Лекция 3

ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ. Постановка задачи

Дисперсионный анализ как метод исследования появился в работах Р. Фишера (1918-1935 гг.) в связи с исследованиями в сельском хозяйстве для выявления условий, при которых испытываемый сорт сельскохозяйственной культуры даёт максимальный урожай. (В агрономических исследованиях первый фактор - сорт почвы, второй фактор - способ обработки.)

Дальнейшее развитие дисперсионный анализ получил в работах Йетса.

Сейчас теорию дисперсионного анализа можно считать в достаточной мере сформировавшейся, но способы организации эксперимента и вычислительные схемы продолжают совершенствоваться.

Дисперсионный анализ

Дисперсионный анализ – это статистический метод анализа результатов наблюдений, зависящий от разных, одновременно действующих факторов, выбора наиболее важных факторов и оценки их влияния.

Идея дисперсионного анализа заключается в разложении общей дисперсии случайной величины на независимые случайные слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия.

Последующее сравнение этих дисперсий позволяет оценить существенность влияния факторов на исследуемую величину.

Дисперсионный анализ. Постановка задачи

В любом ряде испытаний имеется несколько факторов, вызывающих изменчивость средних значений наблюдаемых случайных величин - ***результативных признаков***.

Эти факторы могут принадлежать одному или нескольким источникам изменчивости (например, расположение торговых заведений в центре и на окраине города, изменения в законодательстве, разные климатические условия, разные уровни образования и т. п.).

Даже при самом тщательном исследовании не удается выявить все источники изменчивости, а иногда в этом нет необходимости или смысла.

Но при наличии опыта у эксперта и в зависимости от цели исследования ***всегда можно выдвинуть гипотезу о существовании влияния тех или иных факторов на результативный признак***.

Дисперсионный анализ. Постановка задачи

В дисперсионном анализе используются следующие термины:

фактор (X) – то, что, как мы считаем, должно оказывать влияние на результат (результативный признак) Y;

уровень фактора (или способ обработки, иногда буквально, например - способ обработки почвы) - значения (X_i , i = 1,2,...I), которые может принимать фактор;

отклик – значение измеряемого признака (величина результата Y_i).

Техника дисперсионного анализа меняется в зависимости от числа изучаемых независимых факторов.

Если факторы, вызывающие изменчивость среднего значения признака, принадлежат одному источнику, то мы имеем простую группировку, или однофакторный дисперсионный анализ, и далее, соответственно, двойная группировка - двухфакторный дисперсионный анализ, трехфакторный дисперсионный анализ, m-факторный.

Факторы в многофакторном анализе принято обозначать латинскими буквами: A, B, C и т. д.

Дисперсионный анализ. Постановка задачи

Задача дисперсионного анализа - исследование влияния тех или иных факторов (или уровней факторов) на изменчивость средних значений наблюдаемых случайных величин.

Сущность дисперсионного анализа. Дисперсионный анализ состоит в выделении и оценке отдельных факторов, вызывающих изменчивость.

С этой целью производят разложение общей дисперсии σ^2 наблюдаемой частичной совокупности (общей дисперсии признака), вызванной всеми источниками изменчивости, на составляющие дисперсий, порожденные независимыми факторами. Каждая из этих составляющих дает оценку дисперсии σ^2_A, σ^2_B,..., вызванную конкретным источником изменчивости, в общей совокупности.

Для проверки значимости этих составляющих оценок дисперсии их сравнивают с общей дисперсией в общей совокупности (по критерию Фишера).

Дисперсионный анализ. Постановка задачи

В дисперсионном анализе рассматривается гипотеза:

H_0 – ни один из рассматриваемых факторов не оказывает влияния на изменчивость признака.

Значимость каждой из оценок дисперсии проверяется по величине её отношения к оценке случайной дисперсии и сравнивается с соответствующим критическим значением, при уровне значимости α, с помощью таблиц критических значений **F-распределения Фишера - Снедекора** (табулировано).

Гипотеза H_0 относительно того или иного источника изменчивости отвергается, если $F_{расч.} > F_{кр.}$

Дисперсионный анализ. Постановка задачи

В дисперсионном анализе рассматриваются эксперименты 3-х видов:

а) эксперименты, в которых ***все факторы имеют систематические (фиксированные) уровни***;

б) эксперименты, в которых ***все факторы имеют случайные уровни***;

в) эксперименты, в которых ***есть факторы, имеющие случайные уровни, а также факторы, имеющие фиксированные уровни***.

