Основы дисперсионного анализа Исходные понятия. Однофакторный дисперсионный анализ. Понятие линейных контрастов.

Объектами исследования ***дисперсионного анализа*** являются *стохастические связи* между *откликом* и *факторами*, когда последние носят не количественный, а качественный характер. Примерами таких факторов могут служить:

– способ крепления детали при ее обработке;

– режим функционирования прибора;

– уровень квалификации оператора;

– методика обучения (или лечения) и т.д.

Чтобы подчеркнуть качественный характер факторов, будем их обозначать через $A, B, C, \ldots$, а отклик при этом — через X. Каждый из факторов имеет несколько ***уровней***, или градаций. Так, например, если X — это степень износа покрышки на колесе автомобиля, а выбранные факторы A и B — это тип дороги и тип рисунка протектора, то различные уровни фактора A — различные типы дорог, различные уровни фактора B — различные рисунки протектора.

Пусть наблюдаемый объект обладает таким свойством, которое характеризуется переменным (откликом) X и подвержено влиянию некоторых учитываемых факторов A, B и других, не контролируемых в данном эксперименте факторов. Задача дисперсионного анализа состоит в том, чтобы по результатам наблюдений за этим объектом дать ответ на вопрос: следует ли считать действие факторов A и B существенным (значимым) на фоне остальных (неучтенных) факторов или нет?

Формулировка и проверка соответствующих *статистических гипотез* для ответа на этот вопрос и является содержанием дисперсионного анализа.

В зависимости от числа анализируемых факторов различают ***однофакторный***, ***двухфакторный*** и т.д. ***дисперсионный анализ***. Мы здесь ограничимся рассмотрением однофакторного и двухфакторного дисперсионного анализа с постоянными (неслучайными) факторами.

8.2. Однофакторный дисперсионный анализ

Будем предполагать, что исследователя интересует степень влияния *фактора* A на *отклик* X. Для конкретности, пусть X — долговечность покрышки на колесе автомобиля, а фактор A — тип дорожного покрытия, который имеет l уровней (l — целое число).

Пусть $\mu_0 = \mathrm{M}X$ — среднее значение случайной величины X и пусть x_{ik} — значение X в i-м эксперименте, $i = \overline{1, n_k}$, соответствующем k-му *уровню фактора* A, $k = \overline{1, l}$. Тогда математическую модель *однофакторного дисперсионного анализа* можно представить в виде** (***линейная модель дисперсионного анализа***)

$$X_{ik} = \mu_0 + \alpha_k + \varepsilon_{ik}, \quad i = \overline{1, n_k}, \tag{8.1}$$

где α_k — вклад в величину X_{ik}, обусловленный действием фактора A (α_k — неслучайная величина); ε_{ik} — вклад в X_{ik}, обусловленный действием неучтенных факторов (случайные ошибки эксперимента, т.е. ε_{ik} — случайные величины). При этом $\sum\limits_{k=1}^{n} \alpha_k = 0$.

Относительно случайных величин ε_{ik} сделаем те же предположения, что и в *регрессионном анализе* (см. **7.1**, **7.3**):

– систематическая ошибка отсутствует, т.е. $\mathrm{M}\varepsilon_{ik} = 0$ для любых i и k;

– случайные ошибки эксперимента ε_{ik} не коррелированы между собой и имеют одинаковую (неизвестную) дисперсию, т.е.

$$\mathrm{M}(\varepsilon_{ik}\varepsilon_{jm}) = \begin{cases} \sigma^2, & i = j \text{ и } k = m); \\ 0, & i \neq j \text{ или } k \neq m; \end{cases}$$

– случайные ошибки эксперимента ε_{ik} имеют нормальный закон распределения с нулевым средним и неизвестной дисперсией σ^2, т.е.

$$\varepsilon_{ik} \sim N(0, \sigma^2).$$

Именно последнее допущение и позволит нам проводить проверку *статистических гипотез*, используя уже известные критерии, основанные на нормальном законе распределения наблюдаемых в эксперименте случайных величин. Разумеется, принятые допущения требуют последующей проверки. Однако на первом этапе исследования они являются вполне естественными.

