4 ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ разработан для сельскохозяйственных и биологических исследований Р.А. Фишером на основе открытого им закона распределения отношения средних квадратов (дисперсий)

$$\mathbf{F} = \frac{\mathbf{S_1^2}}{\mathbf{S_2^2}}. \qquad (4.1)$$

В дальнейшем этот метод получил распространение во всех сферах исследований. Дисперсионный метод широко используется для планирования эксперимента и статистической обработки его данных. Если в недалеком прошлом считали, что роль математика состоит лишь в анализе экспериментальных данных, то работы Р.А. Фишера коренным образом изменили эту точку зрения и в настоящее время статистическое планирование опыта в соответствии с требованиями дисперсионного анализа и математическая интерпретация результатов – непременные условия успешного эксперимента. Статистически обоснованный план эксперимента определяет и метод математического анализа результатов. Поэтому современный эксперимент нельзя правильно спланировать, не зная основ дисперсионного анализа.

При дисперсионном анализе одновременно обрабатывают данные нескольких выборок (вариантов) составляющих единый статический комплекс, сформированный в виде специальной рабочей таблицы.

Сущностью дисперсионного анализа является расчленение общей суммы квадратов отклонений и общего числа степеней на части – компоненты, соответствующие структуре эксперимента, и оценка значимости действия и взаимодействия изучаемых факторов по **F**-критерию.

Если обрабатывают однофакторные статистические комплекы состоящие из нескольких независимых выборок, например ℓ - вариантов, то общая изменчивость результа-

тивного признака, измеряемая общей суммой квадратов отклонений $\mathbf{C_V}$ расчленяется на компоненты: варьирование между выборками (вариантов) и внутри выборок $\mathbf{C_Z}$.

Следовательно, в общей форме изменчивость признака может быть представлена выражением

$$\mathbf{C_Y = C_V + C_Z}. \qquad (4.2)$$

Здесь вариация между выборками (вариантами) представляет ту часть общей дисперсии, которая обусловлена действием изучаемых факторов, а дисперсия внутри выборок характеризует случайное варьирование изучаемого признака, то есть ошибку измерения.

Общее число степеней свободы равняется (**N–1**), где **N** – общее количество обрабатываемых данных. Оно также расчленяется на две части- степени свободы для вариантов (ℓ–1) где ℓ - количество вариантов и для случайного варьирования (**N** - ℓ)

$$\mathbf{N\text{-}1=(}\ell\mathbf{-1) + (N-}\ell\mathbf{),} \qquad (4.3)$$

При наличии повторений общая сумма квадратов отклонений дисперсий раскладывается на три части: варьирование повторностей $\mathbf{C_P}$, вариантов $\mathbf{C_V}$ и случайное $\mathbf{C_Z}$

$$\mathbf{C_Y = C_P + C_V + C_Z}, \qquad (4.4)$$

а

$$\mathbf{N-1=(}\ell\mathbf{-1)+(n-1)+(n-1)(}\ell\mathbf{-1)}. \qquad (4.5)$$

Проверка

Пусть в опыте изучалось 5 вариантов ($\ell = \mathbf{5}$) в пятикратной повторности (**n**=5). В этом случае общее число данных **N**=5×5= 25, а число степеней свободы, учитывая, что на ряд дисперсий накладывается одна связь составит

25-1=(5-1)+(5-1)+(5-1)(5-1)=24.

Сумму квадратов отклонений статистического комплекса с ℓ- вариантами и **n**-повторениями находят обычно в следующей последовательности. В исходной таблице определяют

суммы **P** (повторений), суммы **V** (по вариантам) и общую сумму всех наблюдений $\sum \mathbf{x}$, затем вычисляют

1. Общее число всех наблюдений

$$\mathbf{N} = \ell \cdot \mathbf{n}, \tag{4.6}$$

2. Корректирующий фактор

$$\mathbf{C} = \frac{\sum \mathbf{x}^2}{\mathbf{N}}. \tag{4.7}$$

Этот фактор представляет собой второй член формулы при определении суммы квадратов отклонений:

$$\sum (\mathbf{x} - \bar{\mathbf{x}})^2 = \sum \mathbf{x}^2 - \left(\frac{\sum \mathbf{x}^2}{\mathbf{N}} \right) = \sum \mathbf{x}^2 - \mathbf{c}. \tag{4.8}$$

Введение в технику вычислений суммы квадратов отклонений корректирующего фактора значительно упрощает ее.