Случаи а), б), в) соответствуют трем моделям, которые рассматриваются в дисперсионном анализе.

Применение дисперсионного анализа предполагает, что:

$M(\varepsilon_{ij})=0$, $D(\varepsilon_{ij})=\sigma^2=\text{const}$, $\varepsilon_{ij} \to N(0, \sigma^2)$ или $x_{ij} \to N(a, \sigma^2)$.

ε_{ij} - вариация результатов внутри отдельного уровня фактора.

Однофакторный дисперсионный анализ

Предположим, что совокупности случайных величин имеют нормальное распределение и равные дисперсии.

Пусть имеется m таких совокупностей, из которых произведены выборки объемом $n_1,n_2,\ldots,n_m$. Обозначим выборку из i-ой совокупности $(x_{i1},x_{i2},\ldots,x_{in})$.

Тогда все выборки можно записать в виде следующей таблице, которая называется ***матрицей наблюдений***.

Матрица наблюдений

Количество элементов совокупности (n) Количество совокупностей (m)	1	2	…	j	…	n
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n1}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n2}
…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ini}
…	…	…	…	…	…	…
m	x_{m1}	x_{m2}	…	x_{mj}	…	x_{mnm}

Средние выборок обозначим через $\beta_1, \beta_2, \ldots, \beta_m$.

Проверим нулевую гипотезу о равенстве средних:

$$H_0: \beta_1=\beta_2=\ldots=\beta_m.$$

Гипотеза H_0 проверяется сравнением внутригрупповых и межгрупповых дисперсий по F-критерию Фишера.

Если расхождение между ними значительно, то нулевая гипотеза принимается.

В противном случае гипотеза о равенстве средних отвергается и делается заключение о том, что различие в средних обусловлено не только случайностями выборок, но и действием исследуемого фактора.

Рассмотрим структуру межгрупповой и внутригрупповой дисперсии.

Для этого найдем сначала средние арифметические членов каждой совокупности:

$$\bar{x}_{i*} = \frac{\sum_{j=1}^{n_1} x_{ij}}{n_i}; \ldots \bar{x}_{ij} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}; \ldots \bar{x}_{mj} = \frac{\sum_{j=1}^{n_m} x_{mj}}{n_m}.$$

Общую среднюю арифметическую всех m совокупностей рассчитываем по формуле:

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} \bar{x}_{i*}.$$

Найдем сумму квадратов отклонений x_{ij} от $\bar{x}$:

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2}_{Q} = \underbrace{n\sum_{i=1}^{m}(\bar{x}_{i*}-\bar{x})^2}_{Q_1} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})^2}_{Q_2},$$

Слагаемое Q_1 является суммой квадратов разностей между средними отдельных совокупностей и общей средней всей совокупности наблюдений.

Эта сумма называется ***суммой квадратов отклонений между группами*** и характеризует систематическое расхождение между совокупностями наблюдений.

Величину Q_1 называют иногда ***рассеиванием по факторам*** (т.е. за счет исследуемого фактора).

Слагаемое Q_2 представляет собой сумму квадратов разностей между отдельными наблюдениями и средней соответствующей совокупности.

Эта сумма называется ***суммой квадратов отклонений внутри группы.***

Она характеризует ***остаточное рассеивание*** случайной погрешности совокупностей.

Наконец, Q называется ***общей*** или ***полной суммой квадратов отклонений отдельных наблюдений от общей средней .***

Оценим дисперсии:

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x})^2 = \frac{Q_1}{m-1},$$

$$s_2^2 = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})^2 = \frac{Q_2}{m(n-1)},$$

$$s^2 = \frac{1}{mn-1}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2.$$

Произведем оценку различия между дисперсиями по F-критерию:

$$F = \frac{Q_1 / (m-1)}{Q_2 / m(n-1)}.$$

Если полученное значение критерия больше табличного с заданным уровнем значимости α и числом степенями свободы (m-1; m(n-1)), то нулевую гипотезу отвергаем, т.е. фактор влияет на исследуемую величину.

