Лекція 17. Елементи дисперсійного аналізу

Дисперсійний аналіз - статистичний метод, призначений для оцінки впливу різних факторів на результат експерименту, а також для подальшого планування аналогічних експериментів.

Залежно від кількості факторів, включених до аналізу, розрізняють однофакторний, двохфакторну і багатофакторний аналізи.

Суть дисперсійного аналізу полягає у тому, що загальну дисперсію досліджуваної ознаки розділяють на окремі компоненти, які обумовлені впливом певних конкретних факторів:

$S_{заг} = S_{факт} + S_{залиш}$,

де $S_{заг}$ - сума квадратів відхилень від середнього для всієї вибірки;

$S_{факт}$- для досліджуваного фактора (між групова дисперсія);

$S_{залиш}$ - для неврахованих факторів (внутрігрупова дисперсія).

Цю рівність називають основним рівнянням дисперсійного аналізу

Для однофакторного дисперсійного аналізу: $S_{заг} = S_A + S_{залиш}$.

Для двофакторного: $S_{заг} = S_A + S_B + S_{залиш}$, якщо для кожної комбінації рівнів факторів є лише одне виміряне значення та якщо для кожної комбінації рівнів факторів є багаторазово виміряні значення:

$S_{заг} = S_A + S_B + S_{AB} + S_{залиш}$, де S_{AB} − сума квадратів відхилень, викликаних взаємодією факторів A і B.

11.1. Однофакторний діперсійній аналіз

Однофакторний дисперсійний аналіз (ANOVA – analysis of variance) використовується для порівняння середніх значень для трьох і більше вибірок. Фактором називається незалежна змінна, вплив якої вивчається на залежну змінну.

Необхідною умовою для проведення дисперсійного аналізу є те, щоб незалежна змінна була категоріальною, а залежна - метричною.

Набір даних у ANOVA складається з k - незалежних одновимірних вибірок, елементи яких виміряні у однакових одиницях (дол., кг., бали). Припустимі різні обсяги (розміри) вибірок.

Нехай є N нормально розподілених генеральних сукупностей з рівними дисперсіями та, можливо, з різними математичними сподіваннями.

Із кожної сукупності робимо вибірку об'єму $\{n_i\}, i = 1, 2, ..., k$. Тоді $\sum_{i=1}^{k} n_i = n$ - об'єм усієї вибірки.

Процедура виконання однофакторного дисперсійного аналізу:

1. Визначення незалежних і залежних змінних
2. Розкладання повної дисперсії $(S_{заг})$
3. Вимірювання ефекту (η^2)
4. Перевірка значущості (F)
5. Подання результату

Розглянемо алгоритм однофакторного дисперсійного аналізу поетапно

1 етап. Підготовка даних для аналізу виглядає наступним чином:

	Незалежна змінна - фактор (Напр., вид діяльності) (Кількість вибірок $k = 4$)			
	Вибірка 1 - (економісти)	Вибірка 2 - (інженери)	Вибірка 3 - (філологи)	Вибірка k - (хіміки)
Залежна	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{k,1}$
Залежна	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{k,2}$
Залежна	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{k,3}$
Залежна	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{k,4}$
Залежна	$X_{1,5}$	$X_{2,5}$		$X_{2,5}$
Залежна		$X_{2,6}$		$X_{k,6}$
Залежна		$X_{2,7}$		
Об`єм $n = n_1+n_2+n_3+\ldots+n_k$	$n_1 = 5$	$n_2 = 7$	$n_3 = 4$	$n_k = 6$
Середнє	X_1	X_2	X_3	X_k
Ст. відхилення	σ_1	σ_2	σ_3	σ_k

Нульова гіпотеза у однофакторному дисперсійному аналізі стверджує, що всі середні значення з різних генеральних сукупностей (які представлені вибірковими середніми) рівні між собою.

$H_0 : \mu_1 = \mu_k$ (всі рівні), або $\left(X_1 = X_2 = \ldots = X_k \right)$, або $H_0 : S_{факт} = S_{залиш}$.

Альтернативна гіпотеза стверджує, що хоча б два будь-яких середніх не рівні між собою.

