Тема. Дисперсійний аналіз та умови його застосування. Статистична обробка даних з використанням комп'ютерних програм

Дисперсійний аналіз застосовується для дослідження впливу однієї або декількох якісних змінних (факторів) на одну залежну кількісну змінну (відгук).

Аналіз часових рядів застосуємо до одиночних або пов'язаним часових рядах і дозволяє виділяти різні форми періодичності і взаємовпливу часових процесів, а також здійснювати прогнозування майбутньої поведінки часового ряду.

Регресивні процедури дозволяють розрахувати модель, що описується деяким рівнянням і відображає функціональну залежність між експериментальними кількісними змінними, а також перевіряють гіпотезу про адекватність моделі експериментальним даним. За отриманими результатами можна оцінити природу і ступінь залежності змінних і передбачити нові значення залежної змінної.

Кореляційний аналіз - це група статистичних методів, спрямована на виявлення та математичне уявлення структурних залежностей між вибірками.

Кластерний аналіз здійснює розбиття об'єктів на задане число віддалених один від одного класів, а також будує дерево класифікацій об'єктів за допомогою ієрархічного об'єднання їх в групи (кластери).

Основним завданням факторного аналізу є знаходження в багатовимірному просторі первинних змінних (значення яких реєструються в експерименті), скороченою системи вторинних змінних (факторів).

Факторний аналіз – статистичний метод перевірки гіпотез про вплив різних чинників на досліджувану випадкову величину. Розроблено і загальноприйнята модель, при якій вплив фактора представлено в лінійному вигляді.

Дисперсійний аналіз, запропонований Р. Фішером, є статистичним методом, призначеним для виявлення впливу ряду окремих факторів на результати експерименту.

В основі дисперсійного аналізу лежить припущення про те, що одні змінні можуть розглядатися як причини (фактори, незалежні змінні), а інші як слідства (залежні змінні). Незалежні змінні називають іноді регульованими чинниками саме тому, що в експерименті дослідник має можливість варіювати ними і аналізувати виходить результат.

Сутність дисперсійного аналізу полягає в розчленуванні загальної дисперсії досліджуваного ознаки на окремі компоненти, зумовлені впливом конкретних факторів, і перевірці гіпотез про значущість впливу цих факторів на досліджуваний ознака. Порівнюючи компоненти дисперсії один з одним за допомогою F - критерію Фішера, можна визначити, яка частка загальної варіативності результативної ознаки зумовлена дією регульованих факторів.

Вихідним матеріалом для дисперсійного аналізу є дані дослідження трьох і більше вибірок, які можуть бути як рівними, так і нерівними за чисельністю, як зв'язковими, так і незв'язними. За кількістю виявлених регульованих факторів дисперсійний аналіз може бути однофакторний (при цьому вивчається вплив одного фактора на результати експерименту), двофакторна (при вивченні впливу двох чинників) і багатофакторним (дозволяє оцінити не тільки вплив кожного з факторів окремо, але і їх взаємодія).

Дисперсійний аналіз відноситься до групи параметричних методів і тому його слід застосовувати тільки тоді, коли доведено, що розподіл є нормальним.

t – критерій Стьюдента, як відомо, дозволяє перевіряти гіпотези тільки про рівність середніх двох генеральних сукупностей. Однак часто перед дослідником стоїть завдання провести порівняння більшого числа совокупностей. Таке порівняння можна було б здійснити за допомогою спеціалізованих непараметричних критеріїв (H - Крускала-Уоллеса, S - Джонкіра та інших). Використовуючи ж більш потужні параметричні методи, найбільш просто вирішити це завдання, попарно порівнюючи між собою сукупності за допомогою того ж t - критерію Стьюдента, однак такий підхід не є правильним.

Справа в тому, що при зростанні числа груп стрімко збільшується число можливих пар, і ризик неправильного відкидання нульової гіпотези (т. Е. Ймовірність припуститися помилки

першого роду - стверджувати, що відмінності між генеральними сукупностями реально існують, коли це фактично не так) значно перевищує встановлюваний дослідником рівень значимості.

Більш коректним вирішенням цієї проблеми є застосування більш загального методу перевірки гіпотез про середні – так званого дисперсійного аналізу.

Нехай генеральні сукупності X1 X2 Xp. Розподілені нормально і мають однакову, хоча і невідому дисперсію. Математичні очікування також невідомі, але можуть бути різні. Потрібно при заданому рівні значущості за вибірковими середнім перевірити нульову гіпотезу

H0: M(X1)= M(X2)=…= M(Xp)

про рівність всіх математичних очікувань.