С учетом принятых допущений о случайных ошибках эксперимента и на основании принятой модели (8.1) делаем заключение, что случайные величины X_{ik} имеют нормальный закон распределения со средним значением $\mathrm{M}X_{ik} = \mu_0 + \alpha_k$ и дисперсией $\mathrm{D}X_{ik} = \sigma^2$, $k = \overline{1,l}$.

Таким образом, действие фактора A проявляется в том, что для каждого его уровня k $(k = \overline{1,l})$ результаты наблюдений над случайной величиной (откликом) X можно рассматривать как *случайную выборку* $X_{1k}, X_{2k}, \ldots, X_{n_k k}$ *объема* n_k *из генеральной совокупности* X_k, причем каждая случайная величина X_k, $k = \overline{1,l}$, нормально распределена со средним значением $\mu_k = \mu_0 + \alpha_k$ и дисперсией σ^2.

Отсюда следует, что статистическая гипотеза H_0, предполагающая отсутствие влияния фактора A на отклик X, означает, что $\mu_k = \mu_0 + \alpha_k = \mu_0$, или $\alpha_k = 0$, $k = \overline{1,l}$. В качестве *альтернативной гипотезы* H_1 могут выступать различные предположения о значениях величин α_k или их некоторых линейных комбинаций — далее этот вопрос рассмотрен подробно.

Итак, задача проверки влияния фактора A на отклик X по результатам эксперимента сводится к следующей формализованной постановке, если принята модель наблюдений (8.1) и сформулированные выше предположения о случайных ошибках эксперимента.

Пусть $X_1, \ldots, X_l$ — независимые случайные величины и $X_k \sim N(\mu_k, \sigma^2)$, $k = \overline{1,l}$. Пусть для каждого $k = \overline{1,l}$ дана случайная выборка $X_{1k}, \ldots, X_{n_k k}$ из генеральной совокупности случайной величины X_k, которую далее мы будем называть ***k-й случайной выборкой.***

Требуется по этим данным проверить на заданном *уровне значимости* α гипотезу H_0: $\mu_1 = \mu_2 = \ldots = \mu_l = \mu_0$ (или, что то же самое, H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_l = 0$, если $\mu_k = \mu_0 + \alpha_k$, $k = \overline{1, l}$).

Для нашей интерпретации отклика X (долговечность покрышки) и фактора A (тип дорожного покрытия) каждая случайная величина X_k, $k = \overline{1, l}$, характеризует долговечность покрышки на дорогах с k-м типом покрытия. Отсутствие влияния фактора A, т.е. выполнение гипотезы H_0, означает, что на дорогах с любым типом покрытия средняя долговечность одна и та же. Если гипотеза H_0 неверна, то тип покрытия (фактор A) влияет на долговечность покрышки.

Статистику F используют для проверки гипотезы H_0: $\mu_1 = \ldots = \mu_l = \mu_0$. Гипотеза H_0 не противоречит результатам наблюдений, если выборочное значение $F_{\text{в}}$ статистики (8.4) меньше ее критического уровня $F_{\text{кр}} = F_{1-\alpha}(r_A, r_l)$, т.е. если

$$F_{\text{в}} \leqslant F_{\text{кр}} = F_{1-\alpha}(r_A, r_l).$$

Если же

$$F_{\text{в}} > F_{\text{кр}} = F_{1-\alpha}(r_A, r_l),$$

то гипотеза H_0 отклоняется и следует считать, что среди средних значений $\mu_1, \ldots, \mu_l$ имеются хотя бы два, не равных друг другу.

В случае принятия гипотезы H_0 в качестве несмещенных оценок параметров μ_0 и σ^2 можно взять соответственно $\overline{X}$ и $S_l^2 = Q_l(\vec{X}_n)/(n-l)$.

Результаты проверки гипотезы H_0 принято оформлять в виде так называемой ***таблицы дисперсионного анализа*** (табл. 8.1).

