m+1)}{2} \qquad (4.19)$$

Примітка. При малих значеннях m і n критерій χ^2 дає досить грубе наближення і при цьому можливе ухвалення неправильного рішення. Тому критерій χ^2 застосовується у тому випадку, коли виконуються наступні умови: m=3 і n>9 або m=4 і n>4 або m>4, $n\geq$9.

Якщо ці умови не виконуються, то перевірка здійснюється за критерієм Фрідмана.

4.4. Аналіз таблиць зв'язаності (сполучення)

У медико-біологічних дослідженнях велику роль грає аналіз таблиць зв'язаності або по-іншому – „аналіз таблиць часток і пропорцій". Використовувані тут методи призначені для аналізу даних, які описують об'єкти з деякою якістю властивостей, причому часто про властивість можна сказати, що вона є або її немає.

Нульова гіпотеза полягає в тому чи є властивість, чи немає.

4.4.1. Чотирьохклітинні таблиці (ЧТ)

В загальному вигляді ЧТ називають ще таблицями 2*2. Вони мають наступний вигляд (див. таблицю 4.19).

Таблиця 4.19. Загальний вид чотирьохкліткової таблиці зв'язаності.

Вибірка	Наявність ознаки	Відсутність ознаки	Разом
Перша вибірка	A	B	$n_1=A+B$
Друга вибірка	C	D	$n_2=C+D$
Разом	$A+C$	$B+D$	$n=n_1+n_2$

Нульова гіпотеза про приналежність обох вибірок до однієї генеральної сукупності визначається на основі використання критерію χ^2, який розраховується за формулою:

$$\chi^2 = \frac{n(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \qquad (4.20)$$

Для малих вибірок замість n беруть (n–1). Розрахункове значення порівнюється з критичним, взятим з однією степеню свободи і заданим рівнем значущості. Якщо розраховане значення більше критичного, то гіпотезу про однорідність можна відкинути і прийняти гіпотезу про наявність між ознаками, що вивчаються, істотного зв'язку.

Примітка. Правильність отриманих висновків залежить від того, як були вибрані дані: вибірка має бути однорідна по відношенню до аналізованої ознаки. Наприклад, якщо в аналізовану вибірку входять одночасно особи, на яких препарат має позитивний вплив, і особи, на яких він має негативний вплив, то в результаті аналізу може бути прийнята гіпотеза про те, що препарат не має ніякого впливу. А це не відповідає реальному стану речей.

Треба пам'ятати, що об'єм вибірок не має бути дуже малим. Так, для рівня значущості 0,05 необхідне мінімальне значення $n_1=n_2=124$ ($n=248$).

Окрім чотирьохкліткових таблиць зв'язаності існують багатоклітинні таблиці зв'язаності.

4.4.2. Таблиці виду *2*K*

Загальний вигляд таблиця зв'язаності *2*K* наведено в таблиці 4.20.

Ці таблиці використовуються для перевірки нульової гіпотези про однорідність k вибірок за допомогою формули, що запропонована Брандтом і Снедекором:

$$\chi^2 = \frac{n^2}{m(n-m)}\left[\sum_{i=1}^{k}\frac{m_i^2}{n_i} - \frac{m^2}{n}\right] \qquad (4.21)$$

Таблиця 4.20. Загальний вид таблиці зв'язаності типа *2*K*.

Номер вибірки і номер рівня другої ознаки	Ознака 1		Разом
	Є	Відсутня	
1	m_1	n_1-m_1	n_1
2	m_2	n_2-m_2	n_2
...	...	...	...
i	m_i	n_i-m_i	n_i
...	...	...	...
k	m_k	n_k-m_k	n_k
Разом	m	n-m	n

Розрахункове значення порівнюється з критичним, взятим з (k-1) степенями свободи і вибраним рівнем значущості. Якщо розрахункове значення більше критичного, то гіпотезу про однорідність треба відкинути і

прийняти гіпотезу про наявність між ознаками, що вивчаються, істотного зв'язку.

Для оцінки тісноти зв’язку в таких таблицях інколи використовують коефіцієнт бісеріальної кореляції.

4.4.3. Таблиці виду *K*L*

Таблиці виду *K*L* (див. таблицю 4.21) є найбільш загальним видом таблиць зв'язаності. В цьому випадку значеннями першої ознаки можуть бути, наприклад, різноманітні види лікування: симптоматичне специфічне з використанням препаратів в середньо терапевтичних або в підвищених дозах; специфічне з додатковим застосуванням інших препаратів тощо. Значеннями другої ознаки можуть бути, наприклад, видужання за два тижні, видужання за чотири тижні, летальний результат. Друга ознака може бути сукупністю різноманітних вибірок.

Для випадку, коли перший стовпчик (таблиця 4.21) є *K* значеннями рівня другої ознаки, перевіряється гіпотеза про незалежність першої і другої ознаки. Якщо ж перший стовпчик містить *k* різних вибірок, то перевіряється гіпотеза про однорідність цих вибірок (тобто чи можна вважати, що ці вибірки взяті з однієї генеральної сукупності).