3. Общую сумму квадратов

$$\mathbf{C_Y} = \sum \mathbf{x}^2 - \mathbf{C}, \tag{4.9}$$

4. Сумму квадратов для повторений

$$\mathbf{C_P} = \frac{\sum \mathbf{P}^2}{\ell} - \mathbf{C}, \tag{4.10}$$

5. Сумму квадратов для вариантов

$$\mathbf{C_V} = \frac{\sum \mathbf{V}^2}{\mathbf{n}} - \mathbf{C}, \tag{4.11}$$

6. Сумму квадратов для ошибки (остаток)

$$\mathbf{C_Z} = \mathbf{C_Y} - \mathbf{C_p} - \mathbf{C_V}. \tag{4.12}$$

Две последние суммы квадратов отклонений $\mathbf{C_V}$ и $\mathbf{C_Z}$ делят на соответствующие им степени свободы, то есть приводят к сравниваемому виду – одной степени свободы вариации. В результате получаются два средних квадрата отклонений представляющие собой дисперсии:

Вариантов

$$\mathbf{S_V^2} = \frac{\mathbf{C_V}}{\ell - \mathbf{1}}, \tag{4.13}$$

Ошибки

$$\mathbf{S^2_{ош} = \frac{C_Z}{(n-1)(\ell-1)}}. \qquad (4.14)$$

Эти средние квадраты (дисперсии) и используются в дисперсионном анализе для оценки значимости изучаемых факторов. Оценка проводится путем сравнивания дисперсий вариантов $\mathbf{S^2_V}$ с дисперсией ошибки $\mathbf{S^2_{ош}}$ по критерию Фишера

$$\mathbf{F = \frac{S^2_V}{S^2_{ош}}}. \qquad (4.15)$$

Таким образом, за базу – единицу сравнения принимают средний квадрат случайной дисперсии, которая определяет случайную ошибку эксперимента. При этом проверяемой нулевой гипотезой служит предположение, что все выборочные средние являются оценками одной генеральной средней, и, следовательно, различия между ними несущественны.

Если

$$\mathbf{F_{факт} = \frac{S^2_V}{S^2_{ош}} < F_{теор}}, \qquad (4.16)$$

то нулевая гипотеза не отвергается; между всеми выборочными средними нет существенных различий. На этом проверка заканчивается.

Нулевая гипотеза отвергается, когда

$$\mathbf{F_{факт} = \frac{S^2_V}{S^2_{от}} > F_{теор}}. \qquad (4.17)$$

В этом случае дополнительно оценивают существенность частных различий по **НСР** и определяется между какими средними имеются значимые разности.

При обработке данных многофакторных опытов дисперсионный анализ позволяет выявить не только главные эффекты, но и оценить существенность их взаимодействия.

Пример: Трехфакторный опыт с двумя градациями факторов А; В; С в трехкратной повторности. Вначале определя-

ют суммы квадратов отклонений: общую для повторений, вариантов и остатка (ошибки) по формулам

Общее количество наблюдений

$$\mathbf{N = \ell_A \ell_B \ell_C \cdot n = 2 \cdot 2 \cdot 2 \cdot 3 = 24}. \quad (4.18)$$

$$\mathbf{C = \left(\sum x\right)^2 : N}, \quad (4.19)$$

$$\mathbf{C_Y = \sum x^2 - c}, \quad (4.20)$$

$$\mathbf{C_P = \sum P^2 : \ell_A \ell_B \ell_C - C}, \quad (4.21)$$

$$\mathbf{C_V = \sum V^2 : n - c}, \quad (4.22)$$

$$\mathbf{C_Z = C_Y - C_P - C_V}. \quad (4.23)$$

Далее раскладывают сумму квадратов отклонений – $\mathbf{C_V}$ на главные эффекты A; B; C и их взаимодействия. Для этого составляют новую таблицу, куда вместо первичных данных вносят их суммы V и определяют суммы для всех главных эффектов и их взаимодействий. Затем определяют:

$$\mathbf{C_A = \sum A^2 : \ell_B \ell_C \cdot n - c}, \quad (4.24)$$

$$\mathbf{C_B = \sum B^2 : \ell_A \ell_C \cdot n - c}, \quad (4.25)$$

$$\mathbf{C_C = \sum C^2 : \ell_A \ell_B \cdot n - c}, \quad (4.26)$$

$$\mathbf{C_{AB} = \sum AB^2 : \ell_C n - C_A - C_B - C}, \quad (4.26)$$

$$\mathbf{C_{AC} = \sum AC^2 : \ell_B n - C_A - C_C - C}, \quad (4.27)$$

$$\mathbf{C_{BC} = \sum BC^2 : \ell_A n - C_B - C_C - C}. \quad (4.28)$$

Сумму квадратов отклонений для тройного взаимодействия находят по разности:

$$\mathbf{C_{ABC} = C_V - \left(C_A + C_B + C_C + C_{AB} + C_{AC} + C_{BC}\right)}. \quad (4.29)$$

При этом общее число степеней свободы $\mathbf{(N-1)}$ равное 23 раскладывается в следующем порядке: общая – 23; повторений – 2; A; B; C; AB; AC; BC; ABC по одной, остатка (ошибки) =14.