Іншими словами, потрібно встановити, значимо чи незначимо розрізняються вибіркові середні.

Здавалося б, при порівнянні декількох середніх їх можна було б порівнювати попарно, проте зі зростанням числа середніх зростає і найбільша відмінність між ними: середня нової вибірки може виявитися більше найбільшого або менше найменшого із середніх, отриманих до нового іспиту.

Тому для порівняння декількох середніх використовують інший метод, який заснований на порівнянні дисперсій і тому названий дисперсійним аналізом. Метод розвинений англійським статистиком Р. Фішером

На практиці дисперсійний аналіз використовують, щоб встановити, чи істотний вплив деякого якісного фактору F, котрий має *p* рівнів F1 F2 Fp, на досліджувану випадкову величину X, наприклад, потрібно з'ясувати, який вид добрив. Якщо відмінність між цими дисперсіями значима, то фактор має істотний найбільш ефективний для отримання найбільшого врожаю. В цьому випадку якісний фактор F - добриво, а його рівні - види добрив. Основна ідея дисперсійного аналізу полягає в порівнянні «Факторної» дисперсії, яку породжує впливом фактора, і «Залишкової» дисперсії, обумовленої випадковими причинамивплив на випадкову величину X. В цьому випадку середні спостережуваних значень на кожному рівні (групові середні) розрізняються також

значимо. Якщо вже встановлено, що фактор істотно впливає на випадкову величину X, а потрібно з'ясувати, який з рівнів фактора має найбільший вплив, то додатково виконують попарне порівняння середніх.

Дисперсійний аналіз використовують також для встановлення однорідності декількох сукупностей (дисперсії цих сукупностей однакові за припущенням; якщо дисперсійний аналіз покаже, що математичні сподівання однакові, то в цьому сенсі сукупності однорідні).

Однорідні сукупності можна об'єднувати в одну і тим самим отримати про неї більш повну інформацію, отже, і більш надійні висновки. У більш складних випадках досліджують вплив декількох факторів на кількох постійних або випадкових рівнях і з'ясовують вплив окремих рівнів та їх комбінацій (багатофакторний аналіз).

Загальна, факторна і залишкова суми квадратів відхилень

Розглянемо випадок однофакторного аналізу, коли на випадкову величину X впливає тільки один фактор F, котрий має p постійних рівнів. Нехай ознака X розподілена нормально. На неї впливає фактор F, котрий має *p* постійних рівнів. Будемо припускати, що число спостережень на кожному рівні постійно і дорівнює *q*. Таким чином, спостерігалося n = pq значень фактора, яким відповідають значення x_{ij} ознаки X, де *i* - номер випробування, $i = \overline{1, q}$, *j* - номер рівня фактора, $j = \overline{1, p}$. Результати спостережень зведені в таблицю

Номер испытания	Уровни фактора			
	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
q	x_{q1}	x_{q2}	…	x_{qp}
Групповые средние	$\bar{x}_{гр1}$	$\bar{x}_{гр2}$	…	$\bar{x}_{грp}$

Введемо в розгляд величини

$$S_{общ} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \bar{x})^2$$

-загальна сума квадратів відхилень спостережуваних значень від загальної середньої x.

$$S_{факт} = q\sum_{j=1}^{p}(\bar{x}_{грj} - \bar{x})^2$$

- факторная сума квадратів відхилень групових середніх від загальної середньої x. Вона характеризує розсіювання «між групами».

$$S_{ост} = \sum_{i=1}^{q}(x_{i1} - \bar{x}_{гр1})^2 + \sum_{i=1}^{q}(x_{i2} - \bar{x}_{гр2})^2 + ... + \sum_{i=1}^{q}(x_{ip} - \bar{x}_{грp})^2$$

- залишкова сума квадратів відхилень спостережуваних значень групи від своєї групової середньої, яка характеризує розсіювання «всередині групи».

Ці суми пов'язані рівністю:

$$S_{ост} = S_{общ} - S_{факт}.$$

Елементарними перетвореннями можна отримати формули, більш зручні для розрахунків:

$$S_{общ} = \sum_{j=1}^{p}P_j - \left[(\sum_{j=1}^{p}R_j)^2 \Big/ (pq)\right],$$

$$S_{факт} = \left[(\sum_{j=1}^{p}R_j^2) \Big/ q\right] - \left[(\sum_{j=1}^{p}R_j)^2 \Big/ (pq)\right],$$

де

$P_j = \sum_{i=1}^{q}x_{ij}^2$ – сума квадратів значень ознаки, відповідних рівню F_j (або говорять - на рівні F_j),

$R_j = \sum_{i=1}^{q}x_{ij}$ – сума значень ознаки на рівні F_j

Зауваження 1.

