14. Разложение дисперсии результативного признака в дисперсионном анализе. Дисперсионная таблица. Проверка гипотезы о равенстве групповых математичнских ожиданий (гипотеза, используемая статистика, критерий). Модель однофакторного дисперсионного анализа.

чуточку определений

ОПР(Дисперсионный анализ)

Дисперсионный анализ - статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также планнирования аналогичных экспериментов. По кол-ву факторов можно определить:

- Однофакторный Дисперсионный анализ (только его рассматриваем)
- Многофакторный Дисперсионный анализ

ОПР(Фактор)

Фактор - это не случайная переменная, влияющая на результат

ОПР(Уровень фактора)

Уровень фактора - это конкретное значение фактора

- В качестве значения может быть количественная(кол-во работяг в команде) либо качественная переменная(тип бригады, вид упаковки)

ОПР(Отклик)

Отклик - это значение измеряемого признака(результата)

Модель однофакторного дисперсионного анализа

Модель однофакторного дисперсионного анализа имеет вид:

$$Y_{i,j} = \mu + \tau_j + \varepsilon i, j$$

где:

- $i \in \{1, 2, ...n\}$ - номер наблюдения наблюдений
- $j \in \{1, 2, ...\nu\}$ - номер фактора
- $Y_{i,j}$ - значение отклика для для i-ого наблюдения и j-ого фактора
- μ - общеее среднее отклика
  - $\mu = \bar{Y} = \frac{1}{n} \cdot \sum_{j=1}^{\nu} [\sum_{i=1}^{n} Y_{i,j}] = \frac{1}{\nu} \cdot \sum_{j=1}^{\nu} \frac{1}{n_j} \cdot [\sum_{i=1}^{n_j} Y_{i,j}] = \frac{1}{\nu} \sum_{j=1}^{\nu} \bar{Y}_{\cdot j}$
  - Если по-человечески это просто средние число отклика для всех наблюдений
- τ_j - отклонение от общего среднего, вызванное изменением уровня фактора
- $\varepsilon_{i,j}$ - случайная компонента, вызванная влиянением не рассмотренных факторов
  - Условно мы изучаем продолжительность жизни людей и в качестве фактора выбрали курение, т.е курит он или нет
    - Т.К не только курение влияет на продолжительность жизни, но и много других факторов, то предположу, что $\varepsilon_{i,j}$ будет довольно большой

Определив всё это добро, теперь приступаем к сути - к доказанию гиппотезы:

- H_0: $\tau_1 = \tau_2 = ... = \tau_\nu = 0$
  - т.е факторы не оказывают никакого влияния, средние значения во всех группах одинаковые и различия данных вызваны лишь случайностью
- H_1: $\exists \tau_j \neq 0$
  - Найдётся какой-то фактор, оказывающий влияние на данные

Перед проведением анализа нужно убедиться:

- Все наблюдения - независимы
- Ошибки подчиняются нормальному закону
  - $\varepsilon_{i,j} \sim N(0, \sigma^2)$
- Дисперсия для разных уровней фактора постоянна
  - $\sigma_1^2 = \sigma_2^2 = ... = \sigma_\nu^2 = const$

Без этого анализ не даст точных результатов

С дисперсиями разбираемся с помощью критерия Бартлета(это не мой билет, поэтому просто скрины)

Критерий Бартлетта

С помощью критерия Бартлетта проверяется гипотеза:

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_\nu^2 = \sigma_0^2; \\ H_1: &\quad \exists \sigma_j^2 \neq \sigma_0^2. \end{aligned} \tag{7.3}$$

Эта гипотеза проверяется следующим образом:

1. Находим несмещенные оценки s_i^2 групповых дисперсий σ_i^2 по формуле

$$s_j^2 = \frac{n_j}{n_j - 1}\,\widehat{\sigma}_j^2,$$

где $\widehat{\sigma}_j^2$ — выборочные групповые дисперсии, n_j — численность наблюдений в группах.

