Блок-схемы и их применение в анализе неполных данных

Подлеснов Яков Сергеевич, гр. 20.Б04-мм

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доцент Алексеева Н.П.
Рецензент: к.т.н., научный сотрудник Белякова Л. А.

Санкт-Петербург
2024г.

Введение: Постановка задачи

Дисперсионный анализ — метод, позволяющий выявить влияние факторов на зависимую переменную.
Модель дисперсионного анализа имеет вид [Дюге, 1972]:

$$x_{ij} = \mu + v_i + b_j + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),\ i = 1, 2, \dots, v,\ j = 1, 2, \dots, b$$

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

Задача: реализовать алгоритм дисперсионного анализа с помощью блок-схем при неполных данных.

Блок-схема (дизайн) $D(v, b, r, k, \lambda)$ — размещение v элементов по b блокам размера k, что каждый элемент встречается r раз, а каждая пара λ раз.

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

Ниже приведен важный пример симметричной блок-схемы **D(7, 3, 1)**:

$B_1 : 2, 4, 6;$	$B_4 : 1, 2, 3;$	
$B_2 : 1, 4, 5;$	$B_5 : 2, 5, 7;$	$B_7 : 3, 5, 6;$
$B_3 : 3, 4, 7;$	$B_6 : 1, 6, 7;$	

Двойственный дизайн

Если пронумеровать b блоков дизайна и собрать из этих номеров блок-схему таким образом, чтобы в один блок входили номера блоков, содержащие один из v элементов, то мы получим **"двойственный"** дизайн $D^*(v, b, r, k, \lambda)$ состоящий из v блоков, содержащих r элементов.

Пример построения:

Было:

$D(4, 6, 3, 2, 1)$

$B_1 : 1, 3;$

$B_2 : 1, 2;$

$B_3 : 1, 4;$

$B_4 : 3, 4;$

$B_5 : 2, 4;$

$B_6 : 2, 3;$

Стало:

$D^*(4, 6, 3, 2, 1)$

$\Gamma_1 : 1, 2, 3;$

$\Gamma_2 : 1, 5, 6;$

$\Gamma_3 : 2, 4, 6;$

$\Gamma_4 : 3, 4, 5;$

Построение блок-схем: матрица Адамара

Матрица Адамара H — матрица порядка m, элементами которой являются $+1$ и -1, такая, что $HH^T = mE_m$.
Если у H первая строка и столбец состоят из $+1$, то она **нормализованная**.
Теорема [Холл, 1970]
Из H порядка $m = 4t$ можно построить симметричную блок-схему $D(v, k, \lambda)$:

$$v = 4t - 1,\ \ k = 2t - 1,\ \ \lambda = t - 1.$$

[**Конструкция Сильвестра**] Пусть H — нормализованная матрица Адамара порядка n. Тогда разделенная матрица

$$H_{2^k} = \begin{bmatrix} H_{2^{k-1}} & H_{2^{k-1}} \\ H_{2^{k-1}} & -H_{2^{k-1}} \end{bmatrix}, H_1 = [1], k \in \{1, 2, \dots\}$$

Построение $D(7,3,1)$, используя матрицы Адамара

- Получение из нормализованной матрицы Адамара H_8 симметричного дизайна $D(7,3,1)$.

$$H_8 = \begin{array}{cc} \begin{array}{c:ccc|cccc} & B_1 & B_2 & B_3 & B_4 & B_5 & B_6 & B_7 \end{array} & \\ \left[\begin{array}{c:ccc|cccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hdashline 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{array}\right] & \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \end{array} \end{array}$$

Например, $B_1 = (2,4,6)$, $B_2 = (1,4,5)$, $B_3 = (3,4,7)$.

Обобщение конструкции Сильвестра

Образующая матрица A для $q = 3$, $n = 1$, результирующая матрица K.

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{bmatrix} \Rightarrow K = \begin{array}{cc} \begin{array}{ccccccccc} & B_1 & B_2 & B_3 & B_4 & B_5 & B_6 & B_7 & B_8 \end{array} & \\ \left[\begin{array}{c:cccc|cccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 2 & 1 & 0 & 2 & 1 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 1 & 2 & 1 & 2 & 0 & 2 & 0 & 1 \\ \hline 0 & 2 & 1 & 1 & 0 & 2 & 2 & 1 & 0 \\ 0 & 0 & 0 & 2 & 2 & 2 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 0 & 1 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 & 1 & 0 & 1 & 0 & 2 \end{array}\right] & \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{array} \end{array}$$

$B_1 = (3, 6)$, проецируя $B_1 = ((10)^T, (20)^T)$, элементы равны с точностью домножения на 2, оставляем только первый $B_1 = (10)^T$.

