ДИСПЕРСИОННЫЙ АНАЛИЗ

Однофакторный дисперсионный анализ

Во многих практических ситуациях представляет интерес влияние того или иного фактора на рассматриваемый признак.

Пусть, например, оценка качества поверхности детали проводится с помощью l приборов и необходимо исследовать влияние фактора «прибор» на результат измерений. Если приборов два, то проверка нулевой гипотезы о равенстве их средних показаний проводится обычными методами проверки статистических гипотез. Если же $l > 2$**,** то используются методы диспер-тьлсионного анализа.

Проверяется нулевая гипотеза H_0: $m_1 = m_2 = \ldots = m_l$ об отсутствии влияния на результативный признак X фактора A, имеющего l уровней A_k, $k = 1, \ldots, l$. Основная идея дисперсионного анализа состоит в том, чтобы сопоставить дисперсию за счет воздействия фактора A с дисперсией, обусловленной случайными причинами. Если различие между ними не существенно, то влияние фактора A на признак X незначительно. Если же различие между факторной и остаточной дисперсиями значимо, то это говорит о влиянии фактора A на рассматриваемый признак X.

Предполагается, что случайная величина X имеет нормальное распределение с математическим ожиданием m_k, зависящим от уровня фактора A_k, и постоянной дисперсией σ^2. В качества исходных данных используются выборочные значения величины X, полученные для каждого уровня фактора A; число элементов выборки на каждом уровне равно n, тогда общее число наблюдений nl, x_{ik} - результат

номер опыта	Уровни фактора A			
	A1	A2	… Ak …	A*l*
1	X11	X12	… X1k …	X1*l*
2	X12	X22	… X2k …	X2*l*
…	…	…	--------	…
i	Xi1	Xi2	… Xik …	Xi*l*
…	…	…	--------	…
n	Xn1	Xn2	… Xnk …	Xn*l*

i -го наблюдения (i =1, …, n) за k-тым уровнем фактора A (k = 1, …, l).

Выборочная средняя, соответствующая k-му уровню фактора A, (групповая средняя) вычисляется по формуле:

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \; ; \qquad (3.1)$$

общая выборочная средняя есть

$$\bar{x} = \frac{1}{nl}\sum_{k=1}^{l}\sum_{i=1}^{n} x_{ik} = \frac{1}{l}\sum_{k=1}^{l}\bar{x}_k . \qquad (3.2)$$

Для вычисления дисперсии найдем суммы квадратов.

Общая сумма квадратов – это сумма квадратов отклонений наблюдаемых значений x_{ik} от общей выборочной средней:

$$Q = \sum_{k=1}^{l}\sum_{i=1}^{n}(x_{ik} - \bar{x})^2 = \sum_{k=1}^{l}\sum_{i=1}^{n} x_{ik}^2 - nl\bar{x}^2 . \qquad (3.3)$$

Факторная сумма квадратов (обусловленная влиянием фактора A) - это сумма квадратов отклонений групповых средних от общей средней:

$$Q_A = n\sum_{k=1}^{l}(\bar{x}_k - \bar{x})^2 = n\sum_{k=1}^{l}\bar{x}_k^2 - nl\bar{x}^2 . \qquad (3.4)$$

Остаточная сумма квадратов характеризует рассеяние внутри группы:

$$Q_e = \sum_{k=1}^{l}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2 . \qquad (3.5)$$

На практике эта сумма определяется из основного тождества дисперсионного анализа, в соответствии с которым

$$Q = Q_A + Q_e. \qquad (3.6)$$

Разделив суммы квадратов на соответствующее число степеней свободы, найдем соответствующие дисперсии (иногда их называют средними суммами квадратов):

$$S^2 = \frac{Q}{nl-1},$$
$$S_A{}^2 = \frac{Q_A}{l-1}, \qquad (3.7)$$
$$S_e{}^2 = \frac{Q}{l(n-1)}.$$

Если нулевая гипотеза о равенстве средних справедлива, то эти дисперсии являются несмещенными оценками дисперсий генеральной совокупности. Значительное превышение дисперсии S_A^2 над дисперсией S_e^2 можно объяснить различием средних в группах. Поэтому для проверки нулевой гипотезы используется отношение этих средних, которое имеет распределение Фишера

$$F = \frac{S_A}{S_e} = \frac{\frac{1}{l-1}Q_A}{\frac{1}{l(n-1)}Q_e} \qquad (3.8)$$

с числом степеней свободы (l - 1) и $l(n$ - 1). Гипотеза H_0: $m_1 = m_2 = \ldots = m_l$ не противоречит результатам наблюдений при заданном уровне значимости α, если

$$F > F_{1-\alpha}(l-1, l(n-1));$$

в этом случае считается, что фактор A не оказывает существенного влияния на признак X.

Результаты расчета обычно сводятся в таблицу.

Источник дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия	Выборочное значение статистики Фишера
Фактор A	Q_A	$l-1$	S_A^2	F
Остаток	Q_e	$l(n-1)$	S_e^2	
Общая	Q	$ln-1$	S^2	

3.2.
Многофакторный дисперсионный анализ

В двухфакторном дисперсионном анализе проверяется влияние на результативный признак X двух факторов A и B и их взаимодействия. Фактор A имеет l уровней A_j, $j = 1, \ldots, l$; фактор B – r уровней B_k, $k = 1, \ldots, r$. При каждом сочетании уровней A_jB_k делается n наблюдений. Общее число наблюдений nlr.

