***Элементы дисперсионного анализа***

## *2.1. Сравнение нескольких средних. Понятие о дисперсионном анализе*

Дисперсионный анализ – статистический метод, предназначенный для выявления влияния отдельных факторов на результат эксперимента, а также для последующего планирования экспериментов.

Пусть генеральные совокупности $X_1, X_2, \dots, X_p$ распределены нормально и имеют одинаковую, хотя и неизвестную, дисперсию; математические ожидания также неизвестны, но могут быть различными. Требуется при заданном уровне значимости по выборочным средним проверить нулевую гипотезу $H_0: M(X_1) = M(X_2) = \cdots = M(X_p)$ о равенстве всех математических ожиданий. Другими словами, требуется установить, значимо или незначимо различаются выборочные средние. Казалось бы, для сравнения нескольких средних $(p > 2)$ можно сравнить их попарно. Однако с возрастанием числа средних возрастает и наибольшее различие между ними: среднее новой выборки может оказаться больше наибольшего или меньше наименьшего из средних, полученных до нового опыта. По этой причине для сравнения нескольких средних пользуются другим методом, который основан на сравнении дисперсий и поэтому назван *дисперсионным анализом* (в основном развит английским статистиком Р. Фишером).

На практике дисперсионный анализ применяют, чтобы установить, оказывает ли существенное влияние некоторый качественный фактор F, который имеет *p* уровней $F_1, F_2, \dots, F_p$ на изучаемую величину *X*. Например, если требуется выяснить, какой вид удобрений наиболее эффективен для получения наибольшего урожая, то фактор *F* – удобрение, а его уровни – виды удобрений.

Основная идея дисперсионного анализа состоит в сравнении «факторной дисперсии», порождаемой воздействием фактора, и «остаточной дисперсии», обусловленной случайными причинами. Если различие между этими дисперсиями значимо, то фактор оказывает существенное влияние на *X*; в этом случае средние наблюдаемых значений на каждом уровне (групповые средние) различаются также значимо.

Если уже установлено, что фактор существенно влияет на *X*, а требуется выяснить, какой из уровней оказывает наибольшее воздействие, то дополнительно производят попарное сравнение средних.

Иногда дисперсионный анализ применяется, чтобы установить однородность нескольких совокупностей (дисперсии этих совокупностей одинаковы по предположению; если дисперсионный анализ покажет, что и математические ожидания одинаковы, то в этом смысле совокупности однородны). Однородные же совокупности можно объединить в одну и тем самым получить о ней более полную информацию, следовательно, и более надежные выводы.

В более сложных случаях исследуют воздействие нескольких факторов на нескольких постоянных или случайных уровнях и выясняют влияние отдельных уровней и их комбинаций (*многофакторный анализ*).

Мы ограничимся простейшим случаем однофакторного анализа, когда на *X* воздействует только один фактор, который имеет *p* постоянных уровней.

### *2.2.Общая факторная и остаточная суммы квадратов отклонений*

Пусть на количественный нормально распределенный признак *X* воздействует фактор *F*, который имеет *p* постоянных уровней. Будем предполагать, что число наблюдений (испытаний) на каждом уровне одинаково и равно *q*.

Таблица 1

| Номер испытания | Уровни фактора $F_j$ | | | |
|---|---|---|---|---|
| | $F_1$ | $F_2$ | … | $F_p$ |
| 1 | $x_{11}$ | $x_{12}$ | … | $x_{1p}$ |
| 2 | $x_{21}$ | $x_{22}$ | … | $x_{2p}$ |
| … | … | … | … | … |
| q | $x_{q1}$ | $x_{q2}$ | … | $x_{qp}$ |
| Групповая средняя | $\bar{x}_{\text{гр}}$ | $\bar{x}_{\text{гр}2}$ | … | $\bar{x}_{\text{гр}p}$ |

Пусть наблюдалось $n = pq$ значений $x_{ij}$ признака *X*, где *i* – номер испытания ($i = 1,2,\dots q$), j – номер уровня фактора ($j = 1, 2, \dots, p$). Результаты наблюдений приведены в таблице 1.