Значимость действия факторов и их взаимодействий оценивается по **F** критерию. Критерий $\mathbf{F_{теор}}$ для принятого уровня значимости определяют из таблиц, исходя из числа степеней

свободы главных эффектов, их взаимодействий и числа степеней остатка. В данном случае это соответственно 1 и 14. Различия существенны при $\mathbf{F_{фак} > F_{теор.}}$ Расчетное значение $\mathbf{F_{фак}}$ определяется делением среднего квадрата отклонений на средний квадрат остатка (ошибки). Для оценки существенности частных различий определяется $\mathbf{НСР_{05}}$

$$\mathbf{НСР_{05} = t_{05} \cdot \sqrt{\frac{2S^2}{n}}}, \qquad (4.30)$$

где **S** – дисперсия ошибки (остатка);

n – число повторностей.

Для главных эффектов

$$\mathbf{НСР_{05} = t_{05} \sqrt{\frac{2S^2}{\ell_A \cdot n}}}. \qquad (4.31)$$

Для парных взаимодействий

$$\mathbf{НСР_{05} = t_{05} \cdot \sqrt{\frac{2S^2}{\ell_A \ell_B n}}}. \qquad (4.32)$$

Дисперсионный анализ дает возможность получить представление о степени или доле влияния того или иного фактора в общей дисперсии - признаке, которую принимают за единицу или 100%.

Влияние вариантов

$$\mathbf{\eta_V^2 = \frac{C_V}{C_Y}}, \qquad (4.33)$$

Влияние повторений

$$\mathbf{\eta_P^2 = \frac{C_P}{C_Y}}, \qquad (4.34)$$

Влияние случайных факторов

$$\mathbf{\eta_Z^2 = \frac{C_Z}{C_Y}}, \qquad (4.35)$$

$$\mathbf{\eta_Y^2 = \eta_V^2 + \eta_P^2 + \eta_Z^2 = 1 \cdot (100\%)}. \qquad (4.36)$$

Корреляционное отношение, характеризующее частоту связи результативного признака с факториальным, определяется по выражению

$$\eta_{\text{э}} = \sqrt{\eta_V^2} = \sqrt{\frac{C_V}{C_Y}}, \qquad (4.37)$$

где η_V^2 - индекс детерминации, показывающий долю его варьирования.

Дисперсионный анализ имеет следующие преимущества перед методом попарных сравнений по критерию Стьюдента.

1. Вместо индивидуальных ошибок, средних по каждому варианту, здесь используется обобщенная ошибка средних, которая опирается на большее число наблюдений и, следовательно, является более надежной базой для оценок.

2. Методом дисперсионного анализа можно обрабатывать данные простых и сложных, однофакторных и многофакторных опытов.

3. Дисперсионный анализ позволяет компактно в виде существенных разностей представить итоги статистической обработки.

Не следует забывать, что дисперсионный анализ базируется на принципе рендомизации (случайности). Особенно это касается полевых опытов. Теория требует, чтобы все наблюдения были независимы. В этом случае дисперсионный анализ дает правильную несмещенную оценку ошибки эксперимента. Правильное использование дисперсионного анализа для обработки экспериментального материала предполагает однородность дисперсии по вариантам, нормальное или близкое к нему распределение варьирующих величин, значения которых получают независимо одно от другого. Это предположение обычно оправдывается, если для наблюдения используется одна и та же методика, одни и те же приборы. Если стабильность дисперсий вызывает сомнение, нужно провести специальное исследование с помощью критериев Бартле-

та или Кохрана. Наиболее удобным является критерий Кохрана, который предложил рассматривать отношение максимальной дисперсии к сумме всех остальных

$$q = \frac{S^2_{V\cdot max}}{S_1^2 + S_2^2 + S_K^2}, \qquad (4.38)$$

и нашел распределение величины **q**. Оказалось, что оно зависит только от общего количества дисперсий – **K** и от числа степеней свободы, по которым определена каждая дисперсия S_V^2. Имеются таблицы, где для выбранного уровня значимости, числа степеней свободы и числа анализируемых дисперсий – **K** приведены квантили распределения Кохрана $q_1 - \rho$. Если найденное значение **q** окажется больше табличного (для выбранного уровня значимости), то нулевую гипотезу нужно отбросить и расхождение между дисперсиями считать значимым. В этом случае нужно преобразовывать исходные данные. Трансформация данных дает возможность уменьшить пределы варьирования, устранить неоднородность дисперсий и провести сравнение результатов более точно. Чаще всего применяются следующие преобразования:

1. Логарифмические, когда каждое значение **x** трансформируется в $\lg x$ или $\lg(X+1)$, если некоторые данные равны нулю.

2. Трансформация опытных данных путем извлечения $\sqrt{x}$ или $\sqrt{x}+1$ (при нулевых значениях некоторых данных).

Преобразованные данные вновь проверяются на однородность и при ее достижении $(q < q_1 - \rho)$, обрабатываются по схеме дисперсионного анализа, и после оценки существенности частных различий вновь переходят к первоначальным данным.