Проверяются три нулевые гипотезы: об отсутствии влияния на результативный признак X фактора A, фактора B и их взаимодействия AB.

Пусть X_{ijk} – результат i-го наблюдения ($i = 1, \ldots, n$) при j-ом уровне фактора A и k-ом уровне фактора B. Тогда средняя, соответствующая сочетанию уровней A и B:

$$\bar{x}_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ijk};$$

средняя, соответствующая уровню A_j:

$$\bar{x}_{0j} = \frac{1}{nr}\sum_{i=1}^{n}\sum_{k=1}^{r} x_{ijk} = \frac{1}{r}\sum_{k=1}^{r} x_{jk};$$

средняя, соответствующая уровню B_k:

$$x_{k0} = \frac{1}{nl}\sum_{i=1}^{n}\sum_{j=1}^{l} x_{ijk} = \frac{1}{l}\sum_{j=1}^{l} \bar{x}_{jk},$$

общая средняя

$$\bar{x} = \frac{1}{nlr}\sum_{i=1}^{n}\sum_{j=1}^{l}\sum_{k=1}^{r} x_{ijk} = \frac{1}{l}\sum \bar{x}_{0j} = \frac{1}{r}\sum_{k=1}^{r} x_{k0}.$$

По аналогии с однофакторным анализом справедливо тождество

$$Q = Q_A + Q_B + Q_{AB} + Q_e,$$

где общая сумма квадратов:

$$Q = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x})^2,$$

сумма квадратов, учитывающая влияние фактора *A*:

$$Q_A = nr\sum_j (\bar{x}_{0j} - \bar{x})^2;$$

сумма квадратов, учитывающая влияние фактора *B*:

$$Q_B = nl\sum_k (\bar{x}_{0k} - \bar{x})^2;$$

сумма квадратов, учитывающая взаимодействие факторов *A* и *B*:

$$Q_{AB} = n\sum_j \sum_k (\bar{x}_{jk} - x_{0j} - x_{k0} + \bar{x})^2;$$

остаточная сумма квадратов:

$$Q_e = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{jk})^2;$$

соответствующие дисперсии:

$$S^2 = \frac{Q}{nlr-1},$$

$$S_A{}^2 = \frac{Q_A}{l-1},$$

$$S_B{}^2 = \frac{Q_B}{r-1},$$

$$S_{AB}{}^2 = \frac{Q_{AB}}{(l-1)(r-1)},$$

$$S_e{}^2 = \frac{Q_e}{lr(n-1)}.$$

Проверка нулевых гипотез осуществляется с использованием статистик Фишера:

$$F_A = \frac{S_A^2}{S_e^2},$$

$$F_B = \frac{S_B^2}{S_e^2},$$

$$F_{AB} = \frac{S_{AB}^2}{S_e^2},$$

которые сравниваются с соответствующими квантилями. Например, гипотеза H_0 об отсутствии влияния взаимодействия факторов *A* и *B* на результативный признак *X* принимается, если

$$F_{AB} < F_{1-\alpha}[(l-1)(r-1), lr(n-1)].$$

Результаты оформляются в виде таблицы.

Источник дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия	Выборочное значение статистики Фишера
Фактор *A*	Q_A	$l-1$	S_A^2	F_A
Фактор *B*	Q_B	$r-1$	S_B^2	F_B
Взаимодействие *AB*	Q_{AB}	$(l-1)(r-1)$	S_{AB}^2	F_{AB}
Остаток	Q_e	$rl(n-1)$	S_e^2	
Общая	Q	$lrn-1$	S^2	

Алгоритм трехфакторного дисперсионного анализа аналогичен двухфакторному. Оценивается влияние факторов *A*, *B*, *C*, их попарного взаимодействия *AB*, *BC*, *AC* и общего взаимодействия *ABC* на результативный признак *X*. Фактор *A* имеет *l* уровней, фактор *B* – *r* уровней, фактор *C* – *q* уровней. При каждом сочетании уровней проводятся по n измерений, то есть общее число измерений $nlrq$.

Таблица трехфакторного анализа имеет вид:

Источник дисперсии	Сумма квад-ратов	Число степеней свободы	Дис-персия	Выборочное значение статистики Фишера
Фактор A	Q_A	$l-1$	S_A^2	F_A
Фактор B	Q_B	$r-1$	S_B^2	F_B
Фактор C	Q_C	$Q-1$	S_C^2	F_C
Взаимодействие AB	Q_{AB}	$(l-1)(r-1)$	S_{AB}^2	F_{AB}
Взаимодействие BC	Q_{BC}	$(q-1)(r-1)$	S_{BC}^2	F_{BC}
Взаимодействие AC	Q_{AC}	$(l-1)(q-1)$	S_{AC}^2	F_{AC}
Взаимодействие ABC	Q_{ABC}	$(l-1)(r-1)*$ $*(q-1)$	S_{ABC}^2	F_{ABC}
Остаток	Q_e	$lrq(n-1)$	S_e^2	
Общая	Q	$lrqn-1$	S^2	

Для проверки нулевой гипотезы, например, об отсутствии влияния общего взаимодействия ABC значение статистики Фишера

$$F_{ABC}=\frac{S_{ABC}^2}{S_e^2}=\frac{\frac{1}{(l-1)(r-1)(q-1)}Q_{ABC}}{\frac{1}{lrq(n-1)}Q_e}$$

сравнивается с квантилью

$$F_{1-\alpha}[(l-1)(r-1)(q-1),lrq(n-1)].$$