Построение $D(13,4,1)$, $D(21,4,1)$, $D(31,6,1)$

- Получение $D(13,4,1)$ из матрицы Адамара над полем F_3 и $n = 2$,получение $D(31,6,1)$ из матрицы Адамара над полем F_5 и $n = 2$.
- Получение $D(21,4,1)$ из матрицы Адамара над полем F_4 и $n = 2$.

Образующая матрица A для F_4 и F_5 соответственно.

$$A_{F_4} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 \\ 0 & 2 & 3 & 1 \\ 0 & 3 & 1 & 2 \end{bmatrix} \qquad A_{F_5} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 2 & 4 & 1 & 3 \\ 0 & 3 & 1 & 4 & 2 \\ 0 & 4 & 3 & 2 & 1 \end{bmatrix}$$

Исследуемые данные

Для проведения дисперсионного анализа были взяты данные о крысах с различной площадью ожога, измеряемой в течение 36 дней. Объем выборки $n = 25$.

В качестве факторов были взяты:

- Различные виды лечебных препаратов.
- Исходная масса крысы в момент ожога.

Таблица: Дизайн $D(4, 6, 3, 2, 1)$

	≤245	246-247	248	249-251	252-254	≥257
Хитозан йод.	13.4	13.4	—	—	11.1	—
Хитозан кл.	—	—	—	13.9	12.2	15.2
Травотан	8.1	—	7.4	—	—	8.3
Левомеколь	—	7	10.6	9.4	—	—
Блоки	13	14	34	24	12	23

Известные результаты, основные статистики и оценки МНК

Пусть β_i - блоки прямого дизайна, а γ_i - блоки двойственного дизайна, тогда [Дюге, 1972]:

- $V_i = \sum\limits_{j \in \gamma_i} x_{ij}$, $B_j = \sum\limits_{i \in \beta_j} x_{ij}$,
- $T_l = \sum\limits_{j \in \gamma_l} B_j$, $\;j = 1, 2, \ldots, b,\; i, l = 1, 2, \ldots, v.$

С помощью МНК можем получить оценки модели [Дюге, 1972]:

- $\hat{v}_l = \dfrac{kV_l - T_l}{\lambda v}$, $l = 1, 2, \ldots, v,$
- $\hat{\mu} = \dfrac{1}{bk} \sum\limits_{i=1}^{v} \sum\limits_{j \in \gamma_i} x_{ij}.$

Проверка гипотез

- $H_0 : v_i = 0$, то есть нет эффекта фактора v.
- $H_0 : b_j = 0$, то есть нет эффекта фактора b.

Для проверки значимости эффектов используются статистики [Дюге, 1972]:

$F_v = \dfrac{S_v^2/df_v}{S_e^2/df_e} \sim \mathcal{F}(df_v, df_e)$ и $F_b = \dfrac{S_b^2/df_b}{S_e^2/df_e} \sim \mathcal{F}(df_b, df_e)$, где:

- $S_v^2 = \dfrac{\lambda v}{k} \sum\limits_{i=1}^{v} \hat{v_i}^2$, $df_v = v - 1$,
- $S_b^2 = \sum\limits_{j=1}^{b} k \left(\dfrac{B_j}{k} - \hat{\mu} \right)^2$, $df_b = b - 1$
- $S_e^2 = \sum\limits_{(i,j)} \left(x_{ij} - \hat{v}_i + \dfrac{1}{k} \sum\limits_{l \in \beta_j} \hat{v}_l - \dfrac{B_j}{k} \right)^2$, $df_e = bk - v - b + 1$

Подсчет p-value, перестановка индивидов и блоков

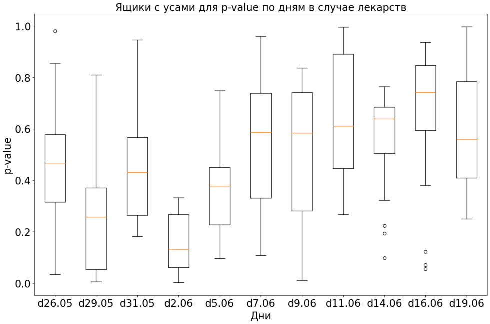
Для формирования дизайна можно использовать разных индивидов в одном и том же блоке или переставить блоки местами, тогда получим разные p-value. Красным - изначальный вариант, синим - новый вариант.

