Глава 5
Дисперсионный анализ

5.1. Основные понятия

Предмет дисперсионный анализа – исследование статистических связей между *случайным откликом* (*реакцией системы*) *X* и *факторами A, B, C, ...*, действующими на систему и носящими не количественный, а качественный характер. Примеры факторов:

- способ крепления обрабатываемой детали;
- режим работы прибора;
- методика лечения;
- уровень квалификации оператора.

Каждый фактор имеет несколько *уровней* (градаций).

Пример 5.1. *X* – пробег шины до полного износа

Фактор	Уровни
A – тип дорожного покрытия	грунт
	гравий
	асфальт
	бетон
B - тип рисунка протектора	рис. 1
	рис. 2
	рис. 3

Задача дисперсионного анализа – по результатам наблюдений (измерений) выяснить: *является ли действие факторов A, B, C, ... существенным (значимым) по сравнению с другими (неучитываемыми) факторами*?

В зависимости от числа факторов различают однофакторный, двухфакторный и т.д. дисперсионный анализ. Мы ограничимся рассмотрением однофакторного анализа.

5.2. Однофакторный дисперсионный анализ

Исследуется влияние фактора *A* на отклик *X*. Обозначим:

$m_0 = M[X]$ *математическое ожидание* случайной величины *X*;

$\{x_{ik}\}_{i=1}^{n_k}$ - *выборка* значений случайной величины *X*, соответствующих *k*-му уровню фактора *A*, т.е. выборка из генеральной совокупности X_k, $k = 1, 2, \ldots, l$;

$\{X_{ik}\}_{i=1}^{n_k}$ - *случайная выборка* из генеральной совокупности X_k, $k = 1, 2, \ldots, l$.

Обычно используют следующую *линейную модель однофакторного дисперсионного анализа*:

$$X_{ik} = m_0 + \alpha_k + \varepsilon_{ik};\ i = 1, 2, \ldots, n_k,\ k = 1, 2, \ldots, l. \qquad (5.1)$$

Здесь:

α_k - *неслучайный* вклад *k*-го уровня фактора *A* в величину X_{ik},

ε_{ik} - *случайная ошибка эксперимента*, вызванная неучитываемыми факторами.

Относительно $\varepsilon_i^{(k)}$ делают те же предположения, что и в линейной модели регрессионного анализа:

- случайные ошибки ε_{ik} независимы;

- $M\left[\varepsilon_{ik}\right]=0$ (нет систематических ошибок);

- дисперсии ошибок ε_{ik} одинаковы: $D\left[\varepsilon_{ik}\right]=\sigma^2$ (измерения равноточны);

- случайные ошибки ε_{ik} имеют нормальное распределение: $\varepsilon_i^{(k)} \sim N\left(0,\sigma\right)$.

Относительно параметров α_k предполагается: $\sum_{k=1}^{l}\alpha_k=0$, поскольку $M\left[X\right]=m_0$.

Все допущения описанной модели требуют проверки, но на начальном этапе исследования они являются естественными.

Замечания. 1. Модель (5.1) позволяет использовать известные критерии проверки статистических гипотез, основанные на нормальности закона распределения исследуемых случайных величин. На основании данной модели имеем:

$$X_{ik} \sim N\left(m_0+\alpha_k,\sigma\right), \text{ т.е. и } X_k \sim N\left(m_0+\alpha_k,\sigma^2\right).$$

2. Отсутствию действия фактора *A* на отклик *X* соответствует гипотеза

$$H_0=\left\{m_k=m_0+\alpha_k=m_0,\ k=1,2,\ldots,l\right\} \text{ или } H_0=\left\{\alpha_k=0,\ k=1,2,\ldots,l\right\}.$$

В качестве альтернативных гипотез используются различные предположения относительно величин α_k или их линейных комбинаций.

3. Если фактор *A* имеет $l=2$ уровня, то гипотеза H_0 сводится к стандартному случаю гипотезы о равенстве математических ожиданий двух генеральных совокупностей. При $l>2$ *применяют однофакторный дисперсионный анализ*.

Пример 5.2. В условиях примера 5.1 для интерпретации отклика *X* и фактора *A* генеральная совокупность X_k характеризует пробег шины на дорогах с *k*-м типом покрытия. Если H_0 верна, то средний пробег покрышек не зависит от типа покрытия. Если неверна, то тип покрытия влияет на долговечность.