2. Находим значение величины s_0^2 - так называемой обобщенной дисперсии:

$$s_0^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_\nu - 1)s_\nu^2}{(n_1 - 1) + (n_2 - 1)\ldots + (n_\nu - 1)}$$

3. Определяем вспомогательный параметр q:

$$q = \left[1 + \frac{1}{3(\nu - 1)}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} + \ldots + \frac{1}{n_\nu - 1} - \frac{1}{(n_1 - 1) + \ldots + (n_\nu - 1)}\right)\right]^{-1}.$$

4. Вычисляем значение критерия Бартлетта:

$$K_B = q\left[(n_1 - 1)\ln\frac{s_0^2}{s_1^2} + (n_2 - 1)\ln\frac{s_0^2}{s_2^2} + \ldots + (n_\nu - 1)\ln\frac{s_0^2}{s_\nu^2}\right]$$

при выполнении условия $n_j > 3 (j = 1, 2, \ldots, \nu)$ и гипотезы (7.3) K_B будет иметь распределение, близкое к χ^2-распределению с $k = \nu - 1$ степенями свободы.

5. Задавшись уровнем значимости α, находим правостороннюю критическую точку $\chi^2_{\text{кр, пр}}$ по следующей схеме:

$$\left.\begin{array}{rcl} \alpha & \to & \gamma = 1 - \alpha \\ \nu & \to & k = \nu - 1 \end{array}\right\} \to \chi^2_\gamma \to \chi^2_{\text{кр, пр}} = \chi^2_\gamma.$$

6. Если K_B попадает в интервал $(\chi^2_{\text{кр, пр}}, +\infty)$, то гипотезу H_0 (7.3) отвергаем. В противном случае считаем, что гипотеза (7.3) не противоречит результатам наблюдений.

Основное дисперсионное равенство (Разложение дисперсии результативного признака в дисперсионном анализе)

Понадобится, чтобы доказать, что средния значения в группах с разными уровнями фактора одинаковы

Для этого:

1. Найдем
   i. $$S_Y^2 = \sum_{j=1}^{\nu} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y})^2.$$
   ii. $\bar{Y}$ - это среднее среди всех значений
2. затем внутри суммы прибавим 0 ($+\bar{Y}_{\cdot j} - \bar{Y}_{\cdot j}$)
   i. $$S_Y^2 = \sum_{j=1}^{\nu} \sum_{i=1}^{n_j} [(Y_{ij} - \overline{Y}_{\cdot j}) + (\overline{Y}_{\cdot j} - \overline{Y})]^2 =$$
   $$= \sum_{j=1}^{\nu} \sum_{i=1}^{n_j} [(Y_{ij} - \overline{Y}_{\cdot j})^2 + 2(Y_{ij} - \overline{Y}_{\cdot j})(\overline{Y}_{\cdot j} - \overline{Y}) + (\overline{Y}_{\cdot j} - \overline{Y})^2]$$
3. Оказывается, что слагаемое $2(Y_{i,j} - \bar{Y}_{\cdot j})(\bar{Y}_{\cdot j} - \bar{Y}) = 0$
   i. $$\sum_{j=1}^{\nu} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_{\cdot j})(\overline{Y}_{\cdot j} - \overline{Y}) = \sum_{j=1}^{\nu} (\overline{Y}_{\cdot j} - \overline{Y}) \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_{\cdot j}) =$$
   $$= \sum_{j=1}^{\nu} (\overline{Y}_{\cdot j} - \overline{Y}) \left\{ \sum_{i=1}^{n_j} Y_{ij} - \sum_{i=1}^{n_j} \overline{Y}_{\cdot j} \right\} = \sum_{j=1}^{\nu} (\overline{Y}_{\cdot j} - \overline{Y}) \left\{ \overline{Y}_{\cdot j} \cdot n_j - n_j \cdot \overline{Y}_{\cdot j} \right\} = 0.$$

4. Получаем

i. $$S_Y^2 = \sum_{j=1}^{\nu}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y})^2 = \sum_{j=1}^{\nu}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\cdot j})^2 + \sum_{j=1}^{\nu}\sum_{i=1}^{n_j}(\overline{Y}_{\cdot j} - \overline{Y})^2.$$

Слагамое

$$S_O^2 = \sum_{j=1}^{\nu}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\cdot j})^2,$$

- характеризует вариацию признака внутри группы

Слагаемое

$$S_\Phi^2 = \sum_{j=1}^{\nu}\sum_{i=1}^{n_j}(\overline{Y}_{\cdot j} - \overline{Y})^2,$$