1) $S_{факт}$ характеризує вплив фактора F. Припустимо, що фактор істотно впливає на ВВ X, тоді група значень, що спостерігаються на одному певному рівні, взагалі кажучи, відрізняється від груп спостережень на інших рівнях. отже, розрізняються і групові середні, причому вони

тим більше розсіяні навколо загальної середньої, чим більшим виявиться вплив фактора. Звідси випливає, що для оцінки впливу фактора доцільно скласти суму квадратів відхилень групових середніх від загальної середньої (відхилення зводять в квадрат, щоб виключити погашення позитивних і негативних відхилень). Помноживши цю суму на *q*, отримаємо величину $S_{факт}$, яка і відображає вплив фактора.

2) $S_{ост}$ характеризує вплив випадкових причин. Здавалося б, спостереження однієї групи не повинні відрізнятися. Однак, оскільки на X крім фактора F впливають і випадкові причини, спостереження однієї групи, взагалі кажучи, різні, а значить, розсіяні навколо своєї груповий середньої. Звідси випливає, що для оцінки впливу випадкових причин доцільно скласти суму квадратів відхилень спостережуваних значень кожної групи від своєї групової середньої, тобто $S_{ост}$.

3) $S_{общ}$ відображає вплив і фактора, і випадкових причин. Будемо розглядати всі спостереження як єдину сукупність. Спостережувані значення ознаки різні внаслідок впливу фактора і випадкових причин. Для оцінки цього впливу доцільно скласти суму квадратів відхилень спостережуваних значень від загальної середньої.

Зауваження 2. Для спрощення обчислень віднімають з кожного спостережуваного значення одне і те ж число C, приблизно рівне загальному середньому. Позначимо $y_{ij} = x_{ij} - C$. тоді

$$S_{общ} = \sum_{j=1}^{p} Q_j - \left[(\sum_{j=1}^{p} T_j)^2 \Big/ (pq) \right],$$

$$S_{факт} = \left[(\sum_{j=1}^{p} T_j^2) \Big/ q \right] - \left[(\sum_{j=1}^{p} T_j)^2 \Big/ (pq) \right],$$

де $Q_j = \sum_{i=1}^{q} y_{ij}^2$, $T_j = \sum_{i=1}^{q} y_{ij}$. $R_j = \sum_{i=1}^{q} x_{ij} = \sum_{i=1}^{q} (y_{ij} + C) = \sum_{i=1}^{q} y_{ij} + qC = T_j + qC$.

Загальна, факторна і залишкова дисперсії

Розділивши суми квадратів відхилень на відповідне число ступенів свободи, отримаємо загальну, факторну і залишкову дисперсії:

$$s_{общ}^2 = \frac{S_{общ}}{pq-1}, \; s_{факт}^2 = \frac{S_{факт}}{p-1}, \; s_{ост}^2 = \frac{S_{ост}}{p(q-1)},$$

де *p* – число спостережень на кожному рівні,

q – число рівнів,

pq - 1 – число ступенів свободи загальної дисперсії,

p - 1 – число ступенів свободи факторної дисперсії,

p (q - 1) – число ступенів свободи залишкової дисперсії.

Якщо нульова гіпотеза про рівність середніх справедлива, то всі ці дисперсії є незміщеними оцінками генеральної дисперсії. З огляду на, наприклад, що обсяг вибірки *n*= *pq*, робимо висновок, що

$s_{общ}^2 = \frac{S_{общ}}{pq-1} = \frac{S_{общ}}{n-1}$ – виправлена вибіркова дисперсія, яка є незміщеною оцінкою генеральної дисперсії.

Зауваження. Число ступенів свободи *p*(*q* - 1) залишкової дисперсії дорівнює різниці між числом ступенів свободи загальної та факторної дисперсій.

Порівняння декількох середніх методом дисперсійного аналізу

При заданому рівні значущості потрібно перевірити нульову гіпотезу про рівність декількох ($p > 2$) середніх для нормальних сукупностей з невідомими, але однаковими дисперсіями.

Покажемо, що рішення цього завдання зводиться до порівняння факторної і залишкової дисперсій за критерієм Фішера.