$H_1 : \mu_1 \neq \mu_k$ (хоча б дві нерівні), або $\left(X_1 \neq X_k \right)$, або $H_1 : S_{факт} > S_{залиш}$.

F – тест складається у розрахунку F - статистики та порівнянні її з табличним значенням (аналогічно з t - тестом).

Оскільки нульова гіпотеза стверджує, що середні всіх генеральних сукупностей рівні, необхідно оцінити це середнє значення за всіма вибірками, тобто розрахувати загальну середню. Загальна середня є середньою всіх значень з усіх вибірок.

Якщо розміри вибірок не рівні, то середнє розраховується як середньозважене

$\overline{x} = \dfrac{\sum_{i=1}^{k} x_i n_i}{n}$ з урахуванням розміру вибірок.

2 етап. Для вивчення відмінностей між залежними змінними проводиться розкладання повної дисперсії: $S_{заг} = S_{факт} + S_{залиш}$, де $S_{факт}$ – міжгрупова варіація и $S_{залиш}$ - внутрігрупова варіація.

Міжгрупова варіація $\left(S_{факт} \right)$ показує, наскільки вибіркові середні відрізняються між собою. Вона дорівнює нулю, якщо середні рівні і тим більше, чим сильніше розрізняються середні.

Розрахунок міжгрупової дисперсії (варіації):

$$S_{факт} = \frac{n_i \cdot \sum_{i=1}^{k} \left(\overline{x}_i - \overline{x} \right)}{k - 1},$$

$\overline{x}_i = \dfrac{\sum\limits_{j=1}^{n} x_{ij}}{n_i}$ - середня арифметична вибірки із i - тої сукупності;

$\overline{x} = \dfrac{\sum\limits_{i=1}^{k} \overline{x}_i}{k}$ - середня усієї вибірки.

Внутрігрупова варіація $\left(S_{залиш}\right)$ показує, наскільки відрізняються між собою значення по кожній вибірці

$$S_{залиш} = \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n}\left(x_{ij} - \overline{x}_i\right)^2}{n-k}.$$

Очевидно, що: $S_{факт} > S_{залиш}$.

В противному випадку немає необхідності застосовувати критерій Фішера, тому що в цьому випадку фактор впливає незначно.

3 етап. Ефект впливу незалежної змінної на залежну змінну розраховується через кореляційне відношення η^2 (ета-квадрат), яке розраховується за формулою:

$$\eta^2 = \frac{S_{факт}\cdot(k-1)}{S_{факт}\cdot(k-1) + S_{залиш}\cdot(n-k)}$$

Значення кореляційного відношення знаходиться в межах від 0 до 1. Воно дорівнює 0, коли всі вибіркові середні рівні, тобто незалежна змінна не впливає на залежну, і, навпаки, вплив збільшується зі зростанням цього значення. Іншими словами, величина η^2 являє собою міру варіації залежної змінної, викликану впливом на неї незалежною змінної.

4 етап фактично зводиться до процедури статистичної перевірки гіпотези про рівність середніх (наявності відмінностей) шляхом розрахунку *F-статистики*:

$$F = \frac{S_{факт}}{S_{залиш}}.$$

5 етап. Для того, щоб зробити остаточний висновок, необхідно звернутися до F - таблиці, що містить критичні значення F - статистики при істинної нульовій гіпотезі. Щоб знайти критичне значення, необхідно врахувати кількість ступенів свободи (df- degree freedom) і відповідний рівень перевірки (за замовчуванням 5%).

Ступінь свободи для груповий варіації становить $k-1$, а для внутрішньо групової варіації $n-k$.

F - тест полягає в порівнянні F - статистики, розрахованої за наявними даними з критичним значенням F – таблиці розподілу Фішера (додаток 7). Результат є значущим, якщо $F_{стат} > F_{крит}$, по-скільки це говорить про наявність істотних відмінностей між середніми значеннями по групах.

Приклад 17.1. Поставки продукції для деякої компанії здійснюються трьома по-постачальниками («Азимут», «Елен» і «Oxamit») в різний час: денні години, нічні зміни і під час перезміни. Цілком очевидно, що контроль за якістю продукції в денний час вище, ніж в інший час. Зібрані дані з оцінками якості (в балах), і необхідно дізнатися, чи є відмінність в якості продукції, яка поставляється в різний час?