Тобто ознаки 1 і 2 можуть бути сукупністю різних вибірок. В цьому випадку застосовується один критерій для перевірки гіпотези про незалежність ознак і про однорідність вибірок.

Таблиця 4.21. Загальний вигляд таблиці зв’язаності (сполучення) *K*L*.

Ознака 2 (К значень рівня)	Ознака 1 (*m* значень рівня)						Суми по рядках
	1	2	.	*j*	.	*m*	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	n_1
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	n_2
...	...	...	...	...	...	...	...
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	n_i
...	...	...	...	...	...	...	...
k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{km}	n_k
Суми по стовпцях	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.m}$	$n..=n$

Значення критерію розраховується за формулою:

$$\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{n_i n_j} - 1\right] \qquad (4.22)$$

Розрахункове значення порівнюється з критичним, взятим з $(k\text{-}1)(m\text{-}1)$ степенями свободи і заданим рівнем значущості. Якщо розрахункове

значення більше критичного, то гіпотезу про однорідність треба відкинути і прийняти гіпотезу про наявність між ознаками, що вивчаються, істотного зв'язку.

4.4.4. Розмір вибірки і рандомізація

Проблема рандомізації виникає у тому випадку, коли дані, які підлягають аналізу, взяті як деяка вибірка з генеральної сукупності. При цьому виникають наступні питання:

– скільки даних необхідно для ухвалення правильного рішення;

– як формулювати вибірку?

Кількість даних, що необхідні для аналізу, залежить від наступних факторів:

– припустимої похибки першого роду, тобто ймовірності встановлення значимої залежності, коли її немає. Це забезпечується за рахунок встановлення рівня значущості α;

– припустимої похибки другого роду, тобто ймовірності встановлення відсутності зв'язку, коли насправді він є. Зазвичай задається через ймовірність похибки другого роду β або потужність критерію $1-\beta$;

– від того, розпочинаючи з якої різниці частот можна вважати їх відмінність за значущу.

Проблема в тому, що зменшення розміру вибірки призводить до зменшення ймовірності виявлення значущої відмінності, а збільшення її (вибірки) – до підвищення ймовірності визнання значущими несуттєвих відмінностей.

Визначення розміру вибірки для випадку підвибірок однакових розмірів

Спочатку задаються рівень значущості α і потужність критерію $1-\beta$. Потім встановлюються частоти, які вважатимемо як такі, що відрізняються. Зазвичай вибирається одна частота і встановлюється частка відмінності. Після цього розраховується друга частота за формулою:

$$P_2=P_1+f(1-P_1) \qquad (4.23)$$

де P_1 – перша частота; f – частка, з якою ми вважаємо відмінність за значущу.

Наприклад, якщо P_1=0,6 і при цьому f=0,25, то P_2=0,6+0,25(1-0,6)=0,7. Це означає, що частоти 0,6 і 0,7 можна вважати як такі, що значущо відрізняються.

Тоді розмір кожної з двох підвибірок розраховується за формулою:

$$n=\frac{n'}{4}\left[1+\sqrt{1+\frac{4}{n'\left|P_2-P_1\right|}}\right] \qquad (4.24)$$

При цьому n' визначається за формулою:

$$n' = \frac{\left(Z_{\frac{\alpha}{2}} \sqrt{2\bar{\bar{P}}\bar{\bar{Q}}} - Z_{1-\beta} \sqrt{P_1 Q_1 + P_2 Q_2} \right)^2}{(P_2 - P_1)^2} \tag{4.25}$$

де $Q=1-P$ (для 1-ої і 2-ої); $\bar{\bar{P}} = (P_1 + P_2)/2$, $\bar{\bar{Q}} = (Q_1 + Q_2)/2$.

Визначення розміру вибірки для випадку підвибірок різних розмірів

Досить поширені ситуації, коли розміри вибірок відрізняються. В такому випадку вважаємо, що розмір однієї вибірки дорівнює m, а другої $N=(r+1)m$, де r в загальному випадку не дорівнює 1. Тоді розмір m обчислюється за формулою:

$$m = \frac{m'}{4} \left[1 + \sqrt{\frac{2(r+1)}{1 + m' r \left| P_2 - P_1 \right|}} \right], \tag{4.26}$$

де

$$m' = \frac{\left(Z_{\frac{\alpha}{2}} \sqrt{(r+1)\bar{\bar{P}}\bar{\bar{Q}}} - Z_{1-\beta} \sqrt{r P_1 Q_1 + P_2 Q_2} \right)^2}{r(P_2 - P_1)^2}. \tag{4.27}$$

Тут $Q=1-P$ (для 1-ої і 2-ої); $\bar{\bar{P}} = (P_1 + P_2)/2$, $\bar{\bar{Q}} = (Q_1 + Q_2)/2$, як і в попередньому випадку.

Рандомізація

Якщо не виконувати рандомізацію, то висновки будуть некоректні внаслідок невипадкового формування вибірок через порушення початкових передумов випадковості і незалежності спостережень. Здійснюється вона, наприклад, шляхом використання набору випадкових чисел або генератора випадкових чисел, які обмежені розмірами вибірок.