

Дисперсионный анализ неполных данных на основе блок-схем с приложениями в медицине

Положиев Роман, группа 18.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к.ф.-м.н., доцент Алексеева Н.П.

Дисперсионный анализ применяется для исследования влияния качественных переменных (факторов) на одну зависимую количественную переменную.

Полный факторный эксперимент — эксперимент, в котором реализуются все возможные сочетания уровней факторов, что трудоёмко либо невозможно.

Используемые данные

Больные COVID-19. Всего 242 пациента. Зависимая переменная — процент поражения легких. Факторы:

- Возраст (a уровней)
- Сатурация (b уровней)

Присутствует проблема трудоёмкости и/или невозможности проведения всех $a \times b$ групп испытаний.

В качестве решения этой проблемы предлагается использование *дизайнов*.

Дизайн $D(v, b, r, k, \lambda)$ — размещение v элементов (методов обработки) по b блокам размера k , что каждый элемент встречается r раз, а каждая пара λ раз.

Симметричный дизайн $D(v, k, \lambda)$ — случай $v = b, r = k$.

Пример

Фактор возраст имеет 6 уровней b_j , а фактор сатурация 4 уровня v_i .

Возраст соответствует блоку, а сатурация методу обработки.

Данные в соответствии с дизайном $D(4, 6, 3, 2, 1)$ отмечены синим:

$v_i b_j$	< 54	54–60	61–67	68–74	75–82	≥ 82
< 88	12	8	10	10	0	0
88–91	0	6	9	6	7	0
92–94	9	11	15	11	9	7
95–100	20	11	8	13	11	12

- Методы построения блок-схем;
- Дисперсионный анализ на блок-схемах с одним наблюдением в ячейке;
- Дисперсионный анализ для одинакового количества наблюдений в каждой ячейке.

Методы построения блок-схем

В рамках работы были рассмотрены три метода построения блок-схем, изложенных в (Дюге, 1972):

- С помощью построения проективной геометрии над полем Галуа;
- С помощью евклидовой геометрии над полем Галуа;
- С помощью абелевой группы.

Были построены следующие блок-схемы:

- $D(7, 3, 1)$ с помощью $PG(2, 2)$;
- $D(4, 6, 3, 2, 1)$ с помощью $EG(2, 2)$;
- $D(9, 12, 4, 3, 1)$ с помощью G_3 .

Известные результаты. Модель дисперсионного анализа

Модель двухфакторного дисперсионного анализа на блок-схемах с одним наблюдением в ячейке (Дюге, 1972):

$$x_{ij} = \mu + v_i + b_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma).$$

- μ — генеральное среднее,
- v_i — дифференциальный эффект фактора v ,
- b_j — дифференциальный эффект фактора b ,
- ε_{ij} — независимые случайные ошибки.

Известные результаты. Суммы квадратов

В (Дюге, 1972) было показано:

Источник вариации	Сумма квадратов	Степени свободы
Фактор v	$SS_v = \frac{k}{\lambda v} \sum_{i=1}^v \left(V_i - \frac{1}{k} T_i \right)^2$	$df_v = v - 1$
Фактор b	$SS_b = k \sum_{j=1}^b \left(\frac{B_j}{k} - \hat{\mu}^2 \right)^2$	$df_b = b - 1$
Остаток (ошибка)	$SS_E = \sum_{ij} \left(x_{ij} - \hat{v}_i + \frac{1}{k} \sum_i \eta_{ij} \hat{v}_i - \frac{B_j}{k} \right)^2$	$df_E = bk - bv + 1$
Общая	$SS_T = \sum_{ij} (x_{ij} - \hat{\mu})^2$	$df_T = bk - 1$

- $\hat{\mu} = \frac{\sum_{ij} x_{ij}}{bk},$
- $\hat{v}_i = \frac{kV_i - T_i}{\lambda v},$
- $\eta_{ij} = \begin{cases} 1, & \text{если } x_{ij} \text{ существует} \\ 0, & \text{иначе} \end{cases}$

- $V_i = \sum_{j=1}^b x_{ij},$
- $B_j = \sum_{i=1}^v x_{ij},$
- $T_i = \sum_{j=1}^b \eta_{ij} B_j.$

Нулевая гипотеза:

- Нет эффекта фактора ν , то есть $H_0: \nu_i = 0$,
- Нет эффекта фактора b , то есть $H_0: b_j = 0$.

Статистические критерии:

$$F = \frac{SS_\nu/df_\nu}{SS_E/df_E} \sim F(df_\nu, df_E),$$

$$F = \frac{SS_b/df_b}{SS_E/df_E} \sim F(df_b, df_E).$$

Рассмотрим данные больных COVID-19.

Из переменной возраста и сатурации сделаем факторы с шестью и четырьмя уровнями соответственно.

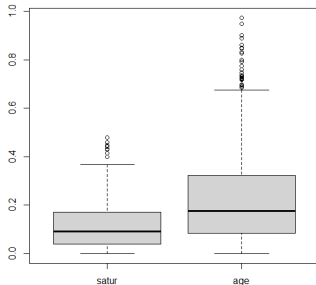
Для каждой из получившихся 24 групп посчитаем среднее значение.