Таблица 8.1

Источник изменчивости	Сумма квадратов (СК)	Степени свободы	Средняя сумма квадратов	Статистика F
Между группами (фактор A)	$Q_A(\vec{X}_n) = \sum_{k=1}^{l} n_k(\overline{X}_k - \overline{X})^2$	$l-1$	$S_A^2(\vec{X}_n) = \frac{Q_A(\vec{X}_n)}{r_A}$	$F_в = S_A^2/S_l^2$
Внутри групп (ошибки)	$Q_l(\vec{X}_n) = \sum_{k=1}^{l}\sum_{i=1}^{n_k}(X_{ik} - \overline{X}_k)^2$	$n-l$	$S_l^2(\vec{X}_n) = \frac{Q_l(\vec{X}_n)}{r_l}$	
Общая сумма квадратов	$Q(\vec{X}_n) = \sum_{k=1}^{l}\sum_{i=1}^{n_k}(X_{ik} - \overline{X})^2$	$n-1$		$F_{кр} = F_{1-\alpha}(r_A, r_l)$

Пример 8.1. Три группы операторов ЭВМ обучались по трем различным методикам. После окончания срока обучения был проведен тестовый контроль случайно отобранных операторов из каждой группы. Получены следующие результаты (табл. 8.2).

Таблица 8.2

Номер группы k	Число ошибок, допущенных операторами, X_{ik}	Сумма $\sum_{i=1}^{n_k} X_{ik}$	Число контролируемых операторов n_k
1	1, 3, 2, 1, 0, 2, 1	10	7
2	2, 3, 2, 1, 4, –, –	12	5
3	4, 5, 3, –, –, –, –	12	3

Требуется на уровне значимости $\alpha = 0{,}05$ проверить гипотезу об отсутствии влияния различных методик обучения на результаты тестового контроля операторов. Предполагается, что выборки получены из независимых нормально распределенных совокупностей с одной и той же дисперсией.

В данном случае фактор A — это тип методики обучения, имеющий $l = 3$ уровня. Объем наблюдений $n = n_1 + n_2 + n_3 = 15$. Проверяется гипотеза H_0: $\mu_1 = \mu_2 = \mu_3$, где μ_k — математическое ожидание числа ошибок, допущенных операторами k-й группы.

Сперва вычисляем суммы

$$x_{\cdot,\cdot} = \sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik} = 10+12+12=34, \qquad \sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik}^2 = 104.$$

Затем, используя (8.2) и (8.3), находим

$$Q = \sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik}^2 - \frac{1}{n}x_{\cdot,\cdot}^2 = 104 - \frac{1}{15}\cdot 34^2 \approx 26{,}93,$$

$$Q_A = \sum_{k=1}^{l}\frac{1}{n_k}x_{\cdot k}^2 - \frac{1}{n}x_{\cdot,\cdot}^2 = 91{,}08 - \frac{1}{15}\cdot 34^2 \approx 14{,}02,$$

$$Q_l = Q - Q_A = 26{,}93 - 14{,}02 = 12{,}91.$$

Теперь вычисляем выборочное значение статистики (8.4):

$$f_{\text{в}} = \frac{Q_A/(l-1)}{Q_l/(n-l)} = \frac{14{,}02/2}{12{,}91} \approx 6{,}52.$$

Из таблицы квантилей распределения Фишера (см. табл. П.5) для уровня значимости $\alpha = 0{,}05$ и степеней свободы $r_A = l-1 = 2$, $r_l = n-l = 12$ находим $F_{\text{кр}} = F_{0{,}95}(2,12) = 3{,}89$. Так как $F_{\text{в}} = 6{,}52 > F_{\text{кр}}$, то гипотеза H_0 о равенстве средних отклоняется. Это означает, что исследуемые методики обучения операторов дают значимо различные результаты тестового контроля.