Пусть $\left\{X_{ik}\right\}_{i=1}^{n_k}$ - *k*-я случайная выборка. Рассмотрим выборочные средние:

$\bar{X}_k=\frac{1}{n_k}\sum_{i=1}^{n_k}X_{ik}$ (среднее *k*-й выборки) и

$\bar{\bar{X}}=\frac{1}{n}\sum_{k=1}^{l}\sum_{i=1}^{n_k}X_{ik}$ (общее среднее), $n=n_1+\ldots+n_l$ - общее число измерений.

Рассмотрим общую сумму квадратов случайных отклонений результатов X_{ik} измерений отклика *X* от общего выборочного среднего. Она представляется в виде

$$\sum_{k=1}^{l}\sum_{i=1}^{n_k}\left(X_{ik}-\bar{\bar{X}}\right)^2=\sum_{k=1}^{l}n_k\left(\bar{X}_k-\bar{\bar{X}}\right)^2+\sum_{k=1}^{l}\sum_{i=1}^{n_k}\left(X_{ik}-\bar{X}_k\right)^2. \qquad (5.2)$$

или

$$Q\left(\mathbf{X}^{(n)}\right)=Q_A\left(\mathbf{X}^{(n)}\right)+Q_\varepsilon\left(\mathbf{X}^{(n)}\right), \qquad (5.3)$$

где:

$\mathbf{X}^{(n)}$ - общая случайная выборка;

$Q\left(\mathbf{X}^{(n)}\right)$ - общая сумма квадратов отклонений отклика *X* от его среднего;

$Q_A\left(\mathbf{X}^{(n)}\right)$ - сумма квадратов отклонений средних по группам от общего среднего;

$Q_\varepsilon\left(\mathbf{X}^{(n)}\right)$ сумма квадратов отклонений результатов наблюдений от средних внутри групп.

◄Возведём в квадрат и просуммируем по *i* и по *k* равенство

$$X_{ik}-\bar{X}=\left(\bar{X}_k-\bar{X}\right)+\left(X_{ik}-\bar{X}_k\right):$$

$$\sum_{k=1}^{l}\sum_{i=1}^{n_k}\left(X_{ik}-\bar{X}\right)^2=\sum_{k=1}^{l}n_k\left(\bar{X}_k-\bar{X}\right)^2+\sum_{k=1}^{l}\sum_{i=1}^{n_k}\left(X_{ik}-\bar{X}_k\right)^2+$$

$$+2\sum_{k=1}^{l}\sum_{i=1}^{n_k}\left(X_{ik}-\bar{X}_k\right)\left(\bar{X}_k-\bar{X}\right).$$

Далее,

$$\sum_{k=1}^{l}\sum_{i=1}^{n_k}\left(X_{ik}-\bar{X}_k\right)\left(\bar{X}_k-\bar{X}\right)=\sum_{k=1}^{l}\left(\bar{X}_k-\bar{X}\right)\sum_{i=1}^{n_k}\left(X_{ik}-\bar{X}_k\right)=$$

$$=\sum_{k=1}^{l}\left(\bar{X}_k-\bar{X}\right)\left[\sum_{i=1}^{n_k}X_{ik}-n_k\bar{X}_k\right]=\sum_{k=1}^{l}\left(\bar{X}_k-\bar{X}\right)\left[n_k\bar{X}_k-n_k\bar{X}_k\right]=0.\blacktriangleright$$

Теорема 5.1. Если верна гипотеза $H_0=\left\{m_k=m_0+\alpha_k=m_0,\,k=1,2,\ldots,l\right\}$, то статистики $Q_A\left(\mathbf{X}^{(n)}\right)$ и $Q_\varepsilon\left(\mathbf{X}^{(n)}\right)$ независимы и

$$Q_A\left(\mathbf{X}^{(n)}\right)\sim\chi^2\left(l-1\right);\ Q_\varepsilon\left(\mathbf{X}^{(n)}\right)\sim\chi^2\left(n-l\right).$$

При этом статистики

$$S_A^2\left(\mathbf{X}^{(n)}\right)=Q_A\left(\mathbf{X}^{(n)}\right)\Big/\left(l-1\right)\ \text{и}\ S_\varepsilon^2\left(\mathbf{X}^{(n)}\right)=Q_\varepsilon\left(\mathbf{X}^{(n)}\right)\Big/\left(n-l\right)$$

являются несмещёнными оценками неизвестной дисперсии σ^2.