Введем, по определению,

$$S_{\text{общ}} = \sum_{j=l}^{p}\sum_{i=l}^{q}\left(x_{ij} - \bar{x}\right)^2$$

(*общая сумма* квадратов отклонений наблюдаемых значений от общей средней $\bar{x}$),

$$S_{\text{ост}} = q\sum_{j=l}^{p}\left(\bar{x}_{\text{гр}j} - \bar{x}\right)^2$$

*(факторная сумма* квадратов отклонений групповых средних от общей средней, которая характеризует рассеяние «между группами»),

$$S_{\text{ост}} = \sum_{j=l}^{q}\left(x_{i1} - \bar{x}_{\text{гр1}}\right)^2 + \sum_{i=l}^{q}\left(x_{i2} - \bar{x}_{\text{гр2}}\right)^2 + \cdots + \sum_{i=l}^{q}\left(x_{ip} - \bar{x}_{\text{грр}}\right)^2$$

*(остаточная сумма* квадратов отклонений наблюдаемых значений группы от своей групповой средней, которая характеризует рассеяние «внутри групп»). Практически остаточную сумму находят по равенству:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}}.$$

Элементарными преобразованиями можно получить формулы, более удобные для расчетов:

$$S_{\text{общ}} = \sum_{i=l}^{p} P_j - \left[\left(\sum_{j=l}^{p} R_j\right)^2 /(pq)\right], \qquad (*)$$

$$(**) \qquad S_{\text{факт}} = \left[\left(\sum_{j=l}^{p} R_l^2\right)/q\right] - \left[\left(\sum_{j=l}^{p} R_j\right)^2 /(pq)\right],$$

где $P_j = \sum_{i=l}^{q} x_{ij}^2$ – сумма квадратов значений признака на уровне $F_j$;

$R_j = \sum_{i=l}^{q} x_{ij}$ – сумма значений признака на уровне $F_j$.

З а м е ч а н и е . Для упрощения вычислений вычитают из каждого наблюдаемого значения одно и то же число С, примерно равное общей средней. Если уменьшенные значения $y_{ij} = x_{ij} - C$, то

$$(***) \qquad S_{\text{общ}} = \sum_{j=l}^{p} Q_j - \left[\left(\sum_{j=l}^{p} T_j\right)^2 /(pq)\right],$$

$$(****) \qquad S_{\text{факт}} = \left[\left(\sum_{j=l}^{p} T_j^2\right)/q\right] - \left[\left(\sum_{j=l}^{p} T_j\right)^2 /(pq)\right],$$

$$Q_j = \sum_{i-l}^{q} y_{ij}^2$$

где – сумма квадратов уменьшенных значений признака на уровне $F_j$;

$T_j = \sum_{i=1}^{q} y_{ij}$ – сумма уменьшенных значений признака на уровне $F_j$.

Для вывода формул $(***)$ и $(****)$ достаточно подставить $x_{ij} = y_{ij} + C$ в соотношение $(*)$ и $R_j = \sum_{i=l}^{q} x_{ij} = \sum_{i=l}^{q} (y_{ij} + C) = \sum_{i=1}^{q} y_{ij} + qC = T_j + qC$ соотношение $(**)$.

П о я с н е н и я . 1. Убедимся, что $S_{\text{факт}}$ характеризует воздействие фактора *F*. Допустим, что фактор оказывает существенное влияние на *X*. Тогда группа наблюдаемых значений при одном определенном уровне, вообще говоря, отличается от групп наблюдений на других уровнях. Следовательно, различаются и групповые средние, причем они тем больше рассеяны вокруг общей средней, чем большим окажется воздействие фактора. Отсюда следует, что для оценки воздействия фактора целесообразно составить сумму квадратов отклонений групповых средних об общей средней (отклонение возводят в квадрат, чтобы исключить погашение положительных и отрицательных отклонений). Умножив эту сумму на *q*, получим $S_{\text{факт}}$. Итак, $S_{\text{факт}}$ характеризует воздействие фактора.

2. Убедимся, что $S_{\text{ост}}$ отражает влияние случайных причин. Казалось бы, наблюдения одной группы не должны различаться. Однако, поскольку на *X,* кроме фактора *F*, воздействуют и случайные причины наблюдения одной и той же группы, вообще говоря, различны и, значит, рассеяны вокруг своей групповой средней. Отсюда следует, что для оценки влияния случайных причин целесообразно составить сумму квадратов отклонений наблюдаемых значений каждой группы от своей групповой средней, т.е. $S_{\text{ост}}$. Итак, $S_{\text{ост}}$ характеризует воздействие случайных причин.

