Практична робота № 7

Тема: Однофакторний дисперсійний аналіз

Мета роботи: Зробити однофакторний дисперсійний аналіз, використовуючи пакет SPSS.

Стислі теоретичні відомості

Дисперсійний аналіз виконується для з'ясування впливу фактора X на вихідну величину Y (результат) у випадках, коли неможливо провести порівняльні спостереження вихідної величини при наявності фактора і без нього (контрольна група) та набрати відповідні статистичні дані. 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 двох або більше 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 груп ми порівнюємо оцінки дисперсій. В основі дисперсійного аналізу лежить представлення загальної дисперсії як суми двох дисперсій – внутрігрупової 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).

Середні двох груп суттєво відрізняються (2 и 6 відповідно). Сума квадратів відхилень у кожної групи дорівнює 2. Додаючи їх, отримуємо 4. Якщо тепер не враховувати наявність окремих груп, тобто впливу фактора на результат, тоді загальна сума квадратів відхилень загального середнього цих двох вибірок дорівнює 28.

Таблиця 5

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

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. Тому внутрігрупова дисперсія буде практично збігатися з загальною дисперсією, підрахованою без обліку групової приналежності. Розглянемо такий набір даних (табл. 6):

Таблиця 6

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

Сума квадратів відхилень у кожної групи знову дорівнює 2, а їхня сума $2 + 2 = 4$. Але загальна сума квадратів відхилень всіх спостережень дорівнює також 4, а сума квадратів, яка обумовлена різницею середніх значень між групами, дорівнює нулю $(2 + 2) – 4 = 0$.

Перевірка значимості в дисперсійному аналізі заснована на порівнянні компоненти дисперсії, обумовленої міжгруповим розкидом, і компоненти дисперсії, обумовленої внутрігруповим розкидом. Отримані компоненти дисперсії можна порівняти за допомогою *F*-критерію, що перевіряє, чи дійсно відношення дисперсій значиме більше 1.

Послідовність дій при дисперсійному аналізі

- Весь діапазон значень фактора розбивається на декілька груп.
- У кожній групі знаходяться середні значення вихідної величини

$$y_{cp_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} .$$

де n_j – число значень вихідної величини у j-групі.

- Висувається нульова гіпотеза, яка полягає у тому, що фактор не впливає на результат, тому дисперсію вихідної величини можна оцінити, використовуючи середні групові значення $y_{\text{ср}\,j}$.
- Обирається рівень довіри нульової гіпотезі $1 - \alpha = (90 - 95)\%$.
- Обчислюється критерій Фішера, який є відношенням двох оцінок дисперсії: D^*, яка обчислена з врахуванням нульової гіпотези, використовуючи середні групові значення $y_{\text{ср}\,j}$, та D^{**}, яка обчислена без врахування нульової гіпотези:

$$F = \frac{D^*}{D^{**}} = \frac{\frac{1}{k-1} * \sum_{i=1}^{k} n_j (y_{\text{ср}j} - y_{\text{заг.ср}})^2}{\frac{1}{N-k} * \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - y_{\text{заг.ср}})^2},$$

де N – загальне число спостережень, k – число рівнів, n_j – число значень вихідної величини в j-групі, $k - 1$ – число ступенів волі чисельника, $N - k$ – число ступенів волі знаменника, $y_{\text{заг.ср.}}$ – загальне середнє.

- Визначається квантиль $F_{1-\alpha,\, k-1,\, N-k}$ (таке значення випадкової величини, підпорядкованої розподілу Фішера з $k - 1$ та $N - k$ ступе-

нями вільності, яке відсікає площу 1 – α від кривої щільності розподілу ймовірності).

- Якщо $F > F_{1-\alpha, k-1, N-k}$, то ймовірність нульової гіпотези менше α і фактор впливає на результат.
- Для виконання дисперсійного аналізу у пакеті SPSS у таблицю додаються стовпчики рівнів кожного фактора. Потім у головному меню обираються пункти:

Statistic → Compare → Means → One-Way ANOVA.

Отримуємо значення критерію Фішера та ймовірність нульової гіпотези.

Приклад виконання

Фактор може бути розбитим на рівні, тобто набувати різних значень не тільки в кількісному плані, але і в якісному. Наступні дані включають результати роботи трьох спостерігачів: Іванов ($y1$), Петров ($y2$), Сидоров ($y3$). Перевіримо, чи є результати їхньої роботи статистично суттєво різними. Нульова гіпотеза, якраз навпаки, полягає у тому, що це не так, тобто математичні сподівання результатів вимірів кожного з них однакові. Якщо її ймовірність буде малою, це дасть нам підставу стверджувати, що нульова гіпотеза невірна, тобто спостереження кожної особи статистично відрізняються (табл. 7).