Таблица однофакторного дисперсионного анализа

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Средний квадрат	Оценка диспер-сий
Межгрупповая	$m\sum_i (\bar{x}_{i*} - \bar{x})^2$	$m-1$	$\frac{1}{m-1}\sum_i (\bar{x}_{i*} - \bar{x})^2$	s_1^2
Внутригруп-повая	$\sum_{ij} (x_{ij} - \bar{x}_{i*})^2$	$m(n-1)$	$\frac{1}{m(n-1)}\sum_{ij} (x_{ij} - \bar{x}_{i*})^2$	s_2^2
Полная (общая)	$\sum_{ij} (x_{ij} - \bar{x})^2$	$mn-1$	$\frac{1}{mn-1}\sum_{ij} (x_{ij} - \bar{x})^2$	s^2

Далее необходимо рассчитать **доли влияния учтенного и неучтенного факторов** как отношения соответствующих сумм квадратов отклонений:

$$\eta_1^2 = Q_1 / Q; \eta_2^2 = Q_2 / Q; \eta_1^2 + \eta_2^2 = 1(100\%),$$

где η_1^2 - доля влияния учтенных факторов;

η_2^2 - доля влияния неучтенных факторов.

Пример. В трех различных местах обитания были собраны жуки-скакуны. У каждого жука измерялась ширина головки. Требуется с помощью дисперсионного анализа выяснить, влияет ли место обитания на ширину головки жука (данные приведены в таблице).

Вначале подсчитаем средние значения:

$$\bar{x}_{1*} = 3{,}76; \bar{x}_{2*} = 3{,}6137; \bar{x}_{3*} = 3{,}5635; \bar{x} = 3{,}6457.$$

m=3, n=10.

Таблица. Значения ширины головки жука в месте обитания

Мест-ность	Измерения ширины головки жука, мм									
	1	2	3	4	5	6	7	8	9	10
A1	3,712	3,732	3,752	3,762	3,769	3,775	3,782	3,787	3,792	3,737
A2	3,602	3,605	3,607	3,61	3,612	3,615	3,617	3,62	3,622	3,627
A3	3,532	3,537	3,542	3,549	3,555	3,562	3,572	3,582	3,592	3,612

Рассчитываем суммы квадратов:

$$Q_1 = n\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x})^2 = 0{,}20845; k_1 = m - 1 = 2;$$

$$Q_2 = \sum_{i=1}^{3}\sum_{j=1}^{10}(x_{ij} - \bar{x}_{i*})^2 = 0{,}01282; k_2 = m(n-1) = 27;$$

$$Q = Q_1 + Q_2 = 0{,}22127; k = mn - 1 = 29.$$

Оценим дисперсии:

$$s_1^2 = Q_1 / k_1 = 0{,}104225; \; s_2^2 = Q_2 / k_2 = 0{,}000475;$$

$$s^2 = Q/k = 0{,}22127/29 = 0{,}00763.$$

Значение критерия Фишера равно

$$F = s_1^2 / s_2^2 = 219{,}421.$$

Табличное значение критерия при уровне значимости $\alpha=0,05$ равно 3,354.

Так как полученное значение критерия Фишера больше табличного (219,421>3,354), то гипотезу о том, что место обитания не влияет на размеры головки жука, отвергаем.

Рассчитаем доли влияния учтенного и неучтенных факторов:

$$\eta_1^2 = Q_1 / Q = 0{,}208453 / 0{,}221274 = 0{,}942;$$

$$\eta_2^2 = Q_2 / Q = 0{,}012821 / 0{,}221274 = 0{,}0579.$$

На долю учтенного фактора – место обитания приходится 94,2% изменчивости, а 5,79% составляют неучтенные факторы.

Таким образом, место обитания оказывает существенное влияет на размеры головки жуков-скакунов.

Многофакторный дисперсионный анализ

Если исследуют действие двух, трех и т.д. факторов, то структура дисперсионного анализа та же, что и при однофакторном анализе, усложняются лишь вычисления. Рассмотрим задачу оценки действия двух одновременно действующих факторов.

Введем некоторые ограничения:

- включаемые в анализ факторы должны быть независимы друг от друга, корреляция между ними не допустима;
- число наблюдений по совокупностям должно быть одинаковым или хотя бы пропорциональным.

Пусть имеется несколько разнотипных участков земли и несколько типов удобрения.

Требуется выяснить, значимо ли влияние качества различных участков земли и качества удобрений на урожайность зерновой культуры.

Пусть фактор А – влияние земли, фактор В – влияние качества удобрения, x_{ij} – урожайность.

Рассмотрим случай, когда для каждого участка земли и для каждого вида удобрения сделано одно наблюдение.