$v_i b_j$	1	2	3	4	5	6
1	66.00	62.00	67.46	73.85	69.38	53.67
2	58.33	55.64	67.22	51.67	51.18	45.56
3	26.38	32.50	42.00	49.75	50.60	34.78
4	34.75	33.56	51.50	40.00	61.33	44.90

Применим дизайн $D(4, 6, 3, 2, 1)$. Заметим, что можно использовать различные подстановки данных.

Применим дисперсионный анализ на дизайне $D(4, 6, 3, 2, 1)$, используя всевозможные подстановки данных.

Распределение p – *value* по фактору сатурации и возраста:



Вывод: сложно судить о наличии влияния факторов.

Полученные результаты. Модель

Модель дисперсионного анализа с T наблюдениями в каждой ячейке:

$$x_{ijt} = \mu + v_i + b_j + \varepsilon_{ijt}, \quad \varepsilon_{ijt} \sim N(0, \sigma),$$

где x_{ijt} — значение переменной x полученной при t -ом повторении эксперимента.

Оценим параметры модели по методу наименьших квадратов:

$$L = \sum_{ijt} (x_{ijt} - \mu - v_i - b_j)^2 \rightarrow \min.$$

Требую, чтобы:

$$\sum_i \hat{v}_i = 0, \quad \sum_j \hat{b}_j = 0.$$

Полученные результаты. Оценка параметров

Находя

$$\frac{\partial L}{\partial \mu} = 0, \quad \frac{\partial L}{\partial v_i} = 0, \quad \frac{\partial L}{\partial b_j} = 0.$$

Получаем оценку параметров:

$$\hat{\mu} = \frac{\sum_{ijt} x_{ijt}}{bkT}, \quad \hat{v}_i = \frac{kV_i - T_i}{\lambda v}, \quad \hat{b}_j = \frac{1}{k}(B_j - \sum_{(j)} \hat{v}_e) - \hat{\mu}.$$

$\sum_{(i)} \hat{b}_l$ — сумма \hat{b}_l по блокам содержащих i -ый метод обработки,

$\sum_{(j)} \hat{v}_e$ — сумма \hat{v}_e по всем методам, встречающимся в j -ом блоке,

$$V_i = \sum_{(i)} \overline{x_{ij*}}, \quad B_j = \sum_{(j)} \overline{x_{ij*}}, \quad T_i = \sum_{(i)} B_j.$$

Утверждение

$$\sum_{ijt} (x_{ijt} - \hat{\mu})^2 = S_e^2 + S_v^2 + S_b^2,$$

где

$$S_e^2 = \sum_{ijt} \left(x_{ijt} - (\hat{v}_i - \frac{1}{k} \sum_{(j)} \hat{v}_e) - \frac{B_j}{k} \right)^2,$$

$$df_e = bkT - v - b + 1,$$

$$S_v^2 = \frac{\lambda v T}{k} \sum_i \hat{v}_i^2,$$

$$df_v = v - 1,$$

$$S_b^2 = kT \sum_j \left(\frac{B_j}{k} - \hat{\mu} \right)^2,$$

$$df_b = b - 1.$$

Статистические критерии:

$$F = \frac{S_v/df_v}{S_e/df_e} \sim F(df_v, df_e),$$

$$F = \frac{S_b/df_b}{S_e/df_e} \sim F(df_b, df_e).$$

Новые результаты. Применение

Рассмотрим данных о лечении больных от COVID-19. Возьмем следующие уровни факторов.

$v_i b_j$	< 54	54–60	61–67	68–74	75–82	≥ 82
< 88	12	8	10	10	3	4
88–91	4	6	9	6	7	5
92–94	9	11	15	11	9	7
95–100	20	11	8	13	11	12

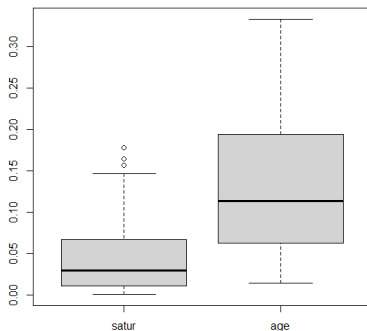
Удалим наблюдения выделенные красным цветом.

Случайным образом выкинем наблюдения из ячейки, если их больше 6.

Проведем дисперсионный анализ на блок-схеме $D(4, 6, 3, 2, 1)$ с 6 наблюдениями в каждой ячейке по всевозможным подстановкам данных.

Новые результаты. Применение

Проведем такую процедуру $N = 100$ раз, беря каждый раз среднее значение $p - value$. Получившееся распределение средних $p - value$ по каждому фактору:



Вывод: снизился разброс значений.

Результаты работы:

- Изучены методы построения блок-схем и построены блок-схемы тремя методами;
- Изучен и применен дисперсионный анализ на блок-схемах с одним наблюдением в ячейке;
- Построена и применена модель дисперсионного анализа для одинакового количества наблюдений в каждой ячейке;
- Реализовано применение дисперсионного анализа на блок-схемах в языке программирования R.