Таблица: Дизайн $D(4,6,3,2,1)$

	≤245	246-247	248	249-251	252-254	≥257
Хит.йод.	13.4/9.8	13.4/—	—	—/13.9	11.1	—
Хит.кл.	—	—/13.4	—	13.9/–	12.2	15.2
Трав.	8.1	—	7.4	—	—	8.3
Лев.	—	7	10.6	9.4	—	—
Блоки	13	14/24	34	24/14	12	23

Значимость лекарств в случае исходной массы

Выделена дата с наибольшим отличием по виду лечения.



Через 10 дней наблюдается лучшее заживление при использовании антибиотиков или при комплексном лечении Хитозаном с йодом, фибробластами и Травотаном по сравнению с аналогичным лечением, но без усиливающего регенерацию Травотана.

Значимость влияния исходной массы

- Ни один из дней не показал, что исходная масса может быть значимой.

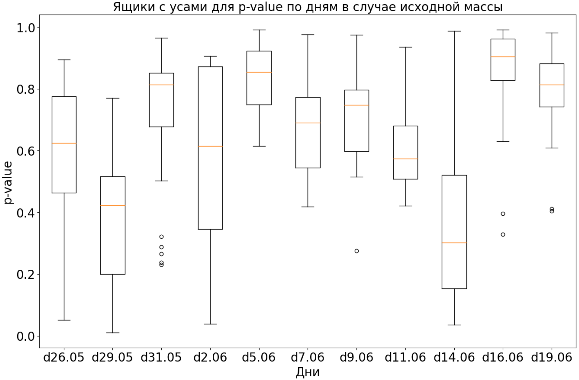


Рис.: p-value для исходной массы

Значимость лекарств в случае фактора текущей массы

- Удалось произвести подсчеты p-value для 5 дней.
- 2 июня вновь оказалось днем, когда лекарства оказали наибольшее воздействие.

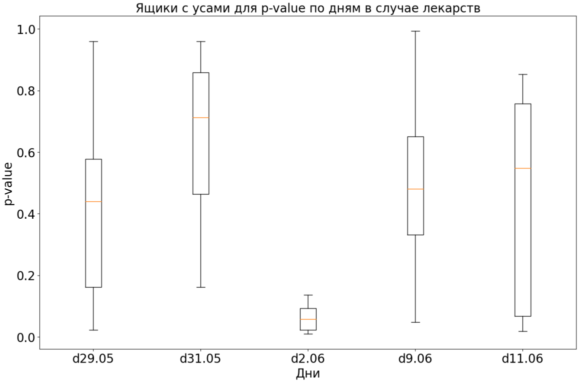
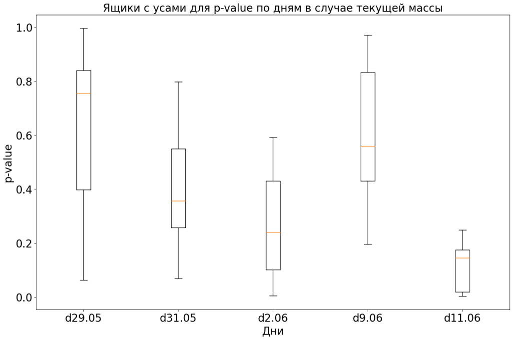


Рис.: p-value для лекарств в случае текущей массы

Значимость текущей массы

- 11.06 текущая масса оказалась значимой.
- При меньшей массе средняя площадь ожога была больше.



Реакция организма на ожог проявилась через 19 дней.

Ковариационный анализ

- В качестве зависимой переменной была взята площадь ожога на 2 июня.
- В качестве независимых переменных были взяты все массы крыс с 24 мая до 2 июня включительно.
- Массы 31 мая и 2 июня оказались значимыми при $\alpha = 0.2$.
- На остатки регрессии был сделан многофакторный (фактор - лекарство) дисперсионный анализ.

Таблица: Значимость лекарств

Препарат	p-value
Хитозан с йодом	0.0013
Хитозан с клетками	0.072
Травотан	0.009
Левомеколь	0.002

Выводы

1. Построены дизайны $D(7,3,1)$, $D(13,4,1)$, $D(21,4,1)$, $D(31,6,1)$, обобщена конструкция Сильвестра с поля характеристики 2 на поля характеристики 3, 4, 5.
2. Реализован алгоритм дисперсионного анализа с помощью блок-схем , который позволяет изучить значимость влияния факторов на зависимую переменную при относительно небольшом объеме данных.
3. Метод применен для сравнения разных методик лечения ожогов у крыс с учетом динамики фактора организма (массы тела).
4. Произведено сравнение полученных результатов со стандартным методом.