8.3. Понятие линейных контрастов

Если *гипотеза* H_0 о равенстве средних значений l нормальных *генеральных совокупностей* отклоняется (т.е. хотя бы в какой-то паре групп средние отличаются друг от друга), то требуется определить, какие именно группы имеют значимое различие средних. Для этой цели используются так называемые ***линейные контрасты***. Линейный контраст L определяется как линейная комбинация

$$L = \sum_{k=1}^{l} c_k \mu_k, \tag{8.5}$$

где c_k — постоянные, однозначно определяемые из формулировки проверяемых гипотез, причем $c_1 + \ldots + c_l = 0$.

Примерами линейных контрастов являются:

$L^{(1)} = \mu_1 - \mu_2$; здесь $c_1 = 1$, $c_2 = -1$, $c_3 = 0$, а выдвигаемая гипотеза $H_0^{(1)}$: $\mu_1 - \mu_2 = 0$;

$L^{(2)} = 0{,}5(\mu_1 + \mu_3) - \mu_2$; здесь $c_1 = c_3 = 0{,}5$, $c_2 = -1$, а выдвигаемая гипотеза $H_0^{(2)}$: $0{,}5(\mu_1 + \mu_3) - \mu_2 = 0$.

Таким образом, если гипотеза H_0: $\mu_1 = \mu_2 = \ldots = \mu_l$ отклоняется, то с помощью линейного контраста можно выдвинуть вспомогательные нулевые гипотезы относительно различных линейных комбинаций средних значений $\mu_1, \ldots, \mu_l$, образующих линейный контраст.

Любая такая гипотеза имеет вид H_0': $L = c_1\mu_1 + \ldots + c_l\mu_l$ при некотором заданном наборе постоянных c_k, для которых $c_1 + \ldots + c_l = 0$.

Нетрудно увидеть, что *несмещенной оценкой* линейного контраста L (при сделанных выше предположениях о *случайных ошибках* эксперимента ε_{ik}) является *оценка*

$$\widehat{L}(\vec{X}_n) = \sum_{k=1}^{l} c_k \overline{X}_k, \tag{8.6}$$

дисперсия которой (с учетом того, что $\mathbf{D}\,\overline{X}_k = \sigma^2/n_k$ и $\overline{X}_k$ — независимые случайные величины) равна

$$\mathbf{D}\,\widehat{L}(\vec{X}_n) = \sigma^2 \sum_{k=1}^{l} \frac{c_k^2}{n_k}. \tag{8.7}$$

При этом *статистика* $\widehat{L}(\vec{X}_n)$ имеет нормальный закон распределения со средним $L = c_1\mu_1 + \ldots + c_l\mu_l$ и дисперсией $\mathbf{D}\,\widehat{L}(\vec{X}_n)$, т.е.

$$\widehat{L}(\vec{X}_n) \sim N\big(L, \mathbf{D}\,\widehat{L}(\vec{X}_n)\big), \tag{8.8}$$

Следовательно,

$$T = \frac{\widehat{L}(\vec{X}_n) - L}{\sqrt{\mathbf{D}\,\widehat{L}(\vec{X}_n)}} \sim N(0,1), \tag{8.9}$$

т.е. статистика T имеет стандартное нормальное распределение.

Последнее утверждение следует из того, что *выборочные средние* $\overline{X}_k$ имеют нормальное распределение, $X_k \sim N(\mu_k, \sigma^2)$, $k = \overline{1,l}$, а линейная комбинация $\widehat{L} = c_1\overline{X}_1 + \ldots + c_l\overline{X}_l$ нормально распределенных случайных величин также распределена по нормальному закону с параметрами $\mathbf{M}\,\widehat{L}(\vec{X}_n) = L$ и $\mathbf{D}\,\widehat{L}(\vec{X}_n) = \sigma^2 c_1^2/n_1 + \ldots + c_l^2/n_l$. Кроме того, статистика $Q_l(\vec{X}_n)/\sigma^2$ име-

ет χ^2-*распределение* с числом степеней свободы $r_l = n - l$, т.е.

$$V = \frac{(n-l)S_l^2(\vec{X}_n)}{\sigma^2} \sim \chi^2(n-l), \tag{8.10}$$

и можно показать, что V и T — независимые случайные величины. На основании (8.9) и (8.10) приходим к следующему *критерию* проверки гипотезы H_0': $L = c_1\mu_1 + \ldots + c_l\mu_l = 0$.