	Денна зміна	Нічна зміна	Перезмінка
«Азимут»	77,06	93,12	77,05
«Елен»	81,14	88,13	78,11
«Oxamit»	82,02	81,18	79,91

Розв’язання: $k=3,\ n_i=3,\ n=k\cdot n_i=3\cdot 3=9.$

1. Обчислимо оцінку математичного очікування кожного з варіантів поставки продукції:

$$\overline{x}_1 = \frac{\sum_{j=1}^{3} x_{1j}}{n_i} = \frac{77,06 + 81,14 + 82,02}{3} = 80,07;$$

$$\overline{x}_2 = \frac{\sum_{j=1}^{3} x_{2j}}{n_i} = \frac{93,12 + 88,13 + 81,18}{3} = 87,48;$$

$$\overline{x}_3 = \frac{\sum_{j=1}^{3} x_{3j}}{n_i} = \frac{77,05 + 78,11 + 79,91}{3} = 78,36.$$

2. Обчислимо оцінку математичного очікування спостережуваних величин:

$$\overline{x} = \frac{\sum_{i=1}^{k} \overline{x}_i}{k} = \frac{80,07 + 87,48 + 78,36}{3} = 81,97\,.$$

3. Обчислимо оцінку груповий дисперсії:

$$S_{факт} = \frac{n_i \cdot \sum_{i=1}^{k} \left(\overline{x}_i - \overline{x}\right)^2}{k-1} =$$

$$= \frac{3 \cdot \left[\left(80,07 - 81,97\right)^2 + \left(87,48 - 81,97\right)^2 + \left(78,36 - 81,87\right)^2 \right]}{2} = 70,47.$$

4. Обчислимо оцінку внутрішньо групової дисперсії:

$$S_{залиш} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} - \overline{x}_i\right)^2}{n-k} = \frac{1}{(9-3)} \times$$

$$\times \Big[\left(77,06 - 80,07\right)^2 + \left(81,14 - 80,07\right)^2 + \left(82,02 - 80,07\right)^2 + \left(93,12 - 87,48\right)^2 +$$

$$+ \left(88,13 - 87,48\right)^2 + \left(81,18 - 87,48\right)^2 + \left(77,05 - 78,36\right)^2 + \left(79,91 - 78,36\right)^2 \Big] = 15,02.$$

5. Обчислимо ефект впливу незалежної змінної на залежну змінну через кореляційне відношення η^2:

$$\eta^2 = \frac{S_{факт} \cdot (k-1)}{S_{факт} \cdot (k-1) + S_{залиш} \cdot (n-k)} = \frac{70,47 \cdot (3-1)}{\left[70,47 \cdot (3-1)\right] + \left[15,02 \cdot (9-3)\right]} = 0,61\,.$$

Отримане кореляційне відношення показує, що 61% варіації залежної змінної пояснюється зміною незалежної змінної і 39% залежною змінною пояснюється іншими факторами, що діють на вибірку вибірково.

6. Обчислимо F - статистику:

$$F_{стат} = \frac{S_{факт}}{S_{залиш}} = \frac{70,47}{15,02} = 4,69.$$

7. При рівні значущості $\alpha = 0,05$ по таблиці критичних точок розподілу Фішера зі ступенями свободи визначимо критичне значення критерію

$$k_1 = k - 1 = 3 - 1 = 2; \quad k_2 = (n - k) = (9 - 3) = 6.$$

$$F_{крит} = F(\alpha; k_1; k_2) = F(0,05; 2; 6) = 5,14.$$

8. Оскільки $F_{стат} < F_{крит}$ $(4,69 < 5,14)$, то немає підстав відкинути нульову гіпотезу H_0.

Таким чином, групові середні відрізняються незначно, тобто фактор A впливає незначно. Іншими словами можна відзначити: результати розрахунку показують, що $F_{стат} < F_{крит}$ $(4,69 < 5,14)$, отже, відмінність в якості продукції, що поставляється в різний час відсутня. Можна вважати доведеним той факт, що якість продукції, що поставляється не залежить від часу поставки і є однаковим в різний час.