- характеризует вариацию признака между группами

Таким образом получаем основное дисперсионное р-во:

$$S_Y^2 = S_O^2 + S_\Phi^2$$

Про Таблицу однофакторного дисперсионного анализа

Таблица однофакторного дисперсионного анализа

Источник вариации результативного признака Y	Сумма квадратов	Число степеней свободы	Средний квадрат
Фактор A	$S^2_{\Phi} = \sum_{j=1}^{\nu} (\overline{Y}_{\cdot j} - \overline{Y})^2 n_j$	$\nu - 1$	$s^2_{\Phi} = \frac{S^2_{\Phi}}{(\nu - 1)}$
Остаточные факторы	$S^2_O = \sum_{j=1}^{\nu} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_{\cdot j})^2$	$n - \nu$	$s^2_O = \frac{S^2_O}{(n - \nu)}$
Общая вариация	$S^2_Y = \sum_{j=1}^{\nu} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y})^2$	$n - 1$	$s^2_Y = \frac{S^2_Y}{(n - 1)}$

В Дисперсионном анализе используются не сами суммы квадратов, а средние квадрантов, которые являются несмещенными оценками соответствующих дисперсий(вариации дисперсий на одну степень свободы)

Можно показать, что $s^2_{\text{ф}}$ и s^2_o - независимы, а их отношение

$$F = \frac{\frac{S^2_{\Phi}}{\nu - 1}}{\frac{S^2_O}{n - \nu}} = \frac{s^2_{\text{ф}}}{s^2_o} \sim F(k_1, k_2)$$

где числа степеней свободы:

- $k_1 = \nu - 1$
- $k_2 = n - \nu$

чем больше $s^2_{\text{ф}}$ тем больше числитель, отсюда можно сделать вывод, что гиппотезе H1 будут соответствовать больше значения F (правосторонняя критическая область)

Далее всё как по стандарту:

1. определяем уровень значимости α
2. ищем критическую точку $F_{\text{кр}} = F(\alpha, k_1, k_2)$
3. ищем критическую точку $F_{\text{выб}}$ по выборке

4. сравниваем $F_{\text{кр}}$ и $F_{\text{выб}}$:

- Если $F_{\text{кр}} > F_{\text{выб}}$, то принимаем гиппотезу H_0
- Если $F_{\text{кр}} < F_{\text{выб}}$, то принимаем гиппотезу H_1

Если приняли H_1, то разбираемся почему фактор имеет значение, а также для каждого фактора оцениваем велечину влияния на результат. Для этого необходимо рассчитать

$$\tau_j = \bar{Y}_{.j} - \bar{Y}$$

для всех факторов $j \in \{1, 2, ..., \nu\}$

Также можно проверить гиппотезу о влиянии 2ух уровней фактора на отклик, мб они окажутся одинаковы

ОПР(Коэфициент детерминации)

$$R^2 = \frac{S^2_{\Phi}}{S^2_O} = 1 - \frac{S^2_O}{S^2_Y}$$

R^2 показывает какая доля общей дисперсии Y объясняется зависимостью Y от фактора. Чем ближе R^2 к одному тем сильнее фактор влияет на отклик

общий алгоритм Дисперсионного анализа

План, опишем алгоритм проведения дисперсионного анализа:

1. Проверяем (если это неизвестно), что данные подчиняются нормальному распределению.

2. С помощью критерия Бартлетта проверяем гипотезу (7.3) о равенстве дисперсий в группах (при разных уровнях фактора).

3. Составляем таблицу дисперсионного анализа (табл. 7.2).

4. Вычисляем выборочное значение статистики критерия по формуле (7.5).

5. Находим критическое значение статистики Фишера (Приложение 4) и делаем вывод о значимости влияния фактора.

6. Если влияние фактора подтверждается, то определяем его влияние на каждом уровне по формуле (7.6), а также оцениваем степень влияния фактора в целом на результат с помощью коэффициента детерминации (7.7).

7. Записываем модель однофакторного дисперсионного анализа:

$$Y_{\cdot j} = \mu + \tau_j + N(0; s_O^2), \quad j = 1, 2, \ldots, \nu.$$

8. Даем интерпретацию полученным результатам исследования.