Если гипотеза H_0' верна, то статистика $t = T/\sqrt{V(n-l)}$ имеет распределение Стьюдента с числом степеней свободы $n-l$, т.е.

$$t = \frac{\sum\limits_{k=1}^{l} c_k \overline{X}_k}{S_l(\vec{X}_n)\sqrt{\sum\limits_{k=1}^{l} \frac{c_k^2}{n_k}}} \sim S(n-l). \tag{8.11}$$

Таким образом, гипотезу H_0 следует отклонить на *уровне значимости* α (т.е. считать значимым отличие от нуля выбранной линейной комбинации средних $\mu_1, \mu_2, \ldots, \mu_l$), если *выборочное значение* $t_в$ статистики (8.11) по абсолютной величине превышает $t_{кр} = t_{1-\alpha/2}(n-l)$:

$$|t_в| > t_{кр} = t_{1-\alpha/2}(n-l).$$

Пример 8.2. В условиях примера 8.1 при двусторонних *альтернативных гипотезах* проверим гипотезы $H_0^{(1)}$: $\mu_1 = \mu_2$, $H_0^{(2)}$: $\mu_1 = \mu_3$, $H_0^{(3)}$: $\mu_2 = \mu_3$, $H_0^{(4)}$: $\frac{1}{2}(\mu_1 + \mu_3) = \mu_2$.

В соответствии с проверяемыми гипотезами $H_0^{(i)}$, $i = \overline{1,4}$, определим линейные контрасты

$$L_1 = \mu_1 - \mu_2 \quad (c_1 = 1,\ c_2 = -1,\ c_3 = 0);$$
$$L_2 = \mu_1 - \mu_3 \quad (c_1 = 1,\ c_2 = 0,\ c_3 = -1);$$
$$L_3 = \mu_2 - \mu_3 \quad (c_1 = 0,\ c_2 = 1,\ c_3 = -1);$$
$$L_4 = \frac{1}{2}(\mu_1 + \mu_2) - \mu_3 \quad \left(c_1 = \frac{1}{2},\ c_2 = \frac{1}{2},\ c_3 = -1\right).$$

Предварительно вычислим значения оценок линейных контрастов L_i, $i = \overline{1,4}$, и их дисперсий. Выборочные средние $\overline{x}_1 = 1{,}43$, $\overline{x}_2 = 2{,}4$, $\overline{x}_3 = 4$. Значение оценки дисперсии

$$S_l^2 = \frac{Q_l}{n-l} = \frac{12{,}91}{15-3} \approx 1{,}08.$$

Значения оценок контрастов и их дисперсий равны:

$$\widehat{L}_1 = 1{,}43 - 2{,}4 = -0{,}97, \qquad \mathbf{D}\,\widehat{L}_1 = 1{,}08\Big(\frac{1}{7} + \frac{1}{5}\Big) \approx 0{,}37;$$

$$\widehat{L}_2 = 1{,}43 - 4 = -2{,}57, \qquad \mathbf{D}\,\widehat{L}_2 = 1{,}08\Big(\frac{1}{7} + \frac{1}{3}\Big) \approx 0{,}51;$$

$$\widehat{L}_3 = 2{,}4 - 4 = -1{,}60, \qquad \mathbf{D}\,\widehat{L}_3 = 1{,}08\Big(\frac{1}{5} + \frac{1}{3}\Big) \approx 0{,}58;$$

$$\widehat{L}_4 = \frac{1}{2}(1{,}43 + 2{,}4) - 4 = -2{,}08,$$

$$\mathbf{D}\,\widehat{L}_4 = 1{,}08\Big(\frac{(1/2)^2}{7} + \frac{(1/2)^2}{5} + \frac{1}{3}\Big) \approx 0{,}45.$$