Технический слайд: Теорема Зингера

Проективная геометрия P_n^q — пространство векторов $(a_0, a_1, ..., a_n)$ размерности n, где $a_i \in F_q$.

Теорема Зингера (Алексеева Н.П., 2012)
Гиперплоскости P_n^q, $q = p^r$, как блоки, и точки, как элементы, образуют

$D(v, k, \lambda)$:

$$v = \frac{q^{n+1} - 1}{q - 1},\ k = \frac{q^n - 1}{q - 1},\ \lambda = \frac{q^{n-1} - 1}{q - 1}.$$

Технический слайд: моделирование

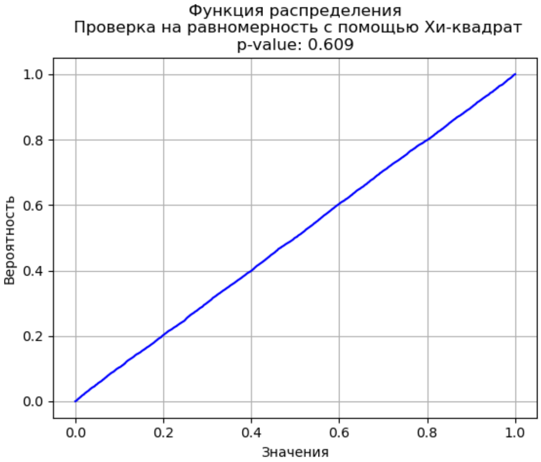
Для корректности статистического теста были смоделированы при условии нулевой гипотезы данные при следующих параметрах:

- $b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = 0$,
- $v_1 = -3.5$, $v_2 = 9.3$, $v_3 = -10.8$, $v_4 = 5$,
- $\mu = 13.5$, $\sigma = 19$.

Равномерность p-value проверяется в случае предложенного метода и классического двухфакторного дисперсионного анализа.

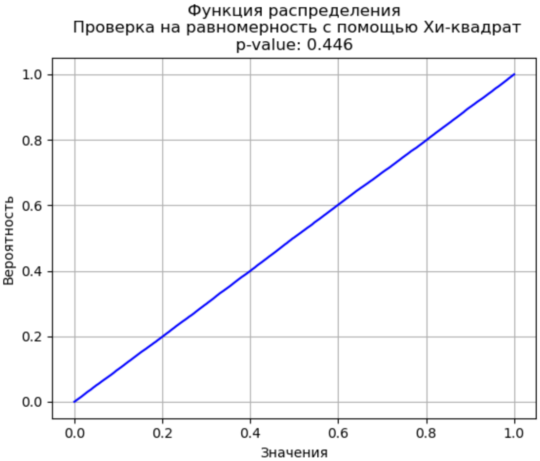
Технический слайд: равномерность p-value 1

Ниже представлена функция распределения p-value, полученная с помощью классического двухфакторного дисперсионного анализа.



Технический слайд: равномерность p-value 2

Ниже представлена функция распределения p-value, полученная с помощью двухфакторного дисперсионного анализа, используя блок-схемы.



Технический слайд: Ковариационный анализ

Модель ковариационного анализа имеет вид, если анализируются n наблюдений $Y_1, \dots, Y_n$ с p сопутствующими переменными $(X = (x^{(1)}, \dots, x^{(p)}))$, k возможными типами условий эксперимента $(F = (f_1, \dots, f_k))$:

$$Y_i = \sum_{j=1}^{k} f_{ij}\theta_j + \sum_{j=1}^{p} \beta_j x_i^{(j)} + \varepsilon_{ij}, \ \ i \in \{1, \dots n\}$$

- f_{ij} — индикаторные переменные f_{ij} равны 1, если j-ое условие эксперимента имело место при наблюдении Y_i, и равны 0 в противном случае,
- θ_j — коэффициенты определяют эффект влияния j-го условия,
- $x_i^{(j)}$ — значение сопутствующей переменной $x^{(j)}$, при котором получено наблюдение Y_i,
- β_j - коэффициенты регрессии Y по $x^{(j)}$.