3. Убедимся, что $S_{\text{общ}}$ отражает влияние и фактора и случайных причин. Будем рассматривать все наблюдения как единую совокупность. Наблюдаемые значения признака различны вследствие воздействия фактора и случайных причин. Для оценки этого воздействия целесообразно составить сумму квадратов отклонений наблюдаемых значений от общей средней, т.е. $S_{\text{общ}}$.

Итак, $S_{\text{общ}}$ характеризует влияние фактора и случайных причин.

Приведем пример, который наглядно показывает, что факторная сумма отражает влияние фактора, а остаточная – влияние случайных причин.

**Пример.** Двумя приборами произведены по два измерения физической величины, истинный размер которой равен *x*. Рассматривая в качестве фактора систематическую ошибку *C*, а в качестве его уровней – систематические ошибки $C_1$ и $C_2$ соответственно первого и второго прибора, показать, что $S_{\text{факт}}$ определяется систематическими, а $S_{\text{ост}}$ – случайными ошибками измерений.

**Решение.** Введем обозначения: $\alpha_1, \alpha_2$ – случайные ошибки первого и второго измерений первым прибором; $\beta_1, \beta_2$ – случайные ошибки первого и второго измерений вторым прибором.

Тогда наблюдения значения результатов измерений соответственно равны (первый индекс при *x* указывает номер измерения, а второй – номер прибора):

$$x_{11} = x + C_1 + \alpha_1, x_{21} = x + C_1 + \alpha_2;\ x_{12} = x + C_2 + \beta_1, x_{22} = x + C_2 + \beta_2.$$

Средние значения измерений первым и вторым приборами соответственно равны:

$$\bar{x}_{\text{гр 1}} = x + C_1 + [(\alpha_1 + \alpha_2)/2] = x + C_1 + \alpha,$$
$$\bar{x}_{\text{гр 2}} = x + C_2 + [(\beta_1 + \beta_2)/2] = x + C_2 + \beta.$$

Общая средняя

$$\bar{x} = \left(\bar{x}_{\text{гр 1}} + \bar{x}_{\text{гр 2}}\right)/2 = x + [(C_1 + C_2)/2] + [(\alpha + \beta)/2],$$

факторная сумма

$$S_{\text{факт}} = \left(\bar{x}_{\text{гр 1}} - \bar{x}\right)^2 + \left(\bar{x}_{\text{гр 2}} - \bar{x}\right)^2.$$

Подставив величины, заключенные в скобках, после элементарных преобразований получим

$$S_{\text{факт}} = [(\mathrm{C}_1 - \mathrm{C}_2)^2/2] + (C_1 - C_2)(\alpha - \beta) + [(\alpha - \beta)^2/2].$$

Мы видим, что $S_{\text{факт}}$ определяется главным образом, первым слагаемым (поскольку случайные ошибки измерений малы) и, следовательно, действительно отражает влияние фактора *C*.

Остаточная сумма

$$S_{\text{ост}} = \left(x_{11} - \bar{x}_{\text{гр }1}\right)^2 + \left(x_{21} - \bar{x}_{\text{гр }1}\right)^2 + \left(x_{12} - \bar{x}_{\text{гр }2}\right)^2 + \left(x_{22} - \bar{x}_{\text{гр }2}\right)^2.$$

Подставив величины, заключенные в скобках, получим

$$S_{\text{ост}} = [(\alpha_1 - \alpha)^2 + (\alpha_2 - \alpha)^2] + [(\beta_1 - \beta)^2 + (\beta_2 - \beta)^2].$$

Мы видим, что $S_{\text{ост}}$ определяются случайными ошибками измерений и, следовательно, действительно отражает влияние случайных причин.

З а м е ч а н и е . То, что $S_{\text{ост}}$ порождается случайными причинами, следует также из равенства:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}}.$$

Действительно, $S_{\text{общ}}$ является результатом воздействия фактора и случайных причин; вычитая $S_{\text{факт}}$, мы исключаем влияние фактора. Следовательно, «оставшаяся часть» отражает влияние случайных причин.