Следовательно, выборочные значения $|t_{\text{в}}^{(i)}|$ статистики (8.11) равны:

– для гипотезы $H_0^{(1)}$: $|t_{\text{в}}^{(1)}| = \left|\dfrac{\widehat{L}_1}{\sqrt{\mathbf{D}\widehat{L}_1}}\right| = \dfrac{0{,}97}{\sqrt{0{,}37}} \approx 1{,}595;$

– для гипотезы $H_0^{(2)}$: $|t_{\text{в}}^{(2)}| = \left|\dfrac{\widehat{L}_2}{\sqrt{\mathbf{D}\widehat{L}_2}}\right| = \dfrac{2{,}57}{\sqrt{0{,}51}} \approx 3{,}598;$

– для гипотезы $H_0^{(3)}$: $|t_{\text{в}}^{(3)}| = \left|\dfrac{\widehat{L}_3}{\sqrt{\mathbf{D}\widehat{L}_3}}\right| = \dfrac{1{,}60}{\sqrt{0{,}58}} \approx 2{,}101;$

– для гипотезы $H_0^{(4)}$: $|t_{\text{в}}^{(4)}| = \left|\dfrac{\widehat{L}_4}{\sqrt{\mathbf{D}\widehat{L}_4}}\right| = \dfrac{2{,}08}{\sqrt{0{,}45}} \approx 3{,}002.$

Критическое значение $t_{\text{кр}} = t_{0{,}975}(12) = 2{,}179$. Так как $|t_{\text{в}}^{(1)}| < t_{\text{кр}}$ и $|t_{\text{в}}^{(3)}| < t_{\text{кр}}$, то гипотезы $H_0^{(1)}$ и $H_0^{(3)}$ принимаются. Гипотезы $H_0^{(2)}$ и $H_0^{(4)}$ отклоняются, ибо $|t_{\text{в}}^{(2)}| > t_{\text{кр}}$ и $|t_{\text{в}}^{(4)}| > t_{\text{кр}}$.

Таким образом, значимо различны средние первой и третьей группы, а также среднее арифметическое средних для первых двух групп и среднее третьей группы.

Условия задач.

8.7. В трех магазинах, продающих товары одного вида, по данным товарооборота (в условных единицах) за 8 месяцев работы была составлена сводка (табл. 8.10). Проверьте на уровне значимости $\alpha = 0{,}01$ гипотезу H_0 о равенстве средних значений товарооборота для магазинов. Если гипотеза принимается, найдите несмещенные оценки для среднего и дисперсии товарооборота для всех трех магазинов.

Таблица 8.10

Магазин	Месяц							
	1	2	3	4	5	6	7	8
1	19	23	26	18	20	20	18	35
2	20	20	32	27	40	24	22	18
3	16	15	18	26	19	17	19	18

О т в е т: гипотезу о равенстве средних значений товарооборота следует принять; $\overline{x} = 22{,}08$, $S_l^2 = \frac{Q_l}{n-l} = 32{,}64$.

8.8. В условиях задачи 8.6 проверьте гипотезы: а) $H_0^{(1)}$: $\mu_1 = \mu_2 = 0$; б) $H_0^{(2)}$: $\mu_4 = \mu_5 = 0$; в) $H_0^{(3)}$: $\mu_3 = \mu_4 = 0$.

О т в е т: а) гипотезу следует принять; б) гипотеза отвергается; в) гипотеза отвергается.

8.9. В табл. 8.11 представлены результаты наблюдений над откликом X на пяти уровнях фактора A и трех уровнях фактора B. На уровне значимости $\alpha = 0{,}05$ проверьте гипотезы: а) H_0^A — фактор A не оказывает влияния на отклик; б) H_0^B — фактор B не оказывает влияния на отклик.

Таблица 8.11

Уровни фактора B (j)	Уровни фактора A (i)				
	1	2	3	4	5
1	3	3	6	6	8
2	8	3	7	6	3
3	6	6	8	7	8

О т в е т: а) гипотеза принимается; б) гипотеза принимается.