Замечания. 1. Из теоремы 5.1 следует, что если гипотеза H_0 верна, то статистика

$$F\left(\mathbf{X}^{(n)}\right)=\frac{Q_A\left(\mathbf{X}^{(n)}\right)\Big/\left(l-1\right)}{Q_l\left(\mathbf{X}^{(n)}\right)\Big/\left(n-l\right)}=\frac{S_A^2\left(\mathbf{X}^{(n)}\right)}{S_l^2\left(\mathbf{X}^{(n)}\right)}\sim F\left(l-1,n-l\right)\Big|H_0. \tag{5.4}$$

2. Оценка $S_A^2\left(\mathbf{X}^{(n)}\right)$ характеризует рассеивание средних $\bar{X}_k$, соответствующих разным уровням фактора *A*, а $S_\varepsilon^2\left(\mathbf{X}^{(n)}\right)$ - рассеивание результатов измерений, вызванное неучтёнными факторами. Поэтому значительное превышение $S_A^2\left(\mathbf{X}^{(n)}\right)$ над $S_\varepsilon^2\left(\mathbf{X}^{(n)}\right)$ говорит о существенном влиянии фактора *A*.

3. Из предыдущего замечания и формулы (5.4) получаем следующий критерий проверки гипотезы H_0 на уровне значимости α:

а) находим выборочное значение $F_{\text{выб}}$ статистики $F\left(\mathbf{X}^{(n)}\right)$ из (5.4);

б) с учётом замечания 2 используем *правосторонний критерий*: если $F_{\text{выб}}<F_{l-1,\,n-l,\,1-\alpha}$, где $F_{k_1,k_2,p}$ - квантиль порядка *p* распределения Фишера $F\left(k_1,k_2\right)$, то гипотеза H_0 принимается, иначе – отвергается.

4. В случае принятия гипотезы H_0 в качестве несмещённых оценок параметров m_0 и σ^2 можно использовать $\bar{X}$ и $S_\varepsilon^2\left(\mathbf{X}^{(n)}\right)$ соответственно.

Пример 5.3. Три группы студентов изучали один и тот же курс по трём разным методикам. После окончания учёбы был проведён контроль случайно выбранных студентов. Получены следующие результаты.

№ группы	Число контролируемых студентов n_k	Число ошибок, допущенных студентами	Сумма $\sum_{i=1}^{n_k} X_i^{(k)}$
1	7	1, 3, 2, 1, 0, 2, 1	10
2	5	2, 3, 2, 1, 4	12
3	3	4, 5, 3	12

На уровне значимости $\alpha = 0,05$ проверить гипотезу об отсутствии влияния выбранной методики обучения на его результаты.

◄В данном случае фактор A – это тип методики обучения, имеющий $l = 3$ уровня. Находим выборочные значения:

$$Q_{A\text{ выб}} \approx 14,02;\ Q_{\varepsilon\text{ выб}} \approx 12,91;\ F_{\text{выб}} = \frac{Q_{A\text{ выб}}/(l-1)}{Q_{\varepsilon\text{ выб}}/(n-l)} \approx \frac{14,02/2}{12,91/12} \approx 6,52.$$

Квантиль $F_{l-1,\,n-l,\,1-\alpha} = F_{2,\,12,\,0,95} = 3,89$. Поскольку $F_{\text{выб}} > 3,89$, гипотеза о равенстве средних отклоняется. Выбор методики обучения существенно влияет на результат обучения.►

5.3. Линейные контрасты

Предположим, что гипотеза о равенстве математических m_k ожиданий генеральных совокупностей X_k, $k = 1, 2, \ldots, l$ отвергнута. Это означает, что хотя бы для одной из пар X_{k_1}, X_{k_2} этих совокупностей величины m_{k_1} и m_{k_2} отличаются существенно. Тогда представляет интерес вопрос о том, какие именно группы совокупностей X_k имеют значимые различия средних. Для ответа на этот вопрос используют линейные контрасты.