Тогда матрица наблюдений будет следующей:

Вид удобрения(j) Участки земли (i)	B_1	B_2	…	B_v …	
A_1 A_2 … A_r	x_{11} x_{21} … x_{r1}	x_{12} x_{22} … x_{r2}	… … … …	x_{1v} x_{2v} x_{rv}	x_{1*} x_{2*} … x_{r*}
$\overline{x}_{*j}$	x_{*1}	x_{*1}		x_{*v}	$\overline{x}$

По каждому столбцу и строке вычислим среднее значение, а также общее среднее.

$$\bar{x}_{i*} = \frac{1}{v}\sum_{j=1}^{v} x_{ij}, \qquad \bar{x}_{*j} = \frac{1}{r}\sum_{i=1}^{r} x_{ij}, \qquad \bar{\bar{x}} = \frac{1}{rv}\sum_{i=1}^{r}\sum_{j=1}^{v} x_{ij}.$$

Основное тождество однофакторного анализа в данном случае принимает вид:

$$\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{\bar{x}})^2 = v\sum_{i=1}^{r}(\bar{x}_{i*}-\bar{\bar{x}})^2 + r\sum_{j=1}^{v}(\bar{x}_{*j}-\bar{\bar{x}})^2 + \sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+\bar{\bar{x}})^2,$$

$$Q=Q_1+Q_2+Q_3.$$

Слагаемое Q_1 представляет собой сумму квадратов разностей между средними по строкам и общим средним и характеризует изменение признака по фактору A.

Слагаемое Q_2 представляет собой сумму квадратов разностей между средними по столбцам и общим средним и характеризует изменение признака по фактору B.

Слагаемое Q_3 называется ***остаточной суммой квадратов*** и характеризует влияние неучтенных факторов.

Сумма Q называется ***общей или полной суммой квадратов отклонений*** отдельных наблюдений от общей средней.

Произведем оценку дисперсий:

$$s^2 = \frac{1}{rv-1}\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x})^2 = \frac{Q}{rv-1},$$

$$s_1^2 = \frac{1}{r-1}v\sum_{i=1}^{r}(\bar{x}_{i*}-\bar{x})^2 = \frac{Q_1}{r-1},$$

$$s_2^2 = \frac{1}{v-1}r\sum_{j=1}^{v}(\bar{x}_{*j}-\bar{x})^2 = \frac{Q_2}{v-1},$$

$$s_3^2 = \frac{1}{(r-1)(v-1)}\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+\bar{x})^2 = \frac{Q_3}{(r-1)(v-1)}.$$

В двухфакторном анализе для выяснения значимости влияния факторов A и B на исследуемый признак сравнивают дисперсии по факторам с остаточной дисперсией.

$$F_A = \frac{Q_1/(r-1)}{Q_3/(r-1)(v-1)} = \frac{s_1^2}{s_3^2},$$

$$F_B = \frac{Q_2/(v-1)}{Q_3/(r-1)(v-1)} = \frac{s_2^2}{s_3^2}.$$

Полученные значения F_A и F_B сравнивают с табличными значениями при выбранном уровне значимости α и соответствующих числах степеней свободы.

При $F_A<F_\alpha$ и $F_B<F_\alpha$ нулевые гипотезы о равенстве средних не отвергается, т.е. влияние факторов A и B на исследуемый признак незначимо.

Для расчета доли влияния учтенных факторов A, B и неучтенного фактора воспользуемся формулами:

$$\eta_A^2 = {s_1^2}/{s^2}; \quad \eta_B^2 = {s_2^2}/{s^2}; \quad \eta^2 = {s_3^2}/{s^2}.$$

Результаты анализа заносятся в таблицу дисперсионного анализа.

Таблица двухфакторного дисперсионного анализа

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсий
Между средними по строкам	$Q_1 = v\sum_{i=1}^{r}(\bar{x}_{i*} - \bar{x})^2$	$r-1$	s_1^2
Между средними по столбцам	$Q_2 = r\sum_{j=1}^{v}(\bar{x}_{*j} - \bar{x})^2$	$v-1$	s_2^2
Остаточная	$Q_3 = \sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})^2$	$(r-1)(v-1)$	s_3^2
Полная (общая)	$Q = \sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij} - \bar{x})^2$	$rv-1$	s^2

Алгоритм расчетов

1. Построение вспомогательной таблицы.
2. Вычисление средних.
3. Вычисление сумм квадратов.
4. Вычисление оценок дисперсий.
5. Проверка гипотезы $H_{0.}$ Если H_0 не отклоняется, то – проверка значимости уровней факторов.