Таблиця 7

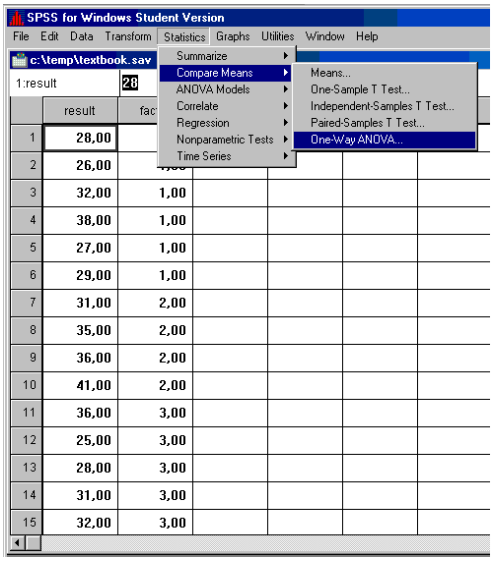
Іванов – y_1	**Петров – y_2**	**Сидоров – y_3**
28	31	36
26	35	25
32	36	28
38	41	31
27		32
29		

Першим кроком є перетворення трьох наборів даних у дві змінні. Перша змінна “result” буде включати всі спостереження. Друга змінна “factor” показує, кому зі спостерігачів належать дані (табл. 8).

Таблиця 8

result	factor
28	1
26	1
32	1
38	1
27	1
29	1
26	2
31	2
35	2
36	2
41	2
37	2
36	3
25	3
28	3
31	3
32	3
34	3

Після того, як ми ввели дані у таблицю SPSS, слід обрати пункт “Statistics”, потім “Compare Means”, нарешті “One-Way ANOVA”.



Заносимо “result” до “Dependent list”, “factor” до “Factor list”, натискуємо ”Define Range” та обираємо мінімальний та максимальний номери груп (для нашого прикладу відповідно 1 та 3), після чого необхідно натиснути “Continue” та “OK” (рис. 22).

Буде отримана така таблиця, яка є таблицею дисперсійного аналізу (рис. 23):

Як видно з таблиці, загальна сума квадратів відхилень 309,33 розбита на складові: суму квадратів, що обумовлена внутрішнім груповим розкидом (217,95; див. другий рядок таблиці), і суму квадратів, яка обумовлена різницею середніх значень між групами (309,33-217,95=91,38; див. перший рядок таблиці). Відмітимо, що оцінки дисперсій: D^*, яка обчислена з врахуванням нульової гіпотези, та D^{**}, яка обчислена без врахування нульової гіпотези, розраховуються як суми квадратів, поділені на відповідні кількості ступенів волі (D.F.):

$D^{*} = 91{,}38/2 = 45{,}69;$ $D^{**} = 217{,}95/12 = 18{,}16.$

F-критерій розраховується як $D^{*} / D^{**} = 45{,}69 / 18{,}16 = 2{,}5157$. Табличне значення розподілу Фішера не видається, але розраховується відповідне значення ймовірності нульової гіпотези $\alpha_t = 0{,}1223$. Якщо необхідний рівень довіри нульовій гіпотезі, скажемо, $1 - \alpha = 1 - 0{,}05 = 0{,}95$, то може бути зроблено висновок, що нульова гіпотеза підтверджується ($\alpha_t = 0{,}1223 > \alpha = 0{,}05$), тобто фактор не впливає на результат.

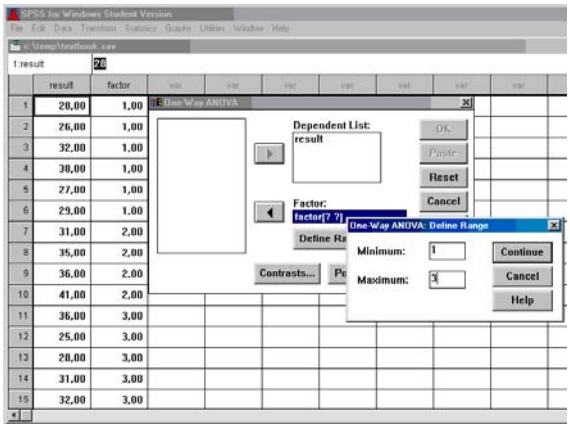


Рис. 23

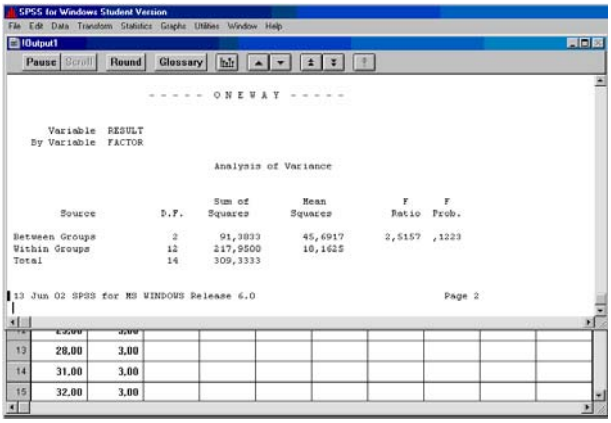


Рис. 24

Контрольні запитання

1. Яке допущення використовується при однофакторному дисперсійному аналізі?
2. В чому полягає основна ідея однофакторного дисперсійного аналізу?
3. На чому заснована перевірка значимості в дисперсійному аналізі?
4. Яка послідовність дій при дисперсійному аналізі?
5. Як формулюється нульова гіпотеза?
6. Як визначається статистика Фішера?
7. Як обирається ступінь довіри?
8. Як визначається квантиль $F_{1-\alpha,\, k-1,\, N-k}$?
9. Як використати пакет SPSS для однофакторного дисперсійного аналізу?