Определение. *Линейным контрастом* называется линейная комбинация

$$L = \sum_{k=1}^{l} c_k m_k,$$

где c_k коэффициенты, определяемые из формулировки проверяемой гипотезы, причём $c_1 + c_2 + \ldots + c_l = 0$.

Пример 5.4. Пусть $l = 3$. Тогда линейными контрастами будут, например:

а) $L_1 = m_2 - m_3$. Здесь $c_1 = 0$, $c_2 = 1$, $c_3 = -1$;

б) $L_2 = (m_1 + m_2)/2 - m_3$. Здесь $c_1 = c_2 = 1/2$, $c_3 = -1$.

Итак, если гипотеза $H_0 = \{m_1 = m_2 = \ldots = m_l\}$ отвергнута, то выдвигаются вспомогательные гипотезы вида

$\tilde{H}_0 = \left\{ L = \sum_{k=1}^{l} c_k m_k = 0 \right\}$ с двухсторонними альтернативами $\tilde{H}_1 = \left\{ L = \sum_{k=1}^{l} c_k m_k \neq 0 \right\}$.

Если гипотеза $\tilde{H}_0$ верна, то статистика

$$T\left(\mathbf{X}^{(n)}\right) = \frac{\sum_{k=1}^{l} c_k \bar{X}_k}{S_l\left(\mathbf{X}^{(n)}\right)\sqrt{\sum_{k=1}^{l} \frac{c_k^2}{n_k}}} \sim St\left(n-l\right). \qquad (5.5)$$

Поэтому, с учётом альтернативы, гипотезу $\tilde{H}_0$ следует принять, если для выборочного значения $T_{\text{выб}}$ статистики $T\left(\mathbf{X}^{(n)}\right)$ из (5.5) выполняется соотношение

$$T_{\text{выб}} \in G_{1-\alpha} = \left(-t_{n-l,\,1-\alpha/2};\, t_{n-l,\,1-\alpha/2}\right) \text{ или } \left|T_{\text{выб}}\right| < t_{n-l,\,1-\alpha/2}, \qquad (5.6)$$

где $t_{k,p}$ - квантиль распределения $St\left(k\right)$. Иначе $\tilde{H}_0$ отвергается.

Пример 5.4. В условиях примера 5.3 проверить гипотезы:

$\tilde{H}_0^{(1)} = \{m_1 = m_2\}$; $\tilde{H}_0^{(2)} = \{m_1 = m_3\}$; $\tilde{H}_0^{(3)} = \{m_2 = m_3\}$; $\tilde{H}_0^{(4)} = \{(m_1 + m_2)/2 = m_3\}$

на уровне значимости $\alpha = 0{,}05$, используя двухсторонние альтернативы.

◄Данным гипотезам соответствуют следующие линейные контрасты:

$L_1 = m_1 - m_2$ ($c_1 = 1$, $c_2 = -1$, $c_3 = 0$); $L_2 = m_1 - m_3$ ($c_1 = 1$, $c_2 = 0$, $c_3 = -1$);

$L_3 = m_2 - m_3$ ($c_1 = 0$, $c_2 = 1$, $c_3 = -1$); $L_4 = (m_1 + m_2)/2 - m_3$ ($c_1 = c_2 = 1/2$, $c_3 = -1$).

Вычисляем выборочные значения статистики T из (5.5) для проверяемых гипотез:

$$\left|T_{\text{выб}}^{(1)}\right| \approx 1{,}595;\ \left|T_{\text{выб}}^{(2)}\right| \approx 3{,}598;\ \left|T_{\text{выб}}^{(3)}\right| \approx 2{,}101;\ \left|T_{\text{выб}}^{(4)}\right| \approx 3{,}002.$$

Из таблиц находим квантиль $t_{n-l,\,1-\alpha/2} = t_{12,\,0{,}975} = 2{,}179$.

В соответствии с (5.6) гипотезы $\tilde{H}_0^{(1)}$ и $\tilde{H}_0^{(3)}$ принимаются, а $\tilde{H}_0^{(2)}$ и $\tilde{H}_0^{(4)}$ отвергаются. Итак, значимо различаются средние первой и третьей групп, а также полусумма средних первых двух групп и среднее третьей группы.►