1) Нехай нульова гіпотеза про рівність декількох середніх (будемо називати їх груповими) справедлива. В цьому випадку факторна і залишкова дисперсії є незміщеними оцінками генеральної дисперсії і, отже, розрізняються незначимо. Якщо порівняти ці оцінки за критерієм Фішера, то, очевидно, критерій вкаже, що нульову гіпотезу про рівність факторної і залишкової дисперсій немає підстав відкинути. Таким чином, якщо гіпотеза про рівність групових середніх правильна, то вірна і гіпотеза про рівність факторної і залишкової дисперсій.

2) Нехай нульова гіпотеза про рівність групових середніх помилкова. В цьому випадку зі зростанням розбіжності між груповими середніми збільшується і факторна дисперсія, а разом з нею і відношення

$$F_{набл} = \frac{s^2_{факт}}{s^2_{ост}}.$$

В результаті $F_{набл}$ виявиться більше $F_{кр}$, отже, гіпотеза про рівність факторної і залишкової дисперсій буде відкинута. Таким чином, якщо гіпотеза про рівність групових середніх помилкова, то помилкова і гіпотеза про рівність факторної і залишкової дисперсій. Від противного можна довести і справедливість зворотних тверджень: з правильності (хибності) гіпотези про рівність факторної і залишкової дисперсій слід правильність (хибність) гіпотези про середні. Отже, щоб перевірити гіпотезу про рівність групових середніх нормальних сукупностей з однаковими дисперсіями, досить перевірити за критерієм Фішера нульову гіпотезу про рівність факторної і залишкової дисперсій. В цьому і полягає метод дисперсійного аналізу.

Зауваження 1. Якщо факторна дисперсія виявиться менше залишкової, то звідси вже випливає справедливість гіпотези про рівність групових середніх, і значить, немає необхідності вдаватися до критерію Фішера.

Зауваження 2. Якщо немає впевненості в справедливості припущення про рівність дисперсій розглянутих сукупностей, то це припущення слід перевірити заздалегідь, наприклад, за критерієм Кочрена.

Приклад 1. Проведено по 6 випробувань на кожному з трьох рівнів фактору А . Методом дисперсійного аналізу при рівні значущості 0,05 перевірити гіпотезу про рівність групових середніх. Передбачається, що вибірки отримані з нормально розподілених генеральних сукупностей з рівними дисперсіями.

Номер измерения	Уровни фактора		
	$Ф_1$	$Ф_2$	$Ф_3$
1	10	17	12
2	15	15	18
3	14	19	20
4	18	22	18
5	20	20	16
6	16	14	22
$\bar{x}_{грj}$	15,5	17,8	17,7

Розв'язок. Обчислимо загальне середнє

$$\bar{x} = \frac{15{,}5 + 17{,}8 + 17{,}7}{3} = 17.$$

Для спрощення приймемо C=17 і перейдемо до умовних варіант. Складемо розрахункову таблицю

Номер измерения	Уровни фактора						Итоговый столбец
	Φ_1		Φ_2		Φ_3		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1	-7	49	0	0	-5	25	
2	-2	4	-2	4	1	1	
3	-3	9	2	4	3	9	
4	1	1	5	25	1	1	
5	3	9	3	9	-1	1	
6	-1	1	-3	9	5	25	
$Q_j = \sum_{i=1}^{q} y_{ij}^2$		73		51		62	$\sum Q_j = 186$
$T_j = \sum_{i=1}^{q} y_{ij}$	-9		5		4		$\sum T_j = 0$
T_j^2	81		25		16		$\sum T_j^2 = 122$

Знайдемо виправлені середньоквадратичні відхилення

$$S_{общ} = \sum_{j=1}^{p} Q_j - \left[(\sum_{j=1}^{p} T_j)^2 \Big/ (pq) \right] = 186 - \frac{0}{3 \cdot 6} = 186,$$

$$S_{факт} = \left[(\sum_{j=1}^{p} T_j^2) \Big/ q \right] - \left[(\sum_{j=1}^{p} T_j)^2 \Big/ (pq) \right] = \frac{122}{6} - 0 \approx 20{,}33,$$

$$S_{ост} = S_{общ} - S_{факт} = 186 - 20{,}33 = 165{,}67.$$

Обчислимо факторну та залишкову дисперсії

$$s_{факт}^2 = \frac{S_{факт}}{p-1} = \frac{20{,}33}{2} \approx 10{,}17, \; s_{ост}^2 = \frac{S_{ост}}{p(q-1)} = \frac{165{,}67}{3(6-1)} \approx 11{,}05.$$

Факторна дисперсія виявилася менше залишкової, отже, навіть не застосовуючи критерій Фішера, можна зробити висновок, що немає підстав відкинути гіпотезу про рівність середніх, а значить, вплив фактора не значимий.