5.2.1. Дисперсійний аналіз

За допомогою дисперсійного аналізу (англійською *analysis of variance*, скорочено ANOVA) досліджують вплив однієї або декількох незалежних змінних (факторів) на одну залежну змінну (одномірний аналіз) або на декілька залежних змінних (багатомірний аналіз). Незалежні змінні мають *тільки дискретні значення* (і відносяться до номінальної або порядкової шкали).

Однофакторний дисперсійний аналіз використовується у тих випадках, коли є в розпорядженні три або більше незалежних вибірок, отриманих з однієї генеральної сукупності шляхом зміни якого-небудь фактора, для якого, з якихось причин, немає кількісних вимірів.

Для цих вибірок припускають, що вони мають різні вибіркові середні й однакові вибіркові дисперсії. Тому необхідно відповісти на питання: зробив цей фактор істотний вплив на зміну вибіркових середніх залежної змінної чи така зміна (розкид) є наслідком випадків, викликаних невеликими обсягами вибірок? Іншими словами, якщо вибірки належать до однієї і тієї ж генеральної сукупності, то розкид даних між вибірками (між групами) повинен бути не більше, ніж розкид даних усередині цих вибірок (усередині груп).

Дисперсійний аналіз виконується при допущенні, що вихідна величина є випадковою і підпорядковується нормальному закону розподілу ймовірності.

Основна ідея дисперсійного аналізу полягає у дослідженні відміни між середніми значеннями окремих груп спостережень. Кожна група характеризується певним рівнем або діапазоном фактора. Може здаватися дивним, що процедура порівняння середніх називається дисперсійним аналізом. Насправді це пов’язано із тим, що при дослідженні статистичної значимості різниці між середніми значеннями груп ми порівнюємо оцінки дисперсій. В основі дисперсійного аналізу лежить представлення загальної дисперсії як суми двох дисперсій – внутрішньогрупової дисперсії та дисперсії, підрахованої без урахування належності до якоїсь групи.

Якщо фактор є впливовим, дисперсія спостережень в окремій групі значно менша, ніж дисперсія всієї вибірки. Причина цього, очевидно, полягає в суттєвій різниці між середніми значеннями у групах. Розглянемо такий набір даних (табл. 5.1).

Таблиця 5.1

	Група 1	**Група 2**
Спостереження 1	2	7

Спостереження 2 **Спостереження 3**	1 3	5 6
Середнє **Сума квадратів відхилень у групах**	(2+1+3)/3=2 $(2-2)^2+(1-2)^2+$ $+(3-2)^2=2$	(7+5+6)/3=6 $(7-6)^2+(5-6)^2+$ $+(6-6)^2=2$
Загальне середнє **Загальна сума квадратів відхилень**	(2+1+3+7+5+6)/6=4 $(2-4)^2+(1-4)^2+(3-4)^2+(7-4)^2+$ $+(5-4)^2+(6-4)^2=28$	

Середні двох груп суттєво відрізняються (2 и 6 відповідно). Сума квадратів відхилень у кожної групи дорівнює 2. Додаючи їх, отримуємо суму квадратів відхилень від середніх у групах:

$$\boldsymbol{D^{**} = 2 + 2 = 4}.$$

Якщо тепер не враховувати наявність окремих груп, тобто впливу фактора на результат, тоді загальна сума квадратів відхилень від загального середнього цих двох вибірок дорівнює:

$$\boldsymbol{D = 28}.$$

Основна тотожність дисперсійного аналізу має такий вигляд:

$$\boldsymbol{D = D^{*} + D^{**}},$$

де $\boldsymbol{D^{*}}$ – сума квадратів, яка обумовлена різницею середніх значень між групами.

$$\boldsymbol{D^{*} = D - D^{**}} = 28 - 4 = 24.$$

Перевірка значимості в дисперсійному аналізі заснована на порівнянні компоненти дисперсії, обумовленої міжгруповим розкидом, і компоненти дисперсії, обумовленої внутрішньо-груповим розкидом. Отримані компоненти дисперсії можна порівняти за допомогою F-критерію, що перевіряє, чи дійсно відношення дисперсій значимо більше 1, тобто дисперсія, що обумовлена міжгруповим розкидом, значимо більше за компоненту дисперсії, обумовлену внутрішньогруповим розки-дом.

Обчислимо критерій Фішера за формулою:

$$F = \frac{\mathrm{D}^{*}/(m-1)}{\mathrm{D}^{**}/(n-m)} = \frac{24/(2-1)}{4/(6-2)} = 24,$$

де m – число груп, n – число даних.

Порівняємо це значення з табличним значенням розподілу Фішера, яке залежить від рівня значимості (альфа), який зазвичай приймають 1-α=1-0,05=0,95 і двох значень ступенів волі $m-1=1$ та $n-m=4$. Тому що $\boldsymbol{F}$

$> F_{0,95;1;4}$, робимо висновок, що наявність груп впливає на результат.

Якщо фактор не є впливовим, сума дисперсій спостережень окремих груп не відрізняється суттєво від дисперсії всієї вибірки. Якщо вірна нульова гіпотеза (рівність середніх у двох групах), то можна чекати порівняно невеликого розходження вибіркових середніх через чисто випадкову мінливість. Тоді внутрішньо-групова дисперсія буде практично збігатися з загальною дисперсією, підрахованою без врахування групової прина-лежності. Розглянемо такий набір даних (табл. 5.2).

Таблиця 5.2

	Група 1	**Група 2**
Спостереження 1 **Спостереження 2** **Спостереження 3**	2 1 3	3 2 1
Середнє **Сума квадратів відхилень**	(2+1+3)/3=2 $(2-2)^2+(1-2)^2+$ $+(3-2)^2 = 2$	(3+2+1)/3=2 $(3-2)^2+(2-2)^2+$ $+(1-2)^2 = 2$
Загальне середнє **Загальна сума квадратів відхилень**	(2+1+3+3+2+1)/6=2 $(2-2)^2+(1-2)^2+(3-2)^2+(3-2)^2+(2-2)^2+$ $+(1-2)^2 = 4$	

Сума квадратів відхилень у кожної групи знову дорівнює 2, а їхня сума $D^{**} = 2 + 2 = 4$. Але сума квадратів, яка обумовлена різницею середніх значень між групами, дорівнює нулю

$$D^{*} = D - D^{**} = 4 - 4 = 0.$$

Значення критерію Фішера теж дорівнює нулю, що дає можливість зробити висновок про відсутність впливу наявності групи.